Absence of Spurious Local Trajectories in Time-Varying Optimization: A Control-Theoretic Perspective

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

***Abstract*—In this paper, we study the landscape of an optimization problem whose input data vary over time. This time-varying problem consists of infinitely-many individual optimization problems, whose solution is a trajectory over time rather than a single point. To understand when it is possible to find a global solution of a time-varying non-convex optimization 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) which, at limit, characterizes the spurious local solutions of the time-varying optimization problem. By building upon this connection, we prove that the absence of spurious local trajectory is closely related to the transient behavior of the proposed ODE. In particular, we show that: (1) if the problem is time-invariant, the spurious local trajectories are ubiquitous since any strict local minimum is a locally stable equilibrium point of the ODE, and (2) if the ODE is time-varying, the data variation may force all ODE trajectories initialized at arbitrary local minima at the initial time to gradually converge to the global solution trajectory. This implies that the natural data variation in the problem may automatically trigger escaping 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. [1]. Other examples include the training of dynamic neural networks [2], dynamic matrix recovery [3], [4], and time-varying multi-armed bandit problem [5]. Indeed, most of these problems are large-scale and should be solved in real-time, which strongly motivates the need for fast 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 [6]–[9]. 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 fast 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 varies over time. First, we introduce the notion of *spurious local trajectory* as a generalization to the point-wise spurious local solutions. 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 trajectory. 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 a class of time-varying univariate optimization problems and present sufficient conditions under which, despite having point-wise spurious local minima at all times, the time-varying problem is free of spurious local trajectory. Finally, by studying the stability of the proposed ODE on feasible manifolds, we prove that every strict local minimum of the time-invariant optimization problem is locally stable on its feasible region. This implies that the time-varying nature of the problem is essential for the absence of spurious local trajectories.

A. Related Works

Nonconvexity is inherent to many problems in machine learning; from the classical compressive sensing and matrix completion/sensing [10]–[12] to the more recent problems on the training of deep neural networks [13], they often

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

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 [6], [7], [14], robust principle component analysis [9], [15], and dictionary recovery [16] have benign landscape, implying that they are free of spurious local minima. It has also been proven recently in [17] 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 [18], [19]. Therefore, it is natural to study the landscape of such time-varying nonconvex optimization problems, taking into account their dynamic nature.

A common framework in machine learning for analyzing sequential decision-making problems is online (convex or non-convex) optimization (see [20] and [21] for a comprehensive survey). In general, the main goal in such problems is to propose a sequential algorithm and measure its performance through the notion of *global regret*, which is defined as the incurred sub-optimality error of the proposed algorithm compared to the optimal fixed algorithm in the hindsight [20], [22]. It is well known that in the nonconvex settings, such notion of global regret is intractable to minimize. Therefore, different works have resorted to the minimization of a surrogate notion of regret, which is called local regret [19], [23], [24]. The local regret measures the sub-optimality compared to a local solution. Similar notions are widely used in learning and signal processing problems. Evidently, most of the existing results on nonconvex online optimization are algorithm-dependent and cannot be used to make general statements on the global landscape of the problem. In particular, they often measure the performance of a specific algorithm in tracking a nearby local solution.

Recently, there has been a growing interest in analyzing the performance of numerical algorithms from a control theoretical perspective [25]–[30]. Roughly speaking, the general idea behind these approaches is to analyze the convergence of a specific algorithm by first modeling its limiting behavior as an autonomous (time-invariant) ODE that describes the evolution of a dynamical system, and then studying its stability properties. As a natural extension, one would generalize this approach to online optimization by modeling its limiting behavior as a non-autonomous ODE corresponding to a time-varying dynamical system. However, the stability analysis of time-varying dynamical systems is highly convoluted, even in the linear case.

II. Motivation: Case Study on Power Systems

In this section, we present an empirical study on the dynamic landscape of the optimal power flow problem to illustrate the notion of spurious trajectory and the role of data variation in online optimization. The objective of this problem 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 [31] and initialize the system from the global minimum, as well as three different spurious local minima at time $t = 0$. 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 [32]. 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.

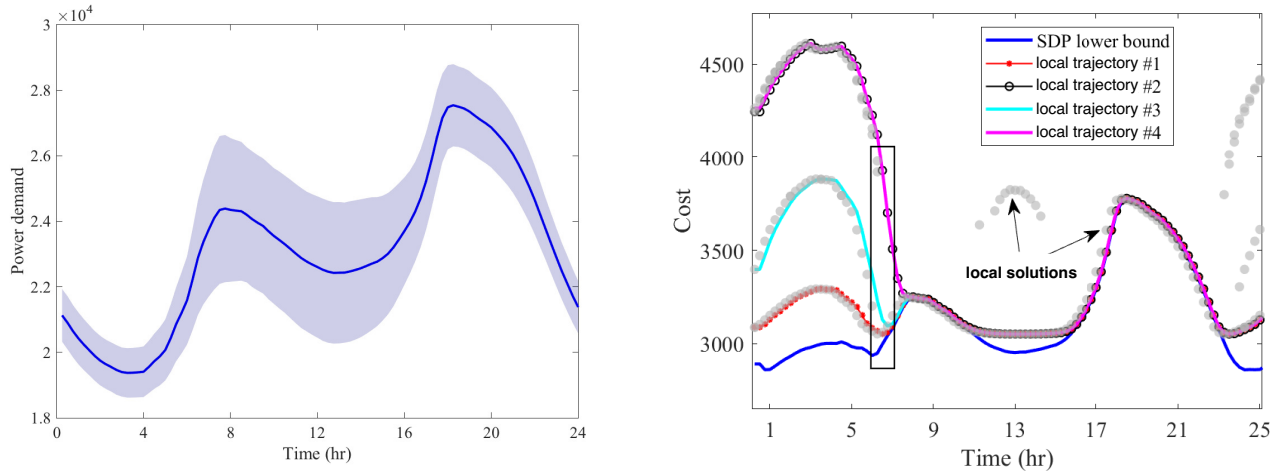
The gray circles in these plots are some of the local solutions that were obtained at different times via a Monte Carlo simulation. Based on Figure 1b, indeed there exist multiple local solutions at almost all time-step (some of them emerge over time). However, surprisingly, the trajectories of the local 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.

III. Notion of Spurious Local Trajectory

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

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

where the objective function $f(x,t)$ and the right-hand side of the equality constraints vary over time $t \in [0,T]$. We assume that $f : \mathbb{R}^n \times [0,T] \longrightarrow \mathbb{R}$, $h_i : \mathbb{R}^n \longrightarrow \mathbb{R}$, $d_i : [0,T] \longrightarrow \mathbb{R}$ for $i = 1,\dots,m$ 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., $f(x,t) \geq M$ for some constant M) and that the problem is feasible for all $t \in [0,T]$.



(a) California average load profile for January 2019.

(b) The solution trajectories of the time-varying optimal power flow.

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

Remark 1. *Inequality constraints can also be included in* (1) *through a reformulation technique. In particular, suppose that* (1) *includes a set of inequality constraints* $g_j(x) \le v_j(t)$ *for* $j = 1, \dots, l$. *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*
$$g_j(x) + s_j = v_j(t), \qquad j = 1, \dots, l$$
2. *Augment the objective function with a penalty* $p(s) = \sum_{j=1}^{l} p_j(s_j)$.

Here, $p_j(s_j)$ *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 [33], [34]. This implies that the previously-introduced optimal power flow problem can be reformulated as* (1).

In practice, one can only hope to sequentially solve the optimization problem (1) at discrete times $0 = t_0 < t_1 < t_2 < \dots < t_N = T$. However, notice that (1) is unregularized. In particular, depending on the properties of the objective function, an arbitrary solution to (1) at time t_k can be arbitrarily far from that of (1) 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 [35].

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

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

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

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 (2). A natural approach to characterizing the global landscape of (1) is to analyze discrete local trajectories of the regularized problem (2). However, notice that the non-convexity of (2) may lead to *bifurcations* in discrete local trajectories. In particular, given a local solution x_k, the regularized problem (2) may possess two local solutions $x_{k+1}^{(1)}$ and $x_{k+1}^{(2)}$, each resulting in a different discrete local trajectory.[1] In what follows, we show that such bifurcations disappear in the ideal scenario, where the regularized problem can be solved 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:

$$\dot{x} = -\frac{1}{\alpha}\eta(x, t) + \theta(x)\dot{d} \tag{3a}$$
$$x(0) = x_0 \tag{3b}$$

[1]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 $t_k \le T/2$ and then, due to the regularization, it bifurcates into two separate discrete trajectories.

where

$$\eta(x,t) := \left[I - \mathcal{J}_h(x)^\top (\mathcal{J}_h(x)\mathcal{J}_h(x)^\top)^{-1}\mathcal{J}_h(x)\right] \times \nabla_x f(x,t), \quad (4a)$$

$$\theta(x) := \mathcal{J}_h(x)^\top (\mathcal{J}_h(x)\mathcal{J}_h(x)^\top)^{-1}. \quad (4b)$$

Above, $\mathcal{J}_h(x)$ 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 $\mathcal{J}_h(x)^\top$.

Later, we will show that the solution to (3) exists, it is unique, and can be used to fully characterize the limiting behavior of every discrete local trajectory of the time-varying problem (1).

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* (2) *at x_k are upper bounded by R_1 and R_2, respectively, for every $k \in \{0, 1, 2, \ldots, N\}$.*

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.*

Note that Assumption 2 implies that linear independence constraint qualification holds at every point of a discrete local trajectory.

Theorem 1 (Existence and Uniqueness)**.** *If x_0 is a local solution to the time-varying optimization* (1) *at $t = 0$, then* (3) *has a unique continuously differentiable solution $x : [0, T] \to \mathbb{R}^n$.*

Theorem 2 (Uniform Convergence)**.** *If x_0 is a local solution to the time-varying optimization* (1) *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*

$$\lim_{N \to +\infty} \sup_{0 \leq k \leq N} \|x_k - x(t_k)\| = 0, \quad (5)$$

where N is the number of points in the discrete local trajectories that are evenly spaced-out in time.

Sketch of the proofs. The proofs of Theorems 1 and 2 are quite involved and hence, they are deferred to the technical report [36]. 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 [37, Theorem 3.1], the Cauchy-Peano Theorem [37, Theorem 1.2], and the Carathéodory Theorem [37, 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 (3). Instead, we take a different approach to prove existence and uniqueness of the solution to (3) (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 (2), we prove that y is a solution to (3) 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 [37, Theorem 2.2], completes the proof of Theorem 1.

Given the existence and uniqueness of the solution to (3), we show the correctness of Theorem 2 by making an extensive use of the so-called backward Euler method [38]. In particular, we show that *all* discrete local trajectories converge to a discretized version of the solution to (3) 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. The details are available in [36]. □

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 $x : [0, T] \longrightarrow \mathbb{R}^n$ is said to be a* ***continuous local trajectory*** *of the time-varying optimization* (1) *if the following holds:*

1) *$x(0)$ is a local solution to the time-varying optimization* (1) *at time $t = 0$;*
2) *x is a solution to* (3)*.*

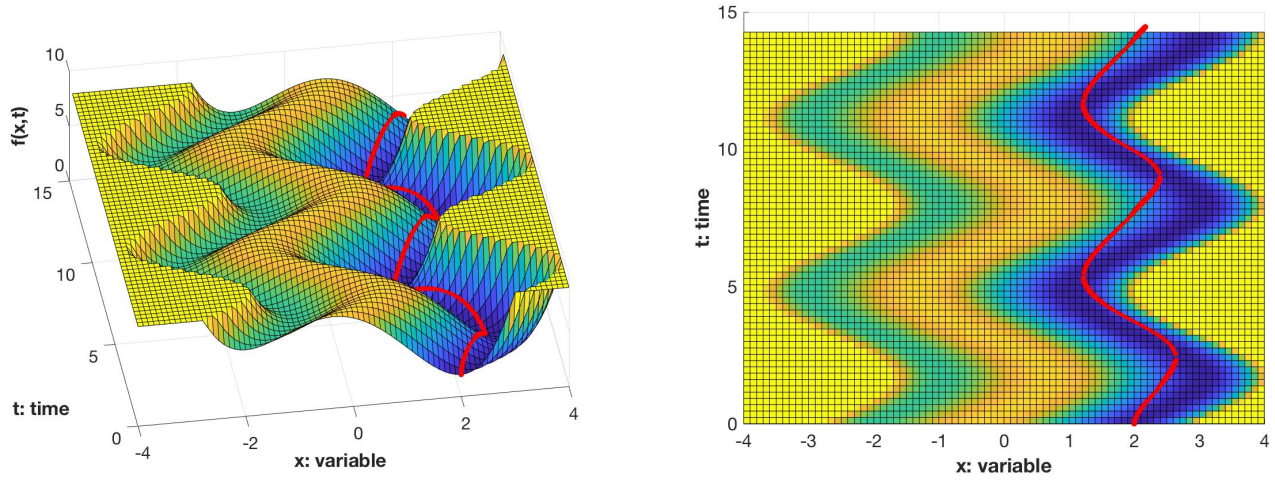
We next introduce the central notion in this paper.

Definition 3. *A continuous local trajectory $x : [0, T] \longrightarrow \mathbb{R}^n$ is said to be a* ***spurious local trajectory*** *if its final state $x(T)$ does not belong to the region of attraction of a global solution to the time-varying optimization* (1) *at time $t = T$. In other words, the trajectory is non-spurious if the initial value problem*

$$\begin{cases} \dot{\bar{x}} &= -\frac{1}{\alpha}\eta(\bar{x}, T), \\ \bar{x}(0) &= x(T). \end{cases} \quad (6)$$

admits a continuously differentiable solution $\bar{x} : \mathbb{R}_+ \longrightarrow \mathbb{R}^n$ such that $\bar{x}(t)$ converges towards a global solution as $t \longrightarrow +\infty$.

One may speculate that the following alternative definition of spurious local trajectory is more natural to the time-varying problem at hand: a continuous local trajectory is non-spurious if the final state is a global solution, or near a global solution upon a small perturbation of the initial condition. However, note that both discrete and continuous local trajectories are defined with respect to the regularized problem (2), as opposed to (1). Indeed, 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



(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. 2: Example of a time-varying optimization.

initialized at the global minimum. See Figures 2a and 2b for an illustration of this phenomenon.

It is worthwhile to note that our definition of spurious local trajectory reveals a novel interplay between time-varying optimization and the theory of switched systems [39]–[41]. Indeed, the question of whether a continuous local trajectory is spurious can be formulated using the following switched system:

$$\dot{x} = -\frac{1}{\alpha}\eta(x,\sigma(t)) \;+\; \theta(x)\;\dot{d}\;\sigma'(t) \tag{7a}$$

$$\sigma(t) := \begin{cases} t & \text{if } \; 0 \leqslant t \leqslant T \\ T & \text{if } \; t > T \end{cases} \tag{7b}$$

where σ is referred to as switching signal in the literature. The fact that its derivative is not defined at $t = T$ poses no problem. Indeed, we are interested in finding continuous solutions in the extended sense

$$\begin{aligned} x(t) =& x(0) \\ &+ \int_0^t \Big[-1/\alpha\;\eta(x(\tau),\sigma(\tau)) \;+\; \theta(x(\tau)\;\dot{d}\;\sigma'(\tau)\Big]\,d\tau. \end{aligned} \tag{8}$$

Therefore, by building upon the contraction analysis of nonlinear systems [42]–[44], the time-varying problem (1) is devoid of spurious local trajectories if all of its local solutions at $t = 0$ belong to a *contraction region* of (7a) that includes a trajectory $\bar{x} : \mathbb{R}_+ \longrightarrow \mathbb{R}^n$ converging towards the global minimum of (1) at time $t = T$. Empirical identification of this region for (7a) is considered as an enticing challenge for future work.

Remark 2. *Combined with the rich literature on the online algorithms such as online gradient and mirror descent [45]–[47], our results imply that if the problem is free of spurious local trajectories, then any KKT-seeking algorithm for the regularized problem* (2) *converges to the basin of attraction of the globally optimal solution at the termination time $t = T$, provided that N is sufficiently large. This is due to the fact that, according to Theorem 2 and Definition 3, the KKT trajectories of* (2) *are well-approximated with the unique solution to the proposed ODE which corresponds to a non-spurious trajectory.*

IV. Stability Analysis of Local Trajectories

In this section, we show that the time-varying nature of (1) is crucial for the absence spurious local trajectories. In particular, we illustrate an intriguing connection between the landscape of the time-varying optimization and the stability of the (3). We show that by starting from an initial spurious local solution to (1) at time $t = 0$, the solution to (3) may be able to escape the basin of attraction of this local minimum over time and converge to the global one. This indeed highlights the premise of our work: an online optimization problem can be devoid of spurious local trajectories, despite possessing point-wise spurious local minimum at all times. In what follows, we formalize this observation by showing that the time-varying natures of (1) and its regularized surrogate (2) are essential for the instability of (3) around the spurious local minima.

We begin by assuming that the time-varying optimization does not change over the time interval $[0, T]$. Then, we may simplify the notations and omit t from (1), as in:

$$\inf_{x\in\mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad h(x) = d \tag{9}$$

where $h(x) = [h_1(x), \ldots, h_m(x)]^T$ and $d = [d_1, \ldots, d_m]^T$. Likewise, we may drop t from the dynamics:

$$\dot{x} = -\frac{1}{\alpha}\left[I - \mathcal{J}_h(x)^\top(\mathcal{J}_h(x)\mathcal{J}_h(x)^\top)^{-1}\mathcal{J}_h(x)\right]\nabla f(x). \tag{10}$$

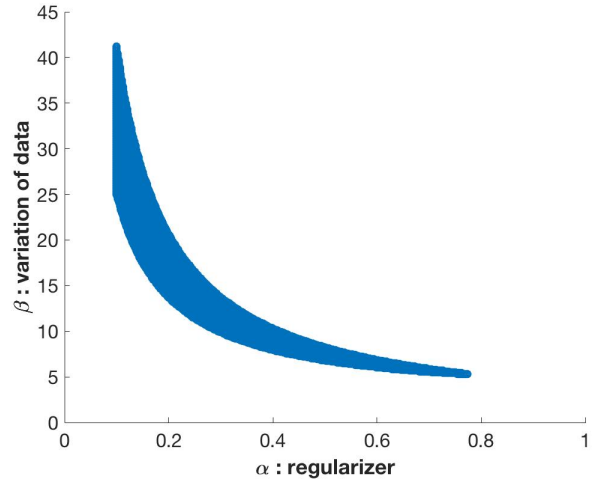
In this case, we show that all continuous local trajectories initialized at strict local minima of (9) are spurious trajectories. This is a direct implication of the following proposition.

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

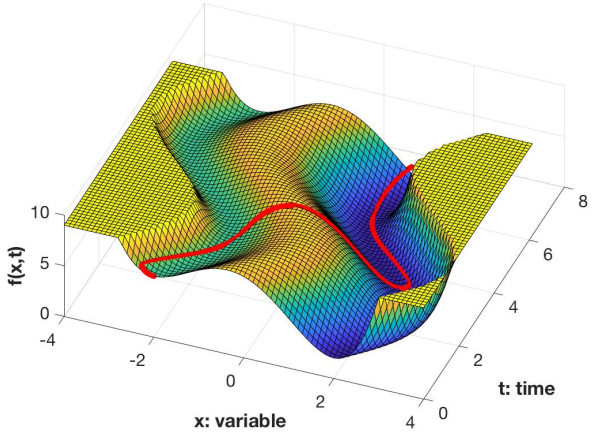
2. $-27\alpha^2\beta^2 + 6885/128\alpha\beta + 61009/256 \geqslant 0$

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

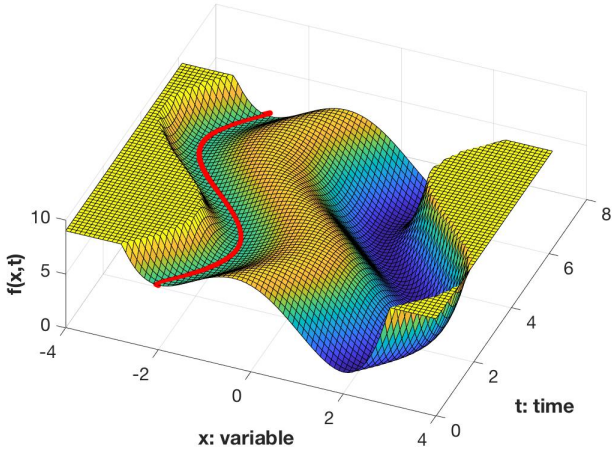
(a) Inequalities in function of α, β guaranteeing absence of spurious trajectories.



(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. 3: Illustration of Proposition 2.

Proposition 1 (Local stability)**.** *The set* $\{x : h(x) = d\}$ *is an invariant manifold for the system* (10)*. Moreover, any strict local minimum* x^* *of the time-invariant optimization* (9) *is locally stable for* (10) *on this manifold in the sense that*

$$\forall \epsilon > 0,\ \exists \delta > 0 : (\|x(0) - x^*\| \leqslant \delta \text{ and } h(x(0)) = d) \implies \forall t \in [0, T],\ \ \|x(t) - x^*\| \leqslant \epsilon \quad (11)$$

where $x : [0, T] \longrightarrow \mathbb{R}^n$ *satisfies the ordinary differential equation* (10)*.*

Proof. The proof is provided in [36]. □

Proposition 1 provides a negative result on the impossibility of escaping spurious local minima in the time-invariant case. However, this proposition does not hold in the time-varying case. As a preliminary step for further study, we show that the strict local minima of (1) may neither be equilibrium nor stable if it is time-varying. In particular, we focus on a class of uni-dimensional time-varying problems in the following form:

$$\inf_{x\in\mathbb{R}} f(x, t) := g(x - \beta\sin(t)) \quad (12)$$

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 2. *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 (12) *has no spurious local trajectories.*

Proof. A continuous local trajectory $x : [0, 2\pi] \longrightarrow \mathbb{R}$ satisfies

$$x(0) \leqslant y_3, \qquad \dot{x} = -\frac{1}{\alpha}\nabla_x f(x, t), \tag{13}$$

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

$$y(0) \leqslant y_3, \qquad \dot{y} = -\frac{1}{\alpha}g'(y) - \beta\cos(t). \tag{14}$$

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

$$\dot{y} = -\frac{1}{\alpha}g'(y) - \beta\cos(t) \geqslant -\frac{C}{\alpha} - \beta\cos(t). \tag{15}$$

Thus, we have

$$y(t_2) \geqslant -\frac{C}{\alpha}(t_2 - t_1) - \beta(\sin(t_2) - \sin(t_1)) + y(t_1). \tag{16}$$

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 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 (16) 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.

Notice that $f(x, T) = g(x)$. We thus consider the initial value problem

$$\dot{\bar{x}} \quad = -\frac{1}{\alpha}g'(\bar{x}) \tag{17a}$$

$$\bar{x}(0) = \quad x(2\pi) \tag{17b}$$

Since g' is continuously differentiable, it is Lipschitz on any interval $[a, b]$ of $\mathbb{R}$. The existence of a local continuously differential solution is then guaranteed by the Picard-Lindelöf Theorem [37, Theorem 3.1]. Consider a maximal solution, that is to say $\bar{x} : [0, \bar{t}) \longrightarrow \mathbb{R}$ where $\bar{t} \in \mathbb{R}$ or $\bar{t} = +\infty$. We next show by contradiction that the latter holds. Without loss of generality, assume that $x(2\pi) < y_3$. We know that $g'(x) < 0$ for all $x(2\pi) \leqslant x < y_3$. As a result, $\bar{x}$ is an increasing function on $[0, \bar{t})$. It is also upper bounded by y_3. Indeed, it is upper bounded by any $y_3 + \epsilon$ for $\epsilon > 0$ small enough, so that $g'(y_3 + \epsilon) > 0$ (and then using the same argument from one of the above paragraphs for the third time). As a result, $\bar{x}$ has limit $\bar{x}(\bar{t})$ as t converges towards $\bar{t}$ from below. Since g' is continuous, $\bar{x} : [0, \bar{t}] \longrightarrow \mathbb{R}$ is a solution to initial value problem, which is a contradiction. As a result, $\bar{t} = +\infty$.

We next show that $\bar{x}(\bar{t}) = y_3$. Since $\bar{x}'(t) = -1/\alpha g'(\bar{x}(t))$ for all $t \geqslant 0$, the derivative $\bar{x}'$ has limit equal to $\bar{x}'(\bar{t}) = -1/\alpha g'(\bar{x}(\bar{t}))$. Since $\bar{x}'(t) \geqslant 0$ for all $t \geqslant 0$, it holds that $\bar{x}'(\bar{t}) \geqslant 0$. Assume that $\bar{x}'(\bar{t}) > 0$. Then there exists $T_0 \geqslant 0$ such that, for all $t \geqslant T_0$, we have $\bar{x}'(t) \geqslant \bar{x}'(\bar{t})/2$. Then $\bar{x}(t) \geqslant \bar{x}(T_0) + \bar{x}'(\bar{t})(t - T_0)/2$ diverges, which is a contradiction. Thus $\bar{x}'(\bar{t}) = -1/\alpha g'(\bar{x}(\bar{t})) = 0$, which implies that $\bar{x}(\bar{t})$ is equal to y_1, y_2 or y_3. Since $\bar{x}(\bar{t}) \geqslant \bar{x}(0) = x(2\pi) \geqslant m_2 > y_2 > y_1$, we conclude that $\bar{x}(\bar{t}) = y_3$. □

We highlight the implications of the above proposition through a numerical example. Consider the objective function $f(x, t) := g(x - \beta\sin(t))$, where

$$g(y) := 1/4y^4 + 1/8y^3 - 2y^2 - 3/2y + 8. \tag{18}$$

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 2 can be brought to bear on this example. They yield three inequalities, as shown in Figure 3a, whose feasible region is represented in Figure 3b. Taking a point in that feasible region, we confirm numerically in Figure 3c 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 3d 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 $f(\cdot, T)$.[2]

We make a few remarks regarding Figure 3a. 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 [48]. Furthermore, the second inequality corresponds to minus the discriminant of a fourth-order polynomial.

V. 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 a case study on power

[2]In order to increase visibility, a maximal threshold is used on the objective function $f(x, t)$ in Figure 3c and Figure 3d (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 3c.

systems and theoretical results, we show that a time-varying optimization can 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. In particular, we prove that any spurious strict local minimum in time-invariant optimization problem is a locally stable equilibrium of its corresponding ODE, thereby giving rise to a spurious local trajectory. However, such undesirable property may disappear for time-varying optimization due to the role of the data variation in the behavior of the underlying ODE.

REFERENCES

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

[20] E. Hazan, "Introduction to Online Convex Optimization," *Foundations and Trends in Optimization*, 2016.

[21] S. Bubeck, "Introduction to online optimization," *Lecture Notes*, pp. 1–86, 2011.

[22] S. Shahrampour and A. Jadbabaie, "Distributed Online Optimization in Dynamic Environments Using Mirror Descent," *IEEE Transactions on Automatic Control*, vol. 63, pp. 714 – 725, 2018.

[23] E. Hazan, K. Singh, and C. Zhang, "Efficient Regret Minimization in Non-Convex Games," *ICML*, 2017.

[24] L. Yang, L. Deng, M. H. Hajiesmaili, C. Tan, and W. S. Wong, "An Optimal Algorithm for Online Non-Convex Learning," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 25, 2018.

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

[26] 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.

[27] 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.

[28] 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.

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

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

[31] 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.

[32] 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.

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

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

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

[36] S. Fattahi, C. Josz, R. Mohammadi, J. Lavaei, and S. Sojoudi, "On the absence of spurious local trajectories in online nonconvex optimization," *Technical Report*, 2020, https://lavaei.ieor.berkeley.edu/Time_Varing_2019_1.pdf.

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

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

[39] R. Shorten, F. Wirth, K. W. O. Mason, and C. King, "Stability criteria for switched and hybrid systems," *SIAM Rev.*, vol. 49, pp. 545–592, 2007.

[40] H. Lin and P. J. Antsaklis, "Hybrid dynamical systems: Stability and stabilization," *Foundations and Trends in Systems and Control*, 2014.

[41] D. Liberzon, "Switching in systems and control. systems & control: Foundations & applications," *Birkhäuser Boston*, 2003.

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

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

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

[45] E. Hazan, A. Rakhlin, and P. L. Bartlett, "Adaptive online gradient descent," in *Advances in Neural Information Processing Systems*, 2008, pp. 65–72.

[46] Z. Allen-Zhu, "Natasha 2: Faster non-convex optimization than SGD," in *Advances in neural information processing systems*, 2018, pp. 2675–2686.

[47] N. Srebro, K. Sridharan, and A. Tewari, "On the universality of online mirror descent," in *Advances in neural information processing systems*, 2011, pp. 2645–2653.

[48] R. Nickalls, "Viète, descartes and the cubic equation," *The Mathematical Gazette*, vol. 90, p. 203208, 2006.