Escaping Spurious Local Minimum Trajectories In Online Time-varying Nonconvex Optimization

Yuhao Ding[1]. Javad Lavaei[1] and Murat Arcak[2]

Abstract— **This paper is concerned with solving online nonconvex optimization problems using simple gradient-based algorithms with an arbitrary initialization. The main objective is to understand how the natural data variation of an online optimization problem affects finding its time-varying global minima. To this end, we investigate the properties of a time-varying gradient flow system with inertia, which can be regarded as the continuous-time limit of the online tracking scheme obtained by working through the optimality conditions for a discretized sequential optimization problem with a proximal regularization. We introduce the notion of the dominant trajectory and show that the inherent temporal variation of the problem could re-shape the landscape and help a proximal algorithm escape the spurious local minimum trajectories if the global minimum trajectory is dominant. By studying the three notions of jumping, tracking and escaping for nonlinear dynamical systems, sufficient conditions are derived to guarantee that no matter how the local search method is initialized, it will find and track a time-varying global solution after some time.**

I. INTRODUCTION

In this paper, we study the following unconstrained online time-varying optimization problem:

$$\min_{x(t)\in\mathbb{R}^n} \quad f(x(t),t) \tag{1}$$

where $t \geq 0$ denotes the time and $x(t)$ is the optimization variable that depends on t. For each time t, the function $f(x(t),t)$ could potentially be nonconvex in $x(t)$ with many local minima. The objective is to solve the above problem in an online fashion under the assumption that at any given time τ the function $f(x,t)$ is known for all $t \leq \tau$ while no knowledge about $f(x,\tau)$ may be available for any $t > \tau$. Therefore, the functions $f(x,\cdot)$'s cannot be minimized off-line and should be solved sequentially. Another issue is that the optimization problem at each time instance could be highly complex due to NP-hardness, which is an impediment to finding its global minima. This paper aims to investigate under what conditions simple local search algorithms can solve

the above online optimization problem to almost global optimality after some finite time.

If $f(x,t)$ does not change over time, the problem reduces to a classic (time-invariant) nonconvex optimization problem. It is known that simple local search methods, such as stochastic gradient descent (SGD) [11], may be able to find a global minimum of such time-invariant problem (under certain conditions) for almost all initializations due to the randomness embedded in SGD [13], [8], [16]. The objective of this paper is to significantly extend the above result from a single optimization problem to infinitely-many problems parametrized by time t. In other words, it is desirable to investigate the following question: **Can the temporal variation in the landscape of time-varying nonconvex optimization problems enable online local search methods to find and track global trajectories?** To answer this question, we study a first-order time-varying ordinary differential equation:

$$\dot{x} = -\frac{1}{\alpha}\nabla_x f(x,t), \quad x(0) = x_0 \tag{ODE}$$

where $\alpha > 0$ is a constant parameter named **inertia** due to a **proximal regularization**. A system of the form (ODE) is called a **time-varying gradient system with inertia α**.

In this work we prove that the natural temporal variation of the time-varying optimization problem encourages the exploration of the state space and re-shaping the landscape of the objective function by making it one-point strongly convex over a large region during some time interval. We introduce the notion of the dominant trajectory and show that if a given spurious local minimum trajectory is dominated by the global minimum trajectory, then the temporal variation of the time-varying optimization would trigger escaping the spurious local minimum trajectory for free. We develop the sufficient conditions under which the ODE solution will jump from a certain local minimum trajectory to a more desirable local minimum trajectory. We then derive sufficient conditions on the inertia α to guarantee that the solution of (ODE) can track a global minimum trajectory. This work generalizes the notion of spurious solutions from static optimization to dynamic optimization, and also its framework can be used to study when stochastic gradient descent is able to escape undesirable local minima.

A. Related work

Online time-varying optimization problems: There are many papers on designing efficient online algorithms for tracking optimizers of time-varying convex problems [21], [7], [2], [20]. With respect to time-varying nonconvex problems, [25], [24], [18] develop algorithms to track the local optimal solution of the time-varying optimization problems. The recent paper [5] poses the question of whether the natural temporal variation in a time-varying nonconvex optimization problem could help a local tracking method escape spurious local minimum trajectories, but it lacks mathematical conditions to guarantee this desirable behavior. The paper [19] also studies this phenomenon in a discrete setting in the context of power systems and verifies on real data for California that the natural load variation enables escaping local minima of the optimal power flow problem. The current work significantly generalizes the results of [5] and [19] by mathematically studying when such an escaping is possible.

Local search methods for global optimization: It has been recently shown that simple local search methods, such as gradient-based algorithms, have a superb performance in solving nonconvex optimization problems. For example, [13], [8] prove that a perturbed gradient descent and SGD could escape the saddle points efficiently. Furthermore, it has been shown that certain nonconvex optimization problems [3], [9], [27], [6], [14], [23] have benign landscape, implying that they are free of spurious local minima. The work [16] proves that SGD could help escape sharp local minima of a loss function. However, these results are all for time-invariant optimization problems. In contrast, many real-world problems should be solved sequentially over time with time-varying data. Therefore, it is essential to study the effect of the temporal variation on the landscape of time-varying nonconvex problems.

Continuous-time interpretation of discrete numerical algorithms: Many iterative numerical optimization algorithms for time-invariant optimization problems can be interpreted as a discretization of a continuous-time process. Then, several new insights have been obtained due to the known results for continuous-time dynamical systems [15], [10]. The recent papers [22], [17], [26] study accelerated gradient methods for convex optimization problems from a continuous-time perspective. It is natural to analyze the continuous-time limit of an online algorithm for tracking a KKT trajectory of time-varying optimization problem [21], [24], [18], [5].

B. Notations

The notation $\|\cdot\|$ shows the Euclidian norm. The interior of the interval I is denoted by $\text{int}(I)$. The symbol $\mathcal{B}_r(h(t)) = \{x \in \mathbb{R}^n : \|x - h(t)\| \leq r\}$ denotes the region centered around a trajectory $h(t)$ with radius r at time t. We denote the solution of $\dot{x} = f(x, t)$ starting from x_0 at the initial time t_0 with $x(t, t_0, x_0)$ or the short-hand notation $x(t)$ if the initial condition (t_0, x_0) is clear from the context.

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 role of data variation in online optimization. Consider the time-varying optimal power flow (OPF) problem, as the most fundamental problem for the operation of electric power grids that aims to match supply with demand while satisfying network and physical constraints. Let $f(x, t)$ be the function to be minimized at time t, which is the sum of the total energy cost and a penalty term taking care of all the constraints of the problem. Assume that the load data corresponds to the California data for August 2019. As discussed in [19], the function $f(x, t)$ has 16 local minima at t=0 and many more for some values of $t > 0$. However, if (ODE) is run from any of these local minima, the 16 trajectories will all converge to the globally optimal trajectory, as shown in Figure 1. This implies that 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. This observation has been made in [19] for a discrete-time version of the problem, but it also holds true for the continuous-time (ODE) model.
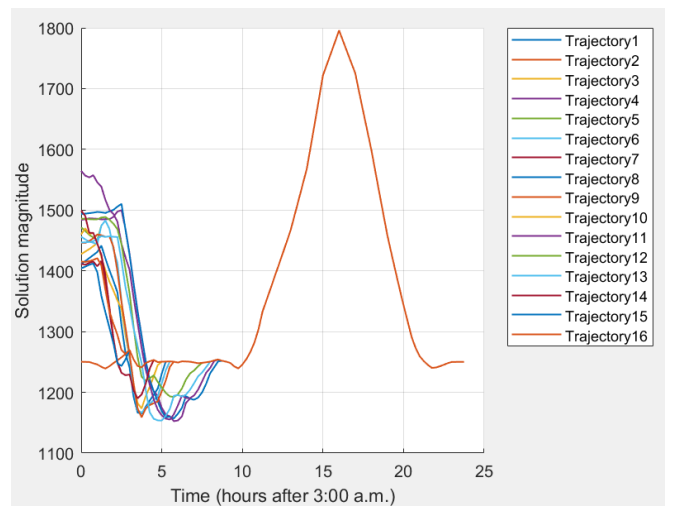

Fig. 1. $|x(t)|$ (magnitude of the solution of (ODE)).

III. Preliminaries and Problem Formulation

A. Time-varying optimization

We assume that $f : \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$ is twice continuously differentiable in x and continuously differentiable in $t \geq 0$. Moreover, suppose that f is uniformly bounded from below. The first-order stationary condition for (1) is as follows:

$$0 = \nabla_x f(x(t), t) \qquad (2)$$

In this work, we assume that the real roots of (2) are all isolated at each time t (i.e., stationary trajectories do not intersect in time). An isolated stationary trajectory can theoretically be a mix of local minima, local maxima and saddle points of the function $f(x, t)$ at different times. However, the goal of this work is to study only isolated local minimum trajectories of the time-varying optimization (1).

Definition 1. *A continuous trajectory $h(t) : [0, \infty) \to \mathbb{R}^n$ is said to be a **local (or global) minimum trajectory** of the time-varying optimization (1) if each point of $h(t)$ is a local (or global) minimum of the time varying optimization (1) at time $t \in [0, \infty)$.*

After freezing the time t in (1) at a particular value, one may use local search methods to minimize $f(x, t)$. The notion of region of attraction is defined by resorting to the continuous-time model of local search algorithms.

Definition 2. *The **region of attraction** of a local minimum point $h(t)$ of $f(\cdot, t)$ at a given time is defined as:*

$$RA(h(t)) = \left\{ x_0 \in \mathbb{R}^n \mid \lim_{\tilde{t} \to \infty} x(\tilde{t}) = h(t), where \right.$$

$$\left. \frac{d\tilde{x}(\tilde{t})}{d\tilde{t}} = -\nabla_x f(\tilde{x}(\tilde{t}), t) \quad and \quad \tilde{x}(0) = x_0 \right\}$$

Definition 3. *Consider arbitrary positive scalars c and r. The function $f(x, t)$ is said to be **locally (c, r)-one-point strongly convex** around the local minimum trajectory $h(t)$ if*

$$\nabla_x f(e + h(t), t)^\top e \geq c \, \|e\|^2, \quad \forall e \in D, \quad \forall t \in [0, \infty) \tag{3}$$

where $D = \{e \in \mathbb{R}^n : \|e\| \leq r\}$. The region $D = \{e \in \mathbb{R}^n : \|e\| \leq r\}$ is called the region of locally (c, r)-one-point strong convexity around $h(t)$.

Note that (3) resembles the (locally) strong convexity condition for the function $f(x, t)$, but it is only expressed around the point $h(t)$. This restriction to a single point constitutes the definition of one-point strong convexity and it does not imply that the function is convex.

B. Derivation of time-varying gradient flow system

In many real-world applications, it is neither practical nor realistic to have solutions that abruptly change over time. To meet this requirement, we impose a soft constraint to the objective function by penalizing the deviation of its solution from the one obtained in the previous time step. This leads to the following sequence of optimization problems with **proximal regularization** (except for the initial optimization problem):

$$\min_{x \in \mathbb{R}^n} \quad f(x, \tau_0), \tag{4a}$$

$$\min_{x \in \mathbb{R}^n} \quad f(x, \tau_i) + \frac{\alpha \, \|x - x_{i-1}^*\|^2}{2(\tau_i - \tau_{i-1})}, i = 1, 2, \dots \tag{4b}$$

where x_{i-1}^* denotes an arbitrary local minimum of the modified optimization problem (4) obtained using a local search method at time iteration $i - 1$. Due to the first-order optimality condition, the local minimum x_i^* of (4) at time step τ_i satisfies the equation:

$$\nabla_x f(x_i^*, \tau_i) + \alpha \frac{x_i^* - x_{i-1}^*}{\tau_i - \tau_{i-1}} = 0 \tag{5}$$

We study the continuous-time limit of (5) as the time step $\tau_{i+1} - \tau_i$ attenuates to zero. This yields the ordinary differential equation:

$$\alpha \dot{x}(t) = -\nabla_x f(x(t), t), \quad x(0) = x_0^* \tag{6}$$

When $\alpha = 0$, the differential equation (6) reduces to the algebraic equation (2), which is indeed the first-order stationary condition for (1). When $\alpha > 0$, we will show that (ODE) has a unique solution defined for all $t \geq 0$ under the assumption that the solutions of (ODE) lie in a compact set[1].

Proposition 1 (Existence and uniqueness). *Suppose that $f(x, t)$ is continuous in t, and that its gradient is locally Lipschitz in x for all $t \geq 0$ and $x \in \mathbb{R}^n$. Let D be a compact subset of $\mathbb{R}^n$ containing x_0 such that every solution of (ODE) lies entirely in D. Then, (ODE) has a unique solution starting from $x_0 \in W$ and is defined for all $t \geq 0$.*

Proof. This results follows from [15, Theorem 3.3]. $\square$

Furthermore, in online optimization, it is desirable to predict the solution at a future time (namely, τ_i) only based on the information at the current time (namely, τ_{i-1}). This can be achieved by implementing the forward Euler method of (ODE):

$$\bar{x}_i^* = \bar{x}_{i-1}^* - \frac{\tau_i - \tau_{i-1}}{\alpha} \nabla_x f(\bar{x}_{i-1}^*, \tau_{i-1}) \tag{7}$$

(note that $\bar{x}_0^*, \bar{x}_1^*, \bar{x}_2^*, \dots$ show the approximate solutions). The following theorem explains the reason behind studying the continuous-time problem (ODE) in the remainder of this paper.

Proposition 2 (Convergence). *Given a local minimum x_0^* of (4a), as the time difference $\Delta_\tau = \tau_{i+1} - \tau_i$ approaches zero, any sequence of discrete local trajectories (x_k^Δ) converges to the (ODE) in the sense that for all fixed $T > 0$:*

$$\lim_{\Delta_\tau \to 0} \max_{0 \leq k \leq \frac{T}{\Delta_\tau}} \left\| x_k^\Delta - x(\tau_k, \tau_0, x_0^*) \right\| = 0 \tag{8}$$

and any sequence of $(\bar{x}_k^\Delta)$ updated by (7) converges to the (ODE) in the sense that for all fixed $T > 0$:

$$\lim_{\Delta_\tau \to 0} \max_{0 \leq k \leq \frac{T}{\Delta_\tau}} \left\| \bar{x}_k^\Delta - x(\tau_k, \tau_0, x_0^*) \right\| = 0 \tag{9}$$

Proof. The first part follows from Theorem 2 in [5]. For the second part, a direct application of the classical

---

[1]Checking the compactness assumption can be done via the Lyapunov's method without solving the differential equation.

results on convergence of the forward Euler method [12] immediately shows that the solution of (ODE) starting at a local minimum of (4a) is the continuous limit of the discrete local trajectory of the sequential regularized optimization (4). □

C. Jumping, tracking and escaping

In this paper, the objective is to study the case where there are at least two local minimum trajectories of the online time-varying optimization problem. Consider two local minimum trajectories $h_1(t)$ and $h_2(t)$.

Definition 4. *It is said that the solution of* (ODE) *(v,u)-jumps from $h_1(t)$ to $h_2(t)$ over the time interval $[t_1, t_2]$ if there exist $u > 0$ and $v > 0$ such that*

$$\mathcal{B}_v(h_1(t_1)) \subseteq RA(h_1(t_1))$$
$$\mathcal{B}_u(h_2(t_2)) \subseteq RA(h_2(t_2))$$
$$\forall x_1 \in \mathcal{B}_v(h_1(t_1)) \implies x(t_2, t_1, x_1) \in \mathcal{B}_u(h_2(t_2))$$

Definition 5. *It is said that $x(t, t_0, x_0)$ u-tracks $h_2(t)$ if there exist a finite time $T > 0$ and a constant $u > 0$ such that*

$$x(t, t_0, x_0) \in \mathcal{B}_u(h_2(t)), \qquad \forall t \geq T$$
$$\mathcal{B}_u(h_2(t)) \subseteq RA(h_2(t)), \quad \forall t \geq T$$

In this paper, the objective is to study the scenario where a solution $x(t, t_0, x_0)$ tracking a poor solution $h_1(t)$ at the beginning ends up tracking a better solution $h_2(t)$ after some time. This needs the notion of "escaping" which is a combination of jumping and tracking.

Definition 6. *It is said that the solution of* (ODE) *(v,u)-escapes from $h_1(t)$ to $h_2(t)$ if there exist $T > 0$, $u > 0$ and $v > 0$ such that*

$$\mathcal{B}_v(h_1(t_0)) \subseteq RA(h_1(t_0))$$
$$\mathcal{B}_u(h_2(t)) \subseteq RA(h_2(t)), \forall t \geq T$$
$$\forall x_0 \in \mathcal{B}_v(h_1(t_0)) \implies x(t, t_0, x_0) \in \mathcal{B}_u(h_2(t)), \forall t \geq T$$
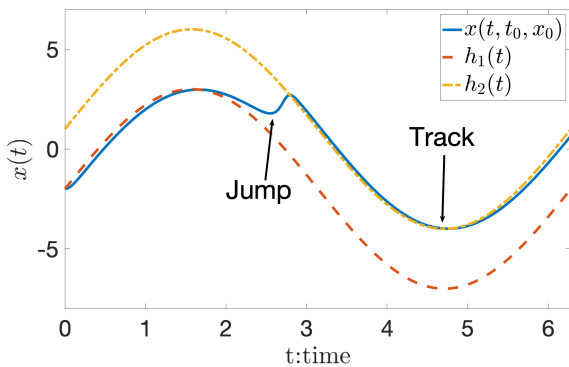

Fig. 2. Illustration of jumping and tracking.

Figure 2 illustrates the definitions of jumping and tracking.

IV. Optimization landscape after a change of variables

Given two isolated local minimum trajectories $h_1(t)$ and $h_2(t)$, one may use the change of variables $x(t, t_0, x_0) = e(t, t_0, e_0) + h_2(t)$ to transform (ODE) into the form

$$\dot{e}(t) = -\frac{1}{\alpha}\nabla_x f(e(t) + h_2(t), t) - \dot{h}_2(t) \qquad (13)$$

We use $e(t, t_0, e_0)$ to denote the solution of this differential equation starting at time $t = t_0$ with the initial point $e_0 = x_0 - h_2(t_0)$ and use $-\frac{1}{\alpha}U(e, t, \alpha)$ to denote the righthand side of (13). Note that $h_1(t)$ and $h_2(t)$ are local solutions of $f(x, t)$ and as long as $f(x, t)$ is time-varying, these functions cannot satisfy (ODE) in general.

A. Inertia creating a one-point strongly convex landscape

The differential equation (13) can be written as

$$\dot{e}(t) = -\frac{1}{\alpha}\nabla_e\left(f(e(t) + h_2(t), t) + \alpha\dot{h}_2(t)^\top e(t)\right) \quad (14)$$

This can be regarded as a time-varying gradient flow system of the original objective function $f(e+h_2(t), t)$ plus a time-varying perturbation $\alpha\dot{h}_2(t)^\top e$. During some time interval $[t_1, t_2]$, the time-varying perturbation $\alpha\dot{h}_2(t)^\top e$ may enable that the time-varying objective function $f(e + h_2(t), t) + \alpha\dot{h}_2(t)^\top e$ over the neighborhood of $h_1(t)$ becomes **one-point strongly convexified** with respect to $h_2(t)$. Under such circumstances, the time-varying perturbation $\alpha\dot{h}_2(t)^\top e$ prompts the solution of (14) starting in a neighborhood of $h_1(t)$ to move to a neighborhood of $h_2(t)$. Before analyzing this phenomenon, we illustrate the concept in an example.

Example 1. Consider $f(x, t) := g(x - b\sin(t))$, where

$$g(y) := 1/4y^4 + 2/3y^3 - 1/2y^2 - 2y$$

This time-varying objective has a spurious local minimum trajectory at $-2 + b\sin(t)$, a local maximum trajectory at $-1 + b\sin(t)$, and a global minimum trajectory at $1 + b\sin(t)$. In Figure 3, we show a bifurcation phenomenon numerically. The red lines are the solutions of (ODE) with the initial point -2. In the case with $\alpha = 0.3$ and $b = 5$, the solution of (ODE) winds up in the region of attraction of the global minimum trajectory. However, for the case with $\alpha = 0.1$ and $b = 5$, the solution of (ODE) remains in the region of attraction of the spurious local minimum trajectory.

In this example, the equation (14) can be expressed as $\dot{e}(t) = -\frac{1}{\alpha}\nabla_e\left(g(1 + e(t)) + 5\alpha\cos(t)e(t)\right)$. The landscape of the new time-varying function $g(1 + e) + 5\alpha\cos(t)e$ with the variable e is shown for two cases $\alpha = 0.3$ and $\alpha = 0.1$ in Figure 4. The red curves are the solutions of (14) starting from $e = -3$. One can observe that when $\alpha = 0.3$, the new landscape becomes one-point strongly convex around $h_2(t)$ over the whole region for some time interval, which provides (14) with the opportunity of escaping from the region around $h_1(t)$ to the region around $h_2(t)$. However, when $\alpha = 0.1$, there are always

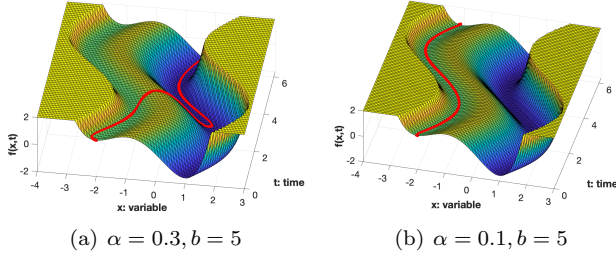(a) $\alpha = 0.3, b = 5$ (b) $\alpha = 0.1, b = 5$

Fig. 3. Illustration of Example 1 (in order to increase visibility, the objective function values are rescaled).

two locally one-point strongly convex regions around $h_1(t)$ and $h_2(t)$ and, therefore, (14) fails to escape the region around $h_1(t)$. To further inspect the case $\alpha = 0.3$,

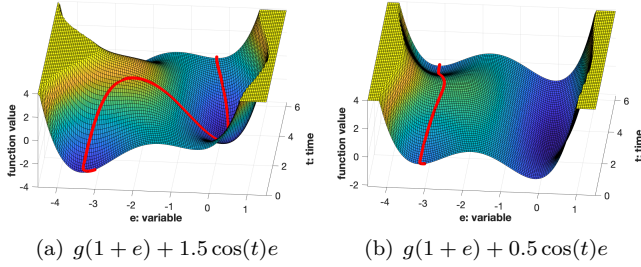(a) $g(1 + e) + 1.5 \cos(t)e$ (b) $g(1 + e) + 0.5 \cos(t)e$

Fig. 4. Illustration of time-varying landscape after change of variables for Example 1.

observe in Figure 5(a) that the landscape of the objective function $g(1 + e) + 1.5 \cos(0.9\pi)e$ shows that the region around the spurious local minimum trajectory $h_1(t)$ is one-point strongly convexified with respect to $h_2(t)$ at time $t = 0.9\pi$. This is consistent with the fact that the solution of $\dot{e} = -\frac{1}{0.3}\nabla_x g(1 + e) - 5 \cos(t)$ starting from $e = -3$ jumps to the neighborhood of 0 around time $t = 0.9\pi$, as demonstrated in Figure 5(c). Furthermore, if the time interval $[t_1, t_2]$ is relatively large enough to allow transition from a neighborhood of $h_1(t)$ to a neighborhood of $h_2(t)$, then the solution of (14) would move to the neighborhood of $h_2(t)$. In contrast, the region around $h_2(t)$ is never one-point strongly convexified with respect to $h_1(t)$, as shown in Figure 5(b). In the next subsection, we introduce the notion of the dominant trajectory after averaging to formally describe when the time-varying linear perturbation $\alpha \dot{h}_2(t)^\top e$ could help re-shape the objective landscape to become one-point strongly convexified.

B. Notion of the dominant trajectory after averaging

To avoid directly analyzing the time-varying system, we first introduce the notion of averaging of a time-varying function over a time interval $[t_1, t_2]$.

Definition 7. *A function $U_{av}(e, \alpha)$ is said to be the* ***average function*** *of $U(e, t, \alpha)$ over the time interval*

$[t_1, t_2]$ *if*

$$U_{av}(e, \alpha) = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} U(e, t, \alpha) d\tau$$

The time-invariant partial interval averaged system of (13) over the time interval $[t_1, t_2]$ can be written as

$$\dot{e} = -\frac{1}{\alpha} U_{av}(e, \alpha) \quad (15)$$

Then, (13) can be regarded as a time-invariant system (15) with the time-varying perturbation term $p(e(t), t, \alpha) = -\frac{1}{\alpha}(U(e(t), t, \alpha) - U_{av}(e, \alpha))$.

In the rest of the paper, we consider two local minimum trajectories $h_1(t)$ and $h_2(t)$ such that the time-varying function $f(x, t)$ is locally (c_2, r_2)-one-point strongly convex with respect to x around $h_2(t)$ in the region $\mathcal{B}_{r_2}(0)$ and that $h_2(t)$ is continuously differentiable. Now, we introduce the notion of the dominant trajectory after averaging.

Definition 8. *It is said that $h_2(t)$ is a (α, w)-****dominant trajectory after averaging**** with respect to $h_1(t)$ during $[t_1, t_2]$ over the region D_{v,ρ,r_2} if the time variation of $h_2(t)$ makes the average function $U_{av}^{h_2}(e.\alpha)$ in (15) become one-point strongly monotone over D_{v,ρ,r_2}, i.e.,*

$$U_{av}(e, \alpha)^\top (e - \bar{e}) \geq w \|e - \bar{e}\|^2, \quad \forall e \in D_{v,\rho,r_2} \quad (16)$$

where $w > 0$ is a constant., $\bar{e}$ is defined in (17) and D_{v,ρ,r_2} is defined as follows:

- *D_{v,ρ,r_2} is a compact positively invariant subset such that*

$$e_1 \in D_{v,\rho,r_2} \Rightarrow e(t, t_1, e_1) \in D_{v,\rho,r_2}, \forall t \in [t_1, t_2].$$

  where $e(t, t_1, e_1)$ is the solution of (13) staring from the initial point e_1 at the initial time t_1.
- *$D_{v,\rho,r_2} \supset D'_v \cup \mathcal{B}_\rho(0)$ where*

$$D'_v = \{e_1 \in \mathbb{R}^n : e_1 + h_2(t_1) \in \mathcal{B}_v(h_1(t_1))$$
$$\subseteq RA(h_1(t_1))\},$$
$$\rho \geq \sup_{\bar{e}:\|\bar{e}\|<r_2, 0 = U_{av}(\bar{e}, \alpha)} \|\bar{e}\|. \quad (17)$$

V. MAIN RESULTS

In this part, we derive different sufficient conditions under which the solution of (ODE) jumps from a poor local minimum trajectory to a better (or global) trajectory.

A. Jumping

In this subsection, we study the jumping property of (ODE) when $h_2(t)$ is a dominant trajectory after averaging.

Theorem 1 (Sufficient conditions for jumping from $h_1(t)$ to $h_2(t)$). *Suppose that $h_2(t)$ is a (α, w)-dominant trajectory after averaging with respect to $h_1(t)$ during $[t_1, t_2]$ over the region D_{v,ρ,r_2}. Assume that the following conditions are satisfied:*

(a) $g(1+e) + 1.5\cos(0.9\pi)e$ (b) $g(1+e) + 1.5\cos(0)e$ (c) solution of $\dot{e} = -\frac{1}{0.3}\nabla_x g(1+e) - 5\cos(t)$ starting from $e_0 = -3$
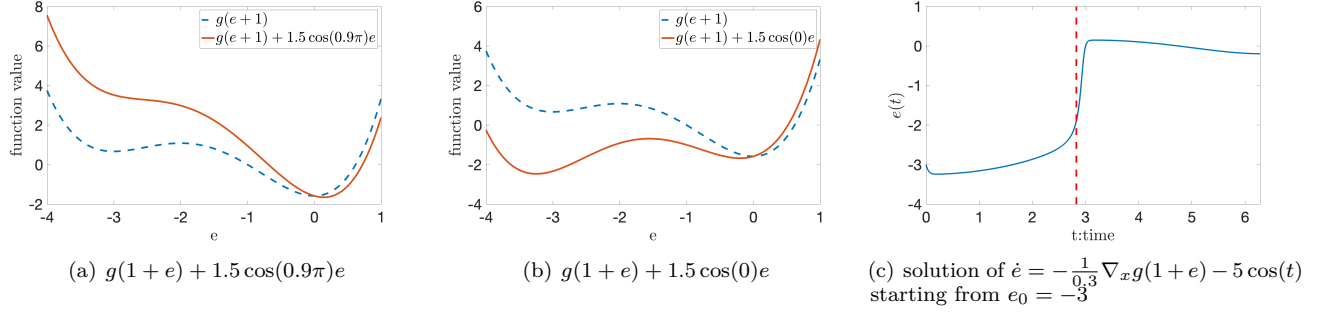
Fig. 5. Illustration of one-point strong convexification for Example 1.

1) *There exist some time-varying scalar functions $\delta_1(\alpha,t)$ and $\delta_1(\alpha,t)$ such that*

$$\|p(e(t),t,\alpha)\| \le \delta_1(\alpha,t)\|e - \bar{e}\| + \delta_2(\alpha,t), \quad (18)$$

for all $t \in [t_1, t_2]$ and there exist some positive constants $\eta_1(\alpha)$ and $\eta_2(\alpha)$ such that:

$$\int_{t_1}^{t} \delta_1(\alpha,\tau)d\tau \le \eta_1(\alpha)(t - t_1) + \eta_2(\alpha). \quad (19)$$

2) *For $\beta_1(\alpha) = \frac{w}{\alpha} - \eta_1(\alpha) > 0$ and $\beta_2(\alpha) = e^{\eta_2(\alpha)} \ge 1$, it holds that:*

$$\beta_2(\alpha)\|e_1 - \bar{e}\| e^{-\beta_1(\alpha)(t_2 - t_1)} + \beta_2(\alpha) \quad (20)$$
$$\int_{t_1}^{t_2} e^{-\beta_1(\alpha)(t_2 - \tau)}\delta_2(\alpha,\tau)d\tau \le r_2 - \rho, \forall e_1 \in D'_v$$

Then, the solution of (ODE) *will (v,r_2)-jump from $h_1(t)$ to $h_2(t)$ over the time interval $[t_1, t_2]$.*

Proof. First, notice that since D_{v,ρ,r_2} is a compact positively invariant set with respect to the dynamics (13), it follows from Proposition 1 that (13) has a unique solution defined for $t \in [t_1, t_2]$ whenever $e_1 \in D_{v,\rho,r_2}$. By using $V(e) = \frac{1}{2}\|e - \bar{e}\|^2 : D_{v,\rho,r_2} \to \mathbb{R}$ as the Lyapunov function for the system (13), the derivative of $V(e)$ along the trajectories of (13) can be obtained as

$$\dot{V}(e) = (e - \bar{e})^\top \left(-\frac{1}{\alpha}U_{av}(e,\alpha) + p(e,\alpha,t) \right)$$
$$\le -\frac{w}{\alpha}\|e - \bar{e}\|^2 + \delta_1(\alpha,t)\|e - \bar{e}\|^2 + \delta_2(\alpha,t)\|e - \bar{e}\|$$

Since $V(e) = \frac{1}{2}\|e - \bar{e}\|^2$, one can derive an upper bound on $\dot{V}$ as

$$\dot{V}(e) \le -\left[\frac{2w}{\alpha} - 2\delta_1(\alpha,t)\right]V(e) + \delta_2(\alpha,t)\sqrt{2V(e)}$$

To obtain a linear differential inequality, we consider $W(t) = \sqrt{V(e(t))}$. When $V(e(t)) \ne 0$, it holds that $\dot{W} = \dot{V}/2\sqrt{V}$ and

$$\dot{W} \le -\left[\frac{w}{\alpha} - \delta_1(\alpha,t)\right]W + \frac{\delta_2(\alpha,t)}{\sqrt{2}} \quad (21)$$

When $V(e(t)) = 0$, we have $e(t) = \bar{e}$. Writing the Tylor expansion of $e(t + \epsilon)$ for a sufficiently small ϵ yields that

$$e(t + \epsilon) = e(t) + \epsilon\left(-\frac{1}{\alpha}U_{av}(e,\alpha) + p(e,\alpha,t) \right) + o(\epsilon)$$
$$= \bar{e} + \epsilon p(\bar{e},\alpha,t) + o(\epsilon)$$

This implies that

$$V(e(t+\epsilon)) = \frac{\epsilon^2}{2}\|p(\bar{e},\alpha,t)\|^2 + o(\epsilon^2).$$

Therefore,

$$D^+W(t) = \lim_{\epsilon \to 0^+} \sup \frac{W(t+\epsilon) - W(t)}{\epsilon}$$
$$= \lim_{\epsilon \to 0^+} \sup \frac{\sqrt{\frac{\epsilon^2}{2}\|p(\bar{e},\alpha,t)\|^2 + o(\epsilon^2)}}{\epsilon} \quad (22)$$
$$= \frac{1}{\sqrt{2}}\|p(\bar{e},\alpha,t)\|$$
$$\le \frac{1}{\sqrt{2}}\delta_2(\alpha,t)$$

Thus, (21) is also satisfied when $V = 0$, and accordingly $D^+W(t)$ satisfies (21) for all values of V. Since W is scalar and the right-hand side of (21) is continuous in t and locally Lipschitz in W for all $t \in [t_1, t_2]$ and $W \ge 0$, the comparison lemma is applicable. In addition, the right-hand side of (21) is linear and a closed-form expression for the solution of the first-order linear differential equation of W can be obtained. Hence, $W(t)$ satisfies

$$W(t) \le \phi(t,t_1)W(t_1) + \frac{1}{\sqrt{2}}\int_{t_1}^{t}\phi(t,\tau)\delta_2(\alpha,\tau)d\tau \quad (23)$$

where the translation function $\phi(t,t_1)$ is given by

$$\phi(t,t_1) = \exp\left[-\frac{w}{\alpha}(t - t_1) + \int_{t_1}^{t}\delta_1(\alpha,\tau)d\tau \right]. \quad (24)$$

$$\|e(t) - \bar{e}\| \le \phi(t,t_1)\|e_1 - \bar{e}\| + \int_{t_1}^{t}\phi(t,\tau)\delta_2(\alpha,\tau)d\tau \quad (25)$$

Since $\int_{t_1}^{t}\delta_1(\alpha,\tau)d\tau \le \eta_1(\alpha)(t - t_1) + \eta_2(\alpha)$, and using $\beta_1(\alpha) = \frac{w}{\alpha} - \eta_1(\alpha) > 0$ and $\beta_2(\alpha) = e^{\eta_2(\alpha)} \ge 1$ in (25), it

holds that

$$\|e(t) - \bar{e}\| \leq \beta_2(\alpha) \|e_1 - \bar{e}\| e^{-\beta_1(\alpha)(t - t_1)}$$
$$+ \beta_2(\alpha) \int_{t_1}^{t} e^{-\beta_1(\alpha)(t - \tau)} \delta_2(\alpha, \tau) d\tau \quad (26)$$

By taking $e_1 \in D_v' \subset D_{v,\rho,r_2}$, since D_{v,ρ,r_2} retains trajectories starting from a feasible initial point with respect to the dynamics (13) for $t \in [t_1, t_2]$, any trajectory of (13) starting from D_v' will stay in D_{v,ρ,r_2}. Thus, the bound in (26) is valid. If t_2 satisfies

$$\beta_2(\alpha) \|e_1 - \bar{e}\| e^{-\beta_1(\alpha)(t_2 - t_1)}$$
$$+ \beta_2(\alpha) \int_{t_1}^{t_2} e^{-\beta_1(\alpha)(t_2 - \tau)} \delta_2(\alpha, \tau) d\tau \leq r_2 - \rho$$

then $\|e(t_2) - \bar{e}\| \leq r_2 - \rho$. Since $\bar{e} \in \mathcal{B}_\rho(0)$, we have $\|e(t_2)\| \leq r_2$. This shows that the solution of (13) jumps from $h_1(t)$ to $h_2(t)$ during the time interval $[t_1, t_2]$. $\quad\square$

Remark 1. *Condition (1) in Theorem 1 means that the original time-varying system is not too distant from the time-invariant averaged system, and Condition (2) means that $[t_1, t_2]$ needs to be large enough to allow the transition of points from a neighborhood of $h_1(t)$ to a neighborhood of $h_2(t)$.*

B. Tracking

In this subsection, we study the tracking property of the local minimum trajectory $h_2(t)$. First, notice that if $h_2(t)$ is not constant, the right-hand side of (ODE) is nonzero while the left-hand side is zero. Therefore, $h_2(t)$ is not a solution of (ODE) in general. This is because the solution of (ODE) approximates the continuous limit of a discrete local trajectory of the sequential regularized optimization problem (4). However, to preserve the optimality of the solution with regards to the original time-varying optimization problem without any proximal regularization, it is required to guarantee that the solution of (ODE) is close to $h_2(t)$. The next theorem shows that every local minimum trajectory can be tracked for a sufficiently small α.

Theorem 2 (Sufficient condition for tracking). *Assume that the time-varying function $f(x, t)$ is locally (c_2, r_2)-one-point strongly convex around $h_2(t)$. Then, $h_2(t)$ can be tracked if α is sufficiently small. In particular, given $0 < \theta' < 1$, $\gamma := \sup_{t \geq 0} \|\dot{h}_2(t)\|$, $u := \frac{\alpha\gamma}{\theta' c_2}$, $\|x_0 - h_2(0)\| \leq r_2$ and $\alpha < \frac{c_2\theta' r_2}{\gamma}$, the solution $x(t, t_0, x_0)$ will u-track $h_2(t)$ exponentially with the convergence rate $(1 - \theta')\frac{c_2}{\alpha}$, namely,*

$$for\ t_0 \leq t \leq t_0 + \frac{\alpha}{c_2(1 - \theta')} \ln(\frac{r_2}{u}):$$
$$\|x(t, t_0, x_0) - h_2(t)\| \leq r_2 \exp{-(1 - \theta')\frac{c_2}{\alpha}(t - t_0)},$$
$$for\ t > t_0 + \frac{\alpha}{c_2(1 - \theta')} \ln(\frac{r_2}{u}):$$
$$\|x(t, t_0, x_0) - h_2(t)\| \leq u.$$

Proof. The proof is based on Lemma 9.2 in [15] and the details of the proof are deferred to the technical report [4] due to the space restriction. $\quad\square$

C. Escaping

Combining Theorem 1 with Theorem 2 immediately yields a sufficient condition on escaping from one local minimum trajectory to the dominant trajectory. The proof is omitted for brevity.

Theorem 3 (Sufficient conditions for escaping). *Suppose that $h_2(t)$ is a (α, w)-dominant trajectory after averaging with respect to $h_1(t)$ during $[t_1, t_2]$ over the region D_{v,ρ,r_2}. Let $\gamma = \sup_{t \geq 0} \|\dot{h}_2(t)\|$, $0 < \theta' < 1$, $\mathcal{B}_v(h_1(t_1)) \subseteq RA(h_1(t_1))$ and $u = \frac{\alpha\gamma}{\theta' c_2}$. Under the conditions of Theorem 1 , if $\alpha < \frac{r_2 c_2 \theta'}{\gamma}$, the solution of (ODE) will (v, r_2)-escape from $h_1(t)$ to $h_2(t)$ after $t \geq t_2$.*

VI. ILLUSTRATIVE EXAMPLE

Example 2. We study a low-dimensional example for which one can visualize the aforementioned conditions. Consider the non-convex function

$$g(x) = 0.5e + 20e^{-d} - 20e^{-\sqrt{0.5(x_1^2 + x_2^2) + d^2}} \quad (27)$$
$$- 0.5e^{(0.5(\cos(2\pi x_1) + \cos(2\pi x_2)))}.$$

This function has a global minimum at $(0, 0)$ with the optimal value 0 and many spurious local minima. Its landscape is shown in Figure 6. When $d = 0$, this function is called the Ackley function [1], which is a benchmark function for global optimization algorithms. To make this function twice continuously differentiable, we take $d = 0.01$. Consider the time-varying objective function $f(x, t) = g(x - z(t))$, where $z(t) = [7\sin(t), 7\cos(t)]^\top$. Two local minimum trajectories are $h_1(t) = [1.95, 0.97]^\top + z(t)$ and $h_2(t) = [0, 0]^\top + z(t)$. It can be shown that $g(x)$ is locally $(3.3, 1.1)$-one-point strongly convex with respect to the origin, which implies that $f(x, t)$ is locally $(3.3, 1.1)$-one-point strongly convex around $h_2(t)$. To ensure that the solution of (ODE) will track $h_2(t)$, we need to take $\alpha \leq \frac{c_2 r_2}{\sup_{t \geq 0} \|\dot{z}(t)\|}$. In this case, $\alpha = 0.5$ simply satisfies the tracking condition. This corresponding averaged system (15) has an equilibrium point at $[-0.0034, 0.0007]^\top$. Then we can take $\rho = 0.01$. Let $\mathcal{B}_\rho(0) = \mathcal{B}_{0.01}(0)$, $D_v' = \{e \in \mathbb{R}^n : e_1 + h_2(t_1) \in \mathcal{B}_{0.1}(h_1(t_1))$ and $D_{v,\rho,r_2} = [-0.2, 2.1] \times [-0.1, 1.1]$. In addition, on the boundary points $e_1 = 2.1$ and $e_1 = -0.2$, the derivative of e_1 along the dynamics (13) is negative and positive, respectively, for all $e_2 \in [-0.1, 1.1]$ and $t \in [0, \frac{\pi}{8}]$. Similarly, on the boundary points $e_2 = 1.1$ and $e_2 = -0.1$, the derivative of e_2 along the dynamics (13) is negative and positive, respectively, for all $e_1 \in [-0.2, 2.1]$ and $t \in [0, \frac{\pi}{8}]$. This implies that the set D_{v,ρ,r_2} retains trajectories with respect to (13) for $t \in [0, \frac{\pi}{8}]$. Then, it can be shown that $h_2(t)$ is a $(0.5, 1.3)$-dominant trajectory with respect to $h_1(t)$ in D_{v,ρ,r_2} during $[0, \frac{\pi}{8}]$. Furthermore, the conditions in the theorem 1 are satisfied. Thus, the conditions of Theorem 3 are all met, and therefore the solution of (13)

will $(0.1, 1.1)$-escape from $h_1(t)$ to $h_2(t)$. Furthermore, we have verified for 1000 runs of random initialization over $x(0) - z(0) \in [-5, 5] \times [-5, 5]$ that all solutions of the corresponding (ODE) will sequentially jump over the local minimum trajectories and end up tracking the global trajectory $[0, 0]^\top + z(t)$ after $t \geq 10\pi$.
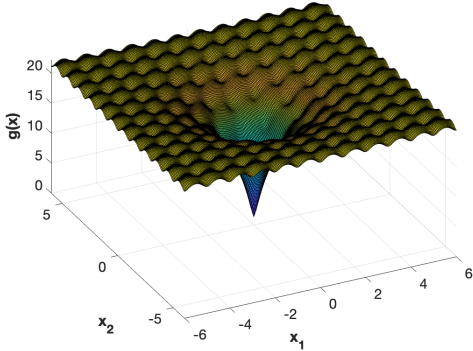

Fig. 6. Illustration of the objective landscape of (27)

VII. Conclusion

In this work, we study the landscape of time-varying nonconvex optimization problems. The objective is to understand when simple local search algorithms can find (and track) time-varying global solutions of the problem over time. We introduce a time-varying gradient flow system with controllable inertia. Via a change of variables, the time-varying gradient flow system is regarded as a composition of a time-varying gradient term and a time-varying perturbation term due to the inertia. We introduce the notion of the dominant trajectory and show that the time-varying perturbation term due to the inertia re-shapes the landscape by potentially making it one-point strongly convex over a large region during some time interval. We also introduce the notions of jumping, tracking and escaping, and use them to develop sufficient conditions under which the time-varying solution escapes from a poor local trajectory when the global minimum trajectory is dominant.

References

[1] David H. Ackley. *A Connectionist Machine for Genetic Hillclimbing.* Kluwer Academic Publishers, Norwell, MA, USA, 1987.

[2] A Bernstein, E Dall'Anese, and A Simonetto. Online optimization with feedback, 2018. *arXiv preprint arXiv:1804.05159*, 2018.

[3] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[4] Yuhao Ding, Javad Lavaei, and Murat Arcak. Escaping spurious local minimum trajectories in online time-varying nonconvex optimization. *arXiv preprint arXiv:1912.00561*, 2019.

[5] S Fattahi, C Josz, R Mohammadi, J Lavaei, and S Sojoudi. Absence of spurious local trajectories in time-varying optimization. *arXiv preprint arXiv:1905.09937*, 2019.

[6] Salar Fattahi and Somayeh Sojoudi. Exact guarantees on the absence of spurious local minima for non-negative robust principal component analysis. *arXiv preprint arXiv:1812.11466*, 2018.

[7] Mahyar Fazlyab, Cameron Nowzari, George J Pappas, Alejandro Ribeiro, and Victor M Preciado. Self-triggered time-varying convex optimization. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 3090–3097. IEEE, 2016.

[8] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[9] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[10] Jack K Hale. *Ordinary differential equations.* Wiley-Interscience, 1980.

[11] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[12] Arieh Iserles. *A first course in the numerical analysis of differential equations.* Cambridge University Press, 2009.

[13] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.

[14] Cedric Josz, Yi Ouyang, Richard Zhang, Javad Lavaei, and Somayeh Sojoudi. A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization. In *Advances in Neural Information Processing Systems*, pages 2441–2449, 2018.

[15] Hassan K Khalil. *Nonlinear systems.* Upper Saddle River, 2002.

[16] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

[17] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems*, pages 2845–2853, 2015.

[18] Olivier Massicot and Jakub Marecek. On-line non-convex constrained optimization. *arXiv preprint arXiv:1909.07492*, 2019.

[19] Julie Mulvaney-Kemp, Salar Fattahi, and Javad Lavaei. Smoothing property of load variation promotes finding global solutions of time-varying optimal power flow, 2020.

[20] Andrea Simonetto. Time-varying convex optimization via time-varying averaged operators. *arXiv preprint arXiv:1704.07338*, 2017.

[21] Andrea Simonetto, Aryan Mokhtari, Alec Koppel, Geert Leus, and Alejandro Ribeiro. A class of prediction-correction methods for time-varying convex optimization. *IEEE Transactions on Signal Processing*, 64(17):4576–4591, 2016.

[22] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[23] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016.

[24] Yujie Tang, Emiliano Dall'Anese, Andrey Bernstein, and Steven Low. Running primal-dual gradient method for time-varying nonconvex problems. *arXiv preprint arXiv:1812.00613*, 2018.

[25] Yujie Tang, Krishnamurthy Dvijotham, and Steven Low. Real-time optimal power flow. *IEEE Transactions on Smart Grid*, 8(6):2963–2973, 2017.

[26] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[27] Richard Y Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *Journal of Machine Learning Research*, 2019.