Continuous-Time Accelerated Methods via a Hybrid Control Lens

ARMAN SHARIFI KOLARIJANI, PEYMAN MOHAJERIN ESFAHANI, TAMÁS KEVICZKY

Abstract. Treating optimization methods as dynamical systems can be traced back centuries ago in order to comprehend the notions and behaviors of optimization methods. Lately, this mind set has become the driving force to design new optimization methods. Inspired by the recent dynamical system viewpoint of Nesterov's fast method, we propose two classes of fast methods, formulated as hybrid control systems, to obtain pre-specified exponential convergence rate. Alternative to the existing fast methods which are parametric-in-time second order differential equations, we dynamically synthesize feedback controls in a state-dependent manner. Namely, in the first class the damping term is viewed as the control input, while in the second class the amplitude with which the gradient of the objective function impacts the dynamics serves as the controller. The objective function requires to satisfy a certain sharpness criterion, the so-called Polyak–Łojasiewicz inequality. Moreover, we establish that both hybrid structures possess Zeno-free solution trajectories. We finally provide a mechanism to determine the discretization step size to attain an exponential convergence rate.

1. Introduction

There is a renewed surge of interest in gradient-based algorithms in many computational communities such as machine learning and data analysis. The following non-exhaustive list of references indicates typical application areas: clustering analysis [21], neuro-computing [4], statistical estimation [34], support vector machines [1], signal and image processing [3], and networked-constrained optimization [11]. This interest primarily stems from low computational and memory loads of these algorithms (making them exceptionally attractive in large-scale problems where the dimension of decision variables can be enormous). As a result, a deeper understating of how these algorithms function has become a focal point of many studies.

One research direction that has been recently revitalized is the application of ordinary differential equations (ODEs) to the analysis and design of optimization algorithms. Consider an iterative algorithm that can be viewed as a discrete dynamical system, with the scalar s as its step size. As s decreases, one can observe that the iterative algorithm in fact recovers a differential equation, e.g., in the case of gradient descent method applied to an unconstrained optimization problem $\min_{X\in\mathbb{R}^n} f(X)$, one can inspect that

$$X^{k+1} = X^k - s\nabla f(X^k) \;\rightsquigarrow\; \dot{X}(t) = -\nabla f\big(X(t)\big)$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a smooth function, X is the decision variable, $k\in\mathbb{Z}_{\geq 0}$ is the iteration index, and $t\in\mathbb{R}_{\geq 0}$ is the time. The main motivation behind this line of research has to do with well-established analysis tools in dynamical systems described by differential equations.

The slow rate of convergence of the gradient descent algorithm ($\mathcal{O}(\frac{1}{t})$ in continuous and $\mathcal{O}(\frac{1}{k})$ in discrete time), limits its application in large-scale problems. In order to address this shortcoming, many researchers resort to the following class of 2nd-order ODEs, which is also the focus of this study:

$$\ddot{X}(t) + \gamma(t)\dot{X}(t) + \nabla f\big(X(t)\big) = 0. \tag{1}$$

Increasing the order of the system dynamics interestingly helps improve the convergence rate of the corresponding algorithms to $\mathcal{O}(\frac{1}{k^2})$ in the discrete-time domain or to $\mathcal{O}(\frac{1}{t^2})$ in the continuous-time domain. Such methods are called *momentum*, *accelerated*, or *fast* gradient-based iterative algorithms in the literature. The

Date: July 23, 2018.

The authors are with the Delft Center for Systems and Control, TU Delft, The Netherlands ({a.sharifikolarijani,p.mohajerinesfahani,t.keviczky}@tudelft.nl).

time-dependent function $\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ is a *damping* or a *viscosity* term, which has also been referred to as the *asymptotically vanishing viscosity* since $\lim_{t\to\infty} \gamma(t) = 0$ [6].

Chronological developments of fast algorithms: It is believed that the application of (1) to speed-up optimization algorithms is originated from [33] in which Polyak was inspired by a physical point of view (i.e., a heavy-ball moving in a potential field). Later on, Nesterov introduced his celebrated accelerated gradient method in [27] using the notion of "estimate sequences" and guaranteeing convergence rate of $\mathcal{O}(\frac{1}{k^2})$. Despite several extensions of Nesterov's method [28, 29, 30], the approach has not yet been fully understood. In this regard, many have tried to study the intrinsic properties of Nesterov's method such as [9, 5, 8, 22]. Recently, the authors in [35] and in details [36] surprisingly discovered that Nesterov's method recovers (1) in its continuous limit, with the time-varying damping term $\gamma(t) = \frac{3}{t}$.

A dynamical systems perspective: Based on the observation suggested by [35], several novel fast algorithms have been developed. Inspired by the mirror descent approach [26], the ODE (1) has been extended to non-Euclidean settings and to higher order methods using the Bregman Lagrangian in [37]. Following [37], a "rate-matching" Lyapunov function is proposed in [39] with its monotonicity property established for both continuous and discrete dynamics. Recently, the authors in [22] make use of an interesting semidefinite programming framework developed by [8] and use tools from robust control theory to analyze the convergence rate of optimization algorithms. More specifically, the authors exploit the concept of integral quadratic constraints (IQCs) [24] to design iterative algorithms under the strong convexity assumption. Later, the authors in [10] extend the results of IQC-based approaches to quasi-convex functions. The authors in [16] use dissipativity theory [38] along with the IQC-based analysis to construct Lyapunov functions enabling rate analyses.

Restarting schemes: A characteristic feature of fast methods is the non-monotonicity in the suboptimality measure $f - f^*$, where f^* refers to the optimal value of function f. The reason behind such an undesirable behavior can be intuitively explained in two ways: (i) a momentum based argument indicating as the algorithm evolves, the algorithm's momentum gradually increases to a level that it causes an oscillatory behavior [32]; (ii) an acceleration-based argument indicating that the asymptotically vanishing damping term becomes so small that the algorithm's behavior drifts from an over-damped regime into an under-damped regime with an oscillatory behavior [36]. To prevent such an undesirable behavior in fast methods, an optimal fixed restart interval is determined in terms of the so-called condition number of function f such that the momentum term is restarted to a certain value, see e.g., [28, 25, 14, 20, 30]. It is worth mentioning that [32] proposes two heuristic adaptive restart schemes. It is numerically observed that such restart rules practically improve the convergence behavior of a fast algorithm.

Regularity for exponential convergence: Generally speaking, exponential convergence rate and the corresponding regularity requirements of the function f are two crucial metrics in fast methods. In what follows, we discuss about these metrics for three popular fast methods in the literature. When the objective functions are strongly convex with a constant σ_f and their gradient is Lipschitz with a constant L_f, [36] proposes the "speed restarting" scheme

$$\sup\Big\{t > 0 :\ \forall \tau \in (0,t), \frac{d\|\dot{X}(\tau)\|^2}{d\tau} > 0\Big\},$$

to achieve the convergence rate of:

$$f\big(X(t)\big) - f^* \leq d_1 e^{-d_2 t}\|X(0) - X^*\|^2.$$

The positive scalars d_1 and d_2 depend on the constants σ_f and L_f. Assuming the convexity of the function f with a certain choice of parameters in their "ideal scaling" condition, [37] guarantees the convergence rate of $\mathcal{O}(e^{-ct})$ for some positive scalar c. However, in this general case, their approach requires to compute a matrix inversion in the Euler-Lagrange equation in the form of:

$$\ddot{X}(t) + c\dot{X}(t) + c^2 e^{ct}\Big(\nabla^2 h\big(X(t) + \frac{1}{c}\dot{X}(t)\big)\Big)^{-1}\nabla f\big(X(t)\big) = 0,$$

where the function h is a distance generating function. Under uniform convexity assumption with a constant ν_f, it is further shown that

$$f\big(X(t)\big) - f^* \leq \Big(f\big(X(0)\big) - f^*\Big)e^{-\nu_f \frac{1}{p-1}t}.$$

where $p-1$ is the order of smoothness of f. The authors in [39] introduce the Lyapunov function

$$\mathcal{E}(t) = e^{\beta(t)}\left(f\big(X(t)\big) - f^* + \frac{\sigma_f}{2}\|X^* - Z(t)\|^2\right),$$

to guarantee the rate of convergence

$$\mathcal{E}(t) \leq \mathcal{E}(0)e^{-\int \dot{\beta}(s)ds},$$

where $Z(t) = X(t) + \frac{1}{\dot{\beta}(t)}\dot{X}$, $\dot{Z}(t) = -\dot{X}(t) - \frac{1}{\sigma_f}\dot{\beta}(t)\nabla f\big(X(t)\big)$, and $\beta(t)$ is a user-defined function.

Contribution: Much of the references reviewed above primarily deal with constructing a time-dependent damping term $\gamma(t)$ that is sometimes tied to a Lyapunov function. Furthermore, due to underlying oscillatory behavior of the corresponding 2nd-order ODE, researchers utilize restarting schemes to over-write the steady-state non-monotonic regime with the transient monotonic regime of the dynamics. In general, notice that these schemes are based on time-dependent schedulers.

Statement of hypothesis: With the above argument in mind, let us view an algorithm as a unit point mass moving in a potential field caused by an objective function f under a parametric (or possibly constant) viscosity, similar to the second order ODE (1). In this view, we aim to address the following two questions:

Is it possible to

(I) synthesize the damping term γ as a state-dependent term (i.e., $\gamma(X, \dot{X})$), or
(II) dynamically control the magnitude of the potential force $\nabla f(X)$,

such that the underlying properties of the optimization algorithm are improved?

In this paper, we answer these questions by amending the 2nd-order ODE (1) in two ways as follows:

$$\text{(I)}\ \ddot{X}(t) + u_{\mathbf{I}}\big(X(t), \dot{X}(t)\big)\ \dot{X}(t) + \nabla f(X(t)) = 0,$$
$$\text{(II)}\ \ddot{X}(t) + \dot{X}(t) + u_{\mathbf{II}}\big(X(t), \dot{X}(t)\big)\ \nabla f(X(t)) = 0,$$

where the indices indicate to which question each structure is related to in the above hypothesis. Evidently, in the first structure, the state-dependent input $u_{\mathbf{I}}$ replaces the time-dependent damping γ in (1). While in the second structure, the feedback input $u_{\mathbf{II}}$ dynamically controls the magnitude with which the potential force enters the dynamics (we assume for simplicity of exposition that $\gamma(t) = 1$, however, one can modify our proposed framework and following a similar path develop the corresponding results for the case $\gamma(t) \neq 1$). Given a positive scalar α, we seek to achieve an exponential rate of convergence $\mathcal{O}(e^{-\alpha t})$ for an unconstrained, smooth optimization problem in the suboptimality measure $f\big(X(t)\big) - f^*$. To do so, we construct the state-dependent feedback laws for each structure as follows:

$$u_{\mathbf{I}}\big(X(t), \dot{X}(t)\big) := \alpha + \frac{\|\nabla f(X(t))\|^2 - \langle \nabla^2 f\big(X(t)\big)\dot{X}(t), \dot{X}(t)\rangle}{\langle \nabla f\big(X(t)\big), -\dot{X}(t)\rangle},$$

$$u_{\mathbf{II}}\big(X(t), \dot{X}(t)\big) := \frac{\langle \nabla^2 f\big(X(t)\big)\dot{X}(t), \dot{X}(t)\rangle + (1-\alpha)\langle \nabla f\big(X(t)\big), -\dot{X}(t)\rangle}{\|\nabla f(X(t))\|^2}.$$

Motivated by restarting schemes, we further extend the class of dynamics to hybrid control systems (see Definition 2.1 for further details) in which both of the above ODE structures play the role of the *continuous flow* in their respective hybrid dynamical extension. We next suggest an admissible control input range $[u_{\min}, u_{\max}]$ that determines the *flow set* of each hybrid system. Based on the model parameters α, $u_{\min}$, and $u_{\max}$, we then construct the *jump map* of each hybrid control system by the mapping $\big(X^\top, -\beta\nabla^\top f(X)\big)^\top$ guaranteeing that the range space of the jump map is contained in its respective flow set. Notice that the velocity restart schemes take the form of $\dot{X} = -\beta\nabla f(X)$.

This paper extends the results of [19] in several ways which are summarized as follows:

- We synthesize a state-dependent gradient coefficient ($u_{\mathbf{II}}(x)$) given a prescribed control input bound and a desired convergence rate (Theorem 3.4). This is a complementary result to our earlier study [30] which is concerned with a state-dependent damping coefficient ($u_{\mathbf{I}}(x)$). Notice that the state-dependent feature of our proposed dynamical systems differs from commonly time-dependent methodologies in the literature.
- We derive a lower bound on the time between two consecutive jumps for each hybrid structure. This ensures that the constructed hybrid systems admit the so-called Zeno-free solution trajectories. It is worth noting that the regularity assumptions required by the proposed structures are different (Theorems 3.2 and 3.5).
- The proposed frameworks are general enough to include a subclass of non-convex problems. Namely, the critical requirement is that the objective function f satisfies the Polyak–Łojasiewicz (PL) inequality (Assumption (A2)), which is a weaker regularity assumption than the strong convexity that is often assumed in this context.
- We utilize the *forward-Euler* method to discretize both hybrid systems (i.e., obtain optimization algorithms). We further provide a mechanism to compute the step size such that the corresponding discrete dynamics have an exponential rate of convergence (Theorem 3.11).

The remainder of this paper is organized as follows. In Section 2, the mathematical notions are represented. The main results of the paper are introduced in Section 3. Section 4 contains the proofs of the main results. We introduce a numerical example in Section 5. This paper is finally concluded in Section 6.

Notations: The sets $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$ denote the n-dimensional Euclidean space and the space of $m \times n$ dimensional matrices with real entries, respectively. For a matrix $M \in \mathbb{R}^{m\times n}$, $M^\top$ is the transpose of M, $M \succ 0$ ($\prec 0$) refers to M positive (negative) definite, $M \succeq 0$ ($\preceq 0$) refers to M positive (negative) semi-definite, and $\lambda_{\max}(M)$ denotes the maximum eigenvalue of M. The $n \times n$ identity matrix is denoted by I_n. For a vector $v \in \mathbb{R}^n$ and $i \in \{1, \cdots, n\}$, v_i represents the i-th entry of v and $\|v\| := \sqrt{\Sigma_{i=1}^n v_i^2}$ is the Euclidean 2-norm of v. For two vectors $x, y \in \mathbb{R}^n$, $\langle x, y\rangle := x^\top y$ denotes the Euclidean inner product. For a matrix M, $\|M\| := \sqrt{\lambda_{\max}(A^\top A)}$ is the induced 2-norm. Given the set $S \subseteq \mathbb{R}^n$, ∂S and $\mathrm{int}(S)$ represent the boundary and the interior of S, respectively.

2. Preliminaries

We briefly recall some notions from hybrid dynamical systems that we will use to develop our results. Then, the problem statement is introduced along with some assumptions related to the optimization problem to be tackled in this paper. We adapt the following definition of a hybrid control system from [12] that is sufficient in the context of this paper.

Definition 2.1 (Hybrid control system)**.** *A time-invariant hybrid control system $\mathcal{H}$ comprises a controlled ODE and a jump (or a reset) rule introduced as:*

$$(\mathcal{H}) \qquad \begin{cases} \dot{x} &= F\big(x, u(x)\big), & x \in \mathcal{C} \\ x^+ &= G(x), & \text{otherwise}, \end{cases}$$

where x^+ is the state of the hybrid system after a jump, the function $u : \mathbb{R}^n \to \mathbb{R}^m$ denotes a feedback signal, the function $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is the flow map, the set $\mathcal{C} \subseteq \mathbb{R}^n$ is the flow set, and the function $G : \partial\mathcal{C} \to \mathrm{int}(\mathcal{C})$ represents the jump map.

In hybrid dynamical systems, the notion of *Zeno behavior* refers to the phenomenon that an infinite number of jumps occur in a bounded time interval. We then call a solution trajectory of a hybrid dynamical system Zeno-free if the number of jumps within any finite time interval is bounded. The existence of a lower bound on the time interval between two consecutive jumps suffices to guarantee the Zeno-freeness of a solution trajectory of a hybrid control system. Nonetheless, there exist solution concepts in the literature that accept Zeno behaviors, see for example [2, 12, 13, 23] and the references therein.

Consider the following class of unconstrained optimization problems:

$$f^* := \min_{X\in\mathbb{R}^n} f(X), \tag{2}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is an objective function. We now formally state the main problem to be addressed in this paper:

Problem 2.2. *Consider the unconstrained optimization problem (2) where the objective function f is twice differentiable. Given a positive scalar α, design a fast gradient-based method in the form of a hybrid control system ($\mathcal{H}$) with α-exponential convergence rate, i.e. for any initial condition $X(0)$ and any $t\geq 0$ we have*

$$f\big(X(t)\big)-f^*\leq e^{-\alpha t}\Big(f\big(X(0)\big)-f^*\Big),$$

where $\{X(t)\}_{t\geq 0}$ denotes the solution trajectory of the system ($\mathcal{H}$).

Assumption 2.3 (Regularity assumptions)**.** *We stipulate that the objective function $f:\mathbb{R}^n\to\mathbb{R}$ is twice differentiable and fulfills the following*

- *(Bounded Hessian) The Hessian of function f, denoted by $\nabla^2 f(x)$, is uniformly bounded, i.e.,*

$$-\ell_f I_n\preceq\nabla^2 f(x)\preceq L_f I_n, \tag{A1}$$

  where ℓ_f and L_f are non-negative constants.

- *(Gradient dominated) The function f satisfies the Polyak-Łojasiewicz inequality with a positive constant μ_f, i.e., for every x in $\mathbb{R}^n$ we have*

$$\frac{1}{2}\big\|\nabla f(x)\big\|^2\geq\mu_f\big(f(x)-f^*\big), \tag{A2}$$

  where f^* is the minimum value of f on $\mathbb{R}^n$.

- *(Lipschitz Hessian) The Hessian of the function f is Lipschitz, i.e., for every x,y in $\mathbb{R}^n$ we have*

$$\big\|\nabla^2 f(x)-\nabla^2 f(y)\big\|\leq H_f\|x-y\|, \tag{A3}$$

  where H_f is a positive constant.

Remark 2.4 (Lipschitz gradient)**.** *Since the function f is twice differentiable, Assumption (A1) implies that the function ∇f is also Lipschitz with a positive constant L_f, i.e., for every x,y in $\mathbb{R}^n$ we have*

$$\big\|\nabla f(x)-\nabla f(y)\big\|\leq L_f\|x-y\|. \tag{3}$$

We now collect two remarks underlining some features of the set of functions that satisfy (A2).

Remark 2.5 (PL functions and invexity)**.** *The PL inequality in general does not imply the convexity of a function but rather the invexity of it. The notion of invexity was first introduced by [15]. The PL inequality (A2) implies that the suboptimality measure $f-f^*$ grows at most as a quadratic function of ∇f.*

Remark 2.6 (Non-uniqueness of stationary points)**.** *While the PL inequality does not require the uniqueness of the stationary points of a function (i.e., $\{x:\nabla f(x)=0\}$), it ensures that all stationary points of the function f are global minimizers [7].*

We close our preliminary section with a couple of popular examples borrowed from [17].

Example 1 (PL functions). The composition of a strongly convex function and a linear function satisfies the PL inequality. This class includes a number of important problems such as least squares, i.e., $f(x)=\|Ax-b\|^2$ (obviously, strongly convex functions also satisfy the PL inequality). Any strictly convex function over a compact set satisfies the PL inequality. As such, the log-loss objective function in logistic regression, i.e., $f(x)=\Sigma_{i=1}^n\log\big(1+\exp(b_i a_i^\top x)\big)$, locally satisfies the PL inequality.

3. Main Results

In this section, the main results of this paper are provided. We begin with introducing two types of structures for the hybrid system ($\mathcal{H}$) motivated by the dynamics of fast gradient methods [36]. Given a positive scalar α, these structures, indexed by **I** and **II**, enable achieving the rate of convergence $\mathcal{O}(e^{-\alpha t})$ in the suboptimality measure $f\big(x_1(t)\big) - f^*$. We then collect multiple remarks highlighting the shared implications of the two structures along with a naive type of time-discretization for these structures. The technical proofs are presented in Section 4. For notational simplicity, we introduce the notation $x := (x_1, x_2)$ such that the variables x_1 and x_2 represent the system trajectories X and $\dot{X}$, respectively.

3.1. Structure I: state-dependent damping coefficient

The description of the first structure follows. We start with the flow map $F_{\mathbf{I}} : \mathbb{R}^{2n} \times \mathbb{R} \to \mathbb{R}^{2n}$ defined as

$$F_{\mathbf{I}}\big(x, u_{\mathbf{I}}(x)\big) = \begin{pmatrix} x_2 \\ -\nabla f(x_1) \end{pmatrix} + \begin{pmatrix} 0 \\ -x_2 \end{pmatrix} u_{\mathbf{I}}(x). \tag{4a}$$

Notice that $F_{\mathbf{I}}(\,\cdot\,,\,\cdot\,)$ is the state-space representation of a 2nd-order ODE. The feedback law $u_{\mathbf{I}} : \mathbb{R}^{2n} \to \mathbb{R}$ is given by

$$u_{\mathbf{I}}(x) = \alpha + \frac{\|\nabla f(x_1)\|^2 - \langle \nabla^2 f(x_1) x_2, x_2 \rangle}{\langle \nabla f(x_1), -x_2 \rangle}. \tag{4b}$$

Next, the candidate flow set $\mathcal{C}_{\mathbf{I}} \subset \mathbb{R}^{2n}$ is characterized by an admissible input interval $[\underline{u}_{\mathbf{I}}\ \overline{u}_{\mathbf{I}}]$, i.e.,

$$\mathcal{C}_{\mathbf{I}} = \big\{x \in \mathbb{R}^{2n} :\ u_{\mathbf{I}}(x) \in [\underline{u}_{\mathbf{I}}, , \overline{u}_{\mathbf{I}}]\big\}, \tag{4c}$$

where the interval bounds $\underline{u}_{\mathbf{I}}, \overline{u}_{\mathbf{I}}$ represent the range of admissible control values. Notice that the flow set $\mathcal{C}_{\mathbf{I}}$ is the domain in which the hybrid system ($\mathcal{H}$) can evolve continuously. Finally, we introduce the jump map $G_{\mathbf{I}} : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ parameterized by a constant $\beta_{\mathbf{I}}$

$$G_{\mathbf{I}}(x) = \begin{pmatrix} x_1 \\ -\beta_{\mathbf{I}} \nabla f(x_1) \end{pmatrix}. \tag{4d}$$

The parameter $\beta_{\mathbf{I}}$ ensures that the range space of the jump map $G_{\mathbf{I}}$ is a strict subset of $\mathrm{int}(\mathcal{C}_{\mathbf{I}})$. By construction, one can inspect that any neighborhood of the optimizer x_1^* has a non-empty intersection with the flow set $\mathcal{C}_{\mathbf{I}}$. That is, there always exist paths in the set $\mathcal{C}_{\mathbf{I}}$ that allow the continuous evolution of the hybrid system to approach arbitrarily close to the optimizer.

We are now in a position to formally present the main results related to the structure **I** given in (4). For the sake of completeness, we borrow the first result from [19]. This theorem provides a framework to set the parameters $\underline{u}_{\mathbf{I}}$, $\overline{u}_{\mathbf{I}}$, and $\beta_{\mathbf{I}}$ in (4c) and (4d) in order to ensure the desired exponential convergence rate $\mathcal{O}(e^{-\alpha t})$.

Theorem 3.1 (Continuous-time convergence rate - **I**). *Consider a positive scalar α and a smooth function $f : \mathbb{R}^n \to \mathbb{R}$ satisfying Assumptions (A1) and (A2). Then, the solution trajectory of the hybrid control system ($\mathcal{H}$) with the respective parameters (4) starting from any initial condition $x_1(0)$ satisfies*

$$f\big(x_1(t)\big) - f^* \le e^{-\alpha t}\Big(f\big(x_1(0)\big) - f^*\Big), \quad \forall t \ge 0, \tag{5}$$

if the scalars $\underline{u}_{\boldsymbol{I}}$, $\overline{u}_{\boldsymbol{I}}$, and $\beta_{\boldsymbol{I}}$ are chosen such that

$$\underline{u}_{\boldsymbol{I}} < \alpha + \beta_{\boldsymbol{I}}^{-1} - L_f \beta_{\boldsymbol{I}}, \tag{6a}$$

$$\overline{u}_{\boldsymbol{I}} > \alpha + \beta_{\boldsymbol{I}}^{-1} + \ell_f \beta_{\boldsymbol{I}}, \tag{6b}$$

$$\alpha \le 2\mu_f \beta_{\boldsymbol{I}}. \tag{6c}$$

The next result establishes a key feature of the solution trajectories generated by the dynamics ($\mathcal{H}$) with the respective parameters (4), that the solution trajectories are indeed *Zeno*-free.

Theorem 3.2 (Zeno-free hybrid trajectories - **I**). *Consider a smooth function $f:\mathbb{R}^n\to\mathbb{R}$ satisfying Assumption 2.3, and the corresponding hybrid control system ($\mathcal{H}$) with the respective parameters (4) satisfying (6). Given the initial condition $\big(x_1(0),-\beta_{\boldsymbol{I}}\nabla f\big(x_1(0)\big)\big)$ the time between two consecutive jumps of the solution trajectory, denoted by $\tau_{\boldsymbol{I}}$, satisfies for any scalar $r>1$*

$$\tau_{\boldsymbol{I}}\geq\log\left(\max\left\{\frac{a_1}{a_2+a_3\Big\|\nabla f\big(x_1(0)\big)\Big\|}+1,r^{1/\delta}\right\}\right), \tag{7}$$

where the constants involved are defined as

$$C:=\frac{(\overline{u}_{\boldsymbol{I}}-\alpha)+\sqrt{(\overline{u}_{\boldsymbol{I}}-\alpha)^2+4L_f}}{2}, \tag{8a}$$

$$\delta:=C+\max\{\overline{u}_{\boldsymbol{I}},-\underline{u}_{\boldsymbol{I}}\}, \tag{8b}$$

$$\mathcal{L}_f:=\max\{\ell_f,L_f\}, \tag{8c}$$

$$a_1:=\min\{\overline{u}_{\boldsymbol{I}}-(\alpha+\beta_{\boldsymbol{I}}^{-1}+\ell_f\beta_{\boldsymbol{I}}),(\alpha+\beta_{\boldsymbol{I}}^{-1}-L_f\beta_{\boldsymbol{I}})-\underline{u}_{\boldsymbol{I}}\}, \tag{8d}$$

$$a_2:=rL_f\delta^{-1}(r\beta_{\boldsymbol{I}}C+1)+\beta_{\boldsymbol{I}}^{-1}+(r^2+r+1)\beta_{\boldsymbol{I}}\mathcal{L}_f, \tag{8e}$$

$$a_3:=r^3\beta_{\boldsymbol{I}}^2H_f\delta^{-1}. \tag{8f}$$

Consequently, the solution trajectories are Zeno-free.

Remark 3.3 (Non-uniform inter-jumps - **I**). *Notice that Theorem 3.2 suggests a lower-bound for the inter-jump interval $\tau_{\boldsymbol{I}}$ that depends on $\|\nabla f\big(x_1\big)\|$. In light of the fact that the solution trajectories converge to the optimal solutions, and as such $\nabla f\big(x_1\big)$ tends to zero, one can expect that the frequency at which the jumps occur reduces as the hybrid control system evolves in time.*

3.2. **Structure II: state-dependent potential coefficient**

In this subsection, we first provide the structure **II** for the hybrid control system ($\mathcal{H}$). We skip the the details of differences with the structure **I** and differ it to Subection 3.3 and Section 4. Consider the flow map $F_{\mathbf{II}}:\mathbb{R}^{2n}\times\mathbb{R}\to\mathbb{R}^{2n}$ given by

$$F_{\mathbf{II}}\big(x,u_{\mathbf{II}}(x)\big)=\begin{pmatrix}x_2\\-x_2\end{pmatrix}+\begin{pmatrix}0\\-\nabla f(x_1)\end{pmatrix}u_{\mathbf{II}}(x), \tag{9a}$$

and the feedback law $u_{\mathbf{II}}:\mathbb{R}^{2n}\to\mathbb{R}$ given by

$$u_{\mathbf{II}}(x)=\frac{\langle\nabla^2f(x_1)x_2,x_2\rangle+(1-\alpha)\langle\nabla f(x_1),-x_2\rangle}{\|\nabla f(x_1)\|^2}. \tag{9b}$$

The candidate flow set $\mathcal{C}_{\mathbf{II}}\subset\mathbb{R}^{2n}$ is parameterized by an admissible interval $[\underline{u}_{\mathbf{II}}\ \overline{u}_{\mathbf{II}}]$ as follows:

$$\mathcal{C}_{\mathbf{II}}=\left\{x\in\mathbb{R}^{2n}:\ u_{\mathbf{II}}(x)\in[\underline{u}_{\mathbf{II}},\overline{u}_{\mathbf{II}}]\right\}. \tag{9c}$$

Parameterized in a constant $\beta_{\mathbf{II}}$, the jump map $G_{\mathbf{II}}:\mathbb{R}^{2n}\to\mathbb{R}^{2n}$ is given by

$$G_{\mathbf{II}}(x)=\begin{pmatrix}x_1\\-\beta_{\mathbf{II}}\nabla f(x_1)\end{pmatrix}. \tag{9d}$$

Theorem 3.4 (Continuous-time convergence rate - **II**). *Consider a positive scalar α and a smooth function $f:\mathbb{R}^n\to\mathbb{R}$ satisfying Assumptions (A1) and (A2). Then, the solution trajectory of the hybrid control system ($\mathcal{H}$) with the respective parameters (9) starting from any initial condition $x_1(0)$ satisfies the inequality (5) if the scalars $\underline{u}_{\boldsymbol{II}}$, $\overline{u}_{\boldsymbol{II}}$, and $\beta_{\boldsymbol{II}}$ are chosen such that*

$$\underline{u}_{\boldsymbol{II}}<-\ell_f\beta_{\boldsymbol{II}}^2+(1-\alpha)\beta_{\boldsymbol{II}}, \tag{10a}$$

$$\overline{u}_{\boldsymbol{II}}>L_f\beta_{\boldsymbol{II}}^2+(1-\alpha)\beta_{\boldsymbol{II}}, \tag{10b}$$

$$\alpha\leq 2\mu_f\beta_{\boldsymbol{II}}. \tag{10c}$$

Theorem 3.5 (Zeno-free hybrid trajectories - **II**). *Consider a smooth function $f : \mathbb{R}^n \to \mathbb{R}$ satisfying Assumptions (A1) and (A2), and the hybrid control system ($\mathcal{H}$) with the respective parameters (9) satisfying (10). Given the initial condition $\big(x_1(0), -\beta_{\boldsymbol{II}} \nabla f\big(x_1(0)\big)\big)$ the time between two consecutive jumps of the solution trajectory, denoted by $\tau_{\boldsymbol{II}}$, satisfies for any scalar $r \in (0,1)$*

$$\tau_{\boldsymbol{II}} \geq \min\left\{r\omega^{-1}, \delta(b_1+b_2)^{-1}\right\}. \tag{11}$$

where the involved scalars are defined as

$$\begin{aligned}
\delta &:= \min\left\{\overline{u}_{\boldsymbol{II}} - (L_f\beta_{\boldsymbol{II}}^2 + (1-\alpha)\beta_{\boldsymbol{II}}), (-\ell_f\beta_{\boldsymbol{II}}^2 + (1-\alpha)\beta_{\boldsymbol{II}}) - \underline{u}_{\boldsymbol{II}}\right\},\\
U &:= \max\{\overline{u}_{\boldsymbol{II}}, -\underline{u}_{\boldsymbol{II}}\},\\
\mathcal{L}_f &:= \max\{\ell_f, L_f\},\\
\omega &:= \mathcal{L}_f(\beta_{\boldsymbol{II}}^2 + \beta_{\boldsymbol{II}}U)^{\frac{1}{2}},\\
b1 &:= \frac{2\mathcal{L}_f\beta_{\boldsymbol{II}}\big(U + \omega(\beta_{\boldsymbol{II}}+U)\big)}{(1-r)^3},\\
b_2 &:= |\alpha-1|\frac{2\omega\beta_{\boldsymbol{II}}}{(1-r)^3} + |\alpha-1|\alpha\beta_{\boldsymbol{II}}(1+r).
\end{aligned}$$

Thus, the solution trajectories are Zeno-free.

Remark 3.6 (Uniform inter-jumps - **II**). *Notice that unlike Theorem 3.2, the derived lower-bound for the inter-jump interval $\tau_{\boldsymbol{II}}$ is uniform in the sense that the bound is independent of $\|\nabla f\big(x_1\big)\|$. Furthermore, the regularity requirement on the function f is weaker than the one used in Theorem 3.2, i.e., the function f is not required to satisfy the Assumption (A3).*

Notice that the main differences between the structures (4), (9) lie in the flow maps and the feedback laws. On the other hand, these structures share the key feature of enabling an α-exponential convergence rate for the hybrid system ($\mathcal{H}$) through their corresponding control inputs. The reason explaining the aforementioned points is deferred until later in Section 4.

3.3. **Further Discussions**

In what follows, we collect several remarks regarding the common features of the proposed structures. Then, we apply the *forward-Euler* method of time-discretization to these structures of the hybrid control system ($\mathcal{H}$). The proposed discretizations guarantee an exponential rate of convergence in the suboptimality measure $f(x_1^k) - f^*$, where k is the iteration index.

Remark 3.7 (Weaker regularity than strong convexity). *The PL inequality is a weaker requirement than the strong convexity, which is often assumed in similar contexts [36, 37, 39]. It is worth noting that such a condition has also been used in the context of 1st-order algorithms [17].*

Remark 3.8 (Hybrid embedding of restarting). *The hybrid frameworks intrinsically capture restarting schemes through the jump map. The schemes are a weighted gradient where the weight factor $\beta_{\boldsymbol{I}}$ or $\beta_{\boldsymbol{II}}$ is essentially characterized by the given data α, μ_f, ℓ_f, and L_f. One may inspect that the constant $\beta_{\boldsymbol{I}}$ or $\beta_{\boldsymbol{II}}$ can be in fact introduced as a state-dependent weight factor to potentially improve the performance. Nonetheless, for the sake of simplicity of exposition, we do not pursue this level of generality in this paper.*

Remark 3.9 (2nd-order information). *Although our proposed frameworks require 2nd-order information, i.e., the Hessian $\nabla^2 f$, this requirement only appears in a mild form as an evaluation in the same spirit as the modified Newton step proposed in [31]. Furthermore, we emphasize that our results still hold true if one replaces $\nabla^2 f(x_1)$ with its upper-bound $L_f I_n$ following essentially the same analysis. For further details we refer the reader to the proof of Theorem 3.4.*

Remark 3.10 (Fundamental limits on control input)**.** *An implication of Theorem 3.4 is that if the desired convergence rate* $\alpha > \big(\frac{2\mu_f}{2\mu_f+\ell_f}\big)$*, it is then required to choose* $\underline{u}_{\boldsymbol{II}} < 0$*, indicating that the system may need to receive energy through a negative damping. On a similar note, Theorem 3.1 asserts that the upper bound requires* $\overline{u}_{\boldsymbol{I}} > \alpha$*, and if* $\alpha > \big(\frac{2\mu_f}{\sqrt{\max\{L_f-2\mu_f,0\}}}\big)$*, we then have to set* $\underline{u}_{\boldsymbol{I}} < 0$ [19, Remark 3.4].

3.4. **Discrete-Time Dynamics**

In the next result, we show that if one applies the forward-Euler method on the two proposed structures properly, the resulting discrete-time hybrid control systems possess exponential convergence rates. Suppose $i \in \{\mathbf{I}, \mathbf{II}\}$ and let us denote by s the time-discretization step size. Consider the discrete-time hybrid control system

$$\mathcal{H}_{d,i} := \begin{cases} x^{k+1} = F_{d,i}\big(x^k, u_{d,i}(x^k)\big), & x^k \in \mathcal{C}_{d,i} \\ x^{k+1} = G_{d,i}(x^k), & \text{otherwise,} \end{cases} \tag{12}$$

where $F_{d,i}$, $G_{d,i}$, and $\mathcal{C}_{d,i}$ are the flow map, the jump map, and the flow set, respectively. The discrete flow map $F_{d,i} : \mathbb{R}^{2n} \times \mathbb{R} \to \mathbb{R}^{2n}$ is given by

$$F_{d,i}\big(x^k, u_{d,i}(x^k)\big) = x^k + sF_i\big(x^k, u_i(x^k)\big),\ i \in \{\mathbf{I}, \mathbf{II}\}, \tag{13a}$$

where F_i and u_i are defined in (4a) and (4b), or (9a) and (9b) based on the considered structure i. The discrete flow set $\mathcal{C}_{d,i} \subset \mathbb{R}^{2n}$ is defined as

$$\mathcal{C}_{d,i} := \big\{(x_1^k, x_2^k) \in \mathbb{R}^{2n} : c_1\|x_2^k\|^2 \le \|\nabla f(x_1^k)\|^2 \le c_2\langle \nabla f(x_1^k), -x_2^k\rangle\big\}, \tag{13b}$$

and, c_1 and c_2 are two positive scalars. The discrete jump map $G_{d,i} : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is given by $G_{d,i}(x^k) = \big((x^k)^\top, -\beta\nabla^\top f(x^k)\big)^\top$.

It is evident in the flow sets $\mathcal{C}_{d,i}$ of the discrete-time dynamics that these sets are no longer defined based on admissible input intervals. The reason has to do with the difficulties that arise from appropriately discretizing the control inputs $u_{\mathbf{I}}$ and $u_{\mathbf{II}}$. Nonetheless, the next result guarantees exponential rate of convergence of the discrete-time control system (12) with either of the respective structure **I** or **II**, by introducing a mechanism to set the scalars c_1, c_2, and β.

Theorem 3.11 (Stable discretization - **I & II**)**.** *Consider a smooth function* $f : \mathbb{R}^n \to \mathbb{R}$ *satisfying Assumptions* (A1) *and* (A2)*. The solution trajectory of the discrete-time hybrid control system (12) with the respective structure* $i \in \{\mathbf{I}, \mathbf{II}\}$ *and starting from any initial condition* x_1^0*, satisfies*

$$f(x_1^{k+1}) - f^* \le \lambda(s, c_1, c_2, \beta)\big(f(x_1^k) - f^*\big), \tag{14}$$

with $\lambda(s, c_1, c_2, \beta) \in (0, 1)$ *given by*

$$\lambda(s, c_1, c_2, \beta) := 1 + 2\mu_f\big(-\frac{s}{c_2} + \frac{L_f}{2c_1}s^2\big), \tag{15}$$

if the parameters s*,* c_1 *,*c_2*, and* β *satisfy*

$$\sqrt{c_1} \le c_2, \tag{16a}$$

$$\beta^2 c_1 \le 1 \le \beta c_2, \tag{16b}$$

$$c_2 L_f s < 2c_1. \tag{16c}$$

Remark 3.12 (Naive discretization)**.** *We would like to emphasize that the exponential convergence of the proposed discretization method solely depends on the dynamics* x_1 *and the properties of the objective function* f*. Thus, we deliberately avoid labeling the scalars* c_1*,* c_2*, and* β *by the structure index* i*. Crucially, the structures of the control laws do not impact the relations* (16) *in Theorem 3.11, see Subsection 4.4 for more details. In light of the above facts, we believe that a more in-depth analysis of the dynamics along with the control structures may provide a more intelligent way to improve the discretization result of Theorem 3.11.*

Algorithm 1 Sate-dependent fast gradient method

Input: data x_1^0, ℓ_f, L_f, μ_f, $\alpha \in \mathbb{R}^+$, $k_{\max} \in \mathbb{N}^+$, $i \in \{\mathbf{I}, \mathbf{II}\}$
Set: $\sqrt{c_1} = c_2 = \beta^{-1} = L_f s$, $x_2^0 = -\beta \nabla f(x_1^0)$
$x^0 = (x_1^0, x_2^0)$
for $k = 1$ **to** $k_{\max}$ **do**
  if $c_1 \|x_2^k\|^2 \leq \|\nabla f(x_1^k)\|^2 \leq c_2 \langle \nabla f(x_1^k), -x_2^k \rangle$ **then**
    $x^{k+1} \leftarrow F_{d,i}(x^k)$
  else
    $x^{k+1} \leftarrow G_{d,i}(x^k)$
  end if
end for

Corollary 3.13 (Optimal guaranteed rate)**.** *The optimal convergence rate guaranteed by Theorem 3.11 for the discrete-time dynamics is* $\lambda^* := \left(1 - \frac{\mu_f}{L_f}\right)$ *and*

$$\sqrt{c_1^*} = c_2^* = \frac{1}{\beta^*} = L_f s^*.$$

The pseudocode to implement the above corollary is presented in Algorithm 1 using the discrete-time dynamics (12) with the respective parameters **I** or **II**.

4. Technical Proofs

4.1. Proof of Theorem 3.2

In this subsection, we first set the stage by providing two intermediate results regarding the properties of dynamics of the hybrid control system ($\mathcal{H}$) with the respective parameters (4). We then employ these facts to formally state the proof of Theorem 3.2. The next lemma reveals a relation between $\nabla f(x_1)$ and x_2 along the trajectories of the hybrid control system. In this subsection, for the sake of brevity we denote $x_1(t)$ and $x_1(0)$ by x_1 and $x_{1,0}$, respectively. We adapt the same change of notation for x_2 and x, as well.

Lemma 4.1 (Velocity lower bound)**.** *Consider the continuous-time hybrid control system ($\mathcal{H}$) with the respective parameters (4) satisfying (6) where the function f satisfies Assumptions (A1) and (A2). Then, we have*

$$\left\|\nabla f(x_1)\right\| \leq C \|x_2\|, \tag{17}$$

where C is given by (8a).

Proof. Notice that, by the definition of the control law and the upper bound condition $u_{\mathbf{I}}(x) \leq \overline{u}_{\mathbf{I}}$, we have

$$\left\|\nabla f(x_1)\right\|^2 - \langle \nabla^2 f(x_1) x_2, x_2 \rangle \leq (\overline{u}_{\mathbf{I}} - \alpha) \langle \nabla f(x_1), -x_2 \rangle \leq (\overline{u}_{\mathbf{I}} - \alpha) \left\|\nabla f(x_1)\right\| \cdot \|x_2\|,$$

where the second inequality follows from the Cauchy-Schwarz inequality. Since the function f satisfies Assumption (A1), one can infer that

$$\left\|\nabla f(x_1)\right\|^2 - L_f \|x_2\|^2 \leq (\overline{u}_{\mathbf{I}} - \alpha) \left\|\nabla f(x_1)\right\| \cdot \|x_2\|,$$

which in turn can be reformulated into

$$\frac{\left\|\nabla f(x_1)\right\|^2}{\|x_2\|^2} - (\overline{u}_{\mathbf{I}} - \alpha) \frac{\left\|\nabla f(x_1)\right\|}{\|x_2\|} - L_f \leq 0. \tag{18}$$

Defining the variable $y := \left\|\nabla f(x_1)\right\| / \|x_2\|$, the inequality (18) becomes the quadratic inequality $y^2 - (\overline{u}_{\mathbf{I}} - \alpha) y - L_f \leq 0$. Taking into account that $y \geq 0$, it then follows from (17) that

$$y = \frac{\left\|\nabla f(x_1)\right\|}{\|x_2\|} \leq \frac{(\overline{u}_{\mathbf{I}} - \alpha) + \sqrt{(\overline{u}_{\mathbf{I}} - \alpha)^2 + 4L_f}}{2} =: C.$$

This concludes the proof of Lemma 4.1. □

In the following, we provide a result that indicates the variation of norms x_1 and x_2, along the trajectories of the hybrid control system, are bounded in terms of time while they evolve according to the continuous mode. Since the hybrid control system is time-invariant, such bounds can be generalized to all inter-jump intervals.

Lemma 4.2 (Trajectory growth rate)**.** *Suppose that the same conditions as specified in Lemma 4.1 hold, and the hybrid control system (H), (4) starts from the initial condition* $\big(x_{1,0}, -\beta_{\boldsymbol{I}}\nabla f(x_{1,0})\big)$ *for some* $x_{1,0} \in \mathbb{R}^n$*. Then*

$$\|x_1 - x_{1,0}\| \leq \delta^{-1}\|x_{2,0}\|\big(e^{\delta t} - 1\big), \tag{19a}$$

$$\|x_2 - x_{2,0}\| \leq \|x_{2,0}\|\big(e^{\delta t} - 1\big), \tag{19b}$$

where δ *is given by (8b).*

Proof. Using the flow dynamics (4a) we obtain

$$\begin{aligned}\frac{d}{dt}\|x_2\| \leq \Big\|\frac{d}{dt}x_2\Big\| &\leq \big\|\nabla f(x_1)\big\| + \big|u_{\mathbf{I}}(x)\big| \cdot \|x_2\| \\ &\leq (C + \max\{\overline{u}_{\mathbf{I}}, -\underline{u}_{\mathbf{I}}\})\|x_2\| = \delta\|x_2\|.\end{aligned} \tag{20}$$

The inequality (20) implies that

$$\|x_2\| \leq \|x_{2,0}\|e^{\delta t}. \tag{21}$$

Furthermore, notice that

$$\frac{d}{dt}\|x_1 - x_{1,0}\| \leq \Big\|\frac{d}{dt}(x_1 - x_{1,0})\Big\| = \|x_2\|.$$

Integrating the two sides of the above inequality leads to

$$\begin{aligned}\|x_1 - x_{1,0}\| &\leq \int_0^t \big\|x_2(s)\big\|\, ds \leq \int_0^t \|x_{2,0}\|e^{\delta s}\, ds \\ &= \frac{\|x_{2,0}\|}{\delta}\big(e^{\delta t} - 1\big),\end{aligned}$$

in which we made use of (21). Hence, the inequality (19a) in Lemma 4.1 is concluded. Next, we shall establish the inequality (19b). Note that

$$\begin{aligned}\frac{d}{dt}\|x_2 - x_{2,0}\| \leq \Big\|\frac{d}{dt}(x_2 - x_{2,0})\Big\| = \Big\|\frac{d}{dt}x_2\Big\| &\leq \delta\big\|x_2\big\| \\ &\leq \delta\|x_2 - x_{2,0}\| + \delta\|x_{2,0}\|.\end{aligned}$$

Applying Grownwall's inequality [18, Lemma A.1] then leads to the desired inequality (19b). The claims in Lemma 4.2 follow. □

Proof of Theorem 3.2: The proof comprises five steps, and the key part is to guarantee that during the first inter-jump interval the quantity $\big|u_{\mathbf{I}}(x) - u_{\mathbf{I}}(x_{,0})\big|$ is bounded by a continuous function $\phi\Big(t, \big\|\nabla f(x_{1,0})\big\|\Big)$, which is exponential in its first argument and linear in its second argument. Then, it follows from the continuity of the function ϕ that the solution trajectories of the hybrid control system are Zeno-free.

Step 1: Let us define $g(t) := \langle \nabla f(x_1), -x_2\rangle$. We now compute the derivative of $g(t)$ along the trajectories of the hybrid control system (H), (4) during the first inter-jump interval, i.e.,

$$\begin{aligned}\frac{d}{dt}g(t) &= \langle \nabla^2 f(x_1)x_2, -x_2\rangle + \langle \nabla f(x_1), u_{\mathbf{I}}(x)x_2 + \nabla f(x_1)\rangle \\ &= -\langle \nabla^2 f(x_1)x_2, x_2\rangle + \big\|\nabla f(x_1)\big\|^2 + u_{\mathbf{I}}(x)\langle \nabla f(x_1), x_2\rangle \\ &= -\alpha\langle \nabla f(x_1), -x_2\rangle = -\alpha\, g(t).\end{aligned}$$

According to the above discussion and considering the initial state $x_{2,0} = -\beta_{\mathbf{I}} \nabla f(x_{1,0})$, it follows that

$$\langle \nabla f(x_1), -x_2 \rangle = \beta_{\mathbf{I}} \big\|\nabla f(x_{1,0})\big\|^2 e^{-\alpha t}. \tag{22}$$

Step 2: The quantity $\Big|e^{\alpha t}\big\|\nabla f(x_1)\big\|^2 - \big\|\nabla f(x_{1,0})\big\|^2\Big|$ is bounded along the trajectories of the hybrid control system ($\mathcal{H}$) with the respective parameters (4) during the first inter-jump interval, i.e.,

$$\begin{aligned}
\Big|e^{\alpha t}\big\|\nabla f(x_1)\big\|^2 - \big\|\nabla f(x_{1,0})\big\|^2\Big| &= \Big|e^{\alpha t}\big\|\nabla f(x_1)\big\|^2 - (e^{\alpha t} - e^{\alpha t} + 1)\big\|\nabla f(x_{1,0})\big\|^2\Big| \\
&\overset{\text{(i)}}{\leq} e^{\alpha t}\Big|\big\|\nabla f(x_1)\big\|^2 - \big\|\nabla f(x_{1,0})\big\|^2\Big| + (e^{\alpha t} - 1)\big\|\nabla f(x_{1,0})\big\|^2 \\
&= e^{\alpha t}\Big|\big\langle \nabla f(x_1) - \nabla f(x_{1,0}), \nabla f(x_1) + \nabla f(x_{1,0})\big\rangle\Big| \\
&\qquad + (e^{\alpha t} - 1)\big\|\nabla f(x_{1,0})\big\|^2 \\
&\overset{\text{(ii)}}{\leq} e^{\alpha t}\big\|\nabla f(x_1) - \nabla f(x_{1,0})\big\| \cdot \big\|\nabla f(x_1) + \nabla f(x_{1,0})\big\| \\
&\qquad + (e^{\alpha t} - 1)\big\|\nabla f(x_{1,0})\big\|^2 \\
&\overset{\text{(iii)}}{\leq} e^{\alpha t} L_f \|x_1 - x_{1,0}\| \cdot \big(\beta_{\mathbf{I}} C e^{\delta t} + 1\big)\frac{\|x_{2,0}\|}{\beta_{\mathbf{I}}} + \big(e^{\alpha t} - 1\big)\frac{\|x_{2,0}\|^2}{\beta_{\mathbf{I}}^2} \\
&\overset{\text{(iv)}}{\leq} e^{\alpha t} L_f \big(e^{\delta t} - 1\big)\frac{\|x_{2,0}\|}{\delta} \cdot \big(\beta_{\mathbf{I}} C e^{\delta t} + 1\big)\frac{\|x_{2,0}\|}{\beta_{\mathbf{I}}} + \big(e^{\alpha t} - 1\big)\frac{\|x_{2,0}\|^2}{\beta_{\mathbf{I}}^2} \\
&= \left(\frac{L_f}{\delta \beta_{\mathbf{I}}} e^{\alpha t}\big(\beta_{\mathbf{I}} C e^{\delta t} + 1\big)\big(e^{\delta t} - 1\big) + \frac{1}{\beta_{\mathbf{I}}^2}\big(e^{\alpha t} - 1\big)\right)\|x_{2,0}\|^2,
\end{aligned}$$

where we made use of the triangle inequality in the inequality (i), the Cauchy-Schwarz inequality in the inequality (ii), Assumption (A1) and its consequence in Remark 2.4 along with the triangle inequality in the inequality (iii), and the inequality (19a) in the inequality (iv), respectively.

Step 3: Observe that

$$\begin{aligned}
&\big|e^{\alpha t}\langle \nabla^2 f(x_1)x_2, x_2\rangle - \langle \nabla^2 f(x_{1,0})x_{2,0}, x_{2,0}\rangle\big| \\
&\quad = \Big|e^{\alpha t}\big\langle \big[\nabla^2 f(x_1) - \nabla^2 f(x_{1,0}) + \nabla^2 f(x_{1,0})\big]x_2, x_2\big\rangle - \big(e^{\alpha t} - e^{\alpha t} + 1\big)\langle \nabla^2 f(x_{1,0})x_{2,0}, x_{2,0}\rangle\Big| \\
&\quad = \Big|e^{\alpha t}\big\langle \big[\nabla^2 f(x_1) - \nabla^2 f(x_{1,0})\big]x_2, x_2\big\rangle + e^{\alpha t}\langle \nabla^2 f(x_{1,0})x_2, x_2\rangle - e^{\alpha t}\langle \nabla^2 f(x_{1,0})x_2, x_2\rangle \\
&\qquad + \big(e^{\alpha t} - 1\big)\langle \nabla^2 f(x_{1,0})x_{2,0}, x_{2,0}\rangle\Big| \\
&\quad \overset{\text{(i)}}{\leq} e^{\alpha t}\Big|\big\langle \big[\nabla^2 f(x_1) - \nabla^2 f(x_{1,0})\big]x_2, x_2\big\rangle\Big| + e^{\alpha t}\Big|\langle \nabla^2 f(x_{1,0})x_2, x_2\rangle - \langle \nabla^2 f(x_{1,0})x_{2,0}, x_{2,0}\rangle\Big| \\
&\qquad + \big(e^{\alpha t} - 1\big)\Big|\langle \nabla^2 f(x_{1,0})x_{2,0}, x_{2,0}\rangle\Big| \\
&\quad \overset{\text{(ii)}}{\leq} e^{\alpha t} H_f \|x_1 - x_{1,0}\| \cdot \|x_2\|^2 + e^{\alpha t}\Big|\big\langle \nabla^2 f(x_{1,0})\big[x_2 - x_{2,0}\big], x_2 + x_{2,0}\big\rangle\Big| + \mathcal{L}_f \|x_{2,0}\|^2\big(e^{\alpha t} - 1\big),
\end{aligned}$$

where the inequality (i) follows from the triangle inequality, and the inequality (ii) is an immediate consequence of Assumptions (A3) and (A1), recalling $\mathcal{L}_f = \max\{\ell_f, L_f\}$. According to the above analysis, one can deduce that

$$\begin{aligned}
&\big|e^{\alpha t}\langle \nabla^2 f(x_1)x_2, x_2\rangle - \langle \nabla^2 f(x_{1,0})x_{2,0}, x_{2,0}\rangle\big| \\
&\overset{\text{(i)}}{\leq} e^{\alpha t} H_f \frac{\|x_{2,0}\|}{\delta}\big(e^{\delta t} - 1\big) \cdot e^{2\delta t}\|x_{2,0}\|^2 + e^{\alpha t}\mathcal{L}_f \|x_2 - x_{2,0}\| \cdot \|x_2 + x_{2,0}\| + \big(e^{\alpha t} - 1\big)\mathcal{L}_f \|x_{2,0}\|^2 \\
&\overset{\text{(ii)}}{\leq} \frac{H_f}{\delta} e^{(\alpha + 2\delta)t}\big\|x_2(0)\big\|^3 \cdot (e^{\delta t} - 1) + e^{\alpha t}\mathcal{L}_f\big(e^{\delta t} - 1\big)\|x_{2,0}\| \cdot \big(e^{\delta t} + 1\big)\|x_{2,0}\| + \mathcal{L}_f \|x_{2,0}\|^2\big(e^{\alpha t} - 1\big) \\
&= \Big((H_f/\delta)\ e^{(\alpha + 2\delta)t}\|x_{2,0}\| \cdot \big(e^{\delta t} - 1\big) + \mathcal{L}_f\big(e^{(\alpha+\delta)t} + e^{\alpha t}\big)\big(e^{\delta t} - 1\big) + \mathcal{L}_f(e^{\alpha t} - 1)\Big)\|x_{2,0}\|^2,
\end{aligned}$$

where we made use of the inequality (19a), the inequality (19b), and the triangle inequality in the inequality (i), and the inequality (19b) and the triangle inequality in the inequality (ii), respectively.

Step 4: We now study the input variation $\big|u_{\mathbf{I}}(x) - u_{\mathbf{I}}(x_{,0})\big|$ along the solution trajectories of the hybrid control system ($\mathcal{H}$), (4) during the first inter-jump interval. Observe that

$$\begin{aligned}
&\big|u_{\mathbf{I}}(x) - u_{\mathbf{I}}(x_{,0})\big| \\
&= \Big|\frac{\big\|\nabla f(x_1)\big\|^2 - \langle \nabla^2 f(x_1)x_2(t), x_2\rangle}{\langle \nabla f(x_1), -x_2\rangle} - \frac{\big\|\nabla f(x_{1,0})\big\|^2 - \langle \nabla^2 f(x_{1,0})x_{2,0}, x_{2,0}\rangle}{\langle \nabla f(x_{1,0}), -x_{2,0}\rangle}\Big| \\
&= \Big|\frac{\big\|\nabla f(x_1)\big\|^2}{\beta_{\mathbf{I}}\big\|\nabla f(x_{1,0})\big\|^2 e^{-\alpha t}} - \frac{\langle \nabla^2 f(x_1)x_2, x_2\rangle}{\beta_{\mathbf{I}}\big\|\nabla f(x_{1,0})\big\|^2 e^{-\alpha t}} - \frac{\big\|\nabla f(x_{1,0})\big\|^2}{\beta_{\mathbf{I}}\big\|\nabla f(x_{1,0})\big\|^2} + \frac{\langle \nabla^2 f(x_{1,0})x_{2,0}, x_{2,0}\rangle}{\beta_{\mathbf{I}}\big\|\nabla f(x_{1,0})\big\|^2}\Big| \\
&\overset{\text{(i)}}{\leq} \frac{1}{\beta_{\mathbf{I}}\big\|\nabla f(x_{1,0})\big\|^2}\Big|e^{\alpha t}\big\|\nabla f(x_1)\big\|^2 - \big\|\nabla f(x_{1,0})\big\|^2\Big| \\
&\quad + \frac{1}{\beta_{\mathbf{I}}\big\|\nabla f(x_{1,0})\big\|^2}\Big|e^{\alpha t}\langle \nabla^2 f(x_1)x_2, x_2\rangle - \langle \nabla^2 f(x_{1,0})x_{2,0}, x_{2,0}\rangle\Big| \\
&\overset{\text{(ii)}}{=} \frac{\beta_{\mathbf{I}}}{\|x_{2,0}\|^2}\Big|e^{\alpha t}\big\|\nabla f(x_1)\big\|^2 - \big\|\nabla f(x_{1,0})\big\|^2\Big| + \frac{\beta_{\mathbf{I}}}{\|x_{2,0}\|^2}\Big|e^{\alpha t}\langle\ \nabla^2 f(x_1)x_2, x_2\rangle - \langle \nabla^2 f(x_{1,0})x_{2,0}, x_{2,0}\rangle\Big|,
\end{aligned}$$

where we made use of the triangle inequality in the inequality (i) and the relation (22) in the equality (ii), respectively. Based on the above discussion, we then conclude that

$$\begin{aligned}
&\big|u_{\mathbf{I}}(x) - u_{\mathbf{I}}(x_{,0})\big| \\
&\overset{\text{(i)}}{\leq} \frac{\beta_{\mathbf{I}}}{\|x_{2,0}\|^2}\left(\frac{L_f}{\delta\beta_{\mathbf{I}}}e^{\alpha t}\big(\beta_{\mathbf{I}}Ce^{\delta t}+1\big)\big(e^{\delta t}-1\big) + \frac{1}{\beta_{\mathbf{I}}^2}\big(e^{\alpha t}-1\big)\right)\|x_{2,0}\|^2 \\
&\quad + \frac{\beta_{\mathbf{I}}}{\|x_{2,0}\|^2}\left(\frac{H_f}{\delta}e^{(\alpha+2\delta)t}\|x_{2,0}\|\cdot\big(e^{\delta t}-1\big) + \mathcal{L}_f\big(e^{(\alpha+\delta)t}+e^{\alpha t}\big)\big(e^{\delta t}-1\big) + \mathcal{L}_f\big(e^{\alpha t}-1\big)\right)\|x_{2,0}\|^2 \\
&\overset{\text{(ii)}}{\leq} \frac{L_f}{\delta}e^{\delta t}(\beta_{\mathbf{I}}Ce^{\delta t}+1)(e^{\delta t}-1) + \frac{1}{\beta_{\mathbf{I}}}(e^{\delta t}-1) \\
&\quad + \beta_{\mathbf{I}}\Big(\beta_{\mathbf{I}}H_f\delta^{-1}\cdot e^{3\delta t}\big\|\nabla f(x_{1,0})\big\|\cdot\big(e^{\delta t}-1\big) + \mathcal{L}_f\big(e^{2\delta t}+e^{\delta t}\big)\big(e^{\delta t}-1\big) + \mathcal{L}_f\big(e^{\delta t}-1\big)\Big) \\
&= \Big(L_f\delta^{-1}\cdot e^{\delta t}(\beta_{\mathbf{I}}Ce^{\delta t}+1) + \frac{1}{\beta_{\mathbf{I}}} + \frac{\beta_{\mathbf{I}}^2H_f}{\delta}e^{3\delta t}\big\|\nabla f(x_{1,0})\big\| + \beta_{\mathbf{I}}\mathcal{L}_f(e^{2\delta t}+e^{\delta t}) + \beta_{\mathbf{I}}\mathcal{L}_f\Big)\big(e^{\delta t}-1\big) \\
&=: \phi\Big(t, \big\|\nabla f(x_{1,0})\big\|\Big),
\end{aligned}$$

where the inequality (i) follows from the implications of Steps 2 and 3, and the equality (ii) is an immediate consequence of the relation $\alpha < \delta$ and the equality $x_{2,0} = -\beta_{\mathbf{I}}\nabla f(x_{1,0})$.

Step 5: Consider a_1 defined in (8d) and recall that $u_{\mathbf{I}}(x_{,0})$ by design lies inside the input interval $[\underline{u}_{\mathbf{I}}, \overline{u}_{\mathbf{I}}]$. The quantity a_1 is a lower bound on the distance of $u_{\mathbf{I}}(x_{,0})$ to the boundaries of the interval $[\underline{u}_{\mathbf{I}}, \overline{u}_{\mathbf{I}}]$. Thus, the inter-jump interval $\tau_{\mathbf{I}}$ satisfies

$$\tau_{\mathbf{I}} \geq \max\big\{t\geq 0:\ \big|u_{\mathbf{I}}(x) - u_{\mathbf{I}}(x_{,0})\big| \leq a_1\big\} \geq \max\Big\{t\geq 0:\ \phi\Big(t, \big\|\nabla f(x_{1,0})\big\|\Big) \leq a_1\Big\},$$

where the second inequality is implied by the analysis provided in Step 4. Consider a positive constant $r > 1$. One can infer for every $t \in \big[0, \delta^{-1}\log r\big]$ that

$$\begin{aligned}
\phi\Big(t, \big\|\nabla f(x_{1,0})\big\|\Big) &\leq \Big(rL_f\delta^{-1}(r\beta_{\mathbf{I}}C+1) + \beta_{\mathbf{I}}^{-1} + r^3\beta_{\mathbf{I}}^2H_f\delta^{-1}\big\|\nabla f(x_{1,0})\big\| \\
&\qquad + (r^2+r)\beta_{\mathbf{I}}\mathcal{L}_f + \beta_{\mathbf{I}}\mathcal{L}_f\Big)(e^{\delta t}-1) \\
&= \Big(a_2 + a_3\big\|\nabla f(x_{1,0})\big\|\Big)(e^{\delta t}-1) \\
&=: \phi'\Big(t, \big\|\nabla f(x_{1,0})\big\|\Big),
\end{aligned}$$

where the constants a_2 and a_3 are defined in (8e), (8f), respectively, and the inequality $e^{\delta t} < r$ is used. Suppose now τ' is the lower bound of the inter jump in (7). Then $\phi'\Big(\tau', \big\|\nabla f(x_{1,0})\big\|\Big) = a_1$, where the constant a_1 is defined in (8d). It is straightforward to establish the assertion made in (7).

In the second part of the assertion, we should show that the proposed lower bound in (7) is uniformly away from zero along any trajectories of the hybrid system. To this end, we only need to focus on the term $\|\nabla f\big(x_1(t)\big)\|$. Recall that Theorem 3.1 effectively implies that $\lim_{t\to\infty} \|\nabla f\big(x_1(t)\big)\| = 0$, possibly not in a monotone manner though. This observation allows us to deduce that $M := \sup_{t\geq 0} \|\nabla f\big(x_1(t)\big)\| < \infty$. Using the uniform bound M, we have a minimum non-zero inter-jump interval, giving rise to a Zeno-free behavior for all solution trajectories.

4.2. **Proof of Theorem 3.4**

The proof follows a similar idea as in [19, Theorem 3.1] but the required technical steps are somewhat different, leading to another set of technical assumptions. In the first step, we begin with describing on how the chosen input $u_{\mathbf{II}}(x)$ in (9b) ensures achieving the desired exponential convergence rate $\mathcal{O}\big(e^{-\alpha t}\big)$. Let us define the set $\mathcal{E}_\alpha := \Big\{x \in \mathbb{R}^{2n} : \alpha\big(f(x_1) - f^*\big) < \langle \nabla f(x_1), -x_2\rangle\Big\}$. We demonstrate that as long as a solution trajectory of the continuous flow (9a) is contained in the set $\mathcal{E}_\alpha$, the function f obeys the exponential decay (5). To this end, observe that if $\big(x_1(t), x_2(t)\big) \in \mathcal{E}_\alpha$,

$$\frac{d}{dt}\Big(f\big(x_1(t)\big) - f^*\Big) = \big\langle \nabla f\big(x_1(t)\big), x_2(t)\big\rangle \leq -\alpha\big(f(x_1) - f^*\big).$$

The direct application of Gronwall's inequality, see [18, Lemma A.1], to the above inequality yields the desired convergence claim (5). Hence, it remains to guarantee that the solution trajectory renders the set $\mathcal{E}_\alpha$ invariant. Let us define the quantity

$$\sigma(t) := \langle \nabla f\big(x_1(t)\big), x_2(t)\rangle + \alpha\Big(f\big(x_1(t)\big) - f^*\Big).$$

By construction, if $\sigma(t) < 0$, it follows that $\big(x_1(t), x_2(t)\big) \in \mathcal{E}_\alpha$. As a result, if we synthesize the feedback input $u_{\mathbf{II}}(x)$ such that $\dot\sigma(t) \leq 0$ along the solution trajectory of (9a), the value of $\sigma(t)$ does not increase, and as such

$$\big(x_1(t), x_2(t)\big) \in \mathcal{E}_\alpha,\ \forall t \geq 0 \iff \big(x_1(0), x_2(0)\big) \in \mathcal{E}_\alpha.$$

To ensure non-positivity property of $\dot\sigma(t)$, note that we have

$$\begin{aligned}
\dot\sigma(x) &= \langle \nabla^2 f(x_1)x_2, x_2\rangle + \langle \nabla f(x_1), \dot x_2\rangle + \alpha\langle \nabla f(x_1), x_2\rangle \\
&= \langle \nabla^2 f(x_1)x_2, x_2\rangle + \langle \nabla f(x_1), -x_2 - u_{\mathbf{II}}(x)\nabla f(x_1)\rangle + \alpha\langle \nabla f(x_1), x_2\rangle \\
&= \langle \nabla^2 f(x_1)x_2, x_2\rangle + \langle \nabla f(x_1), -x_2\rangle - u_{\mathbf{II}}(x)\|\nabla f(x_1)\|^2 - \alpha\langle \nabla f(x_1), -x_2\rangle \\
&= \langle \nabla^2 f(x_1)x_2, x_2\rangle + (1-\alpha)\langle \nabla f(x_1), -x_2\rangle - u_{\mathbf{II}}(x)\|\nabla f(x_1)\|^2 = 0,
\end{aligned}$$

where the last equality follows from the definition of the proposed control law (9b). It is worth noting that one can simply replace the information of the Hessian $\nabla^2 f\big(x_1(t)\big)$ with the upper bound L_f and still arrive at the desired inequality, see also Remark 3.9 with regards to the 1st-order information oracle. Up to now, we showed that the structure of the control feedback guarantees the α-exponential convergence. It then remains to ensure that $x(0) \in \mathcal{E}_\alpha$. Consider the initial state $x_2(0) = -\beta_{\mathrm{II}}\nabla f\big(x_1(0)\big)$. Notice that

$$\begin{aligned}
\alpha\Big(f\big(x_1(0)\big) - f^*\Big) &\leq \frac{\alpha}{2\mu_f}\big\|\nabla f\big(x_1(0)\big)\big\|^2 \\
&= \frac{\alpha}{2\mu_f\beta_{\mathrm{II}}}\langle -x_2(0), \nabla f\big(x_1(0)\big)\rangle \\
&\leq \langle \nabla f\big(x_1(0)\big), -x_2(0)\rangle,
\end{aligned}$$

where in the first line we use (A2), and in the last line the condition (10c) is employed. Suppose $\big(x_1^\top(0), x_2^\top(0)\big)^\top$ as the jump state x^+. It is evident that the range space of the jump map (9d) lies inside the set $\mathcal{E}_\alpha$. At

last, it is required to show that the jump policy is well-defined in the sense that the trajectory lands in the interior of the flow set $\mathcal{C}_{\mathbf{I}}$ (9c), i.e., the control values also belong to the admissible set $[\underline{u}_{\mathbf{II}}, \overline{u}_{\mathbf{II}}]$. To this end, we only need to take into account the initial control value since the switching law is continuous in the states and serves the purpose by design. Suppose that $x^+ \in \mathcal{C}_{\mathbf{II}}$, we then have the sufficient requirements

$$\underline{u}_{\mathbf{II}} < \frac{-\ell_f \beta_{\mathrm{II}}^2 \|\nabla f(x_1^+)\|^2 + (1-\alpha)\beta_{\mathrm{II}}\|\nabla f(x_1^+)\|^2}{\|\nabla f(x_1^+)\|^2} \leq u_{\mathbf{II}}(x^+) \leq \frac{L_f \beta_{\mathrm{II}}^2 \|\nabla f(x_1^+)\|^2 + (1-\alpha)\beta_{\mathrm{II}}\|\nabla f(x_1^+)\|^2}{\|\nabla f(x_1^+)\|^2} < \overline{u}_{\mathbf{II}},$$

where the relations (9b) and (A1) are considered. Factoring out the term $\|\nabla f(x_1^+)\|^2$ leads to the sufficiency requirements given in (10a) and (10b). Hence, the claim of Theorem 3.4 follows.

4.3. **Proof of Theorem 3.5**

In order to facilitate the argument regarding the proof of Theorem 3.5, we begin with providing a lemma describing the norm-2 behaviors of $\langle \nabla f(x_1), -x_2\rangle$, x_2, and $\nabla f(x_1)$. For the sake of brevity, we employ the same notations used in Subsection 4.1, as well.

Lemma 4.3 (Growth bounds)**.** *Consider the continuous-time hybrid control system (H) with the respective parameters (9) satisfying (10) where the function f satisfies Assumptions (A1) and (A2). Suppose the hybrid control system is initiated from $\big(x_{1,0}, \beta_{\boldsymbol{II}}\nabla f(x_{1,0})\big)$ for some $x_{1,0} \in \mathbb{R}^n$. Then,*

$$\langle \nabla f(x_1), -x_2\rangle = \beta_{\boldsymbol{II}} e^{-\alpha t}\|\nabla f(x_{1,0})\|^2, \tag{23a}$$

$$\|x_2\| \leq D(t)\|\nabla f(x_{1,0})\|, \tag{23b}$$

$$\underline{\eta}(t)\|\nabla f(x_{1,0})\| \leq \|\nabla f(x_1)\| \leq \overline{\eta}(t)\|\nabla f(x_{1,0})\|, \tag{23c}$$

with the time-varying scalars D, $\underline{\eta}$, and $\overline{\eta}$ given by

$$D(t) := \Big(\beta_{\boldsymbol{II}}^2 e^{-2t} + \beta_{\boldsymbol{II}} U\big(1 - e^{-2t}\big)\Big)^{\frac{1}{2}}, \tag{24a}$$

$$\underline{\eta}(t) := 1 - \mathcal{L}_f(\beta_{\boldsymbol{II}}^2 + \beta_{\boldsymbol{II}} U)^{\frac{1}{2}} t, \tag{24b}$$

$$\overline{\eta}(t) := 1 + \mathcal{L}_f(\beta_{\boldsymbol{II}}^2 + \beta_{\boldsymbol{II}} U)^{\frac{1}{2}} t, \tag{24c}$$

respectively, where $U := \max\{\overline{u}_{\boldsymbol{II}}, -\underline{u}_{\boldsymbol{II}}\}$ and $\mathcal{L}_f := \max\{\ell_f, L_f\}$.

Proof. Considering the flow dynamics (9a) and the feedback input (9b), one obtains

$$\begin{aligned}
\frac{d}{dt}\langle \nabla f(x_1), -x_2\rangle &= \langle \nabla^2 f(x_1)x_2, -x_2\rangle + \langle \nabla f(x_1), -\dot{x}_2\rangle \\
&= \langle \nabla^2 f(x_1)x_2, -x_2\rangle + \langle \nabla f(x_1), x_2 + u_{\mathbf{II}}(x)\nabla f(x_1)\rangle \\
&= \langle \nabla^2 f(x_1)x_2, -x_2\rangle + \langle \nabla f(x_1), x_2\rangle + u_{\mathbf{II}}(x)\|\nabla f(x_1)\|^2 \\
&= \langle \nabla^2 f(x_1)x_2, -x_2\rangle + \langle \nabla f(x_1), x_2\rangle + \langle \nabla^2 f(x_1)x_2, x_2\rangle - (1-\alpha)\langle \nabla f(x_1), x_2\rangle \\
&= -\alpha\langle \nabla f(x_1), -x_2\rangle,
\end{aligned}$$

and as a result given the initial state $\big(x_{1,0}, -\beta_{\mathbf{II}}\nabla f(x_{1,0})\big)$, the equality given in (23a) is valid. We next turn to establish that (23b) holds. Let us define $h(t) = \|x_2\|^2$. Hence,

$$\begin{aligned}
\frac{d}{dt}h(t) &\overset{\text{(i)}}{=} 2\langle x_2, -x_2 - u_{\mathbf{II}}(x)\nabla f(x_1)\rangle = -2\|x_2\|^2 + 2u_{\mathbf{II}}(x)\langle \nabla f(x_1), -x_2\rangle \\
&\overset{\text{(ii)}}{=} -2h(t) + 2u_{\mathbf{II}}(x)\beta_{\mathbf{II}} e^{-\alpha t}\|\nabla f(x_{1,0})\|^2 \leq -2h(t) + 2U\beta_{\mathbf{II}}\|\nabla f(x_{1,0})\|^2,
\end{aligned}$$

where we made use of the flow dynamics (9a) in the inequality (i) and the equation (23a) in the equality (ii). We then use the Gronwall's inequality to infer that

$$\begin{aligned}\|x_2\|^2 &\le e^{-2t}\|x_{2,0}\|^2 + \int_0^t e^{-2(t-\tau)} 2U\beta_{\mathbf{II}} \big\|\nabla f(x_{1,0})\big\|^2 d\tau \\ &= e^{-2t}\beta_{\mathbf{II}}^2 \big\|\nabla f(x_{1,0})\big\|^2 + e^{-2t} 2U\beta_{\mathbf{II}} \big\|\nabla f(x_{1,0})\big\|^2 \int_0^t e^{2\tau} d\tau \\ &= e^{-2t}\big\|\nabla f(x_{1,0})\big\|^2 \Big(\beta_{\mathbf{II}}^2 e^{-2t} + \beta_{\mathbf{II}} U\big(1 - e^{-2t}\big)\Big) \\ &=: D^2(t)\big\|\nabla f(x_{1,0})\big\|^2,\end{aligned}$$

where $D(t)$ is defined in (24a). As a result, the claim in (23b) holds. The argument to show the last claim in Lemma 4.3 is discussed now. Let us define $g(t) := \big\|\nabla f(x_1)\big\|^2$. Observe that

$$\frac{d}{dt} g(t) = 2\langle \nabla^2 f(x_1) x_2, \nabla f(x_1)\rangle,$$

and as a result

$$\left|\frac{d}{dt} g(t)\right| \overset{\text{(i)}}{\le} 2\mathcal{L}_f \|x_2\| \cdot \big\|\nabla f(x_1)\big\| = 2\mathcal{L}_f \|x_2\| \sqrt{g(t)} \overset{\text{(ii)}}{\le} 2\mathcal{L}_f D(t) \big\|\nabla f(x_{1,0})\big\| \sqrt{g(t)},$$

where the inequalities (i) and (ii) are implied by Assumption (A1) and the inequality (23b), respectively. Hence, we deduce that

$$\frac{d}{dt} g(t) \ge -2\mathcal{L}_f D(t) \big\|\nabla f(x_{1,0})\big\| \sqrt{g(t)},$$

and as a consequence

$$\frac{dg(t)}{\sqrt{g(t)}} \ge -2\mathcal{L}_f D(t) \big\|\nabla f(x_{1,0})\big\| dt.$$

Integrating the two sides of the above inequality results in

$$\begin{aligned}\sqrt{g(t)} - \sqrt{g(0)} &\ge -\mathcal{L}_f \big\|\nabla f(x_{1,0})\big\| \int_0^t D(\tau) d\tau \\ &= -\mathcal{L}_f \big\|\nabla f(x_{1,0})\big\| \int_0^t \Big(\beta_{\mathbf{II}}^2 e^{-2\tau} + \beta_{\mathbf{II}} U\big(1 - e^{-2\tau}\big)\Big)^{\frac{1}{2}} d\tau \\ &\ge -\mathcal{L}_f \big\|\nabla f(x_{1,0})\big\| \int_0^t \big(\beta_{\mathbf{II}}^2 + \beta_{\mathbf{II}} U\big)^{\frac{1}{2}} d\tau \\ &= -\mathcal{L}_f \big\|\nabla f(x_{1,0})\big\| \big(\beta_{\mathbf{II}}^2 + \beta_{\mathbf{II}} U\big)^{\frac{1}{2}} t.\end{aligned}$$

Based on the above analysis and the definition of $g(t)$, it follows that

$$\big\|\nabla f(x_1)\big\| \ge \underline{\eta}(t) \big\|\nabla f(x_{1,0})\big\|,$$

where $\underline{\eta}(t)$ is given in (24b). Proceeding with a similar approach to the one presented above, one can use the inequality

$$\frac{d}{dt} g(t) \le 2\mathcal{L}_f D(t) \big\|\nabla f(x_{1,0})\big\| \sqrt{g(t)},$$

and infer that

$$\big\|\nabla f(x_1)\big\| \le \overline{\eta}(t) \big\|\nabla f(x_{1,0})\big\|,$$

where $\overline{\eta}(t)$ is defined in (24c). Thus, the last claim in Lemma 4.3 also holds. □

Proof of Theorem 3.5: We are now in a position to formally state the proof of Theorem 3.5. Consider the parameter δ as defined in Theorem 3.5. Intuitively, this quantity represents a lower bound on the distance of $u_{\mathbf{II}}(0)$ from the endpoints of the flow set interval. Thus, one can obtain a lower bound on the inter-jump interval $\tau_{\mathbf{II}}$ as follows

$$\tau_{\mathbf{II}} \geq \sup\, \{t > 0 :\ |u_{\mathbf{II}}(t) - u_{\mathbf{II}}(0)| \leq \delta\}. \tag{25}$$

On the other hand, given the structure of $u_{\mathbf{II}}$ in (9b),

$$-\frac{\ell_f \|x_2\|^2}{\|\nabla f(x_1)\|^2} + (1-\alpha)\frac{\beta_{\mathbf{II}} e^{-\alpha t}\|\nabla f(x_{1,0})\|^2}{\|\nabla f(x_1)\|^2} \leq u_{\mathbf{II}}(t) \leq \frac{L_f \|x_2\|^2}{\|\nabla f(x_1)\|^2} + (1-\alpha)\frac{\beta_{\mathbf{II}} e^{-\alpha t}\|\nabla f(x_{1,0})\|^2}{\|\nabla f(x_1)\|^2},$$

since the function f satisfies Assumption (A1). In light of Lemma 4.3 and considering the above relation, one can infer that for $\alpha \leq 1$, we name Case(i),

$$\underline{e}(t) := -\frac{\ell_f D(t)^2}{\underline{\eta}(t)^2} + (1-\alpha)\frac{\beta_{\mathbf{II}} e^{-\alpha t}}{\overline{\eta}(t)^2} \leq u_{\mathbf{II}}(t) \leq \frac{L_f D(t)^2}{\underline{\eta}(t)^2} + (1-\alpha)\frac{\beta_{\mathbf{II}} e^{-\alpha t}}{\underline{\eta}(t)^2} =: \overline{e}(t), \tag{26a}$$

and that for $\alpha > 1$, we denote by Case (ii),

$$\underline{p}(t) := -\frac{\ell_f D(t)^2}{\underline{\eta}(t)^2} + (1-\alpha)\frac{\beta_{\mathbf{II}} e^{-\alpha t}}{\underline{\eta}(t)^2} \leq u_{\mathbf{II}}(t) \leq \frac{L_f D(t)^2}{\underline{\eta}(t)^2} + (1-\alpha)\frac{\beta_{\mathbf{II}} e^{-\alpha t}}{\overline{\eta}(t)^2} =: \overline{p}(t). \tag{26b}$$

According to the above discussion, we employ (26) to obtain a lower bound $\tau_{\mathbf{II}}$ instead of using (25). Consider a time instant $t_\circ$ such that $t_\circ < 1/\omega$ where ω is defined in Theorem 3.5.

Case (i) ($\alpha \leq 1$): Let us denote $\sup_{t\in[0,t_\circ]} \dot{\overline{e}}(t)$ by b_1. Observe that

$$\begin{aligned}
\dot{\overline{e}}(t) &= \frac{2L_f\beta_{\mathbf{II}} e^{-2t}(-\beta_{\mathbf{II}} + U)(1-\omega t)^2 + 2\omega(1-\omega t)L_f\beta_{\mathbf{II}}\big(\beta_{\mathbf{II}} e^{-2t} + U(1-e^{-2t})\big)}{(1-\omega t)^4} \\
&\quad + (1-\alpha)\frac{-\alpha\beta_{\mathbf{II}} e^{-\alpha t}(1-\omega t)^2 + 2\omega(1-\omega t)\beta_{\mathbf{II}} e^{-2t}}{(1-\omega t)^4} \\
&\leq \frac{2L_f\beta_{\mathbf{II}} U e^{-2t}(1-\omega t)^2 + 2\omega(1-\omega t)L_f\beta_{\mathbf{II}}\big(\beta_{\mathbf{II}} e^{-2t} + U(1-e^{-2t})\big)}{(1-\omega t)^4} \\
&\quad + (1-\alpha)\frac{2\omega(1-\omega t)\beta_{\mathbf{II}} e^{-2t}}{(1-\omega t)^4} \\
&\leq \frac{2L_f\beta_{\mathbf{II}}\big(U + \omega(\beta_{\mathbf{II}} + U)\big)}{(1-\omega t)^3} + (1-\alpha)\frac{2\omega\beta_{\mathbf{II}}}{(1-\omega t)^3} \\
&\leq \frac{2L_f\beta_{\mathbf{II}}\big(U + \omega(\beta_{\mathbf{II}} + U)\big)}{(1-\omega t_\circ)^3} + (1-\alpha)\frac{2\omega\beta_{\mathbf{II}}}{(1-\omega t_\circ)^3} =: b_1,
\end{aligned}$$

considering (26a). Hence, $\overline{e}(t) \leq b_1 t + \overline{e}(0)$ and as a result

$$\tau_{\mathbf{II}} \geq \max\{t \in (0, t_\circ] :\ b_1 t + \overline{e}(0) - \overline{e}(0) \leq \delta\} = \min\{t_\circ, \delta/b_1\}, \tag{27}$$

by virtue of the fact that $b_1 t + \overline{e}(0)$ is a monotonically increasing function that upper bounds $u_{\mathbf{II}}(t)$. Now, let us define $b_2 := \inf_{t\in(0,t_\circ]} \dot{\underline{e}}(t)$. Notice that

$$\begin{aligned}
\dot{\underline{e}}(t) &= \frac{2\ell_f\beta_{\mathbf{II}} e^{-2t}(\beta_{\mathbf{II}} - U)(1-\omega t)^2 - 2\omega(1-\omega t)\ell_f\beta_{\mathbf{II}}\big(\beta_{\mathbf{II}} e^{-2t} + U(1-e^{-2t})\big)}{(1-\omega t)^4} \\
&\quad + (1-\alpha)\frac{-\alpha\beta_{\mathbf{II}} e^{-\alpha t}(1+\omega t)^2 - 2\omega(1+\omega t)\beta_{\mathbf{II}} e^{-2t}}{(1+\omega t)^4} \\
&\geq \frac{-2\ell_f\beta_{\mathbf{II}} e^{-2t} U(1-\omega t)^2 - 2\omega(1-\omega t)\ell_f\beta_{\mathbf{II}}\big(\beta_{\mathbf{II}} e^{-2t} + U(1-e^{-2t})\big)}{(1-\omega t)^4} \\
&\quad - (1-\alpha)\frac{\alpha\beta_{\mathbf{II}} e^{-\alpha t}(1+\omega t)^2 + 2\omega(1+\omega t)\beta_{\mathbf{II}} e^{-2t}}{(1+\omega t)^4}
\end{aligned}$$

$$\geq -\frac{2\ell_f\beta_{\mathbf{II}}\big(U+\omega(\beta_{\mathbf{II}}+U)\big)}{(1-\omega t_\circ)^3} - (1-\alpha)\frac{\alpha\beta_{\mathbf{II}}(1+\omega t_\circ)+2\omega\beta_{\mathbf{II}}}{1} =: -b_2.$$

Thus, $\underline{e}(t) \geq -b_2 t + \underline{e}(0)$ and as a consequence

$$\tau_{\mathbf{II}} \geq \max\{t\in(0,t_\circ]:\ \underline{e}(0) - \big(-b_2 t + \underline{e}(0)\big) \leq \delta\} = \min\{t_\circ, \delta/b_2\}, \tag{28}$$

because the function $-b_2 t + \underline{e}(0)$ is a monotonically decreasing function that lower bounds $u_{\mathbf{II}}(t)$.

Case (ii) ($\alpha > 1$): Much of this case follows the same line of reasoning used in Case (i). We thus provide only main mathematical derivations and refer the reader to the previous case for the argumentation. Define $b_3 := \sup_{t\in(0,t_\circ]} \dot{\overline{p}}(t)$. One can deduce from (26b) that

$$\begin{aligned}\dot{\overline{p}}(t) &= \frac{2L_f\beta_{\mathbf{II}}e^{-2t}(-\beta_{\mathbf{II}}+U)(1-\omega t)^2 + 2\omega(1-\omega t)L_f\beta_{\mathbf{II}}\big(\beta_{\mathbf{II}}e^{-2t}+U(1-e^{-2t})\big)}{(1-\omega t)^4}\\ &\quad + (1-\alpha)\frac{-\alpha\beta_{\mathbf{II}}e^{-\alpha t}(1+\omega t)^2 - 2\omega(1+\omega t)\beta_{\mathbf{II}}e^{-2t}}{(1+\omega t)^4}\\ &\leq \frac{2L_f\beta_{\mathbf{II}}\big(U+\omega(\beta_{\mathbf{II}}+U)\big)}{(1-\omega t_\circ)^3} + (\alpha-1)\frac{\alpha\beta_{\mathbf{II}}(1+\omega t_\circ)+2\omega\beta_{\mathbf{II}}}{1} =: b_3.\end{aligned}$$

Hence, $\overline{p}(t) \leq b_4 t + \overline{p}(0)$ and as a result

$$\tau \geq \min\{t_\circ, \delta/b_3\}. \tag{29}$$

Finally, define $\dot{\underline{p}}(t) := \inf_{t\in(0,t_\circ]} \underline{p}(t)$ from which it follows that

$$\begin{aligned}\dot{\underline{p}}(t) &= \frac{2\ell_f\beta_{\mathbf{II}}e^{-2t}(\beta_{\mathbf{II}}-U)(1-\omega t)^2 - 2\omega(1-\omega t)\ell_f\beta_{\mathbf{II}}\big(\beta_{\mathbf{II}}e^{-2t}+U(1-e^{-2t})\big)}{(1-\omega t)^4}\\ &\quad + (1-\alpha)\frac{-\alpha\beta_{\mathbf{II}}e^{-\alpha t}(1-\omega t)^2 + 2\omega(1-\omega t)\beta_{\mathbf{II}}e^{-2t}}{(1-\omega t)^4}\\ &\geq -\frac{2\ell_f\beta_{\mathbf{II}}\big(U+\omega(\beta_{\mathbf{II}}+U)\big)}{(1-\omega t_\circ)^3} - (\alpha-1)\frac{2\omega\beta_{\mathbf{II}}}{(1-\omega t_\circ)^3} =: -b_4,\end{aligned}$$

considering (26b). Now, since $\underline{p}(t) \geq -b_4 t + \underline{p}(0)$, it is implied that

$$\tau_{\mathbf{II}} \geq \min\{t_\circ, \delta/b_4\}. \tag{30}$$

Notice that based on the relations derived in (28)-(30),

$$\tau_{\mathbf{II}} \geq \min\Big\{t_\circ, \frac{2\mathcal{L}_f\beta_{\mathbf{II}}\big(U+\omega(\beta_{\mathbf{II}}+U)\big)}{(1-\omega t_\circ)^3} + |\alpha-1|\frac{2\omega\beta_{\mathbf{II}}}{(1-\omega t_\circ)^3} + |\alpha-1|\alpha\beta_{\mathbf{II}}(1+\omega t_\circ)\Big\}.$$

Suppose now for some scalar $r\in(0,1)$, $t_\circ$ is chosen such that $t_\circ \leq \frac{r}{\omega}$. It is evident that

$$\tau_{\mathbf{II}} \geq \min\Big\{\frac{r}{\omega}, \delta\Big/\Big(\frac{2\mathcal{L}_f\beta_{\mathbf{II}}\big(U+\omega(\beta_{\mathbf{II}}+U)\big)}{(1-r)^3} + |\alpha-1|\frac{2\omega\beta_{\mathbf{II}}}{(1-r)^3} + |\alpha-1|\alpha\beta_{\mathbf{II}}(1+r)\Big)\Big\}.$$

It turns out that the relation (11) in Theorem 3.5 is valid and this concludes the proof.

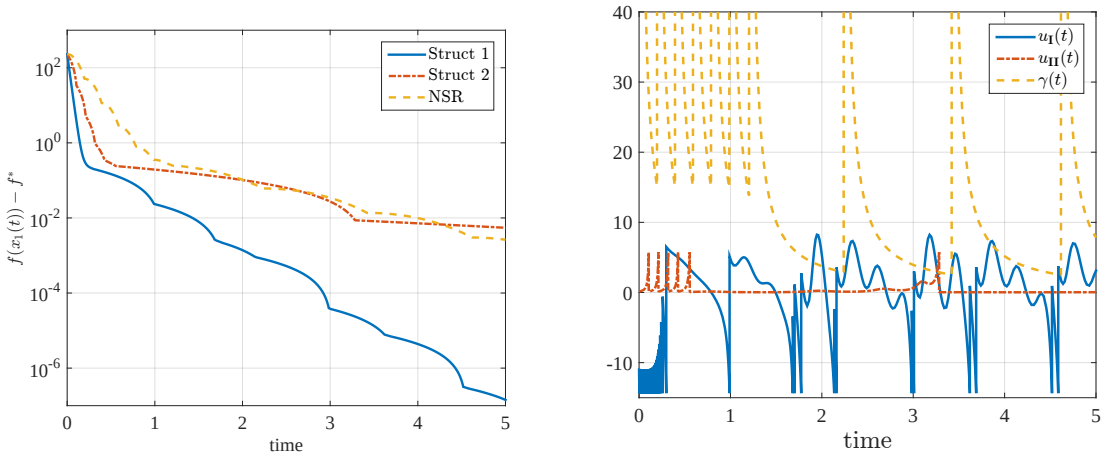
4.4. **Proof of Theorem 3.11**

In what follows, we provide the proof for the structure **II** and refer the interested reader to [19, Theorem 3.7] for the structure **I**. We emphasize that the technical steps to establish a stable discretization for both structures are similar.

According to the forward-Euler method, the velocity $\dot{x}_1$ and the acceleration $\dot{x}_2$ in the dynamics ($\mathcal{H}$) with (9) are discretized as follows:

$$\frac{x_1^{k+1}-x_1^k}{s} = x_2^k, \tag{31a}$$

$$\frac{x_2^{k+1}-x_2^k}{s} = -u_{d,\mathbf{II}}(x^k)\nabla f(x_1^k) - x_2^k, \tag{31b}$$



(a) Objective value along system trajectories.

(b) State-dependent and time-varying coefficients.

FIGURE 1. Continuous-time dynamics of **Struct I**, **Struct II**, **NSR**.

where the discrete input $u_{d,\mathbf{II}}(x^k) = u_{\mathbf{II}}(x^k)$. Now, observe that the definition of the flow set $\mathcal{C}_{d,\mathbf{II}}$ (13b) implies

$$c_1\|x_2^k\|^2 \leq \|\nabla f(x_1^k)\|^2 \leq c_2\langle \nabla f(x_1^k), -x_2^k\rangle \leq c_2\|\nabla f(x_1^k)\| \cdot \|x_2^k\|,$$

where the extra inequality follows from the Cauchy-Schwarz inequality ($\forall\ a, b \in \mathbb{R}^n$, $\langle a, b\rangle \leq \|a\| \cdot \|b\|$). In order to guarantee that the flow set $\mathcal{C}_{d,\mathbf{II}}$ is non-empty the relation (16a) should hold between the parameters c_1 and c_2 since $\sqrt{c_1} \leq \frac{\|\nabla f(x_1^k)\|}{\|x_2^k\|} \leq c_2$. Next, suppose that the parameters c_1, c_2, and β satisfy (16b). Multiplying (16b) by $\|\nabla f(x_1^k)\|$, one can observe that the range space of the jump map $G_{d,\mathbf{II}}(x^k) = \big((x^k)^\top, -\beta\nabla^\top f(x^k)\big)^\top$ is inside the flow set $\mathcal{C}_{d,\mathbf{II}}$ (13b). From the fact that the discrete dynamics (12) evolves respecting the flow set $\mathcal{C}_{d,\mathbf{II}}$ defined in (13b), we deduce

$$\begin{aligned}
f(x_1^{k+1}) - f(x_1^k) &\leq \langle \nabla f(x_1^k), x_1^{k+1} - x_1^k\rangle + \frac{L_f}{2}\|x_1^{k+1} - x_1^k\|^2\\
&\leq -s\langle \nabla f(x_1^k), -x_2^k\rangle + \frac{L_f s^2}{2}\|x_2^k\|^2\\
&< -\frac{s}{c_2}\|\nabla f(x_1^k)\|^2 + \frac{L_f s^2}{2c_1}\|\nabla f(x_1^k)\|^2\\
&= \big(-\frac{s}{c_2} + \frac{L_f}{2c_1}s^2\big)\|\nabla f(x_1^k)\|^2\\
&\leq 2\mu_f\big(-\frac{s}{c_2} + \frac{L_f}{2c_1}s^2\big)\big(f(x_1^k) - f^*\big),
\end{aligned}$$

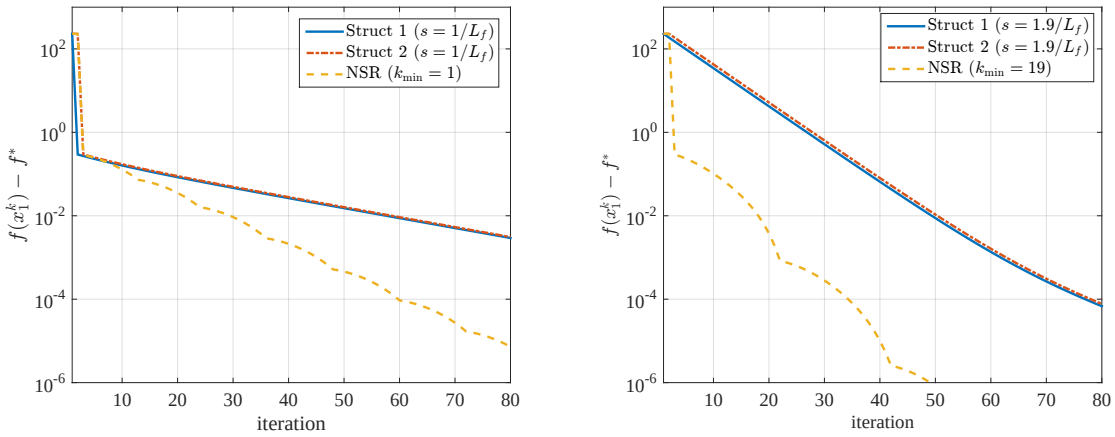
where we made use of the relation (3), the definition (31a), the relation (13b), and the assumption (A2), respectively. Then, considering the inequality implied by the first and last terms given above and adding $f(x_1^k) - f^*$ to both sides of the considered inequality, we arrive at

$$f(x_1^{k+1}) - f^* \leq \lambda(s, c_1, c_2, \beta)\,\big(f(x_1^k) - f^*\big)$$

where $\lambda(s, c_1, c_2, \beta)$ is given by (15). As a result, if the step size s is chosen such that $s < \frac{2c_1}{c_2 L_f}$ then $\lambda(s, c_1, c_2, \beta) \in (0, 1)$. The claim of Theorem 3.11 follows.

5. NUMERICAL EXAMPLES

In this section a numerical example illustrating the results in this paper is represented. The example is a least mean square error (LMSE) problem $f(x_1) = \|Ax_1 - b\|^2$ where $x \in \mathbb{R}^5$ denotes the decision variable,



(a) Standard tuning parameters.

(b) Example-based optimal tuning parameters.

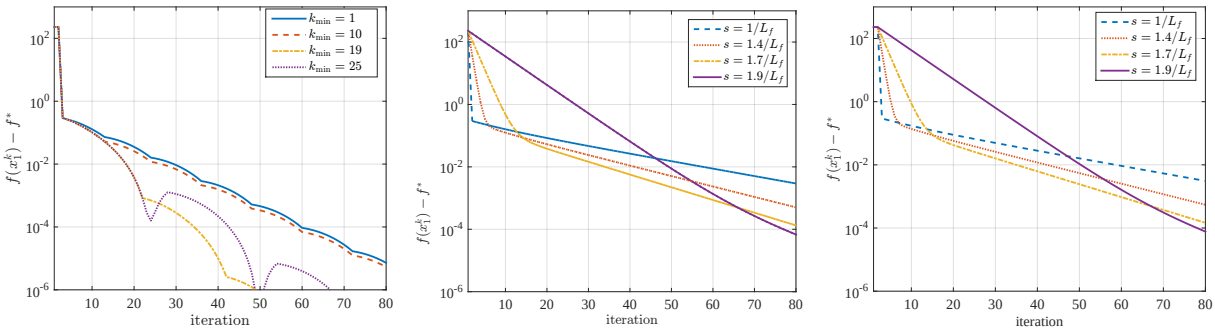
FIGURE 2. Discrete-time dynamics of **Struct I**, **Struct II**, **NSR**.

$A \in \mathbb{R}^{50\times 5}$ with $L_f = 2\lambda_{\max}(A^\top A) = 136.9832$ and $\mu_f = 2\lambda_{\min}(A^\top A) = 3.6878$, and $b \in \mathbb{R}^{50}$. Since the LMSE function is convex (in our case, this function is strongly convex), we take $\ell_f = 0$. In what follows, we compare the behaviors of the proposed structures **I** and **II** (denoted by **Struct I** and **Struct II**, respectively) along with Nesterov's fast method with the speed restarting scheme proposed by [36] (denoted by **NSR**). We begin with providing the results concerning the continuous-time case. Then, the discrete-time case's results are shown where we employ Algorithm 1 for **Struct I** and **Struct II**.

Continuous-time case: The corresponding parameters of **Struct I** and **Struct II** are as follows: $\alpha_I = 0.2$, $\beta_I = 0.1356$, $\underline{u}_{\mathbf{I}} = -14.352$, $\overline{u}_{\mathbf{I}} = 15.1511$; $\alpha_{II} = 0.2$, $\beta_{II} = 0.0298$, $\underline{u}_{\mathbf{II}} = -0.1861$, $\overline{u}_{\mathbf{II}} = 5.7457$. In Figure 1(a), the behaviors of the suboptimality measure $f\big(x_1(t)\big) - f^*$ of **Struct I**, **Struct II**, and **NSR** are depicted. With regards to Theorem 3.2, observe that the length of inter-jump intervals is small during the early stages of simulation. As time progresses and the value of $\nabla f(x_1)$ decreases, the length of inter-jump intervals relatively increases (echoing the same message conveyed in Theorem 3.2). The corresponding control inputs are represented in Figure 1(b). Furthermore, in the case of **Struct I** where $u_{\mathbf{I}}$ plays the role of damping, the input $u_{\mathbf{I}}$ admits a negative range unlike most of the approaches in the literature.

Discrete-time case: Figure 2(a) shows the discrete-time counterparts of the previously mentioned continuous-time dynamics in Figure 1. It is evident that the discrete counterparts of our proposed structures perform poorly compared to the NSR's discrete counterpart, reinforcing the assertion of Remark 3.12 calling for a smarter discretization technique. The results depicted in Figure 2(a) correspond to the standard parameters involved in each of the algorithm, i.e., the step size $s = 1/L_f$ for the proposed methods in Corollary 3.13, and the parameter $k_{\min} = 1$ in NSR. However, these parameters can also be tuned depending on the application at hand. In case of NSR, the role of the parameter $k_{\min}$ is to prevent unnecessary restarting instants that may degrade the overall performance. On the other hand, setting $k_{\min} > 1$ may potentially cause the algorithm to lose its monotonicity property. In Figure 2(b), we illustrate the best behavior of the three methods with respect to these parameters for this numerical example.

Finally, Figure 3(a) shows how changing $k_{\min}$ affects the performance. The best performance is achieved by setting $k_{\min} = 19$ and the algorithm becomes non-monotonic for $k_{\min} > 19$. With regards to our proposed methods we observe that if one increases the step size s, the performance improves, see Figure 3(b) for **Struct I** and Figure 3(c) for **Struct II**. Moreover, it is obvious that the discrete-time couterparts of **Struct I** and **Struct II** behave in a very similar fashion that has to do with the lack of a proper discretization that can fully exploit the properties of the corresponding feedback input, see Remark 3.12.



(a) **NSR** (b) **Struct I** (c) **Struct II**

FIGURE 3. Discrete-time dynamics under different tuning parameters.

6. CONCLUSIONS

Inspired by a control-oriented viewpoint, we proposed two hybrid dynamical structures to achieve exponential convergence rates for a certain class of unconstrained optimization problems, in a continuous-time setting. The distinctive feature of our methodology is the synthesis of certain inputs in a state-dependent fashion compared to a time-dependent approach followed by most results in the literature. Due to the state-dependency of our proposed methods, the time-discretization of continuous-time hybrid dynamical systems is in fact difficult (and to some extent even more involved than the time-varying dynamics that is commonly used in the literature). In this regard, we have been able to show that one can apply the the forward-Euler method to discretize the continuous-time dynamics and still guarantee exponential rate of convergence. Thus, a more in-depth analysis is due. We expect that because of the state-dependency of our methods a proper venue to search is geometrical types of discretization.

REFERENCES

[1] Z. ALLEN-ZHU, *Katyusha: The first direct acceleration of stochastic gradient methods*, arXiv preprint arXiv:1603.05953, (2016).

[2] J.-P. AUBIN, J. LYGEROS, M. QUINCAMPOIX, S. SASTRY, AND N. SEUBE, *Impulse differential inclusions: a viability approach to hybrid systems*, IEEE Transactions on Automatic Control, 47 (2002), pp. 2–20.

[3] S. BECKER, J. BOBIN, AND E. J. CANDÈS, *Nesta: A fast and accurate first-order method for sparse recovery*, SIAM Journal on Imaging Sciences, 4 (2011), pp. 1–39.

[4] L. BOTTOU, *Stochastic gradient learning in neural networks*, Proceedings of Neuro-Nımes, 91 (1991).

[5] S. BUBECK, Y. T. LEE, AND M. SINGH, *A geometric alternative to Nesterov's accelerated gradient descent*, arXiv preprint arXiv:1506.08187, (2015).

[6] A. CABOT, *The steepest descent dynamical system with control. applications to constrained minimization*, ESAIM: Control, Optimisation and Calculus of Variations, 10 (2004), pp. 243–258.

[7] B. D. CRAVEN AND B. M. GLOVER, *Invex functions and duality*, Journal of the Australian Mathematical Society, 39 (1985), pp. 1–20.

[8] Y. DRORI AND M. TEBOULLE, *Performance of first-order methods for smooth convex minimization: a novel approach*, Mathematical Programming, 145 (2014), pp. 451–482.

[9] D. DRUSVYATSKIY, M. FAZEL, AND S. ROY, *An optimal first order method based on optimal quadratic averaging*, arXiv preprint arXiv:1604.06543, (2016).

[10] M. FAZLYAB, A. RIBEIRO, M. MORARI, AND V. M. PRECIADO, *Analysis of optimization algorithms via integral quadratic constraints: Non-strongly convex problems*, arXiv preprint arXiv:1705.03615, (2017).

[11] E. GHADIMI, I. SHAMES, AND M. JOHANSSON, *Multi-step gradient methods for networked optimization*, IEEE Transactions on Signal Processing, 61 (2013), pp. 5417–5429.

[12] R. GOEBEL, R. G. SANFELICE, AND A. R. TEEL, *Hybrid dynamical systems: modeling, stability, and robustness*, Princeton University Press, 2012.

[13] R. GOEBEL AND A. R. TEEL, *Solutions to hybrid inclusions via set and graphical convergence with stability theory applications*, Automatica, 42 (2006), pp. 573–587.

[14] M. Gu, L.-H. Lim, and C. J. Wu, *Parnes: a rapidly convergent algorithm for accurate recovery of sparse and approximately sparse signals*, Numerical Algorithms, 64 (2013), pp. 321–347.
[15] M. A. Hanson, *On sufficiency of the Kuhn-Tucker conditions*, Journal of Mathematical Analysis and Applications, 80 (1981), pp. 545–550.
[16] B. Hu and L. Lessard, *Dissipativity theory for Nesterov's accelerated method*, arXiv preprint arXiv:1706.04381, (2017).
[17] H. Karimi, J. Nutini, and M. Schmidt, *Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition*, Springer International Publishing, 2016, pp. 795–811.
[18] H. S. Khalil, *Nonlinear systems*, Prentice Hall, 3rd ed., 2002.
[19] A. S. Kolarijani, P. Mohajerin Esfahani, and T. Keviczky, *Fast gradient-based methods with exponential rate: A hybrid control framework*, in Proceedings of the 35th International Conference on Machine Learning (ICML 2018), 2018.
[20] G. Lan and R. Monteiro, *Iteration-complexity of first-order penalty methods for convex programming*, Mathematical Programming, 138 (2013), pp. 115–139.
[21] D. Lashkari and P. Golland, *Convex clustering with exemplar-based models*, in Advances in Neural Information Processing Systems (NIPS 2008), 2008, pp. 825–832.
[22] L. Lessard, B. Recht, and A. Packard, *Analysis and design of optimization algorithms via integral quadratic constraints*, SIAM Journal on Optimization, 26 (2016), pp. 57–95.
[23] J. Lygeros, K. H. Johansson, S. N. Simic, J. Zhang, and S. Sastry, *Dynamical properties of hybrid automata*, IEEE Transactions on automatic control, 48 (2003), pp. 2–17.
[24] A. Megretski and A. Rantzer, *System analysis via integral quadratic constraints*, IEEE Transactions on Automatic Control, 42 (1997), pp. 819–830.
[25] A. Nemirovski, *Efficient methods in convex programming*, (2005).
[26] A. Nemirovskii, D. B. Yudin, and E. R. Dawson, *Problem complexity and method efficiency in optimization*, (1983).
[27] Y. Nesterov, *A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$*, in Soviet Mathematics Doklady, vol. 27, 1983, pp. 372–376.
[28] ———, *Introductory lectures on convex optimization: a basic course*, Springer Science and Business Media, 2004.
[29] ———, *Smooth minimization of non-smooth functions*, Mathematical Programming, 103 (2005), pp. 127–152.
[30] ———, *Gradient methods for minimizing composite functions*, Mathematical Programming, 140 (2013), pp. 125–161.
[31] Y. Nesterov and B. T. Polyak, *Cubic regularization of newton method and its global performance*, Mathematical Programming, 108 (2006), pp. 177–205.
[32] B. O'Donoghue and E. Candès, *Adaptive restart for accelerated gradient schemes*, Foundations of Computational Mathematics, 15 (2015), pp. 715–732.
[33] B. T. Polyak, *Some methods of speeding up the convergence of iteration methods*, USSR Computational Mathematics and Mathematical Physics, 4 (1964), pp. 1–17.
[34] R. Salakhutdinov, S. T. Roweis, and Z. Ghahramani, *Optimization with em and expectation-conjugate-gradient*, in Proceedings of the 20th International Conference on Machine Learning (ICML 2003), 2003, pp. 672–679.
[35] W. Su, S. Boyd, and E. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, in Advances in Neural Information Processing Systems (NIPS 2014), 2014, pp. 2510–2518.
[36] ———, *A differential equation for modeling Nesterov's accelerated gradient method: theory and insights*, Journal of Machine Learning Research, 17 (2016), pp. 1–43.
[37] A. Wibisono, A. C. Wilson, and M. I. Jordan, *A variational perspective on accelerated methods in optimization*, Proceedings of the National Academy of Sciences, 113 (2016), pp. E7351–E7358.
[38] J. C. Willems, *Dissipative dynamical systems part i: General theory*, Archive for Rational Mechanics and Analysis, 45 (1972), pp. 321–351.
[39] A. C. Wilson, B. Recht, and M. I. Jordan, *A Lyapunov analysis of momentum methods in optimization*, arXiv preprint arXiv:1611.02635, (2016).