Time-variation in online nonconvex optimization enables escaping from spurious local minima

Yuhao Ding, Javad Lavaei, and Murat Arcak

Abstract—**A major limitation of online algorithms that track the optimizers of time-varying nonconvex optimization problems is that they focus on a specific local minimum trajectory, which may lead to poor spurious local solutions. In this paper, we show that the natural temporal variation may help simple online tracking methods find and track time-varying global minima. To this end, we investigate the properties of a time-varying projected gradient flow system with inertia, which can be regarded as the continuous-time limit of (1) the optimality conditions for a discretized sequential optimization problem with a proximal regularization and (2) the online tracking scheme. We introduce the notion of the dominant trajectory and show that the inherent temporal variation could reshape the landscape of the Lagrange functional and help a proximal algorithm escape the spurious local minimum trajectories if the global minimum trajectory is dominant. For a problem with twice continuously differentiable objective function and constraints, sufficient conditions are derived to guarantee that no matter how a local search method is initialized, it will track a time-varying global solution after some time. The results are illustrated on a benchmark example with many local minima.**

Index Terms—**Time-varying optimization, nonconvex optimization, stability analysis.**

I. Introduction

$\mathbf{I}$N this paper, we study the following equality-constrained time-varying optimization problem:

$$\min_{x(t) \in \mathbb{R}^n} \quad f(x(t), t) \tag{1}$$
$$\text{s.t.} \quad g(x(t), t) = 0$$

where $t \geq 0$ denotes the time and $x(t)$ is the optimization variable that depends on t. Moreover, the objective function $f : \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$ and the constraint function $g(x, t) = (g_1(x, t), \dots, g_m(x, t))$ with $g_k : \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$

Y. Ding and J. Lavaei are with the Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94709 USA (e-mail: yuhao_ding@berkeley.edu; lavaei@berkeley.edu).

M. Arcak is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94709 USA (e-mail: arcak@berkeley edu).

for $k = 1, ..., m$ are assumed to be twice continuously differentiable in state x and continuously differentiable in time t. For each time t, the function $f(x, t)$ could potentially be nonconvex in x with many local minima and the function $g(x, t)$ could also potentially be nonlinear in x, leading to a nonconvex feasible set. The objective is to solve the above problem online under the assumption that at any given time t the function $f(x, t')$ and $g(x, t')$ are known for all $t' \leq t$ while no knowledge about $f(x, t')$ or $g(x, t')$ may be available for any $t' > t$. Therefore, the problem (1) 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. More precisely, the goal is to devise an algorithm that can track a global solution of (1) as a function of time t with some error at the initial time and a diminishing error after some time.

If $f(x, t)$ and $g(x, t)$ do not change over time, the problem reduces to a classic (time-invariant) optimization problem. It is known that simple local search methods, such as stochastic gradient descent (SGD) [2], may be able to find a global minimum of such time-invariant problems (under certain conditions) for almost all initializations due to the randomness embedded in SGD [3]–[5]. 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 (ODE), which is the counterpart of the classic projected gradient flow system for time-invariant optimization problems [6] and serves as a continuous-time limit of the discrete online tracking method for (1) with the proximal regularization. This ODE is given as

$$\dot{x}(t) = -\frac{1}{\alpha} \mathcal{P}(x(t), t) \nabla_x f(x(t), t) - \mathcal{Q}(x(t), t) g'(x(t), t) \tag{P-ODE}$$

where $\alpha > 0$ is a constant parameter named **inertia** due to a **proximal regularization**, $g'(z, t) = \frac{\partial g(z, t)}{\partial t}$, $\mathcal{P}(x(t), t)$ and $\mathcal{Q}(x(t), t)$ are matrices related to the Jacobian of $g(x, t)$ that will be derived in detail later. A system of

(a) $\alpha = 0.3, b = 5$ (b) $\alpha = 0.1, b = 5$ (c) $\alpha = 0.8, b = 5$ (d) $\alpha = 0.1, b = 10$
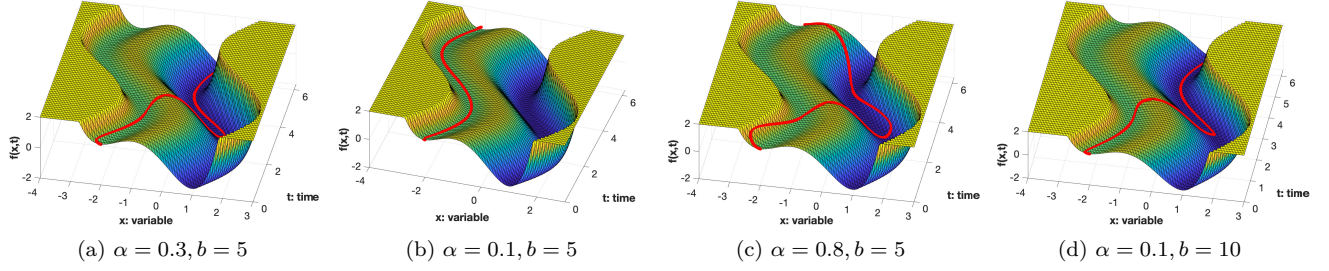
Fig. 1. Illustration of Example 1 (in order to increase visibility, the objective function values are rescaled). Jumping from a spurious local minimum trajectory to a global minimum trajectory occurs in Figure 1a and 1d when the inertia α and the change (controlled by the parameter b) of local minimum trajectory are appropriate.

the form (P-ODE) is called a **time-varying projected gradient system with inertia** α. The behavior of the solutions of this system initialized at different points depends on the value of α. In the unconstrained case, this ODE reduces to the **time-varying gradient system with inertia** α given as

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

In what follows, we offer a motivating example without constraints (to simplify the visualization) before stating the goals of this paper.

A. Motivating example

Example 1. *Consider $f(x, t) := \bar{f}(x - b\sin(t))$, where*

$$\bar{f}(y) := \frac{1}{4}y^4 + \frac{2}{3}y^3 - \frac{1}{2}y^2 - 2y$$

This time-varying objective has a spurious (non-global) 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 1, we show a bifurcation phenomenon numerically. The red lines are the solutions of (P-ODE) with the initial point -2. In the case with $\alpha = 0.3$ and $b = 5$, the solution of (P-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 (P-ODE) remains in the region of attraction of the spurious local minimum trajectory. In the case with $\alpha = 0.8$ and $b = 5$, the solution of (P-ODE) fails to track any local minimum trajectory. In the case with $\alpha = 0.1$ and $b = 10$, the solution of (P-ODE) winds up in the region of attraction of the global minimum trajectory.

Two observations can be made here. First, jumping from a local minimum trajectory to a better trajectory tends to occur with the help of a relatively large inertia when the local minimum trajectory changes the direction abruptly and there happens to exist a better local minimum trajectory in the direction of the inertia. Second, when the inertia α is relatively small, the solution of (P-ODE) tends to track a local (or global) minimum trajectory closely and converges to that trajectory quickly.

Example 2. *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 inequality constraints of the problem. Let $g(x, t) = 0$ describe the time-varying demand constraint. Assume that the load data corresponds to the California data for August 2019. As discussed in [7], this time-varying OPF 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 2. This observation has been made in [7] for a discrete-time version of the problem, but it also holds true for the continuous-time (ODE) model.
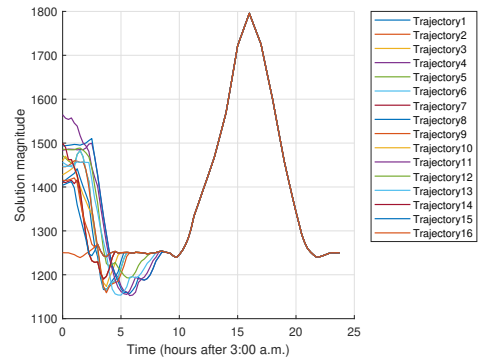

Fig. 2. $|x(t)|$ (magnitude of the solution of (ODE)).

B. Our contributions

To mathematically study the observations made in Example 1 and Example 2 for a general time-varying nonconvex optimization problem with equality constraints, we focus on the aforementioned time-varying projected gradient flow system with inertia α as a continuous-time limit of an online updating scheme for (1). We first introduce a time-varying Lagrange functional to unify the analysis of unconstrained problems and equality-constrained problems, and make the key assumption that the time-varying Lagrange functional is locally one-point strongly convex around each local minimum trajectory. This assumption is justified by

the second-order sufficient optimality conditions. A key property of (P-ODE) is that its solution will remain in the time-varying feasible region if the initial point is feasible for (1), which allows us to use the Lyapunov technique without worrying about the feasibility of the solution. Then, we show that the time-varying projected gradient flow system with inertia α is a continuous-time limit of the Karush–Kuhn–Tucker (KKT) optimality conditions for a discretized sequential optimization problem with a proximal regularization. The existence and uniqueness of the solution for such ODE is proven.

As a main result of this work, it is proven 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 (in the unconstrained case) or the Langrange functional (in the constrained case) 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 two 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 (P-ODE) can track a global minimum trajectory. To illustrate how the time variation nature of an online optimization problem promotes escaping a spurious minimum trajectory, we offer a case study with many shallow minimum trajectories.

C. Related work

Online time-varying optimization problems: Time-varying optimization problems of the form (1) arise in the real-time optimal power flow problem [8], [9] for which the power loads and renewable generations are time-varying and operational decisions should be made every 5 minutes, as well as in the real-time estimation of the state of a nonlinear dynamic system [10]. Other examples include model predictive control [11], time-varying compressive sensing [12], [13] and online economic optimization [14], [15]. There are many researches on the design of efficient online algorithms for tracking the optimizers of time-varying convex optimization problems [16]–[19]. With respect to time-varying nonconvex optimization problems, the work [20] presents a comprehensive theory on the structure and singularity of the KKT trajectories for time-varying optimization problems. On the algorithm side, [8] provides regret-type results in the case where the constraints are lifted to the objective function via penalty functions. [21] develops a running regularized primal-dual gradient algorithm to track a KKT trajectory, and offers asymptotic bounds on the tracking error. [22] obtains an ODE to approximate the KKT trajectory and derives an algorithm based on a predictor-corrector method to track the ODE solution.

Recently, [23] proposed 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. It developed a differential equation to characterize this phenomenon (which is the basis of the current work), but it lacked mathematical conditions to guarantee this desirable behavior. The paper [7] also studies this phenomenon 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 [23] and [7] by mathematically studying when such an escaping is possible.

Local search methods for global optimization: Nonconvexity is inherent in many real-world problems: the classical compressive sensing and matrix completion/sensing [24]–[26], training of deep neural networks [27], the optimal power flow problem [28], and others. From the classical complexity theory, this nonconvexity is perceived to be the main contributor to the intractability of these problems. However, 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, [29] shows that the gradient descent with a random initialization could avoid the saddle points almost surely, and [3] and [4] prove that a perturbed gradient descent and SGD could escape the saddle points efficiently. Furthermore, it has been shown that nearly-isotropic classes of problems in matrix completion/sensing [30]–[32], robust principle component analysis [33], [34], and dictionary recovery [35] have benign landscape, implying that they are free of spurious local minima. The work [5] proves that SGD could help escape sharp local minima of a loss function by taking the alternative view that SGD works on a convolved (thus smoothed) version of the loss function. However, these results are all for time-invariant optimization problems for which the landscape is time-invariant. 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 optimization 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 [36], [37]. Perhaps, the simplest and oldest example is the gradient flow system for the gradient descent algorithm with an infinitesimally small step size. The recent papers [38]–[40] study accelerated gradient methods for convex optimization problems from a continuous-time perspective. In addition, the continuous-time limit of the gradient descent is also employed to analyze various non-convex optimization problems, such as deep linear neural networks

[41] and matrix regression [42]. It is natural to analyze the continuous-time limit of an online algorithm for tracking a KKT trajectory of time-varying optimization problem [16], [21]–[23].

D. Paper organization

This paper is organized as follows. Section II presents some preliminaries for time-varying optimization with equality constraints and the derivation of time-varying projected gradient flow with inertia. Section III offers an alternative view on the landscape of time-varying nonconvex optimization problems after a change of variables and explains the role of the time variation of the constraints. Section IV analyzes the jumping, tracking and escaping behaviors of local minimum trajectories. Section V illustrates the phenomenon that the time variation of an online optimization problem can assist with escaping spurious local minimum trajectories, by working on a benchmark example with many shallow minimum trajectories. Concluding remarks are given in Section VI.

E. Notation

The notation $\|\cdot\|$ represents the Euclidean norm. The interior of the interval $\bar{I}_{t,2}$ is denoted by $\text{int}(\bar{I}_{t,2})$. 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 shorthand notation $x(t)$ if the initial condition (t_0, x_0) is clear from the context.

II. Preliminaries and Problem Formulation

A. Time-varying optimization with equality constraints

The first-order KKT conditions for the time-varying optimization (1) are as follows:

$$0 = \nabla_x f(x(t), t) + \mathcal{J}_g(x(t), t)^\top \lambda(t) \tag{2a}$$
$$0 = g(x(t), t) \tag{2b}$$

where $\mathcal{J}_g(z, t) := \frac{\partial g(z,t)}{\partial z}$ denotes the Jacobian of $g(\cdot, \cdot)$ with respect to the first argument and $\lambda(t) \in \mathbb{R}^m$ is a Lagrange multiplier associated with the equality constraint. We first make some assumptions below.

Assumption 1. $f : \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$ is twice continuously differentiable in $x \in \mathbb{R}^n$ and continuously differentiable in $t \geq 0$. $g_k : \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$ is twice continuously differentiable in $x \in \mathbb{R}^n$ and twice continuously differentiable in $t \geq 0$ for $k = 1, \dots, m$. Moreover, at any given time t, $f(x,t)$ is uniformly bounded from below over the set $\{x \in \mathbb{R}^n : g(x,t) = 0\}$, meaning that there exists a constant M such that $f(x,t) \geq M$ for all $x \in \{x \in \mathbb{R}^n : g(x,t) = 0\}$ and $t \geq 0$.

Assumption 2. The feasible set at t defined as

$$\mathcal{M}(t) := \{x \in \mathbb{R}^n : g(x,t) = 0\}$$

is nonempty for all $t \geq 0$.

Assumption 3. For all $t \geq 0$ and $x \in \mathcal{M}(t)$, the matrix $\mathcal{J}_g(x, t)$ has full row-rank.

Remark 1. Although Assumption 3 is somewhat stronger than the Linear independence constraint qualification [43], it is necessary for our following analysis because with different values of α and different initial points, the solution of (P-ODE) may land anywhere in the feasible region. Furthermore, Sard's theorem [44] ensures that if the constraint function $g(\cdot, t)$ is sufficiently smooth, then the set of values of $g(\cdot, t)$, denoted as $\mathcal{S}(t)$, for which $\mathcal{J}_g(x, t)$ is not full row-rank has measure 0. Thus, Assumption 3 is satisfied if $0 \notin \mathcal{S}(t)$ where $\mathcal{S}(t)$ is only a set with measure 0. Finally, if the inertia parameter α is fixed and the initial point of (P-ODE) is a local solution, then the work [23] provides a sophisticated proof for the existence and uniqueness of the solution for a special class of (P-ODE) under a minor assumption that the Jacobian has full-row rank only at the discrete local trajectories (which is defined in the paragraph after equation (10) in our work). However, to be able to study the solution of (P-ODE) for all $\alpha > 0$ and any initial feasible point and keep the focus of the paper on studying the escaping behavior, we made Assumption 3.

Under Assumption 3, the matrix $\mathcal{J}_g(x(t), t)\mathcal{J}_g(x(t), t)^\top$ is invertible and therefore $\lambda(t)$ in (2a) can be written as

$$\lambda(t) = -(\mathcal{J}_g(x(t), t)\mathcal{J}_g(x(t), t)^\top)^{-1}\mathcal{J}_g(x(t), t)\nabla_x f(x(t), t) \tag{3}$$

Since $\lambda(t)$ is written as a function of $x(t)$ in (3), we also denote it as $\lambda(x(t), t)$. Now, (2a) can be written as

$$0 = \Big[I_n - \mathcal{J}_g(x(t), t)^\top (\mathcal{J}_g(x(t), t)\mathcal{J}_g(x(t), t)^\top)^{-1}$$
$$\mathcal{J}_g(x(t), t)\Big]\nabla_x f(x(t), t) \tag{4}$$

where I_n is the identity matrix in $\mathbb{R}^{n \times n}$. For the sake of readability, we introduce the symbolic notation

$$\mathcal{P}(x(t), t) := I_n - \mathcal{J}_g(x(t), t)^\top (\mathcal{J}_g(x(t), t)\mathcal{J}_g(x(t), t)^\top)^{-1}$$
$$\mathcal{J}_g(x(t), t)$$

which is the orthogonal projection operation onto T_x^t, where T_x^t denotes the tangent plane of $g(x(t), t)$ at the point $x(t)$ and the time t. It is convenient and conventional to introduce the time-varying Lagrange functional

$$L(x, \lambda, t) = f(x, t) + \lambda g(x, t) \tag{5}$$

In terms of this functional, (4) can be written as

$$0 = \nabla_x L(x, \lambda, t) \tag{6}$$

where λ is given in (3). Here, $\nabla_x L(x, \lambda, t)$ means first taking the partial gradient with respect to the first argument and then using the formula (3) for λ. Since the solution is time-varying, we define the notion of the local (or global) minimum trajectory below.

Definition 1. A continuous trajectory $h : I_t \to \mathbb{R}^n$, where $I_t \subseteq [0, \infty)$, 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) for every $t \in I_t$.

In this paper, we focus on the case when the local minimum trajectories will not cross, bifurcate or disappear by assuming the following uniform regularity condition.

Assumption 4. For each local minimum trajectory $h(t)$, its domain I_t is $[0, \infty)$ and $h(t)$ satisfies the second-order sufficient optimality conditions uniformly, meaning that $\nabla^2_{xx} L(h(t), \lambda, t)$ is positive definite on $T^t_{h(t)} = \{y : \mathcal{J}_g(h(t), t)^\top y = 0\}$ for all $t \in [0, \infty)$.

Lemma 1. Under Assumptions 1-4, each local minimum trajectory $h(t)$ is differentiable and isolated, and therefore it can not bifurcate or merge with other local minimum trajectories.

Proof. Under Assumptions 1-4, one can apply the inverse function theorem to (2) (see [45, Theorem 4.4, Example 4.7]) to conclude that for every $h(\bar{t})$ and $\bar{t}$, there exist an open set $\mathcal{S}_{h(\bar{t})}$ containing $h(\bar{t})$ and an open set $\mathcal{S}_{\bar{t}}$ containing $\bar{t}$ such that there exist a unique differentiable function $\bar{x}(t)$ in $\mathcal{S}_{h(\bar{t})}$ for all $t \in \mathcal{S}_{\bar{t}}$ where $x(t)$ is the isolated local minimizer of the time-varying optimization problem (1). Because of this uniqueness property and the continuity of the local minimum trajectory $h(t)$, $x(t)$ must coincide with $h(t)$ for all $t \in \mathcal{S}_{\bar{t}}$. Then, because the above property holds uniformly for every $t \in [0, \infty)$, $h(t)$ must be a differentiable isolated minimum trajectory. $\qquad\square$

After freezing the time t in (1) at a particular value, one may use local search methods, like Rosen's gradient projection method [46], to minimize $f(x, t)$ over the feasible region $\mathcal{M}(t)$. If the initial point is feasible and close enough to a local solution and the step size is small enough, the algorithm will converge to the local minimum. This leads to the notion of region of attraction defined by resorting to the continuous-time model of Rosen's gradient projection method [6] (for which the step size is not important anymore).

Definition 2. The **region of attraction** of a local minimum point $h(t)$ of $f(\cdot, t)$ in the feasible set $\mathcal{M}(t)$ at a given time t is defined as:

$$RA^{\mathcal{M}(t)}(h(t)) = \left\{ x_0 \in \mathcal{M}(t) \mid \lim_{\tilde{t} \to \infty} \tilde{x}(\tilde{t}) = h(t) \quad \text{where} \right.$$

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

In the unconstrained case, the notion of the locally one-point strong convexity can be defined as follows:

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

$$\nabla_x f(e + h(t), t)^\top e \geq c \|e\|^2, \quad \forall e \in D, \quad \forall t \in [0, \infty) \quad (7)$$

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 $h(t)$.

This definition 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. The following result paves the way for the generalization of the notion of the locally one-point strong convexity from the unconstrained case to the equality constrained case.

Lemma 2. Consider an arbitrary local minimum trajectory $h(t)$ satisfying Assumption 4, there exist positive constants $\hat{r}$ and $\hat{c}$ such that

$$e(t)^\top \nabla_x L(e(t) + h(t), \lambda(e(t) + h(t), t), t) \geq \hat{c} \|e(t)\|^2$$

for all $e(t) \in \{e + h(t) \in \mathcal{M}(t) : \|e\| \leq \hat{r}\}$.

Proof. Due to the second-order sufficient conditions for the equality constrained minimization problem, $\nabla^2_{xx} L(h(t), \lambda(h(t), t), t)$ is positive definite on $T^t_{h(t)}$ for all $t \in [0, \infty)$, meaning that for every nonzero vector $y \in T^t_{h(t)}$, there exists a positive constant $\bar{c}$ such that $y \nabla^2_{xx} L(h(t), \lambda, t) y > \bar{c} \|y\|^2$. Since $\mathcal{P}(h(t), t)$ is the orthogonal projection matrix onto the tangent plane $T^t_{h(t)}$, we have $y \nabla^2_{xx} L(h(t), \lambda(h(t), t), t) \mathcal{P}(h(t), t) y > \bar{c} \|y\|^2$ for all $y \in T^t_{h(t)}$ and $y \neq 0$, and $y \nabla^2_{xx} L(h(t), \lambda(h(t), t), t) \mathcal{P}(h(t), t) y = 0$ for all $y \notin T^t_{h(t)}$. Taking the first-order Taylor expansion of $\nabla_x L(x, \lambda(x, t), t)$ with respect to x around $h(t)$ and using the following result from [47, Corollary 1]:

$$\frac{\partial}{\partial x} \nabla_x L(x, \lambda(x, t), t) \big|_{x=h(t)} = \nabla^2_{xx} L(h(t), \lambda(h(t), t), t)$$
$$\mathcal{P}(h(t), t),$$

it yields that

$$e(t)^\top \nabla_x L(e(t) + h(t), \lambda, t) = e(t)^\top \nabla_x L(h(t), \lambda, t)$$
$$+ e(t)^\top \nabla^2_{xx} L(h(t), \lambda, t) \mathcal{P}(h(t), t) e(t) + o(e(t)^3)$$
$$= e(t)^\top \nabla^2_{xx} L(h(t), \lambda, t) \mathcal{P}(h(t), t) e(t) + o(e(t)^3)$$

From Lemma 6 in Appendix A, we know that $\nabla^2_{xx} L(x, \lambda, t) \mathcal{P}(x, t)$ is continuous in x and t. In addition, $g(x, t)$ is also continuous in x and t. As a result, there exist positive constants $\hat{r}$ and $\hat{c}$ such that

$$e(t)^\top \nabla_x L(e(t) + h(t), \lambda, t) \geq \hat{c} \|e(t)\|^2$$

for all $e(t) \in \{e + h(t) \in \mathcal{M}(t) : \|e\| \leq \hat{r}\}$ $\qquad\square$

Definition 4. Consider arbitrary positive scalars c and r. The Lagrange function $L(x, \lambda, t)$ with λ given in (3) is said to be **locally (c, r)-one-point strongly convex** with respect to x around the local minimum trajectory $h(t)$ in the feasible set $\mathcal{M}(t)$ if:

$$e^\top \nabla_x L(e + h(t), \lambda(e + h(t), t), t) \geq c \|e\|^2 \quad (8)$$

for all $e \in D^{\mathcal{M}(t)}$ and $t \in [0, \infty)$, where $D^{\mathcal{M}(t)} = \{e + h(t) \in \mathcal{M}(t) : \|e\| \leq r\}$. The region $D^{\mathcal{M}(t)} = \{e + h(t) \in$

$\mathcal{M}(t) : \|e\| \leq r\}$ is called the region of locally (c,r)-one-point strong convexity of the Lagrange function $L(x,\lambda,t)$ around $h(t)$ in the feasible set $\mathcal{M}(t)$.

Remark 2. The Lagrange function $L(x,\lambda,t)$ with λ given in (3) being locally (c,r)-one-point strongly convex with respect to x around $h(t)$ is equivalent to the vector field $\mathcal{P}(x,t)\nabla_x f(x(t),t)$ being **locally (c,r)-one-point strongly monotone** with respect to x around $h(t)$.

B. Derivation of time-varying projected gradient flow system

In practice, one can only hope to sequentially solve the time-varying optimization problem (1) at some discrete time instances $0 = \tau_0 < \tau_1 < \tau_2 < \tau_3 < \ldots$ as follows:

$$\min_{x \in \mathbb{R}^n} \quad f(x,\tau_i), \quad \text{s.t.} \quad g(x,\tau_i) = 0, \quad i = 1,2,\ldots \quad (9)$$

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):

$$\min_{x \in \mathbb{R}^n} \quad f(x,\tau_0), \quad (10a)$$
$$\text{s.t.} \quad g(x,\tau_0) = 0,$$
$$\min_{x \in \mathbb{R}^n} \quad f(x,\tau_i) + \frac{\alpha}{2(\tau_i - \tau_{i-1})} \left\| x - x_{i-1}^* \right\|^2, \quad (10b)$$
$$\text{s.t.} \quad g(x,\tau_i) = 0, \quad i = 1,2,\ldots$$

where x_{i-1}^* denotes an arbitrary local minimum of the modified optimization problem (10) obtained using a local search method at time iteration $i-1$. A local optimal solution sequence $x_0^*, x_1^*, x_2^*, \ldots$ is said to be a **discrete local trajectory** of the sequential regularized optimization (10). The parameter α is called inertia because it acts as a resistance to changes x at time step τ_i with respect to x at the previous time step τ_{i-1}. Note that α could be time-varying (and adaptively changing) in the analysis of this paper, but we restrict our attention to a fixed regularization term to simplify the presentation.

Under Assumption 3, all solutions x^* of (10b) must satisfy the KKT conditions:

$$0 = \nabla_x f(x_i^*,\tau_i) + \alpha \frac{x_i^* - x_{i-1}^*}{\tau_i - \tau_{i-1}} + \mathcal{J}_g(x_i,\tau_i)^\top \bar{\lambda}_i, \quad (11a)$$

$$0 = g(x_i,\tau_i), \quad (11b)$$

where $\bar{\lambda}_i$'s are the Lagrange multipliers for the sequence of optimization problems with proximal regularization in (10). Similar to [22], we can write the right-hand side of the constraint (11b) as:

$$\frac{g(x_i,\tau_i) - g(x_i,\tau_{i-1}) + g(x_i,\tau_{i-1}) - g(x_{i-1},\tau_{i-1})}{\tau_i - \tau_{i-1}} \quad (12)$$

Since the function $f(x,t)$ and $g(x,t)$ are nonconvex in general, the problem (10) may not have a unique solution x_i^*.

In order to cope with this issue, we study the continuous-time limit of (11) as the time step $\tau_{i+1} - \tau_i$ diminishes to zero. This yields the following time-varying ordinary differential equations:

$$0 = \nabla_x f(x(t),t) + \alpha \dot{x}(t) + \mathcal{J}_g(x(t),t)^\top \bar{\lambda}(t), \quad (13a)$$
$$0 = \mathcal{J}_g(x(t),t)\dot{x}(t) + g'(x(t),t), \quad (13b)$$

where $g' = \frac{\partial g(x,t)}{\partial t}$ denotes the partial derivative of g with respect to t. Since $\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top$ is invertible, we have

$$0 = (\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1}\mathcal{J}_g(x(t),t)\nabla_x f(x(t),t)$$
$$- \alpha(\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1}g'(x(t),t) + \bar{\lambda}(t). \quad (14)$$

Therefore, $\bar{\lambda}(t)$ can be written as a function of x, t and α:

$$\bar{\lambda}(t) = -(\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1}\mathcal{J}_g(x(t),t)\nabla_x f(x(t),t)$$
$$+ \alpha(\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1}g'(x(t),t)$$
$$= \lambda(x(t),t) + \alpha(\mathcal{J}_g(x,t)\mathcal{J}_g(x,t)^\top)^{-1}g'(x,t) \quad (15)$$

We alternatively denote $\bar{\lambda}(t)$ as $\bar{\lambda}(x(t),t,\alpha)$. When $\alpha = 0$, we have $\bar{\lambda}(x(t),t,\alpha) = \lambda(x(t),t)$ and the differential equation (13) reduces to the algebraic equation (2), which is indeed the first-order KKT condition for the unregularized time-varying optimization (1). When $\alpha > 0$, substituting $\bar{\lambda}(x(t),t,\alpha)$ into (13a) yields the following time-varying ODE:

$$\dot{x}(t) = -\frac{1}{\alpha}\mathcal{P}(x(t),t)\nabla_x f(x(t),t) - \mathcal{Q}(x(t),t)g'(x(t),t),$$
$$\text{(P-ODE)}$$

where $\mathcal{Q}(x(t),t) = \mathcal{J}_g(x(t),t)^\top(\mathcal{J}_g(x(t),t)\mathcal{J}_g(x(t),t)^\top)^{-1}$. In terms of the Lagrange functional, (P-ODE) can be written as

$$\dot{x} = -\frac{1}{\alpha}\nabla_x L(x,\bar{\lambda},t) = -\frac{1}{\alpha}\nabla_x L(x,\lambda,t) - \mathcal{Q}(x,t)g'(x,t). \quad (16)$$

Here, $\nabla_x L(x,\bar{\lambda},t)$ means first taking the partial gradient with respect to the first argument and then using the formula (15) for $\bar{\lambda}$. It can be shown that if the initial point of (P-ODE) is in the feasible set $M(t_0)$, the solution of (P-ODE) will stay in the feasible set $M(t)$.

Lemma 3. Suppose that the solution $x(t,t_0,x_0)$ of (P-ODE) is defined in $[t_0,\infty)$ with the initial point x_0. If $x_0 \in \mathcal{M}(t_0)$, then the solution $x(t,t_0,x_0)$ belongs to $\mathcal{M}(t)$ for all $t \geq t_0$.

Proof. On examining the evolution of $g(x(t),t)$ along the flow of the system (P-ODE), we obtain

$$\dot{g}(x(t),t) = \mathcal{J}_g(x(t),t)\dot{x}(t) + g'(x(t),t) = 0$$

Hence, $g(x(t_0),t_0) = g(x(t,t_0,x_0),t)$ for all $t \geq t_0$. □

Therefore, as long as the initial point of (P-ODE) is in the feasible set $M(t_0)$, the above lemma guarantees that we can analyze the stability of (P-ODE) using the standard Lyapunov's theorem without worrying about the feasibility of the solution. When $\alpha > 0$, we will show that for any initial point x_0, (P-ODE) has a unique solution defined for all $t \in I_t \subseteq [0,\infty)$ if there exists a

local minimum trajectory $h(t)$ such that the solutions of (P-ODE) lie in a compact set around $h(t)$ [1].

Theorem 1 (Existence and uniqueness)**.** Under Assumptions 1-4 and given any initial point $x_0 \in \mathcal{M}(t_0)$, suppose that there exists a local minimum trajectory $h(t)$ with the property that $x(t) - h(t)$ lies entirely in D for all $t \in I_t \subseteq [0, \infty)$ where D is a compact subset of $\mathbb{R}^n$ containing $x_0 - h(t_0)$ and $x(t)$ denotes the solution of (P-ODE) with the initial point x_0. Then, (P-ODE) has a unique solution starting from x_0 that is defined for all $t \geq 0$.

Proof. Since $h(t)$ is differentiable by Lemma 1, we can use the change of variables $e(t) = x(t) - h(t)$ to rewrite (P-ODE) as:

$$\dot{e}(t) = -\frac{1}{\alpha}\mathcal{P}(e(t)+h(t),t)\nabla_x f(e(t)+h(t),t)-$$
$$\mathcal{Q}(e(t)+h(t),t)g'(e(t)+h(t),t) - \dot{h}(t) \qquad (17)$$

In light of the conditions in Theorem 1, the solution of (17) stays in a compact set. Then, by Lemma 3 and [36, Theorem 3.3], the equation (17) has a unique solution. Thus, (P-ODE) must also have a unique solution. $\qquad \square$

In online optimization, it is sometimes 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 to obtain a numerical approximation to the solution of (P-ODE):

$$\bar{x}_i^* = \bar{x}_{i-1}^* - (\tau_i - \tau_{i-1})\Big(\frac{1}{\alpha}\mathcal{P}(\bar{x}_{i-1}^*,\tau_{i-1})\nabla_x f(\bar{x}_{i-1}^*,\tau_{i-1})$$
$$+ \mathcal{Q}(\bar{x}_{i-1}^*,\tau_{i-1})g'(\bar{x}_{i-1}^*,\tau_{i-1})\Big) \qquad (18)$$

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

Theorem 2 (Convergence)**.** Under Assumptions 1-4 and given a local minimum x_0^* of (10a), 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 (P-ODE) in the sense that for all fixed $T > 0$:

$$\lim_{\Delta_\tau \to 0} \max_{0 \leq k \leq \frac{T}{\Delta_\tau}} \left\| x_k^\Delta - x(\tau_k, \tau_0, x_0^*) \right\| = 0 \qquad (19)$$

and any sequence of $(\bar{x}_k^\Delta)$ updated by (18) converges to the (P-ODE) in the sense that for all fixed $T > 0$:

$$\lim_{\Delta_\tau \to 0} \max_{0 \leq k \leq \frac{T}{\Delta_\tau}} \left\| \bar{x}_k^\Delta - x(\tau_k, \tau_0, x_0^*) \right\| = 0 \qquad (20)$$

Proof. The first part follows from Theorem 2 in [23]. For the second part, a direct application of the classical

results on convergence of the forward Euler method [48] immediately shows that the solution of (P-ODE) starting at a local minimum of (10a) is the continuous limit of the discrete local trajectory of the sequential regularized optimization (10). $\qquad \square$

Theorem 2 guarantees that the solution of (P-ODE) is a reasonable approximation in the sense that it is the continuous-time limit of both the solution of the sequential regularized optimization problem (10) and the solution of the online updating scheme (18). For this reason, we only study the continuous-time problem (P-ODE) in the remainder of this paper.

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 $h_2(t)$. We provide the definitions of jumping, tracking and escaping below.

Definition 5. It is said that the solution of (P-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

$$\mathcal{B}_v(h_1(t_1)) \cap \mathcal{M}(t_1) \subseteq RA^{\mathcal{M}(t_1)}(h_1(t_1)) \qquad (21a)$$
$$\mathcal{B}_u(h_2(t_2)) \cap \mathcal{M}(t_2) \subseteq RA^{\mathcal{M}(t_2)}(h_2(t_2)) \qquad (21b)$$
$$\forall x_1 \in \mathcal{B}_v(h_1(t_1)) \cap \mathcal{M}(t_1)$$
$$\implies x(t_2,t_1,x_1) \in \mathcal{B}_u(h_2(t_2)) \cap \mathcal{M}(t_2) \qquad (21c)$$

Definition 6. Given $x_0 \in \mathcal{M}(t_0)$, 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

$$x(t,t_0,x_0) \in \mathcal{B}_u(h_2(t)) \cap \mathcal{M}(t), \qquad \forall t \geq T \qquad (22a)$$
$$\mathcal{B}_u(h_2(t)) \cap \mathcal{M}(t) \subseteq RA^{\mathcal{M}(t)}(h_2(t)), \quad \forall t \geq T \qquad (22b)$$

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 7. 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

$$\mathcal{B}_v(h_1(t_0)) \cap \mathcal{M}(t_0) \subseteq RA^{\mathcal{M}(t_0)}(h_1(t_0)) \qquad (23a)$$
$$\mathcal{B}_u(h_2(t)) \cap \mathcal{M}(t) \subseteq RA^{\mathcal{M}(t)}(h_2(t)), \ \forall t \geq T \qquad (23b)$$
$$\forall x_0 \in \mathcal{B}_v(h_1(t_0)) \cap \mathcal{M}(t_0) \implies$$
$$x(t,t_0,x_0) \in \mathcal{B}_u(h_2(t)) \cap \mathcal{M}(t), \ \forall t \geq T \qquad (23c)$$

Figure 3 illustrates the definitions of jumping and tracking for Example 1 with $\alpha = 0.3$ and $b = 5$. The objective of this paper is to study when the solution of (P-ODE) started at a poor local minimum at the initial time jumps to and tracks a better (or global) minimum of the problem after some time. In other words, it is desirable to investigate the escaping property from $h_1(t)$ and $h_2(t)$.
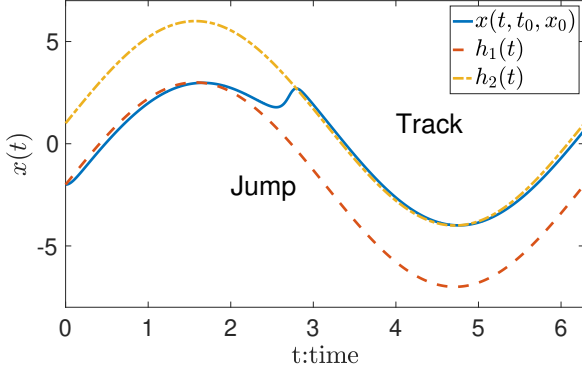
---

[1] In Theorems 3 and 4, the compactness assumption is included in the definition of the dominant trajectory. In Theorem 5, checking the compactness assumption can be carried out via the Lyapunov's method without solving the differential equation due to the one-point strong convexity condition around h.

Fig. 3. Illustration of jumping and tracking.

III. CHANGE OF VARIABLES

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

$$\dot{e}(t) = -\frac{1}{\alpha}\mathcal{P}(e(t) + h_2(t), t)\nabla_x f(e(t) + h_2(t), t) - $$
$$\mathcal{Q}(e(t) + h_2(t), t)g'(e(t) + h_2(t), t) - \dot{h}_2(t) \quad (24a)$$
$$= -\frac{1}{\alpha}\nabla_x\Big(L\big(e(t) + h_2(t), \bar{\lambda}(e(t) + h_2(t), t, \alpha), t\big)$$
$$+ \alpha\dot{h}_2(t)^\top e(t)\Big) \quad (24b)$$

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 $-\frac{1}{\alpha}U(e(t), t, \alpha)$ to denote the right-hand side of (24). Note that $h_1(t)$ and $h_2(t)$ are local solutions of (1) and as long as (1) is time-varying, these functions cannot satisfy (P-ODE) in general. We denote $\mathcal{M}^h(t) := \{e \in \mathbb{R}^n : g(e + h(t), t) = 0\}$.

A. Unconstrained optimization landscape after a change of variables

In this subsection, we study the unconstrained case to enable a better visualization of the optimization landscape. In the unconstrained case, (24) is reduced to

$$\dot{e}(t) = -\frac{1}{\alpha}\nabla_x f(e(t) + h_2(t), t) - \dot{h}_2(t). \quad (25)$$

1) Inertia encouraging the exploration: The first term $\nabla_x f(e + h_2(t), t)$ in (25) can be understood as a time-varying gradient term that encourages the solution of (25) to track $h_2(t)$, while the second term $\dot{h}_2(t)$ represents the inertia from this trajectory. In particular, if $\dot{h}_2(t)$ points toward outside of the region of attraction of $h_2(t)$ during some time interval, the term $\dot{h}_2(t)$ acts as an **exploration** term that encourages the solution of (ODE) to leave the region of attraction of $h_2(t)$. The parameter α balances the roles of the gradient and the inertia.

In the extreme case where α goes to infinity, $e(t)$ converges to $-h_2(t)$ and $x(t)$ approaches a constant trajectory determined by the initial point x_0; when α is sufficiently

small, the time-varying gradient term dominates the inertia term and the solution of (ODE) would track $h_2(t)$ closely. With an appropriate proximal regularization α that keeps the balance between the time-varying gradient term and the inertia term, the solution of (ODE) could temporarily track a local minimum trajectory with the potential of exploring other local minimum trajectories.

2) Inertia creating a one-point strongly convex landscape: The differential equation (25) can be written as

$$\dot{e}(t) = -\frac{1}{\alpha}\nabla_e\Big(f(e(t) + h_2(t), t) + \alpha\dot{h}_2(t)^\top e(t)\Big) \quad (26)$$

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 the time-varying objective function $f(e + h_2(t), t) + \alpha\dot{h}_2(t)^\top e$ over a neighborhood of $h_1(t)$ to become 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 (26) starting in a neighborhood of $h_1(t)$ to move towards a neighborhood of $h_2(t)$. Before analyzing this phenomenon, we illustrate the concept in an example.

Consider again Example 1 and recall that $\bar{f}(x)$ has 2 local minima at $x = -2$ and $x = 1$. By taking $b = 5$, $h_1(t) = -2 + 5\sin(t)$ and $h_2(t) = 1 + 5\sin(t)$, the differential equation (26) can be expressed as $\dot{e}(t) = -\frac{1}{\alpha}\nabla_e\Big(\bar{f}(1 + e(t)) + 5\alpha\cos(t)e(t)\Big)$. The landscape of the new time-varying function $\bar{f}(1 + e) + 5\alpha\cos(t)e$ 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 (26) 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 (26) 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 two locally one-point strongly convex regions around $h_1(t)$ and $h_2(t)$ and, therefore, (26) fails to escape the region around $h_1(t)$.

To further inspect the case $\alpha = 0.3$, observe in Figure 5a that the landscape of the objective function $\bar{f}(1 + e) + 1.5\cos(0.85\pi)e$ 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 $t = 0.85\pi$. This is consistent with the fact that the solution of $\dot{e} = -\frac{1}{0.3}\nabla_x\bar{f}(1 + e) - 5\cos(t)$ starting from $e = -3$ jumps to the neighborhood of 0 around time $t = 0.85\pi$, as demonstrated in Figure 5c. Furthermore, if the time interval $[t_1, t_2]$ is large enough to allow transitioning from a neighborhood of $h_1(t)$ to a neighborhood of $h_2(t)$, then the solution of (26) would move to the neighborhood of $h_2(t)$. In contrast, the region around $1 + b\sin(t)$ is never one-point strongly convexified with respect to $-2 + b\sin(t)$, as shown in Figure 5b.

From the right-hand side of (26), it can be inferred that if the gradient of $f(\cdot, t)$ is relatively small around
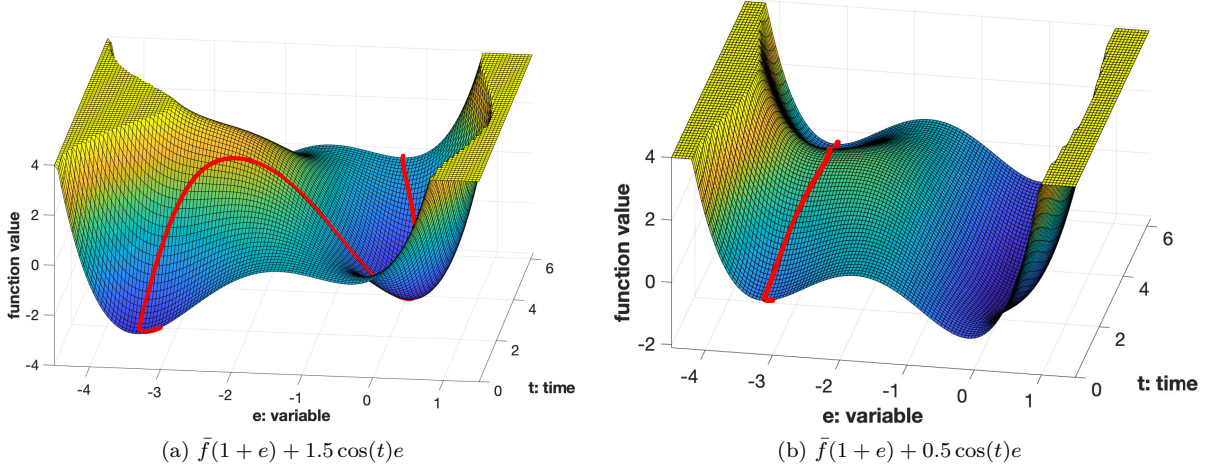
(a) $\bar{f}(1+e) + 1.5\cos(t)e$

(b) $\bar{f}(1+e) + 0.5\cos(t)e$

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

(a) $\bar{f}(1+e) + 1.5\cos(0.85\pi)e$

(b) $\bar{f}(1+e) + 1.5\cos(0)e$

(c) solution of $\dot{e} = -\frac{1}{0.3}\nabla_x\bar{f}(1+e) - 5\cos(t)$ starting from $e_0 = -3$
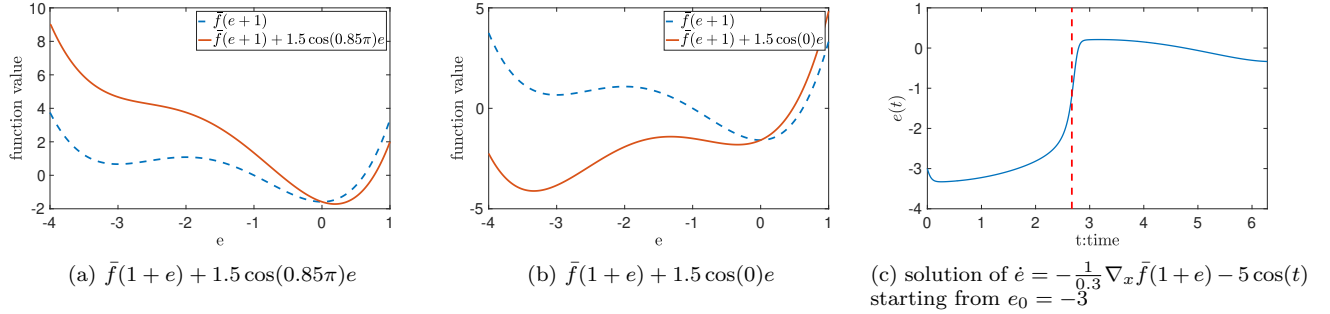
Fig. 5. Illustration of one-point strong convexification for Example 1.

TABLE I
A UNIFIED VIEW FOR UNCONSTRAINED AND EQUALITY-CONSTRAINED PROBLEMS

	Unconstrained problem	Equality-constrained problem
First-order optimality condition(FOC)	$0 = \nabla_x f(x,t)$	$0 = \nabla_x L(x,\lambda,t)$
ODE (continuous time limit of FOC for regularized problem)	$\dot{x} = -\frac{1}{\alpha}\nabla_x f(x,t)$	$\dot{x} = -\frac{1}{\alpha}\nabla_x L(x,\bar{\lambda},t)$
Change of variables: $x = h + e$	$\dot{e} = -\frac{1}{\alpha}\nabla_e f(e+h,t) - \dot{h}$	$\dot{e} = -\frac{1}{\alpha}\nabla_e L(e+h,\bar{\lambda},t) - \dot{h}$
Key assumption: one-point strong convexity	$e^\top \nabla_e f(e+h,t) \geq c\,\|e\|^2$	$e^\top \nabla_e L(e+h,\lambda,t) \geq c\,\|e\|^2$
Reshaping of the landscape: one-point strong convexification	$e^\top\left(\nabla_e f(e+h,t) + \alpha\dot{h}\right) \geq w\,\|e\|^2$	$e^\top\left(\nabla_e L(e+h,\bar{\lambda},t) + \alpha\dot{h}\right) \geq w\,\|e\|^2$

some local minimum trajectory, then its landscape is easier to be re-shaped by the time-varying linear perturbation $\alpha\dot{h}_2(t)^\top e$. The local minimum trajectory in a neighborhood with small gradients usually corresponds to a shallow minimum trajectory in which the trajectory has a relatively flat landscape and a relatively small region of attraction. Thus, the one-point strong convexication introduced by the time-varying perturbation could help escape the shallow minimum trajectories.

B. Dominant trajectory

In this subsection, we will formalize the intuitions discussed in Section III-A. We first define the notion of the shallow local minimum trajectory.

Definition 8. Consider a positive number α and assume that $\dot{h}_1(t)$ is L-Lipschitz continuous. It is said that the local minimum trajectory $h_1(t)$ is α-**shallow** during the time period $[t_0, t_0 + \delta]$ if $\epsilon > E(\alpha) + L\delta$ and $r \leq \frac{1}{2}\delta(\epsilon - E(\alpha) - L\delta)$, where $\epsilon = \sup_{t\in[t_0,t_0+\delta]}\left\|\dot{h}_1(t)\right\|$, $r = \sup_{t\in[t_0,t_0+\delta]}\sup_{x(t)\in RA^{\mathcal{M}(t)}(h_1(t))}\|x(t) - h_1(t)\|$, $E(\alpha) = \sup_{t\in[t_0,t_0+\delta]}\sup_{x(t)\in RA^{\mathcal{M}(t)}(h_1(t))}\left\|\frac{1}{\alpha}\nabla_x L(x,\bar{\lambda},t)\right\|$

and $\frac{1}{\alpha}\nabla_x L(x,\bar{\lambda},t)$ is defined in (16).

In other words, a local minimum trajectory is shallow if it has a large time variation but a small region of attraction. We next show that whenever a local minimum trajectory $h_1(t)$ is shallow during some time interval, the solution of (P-ODE) starting anywhere in the region of attraction of $h_1(t)$ will leave its region of attraction at some time.

Lemma 4. If the local minimum trajectory $h_1(t)$ is α-shallow during $[t_0, t_0 + \delta]$, then for any $x(t_0) \in RA^{\mathcal{M}(t_0)}(h_1(t_0))$, then there exists a time $t \in [t_0, t_0 + \delta]$ such that $x(t) \notin RA^{\mathcal{M}(t)}(h_1(t))$.

Proof. Let $b(t_0)$ be the unit vector $-\frac{\dot{h}_1(t_0)}{\|\dot{h}_1(t_0)\|}$. One can write

$$-\dot{h}_1(t)^\top b(t_0) \geq -\dot{h}_1(t_0)^\top b(t_0) - L|t - t_0| \geq \epsilon - L\delta := \epsilon'$$

For any $t \in [t_0, t_0 + \delta]$ and $e(t) \in RA^{\mathcal{M}(t)}(h_1(t))$, we have

$$(\dot{x}(t) - \dot{h}_1(t))^\top b(t_0) = -\frac{1}{\alpha}\nabla_x L(x,\bar{\lambda},t)^\top b(t_0) - \dot{h}_1(t)^\top b(t_0)$$
$$\geq \epsilon' - \left\|\frac{1}{\alpha}\nabla_x L(x,\bar{\lambda},t)\right\| \geq \epsilon' - E$$

Hence,

$$r \geq \|x(t_0 + \delta) - h_1(t_0 + \delta)\|$$
$$\geq (x(t_0 + \delta) - h_1(t_0 + \delta))^\top b(t_0)$$
$$\geq (x(t_0) - h_1(t_0))^\top b(t_0) + \int_{t_0}^{t_0+\delta} (\epsilon' - E)dt$$
$$\geq -r + (\epsilon' - E)\delta$$

The above contradiction completes the proof. $\qquad\square$

On the one hand, Lemma 4 shows that any shallow local minimum trajectory is unstable in the sense that the time-variation in the minimum trajectory will force the solution of (P-ODE) to leave its region of attraction. If the shallow local minimum trajectory happens to be a non-global local solution, then the solution of (P-ODE), acting as a tracking algorithm, will help avoid the bad local solutions for free. On the other hand, Lemma 4 does not specify where the solution of (P-ODE) will end up after leaving the region of attraction of a shallow local minimum trajectory. Simulations (such as those provided in Sections III-A and V) suggest that, with some appropriate α, the solution of (P-ODE) may move towards a nearby local minimum trajectory that has an enlarged region of one-point strong convexity. This leads to the following definition of the region of the domination and the dominant local minimum trajectory.

Definition 9. Given two local minimum trajectories $h_1(t)$ and $h_2(t)$, suppose that the time-varying Lagrange function $L(x,\lambda,t)$ with λ given in (3) is locally (c_2, r_2)-one-point strongly convex with respect to x around $h_2(t)$ in the region $\mathcal{M}^{h_2}(t) \cap \mathcal{B}_{r_2}(0)$. A set D_{v,ρ,r_2} is said to be **the region of domination** for $h_2(t)$ with respect to $h_1(t)$ if it satisfies the following properties:

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

$$e_1 \in D_{v,\rho,r_2} \Rightarrow e(t, t_1, e_1) \in D_{v,\rho,r_2}, \forall t \in [t_1, t_2] \quad (27)$$

  where $e(t, t_1, e_1)$ is the solution of (24) staring from the feasible initial point $e_1 \in \mathcal{M}^{h_2}(t_1)$ at the initial time t_1.

- $D_{v,\rho,r_2} \supseteq D'_v \cup \mathcal{B}_\rho(0)$ where

$$D'_v = \{e_1 \in \mathbb{R}^n : e_1 + h_2(t_1) \in \mathcal{M}(t_1) \cap \mathcal{B}_v(h_1(t_1))$$
$$\subseteq RA^{\mathcal{M}(t_1)}(h_1(t_1))\},$$
$$\rho \geq \sup_{\substack{t \in [t_1,t_2]}} \sup_{\substack{\bar{e}(t):\|\bar{e}(t)\|<r_2, \\ 0 = U(\bar{e}(t),t,\alpha)}} \|\bar{e}(t)\|. \quad (28)$$

The condition (27) is a set invariance property, which requires that the solution of (24) starting from an initial point in D_{v,ρ,r_2} stays in D_{v,ρ,r_2} during the time period $[t_1, t_2]$. For the visualization of D_{v,ρ,r_2}, $\mathcal{B}_\rho$ and D'_v in Definition 9, we consider again Example 1. In Fig 6, the red curve corresponds to the landscape of the function $\bar{f}(1 + e) + 1.5\cos(0.85\pi)e$, $e = 0$ corresponds to $h_2(t)$ and $e = -3$ corresponds to $h_1(t)$. $\mathcal{B}_\rho$ is a region around $h_2(t)$ containing all zeros of $0 = U(\cdot, t, \alpha)$ during a time period around 0.85π and D'_v is a neighborhood around $h_1(t)$. In this example, the region of domination for $h_2(t)$ with respect to $h_1(t)$ is $D_{v,\rho,r_2} = [-4, 1]$ which contains $\mathcal{B}_\rho$ and D'_v if $h_1(t)$ if it also satisfies (27).
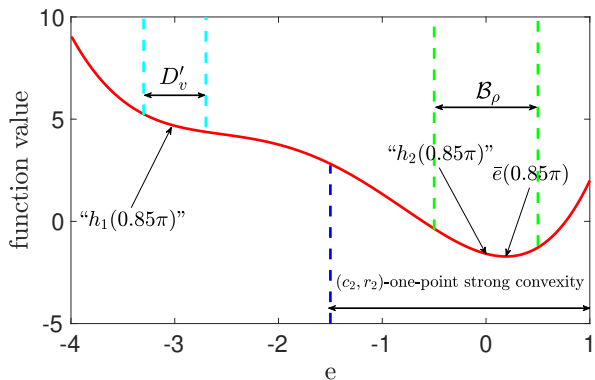

Fig. 6. Illustration of Definition 9: the region of domination.

Definition 10. It is said that $h_2(t)$ is a (α, w)-**dominant trajectory** with respect to $h_1(t)$ during the time period $[t_1, t_2]$ over the region D_{v,ρ,r_2} if the time variation of $h_2(t)$ makes the time-varying function $U(e(t), t, \alpha)$ become one-point strongly monotone over D_{v,ρ,r_2}, i.e.,

$$U(e(t),t,\alpha)^\top \left(e(t) - \bar{e}(t)\right) \geq w\|e(t) - \bar{e}(t)\|^2,$$
$$\forall e(t) \in D_{v,\rho,r_2} \cap \mathcal{M}(t), t \in [t_1, t_2], \quad (29)$$

where $w > 0$ is a constant and $\bar{e}(t)$ is defined in (28).

Note that $h_2(t)$ being a dominant trajectory with respect to $h_1(t)$ is equivalent to the statement that the inertia of $h_2(t)$ creates a strongly convex landscape over D_{v,ρ,r_2}, as discussed in Section III-A.

Remark 3. The intuition behind Definition 10 is that if the time variation in the time-varying optimization could make the landscape after the change of variables become one-point strongly convex with respect to the minimum trajectory $h_2(t)$ in a neighborhood including both trajectories $h_1(t)$ and $h_2(t)$, then minimum trajectory $h_2(t)$ is dominant (with respect to minimum trajectory $h_1(t)$). This definition only depends on the landscape of the time-varying objective and is independent of the algorithm.

C. The role of temporal variations of the constraints

From the perspective of the landscape of the Lagrange functional, (24b) can be regarded as a time-varying gradient flow system of the Lagrange functional $L\big(e(t) + h_2(t), \bar{\lambda}(e(t) + h_2(t), t, \alpha), t\big)$ (the partial gradient is taken with respect to the first argument of L) plus a linear time-varying perturbation $\alpha \dot{h}_2^g(t)^\top e(t)$. Besides the linear time-varying perturbation $\alpha \dot{h}_2^g(t)^\top e(t)$ induced by the inertia of the minimum trajectory similar to the unconstrained case, the constraints' temporal variation $g'(\cdot, t)$ plays the role of shifting the Lagrange multiplier from λ in (3) to $\bar{\lambda}$ in (15), which results in a nonlinear time-varying perturbation of the landscape of the Lagrange functional.

From the perspective of the perturbed gradient, the constraints' temporal variation $g'(\cdot, t)$ perturbs the projected gradient $\mathcal{P}(\cdot, t) \nabla_x f(\cdot, t)$ in an orthogonal direction $\mathcal{Q}(\cdot, t) g'(\cdot, t)$ to drive the trajectory of (24a) towards satisfying the time-varying constraints.

Lemma 5. At any given time t, the vector $\mathcal{P}(x, t) \nabla_x f(x, t)$ is orthogonal to the vector $\mathcal{Q}(x, t) g'(x, t)$.

Proof. Recall that $\mathcal{P}(x, t)$ is the orthogonal projection matrix on the tangent plane of $g(x(t), t)$ at the point $x(t)$ after the freezing time t. Thus, we have $\mathcal{P}(x, t) \nabla_x f(x, t) \in T_x^t$. For the vector $\mathcal{Q}(x, t) g'(x, t)$, it can be shown that

$$\mathcal{P}(x, t) \mathcal{Q}(x, t) g'(x, t) = 0$$

This implies that the orthogonal projection of the vector $\mathcal{Q}(x, t) g'(x, t)$ onto the tangent plane T_x^t is 0. Thus, $\mathcal{Q}(x, t) g'(x, t)$ must be orthogonal to T_x^t. $\qquad\square$

Therefore, in the equality-constrained problem, the time-varying projected gradient flow system after a change of variables in (24a) can be regarded as a composition of a time-varying projected term $\mathcal{P}(e + h_2(t), t) \nabla_x f(e + h_2(t), t)$, a time-varying constraint-driven term $\mathcal{Q}(e + h_2(t), t) g'(e + h_2(t), t)$ and an inertia term $\dot{h}_2(t)$ due to the time variation of the local minimum trajectory.

D. A unified view for unconstrained and equality-constrained problems

By introducing the Lagrange functional in (5) and (16), we can unify the analysis of how the temporal variation and the proximal regularization help reshape the optimization landscape and potentially make the landscape become one-point strongly convex over a larger region, for both unconstrained and equality constrained problems. This unified view is illustrated in Table I.

IV. MAIN RESULTS

In this section, we study the jumping, tracking and escaping properties for the time-varying nonconvex optimization.

A. Jumping

The following theorem shows that the solution of (P-ODE) could jump to the dominant trajectory as long as the time-interval of such domination is large enough.

Theorem 3 (Sufficient conditions for jumping from $h_1(t)$ to $h_2(t)$)**.** Suppose that the local minimum trajectory $h_2(t)$ is a (α, w)-dominant trajectory with respect to $h_1(t)$ during $[t_1, t_2]$ over the region D_{v, ρ, r_2}. Let $e_1 \in D_v'$ be the initial point of (24), and consider $\bar{e}(t)$ defined in (28). Assume that $U(e, t, \alpha)$ is non-singular for all $t \in [t_1, t_2]$ and $e \in D_{v, \rho, r_2}$ and there exists a constant $\theta \in (0, 1)$ such that

$$t_2 - t_1 \geq \max \left\{ \frac{\alpha \rho}{(r_2 - \rho) \theta w}, \frac{\alpha \ln \left(\frac{\|e_1 - \bar{e}(t_1)\|}{r_2 - \rho} \right)}{(1 - \theta) w} \right\}. \quad (30)$$

Then, the solution of (P-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 if $U(e, t, \alpha)$ is uniformly non-singular for all $t \in [t_1, t_2]$ and $e \in D_{v, \rho, r_2}$, then $\bar{e}(t)$ defined in (28) is continuously differentiable for $t \in [t_1, t_2]$. Then, notice that every solution of (24) with an initial point in $D_{v, \rho, r_2} \cap \mathcal{M}(t_1)$ will remain in D_{v, ρ, r_2}. It follows from Theorem 1 that (24) has a unique solution defined for all $t \in [t_1, t_2]$ whenever $e_1 \in D_{v, \rho, r_2} \cap \mathcal{M}(t_1)$.

We take $V(e(t), t) = \frac{1}{2} \|e(t) - \bar{e}(t)\|^2$ as the Lyapunov function for the system (24). Because of Lemma 3, any solution of (24) stating in $\mathcal{M}(t_1)$ will remain in $\mathcal{M}(t)$ for all $t \geq t_1$. Therefore, the derivative of $V(e(t), t)$ along the trajectories of (24) in $\mathcal{M}(t)$ can be expressed as

$$\begin{aligned}
\dot{V} =& (e(t) - \bar{e}(t))^\top \left(-\frac{1}{\alpha} U(e(t), t, \alpha) \right) - \\
& (e(t) - \bar{e}(t))^\top \dot{\bar{e}}(t), \ \forall e(t) \in D_{v, \rho, r_2} \cap \mathcal{M}^{h_2}(t) \\
\leq& -\frac{w}{\alpha} \|e(t) - \bar{e}(t)\|^2 + \|\dot{\bar{e}}(t)\| \|e(t) - \bar{e}(t)\|, \\
& \forall e(t) \in D_{v, \rho, r_2} \cap \mathcal{M}^{h_2}(t) \\
\leq& -(1 - \theta) \frac{w}{\alpha} \|e(t) - \bar{e}(t)\|^2 - \theta \frac{w}{\alpha} \|e(t) - \bar{e}(t)\|^2 \\
& + \delta \|e(t) - \bar{e}(t)\|, \forall e(t) \in D_{v, \rho, r_2} \cap \mathcal{M}^{h_2}(t) \\
\leq& -(1 - \theta) \frac{w}{\alpha} \|e(t) - \bar{e}(t)\|^2, \forall e(t) \in \\
& \{ e(t) \in D_{v, \rho, r_2} \cap \mathcal{M}^{h_2}(t) : \|e(t) - \bar{e}(t)\| \geq \frac{\alpha \delta}{\theta w} \} \quad (31)
\end{aligned}$$

where $\delta := \sup_{t \in [t_1, t_2]} \|\dot{\bar{e}}(t)\|$. By taking $e_1 \in D_v' \cap \mathcal{M}(t_1)$, since D_{v, ρ, r_2} satisfies the condition (27), the solution of (24) starting from e_1 will stay in D_{v, ρ, r_2}. Thus, the bound in (31) is valid. To ensure that the trajectory of (24) enters the time-varying set $\mathcal{B}_{r_2 - \rho}(\bar{e}(t))$, it is sufficient to have $\frac{\alpha \delta}{\theta w} \leq r_2 - \rho$ or $\alpha \leq \frac{(r_2 - \rho) \theta w}{\delta}$. Since $\delta = \sup_{t \in [t_1, t_2]} \|\dot{\bar{e}}(t)\| \geq$

$\frac{\rho}{t_2-t_1}$. We can further bound α as $\alpha \leq \frac{(r_2-\rho)\theta w(t_2-t_1)}{\rho}$ which is equivalent to $t_2 - t_1 \geq \frac{\alpha\rho}{(r_2-\rho)\theta w}$.

Now, it is desirable to show that if the time interval $[t_1, t_2]$ is large enough, the solution of (24a) will enter the time-varying set $\mathcal{B}_{r_2-\rho}(\bar{e}(t))$ with an exponential convergence rate. Since $\dot{V}(\cdot, \cdot)$ is negative in $\Gamma(t) := \{e \in D_{v,\rho,r_2} \cap \mathcal{M}^{h_2}(t) : \|e - \bar{e}(t)\| \geq \frac{\alpha\delta}{\theta w}\}$ and because of (27), a trajectory starting from $\Gamma(t_1)$ must stay in D_{v,ρ,r_2} and move in a direction of decreasing $V(e, t)$. The function $V(e, t)$ will continue decreasing until the trajectory enters the set $\{e \in D_{v,\rho,r_2} \cap \mathcal{M}^{h_2}(t) : \|e - \bar{e}(t)\| \leq \frac{\alpha\delta}{\theta w}\}$ or until time t_2. Let us show that the trajectory enters $\mathcal{B}_{r_2-\rho}(\bar{e}(t))$ before t_2 if $t_2 - t_1 > \frac{\alpha}{w(1-\theta)} \ln(\frac{\|e_1-\bar{e}(t_1)\|}{r_2-\rho})$. Since $V(e(t), t) = \frac{1}{2} \|e(t) - \bar{e}(t)\|^2$, (31) can be written as

$$\dot{V}(e(t), t) \leq -(1-\theta)\frac{2w}{\alpha}V(e(t), t),$$
$$\forall e \in \left\{e \in D_{v,\rho,r_2} \cap \mathcal{M}^{h_2}(t) : \|e(t) - \bar{e}(t)\| \geq \frac{\alpha\delta}{\theta w}\right\},$$

By the comparison lemma [36, Lemma 3.4],

$$V(e(t), t) \leq \exp\{-(1-\theta)\frac{2w}{\alpha}(t - t_1)\}V(e_1, t_1)$$

Hence,

$$\|e(t) - \bar{e}(t)\| \leq \exp\{-(1-\theta)\frac{w}{\alpha}(t - t_1)\} \|e_1 - \bar{e}(t_1)\|.$$

The inequality $\|e(t_2) - \bar{e}(t_2)\| \leq r_2 - \rho$ holds if $t_2 - t_1 \geq \frac{\alpha}{w(1-\theta)} \ln(\frac{\|e_1-\bar{e}(t_1)\|}{r_2-\rho})$. $\square$

We also offer an approach based on the time-averaged dynamics over a small time interval and name it "small interval averaging"[2]. This technique guarantees that the solution of the time-varying differential equation (or system) will converge to a residual set of the origin of (25), provided that: (i) there is a time interval $[t_1, t_2]$ such that the temporal variation makes the averaged objective function during this interval locally one-point strongly convex around $h_2(t)$ not only just over a neighborhood of $h_2(t)$ but also over a neighborhood of $h_1(t)$, (ii) the original time-varying system is not too distant from the time-invariant averaged system, (iii) $[t_1, t_2]$ is large enough to allow the transition of points from a neighborhood of $h_1(t)$ to a neighborhood of $h_2(t)$. Therefore, the time interval $[t_1, t_2]$ and the time-averaged dynamics over this time interval serve as a certificate for jumping from $h_1(t)$ to $h_2(t)$. In what follows, we introduce the notion of averaging a time-varying function over a time interval $[t_1, t_2]$.

Definition 11. A function $U_{\mathrm{av}}(e, \alpha)$ is said to be the **average function** of $U(e, t, \alpha)$ over the time interval $[t_1, t_2]$ if

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

[2]Our averaging approach distinguishes from classic averaging methods [36], [37], [49], [50] and the partial averaging method [51] in the sense that: (1) it is averaged over a small time interval instead of the entire time horizon, and (2) there is no two-time-scale behavior because there is no parameter in (25) that can be taken sufficiently small.

The averaged system of (24) over the time interval $[t_1, t_2]$ can be written as

$$\dot{e} = -\frac{1}{\alpha}U_{\mathrm{av}}(e, \alpha) \qquad (32)$$

Then, (24) can be regarded as a time-invariant system (32) with the time-varying perturbation term $p(e(t), t, \alpha) = -\frac{1}{\alpha}(U(e(t), t, \alpha) - U_{\mathrm{av}}(e(t), \alpha))$. For the averaged system, we can define the on-average region of domination D_{v,ρ,r_2} for $h_2(t)$ with respect to $h_1(t)$ similarly as Definition 9 by replacing (28) with

$$\rho \geq \sup_{\bar{e}: \|\bar{e}\| < r_2, 0 = U_{\mathrm{av}}(\bar{e}, \alpha)} \|\bar{e}\|. \qquad (33)$$

The corresponding on-average (α, w)-dominant trajectory with respect to $h_1(t)$ during $[t_1, t_2]$ over the region D_{v,ρ,r_2} can also be defined similarly as Definition 10 by replacing (29) with

$$U_{\mathrm{av}}(e, \alpha)^\top (e - \bar{e}) \geq w \|e - \bar{e}\|^2, \qquad (34)$$
$$\forall e \in D_{v,\rho,r_2} \cup \left(\cup_{[t_1,t_2]}\mathcal{M}(t)\right)$$

where $\bar{e}$ is defined in (33).

Theorem 4 (Sufficient conditions for jumping from $h_1(t)$ to $h_2(t)$ using averaging)**.** Suppose that the local minimum trajectory $h_2(t)$ is a on-average (α, w)-dominant trajectory 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:

1) There exist some time-varying scalar functions $\delta_1(\alpha, t)$ and $\delta_2(\alpha, t)$ such that

$$\|p(e(t), t, \alpha)\| \leq \delta_1(\alpha, t) \|e - \bar{e}\| + \delta_2(\alpha, t), \qquad (35)$$

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

$$\int_{t_1}^{t} \delta_1(\alpha, \tau)d\tau \leq \eta_1(\alpha)(t - t_1) + \eta_2(\alpha). \qquad (36)$$

2) The inequality

$$\beta_2(\alpha) \|e_1 - \bar{e}\| e^{-\beta_1(\alpha)(t_2-t_1)} + \beta_2(\alpha) \qquad (37)$$
$$\int_{t_1}^{t_2} e^{-\beta_1(\alpha)(t_2-\tau)}\delta_2(\alpha, \tau)d\tau \leq r_2 - \rho, \forall e_1 \in D'_v$$

holds, where $\beta_1(\alpha) = \frac{w}{\alpha} - \eta_1(\alpha) > 0$ and $\beta_2(\alpha) = e^{\eta_2(\alpha)} \geq 1$.

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

Proof. As shown in the proof of Theorem 3, the differential equation (24) has a unique solution defined for all $t \in [t_1, t_2]$ that stays in $\mathcal{M}^{h_2}(t)$ whenever $e_1 \in D_{v,\rho,r_2} \cap \mathcal{M}^{h_2}(t_1)$. 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 (24), the derivative of $V(e)$ along the trajectories of (24) can be obtained as

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

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

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

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

$$\dot{W} \leq -\left[\frac{w}{\alpha} - \delta_1(\alpha, t)\right]W + \frac{\delta_2(\alpha, t)}{\sqrt{2}} \quad (38)$$

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

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

This implies that

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

Therefore,

$$D^+ W(t) = \lim_{\epsilon \to 0^+} \sup \frac{W(t + \epsilon) - W(t)}{\epsilon}$$
$$= \lim_{\epsilon \to 0^+} \sup \frac{\sqrt{\frac{\epsilon^2}{2}\|p(\bar{e}, \alpha, t)\|^2 + o(\epsilon^2)}}{\epsilon} \quad (39)$$
$$= \frac{1}{\sqrt{2}}\|p(\bar{e}, \alpha, t)\|$$
$$\leq \frac{1}{\sqrt{2}}\delta_2(\alpha, t)$$

Thus, (38) is also satisfied when $V = 0$, and accordingly $D^+ W(t)$ satisfies (38) for all values of V. Since W is scalar and the right-hand side of (38) is continuous in t and locally Lipschitz in W for all $t \in [t_1, t_2]$ and $W \geq 0$, the comparison lemma is applicable. In addition, the right-hand side of (38) 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

$$W(t) \leq \phi(t, t_1)W(t_1) + \frac{1}{\sqrt{2}}\int_{t_1}^{t}\phi(t, \tau)\delta_2(\alpha, \tau)d\tau \quad (40)$$

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

$$\phi(t, t_1) = \exp\left[-\frac{w}{\alpha}(t - t_1) + \int_{t_1}^{t}\delta_1(\alpha, \tau)d\tau\right]. \quad (41)$$

$$\|e(t) - \bar{e}\| \leq \phi(t, t_1)\|e_1 - \bar{e}\| + \int_{t_1}^{t}\phi(t, \tau)\delta_2(\alpha, \tau)d\tau \quad (42)$$

Since $\int_{t_1}^{t}\delta_1(\alpha, \tau)d\tau \leq \eta_1(\alpha)(t - t_1) + \eta_2(\alpha)$, and using $\beta_1(\alpha) = \frac{w}{\alpha} - \eta_1(\alpha) > 0$ and $\beta_2(\alpha) = e^{\eta_2(\alpha)} \geq 1$ in (42), it holds that

$$\|e(t) - \bar{e}\| \leq \beta_2(\alpha)\|e_1 - \bar{e}\|e^{-\beta_1(\alpha)(t - t_1)}$$
$$+ \beta_2(\alpha)\int_{t_1}^{t}e^{-\beta_1(\alpha)(t - \tau)}\delta_2(\alpha, \tau)d\tau \quad (43)$$

By taking $e_1 \in D_v' \subseteq D_{v,\rho,r_2}$, since D_{v,ρ,r_2} retains trajectories starting from a feasible initial point with respect to

the dynamics (24) for $t \in [t_1, t_2]$, any trajectory of (24) starting from D_v' will stay in D_{v,ρ,r_2} and remain in the feasible set $\mathcal{M}^{h_2}(t)$. Thus, the bound in (43) 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 (25) jumps from $h_1(t)$ to $h_2(t)$ during the time interval $[t_1, t_2]$. $\quad\square$

Remark 4. If the global minimum trajectory is the dominant trajectory with respect to the spurious local minimum trajectories, then Theorems 3 and 4 guarantee that the solution of (P-ODE) will jump to the neighborhood of the global minimum trajectory.

Remark 5. The condition in Theorem 3 and Condition 2 in Theorem 4 mean 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 $h_2(t)$. Condition 1 in Theorem 4 means that the original time-varying system should not be too distant from the time-invariant averaged system.

Remark 6. To make the one-point strong monotonicity conditions (29) and (34) hold, the inertia parameter α cannot be too small.

Remark 7. The locally one-point strongly convex parameter w in (29) and (34) determines the convergence rate during $[t_1, t_2]$, which is reflected in (30) and (37).

Remark 8. In Theorem 4, to ensure that the time-invariant partial interval averaged system is a reasonable approximation of the time-varying system, the time interval $[t_1, t_2]$ should not be very large. On the other hand, to guarantee that the solution of (24) has enough time to jump, the time interval $[t_1, t_2]$ should not be very small. This trade-off is reflected in (37).

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 (P-ODE) is nonzero while the left-hand side is zero. Therefore, $h_2(t)$ is not a solution of (P-ODE) in general. This is because the solution of (P-ODE) approximates the continuous limit of a discrete local trajectory of the sequential regularized optimization problem (10). 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 (P-ODE) is close to $h_2(t)$.

If the solution of (24) can be shown to be in a small residual set around 0 on the time-varying manifold $\mathcal{M}(t)$, then it is guaranteed that $x(t, t_0, x_0)$ tracks its nearby local minimum trajectory. Notice that (24) can be regarded as a time-varying perturbation of the system

$$\dot{e} = -\frac{1}{\alpha}\mathcal{P}(e + h_2(t), t)\nabla_x f(e + h_2(t), t), \quad \forall t \geq t_0 \quad (44)$$

Since $h_2(t)$ is a local minimum trajectory, it is obvious that $e(t) \equiv 0$ is an equilibrium point of (44). In addition, if the time-varying Lagrange function $L(x, \lambda, t)$ with λ given in (3) is locally one-point strongly convex with respect to x around $h_2(t)$ in the time-varying feasible set $\mathcal{M}(t)$, after noticing the fact that the solution of (24) will remain in $\mathcal{M}^{h_2}(t)$ if the initial point $e_0 \in \mathcal{M}^{h_2}(t_0)$ from Lemma 3, one would expect that the solution of (24) stays in a small residual set of $e = 0$ if the perturbation $\mathcal{Q}(e(t) + h_2(t), t)g'(e(t) + h_2(t), t) + \dot{h}_2(t)$ is relatively small. The perturbation $\mathcal{Q}(e(t) + h_2(t), t)g'(e(t) + h_2(t), t) + \dot{h}_2(t)$ being small is equivalent to α being small. The next theorem shows that every local minimum trajectory can be tracked for a relatively small α.

Theorem 5 (Sufficient condition for tracking). Assume that the time-varying Lagrange function $L(x, \lambda, t)$ with λ given in (3) is locally (c_2, r_2)-one-point strongly convex with respect to x around $h_2(t)$. Given $\gamma(t)$ such that $\left\| \dot{h}_2(t) \right\| \leq \gamma(t)$, suppose that there exist time-varying scalar functions $\delta_1(t)$ and $\delta_2(t)$ such that the perturbed gradient due to the time-variation of constraints satisfies the inequality

$$\|\mathcal{Q}(e(t) + h_2(t), t)g'(e(t) + h_2(t), t)\| \leq \delta_1(t) \|e\| + \delta_2(t), \tag{45}$$

and there exist some positive constants η_1 and η_2 such that

$$\int_{t_1}^{t} \delta_1(\tau)d\tau \leq \eta_1(t - t_1) + \eta_2. \tag{46}$$

If $\sup_{t \geq t_1}(\delta_2(t) + \gamma(t))$ is bounded and the following conditions hold

$$\|x_0 - h_2(0)\| \leq \frac{r_2}{e^{\eta_2}}, \tag{47a}$$

$$\alpha \leq \frac{c_2 r_2}{e^{\eta_2} \sup_{t \geq t_1}(\delta_2(t) + \gamma(t)) + \eta_1 r_2}, \tag{47b}$$

then the solution $x(t, t_0, x_0)$ will r_2-track $h_2(t)$. More specifically, we have

$$\|x(t, t_0, x_0) - h_2(t)\| \leq e^{\eta_2} \|e_1\| e^{-(\frac{c_2}{\alpha} - \eta_1)(t - t_1)}$$
$$+ e^{\eta_2} \int_{t_1}^{t} e^{-(\frac{c_2}{\alpha} - \eta_1)(t - \tau)}(\delta_2(t) + \gamma(t))d\tau \leq r_2. \tag{48}$$

Proof. Consider $V(e) = \frac{1}{2} \|e\|^2 : \mathcal{B}_{r_2}(0) \to \mathbb{R}$ as the Lyapunov function for the system (24). Because of Lemma 3, any solution of (24) stating in $\mathcal{M}(t_1)$ will remain in $\mathcal{M}(t)$ for all $t \geq t_1$. The derivative of $V(e)$ along the trajectories of (24) can be obtained as

$$\dot{V} = e(t)^\top \Big(-\frac{1}{\alpha}\mathcal{P}(e(t) + h_2(t), t)\nabla_x f(e(t) + h_2(t), t)$$
$$- \mathcal{Q}(e(t) + h_2(t), t)g'(t)(e(t) + h_2(t), t) - \dot{h}_2^g(t) \Big),$$
$$\leq -\frac{c}{\alpha} \|e(t)\|^2 + \delta_1(t) \|e(t)\|^2 + (\delta_2(t) + \gamma(t)) \|e(t)\|$$

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

$$\dot{V} \leq -\Big[\frac{2c}{\alpha} - 2\delta_1(t)\Big]V + (\delta_2(t) + \gamma(t))\sqrt{2V}$$

Using the same proof procedure as in Theorem 4 and by taking $\beta_1(\alpha) = \frac{c}{\alpha} - \eta_1 > 0$ and $\beta_2 = e^{\eta_2} \geq 1$, it can be shown that

$$\|e(t)\| \leq \beta_2 \|e_1\| e^{-\beta_1(\alpha)(t - t_1)}$$
$$+ \beta_2 \int_{t_1}^{t} e^{-\beta_1(\alpha)(t - \tau)}(\delta_2(t) + \gamma(t))d\tau \tag{49}$$

To make the bound in (49) valid, we must ensure that $e(t) \in \mathcal{B}_{r_2}(0)$ for all $t \geq t_1$. Note that

$$\|e(t)\| \leq \beta_2 \|e_1\| e^{-\beta_1(\alpha)(t - t_1)} + \frac{\beta_2}{\beta_1(\alpha)}(1 - e^{-\beta_1(\alpha)(t - \tau)})$$
$$\sup_{t \geq t_0}(\delta_2(t) + \gamma(t))$$
$$\leq \max \Big\{ \beta_2 \|e_1\|, \frac{\beta_2}{\beta_1(\alpha)} \sup_{t \geq t_0}(\delta_2(t) + \gamma(t)) \Big\}$$

It can be verified that the condition $e(t) \in \mathcal{B}_{r_2}(0)$ will be satisfied if (47) holds. Furthermore, by $e(t) \in \mathcal{B}_{r_2}(0)$ and Theorem 1, there must exist a unique solution for (P-ODE) for all $t \geq t_1$. □

Remark 9. The inequality (48) implies that the smaller the regularization parameter α is, the smaller the tracking error $x(t, t_0, x_0) - h_2(t)$ is and the faster $x(t, t_0, x_0)$ converges to the neighbourhood of $h_2(t)$.

Remark 10. In the case that the local minimum trajectory $h_2(t)$ is a constant, the upper bound on α simply becomes $\alpha < \infty$. This implies that if $h_2(t)$ is constant, then it will be perfectly tracked with any regularization parameter and can not be escaped by tuning the regularization parameter.

Remark 11. In the unconstrained case or the case with the time-invariant constraints, $\delta_1(t)$ and $\delta_2(t)$ in (45) simply become zero. Then, the tracking conditions in (47) become $\|x_0 - h_2(0)\| \leq r_2$ and $\alpha \leq \frac{c_2 r_2}{\sup_{t \geq t_0} \gamma(t)}$, and the tracking error bound in (48) becomes

$$\|e(t)\| \leq \|e_1\| e^{-\frac{c_2}{\alpha}(t - t_1)} + \int_{t_1}^{t} e^{-\frac{c_2}{\alpha}(t - \tau)}\gamma(t)d\tau$$
$$\leq \frac{\alpha \sup_{t \geq t_1} \gamma(t)}{c_2}$$

Remark 12. After the solution of (P-ODE) has escaped the spurious local trajectories and started tracking the globally minimum trajectory, one may use the state-of-the-art tracking methods in [21] and [15] to improve the tracking of the globally minimum trajectory.

C. Escaping

Combining the results of jumping and tracking immediately yields a sufficient condition on escaping from one local minimum trajectory to a more desirable local (or global) minimum trajectory. The proof is omitted for brevity.

Theorem 6 (Sufficient conditions for escaping from $h_1(t)$ to $h_2(t)$). Given two local minimum trajectories $h_1(t)$ and $h_2(t)$, suppose that the Lagrange function $L(x, \lambda, t)$

with λ given in (3) is locally (c_2, r_2)-one-point strongly convex with respect to x around $h_2(t)$ in the time-varying feasible set $\{e \in \mathbb{R}^n : e + h_2(t) \in \mathcal{M}(t), \|e\| \leq r_2\}$ and let $\mathcal{B}_v(h_1(t_1)) \subseteq RA^{\mathcal{M}(t_1)}(h_1(t_1))$. Under the conditions of Theorem 3 or 4, if (45)-(47) hold, then the solution of (P-ODE) will (v, r_2)-escape from $h_1(t)$ to $h_2(t)$ after $t \geq t_2$.

D. Discussions

Adaptive inertia: To leverage the potential of the time-varying perturbation $\alpha \mathcal{Q}(e(t) + h_2(t), t)g'(e(t) + h_2(t), t) + \alpha \dot{h}_2(t)$ in re-shaping the landscape of the Lagrange function or the objective function to become locally one-point strongly convex in x over a large region, the regularization parameter α should be selected relatively large. On the other hand, to ensure that the solution of (24) and (26) will end up tracking a desirable local (or global) minimum trajectory, Theorem 5 prescribes small values for α. In practice, especially when the time-varying objective function has many spurious shallow minimum trajectories, this suggests using a relatively large regularization parameter α at the beginning of the time horizon to escape spurious shallow minimum trajectories and then switching to a relative small regularization parameter α for reducing the ultimate tracking error bound.

Sequential jumping: When the time-varying optimization problem has many local minimum trajectories, the solution of (P-ODE) or (ODE) may sequentially jump from one local minimum trajectory to a better local minimum trajectory. To illustrate this concept, consider the local minimum trajectories $h_1(t), h_2(t), ..., h_m(t)$, where $h_m(t)$ is a global trajectory. Assume that there exists a sequence of time intervals $[t_1^i, t_2^i]$ for $i = 1, 2, \ldots, m - 1$ such that the conditions of Theorem 3 or 4 are satisfied for $h_i(t)$ and $h_{i+1}(t)$ during each time interval. Then, by sequentially deploying Theorem 3 or 4, it can be concluded that the solution of (P-ODE) or (ODE) will jump from $h_1(t)$ to $h_m(t)$ after $t \geq t_2^m$. Furthermore, if $h_m(t)$ can be tracked with the given α, the solution of (P-ODE) or (ODE) will escape from $h_1(t)$ to $h_m(t)$ after $t \geq t_2^m$.

V. Numerical Examples

Example 3. *Consider the non-convex function*

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

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 7. When $d = 0$, this function is called the Ackley function [52], which is a benchmark function for global optimization algorithms. To make this function twice continuously differentiable, we choose $d = 0.01$.

Consider the time-varying objective function $f(x, t) = \bar{f}(x - z(t))$ and the time-varying constraint $g(x, t) = (x_1 - z_1(t)) - 1/2(x_2 - z_2(t))^2 = 0$, where $z(t) =$
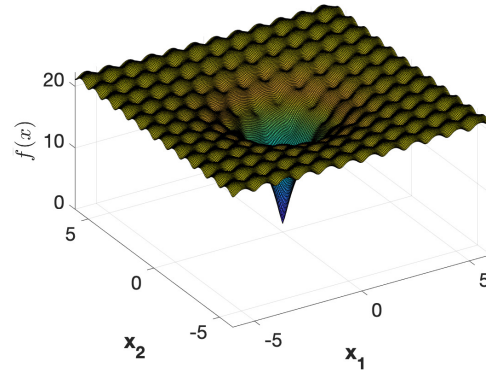

Fig. 7. Illustration of Example 3.

$[24\sin(t), \cos(t)]^\top$. This constrained time-varying optimization problem has the global minimum trajectory $[0, 0]^\top + z(t)$ and many spurious local minimum trajectories. Two local minimum trajectories are $h_1(t) = [1.92, 1.96]^\top + z(t)$ and $h_2(t) = [0, 0]^\top + z(t)$. It can be shown that $L(x, \lambda, t)$ is locally $(20, 0.5)$-one-point strongly convex with respect to $h_2(t)$.

We take $D_{v,\rho,r_2} = D_{0.04,0.01,1} = [-0.1, 2] \times [-0.1, 2]$ in Definition 10. The condition in (27) can be verified by checking the signs of the derivatives of $e_1(t)$ and $e_2(t)$ along the dynamics (24) on the boundary points of $D_{0.04,0.01,1} \cap \mathcal{M}^{h_2}(t)$. Furthermore, (34) is satisfied for $w = 1$. Thus, $h_2(t)$ is a $(0.2, 1)$-dominant trajectory with respect to $h_1(t)$ during $[0, \frac{\pi}{8}]$ over the region $D_{0.04,0.01,1}$.

Regarding Theorem 3, if we select $\theta = 0.2$, the inequality (37) is satisfied for $\alpha = 0.2$ and $t_2 - t_1 = \pi/8$. Thus, the solution of (P-ODE) will $(0.04, 0.5)$-jump from $h_1(t)$ to $h_2(t)$. Regarding Theorem 5, δ_1 and δ_2 in the inequality (45) can be taken as 0 and $24\sqrt{2}\cos(t) + \sqrt{2}\sin(t)$, respectively. Then the inequality (47b) reduces to $\alpha \leq \frac{10}{\sqrt{2(24^2 + 1)}} \approx 0.29$, which is satisfied by $\alpha = 0.2$. Thus, the solution of (P-ODE) will 0.5-track $h_2(t)$. Putting the above findings together, we can conclude that the solution of (24) will $(0.04, 0.5)$-escape from $h_1(t)$ to $h_2(t)$.

In addition, by choosing the inertia parameter $\alpha = 0.2$, the simulation shows that for 1000 runs of random initialization with $x_2(0) - z(0) \in [-5, 5]$ and $x_1(0)$ determined by the equality constraint, all solutions of the corresponding (P-ODE) will sequentially jump over the local minimum trajectories and end up tracking the global trajectory after $t \geq 5\pi$.

VI. 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 projected gradient flow system with controllable inertia as a continuous-time limit of the optimality conditions for discretized sequential optimization problems with proximal regularization and

online updating scheme. Via a change of variables, the time-varying projected gradient flow system is regarded as a composition of a time-varying projected gradient term, a time-varying constraint-driven term and an inertia term due to the time variation of the local minimum trajectory. We show that the time-varying perturbation term due to the inertia encourages the exploration of the state space and re-shapes the landscape by potentially making it one-point strongly convex over a large region during some time interval. We introduce the notions of jumping and escaping, and use them to develop sufficient conditions under which the time-varying solution escapes from a poor local trajectory to a better (or global) minimum trajectory over a finite time interval. We illustrate in a benchmark example with many shallow minimum trajectories that the natural time variation of the problem enables escaping spurious local minima over time. Avenues for future work include the characterization of the class of problems in which all spurious local minimum trajectories are shallow compared with the global minimum trajectory.

Appendix A

Lemma 6. Under Assumptions 1-3, if $f(x,t)$, $g_1(x,t), \ldots, g_m(x,t)$ are twice continuously differentiable in x on $D \times [a,b]$ for some domain $D \in \mathbb{R}^n$, then the function $\frac{1}{\alpha}\mathcal{P}(x,t)\nabla_x f(x,t) + \mathcal{Q}(x,t)g'(x,t)$ is locally Lipschitz continuous in x on $\{(x,t) \in D \times [a,b] : x \in \mathcal{M}(t)\}$.

Proof. We first show that the matrix $\mathcal{J}_g(x,t)\mathcal{J}_g(x,t)^\top$ is positive definite over the time-varying feasible region $\mathcal{M}(t)$ for all $t \in [a,b]$. Since, under Assumption 3, $\mathcal{J}_g(x,t)$ is full row rank for $x \in \mathcal{M}(t)$ for all $t \in [a,b]$, the null space of $\mathcal{J}_g(x,t)$ is 0. Thus, $\mathcal{J}_g(x,t)x = 0$ if and only if $x = 0$. Therefore, $x\mathcal{J}_g(x,t)\mathcal{J}_g(x,t)^\top x^\top = (x\mathcal{J}_g(x,t))(x\mathcal{J}_g(x,t))^\top > 0$ for all $x \neq 0$. Therefore, $\mathcal{J}_g(x,t)\mathcal{J}_g(x,t)^\top$ is positive definite. Denote $\sigma_{\min}(\mathcal{J}_g(x,t)\mathcal{J}_g(x,t)^\top)$ as the minimum eigenvalue of $\mathcal{J}_g(x,t)\mathcal{J}_g(x,t)^\top$. Then, there exists a positive constant c such that $\sigma_{\min}(\mathcal{J}_g(x,t)\mathcal{J}_g(x,t)^\top) \geq c$. By the chain rule and the twice continuously differentiability of $g_k(x,t)$'s in $x \in D$ for $t \in [a,b]$, we know that $G(x,t) := \mathcal{J}_g(x,t)\mathcal{J}_g(x,t)^\top$ is continuously differentiabile in $x \in D$ for $t \in [a,b]$. Next, we show that G^{-1} is also continuously differentiabile in x on $\{(x,t) \in D \times [a,b] : x \in \mathcal{M}(t)\}$. Let x_i be the i-th component of the vector x. By taking the derivative of the identity $I = G(x,t)G(x,t)^{-1}$ with respect to x_i, we obatin

$$\frac{\partial}{\partial x_i}G(x,t)^{-1} = -G(x,t)^{-1}\Big(\frac{\partial}{\partial x_i}G(x,t)\Big)G(x,t)^{-1}$$

and

$$\left\|\frac{\partial}{\partial x_i}G(x,t)^{-1}\right\| \leq \left\|G(x,t)^{-1}\right\|^2 \left\|\frac{\partial}{\partial x_i}G(x,t)\right\|$$
$$\leq \frac{1}{c^2}\left\|\frac{\partial}{\partial x_i}G(x,t)\right\| < \infty$$

Since the inversion operator is continuous by the Cayley–Hamilton theorem, $G(x,t)$ is continuously differentiable in x and $\left\|\frac{\partial}{\partial x_i}G(x,t)^{-1}\right\|$ is bounded, we know that

$\frac{\partial}{\partial x_i}G(x,t)^{-1}$ is well-defined and continuous over D for all $i = 1, \ldots, n$. Therefore, $G(x,t)^{-1}$ is continuously differentiable in x on $\{(x,t) \in D \times [a,b] : x \in \mathcal{M}(t)\}$. Consequently, $\mathcal{P}(x,t)$ and $\mathcal{Q}(x,t)$ are continuously differentiable in x on $\{(x,t) \in D \times [a,b] : x \in \mathcal{M}(t)\}$. Since $f(x,t)$ is twice continuously differentiable in x and $g_k(x,t)$'s are twice continuously differentiable in $x \in D$ for $t \in [a,b]$, it holds that the function $\frac{1}{\alpha}\mathcal{P}(x,t)\nabla_x f(x,t) + \mathcal{Q}(x,t)g'(x,t)$ is continuously differentiable in x. Hence, because of [36, Theorem 3.2], it is locally Lipschitz continuous in x on $\{(x,t) \in D \times [a,b] : x \in \mathcal{M}(t)\}$. $\square$
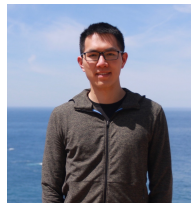
References

[1] Y. Ding, J. Lavaei, and M. Arcak, "Escaping spurious local minimum trajectories in online time-varying nonconvex optimization," *submitted for conference publication*, 2020.

[2] E. Hazan *et al.*, "Introduction to online convex optimization," *Foundations and Trends® in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[3] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1724–1732.

[4] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—online stochastic gradient for tensor decomposition," in *Conference on Learning Theory*, 2015, pp. 797–842.

[5] R. Kleinberg, Y. Li, and Y. Yuan, "An alternative view: When does SGD escape local minima?" *arXiv preprint arXiv:1802.06175*, 2018.

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

[7] J. Mulvaney-Kemp, S. Fattahi, and J. Lavaei, "Smoothing property of load variation promotes finding global solutions of time-varying optimal power flow," *https://lavaei.ieor.berkeley.edu/ DOPF_2020_2.pdf*, 2020.

[8] Y. Tang, K. Dvijotham, and S. Low, "Real-time optimal power flow," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2963–2973, 2017.

[9] A. Hauswirth, I. Subotić, S. Bolognani, G. Hug, and F. Dörfler, "Time-varying projected dynamical systems with applications to feedback optimization of power systems," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 3258–3263.

[10] C. V. Rao, J. B. Rawlings, and D. Q. Mayne, "Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations," *IEEE Transactions on Automatic Control*, vol. 48, no. 2, pp. 246–258, 2003.

[11] T. Binder, L. Blank, H. G. Bock, R. Bulirsch, W. Dahmen, M. Diehl, T. Kronseder, W. Marquardt, J. P. Schlöder, and O. von Stryk, "Introduction to model based optimization of chemical processes on moving horizons," in *Online optimization of large scale systems*. Springer, 2001, pp. 295–339.

[12] M. Salman Asif and J. Romberg, "Sparse recovery of streaming signals using l1-homotopy," *arXiv preprint arXiv:1306.3331*, 2013.

[13] A. Balavoine, C. J. Rozell, and J. Romberg, "Discrete and continuous-time soft-thresholding with dynamic inputs," *arXiv preprint arXiv:1405.1361*, 2014.

[14] J. Kadam, W. Marquardt, M. Schlegel, T. Backx, O. Bosgra, P. Brouwer, G. Dünnebier, D. Van Hessem, A. Tiagounov, and S. De Wolf, "Towards integrated dynamic real-time optimization and control of industrial processes," *Proceedings foundations of computer-aided process operations (FOCAPO2003)*, pp. 593–596, 2003.

[15] V. M. Zavala, E. M. Constantinescu, T. Krause, and M. Anitescu, "On-line economic optimization of energy systems using weather forecast information," *Journal of Process Control*, vol. 19, no. 10, pp. 1725–1736, 2009.

[16] A. Simonetto, A. Mokhtari, A. Koppel, G. Leus, and A. Ribeiro, "A class of prediction-correction methods for time-varying convex optimization," *IEEE Transactions on Signal Processing*, vol. 64, no. 17, pp. 4576–4591, 2016.

[17] M. Fazlyab, C. Nowzari, G. J. Pappas, A. Ribeiro, and V. M. Preciado, "Self-triggered time-varying convex optimization," in *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 3090–3097.

[18] A. Bernstein, E. Dall'Anese, and A. Simonetto, "Online primal-dual methods with measurement feedback for time-varying convex optimization," *IEEE Transactions on Signal Processing*, vol. 67, no. 8, pp. 1978–1991, 2019.

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

[20] J. Guddat, F. G. Vazquez, and H. T. Jongen, *Parametric optimization: singularities, pathfollowing and jumps*. Springer, 1990.

[21] Y. Tang, E. Dall'Anese, A. Bernstein, and S. Low, "Running primal-dual gradient method for time-varying nonconvex problems," *arXiv preprint arXiv:1812.00613*, 2018.

[22] O. Massicot and J. Marecek, "On-line non-convex constrained optimization," *arXiv preprint arXiv:1909.07492*, 2019.

[23] 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.

[24] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_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.

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

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

[27] 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.

[28] 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.

[29] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, "First-order methods almost always avoid saddle points," *arXiv preprint arXiv:1710.07406*, 2017.

[30] 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.

[31] 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.

[32] 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*, 2019.

[33] 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.

[34] 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.

[35] 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, 2016.

[36] H. K. Khalil, *Nonlinear systems*. Upper Saddle River, 2002.

[37] J. K. Hale, *Ordinary differential equations*. Wiley-Inter-science, 1980.

[38] 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.

[39] W. Krichene, A. Bayen, and P. L. Bartlett, "Accelerated mirror descent in continuous and discrete time," in *Advances in Neural Information Processing Systems*, 2015, pp. 2845–2853.

[40] 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.

[41] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[42] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Implicit regularization in matrix factorization," in *Advances in Neural Information Processing Systems*, 2017, pp. 6151–6159.

[43] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 2016.

[44] A. Sard, "The measure of the critical values of differentiable maps," *Bulletin of the American Mathematical Society*, vol. 48, no. 12, pp. 883–890, 1942.

[45] G. Still, "Lectures on parametric optimization: An introduction," *Optimization Online*, 2018.

[46] J. Rosen, "The gradient projection method for nonlinear programming. part ii. nonlinear constraints," *Journal of the Society for Industrial and Applied Mathematics*, vol. 9, no. 4, pp. 514–532, 1961.

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

[48] A. Iserles, *A first course in the numerical analysis of differential equations*. Cambridge University Press, 2009, no. 44.

[49] A. R. Teel, J. Peuteman, and D. Aeyels, "Semi-global practical asymptotic stability and averaging," *Systems & Control Letters*, vol. 37, no. 5, pp. 329–334, 1999.

[50] D. Aeyels and J. Peuteman, "On exponential stability of nonlinear time-varying differential equations," *Automatica*, vol. 35, no. 6, pp. 1091–1100, 1999.

[51] J. Peuteman and D. Aeyels, "Exponential stability of nonlinear time-varying differential equations and partial averaging," *Mathematics of Control, Signals and Systems*, vol. 15, no. 1, pp. 42–70, 2002.

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

Yuhao Ding is currently working toward the Ph.D. degree in Industrial Engineering and Operations Research at the University of California, Berkeley, CA, USA. He obtained the B.E. degree in Aerospace Engineering from Nanjing University of Aeronautics and Astronautics in 2016, and the M.S. degree in Electrical and Computer Engineering from University of Michigan, Ann Arbor in 2018.

Javad Lavaei is currently an Associate Professor in the Department of Industrial Engineering and Operations Research at the University of California, Berkeley, CA, USA. He has worked on different interdisciplinary problems in power systems, optimization theory, control theory, and data science. He is an associate editor of the IEEE Transactions on Automatic Control, the IEEE Transactions on Smart Grid, and the IEEE Control System Letters.

Murat Arcak is currently a Professor with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA. His research interests include dynamical systems and control theory with applications to synthetic biology, multiagent systems, and transportation. Prof. Arcak was a recipient of the CAREER Award from the National Science Foundation in 2003 and the Donald P. Eckman Award from the American Automatic Control Council in 2006.