On the Absence of Spurious Local Trajectories in Time-varying Nonconvex Optimization

Salar Fattahi, Cedric Josz, Yuhao Ding, Reza Mohammadi, Javad Lavaei, and Somayeh Sojoudi

***Abstract*—In this paper, we study the landscape of an online nonconvex optimization problem, for which the input data vary over time and the solution is a trajectory rather than a single point. To understand the complexity of finding a global solution of this problem, we introduce the notion of *spurious (i.e., non-global) local trajectory* as a generalization to the notion of spurious local solution in nonconvex (time-invariant) optimization. We develop an ordinary differential equation (ODE) associated with a time-varying nonlinear dynamical system which, at limit, characterizes the spurious local solutions of the time-varying optimization problem. We prove that the absence of spurious local trajectory is closely related to the transient behavior of the developed system. In particular, we show that if the problem is time-varying, the data variation may force all of the ODE trajectories initialized at arbitrary local minima at the initial time to gradually converge to the global solution trajectory. We study the Jacobian of the dynamical system along a local minimum trajectory and show how its eigenvalues are manipulated by the natural data variation in the problem, which may consequently trigger escaping poor local minima over time.**

I. Introduction

Sequential decision making with time-varying data is at the core of most of today's problems. For example, the optimal power flow (OPF) problem in the electrical grid should be solved every 5 minutes in order to match the supply of electricity with a demand profile that changes over time [2]. Other examples include the training of dynamic neural networks [3], dynamic matrix recovery [4], [5], time-varying multi-armed bandit problem [6], robot navigation and obstacle avoidance [7], and many other applications [8]. Indeed, most of these problems are large-scale and should be solved in real-time, which strongly motivates the need for practical algorithms in such optimization frameworks.

A recent line of work has shown that a surprisingly large class of data-driven and nonconvex optimization problems—including matrix completion/sensing, phase retrieval, and dictionary learning, robust principal component analysis—has a *benign landscape*, i.e., every local solution is also global [9]–[12].[1] A local solution that is not globally optimal is called *spurious*. At the crux of the results on the absence of spurious local minima is the assumption on the static and time-invariant nature of the optimization. Yet, in practice, many real-world and data-driven problems are time-varying and require online optimization. This observation naturally gives rise to the following question:

Would simple local-search algorithms escape spurious local minima in online nonconvex optimization, similar to their time-invariant counterparts?

In this paper, we attempt to address this question by developing a control-theoretic framework for analyzing the landscape of online and time-varying optimization. In particular, we demonstrate that even if a time-varying optimization may have undesired point-wise local minima at almost all times, the variation of its landscape over time would enable simple local-search algorithms to escape these spurious local minima. Inspired by this observation, this paper provides a new machinery to analyze the global landscape of online decision-making problems by drawing tools from optimization and control theory.

We consider a class of nonconvex and online optimization problems where the input data vary over time. First, we introduce the notion of *spurious local trajectory* as a generalization to the point-wise spurious local solutions. Roughly speaking, a solution trajectory is called spurious if it does not belong to the region of attraction of a global solution of the problem (see Section III for a formal definition). We show that a time-varying optimization can have point-wise spurious local minima at every time step, and yet, it can be free of spurious local trajectories. By building upon this notion, we consider a general class of nonconvex optimization problems and model their local trajectories via an ordinary differential equation (ODE) representing a time-varying nonlinear dynamical system. We show that the absence of the spurious local trajectories in this time-varying optimization is equivalent to the convergence of all solutions in its corresponding ODE. Based on this equivalence, we analyze different classes of time-varying optimization problems and present sufficient conditions under which, despite possibly having point-wise spurious local minima at all times, the time-varying problem is free of spurious local trajectories. This implies that the time-varying nature of the problem is essential for the absence of spurious local trajectories. Finally, we analyze the Jacobian of the ODE along a local minimum trajectory and show how its eigenvalues are manipulated by the data variation.

A. Related Works

Benign landscape: Nonconvexity is inherent to many problems in machine learning; from the classical compressive

A shorter version of this paper has been submitted for conference publication [1]. Sections II-B, V and VI are the new additions to the paper.

Salar Fattahi is with University of Michigan, Ann Arbor (email: fattahi@umich.edu). Cedric Josz is with Columbia University (email: cj2638@columbia.edu). Yuhao Ding, Reza Mohammadi, Javad Lavaei, and Somayeh Sojoudi are with the University of California, Berkeley (email: yuhao_ding@berkeley.edu, mohammadi@berkeley.edu, lavaei@berkeley.edu, sojoudi@berkeley.edu).

[1]A local solution is a point that satisfies the first-order optimality conditions. Moreover, a global solution is a point that has the best overall objective value; see Sections III and V for more details.

sensing and matrix completion/sensing [13]–[15] to the more recent problems on the training of deep neural networks [16], they often possess nonconvex landscapes. Reminiscent from the classical complexity theory, this nonconvexity is perceived to be the main contributor to the intractability of these problems. In many (albeit not all) cases, this intractability implies that in the worst-case instances of the problem, spurious local minima exist and there is no efficient algorithm capable of escaping them. However, a lingering question remains unanswered: are these worst-case instances common in practice or do they correspond to some pathological or rare cases?

Answering this question has been the subject of many recent studies. In particular, it has been shown that nearly-isotropic classes of problems in matrix completion/sensing [9], [10], [17], robust principle component analysis [12], [18], and dictionary recovery [19] have benign landscape, implying that they are free of spurious local minima. It has also been proven recently in [20] that under some conditions, the stochastic gradient descent may escape the sharp local minima in the landscape. At the core of the aforementioned results is the assumption on the static and time-invariant nature of the landscape. In contrast, many real-world problems should be solved sequentially over time with time-varying input data. For instance, in the optimal power flow problem, the electricity consumption of the consumers changes hourly [21], [22]. Therefore, it is natural to study the landscape of such time-varying nonconvex optimization problems, by taking into account their dynamic nature.

Time-varying dynamical systems: Recently, there has been a growing interest in analyzing the performance of numerical algorithms from a control-theoretical perspective [23]–[28]. Roughly speaking, the general idea behind these approaches is to analyze the convergence of a specific algorithm by first modeling its limiting behavior as a specific ODE that describes the evolution of the algorithm, and then studying its stability properties. As a natural extension, one would generalize this approach to a general class of time-varying optimization by modeling its KKT points as a general non-autonomous ODE corresponding to a time-varying dynamical system. However, the stability analysis of time-varying dynamical systems is highly convoluted in the general nonlinear settings. We note that several necessary and sufficient conditions for the stability of linear time-varying systems were proposed in [29]. A generalized time-varying Lyapunov function was proposed in [30] and has been applied in [31] to study the stability of an averaged system. Furthermore, slowly time-varying systems are investigated in [32].

II. Case Studies

In this section, we present empirical studies on the dynamic landscapes of two problems in power systems and machine learning: optimal power flow and dynamic matrix recovery.

A. Electrical Power Systems

In the optimal power flow problem, the goal is to match the supply of electricity with a time-varying demand profile, while satisfying the network, physical, and technological constraints. In practice, the problem is solved sequentially over time with the constraint that at every time-step, the solution cannot be significantly different from the one obtained in the previous time-step due to the so-called ramping constraints of the generators. We consider the IEEE 9-bus system [33] and initialize the system from the global solution, as well as three different spurious local solutions. We then change the load over time based on the California average load profile for the month of January 2019 (Figure 1a). The optimal power flow problem is then solved sequentially using local search every 15 minutes for the period of 24 hours, while taking into account the temporal couplings between solutions via the ramping constraints. The trajectories of the solutions for the optimal power flow problem with different initial points appear in Figure 1b. In this figure, the solid blue line represents the cost obtained by the semidefinite programming (SDP) relaxation of the optimal power flow [34]. This curve is a lower bound to the globally optimal cost and serves as a certificate of the global optimality whenever it touches other trajectories.

The gray circles in these plots are some of the local solutions that were obtained via a Monte Carlo simulation. Based on Figure 1b, indeed there exist multiple local solutions at almost all time-step (some of them emerge over time). However, surprisingly, the trajectories of the local solutions that are initialized at different points all converge towards the global solution. This implies that there is no spurious local trajectory, and therefore local search methods are able to find global minima of the optimal power flow problem at future times even when they start from poor local minima at the initial time.

B. Dynamic Matrix Recovery

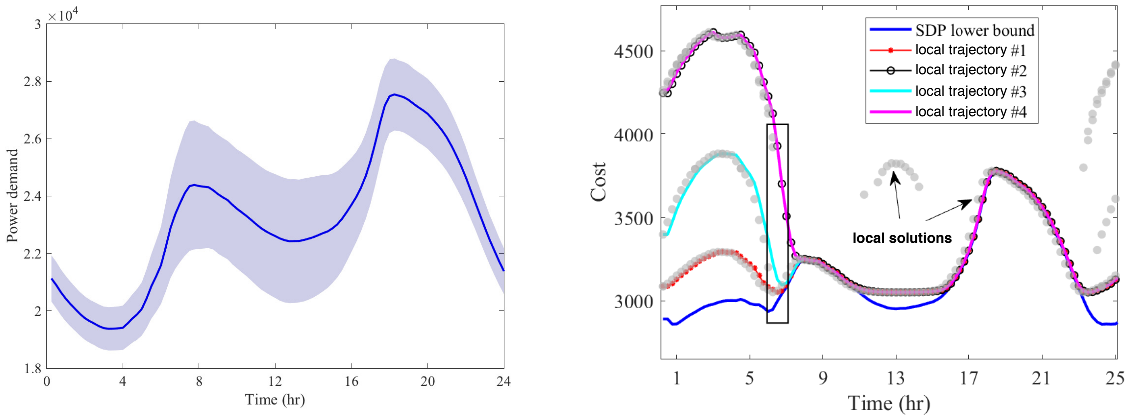
In the dynamic matrix recovery problem, the goal is to recover a time-varying low-rank matrix, based on a limited number of linear observations [4], [5]. This problem can be formulated as follows:

$$\inf_{X\in\mathbb{R}^{n\times r}} \sum_{i=1}^{m} \left(\langle A_i, XX^\top\rangle - d_i(t)\right)^2 \tag{1}$$

where $\langle\cdot,\cdot\rangle$ is the inner product operator, $\{A_i\}_{i=1}^m$ are the sensing matrices, and $d(t)$ is the time-varying measurements vector. Equivalently, (1) can be re-written as

$$\begin{aligned} \inf_{X\in\mathbb{R}^{n\times r},\epsilon\in\mathbb{R}^m} & \sum_{i=1}^{m} \epsilon_i^2 \\ \text{s.t.} \quad & \langle A_i, XX^\top\rangle - \epsilon_i = d_i(t)\ , \quad i=1,\ldots,m \end{aligned} \tag{2}$$

Assuming that $d(t)$ does not change over time, it is well-known that the above optimization problem has no spurious local minima if the sensing matrices $\{A_i\}_{i=1}^m$ satisfy a certain *restricted isometry property* (RIP). In particular, it is said that the sensing matrices $\{A_i\}_{i=1}^m$ satisfy RIP with a constant $\delta\in[0,1)$ if the inequality $(1-\delta)\|X\|_F^2 \leq \frac{1}{m}\sum_{i=1}^m\langle A_i, X\rangle \leq (1+\delta)\|X\|_F^2$ is satisfied for every $X\in\mathbb{R}^{n\times n}$ whose rank is upper bounded by $2r$ ($\|X\|_F$ is the Frobenious norm of the matrix X). Recently, [11] showed that if $r=1$, an RIP constant



(a) California average load profile for January 2019.

(b) Solution trajectories of time-varying optimal power flow.

Fig. 1: Case study in power systems (data collected from http://www.caiso.com).

of $\delta < 1/2$ is both necessary and sufficient for the benign landscape of the time-invariant matrix recovery problem.

Consider the sensing matrices

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0 & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & 0 \end{bmatrix} \tag{3}$$
$$A_3 = \begin{bmatrix} 1 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 0 & 0 \\ 0 & \frac{\sqrt{3}}{2} \end{bmatrix}$$

with the time-invariant measurement vector $d = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}^\top$ and $r = 1$. The paper [11] proved that the RIP constant for the above sensing matrices is equal to $1/2$. This implies that the matrix recovery problem with the aforementioned sensing matrices is prone to having spurious local minima. In fact, [11] showed that the above problem has one global solution at $Z = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$ and one spurious local solution at $X = \left[0, 1/\sqrt{2}\right]^\top$. Now, consider the time-varying version of the above instance, where the measurement vector changes over time, as in:

$$d(t) = \begin{bmatrix} (0.8 + 0.2\cos t)^2 + \frac{1}{2}(0.2\sin t)^2 \\ \sqrt{3}(0.2\sin t)(0.8 + 0.2\cos t) \\ 0 \\ \frac{\sqrt{3}}{2}(0.2\sin t)^2 \end{bmatrix}$$

It is easy to see that $Z = \begin{bmatrix} 0.8 + 0.2\cos t & 0.2\sin t \end{bmatrix}^\top$ is the trajectory of the globally optimal solution to the defined dynamic matrix recovery problem. Moreover, using a Gradient descent algorithm initialized at the spurious local solution at time $t = 0$, we solve (2) sequentially over time with an appropriate regularization (to be defined later). Figures 2a and 2b show that, despite the fact that the problem has a spurious local minimum at $t = 0$ and future times, its local trajectory gradually converges to the global one.

III. Notion of Spurious Local Trajectory

Inspired by the above case studies, we consider the effect of the variation in the input data on the landscape of the optimization problem. We focus on the following time-varying nonconvex optimization:

$$\inf_{x(t) \in \mathbb{R}^n} \; f(x(t), t) \quad \text{s.t.} \quad h_i(x(t)) = d_i(t), \; i = 1, \ldots, m \tag{4}$$

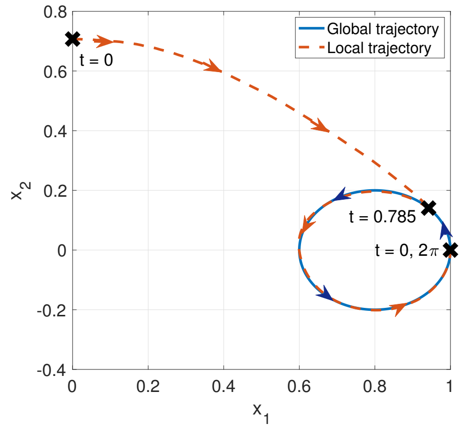
where the objective function $f(x(t), t)$ and the right-hand side of the equality constraints vary over time $t \in [0, T]$. We assume that $f : \mathbb{R}^n \times [0, T] \longrightarrow \mathbb{R}$ is a continuously differentiable function. Moreover, $h_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ and $d_i : [0, T] \longrightarrow \mathbb{R}$ for $i = 1, \ldots, m$ are twice continuously differentiable functions, and that $T > 0$ is a finite time horizon. Moreover, we assume that f is uniformly bounded from below (i.e., $f(x(t), t) \geq M$ for some constant M) and that the problem is feasible for all $t \in [0, T]$. The objective function $f(x, t)$ may be nonconvex in x and the constraint function $h(x) = (h_1(x), \ldots, h_m(x))$ may be nonlinear in x. Note that the dynamic matrix recovery problem (2) is a special case of (4).

Remark 1. *Inequality constraints can also be included in* (4) *through a reformulation technique. In particular, suppose that* (4) *includes a set of inequality constraints $g_j(x) \leq v_j(t)$ for $j = 1, \ldots, l$. Then, one can reformulate them as equality constraints through the following procedure:*

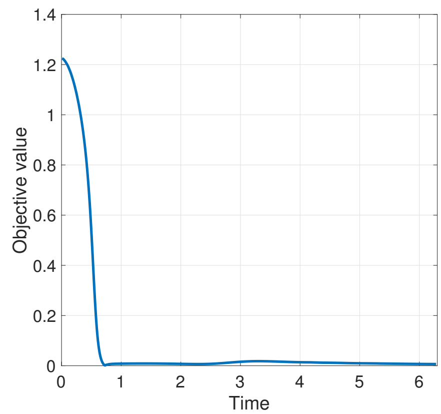
1. *Rewrite the inequality constraints by introducing a slack variable $s \in \mathbb{R}^l$, as in*
$$g_j(x(t)) + s_j(t) = v_j(t), \qquad j = 1, \ldots, l$$
2. *Augment the objective function with a penalty $p(s(t)) = \sum_{j=1}^{l} p_j(s_j(t))$.*

Here, $p_j(s_j(t))$ are nonsmooth loss functions for an exact reformulation. Furthermore, they can be relaxed to continuously differentiable loss functions at the expense of incurring some (controllable) approximation errors; see [35], [36]. This implies that the previously-introduced optimal power flow problem can be reformulated as (4).

In practice, one can only hope to sequentially solve this problem at discrete times $0 = t_0 < t_1 < t_2 < \ldots < t_N = T$. However, notice that (4) is un-regularized. In particular, depending on the properties of the objective function, an arbitrary



(a) The trajectories of local and global solutions over time



(b) The objective value of the local trajectory over time

Fig. 2: Case study in matrix recovery.

solution to (4) at time t_k can be arbitrarily far from that of (4) at time t_{k-1}. However—as elucidated in our case study on the optimal power flow problem— it is neither practical nor realistic to have solutions that change abruptly over time in many real-world problems. One way to circumvent this issue is to regularize the problem at time t_{k+1} by penalizing the deviation of its solution from the one obtained at time t_k. Precisely, we employ a quadratic proximal regularization as is done in online learning [37].

Definition 1. *Given evenly spaced-out time steps $0 = t_0 < t_1 < t_2 < \ldots < t_N = T$ for some integer N, a sequence $x_0, x_1, x_2, \ldots, x_N$ is said to be a* ***discrete local trajectory*** *of the time-varying optimization* (4) *if the following holds:*

1) *x_0 is a local solution to the time-varying optimization* (4) *at time $t_0 = 0$;*
2) *for $k = 0, 1, 2, \ldots, N-1$, x_{k+1} is local solution to the regularized problem*

$$\begin{aligned} \inf_{x\in\mathbb{R}^n} \quad & f(x, t_{k+1}) \ + \ \alpha \ \frac{\|x-x_k\|^2}{2(t_{k+1}-t_k)} \\ \text{s.t.} \quad & h_i(x) = d_i(t_{k+1}) \ , \quad i = 1, \ldots, m. \end{aligned} \tag{5}$$

Above, $\alpha > 0$ is a fixed regularization parameter and $\|\cdot\|$ denotes the Euclidian norm.

Note that in the above definition, the term *local solution* refers to any feasible point that satisfies the Karush-Kuhn-Tucker (KKT) conditions for (5). A natural approach to characterizing the global landscape of (4) is to analyze discrete local trajectories of the regularized problem (5). However, notice that the non-convexity of (5) may lead to *bifurcations* in discrete local trajectories. In particular, given a local solution x_k, the regularized problem (5) may possess two local solutions $x_{k+1}^{(1)}$ and $x_{k+1}^{(2)}$, each resulting in a different discrete local trajectory.[2] The non-uniqueness of the discrete local trajectories due to the bifurcation will make the analysis inconclusive. This is because the next solution of the problem given the current solution is not well-defined and due to the number of possibilities at each step, the solution trajectory is not unique and can take an exponential number of possibilities depending on the settings of the numerical algorithm (the choice of descent directions and step sizes). However, in what follows, we show that such bifurcations disappear in the ideal scenario, where the regularized problem can be sampled arbitrarily fast, or equivalently, as we increase N to infinity. In particular, given a fixed initial local solution x_0, we show that any discrete local trajectory starting from x_0 converges uniformly to the unique solution to a well-defined ODE that is initialized at x_0. By building upon this result, we introduce the notion of spurious local trajectory as a generalization to the notion of spurious local minima.

Given an initial local solution x_0, consider the following initial value problem:

$$\dot{x} = -\frac{1}{\alpha}\eta(x,t) + \theta(x)\dot{d} \tag{6a}$$

$$x(0) = x_0 \tag{6b}$$

where

$$\begin{aligned} \eta(x,t) := \quad & \left[I - \mathcal{J}(x)^\top(\mathcal{J}(x)\mathcal{J}(x)^\top)^{-1}\mathcal{J}(x)\right] \\ & \times \nabla_x f(x,t), \end{aligned} \tag{7a}$$

$$\theta(x) \ := \ \mathcal{J}(x)^\top(\mathcal{J}(x)\mathcal{J}(x)^\top)^{-1}. \tag{7b}$$

Above, $\mathcal{J}(x)$ denotes the Jacobian of the left-hand side of the constraints $h(x) = [h_1(x), \ldots, h_m(x)]^\top$ and $d(t)$ denotes the right-hand side of the constraints, that is to say $d(t) = [d_1(t), \ldots, d_m(t)]^\top$. The term $\theta(x)\dot{d}$ captures the effect of data variation in the dynamics, and the function $\eta(x,t)$ can be interpreted as the orthogonal projection of the gradient $\nabla_x f(x,t)$ on the Kernel of $\mathcal{J}(x)^\top$.

Later, we will show that the solution to (6) exists, it is unique, and can be used to fully characterize the limiting behavior of every discrete local trajectory of the time-varying problem (4).

[2]For example, there exist two discrete trajectories starting at $x_0 = 0$ and at time $t_0 = 0$ for the time-varying objective function $f(x,t) := x^2(T/2-t)$. Indeed, the discrete trajectory stays at $x_k = 0$ for $t_k \leq T/2$ and then, due to the regularization, it bifurcates into two separate discrete trajectories.

Assumption 1 (Uniform Boundedness)**.** *There exist constants $R_1 > 0$ and $R_2 > 0$ such that, for any discrete local trajectory $x_0, x_1, x_2, \ldots$, the parameter $\|x_k\|$ and the objective function of* (5) *at x_k are upper bounded by R_1 and R_2, respectively, for every $k \in \{0, 1, 2, \ldots, N\}$.*

Assumption 1 is a common assumption made in the optimization literature [38], [39], and can be guaranteed by requiring the feasible region to be compact. This condition can also be explicitly imposed via an inequality constraint (such as box or norm constraint). According to Remark 1, such inequality constraint can be moved to the objective function via an (exact/inexact) penalty method. Moreover, the uniform boundedness assumption on the variables is crucial from a practical standpoint. For instance, in the time-varying OPF, the variables, i.e., active and reactive power, voltage magnitudes, and their angles, are restricted to bounded sets implied by the laws of physics and technological constraints on physical devices. It is worth noting that the main results of the paper do not depend on the explicit values of the constants R_1 and R_2.

Assumption 2 (Non-singularity)**.** *There exists a constant $c > 0$ such that, for any discrete local trajectory $x_0, x_1, x_2, \ldots$, it holds that $\sigma_{\min}(\mathcal{J}(x_k)) \geqslant c$ for all $k \in \{0, 1, 2, \ldots\}$, where $\sigma_{\min}$ denotes the minimal singular value.*

Assumption 2 implies that linear independence constraint qualification (LICQ) holds at every point of a discrete local trajectory, which in turn implies that the constraints are non-degenerate. The LICQ is a simple sufficient condition to guarantee the well-definedness of the KKT points [40], and is the most standard assumption in the optimization literature [36], [41], [42]. Indeed, most of the off-the-shelf solvers, such as IPOPT [43], only converge to solutions that automatically satisfy LICQ.

Theorem 1 (Existence and Uniqueness)**.** *Let Assumption 1 and Assumption 2 hold. Suppose that x_0 is an arbitrary local solution to the time-varying optimization* (4) *at $t = 0$. Then, the ODE* (6) *with the initial value condition $x(0) = x_0$ has a unique continuously differentiable solution $x : [0, T] \to \mathbb{R}^n$.*

Theorem 1 states that the proposed ODE is well-defined and has a unique solution, provided that its initial value is a local solution, i.e., it satisfies the KKT conditions for the original time-varying optimization problem. As will be shown later, this assumption is crucial and cannot be relaxed in general. Given the unique solution to the proposed ODE, the next theorem precisely characterizes its relationship to *any* discrete local trajectory of (5) starting at x_0.

Theorem 2 (Uniform Convergence)**.** *Let Assumption 1 and Assumption 2 hold. If x_0 is a local solution to the time-varying optimization* (4) *at $t = 0$, then any discrete local trajectory initialized at x_0 converges towards the solution $x : [0, T] \to \mathbb{R}^n$ with $x(0) = x_0$, in the sense that*

$$\lim_{N \to +\infty} \sup_{0 \leq k \leq N} \|x_k - x(t_k)\| = 0, \tag{8}$$

where N is the number of points in the discrete local trajectories, and $0 = t_0 < t_1 < t_2 < \ldots < t_N = T$ are evenly spaced-out time-steps.

Sketch of the proofs. The proofs for Theorems 1 and 2 are quite involved and hence, they are deferred to the next section. In what follows, we provide the high-level ideas of our developed proof techniques. Note that most of the classical results on ordinary differential equations, namely the Picard-Lindelöf Theorem [44, Theorem 3.1], the Cauchy-Peano Theorem [44, Theorem 1.2], and the Carathéodory Theorem [44, Theorem 1.1], can only guarantee the existence of a solution in a local region, i.e., a neighborhood $[0, \tau]$ where $\tau < T$ is potentially very small. On the other hand, the global version of Picard-Lindelöf Theorem only holds under a restrictive Lipschitz condition, which is not satisfied for (6). Instead, we take a different approach to prove existence and uniqueness of the solution to (6) (Theorem 1). The proof consists of three general steps:

1) By building upon the Arzelà-Ascoli Theorem, we show that, among all the discrete local trajectories that are initialized at x_0, there exists at least one that is uniformly convergent to a continuously differentiable function $y : [0, T] \to \mathbb{R}^n$.
2) By fully characterizing the KKT points of (5), we prove that y is a solution to (6) when $N \to +\infty$.
3) The uniqueness of the solution is then proved by showing the existence of an open and connected set $\mathcal{D}$ such that the proposed ODE is locally Lipschitz continuous on $\mathcal{D}$ and $(y(t), t) \in \mathcal{D}$ for every $t \in [0, T]$. This, together with [44, Theorem 2.2], completes the proof of Theorem 1.

Given the existence and uniqueness of the solution to (6), we show the correctness of Theorem 2 by making an extensive use of the so-called backward Euler method [45]. In particular, we show that all of the discrete local trajectories converge to a discretized version of the solution to (6) that is obtained by the backward Euler method. This, together with the existing convergence results on the backward Euler iterations, completes the proof of Theorem 2. □

Now that we have established the connection between the discrete local trajectories and their continuous limit, we naturally propose the following definition.

Definition 2. *A continuously differentiable function $x(t) : [0, T] \longrightarrow \mathbb{R}^n$ is said to be a* ***continuous local trajectory*** *of the time-varying optimization* (4) *if the following holds:*

1) *$x(0)$ is a local solution to the time-varying optimization* (4) *at time $t = 0$;*
2) *$x(t)$ is a solution to* (6)*.*

The next definition will be at the core of our subsequent definition of spurious local trajectories.

Definition 3. *The* ***region of attraction*** *of a local minimum $x^*(t)$ of $f(\cdot, t)$ in the feasible set $\mathcal{F}(t) = \{x \in \mathbb{R}^n : h(x) =$

$d(t)\}$ *at a given time* t *is defined as:*

$$\Big\{x_0 \in \mathcal{F}(t) \;\Big|\; \lim_{s\to\infty} \tilde{x}(s) = x^*(t) \quad \textit{where}$$
$$\frac{d\tilde{x}(s)}{ds} = -\frac{1}{\alpha}\eta(\tilde{x}(s), t) + \theta(\tilde{x}(s))\dot{d}(t) \quad \textit{and} \quad \tilde{x}(0) = x_0\Big\}$$

Intuitively, the basin of attraction for a local solution $x^*(t)$ is defined as the set of initial points for which an alternative (time-invariant) ODE has a solution whose limit corresponds to $x^*(t)$ (for fixed t). This alternative ODE is akin to the classical Riemannian gradient flow, which is well-studied in the literature with rigorous convergence results [46]–[48]. We next introduce the central notion in this paper.

Definition 4. *A continuous local trajectory* $x(t)$ *is said to be* ***spurious*** *if for all* $\bar{T} < T$*, there exists a time* $t \in [\bar{T}, T]$ *such that* $x(t)$ *does not belong to the region of attraction of a global solution of* $f(\cdot, t)$*. Accordingly, the time-varying optimization problem* (4) *is said to* ***have no spurious local trajectories*** *if, when initialized at a local solution, any continuous local trajectory* $x(t)$ *belongs to the region of attraction of a global solution of* $f(\cdot, t)$ *at all times* $t \in [\bar{T}, T]$ *for some constant* $\bar{T} < T$*.*

So far, we have taken the time horizon T to be finite. However, the above definition naturally applies to problems with an infinite time horizon $T = +\infty$. In Theorem 3, we will provide a sufficient condition under which the above non-spurious trajectory property holds for a general objective function with a damping sinusoidal time-varying perturbation.

It may be speculated that a spurious local trajectory could have been simply defined as a trajectory that does not converge towards a global solution. To understand why the latter definition is not meaningful, notice that both discrete and continuous local trajectories are defined with respect to the regularized problem (5), as opposed to (4). The regularization term acts as an *inertia* in the continuous local trajectory, forcing it to "lag behind" the global solution when it changes rapidly over time. Therefore, under this alternative definition, all trajectories would be considered spurious. This would be true even for the trajectory initialized at the global minimum. See Figures 3a and 3b for an illustration of this phenomenon.

The notion introduced in Definition 4, while it deals with continuous local trajectories, naturally has implications for discrete local trajectories. With sufficiently small time steps, the discrete trajectory will eventually converge to the region of attraction of a global solution if the corresponding continuous trajectory is not spurious.

IV. Conditions for the Absence of Spurious Local Trajectories

In this section, we analyze the role of data variation on the behavior of the solution trajectories. Observe that without data variation, strict spurious local minima cannot not be escaped. This is a consequence of classical results on the local stability of time-invariant ODEs (see for instance [49, Corollary 10]). In contrast, we show that data variation can enable escaping spurious local solutions over time. In particular, we prove that even a simple periodic variation in the data can induce continuous local trajectories to escape non-global minima and eventually track the global minima.

To better illustrate the main idea, we start with a class of uni-dimensional time-varying problems, and provide sufficient conditions for the absence of spurious local trajectories. Then, we extend our results to a general class of multi-dimensional problems. Consider the function

$$\inf_{x\in\mathbb{R}} f(x,t) := g(x - \beta\sin(t)) \tag{9}$$

where $g : \mathbb{R} \longrightarrow \mathbb{R}$ is continuously twice differentiable and $\beta > 0$ models the variation of the data over time. Only the right-hand side varies over time, and therefore, this problem fits well in our introduced framework. We assume that $g(\cdot)$ admits only three stationary points $g'(y_1) = g'(y_2) = g'(y_3)$ with $y_1 < y_2 < y_3$. We assume also that y_1 and y_3 are local minima such that $g(y_1) > g(y_3)$, while y_2 is a local maximum. Finally, we assume that g is coercive (its limit at $\pm\infty$ is $+\infty$). Thus, its global infimum is reached in y_3.

The motivation behind studying this class of functions $f(\cdot)$ is as follows. Since $g(y)$ has a global minimum as well as a spurious solution, when it is minimized by a gradient descent algorithm initialized at the spurious solution, it will become stuck there. This means that using gradient descent for such function is inefficient. However, one can oscillate the function to arrive at the time-varying function $f(x,t)$ and then study it in the context of online optimization. The following result identifies sufficient conditions for the absence of spurious local trajectories, which implies that if α and β are selected appropriately, gradient descent will always find the global solution.

Proposition 1. *If* $\alpha, \beta > 0$ *are such that*

1) $\alpha\beta \geqslant C := \max_{y_1 \leqslant y \leqslant y_3} g'(y)$,
2) $\exists m_1, m_2 \in \mathbb{R} : \quad m_1 < y_1 < m_2 \quad$ *and* $\quad g'(m_1) = g'(m_2) = -\alpha\beta$,
3) $-C/\alpha(t_2 - t_1) - \beta(\sin(t_2) - \sin(t_1)) + m_1 \geqslant m_2$ *where* $0 < t_1 \leqslant t_2$ *satisfy* $\cos(t_1) = \cos(t_2) = -C/(\alpha\beta)$,

then the time-varying problem (9) *has no spurious local trajectories for all time horizon* $T \in [2\pi, +\infty)$*.*

Proof. A continuous local trajectory $x : [0, T] \longrightarrow \mathbb{R}$ satisfies

$$x(0) \leqslant y_3, \qquad \dot{x} = -\frac{1}{\alpha}\nabla_x f(x,t), \tag{10}$$

which, after the change of variable $y := x - \beta\sin(t)$, reads

$$y(0) \leqslant y_3, \qquad \dot{y} = -\frac{1}{\alpha}g'(y) - \beta\cos(t). \tag{11}$$

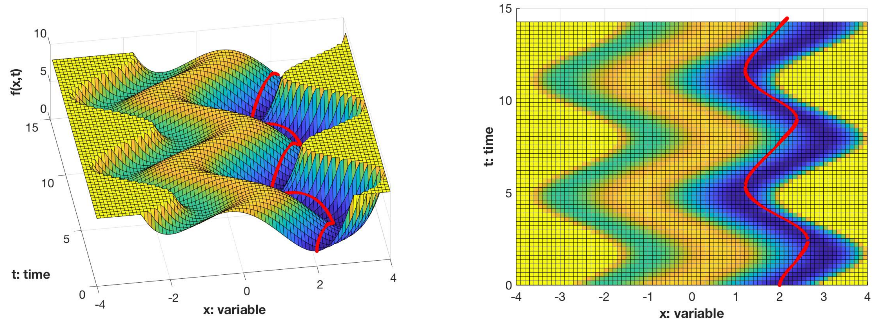
We first show by contradiction that there exists $t \in [0, 2\pi]$ such that $y(t) \geqslant m_2$. Assume that $y(t) < m_2$ for all $t \in [0, 2\pi]$. Then, for all $t \in [0, 2\pi]$, it holds that

$$\dot{y} = -\frac{1}{\alpha}g'(y) - \beta\cos(t) \geqslant -\frac{C}{\alpha} - \beta\cos(t). \tag{12}$$

Thus, we have

$$y(t_2) \geqslant -\frac{C}{\alpha}(t_2 - t_1) - \beta(\sin(t_2) - \sin(t_1)) + y(t_1). \tag{13}$$

We next show by contradiction that $y(t_1) \geqslant m_1$. Assume that $y(t_1) < m_1$. Thus $y(t_1) < m_1 < y_1 \leqslant y(0)$. Let t_3 denote the



(a) Graph of a time-varying optimization $\inf_{x\in\mathbb{R}} f(x,t)$ showing that the final state of the trajectory belongs to the region of attraction of the global minimum.

(b) Graph of the same time-varying optimization $\inf_{x\in\mathbb{R}} f(x,t)$ from above showing that the trajectory can never stay in a neighborhood of the global minimum of arbitrarily small size.

Fig. 3: Example of a time-varying optimization.

maximal element of the compact set $[0,t_1]\cap y^{-1}(m_1)$, where $y^{-1}(b) := \{a \in \mathbb{R} \mid y(a) = b\}$. Thus $y(t) \leqslant y(t_3)$ for all $t \in [t_3, t_1]$. As a result, $y'(t_3) \leqslant 0$. Together with $y'(t_3) = -1/\alpha g'(m_1) - \beta\cos(t_3) = \beta(1-\cos(t_3))$, this implies that $t_3 = 0$ or $t_3 = 2\pi$. This is in contradiction with $0 < t_3 < t_1 < \pi$.

Now that we have proven that $y(t_1) \geqslant m_1$, equation (13) implies that $y(t_2) \geqslant m_2$. This is a contradiction. Therefore there exists $t \in [0, 2\pi]$ such that $y(t) \geqslant m_2$. Using the same argument as in the previous paragraph, we obtain $y(2\pi) \geqslant m_2$. As a result, $x(2\pi) = y(2\pi) - \beta\sin(2\pi) \geqslant m_2$ as well. Finally, using standard arguments in Lyapunov theory[3], there exists $\bar{T} < T$ such that $x(t)$ belongs to the region of attraction of y_3 for all $t \in [\bar{T}, T]$. □

We highlight the implications of the above proposition through a numerical example.

Example 1. *Consider the objective function* $f(x,t) := g(x - \beta\sin(t))$ *where*

$$g(y) := 1/4y^4 + 1/8y^3 - 2y^2 - 3/2y + 8. \tag{14}$$

The time-varying objective $f(x,t)$ *has the following stationary points: it admits a spurious local minimum at* $-2+\beta\sin(t)$*, a local maximum at* $-3/8+\beta\sin(t)$*, and a global minimum at* $2+\beta\sin(t)$*. The three sufficient conditions of Proposition 1 can be brought to bear on this example. They yield three inequalities, as shown in Figure 4a, whose feasible region is represented in Figure 4b. Taking a point in that feasible region, we confirm numerically in Figure 4c that a trajectory initialized at a local minimum of* $f(\cdot,0)$ *winds up in the region of attraction of the global solution to* $f(\cdot,T)$ *at the final time* $T = 2\pi$*. In contrast, taking a point outside the feasible region, we observe in Figure 4d that a trajectory initialized at a local minimum of* $f(\cdot,0)$ *does not end up in the region of attraction of the global solution to* $f(\cdot,T)$*.*[4]

We make a few remarks regarding Figure 4a. Note that k_1 *and* k_2 *are integers in* $\{0,1,2\}$ *such that* k_1 *minimizes the line it appears in, and* k_2 *minimizes the line it appears in while not being equal to* k_1*. These numbers come from Viète's solution to a cubic equation [50]. Furthermore, the second inequality corresponds to minus the discriminant of a fourth-order polynomial.*

Next, we will extend the aforementioned result to a general class of multi-dimensional optimization problems. The goal is to show that certain non-global local solutions of an arbitrary time-invariant function $g(x)$ that cannot be escaped using deterministic local search methods can indeed be escaped via the conversion of the problem to a time-varying function $f(x,t)$ for which there is no spurious trajectory. Consider the time-varying optimization problem

$$\inf_{x\in\mathbb{R}^n} f(x,t) := \inf_{x\in\mathbb{R}^n} g(x - \beta e^{-\lambda t}\sin(\omega t)u) \tag{15}$$

where $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ is continuously twice differentiable, coercive (its limit as $\|y\| \to +\infty$ is $+\infty$). The amplitude $\beta > 0$ and the pulsation $\omega > 0$ model the sinusoidal variation of data over time with a damping factor of $\lambda > 0$. The variation occurs along a direction $u \in \mathbb{R}^n$ of norm 1. Let $\{y_i\}_{i\in\mathcal{I}}$ denote the set of spurious local minima of $g(x)$. Moreover, let $B(a,r)$ (respectively $S(a,r)$) denote the Euclidian ball (respectively sphere) in $\mathbb{R}^n$ centered at a and of radius r. Given a fixed $R > 0$, we define the following constants

$$\begin{aligned} C_1 &:= \max_{y\in\bigcup\limits_{i\in\mathcal{I}} B(y_i,R)} \|\nabla g(y)\|, \\ C_2 &:= \min_{\substack{d\,\in\, S(0,1)\\ i\,\in\,\mathcal{I}}} \langle \nabla g(y_i - Rd), d\rangle. \end{aligned} \tag{16}$$

[3]Details can be found in the first paragraph of page 24 of https://arxiv.org/pdf/1905.09937v1.pdf.

[4]In order to increase visibility, a maximal threshold is used on the objective function $f(x,t)$ in Figure 4c and Figure 4d (hence the flat parts). For the same reason, a non-linear scaling is used. Precisely, $(x,t) \longrightarrow f(x+(\beta-1)\sin(t),t)$ and $t \longrightarrow x(t)-(\beta-1)\sin(t)$ are represented in the figures. This explains why $x(t)$ appears to decrease for small $0 \leqslant t \leqslant 2\pi$ in Figure 4c.

These constants enable us to control fluctuations of $g(x)$ in the vicinity of its local minima. A small constant C_1 corresponds to spurious local minima that tend to be flat, while larges values are associated with local minima that are sharper [51, Metric 2.1]. For the sake of clarity, we assume that $g(x)$ has no saddle points and local maxima outside of $\cup_{i\in\mathcal{I}} B(y_i,R)$ (for more on this, see Remark 2). Notice that $C_1 \geqslant C_2$ due to the Cauchy-Schwarz inequality. Theorem 3 below shows that if C_1 is not too large, then one can escape spurious local minima, and if C_2 is not too small, then one will never return to the vicinity of any spurious local minima after some time.

Theorem 3. *If* $2\alpha\omega(\beta e^{-\lambda\pi/(2\omega)} - R)/\pi > C_1$ *and* $\alpha\beta e^{-\lambda R\alpha/(C_1+\alpha\beta\omega)}\sqrt{\lambda^2+\omega^2} < C_2$*, then the time-varying optimization* (15) *has no spurious trajectories.*

Proof. First, we show that the spurious local minimum is initially escaped. A continuous local trajectory $x(t)$ satisfies

$$x(0) \in \{y_i\}_{i\in\mathcal{I}}, \qquad x'(t) = -\frac{1}{\alpha}\nabla_x f(x(t),t), \tag{17}$$

which, after the change of variables $y(t) := x(t) - \beta e^{-\lambda t}\sin(\omega t)u$, reads

$$\begin{aligned} &y'(t) = -\nabla g(y(t))/\alpha - \beta e^{-\lambda t}[-\lambda\sin(\omega t) + \omega\cos(\omega t)]u, \\ &y(0) \in \{y_i\}_{i\in\mathcal{I}}, \end{aligned} \tag{18}$$

We first show by contradiction that there exists some time $t \in [0,T]$ such that $\|y(t) - y(0)\| > R > 0$. Assume that $\|y(t)-y(0)\| \leqslant R$ for all $t \geqslant 0$. Then, for all $t \geqslant 0$, it holds that

$$\begin{aligned} &\langle y'(t), u\rangle \\ &= \langle -\nabla g(y(t))/\alpha - \beta e^{-\lambda t}[-\lambda\sin(\omega t)+\omega\cos(\omega t)]u, u\rangle \\ &= -\langle \nabla g(y(t)), u\rangle/\alpha - \beta e^{-\lambda t}[-\lambda\sin(\omega t)+\omega\cos(\omega t)]\langle u,u\rangle \\ &\leqslant \|\nabla g(y(t))\|/\alpha - \beta e^{-\lambda t}[-\lambda\sin(\omega t)+\omega\cos(\omega t)] \\ &\leqslant \{C_1 - \alpha\beta e^{-\lambda t}[-\lambda\sin(\omega t)+\omega\cos(\omega t)]\}/\alpha, \end{aligned} \tag{19}$$

from which we deduce that

$$\begin{aligned} \langle y(t)-y(0), u\rangle = \left\langle \int_0^t y'(s)ds, u\right\rangle = \int_0^t \langle y'(s), u\rangle ds \\ \leqslant [C_1 t - \alpha\beta e^{-\lambda t}\sin(\omega t)]/\alpha. \end{aligned} \tag{20}$$

Our assumption that $2\alpha\omega(\beta e^{-\lambda\pi/(2\omega)} - R)/\pi > C_1$ implies that the upper bound in (20) is negative when $t = \pi/(2\omega)$. Using the Cauchy-Schwarz inequality, we then obtain

$$\begin{aligned} &\|y(\pi/(2\omega)) - y(0)\| \geqslant |\langle y(\pi/(2\omega)) - y(0), u\rangle| \\ &\qquad\qquad \geqslant [\alpha\beta e^{-\lambda\pi/(2\omega)} - C_1\pi/(2\omega)]/\alpha > R. \end{aligned}$$

This yields a contradiction. We conclude that there exists $t_1 \geqslant 0$ such that $\|y(t_1) - y(0)\| > R$. Observe that

$$\begin{aligned} \|y(t_1)-y(0)\| &= \left\| \int_0^{t_1} \nabla g(y(t))dt - \beta e^{-\lambda t_1}\sin(\omega t_1)u \right\| \\ &= \int_0^{t_1} \|\nabla g(y(t))dt\| + \beta e^{-\lambda t_1}\sin(\omega t_1) \\ &\leqslant C_1 t_1/\alpha + \beta e^{-\lambda t_1}\sin(\omega t_1) \\ &\leqslant (C_1/\alpha + \beta\omega)t_1. \end{aligned} \tag{21}$$

As a result, $t_1 > R\alpha/(C_1+\alpha\beta\omega)$. We have thus identified a minimum time taken by the trajectory to exit the ball of radius R centered at $y(0)$. Second, we show that, after some time, the continuous trajectory never returns to the vicinity of any spurious local minimum. To reason by contradiction, assume that there exist $i \in \mathcal{I}$ and $t_1 < t_3$ such that $\|y(t_3) - y_i\| < R$. Since the trajectory is continuous, there exists $t_2 \in (t_1, t_3)$ such that $\|y(t_2) - y_i\| = R$, that is to say, there exists $d \in \mathbb{R}^n$ such that $\|d\| = 1$ and $y(t_2) = y_i + Rd$. Take t_2 to be the largest such instance in the interval (t_1, t_3). We then have

$$\begin{aligned} &\langle y'(t_2), d\rangle \\ &= \langle -\nabla g(y(t_2))/\alpha - \beta e^{-\lambda t_2}[-\lambda\sin(\omega t_2)+\omega\cos(\omega t_2)]u, d\rangle \\ &= \langle \nabla g(y_i + Rd), -d\rangle/\alpha \\ &\quad - \beta e^{-\lambda t_2}[-\lambda\sin(\omega t_2)+\omega\cos(\omega t_2)]\langle u, d\rangle \\ &\geqslant C_2/\alpha - \beta e^{-\lambda t_2}[-\lambda\sin(\omega t_2)+\omega\cos(\omega t_2)]\langle u, d\rangle \\ &= \left\{ C_2 - \alpha\beta e^{-\lambda t_2}\sqrt{\lambda^2+\omega^2}\cos(\omega t_2 + \arctan(\lambda/\omega)) \right\}/\alpha \\ &\geqslant (C_2 - \alpha\beta e^{-\lambda t_2}\sqrt{\lambda^2+\omega^2})/\alpha \\ &\geqslant (C_2 - \alpha\beta e^{-\lambda R\alpha/(C_1+\alpha\beta\omega)}\sqrt{\lambda^2+\omega^2})/\alpha \end{aligned} \tag{22}$$

where in the last inequality we used the fact that $R\alpha/(C_1+\alpha\beta\omega) \leqslant t_1 < t_2$. The Taylor expansion for $t > t_2$ in a neighborhood of t_2 reads

$$y(t) - y(t_2) = y'(t_2)(t-t_2) + o(t-t_2), \tag{23}$$

from which we deduce that

$$\begin{aligned} &\left\langle \frac{y(t)-y(t_2)}{t-t_2}, d \right\rangle \\ &= \langle y'(t_2), d\rangle + o(1) \\ &> (C_2 - \alpha\beta e^{-\lambda R\alpha/(C_1+\alpha\beta\omega)}\sqrt{\lambda^2+\omega^2})/(2\alpha) > 0 \end{aligned} \tag{24}$$

where we used $\alpha\beta e^{-\lambda R\alpha/(C_1+\alpha\beta\omega)}\sqrt{\lambda^2+\omega^2} < C_2$. Hence

$$\begin{aligned} \|y(t)-y_i\| \geqslant \langle y(t)-y_i, d\rangle &= \langle y(t)-y(t_2)+y(t_2)-y_i, d\rangle \\ &= \langle y(t)-y(t_2), d\rangle + \langle Rd, d\rangle \\ &> R. \end{aligned} \tag{25}$$

Recall that $\|y(t_3) - y(0)\| \leqslant R$. By continuity of the trajectory, there exists $t \in (t_2, t_3]$ such that $\|y(t) - y_i\| = R$, which contradicts the maximality of t_2. Hence, for all $t \geqslant t_1$ and $i \in \mathcal{I}$, we have that $\|y(t) - y_i\| \geqslant R$.

Third, we show that $x(t_1) = y(t_1) + \beta e^{-\lambda t_1}\sin(wt_1)u$ is in the region of attraction of a global minimum of the function $f(x, t_1)$. Now, we freeze the time at t_1. Consider the set $D = \{x \in \mathbb{R}^n : f(x,t_1) \leq f(x(t_1),t_1)\}$ and choose D_1 as the connected component of D which contains the point $x(t_1)$. Because $f(x,t_1)$ is coercive, D_1 is a compact set. In addition, D_1 is a positively invariant set with respect to the gradient flow system

$$\dot{\tilde{x}}(s) = -\nabla_{\tilde{x}} f(\tilde{x}(s), t_1) \tag{26}$$

for the fixed time t_1 because the gradient flow system will not increase the function value. Denote $f^*(t_1)$ as the global minimum value of $f(\tilde{x}, t_1)$ and take $V(\tilde{x}) = f(\tilde{x}, t_1) - f^*(t_1)$. Then, $V(\tilde{x})$ is a Lyapunov function for (26) such that $\dot{V}(\tilde{x}) = -\|\nabla_{\tilde{x}} f(\tilde{x}, t_1)\|^2 \leq 0$ in D_1. Let E be the points in D_1 such that $\nabla_{\tilde{x}} f(\tilde{x}, t_1) = 0$. Since $g(x)$ has no saddle points and local maxima outside of $\cup_{i\in\mathcal{I}} B(y_i, R)$, then $f(\cdot, t_1)$ has no saddle points and local maxima outside of $\cup_{i\in\mathcal{I}} B(y_i + \beta e^{-\lambda t_1}\sin(wt_1)u, R)$. Thus, the set E only contains the global minima of $f(\tilde{x}, t_1)$. Furthermore, the set E is also an invariant set with respect to (26). Then, by LaSalle's theorem in [52, Theorem 4.4], the solution of (26) starting at $x(t_1)$ converges to the global minimum as $s \to \infty$. This implies that $x(t_1)$ is in the region of attraction of a global minimum of the function $f(x, t_1)$. Finally, we show that the trajectory remains in the region of attraction of the set of global minima after some time. This follows immediately from the assumption that $g(x)$ has no saddle points and local maxima outside of $\cup_{i\in\mathcal{I}} B(y_i, R)$ and the fact that the trajectory will never returns to the vicinity of any spurious local minimum, that is, $\cup_{i\in\mathcal{I}} B(y_i, R)$. □

Observe that a necessary condition for the absence of spurious trajectories readily follows from the proof of Theorem 3, namely that $\alpha\beta\sqrt{\omega^2+\lambda^2} \geqslant -C_2$. Indeed, if $\alpha\beta\sqrt{\omega^2+\lambda^2} < -C_2$, then the spurious local minima cannot be escaped, using the same argument as in (23) and (24).

Remark 2. *Spurious local minima are much more challenging to be escaped than saddle points and local maxima. In Theorem 3, we assume that there are no saddle points or maxima outside of a certain region containing the local minima (i.e. $\cup_{i\in\mathcal{I}} B(y_i, R)$). We do so in order to focus on the main contribution of this work, which is that time variation can lead to the absence of spurious local trajectories. Without this assumption, a significant part of the proof would deal with escaping saddle points, a subject which has already been treated in various papers [38], [53]–[56]. If the variation of the data occurs along a direction u chosen randomly, then it may be argued that the trajectory would escape saddle points with probability 1, using the stable manifold theorem [57] as in [38], [53]–[56]. Theorem 3 would then hold almost surely.*

Remark 3. *Theorem 3 offers the first result in the literature about when spurious minima of a time-invariant function can be escaped via a time-varying deterministic local search method. The existing results are focused on stochastic gradient descent that offers a weaker result in a probabilistic sense [20]. This theorem can be used to define the notion of escapable local minima through the parameters C_1 and C_2, and indeed if C_1 is small enough and C_2 is large enough, the spurious local minima can always be escaped based on the results of this theorem.*

Although Theorem 3 is focused on a certain class of time-varying functions, similar results can be obtained for other classes of functions. The time-varying problem (4) is devoid of spurious local trajectories if one can show that all solutions of (6) with the initial point at any local solutions at $t = 0$ are contractive and the converging trajectory is inside the region of attraction of the global minimum trajectory of (4) after some finite time. This can be studied via the contraction analysis of nonlinear systems [58]–[60].

V. Fundamental Properties of ODE

In this section, we provide the formal versions of Theorems 1 and 2 together with their proofs. We refer to the optimization problem (5) as $\mathrm{OPT}(k, \Delta t, x_{k-1})$. Let the Jacobian of the constraint set be defined as

$$\mathcal{J}(x) = \begin{bmatrix} \nabla_x h_1(x)^\top \\ \nabla_x h_2(x)^\top \\ \vdots \\ \nabla_x h_r(x)^\top \end{bmatrix} \tag{27}$$

Definition 5. *Given a feasible initial point x_0, we say that the tuple $\left(x_0, \Delta t, \{x_k^{\Delta t}\}_{k=0}^\infty\right)$ is an **admissible KKT (AKKT) tuple** if $x_0^{\Delta t} = x_0$ and for every $k \in \{0, 1, ...\}$, $x_k^{\Delta t}$ is a feasible solution of $\mathrm{OPT}(k, \Delta t, x_{k-1}^{\Delta t})$, it satisfies the KKT conditions, and $\mathcal{J}(x_k^{\Delta t})$ is non-singular.*

Assumption 3. *There exists $\bar{t} > 0$ such that any $0 < \Delta t \leq \bar{t}$ is endowed with at least one AKKT tuple $\left(x_0, \Delta t, \{x_k^{\Delta t}\}_{k=0}^\infty\right)$. Furthermore, for any AKKT tuple $\left(x_0, \Delta t, \{x_k^{\Delta t}\}_{k=0}^\infty\right)$, the sequence $\left\{x_0, \{x_k^{\Delta t}\}_{k=0}^\infty\right\}$ is uniformly bounded.*

Roughly speaking, Assumption 3 implies that, for sufficiently small time steps, the regularized problem remains feasible with non-degenerate and bounded solutions.

According to Definition 5, the Jacobian matrix $\mathcal{J}(x_k^{\Delta t})$ is non-singular for every k and every AKKT tuple $\left(x_0, \Delta t, \{x_k^{\Delta t}\}_{k=1}^\infty\right)$. In this work, we impose a slightly stronger condition on the singular values of $\mathcal{J}(x_k^{\Delta t})$.

Assumption 4. *There exists a universal constant $c > 0$ such that $\sigma_{\min}(\mathcal{J}(x_k^{\Delta t})) \geq c$ for every k and every AKKT tuple $\left(x_0, \Delta t, \{x_k^{\Delta t}\}_{k=0}^\infty\right)$.*

Similar to Assumption 2, this assumption requires the constraints to be non-degenerate. Now, we are ready to present our main theorem.

Theorem 4. *Consider the ODE (6) with the condition $x(0) = x_0$, where x_0 is a local solution to the time-varying optimization (4) at $t = 0$. The following statements hold:*

1. *(Existence and uniqueness) (4) has a continuously differentiable and unique solution $x : [0, T] \to \mathbb{R}^n$.*
2. *(Convergence) Any AKKT tuple $\left(x_0, \Delta t, \{x_k^{\Delta t}\}_{k=0}^{\lceil T/\Delta t \rceil}\right)$ satisfies*

$$\lim_{\Delta t \to 0^+} \sup_{0 \leq k \leq \lceil T/\Delta t \rceil} \|x_k^{\Delta t} - x(k\Delta t)\| = 0, \tag{28}$$

We will regularly refer to the following lemma in our subsequent analysis.

Lemma 1 (Lipschitz property on a ball)**.** *Given a continuously differentiable function $p(x) : \mathbb{R}^n \to \mathbb{R}^m$, we have*

$$\|p(x) - p(y)\| \leq L(\epsilon)\|x - y\| \qquad \textit{for every } x, y \in \mathcal{B}(\epsilon)$$

where $L(\epsilon)$ is a universal constant independent of x and y, and $\mathcal{B}(\epsilon)$ is the Euclidean ball centered at zero with radius ϵ.

Proof. The proof is straightforward and omitted. □

A. Proof of Existence and Uniqueness.

Next, we show the existence and uniqueness of the solution to the proposed ODE. Without loss of generality, we assume that $t_k - t_{k-1} = \Delta t$ for every $k = 1, \ldots, \lceil T/\Delta t \rceil$. Furthermore, to simplify the notation, we may use the same symbols to refer to different universal constants throughout the proofs. The next three lemmas will be useful in proving the existence of a solution (6).

Lemma 2. *There exist constants $\bar{t}$ and $c > 0$ such that for every AKKT tuple $\left(x_0, \Delta t, \{x_k^{\Delta t}\}_{k=0}^{\lceil T/\Delta t \rceil}\right)$ with $\Delta t \leq \bar{t}$, we have $\|x_k^{\Delta t} - x_{k-1}^{\Delta t}\| \leq c\Delta t$ for $k = 1, \ldots, \infty$.*

Proof. The proof is provided in the appendix. □

Lemma 3. *Given an initial feasible point x_0, there exist*

1. *$\{s_n\}_{n=1}^{\infty}$ with $\lim_{n\to\infty} s_n = 0$ such that each s_n is endowed with an AKKT tuple $(x_0, s_n, \{x_k^{s_n}\}_{k=0}^{\infty})$, and*
2. *a continuously differentiable and uniformly bounded function $\bar{x} : [0, T] \to \mathbb{R}^n$ that satisfies $\bar{x}(0) = x_0$,*

with the following properties:

$$\lim_{n\to\infty} \sup_{1\leq k\leq \frac{T}{s_n}} \|x_k^{s_n} - \bar{x}(ks_n)\| = 0, \tag{29a}$$

$$\lim_{n\to\infty} \sup_{1\leq k\leq \frac{T}{s_n}} \left\| \frac{x_k^{s_n} - x_{k-1}^{s_n}}{s_n} - \dot{\bar{x}}(ks_n) \right\| = 0. \tag{29b}$$

Moreover, there exists a universal constant $c > 0$ such that $\sigma_{\min}(\mathcal{J}(\bar{x}(t))) \geq c$ for every $t \in [0, T]$.

Proof. The proof is provided in the appendix. □

Lemma 4. *Consider two continuous functions $g_1 : [0, T] \to \mathbb{R}^n$ and $g_2 : [0, T] \to \mathbb{R}^n$. We have $g_1 = g_2$ if and only if*

$$\lim_{\Delta t\to 0^+} \sup_{0\leq k\leq \lceil \frac{T}{\Delta t}\rceil} \|g_1(k\Delta t) - g_2(k\Delta t)\| = 0 \tag{30}$$

Proof. The proof is straightforward and can be found in standard references, e.g., [61]. □

We now provide the proof for the existence and uniqueness of the solution for (6).

Proof of existence and uniqueness: Consider the sequence $\{s_n\}_{n=1}^{\infty}$ and its corresponding AKKT tuple $\left\{(x_0, s_n, \{x_k^{s_n}\}_{k=0}^{T/s_n})\right\}_{n=1}^{\infty}$ that is introduced in Lemma 3. Due to Assumption 4, the linear independence constraint qualification (LICQ) holds at $x_k^{s_n}$ for $k = 0, \ldots, T/s_n$ and $n = 1, \ldots, \infty$. Therefore, for every n, there exists a sequence of Lagrangian vectors $\{\mu_k^{s_n}\}_{k=0}^{T/s_n}$ such that $\left(\{x_k^{s_n}\}_{k=0}^{T/s_n}, \{\mu_k^{s_n}\}_{k=0}^{T/s_n}\right)$ satisfies the KKT conditions:

$$\nabla_x f_k(x_k^{s_n}) + \mathcal{J}(x_k^{s_n})^\top \mu_k^{s_n} + \frac{\alpha}{s_n}(x_k^{s_n} - x_{k-1}^{s_n}) = 0 \tag{Stationarity}$$

$$h_i(x_k^{s_n}) = d_{i,k} \tag{feasibility}$$

for $k = 1, \ldots, T/s_n$, where $f_k(x_k^{s_n}) = f(x_k^{s_n}, ks_n)$ and $d_{i,k} = d_i(ks_n)$. The feasibility condition implies that for every i, we have

$$\begin{aligned}
&\frac{1}{s_n}\left(h_i(x_k^{s_n}) - h_i(x_{k-1}^{s_n})\right) = \frac{d_{i,k} - d_{i,k-1}}{s_n} \\
&\implies \nabla h_i(\tilde{x}_{i,k}^{s_n})^\top \left(\frac{x_k^{s_n} - x_{k-1}^{s_n}}{s_n}\right) = \frac{d_{i,k} - d_{i,k-1}}{s_n}
\end{aligned} \tag{31}$$

for some $\tilde{x}_{i,k}^{s_n} = (1-\alpha_i)x_k^{s_n} + \alpha_i x_{k-1}^{s_n}$ with $\alpha_i \in [0, 1]$, where the last implication is due to the differentiability of $h_i(x)$ and the Mean Value Theorem. For simplicity and with a slight abuse of notation, define

$$\mathcal{J}(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r) = \begin{bmatrix} \nabla h_1(\tilde{x}_{1;k}^{s_n})^\top \\ \vdots \\ \nabla h_r(\tilde{x}_{r;k}^{s_n})^\top \end{bmatrix}, \quad d_k = \begin{bmatrix} d_{1,k} \\ \vdots \\ d_{r,k} \end{bmatrix} \tag{32}$$

This implies that

$$\mathcal{J}(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r)\left(\frac{x_k^{s_n} - x_{k-1}^{s_n}}{s_n}\right) = \frac{d_k - d_{k-1}}{s_n} \tag{33}$$

Combining this equality with the stationarity condition leads to

$$\begin{aligned}
&\mathcal{J}(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r)\nabla_x f_k(x_k^{s_n}) + \mathcal{J}(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r)\mathcal{J}(x_k^{s_n})^\top \lambda_k^{s_n} \\
&\qquad + \alpha\left(\frac{d_k - d_{k-1}}{s_n}\right) = 0
\end{aligned} \tag{34}$$

Now, note that, due to Assumption 4, $\sigma_{\min}(\mathcal{J}(x_k^{s_n})) \geq c$ for some universal constant $c > 0$. Therefore, for every y sufficiently close to $x_k^{s_n}$, $\mathcal{J}(y)$ remains full-row rank. Together with the definition of $\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r$ and Lemma 7 in the appendix, this implies that $\mathcal{J}(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r)\mathcal{J}(x_k)^\top$ is invertible for sufficiently small Δt. Therefore,

$$\begin{aligned}
\lambda_k^{s_n} = &-\left(\mathcal{J}(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r)\mathcal{J}(x_k^{s_n})^\top\right)^{-1} \\
&\times \left(\mathcal{J}(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r)\nabla_x f_k(x_k^{s_n}) + \alpha\left(\frac{d_{i;k} - d_{i;k-1}}{s_n}\right)\right)
\end{aligned} \tag{35}$$

Substituting this into the stationarity condition and performing the necessary simplifications lead to

$$\begin{aligned}
&\frac{x_k^{s_n} - x_{k-1}^{s_n}}{s_n} = -\frac{1}{\alpha}\Bigg(I - \mathcal{J}(x_k^{s_n})^\top \\
&\times \left(\mathcal{J}(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r)\mathcal{J}(x_k^{s_n})^\top\right)^{-1} \mathcal{J}(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r)\Bigg)\nabla_x f_k(x_k^{s_n}) \\
&+ \mathcal{J}(x_k^{s_n})^\top \left(\mathcal{J}(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r)\mathcal{J}(x_k^{s_n})^\top\right)^{-1}\left(\frac{d_k - d_{k-1}}{s_n}\right) \\
&:= g\left(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r, x_k^{s_n}, \left(\frac{d_k - d_{k-1}}{s_n}\right)\right)
\end{aligned} \tag{36}$$

Consider the continuously differentiable function $\bar{x}(t)$ that is introduced in Lemma 3. The above equality together with (29b) implies that

$$\begin{aligned}
\lim_{n\to\infty} \sup_{0\leq k\leq \lceil \frac{T}{s_n}\rceil} \Bigg\| &\dot{\bar{x}}(ks_n) \\
&- g\Big(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r, x_k^{s_n}, \Big(\frac{d_k - d_{k-1}}{s_n}\Big)\Big)\Bigg\| = 0
\end{aligned} \tag{37}$$

Therefore, one can write

$$\lim_{n\to\infty}\sup_{0\le k\le\lceil\frac{T}{s_n}\rceil}\left\|\dot{\bar{x}}(ks_n) - g\left(\{\bar{x}(ks_n)\}_{i=1}^r,\bar{x}(ks_n),\dot{d}(ks_n)\right)\right\| \tag{38}$$
$$\le \lim_{n\to\infty}\sup_{0\le k\le\lceil\frac{T}{s_n}\rceil}\left\|\dot{\bar{x}}(ks_n) - g\left(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r,x_k^{s_n},\left(\frac{d_k-d_{k-1}}{s_n}\right)\right)\right\|$$
$$+\lim_{n\to\infty}\sup_{0\le k\le\lceil\frac{T}{s_n}\rceil}\left\|g\left(\{\bar{x}(ks_n)\}_{i=1}^r,\bar{x}(ks_n),\dot{d}(ks_n)\right) - g\left(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r,x_k^{s_n},\left(\frac{d_k-d_{k-1}}{s_n}\right)\right)\right\|$$

We present the following lemma.

Lemma 5. *Given* $(\{\bar{x}_i\}_{i=1}^r,\bar{y},\bar{z})$ *with* $(\sum_{i=1}^r\|\bar{x}_i\|)+\|\bar{y}\|+\|\bar{z}\|\le c_1$ *for some* $c_1>0$*, suppose that* $\sigma_{\min}\left(\mathcal{J}(\{\bar{x}_i\}_{i=1}^r)\mathcal{J}(\bar{y})^\top\right)\ge c_2$ *for some* $c_2>0$*. Then, there exist constants* $L,r>0$*, such that* $g(\{\bar{x}_i\}_{i=1}^r,\bar{y},\bar{z})$ *is locally* L*-Lipschitz continuous in* $\mathcal{B}=\{(\{x_i\}_{i=1}^r,y,z)\mid(\sum_{i=1}^r\|\bar{x}_i-x_i\|)+\|\bar{y}-y\|+\|\bar{z}-z\|\le r\}$.

Proof. Due to the continuous differentiability of $\mathcal{J}(x)$ and Lemma 1, it is easy to see that r can be chosen such that $\sigma_{\min}\left(\mathcal{J}(\{x_i\}_{i=1}^r)\mathcal{J}(y)^\top\right)\ge c_2/2$ for every $(\{x_i\}_{i=1}^r,y,z)\in\mathcal{B}(r)$. This observation, together with the definition of $g(\cdot,\cdot,\cdot)$ in (36), can be used to complete the proof. The details are omitted for brevity. $\square$

According to Lemma 5, the function $g(\cdot,\cdot,\cdot)$ is locally Lipschitz continuous on a ball with nonzero radius and centered at $\left(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r,x_k^{s_n},\left(\frac{d_k-d_{k-1}}{s_n}\right)\right)$ for every $0\le k\le\lceil\frac{T}{\Delta t}\rceil$ and $n=1,\dots,\infty$. This together with the definition of $\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r$, the differentiability of $d(t)$, and Lemma 3 implies that for sufficiently large n (or, equivalently, for sufficiently small s_n), there exists a Lipschitz constant L such that

$$\left\|g\left(\{\bar{x}(ks_n)\}_{i=1}^r,\bar{x}(ks_n),\dot{d}(ks_n)\right) - g\left(\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r,x_k^{s_n},\left(\frac{d_k-d_{k-1}}{s_n}\right)\right)\right\| \tag{39}$$
$$\le L\left(\sum_{i=1}^r\|\bar{x}(ks_n)-\tilde{x}_{i,k}^{s_n}\|+\|\bar{x}(ks_n)-x_k^{s_n}\| + \left\|\dot{d}(ks_n)-\left(\frac{d_k-d_{k-1}}{s_n}\right)\right\|\right)$$
$$\le L\Bigg((r+1)\|\bar{x}(ks_n)-x_k^{s_n}\|+r\|\bar{x}((k-1)s_n)-x_{k-1}^{s_n}\|+ r\|\bar{x}(ks_n)-\bar{x}((k-1)s_n)\|+\left\|\dot{d}(ks_n)-\left(\frac{d_k-d_{k-1}}{s_n}\right)\right\|\Bigg)$$

where we used the definition of $\{\tilde{x}_{i,k}^{s_n}\}_{i=1}^r$ and triangle inequality. According to Lemmas 2 and 3, the right-hand side of (39) converges to zero as $n\to\infty$. Therefore, combining (39) and (37) with (38) implies that

$$\lim_{n\to\infty}\sup_{0\le k\le\lceil\frac{T}{s_n}\rceil}\left\|\dot{\bar{x}}(ks_n) - g\left(\{\bar{x}(ks_n)\}_{i=1}^r,\bar{x}(ks_n),\dot{d}(ks_n)\right)\right\|=0 \tag{40}$$

Furthermore, due to Lemma 3, $\mathcal{J}(\bar{x}(t))$ is full-row rank at every $t\in[0,T]$ and therefore, $g(\{\bar{x}(t)\}_{i=1}^r,\bar{x}(t),\dot{d}(t))$ is continuous as a function of t in $[0,T]$. Invoking Lemma 4 then leads to

$$\dot{\bar{x}}(t)=g(\{\bar{x}(t)\}_{i=1}^r,\bar{x}(t),\dot{d}(t)) \tag{41}$$

at every $t\in[0,T]$. This shows that $\bar{x}:[0,T]\to\mathbb{R}^n$ is a solution to (6). Finally, due to Lemma 3, we have $\sigma_{\min}(\mathcal{J}(\bar{x}(t)))\ge c$ for a universal constant $c>0$. Therefore, Lemma 5 can be used to verify the existence of an open and connected set $\mathcal{D}$ such that $g(\cdot,\cdot,\cdot)$ is locally L-Lipschitz continuous on $\mathcal{D}$ and $(\bar{x}(t),t)\in\mathcal{D}$ for every $t\in[0,T]$. Therefore, Theorem 2.2 in [44] can be used to show that $\bar{x}:[0,T]\to\mathbb{R}^n$ is the unique solution to (6). $\square$

B. Proof of Convergence

Next, we show the validity of the second statement in Theorem 4.

Lemma 6 (Backward Euler Iterations)**.** *There exists a universal constant* $\bar{t}$ *such that for every* $\Delta t\le\bar{t}$*, there exists a sequence* $\{y_k^{\Delta t}\}_{k=0}^{\lceil T/\Delta t\rceil}$ *that satisfies the following statements:*

- *We have* $y_0^{\Delta t}=x_0$ *and*
$$y_k^{\Delta t}=y_{k-1}^{\Delta t}+\Delta t\cdot g\left(\{y_k^{\Delta t}\}_{i=1}^r,y_k^{\Delta t},\dot{d}(s_k)\right) \tag{42}$$
for $k=1,\dots,\lceil T/\Delta t\rceil$.
- *There exists a universal constant* $c_2>0$ *such that* $\|y_k^{\Delta t}-y_{k-1}^{\Delta t}\|\le c_2\Delta t$ *for* $k=1,\dots,\lceil T/\Delta t\rceil$.
- *We have*
$$\lim_{\Delta t\to 0^+}\sup_{0\le k\le\lceil T/\Delta t\rceil}\|y_k^{\Delta t}-x(s_k)\|=0 \tag{43}$$
where $x:[0,T]\to\mathbb{R}^n$ *is the unique solution to* (6)*.*
- *We have* $\sigma_{\min}(\mathcal{J}(y_k^{\Delta t}))\ge c_1$ *for some universal* c_1 *and every* $k=1,\dots,\lceil T/\Delta t\rceil$.

Proof. Note that (42) is the backward Euler iterations for (6) [62]. Furthermore, we have already shown the existence of a continuously differentiable and uniformly bounded solution to (6). The proof of the first three statements is immediately followed by the classical results on convergence of the backward Euler method; see [62] for more details. To verify the correctness of the last statement, note that we have shown in the previous subsection that the function $\bar{x}:[0,T]\to\mathbb{R}^n$ introduced in Lemma 3 is indeed the unique solution to the proposed ODE and we have $\mathcal{J}(\bar{x}(t))\ge c$ for some universal $c>0$ and every $t\in[0,T]$. This together with (43) and Lemma 1 concludes the proof. $\square$

Proof of convergence: The main idea behind the proof is to show that, given any AKKT tuple $\left(x_0, \Delta t, \{x_k^{\Delta t}\}_{k=1}^{\lceil T/\Delta t \rceil}\right)$, we have

$$\lim_{\Delta t \to 0^+} \sup_{0 \le k \le \lceil T/\Delta t \rceil} \|y_k^{\Delta t} - x_k^{\Delta t}\| = 0 \tag{44}$$

Establishing this equality together with Lemma 6 is enough to complete the proof.

It is evident from (36) that the AKKT tuple $\left(x_0, \Delta t, \{x_k^{\Delta t}\}_{k=1}^{\lceil T/\Delta t \rceil}\right)$ should satisfy

$$x_k^{\Delta t} = x_{k-1}^{\Delta t} + \Delta t g\left(\{\tilde{x}_{i,k}^{\Delta t}\}_{i=1}^{r}, x_k^{\Delta t}, \left(\frac{d_k - d_{k-1}}{\Delta t}\right)\right) \tag{45}$$

where $\tilde{x}_{i;k}^{t_n} = (1-\alpha_i)x_k^{t_n} + \alpha_i x_{k-1}^{t_n}$ with $\alpha_i \in [0,1]$ for $i = 1, \ldots, n$. Combined with the first statement of Lemma 6, this implies that

$$\begin{aligned} x_k^{\Delta t} - y_k^{\Delta t} &= x_{k-1}^{\Delta t} - y_{k-1}^{\Delta t} \\ &+ \Delta t \Bigg(g\left(\{\tilde{x}_{i,k}^{\Delta t}\}_{i=1}^{r}, x_k^{\Delta t}, \left(\frac{d_k - d_{k-1}}{\Delta t}\right)\right) \\ &- g\left(\{y_k^{\Delta t}\}_{i=1}^{r}, y_k^{\Delta t}, \dot{d}(s_k)\right) \Bigg) = x_{k-1}^{\Delta t} - y_{k-1}^{\Delta t} + A + B \end{aligned} \tag{46}$$

where

$$\begin{aligned} A = &\Delta t \times \Bigg(g\left(\{\tilde{x}_{i,k}^{\Delta t}\}_{i=1}^{r}, x_k^{\Delta t}, \left(\frac{d_k - d_{k-1}}{\Delta t}\right)\right) \\ &- g\left(\{y_{k-1}^{\Delta t}\}_{i=1}^{r}, y_{k-1}^{\Delta t}, \dot{d}(s_k)\right) \Bigg), \end{aligned} \tag{47a}$$

$$\begin{aligned} B = &\Delta t \times \Bigg(g\left(\{y_{k-1}^{\Delta t}\}_{i=1}^{r}, y_{k-1}^{\Delta t}, \dot{d}(s_k)\right) \\ &- g\left(\{y_k^{\Delta t}\}_{i=1}^{r}, y_k^{\Delta t}, \dot{d}(s_k)\right) \Bigg). \end{aligned} \tag{47b}$$

Define $E_k = \|x_k^{\Delta t} - y_k^{\Delta t}\|$ as the error at time-step k. Note that, due to the Lemmas 3, 6, and 5, as well as the construction of $\{\tilde{x}_{i,k}^{\Delta t}\}_{i=1}^{r}$, there exist universal constants $L, \bar{c}, \bar{t} > 0$ such that, for every $\Delta t \le \bar{t}$, $g(\cdot,\cdot,\cdot)$ is locally L-Lipschitz continuous in the balls

$$\begin{aligned} \mathcal{B}_1 = \Bigg\{ (\{x_i\}_{i=1}^{r}, y, z) \;\Bigg|\; &\left(\sum_{i=1}^{r} \|\tilde{x}_{i,k}^{\Delta t} - x_i\|\right) \\ &+ \|\tilde{x}_k^{\Delta t} - y\| + \left\|\left(\frac{d_k - d_{k-1}}{\Delta t}\right) - z\right\| \le \bar{c} \Bigg\}, \end{aligned} \tag{48}$$

and

$$\begin{aligned} \mathcal{B}_2 = \Bigg\{ (\{x_i\}_{i=1}^{r}, y, z) \;\Bigg|\; &\left(\sum_{i=1}^{r} \|y_k^{\Delta t} - x_i\|\right) \\ &+ \|y_k^{\Delta t} - y\| + \left\|\dot{d}(s_k) - z\right\| \le \bar{c} \Bigg\}. \end{aligned} \tag{49}$$

To simplify the notation, we denote $\left\|\left(\frac{d_k - d_{k-1}}{\Delta t}\right) - \dot{d}(s_k)\right\|$ as D. The following chain of inequalities will be useful in bounding the expression A in (46):

$$\begin{aligned} &\left(\sum_{i=1}^{r} \left\|\tilde{x}_{i,k}^{\Delta t} - y_{k-1}^{\Delta t}\right\|\right) + \left\|x_k^{\Delta t} - y_{k-1}^{\Delta t}\right\| + D \\ &\le r\left\|x_{k-1}^{\Delta t} - y_{k-1}^{\Delta t}\right\| + (r+1)\left\|x_k^{\Delta t} - y_{k-1}^{\Delta t}\right\| + D \\ &\le r\left\|x_{k-1}^{\Delta t} - y_{k-1}^{\Delta t}\right\| + (r+1)\left\|x_k^{\Delta t} - x_{k-1}^{\Delta t}\right\| \\ &\quad + (r+1)\left\|x_{k-1}^{\Delta t} - y_{k-1}^{\Delta t}\right\| + D \\ &= (2r+1)E_{k-1} + (r+1)\left\|x_k^{\Delta t} - x_{k-1}^{\Delta t}\right\| + D \\ &\le (2r+1)E_{k-1} + (r+1)c_1\Delta t + c_2\Delta t^2 \\ &\le (2r+1)E_{k-1} + ((r+1)c_1 + c_2)\Delta t \end{aligned} \tag{50}$$

provided that $\Delta t \le \bar{t}_1$, where $\bar{t}_1, c_1, c_2 > 0$ are constants. Note that the last two inequalities are due to Lemma 2 and the twice differentiability of $d(t)$.

Subsequently, the next inequality will be used to bound the expression B in (46). In particular, Lemma 6 can be used to show the existence of constants $c_3, \bar{t}_2 > 0$ such that

$$(r+1)\|y_{k-1}^{\Delta t} - y_k^{\Delta t}\| \le c_3\Delta t \tag{51}$$

provided that $\Delta t \le \bar{t}_2$. Given the inequalities (50) and (51), we prove the validity of (28) by proving the following statements:

1. There exists a universal constant $\bar{t}_3$ such that for every $\Delta t \le \bar{t}_3$ and $k = 0, \ldots, T/\Delta t$, (50) and (51) will be upper bounded by $\bar{c}$ which is defined as the radius of the balls (48) and (49). This together with the locally L-Lipschitz continuity of $g(\cdot,\cdot,\cdot)$ within the balls $\mathcal{B}_1$ and $\mathcal{B}_2$ leads to

$$\|A\| \le (2r+1)L\Delta t E_{k-1} + ((r+1)c_1 + c_2)L\Delta t^2 \tag{52a}$$

$$\|B\| \le c_3 L\Delta t^2 \tag{52b}$$

Combining these inequalities with (46) results in the following recursive inequality:

$$\begin{aligned} E_k \le &(1 + (2r+1)L\Delta t)E_{k-1} \\ &+ ((r+1)c_1 + c_2 + c_3)\, L\Delta t^2 \end{aligned} \tag{53}$$

2. We have $\lim_{\Delta t \to 0^+} \sup_{0 \le k \le T/\Delta t} E_k = 0$.

We prove the first statement using an inductive argument on k. In particular, we show that if the following inequality holds

$$\begin{aligned} \Delta t \le \min\Bigg\{ &\bar{t}_1, \bar{t}_2, \frac{(2r+1)\bar{c}}{((r+1)c_1 + c_2 + c_3)(e^{(2r+1)TL} - 1)}, \\ &, \sqrt{\frac{\bar{c}}{((r+1)c_1 + c_2 + c_3)\, L}}, \Bigg\} = \bar{t}_3 \end{aligned} \tag{54}$$

then (50) and (51) remain in the balls $\mathcal{B}_1$ and $\mathcal{B}_2$, respectively and hence, (53) holds for $k = 0, \ldots, T/\Delta t$.

Base case: $\mathbf{k = 1}$. Note that in this case, $E_0 = 0$ and therefore, based on (54), we have $\Delta t \le \bar{t}_1$ and $\Delta t \le \bar{t}_2$. This implies that both (50) and (51) are upper bounded by $\bar{c}$ and, based on (53), we have

$$\begin{aligned} E_1 &\le (1 + (2r+1)L\Delta t)E_0 + ((r+1)c_1 + c_2 + c_3)\, L\Delta t^2 \\ &= ((r+1)c_1 + c_2 + c_3)\, L\Delta t^2 \le \bar{c} \end{aligned} \tag{55}$$

where the last inequality is due to (54).

Inductive step. Suppose that we have

$$(2r+1)E_{k-1} + ((r+1)c_1 + c_2)\Delta t \le \bar{c}, \quad c_3\Delta t \le \bar{c} \tag{56a}$$

for $k = 0, \ldots, m-1$. This implies that (53) holds for $k = 1, \ldots, m$. With some algebra, one can verify that

$$\begin{aligned} E_m &\le ((r+1)c_1 + c_2 + c_3)\, L\Delta t^2 \sum_{i=0}^{m-1} (1 + (2r+1)L\Delta t)^i \\ &\le ((r+1)c_1 + c_2 + c_3)\, L\Delta t^2 \cdot \frac{(1+(2r+1)L\Delta t)^m - 1}{(2r+1)L\Delta t} \\ &\le \frac{(r+1)c_1 + c_2 + c_3}{2r+1} \left((1+(2r+1)L\Delta t)^{T/\Delta t} - 1 \right) \Delta t \\ &\le \frac{(r+1)c_1 + c_2 + c_3}{2r+1} \left(e^{(2r+1)LT} - 1 \right) \Delta t \le \bar{c} \end{aligned} \tag{57}$$

which completes the proof of the first statement. To prove the second statement, note that the above analysis leads to

$$\sup_{0 \le k \le T/\Delta t} E_k \le \frac{(r+1)c_1 + c_2 + c_3}{2r+1} \left(e^{(2r+1)LT} - 1 \right) \Delta t$$

assuming that $\Delta t \le \bar{t}_3$. Due to the fact that $\bar{t}_3 > 0$ and is independent of Δt, we have

$$\lim_{\Delta t \to 0^+} \sup_{0 \le k \le T/\Delta t} E_k = 0 \tag{58}$$

thereby completing the proof of the convergence. □

VI. Properties of System's Jacobian

In this section, we additionally assume that the objective function $f(x,t)$ is twice continuously differentiable in x. For the constraint functions $h = (h_1, h_2, \ldots, h_m)$, the corresponding Hessian matrices $H_1, H_2, \ldots, H_m \in \mathbb{R}^{n\times n}$ are the second partial derivative of h with respect to x. The second-order derivative operator of h, denoted by H, is now regarded as the m-tuple $H = (H_1, \ldots, H_m)$. For $\mu \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$, μH denotes $\mu_1 H_1 + \ldots + \mu_m H_m$ and $x^\top H x$ denotes $x^\top H_1 x + \ldots + x^\top H_m x$. For $M_1, M_2 \in \mathbb{R}^{n\times n}$, $M_1 H M_2 x$ denotes $[M_1 H_1 M_2 x, \ldots, M_1 H_m M_2 x]$. In addition, we have the identity $\mu x^\top H x = x^\top \mu H x$.

Consider the time-invariant optimization problem:

$$\inf_{x \in \mathbb{R}^n} f(x) \quad \text{s.t.} \quad h(x) = d \tag{59}$$

where $h(x) = [h_1(x), \ldots, h_m(x)]^T$ and $d = [d_1, \ldots, d_m]^T$. The corresponding ODE is given by

$$\dot{x} = -\frac{1}{\alpha}\left[I - \mathcal{J}(x)^\top (\mathcal{J}(x)\mathcal{J}(x)^\top)^{-1}\mathcal{J}(x) \right] \nabla f(x). \tag{60}$$

The above ODE is known as the Riemannian gradient flow, and it is well-studied in the literature [46]–[48]. Let z be a local minimum of (59) satisfying the first-order necessary and second-order sufficient optimality conditions:

$$h(z) = d,\ \mathcal{J}(z)\mathcal{J}(z)^\top \text{ is invertible} \tag{61a}$$

$$\nabla f(z) + \mu \mathcal{J}(z) = 0,\ w^\top \left(\nabla^2 f(z) + H(z) \right) w > 0 \tag{61b}$$

for some $\mu \in \mathbb{R}^m$ and every nonzero vector w such that $\mathcal{J}(z)^\top w = 0$. Note that z is an equilibrium point of the system (60). Let the right-hand side of (60) be denoted by $p(x)$:

$$p(x) := -\frac{1}{\alpha}\mathcal{P}(x)\nabla_x f(x) \tag{62}$$

where $\mathcal{P}(x) = I - \mathcal{J}(x)^\top(\mathcal{J}(x)\mathcal{J}(x)^\top)^{-1}\mathcal{J}(x)$ and let $\mathcal{J}_p(z)$ denote the Jacobian of $p(x)$.

Theorem 5. *It holds that*

$$\mathcal{J}_p(z) = -\frac{1}{\alpha}\left(\nabla^2 f(z) + \mu H(z) \right) \mathcal{P}(z). \tag{63}$$

Moreover, $\mathcal{J}_p(z)$ has $n-m$ eigenvalues with negative real parts and m zero eigenvalues.

Proof. The equation (63) follows from Corollary 1 in [63]. To study the eigenvalues of $\mathcal{J}_p(z)$, note that $\mathcal{J}_p(z)\mathcal{J}(z)^\top = 0$. Therefore, $\mathcal{J}_p(z)$ has at least m zero eigenvalues. Let $w \in \mathbb{R}^n$ be an arbitrary nonzero vector in the tangent plane of the manifold $\{x\ :\ h(x) = d\}$ at the point $x = z$. This means that $\mathcal{J}(z)^\top w = 0$. On the other hand, the second-order sufficient optimality condition states that $w^\top \left(\nabla^2 f(z) + \mu H(z) \right) w > 0$. Therefore, we have $w^\top \Omega w > 0$, where

$$\Omega = \mathcal{P}(z) \left(\nabla^2 f(z) + \mu H(z) \right) \mathcal{P}(z). \tag{64}$$

Since $\mathcal{J}(z)$ is in the null space of the symmetric matrix Ω and $w^\top \Omega w > 0$ for every w that is orthogonal of $\mathcal{J}(z)$, it can be concluded that Ω has $n-m$ eigenvalues with positive real parts. On the other hands, the eigenvalues of Ω are the same of the eigenvalues of the matrix

$$\left(\nabla^2 f(z) + \mu H(z) \right) \mathcal{P}^2(x) = \left(\nabla^2 f(z) + \mu H(z) \right) \mathcal{P}(z) \tag{65}$$

which is the identical to $-\alpha \mathcal{J}_p(z)$. □

As shown above, the eigenvalues of the Jacobian only have non-positive real parts. This explains why spurious solutions of a time-invariant optimization problem cannot be escaped using gradient-based methods, such as the ODE (60). Now, consider its time-varying counterpart problem (4) and associated ODE (6). Let $z(t) : [0,T] \to \mathbb{R}^n$ be a local solution of (4) that satisfies the first-order necessary and second-order sufficient optimality conditions for all $t \in [0,T]$. Let $\mu(t)$ denote the corresponding Lagrange multiplier and $\mathcal{Q}(z(t))$ denote $\mathcal{J}(z(t))^\top(\mathcal{J}(z(t))\mathcal{J}(z(t))^\top)^{-1}$. Since $z(t)$ is generally not the solution of the ODE (6), we make a change of variables $e(t) = x(t) - z(t)$ to measure the distance between $x(t)$ and $z(t)$. Then, the ODE (6) can be rewritten as

$$\dot{e}(t) = -\frac{1}{\alpha}\eta(e(t) + z(t), t) + \theta(e(t) + z(t))\dot{d} - \dot{z}(t) \tag{66}$$

Let $\mathcal{J}_q(z(t))$ denote the Jacobian of the right-hand side of (66) at the point $e(t) = 0$. By taking the first-order approximation of (66) around $z(t)$, we have

$$\dot{e}(t) = \mathcal{J}_q(z(t))e(t) + O(e^2(t)) - \dot{z}(t). \tag{67}$$

Theorem 6. *It holds that*

$$\mathcal{J}_q(z(t)) = K_1(t) + K_2(t) \tag{68}$$

where

$$K_1(t) = -\frac{1}{\alpha}\left(\nabla^2 f(z(t)) + \mu(t)H(z(t))\right)\mathcal{P}(z(t)), \quad (69a)$$

$$K_2(t) = \Big(P(z(t))H(z(t))\left(\mathcal{J}(z(t))\mathcal{J}(z(t))^\top\right)^{-1} \quad (69b)$$

$$- \mathcal{Q}(z(t))H(z(t))\mathcal{Q}(z(t))\Big)\dot{d}(t). \quad (69c)$$

Proof. The computation of $K_1(t)$ is similar to that of Theorem 5. Because of the tensor nature of H it is convenient to differentiate with respect to each component separately. For the component $z_1(t)$, we have

$$\begin{aligned}
&\frac{d}{dz_1(t)}\mathcal{Q}(z(t))\dot{d}(t) = H_1(z(t))\left(\mathcal{J}(z(t))\mathcal{J}(z(t))^\top\right)^{-1}\dot{d}(t) \\
&- \mathcal{J}(z(t))^\top\left(\mathcal{J}(z(t))\mathcal{J}(z(t))^\top\right)^{-1}\Big(H_1(z(t))\mathcal{J}(z(t))^\top \\
&+ \mathcal{J}(z(t))H_1(z(t))\Big)\left(\mathcal{J}(z(t))\mathcal{J}(z(t))^\top\right)\dot{d}(t) \\
&= \Big(P(z(t))H_1(z(t))\left(\mathcal{J}(z(t))\mathcal{J}(z(t))^\top\right)^{-1} \\
&- \mathcal{Q}(z(t))H_1(z(t))\mathcal{Q}(z(t))\Big)\dot{d}(t).
\end{aligned}$$

Similar expressions apply to derivatives with respect to other components. These columns can be combined into the matrix

$$\Big[\frac{d}{dz_1(t)}\mathcal{Q}(z(t)), \ldots, \frac{d}{dz_n(t)}\mathcal{Q}(z(t))\Big]\dot{d}(t).$$

This matrix is $K_2(t)$. □

Notice that $K_1(t)$ has only eigenvalues with non-positive reals (due to Theorem 5) but $K_2(t)$ may have eigenvalues with positive reals depending on the time-variation. Thus, the time variation could potentially make the linear system $\dot{\bar{e}}(t) = \mathcal{J}_q(z(t))\bar{e}(t)$ unstable. If $O(e^2(t)) - \dot{z}(t)$ is not large, we may expect that the solution of (67) will behavior similarly to $\dot{\bar{e}}(t) = \mathcal{J}_q(z(t))\bar{e}(t)$ and cannot stay around the point 0. Thus, the time-variation may provide the opportunity to escape the spurious local trajectory $z(t)$. Note that the linearization does not always provide a concrete answer for time-varying ODEs, but this result offers an insight into how the data variation changes the eigenvalues of the Jacobian along a trajectory close to a KKT trajectory.

VII. Conclusion

In this work, we study the landscape of time-varying nonconvex optimization problems. We introduce the notion of spurious local trajectory as a counterpart to the notion of spurious local minima in the time-invariant optimization. The key insight to this new notion is the fact that a regularized version of the time-varying optimization problem is naturally endowed with an ordinary differential equation (ODE) at its limit. This close interplay enables us to study the solutions of this ODE to certify the absence of the spurious local trajectories in the problem. Through different case studies and theoretical results, we show that a time-varying optimization may have multiple spurious local minima, and yet its landscape can be free of spurious local trajectories. We further show that the variation of the landscape over time is the main reason behind the absence of spurious local trajectories.

As a future research direction, we will study the robustness of the solution trajectories against perturbations, along the same lines as [64]. Furthermore, it would be worthwhile to extend the notion of spurious local trajectories to time-varying optimization over an infinite-time horizon.

Appendix

Lemma 7. *We have* $\|x_k^{\Delta t} - x_{k-1}^{\Delta t}\| = O(\sqrt{\Delta t})$ *for every* $k = 0, \ldots, \lceil T/\Delta t\rceil$.

Proof. Note that $f(x,t)$ is uniformly bounded from below. Furthermore, for every AKKT tuple $\left(x_0, \Delta t, \{x_k^{\Delta t}\}_{k=0}^{\lceil T/\Delta t\rceil}\right)$, the sequence $\{x_k^{\Delta t}\}_{k=0}^{\lceil T/\Delta t\rceil}$ is assumed to be uniformly bounded. This together with Assumption 1 implies that

$$f(x_k^{\Delta t}, t_k) + \frac{\alpha}{2\Delta t}\|x_k^{\Delta t} - x_{k-1}^{\Delta t}\|^2 \leq R \quad (70)$$

for some $R < \infty$. Since $f(x_k^{\Delta t}, t_k)$ is assumed to be uniformly bounded from below, this leads to $\frac{\alpha}{2\Delta t}\|x_k^{\Delta t} - x_{k-1}^{\Delta t}\|^2 \leq R'$ for some $R' < \infty$, which in turn yields $\|x_k^{\Delta t} - x_{k-1}^{\Delta t}\| = O(\sqrt{\Delta t})$. □

Proof of Lemma 2. Due to Lemma 7 and the fact that $\mathcal{J}(x)$ is continuously differentiable, one can invoke Lemma 1 to show that there exist constants $\bar{t}, c_1, c_2 > 0$ such that the following statements hold, provided that $\Delta t \leq \bar{t}$:

1. Consider a sequence $\{\tilde{x}_{i,k}^{\Delta t}\}_{i=1}^r$ constructed similar to (33). Due to Assumption 4 and Lemma 7, it can be verified that there exist $\bar{t}, c_1 > 0$ such that $\sigma_{\min}(\mathcal{J}(\{\tilde{x}_{i,k}^{\Delta t}\}_{i=1}^r)\mathcal{J}(x_k^{\Delta t})^\top) \geq c_1$ for all $\Delta t \leq \bar{t}$. This implies that the function $g\left(\{\tilde{x}_{i,k}^{\Delta t}\}_{i=1}^r, x_k^{\Delta t}, \left(\frac{d_k - d_{k-1}}{\Delta t}\right)\right)$ introduced in (36) is well-defined and continuous for all $\Delta t \leq \bar{t}$.
2. Assumption 3 and twice differentiability of d with respect to t imply that $\left\{\{\tilde{x}_{i,k}^{\Delta t}\}_{i=1}^r, x_k^{\Delta t}\right\}$ and $\left(\frac{d_k - d_{k-1}}{\Delta t}\right)$ belong to a compact set. Combined with the continuity of $g(\cdot)$, this implies that
$$\left\|g\left(\{\tilde{x}_{i,k}^{\Delta t}\}_{i=1}^r, x_k^{\Delta t}, \left(\frac{d_k - d_{k-1}}{\Delta t}\right)\right)\right\| \leq c_2 \quad (71)$$
for some $c_2 > 0$.
3. Similar to (36), one can verify that the following equality holds:
$$\frac{x_k^{\Delta t} - x_{k-1}^{\Delta t}}{\Delta t} = g\left(\{\tilde{x}_{i,k}^{\Delta t}\}_{i=1}^r, x_k^{\Delta t}, \left(\frac{d_k - d_{k-1}}{\Delta t}\right)\right)$$

Combined with (71), this implies that $\|x_k^{\Delta t} - x_{k-1}^{\Delta t}\| \leq c_2\Delta t$ and the proof is complete. □

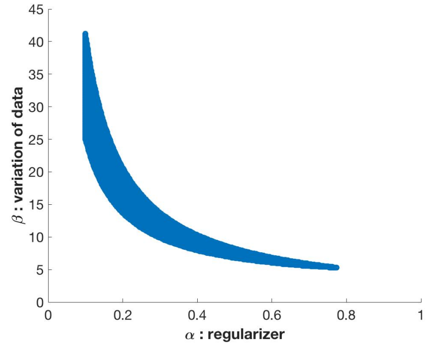
Proof of Lemma 3. Consider a sequence $\{s_n\}_{n=1}^\infty$ such that $s_n > 0$ and $\lim_{n\to\infty} s_n = 0$. Furthermore, without loss of generality, we assume that T/s_n is a natural number for every $n = 1, \ldots, \infty$. Given any n, consider a AKKT tuple $(x_0, s_n, \{x_k^{s_n}\}_{k=0}^\infty)$ and define a vector-valued function $\tilde{x}_{s_n} : [0,T] \to \mathbb{R}^n$ whose i^{th} element is the spline interpolation of the i^{th} elements of the vectors $\{x_0^{s_n}, x_1^{s_n}, \ldots, x_{T/s_n}^{s_n}\}$. Notice that this interpolation can be made in such a way that $\tilde{x}_{s_n}$ is continuously differentiable. We prove this lemma by showing that there exist a continuously differentiable function $\bar{x}$ and a

1. $$\alpha\beta \geqslant C := -255/256 + 259\sqrt{201}/768 - \sqrt{201}^3/1728$$

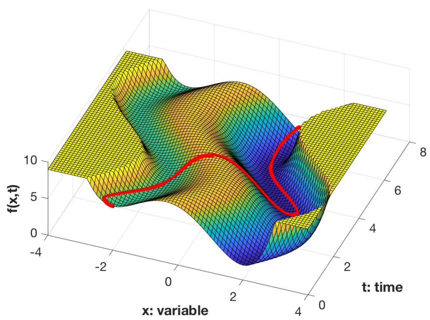
2. $$-27\alpha^2\beta^2 + 6885/128\alpha\beta + 61009/256 \geqslant 0$$

3. $$\begin{aligned} &-\tfrac{C}{\alpha}\left[2\pi - 2\arccos\left(-\tfrac{C}{\alpha\beta}\right)\right] + \dots \\ &-\beta\sin\left[2\pi - \arccos\left(-\tfrac{C}{\alpha\beta}\right)\right] + \beta\sin\left[\arccos\left(-\tfrac{C}{\alpha\beta}\right)\right] + \dots \\ &2\sqrt{\tfrac{259}{192}}\cos\left[\tfrac{1}{3}\arccos\left(\tfrac{48960-49152\alpha\beta}{132608}\sqrt{\tfrac{192}{259}}\right) - \tfrac{2k_1\pi}{3}\right] + \dots \\ &-2\sqrt{\tfrac{259}{192}}\cos\left[\tfrac{1}{3}\arccos\left(\tfrac{48960-49152\alpha\beta}{132608}\sqrt{\tfrac{192}{259}}\right) - \tfrac{2k_2\pi}{3}\right] \geqslant 0 \end{aligned}$$

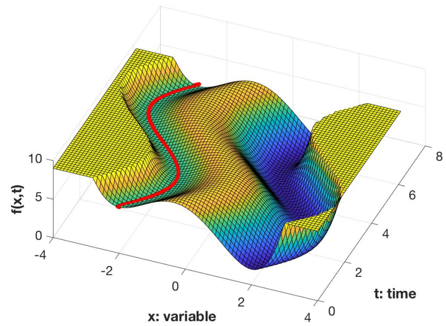
(a) Inequalities in function of α, β guaranteeing absence of spurious trajectories.



(b) Sufficient condition in blue in function of α, β for absence of spurious trajectories.



(c) Non-spurious trajectory for $\alpha = 0.4$ and $\beta = 10$.



(d) Spurious trajectory for $\alpha = 0.2$ and $\beta = 5$.

Fig. 4: Analysis of Example 1.

subsequence $\{\tilde{x}_{t_{n_r}}\}_{r=1}^{\infty}$ of $\{\tilde{x}_{s_n}\}_{n=1}^{\infty}$ such that $\{\tilde{x}_{t_{n_r}}\}_{r=1}^{\infty}$ and $\{\dot{\tilde{x}}_{t_{n_r}}\}_{r=1}^{\infty}$ converge uniformly to $\bar{x}$ and $\dot{\bar{x}}$, respectively. Note that $\tilde{x}_{s_n}$ is continuous for $n = 1, \dots, \infty$, due to Lemma 2. Consider the class of functions $\mathcal{X} = \{\tilde{x}_{s_n} \mid n = 1, \dots, \infty\}$. $\mathcal{X}$ is uniformly bounded (due to Assumption 4) and equicontinuous. Therefore, the Arzelà-Ascoli theorem can be invoked to show the existence of a uniformly convergent subsequence $\{\tilde{x}_{t_{n_k}}\}_{k=1}^{\infty}$. Let $\bar{x} : [0, T] \to \mathbb{R}^n$ be the limit of $\{\tilde{x}_{t_{n_k}}\}_{k=1}^{\infty}$. Now, consider the sequence $\{\dot{\tilde{x}}_{t_{n_k}}\}_{k=1}^{\infty}$. Notice that, due to the construction, $\{\dot{\tilde{x}}_{t_{n_k}}\}_{k=1}^{\infty}$ is continuous. Consider the class of functions $\bar{\mathcal{X}} = \{\dot{\tilde{x}}_{t_{n_k}} \mid k = 1, \dots, \infty\}$. Similar to the previous case, $\bar{\mathcal{X}}$ is uniformly bounded and equicontinuous. Therefore, another application of Arzelà - Ascoli theorem implies that $\{\dot{\tilde{x}}_{t_{n_k}}\}_{k=1}^{\infty}$ has a subsequence $\{\dot{\tilde{x}}_{t_{n_r}}\}_{r=1}^{\infty}$ that converges uniformly to a function $y : [0, T] \to \mathbb{R}^n$. Since $\{n_r\}_{r=1}^{\infty} \subseteq \{n_k\}_{k=1}^{\infty}$, we have that $\{\tilde{x}_{t_{n_r}}\}_{r=1}^{\infty}$ converges uniformly to $\bar{x}$. Therefore, due to Theorem 7.17 of [62], we have $\dot{\bar{x}} = y$. Finally, recall that $\{x_k^{s_n}\}_{n=1}^{\infty}$ is uniformly bounded and there exists a universal constant c such that $\mathcal{J}(x_k^{s_n}) \geq c$ for $k = 0, \dots, T/s_n$ and $n = 1, \dots, \infty$. This implies that the function sequence $\{\tilde{x}_{t_{n_r}}\}_{r=1}^{\infty}$ is also uniformly bounded and since they converge uniformly to $\bar{x}$, one can invoke Lemma 1 to verify the existence of a universal $c' > 0$ such that $c \geq c'$ and $\mathcal{J}(\bar{x}(t)) \geq c'$ for every $t \in [0, T]$. $\square$

References

[1] S. Fattahi, C. Josz, R. Mohammadi, J. Lavaei, and S. Sojoudi, "Absence of spurious local trajectories in time-varying optimization: A control-theoretic perspective," *Under review for Conference on Control Technology and Applications*, 2020.

[2] S. H. Low, "Convex relaxation of optimal power flow—part I: Formulations and equivalence," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 1, pp. 15–27, 2014.

[3] M. Gupta, L. Jin, and N. Homma, *Static and dynamic neural networks: from fundamentals to advanced theory*. John Wiley & Sons, 2004.

[4] L. Xu and M. Davenport, "Dynamic matrix recovery from incomplete observations under an exact low-rank constraint," in *Advances in Neural Information Processing Systems*, 2016, pp. 3585–3593.

[5] L. Xu and M. A. Davenport, "Simultaneous recovery of a series of low-rank matrices by locally weighted matrix smoothing," in *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2017, pp. 1–5.

[6] C. Zeng, Q. Wang, S. Mokhtari, and T. Li, "Online context-aware recommendation with time varying multi-armed bandit," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 2025–2034.

[7] F. Belkhouche, B. Belkhouche, and P. Rastgoufard, "Autonomous navigation and obstacle avoidance using navigation laws with time-varying deviation functions," *Advanced robotics*, vol. 21, no. 5-6, pp. 555–581, 2007.

[8] A. Simonetto, E. Dall'Anese, S. Paternain, G. Leus, and G. B. Giannakis, "Time-varying convex optimization: Time-structured algorithms and applications," *Proceedings of the IEEE*, 2020.

[9] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Advances in Neural Information Processing Systems*, 2016, pp. 3873–3881.

[10] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[11] R. Y. Zhang, S. Sojoudi, and J. Lavaei, "Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery," *Journal of Machine Learning Research*, vol. 20, pp. 1–34, 2019.

[12] S. Fattahi and S. Sojoudi, "Exact guarantees on the absence of spurious

local minima for non-negative robust principal component analysis," *arXiv preprint arXiv:1812.11466*, 2018.

[13] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 6, pp. 797–829, 2006.

[14] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[15] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[16] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Advances in Neural Information Processing Systems*, 2018, pp. 6389–6399.

[17] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1233–1242.

[18] C. Josz, Y. Ouyang, R. Zhang, J. Lavaei, and S. Sojoudi, "A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization," in *Advances in neural information processing systems*, 2018, pp. 2441–2449.

[19] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere i: Overview and the geometric picture," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 853–884, 2017.

[20] B. Kleinberg, Y. Li, and Y. Yuan, "An alternative view: When does SGD escape local minima?" in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 2698–2707.

[21] Y. Tang, K. Dvijotham, and S. Low, "Real-time optimal power flow," *IEEE Transactions on Smart Grid*, vol. 8, pp. 2963–2973, 2017.

[22] Y. Tang, E. Dall'Anese, A. Bernstein, and S. L. , "Running Primal-Dual Gradient Method for Time-Varying Nonconvex Problems," *https://arxiv.org/pdf/1812.00613.pdf*, 2019.

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

[24] A. Wibisono, A. C. Wilson, and M. I. Jordan, "A variational perspective on accelerated methods in optimization," *proceedings of the National Academy of Sciences*, vol. 113, no. 47, pp. E7351–E7358, 2016.

[25] J. Zhang, A. Mokhtari, S. Sra, and A. Jadbabaie, "Direct runge-kutta discretization achieves acceleration," in *Advances in Neural Information Processing Systems*, 2018, pp. 3900–3909.

[26] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *Advances in Neural Information Processing Systems*, 2018, pp. 6571–6583.

[27] D. Scieur, V. Roulet, F. Bach, and A. d'Aspremont, "Integration Methods and Optimization Algorithms," *NeurIPS*, 2017.

[28] P. Xu, J. Chen, D. Zou, and Q. Gu, "Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization," *NeurIPS*, 2018.

[29] B. Zhou, "On asymptotic stability of linear time-varying systems," *Journal Automatica, IFAC*, vol. 68, pp. 266–276, 2016.

[30] D. Aeyels and J. Peuteman, "A new asymptotic stability criterion for nonlinear time-variant differential equations," *IEEE Trans. Automat. Control*, vol. 43, pp. 968—-971, 1998.

[31] A. R. Teel, J. Peuteman, and D. Aeyels, "Semi-global practical asymptotic stability and averaging," *Systems & Control Letters*, vol. 37, pp. 329—-334, 1999.

[32] J. J. DaCunha, "Instability results for slowly time varying linear dynamic systems on time scales," *Journal of Mathematical Analysis and Applications*, vol. 328, pp. 1278–1289, 2007.

[33] W. A. Bukhsh, A. Grothey, K. McKinnon, and P. Trodden, "Local solutions of optimal power flow," *IEEE Transactions on Power Systems*, vol. 28, pp. 4780–4788, 2013.

[34] J. Lavaei and S. H. Low, "Zero duality gap in optimal power flow problem," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 92–107, 2012.

[35] W. I. Zangwill, "Non-linear programming via penalty functions," *Management science*, vol. 13, no. 5, pp. 344–358, 1967.

[36] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.

[37] C. B. Do, Q. V. Le, and C. S. Foo, "Proximal regularization for online and batch learning," *ICML*, 2009.

[38] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient Descent Only Converges to Minimizers," *COLT*, 2016.

[39] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, "Stochastic subgradient method converges on tame functions," *Foundations of computational mathematics*, vol. 20, no. 1, pp. 119–154, 2020.

[40] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.

[41] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[42] D. G. Luenberger, Y. Ye *et al.*, *Linear and nonlinear programming*. Springer, 1984, vol. 2.

[43] A. Waechter, C. Laird, F. Margot, and Y. Kawajir, "Introduction to ipopt: A tutorial for downloading, installing, and using ipopt," in *Revision*, 2009.

[44] E. A. Coddington and N. Levinson, *Theory of ordinary differential equations*. Tata McGraw-Hill Education, 1955.

[45] J. C. Butcher, *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2016.

[46] H. Yamashita, "A differential equation approach to nonlinear programming," *Mathematical Programming*, vol. 18, no. 1, pp. 155–168, 1980.

[47] Y. G. Evtushenko and V. G. Zhadan, "Stable barrier-projection and barrier-newton methods in nonlinear programming," *Optimization Methods and Software*, vol. 3, no. 1-3, pp. 237–256, 1994.

[48] F. Alvarez, J. Bolte, and O. Brahic, "Hessian riemannian gradient flows in convex programming," *SIAM journal on control and optimization*, vol. 43, no. 2, pp. 477–501, 2004.

[49] K. Tanabe, "An algorithm for constrained maximization in nonlinear programming," *J. Oper. Res. Soc. Jpn*, vol. 17, pp. 184–201, 1974.

[50] R. Nickalls, "Viète, descartes and the cubic equation," *The Mathematical Gazette*, vol. 90, p. 203–208, 2006.

[51] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *ICRL*, 2017.

[52] H. K. Khalil and J. W. Grizzle, *Nonlinear systems*. Prentice hall Upper Saddle River, NJ, 2002, vol. 3.

[53] I. Panageas and G. Piliouras, "Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions," *ITCS*, 2017.

[54] D. Davis and D. Drusvyatskiy, "Active strict saddles in nonsmooth optimization," *https://arxiv.org/pdf/1912.07146.pdf*, 2019.

[55] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to Escape Saddle Points Efficiently," *ICML*, 2017.

[56] C. Criscitiello and N. Boumal, "Efficiently escaping saddle points on manifolds," *NeurIPS*, 2019.

[57] C. Chicone, *Ordinary differential equations with applications*. Springer Science & Business Media, 2006, vol. 34.

[58] W. Lohmiller and J.-J. E. Slotine, "On contraction analysis for non-linear systems," *Automatica*, vol. 34, no. 6, pp. 683–696, 1998.

[59] W. Lohmiller and J.-J. E. Slotine, "Nonlinear process control using contraction theory," *AIChE journal*, vol. 46, no. 3, pp. 588–596, 2000.

[60] W. Lu and M. Di Bernardo, "Contraction and incremental stability of switched carathéodory systems using multiple norms," *Automatica*, vol. 70, pp. 1–8, 2016.

[61] W. Rudin, *Real and complex analysis*. Tata McGraw-hill education, 2006.

[62] W. Rudin *et al.*, *Principles of mathematical analysis*. McGraw-hill New York, 1964, vol. 3.

[63] D. G. Luenberger, "The gradient projection method along geodesics," *Management Science*, vol. 18, no. 11, pp. 620–631, 1972.

[64] J. Mulvaney-Kemp, S. Fattahi, and J. Lavaei, "Smoothing property of load variation promotes finding global solutions of time-varying optimal power flow," *IEEE Transactions on Control of Network Systems*, 2021.

Salar Fattahi is an assistant professor at University of Michigan. He has received the INFORMS Data Mining Best Paper Award and been a best paper finalist for ACC 2018.

Cédric Josz is an assistant professor at Columbia University. He received the 2016 Best Paper Award in Springer Optimization Letters and was a finalist of the competition for best PhD thesis of 2017 organized by French Agency for Mathematics in Interaction with Industry and Society.

Yuhao Ding is a Ph.D. candidate at UC Berkeley. His research interests are nonlinear optimization and machine learning.

Reza Mohammadi is a postdoctoral scholar at UC Berkeley. His research interests are statistical learning, nonlinear optimization and energy.

Javad Lavaei is an associate professor at UC Berkeley. He is an associate editor for IEEE Transactions on Automatic Control, IEEE Transactions on Smart Grid and IEEE Control Systems Letters.

Somayeh Sojoudi is an assistant professor at UC Berkeley. She is an associate editor for IEEE Transactions on Smart Grid, IEEE Access, and Systems & Control Letters.