Escaping spurious local minimum trajectories in online time-varying nonconvex optimization

Yuhao Ding [1] **Javad Lavaei** [1] **Murat Arcak** [2]

Abstract

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

1. Introduction

In this paper, we study the following unconstrained online optimization problem at times $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$:

$$\min_{x(\tau_i)\in\mathbb{R}^n} \quad f(x(\tau_i), \tau_i), \quad i = 0, 1, 2, \ldots \tag{1}$$

where τ_i denotes the time index and $x(\tau_i)$ is the optimization variable at time τ_i. For each time τ_i, the function $f(x(\tau_i), \tau_i)$ could potentially be nonconvex in $x(\tau_i)$ with many local minima. The objective is to solve the above problem in an online fashion under the assumption that at any given time τ_i the function $f(x, \tau)$ is known for all $\tau \leq \tau_i$ while no knowledge about $f(x, \tau)$ may be available for any $\tau > \tau_i$. Therefore, the functions $f(x, \cdot)$'s cannot be minimized off-line and should be solved sequentially. Another issue is that the optimization problem at each time instance could be highly complex due to NP-hardness, which is an impediment to finding its global minima. Note that problem (1) is stated as a deterministic optimization, but the function $f(x, \tau_i)$ could be equal to the expected value of a stochastic function at time τ_i, where the stochasticity is due to uncertain environment, unknown model, or noise (Roy et al., 2019). 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.

To mathematically analyze the online optimization (1), we first assume that there is a function $f(x, t)$ that is continuous in t whose samples include the sequential functions $\{f(x, \tau_i), i = 0, 1, 2, \ldots\}$. In other words, the original problem (1) can be regarded as a discretization of the time-varying continuous optimization problem:

$$\min_{x(t)\in\mathbb{R}^n} \quad f(x(t), t) \tag{2}$$

If $f(x, t)$ does not change over time, the problem reduces to a classic (time-invariant) nonconvex optimization problem. It is known that simple local search methods, such as stochastic gradient descent (SGD) (Hazan et al., 2016), 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 (Jin et al., 2017; Ge et al., 2015; Kleinberg et al., 2018). 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 gradient flow system for time-invariant optimization problems and serves as a continuous-time limit of the discrete online

[1]Department of Industrial Engineering and Operations Research, University of California, Berkeley, USA [2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA. Correspondence to: Yuhao Ding <yuhao_ding@berkeley.edu>, Javad Lavaei <lavaei@berkeley.edu>, Murat Arcak <arcak@berkeley.edu>.

This paper is submitted for the conference publication.

tracking method for (1). This ODE is given as

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

where $\alpha > 0$ is a constant parameter named **inertia** due to a **proximal regularization**. A system of the form (ODE) is called a **time-varying gradient system with inertia** α. The behavior of the solutions of this system initialized at different points depends on the value of α. We offer a motivating example below before stating the goals of this paper.

1.1. Motivating example

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

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

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 (ODE) with the initial point -2. In the case with $\alpha = 0.3$ and $b = 5$, the solution of (ODE) winds up in the region of attraction of the global minimum trajectory. However, for the case with $\alpha = 0.1$ and $b = 5$, the solution of (ODE) remains in the region of attraction of the spurious local minimum trajectory. In the case with $\alpha = 0.8$ and $b = 5$, the solution of (ODE) fails to track any local minimum trajectory. In the case with $\alpha = 0.1, b = 10$, the solution of (ODE) winds up in the region of attraction of the global minimum trajectory.

The observations in this example can be summarized as: (1) 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; (2) When the inertia α is relatively small, the solution of (ODE) tends to track a local (or global) minimum trajectory closely and converges to that trajectory quickly.

1.2. Our contributions

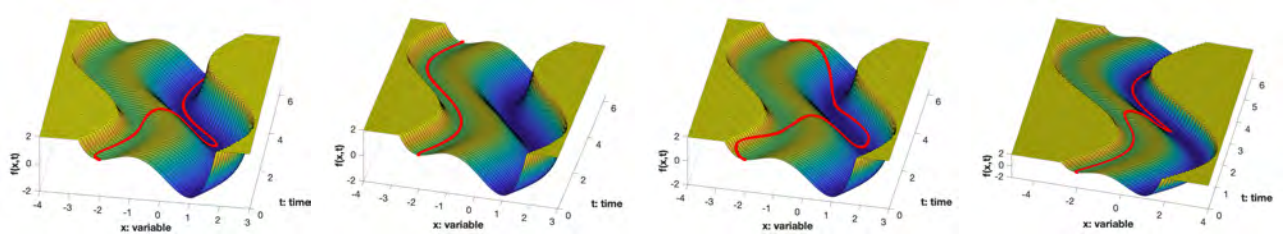
In order to mathematically study the observations made in Example 1 for a general online optimization problem, we focus on the aforementioned time-varying gradient flow system with inertia α as a continuous-time limit of the stationary condition for a discretized sequential optimization problem with a proximal regularization and its online updating scheme. 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 by potentially making it one-point strongly convex over a large region during some time interval. We show that if a given spurious local minimum trajectory is a shallow minimum trajectory compared to 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 sufficient conditions under which the ODE solution will jump from a certain local minimum trajectory to another local minimum trajectory. We then derive a sufficient condition on the inertia α to guarantee that the solution of (ODE) can track a global minimum trajectory. We also provide an ultimate tracking error bound and estimate the time for reaching the tracking error bound. To illustrate the technical results on how the time variation nature of an online optimization problem enables escaping a spurious minimum trajectory, we offer a case study with many shallow minimum trajectories.

1.3. Related work

Online time-varying optimization problems: There are many papers on designing efficient online algorithms for tracking the optimizers of time-varying convex optimization problems (Simonetto et al., 2016; Fazlyab et al., 2016; Bernstein et al., 2018; Simonetto, 2017). With respect to time-varying nonconvex optimization problems, Guddat et al. (1990) presents a comprehensive theory on the structure and singularity of the KKT trajectories for time-varying optimization problems. Tang et al. (2017; 2018); Massicot & Marecek (2019) develop algorithms to track the KKT trajectory or the local optimal solution of the time-varying optimization problems. Recently, Fattahi et al. (2019) asked the question of whether the natural temporal variation in a time-varying nonconvex optimization problem could help a local tracking method escape spurious local minimum trajectories, but that work lacked theoretical results on this phenomenon.

Online optimization for machine learning: A common framework in machine learning for analyzing a time-varying optimization problem is online optimization (Hazan et al., 2016). In general, the main goal in such online convex optimization is to propose a sequential algorithm and measure its performance through the notion of stationary regret (Zinkevich, 2003) or dynamic regret (Besbes et al., 2015; Jadbabaie et al., 2015), depending on whether there is any regularity condition on the temporal variability of the sequence of objective functions or minimizers. With respect to the general online nonconvex optimization, Hazan et al. (2017); Roy et al. (2019) propose to minimize a surrogate notion of regret based on the first-order or second-order stationary conditions, which measures the sub-optimality compared to a local point-wise solution to the problem. Contrary to this line of research, we focus on the global



(a) $\alpha = 0.3, b = 5$ (b) $\alpha = 0.1, b = 5$ (c) $\alpha = 0.8, b = 5$ (d) $\alpha = 0.1, b = 10$

Figure 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 1(a) and 1(d) when the inertia α and the change (controlled by the parameter b) of local minimum trajectory are appropriate.

landscape and the ultimate tracking error bound of general time-varying nonconvex optimization problems.

Local search methods for global optimization: It has been recently shown that simple local search methods, such as gradient-based algorithms, have a superb performance in solving nonconvex optimization problems. For example, Jin et al. (2017); Ge et al. (2015) 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 (Bhojanapalli et al., 2016; Ge et al., 2016; Zhang et al., 2019), robust principle component analysis (Fattahi & Sojoudi, 2018; Josz et al., 2018), and dictionary recovery (Sun et al., 2016) have benign landscape, implying that they are free of spurious local minima. The work Kleinberg et al. (2018) proves that SGD could help escape sharp local minima of a loss function. However, these results are all for time-invariant optimization problems. In contrast, many real-world problems should be solved sequentially over time with time-varying data. Therefore, it is essential to study the effect of the temporal variation on the landscape of time-varying nonconvex 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 (Khalil, 2002; Hale, 1980). The recent papers Su et al. (2014); Krichene et al. (2015); Wibisono et al. (2016) 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 (Saxe et al., 2013) and matrix regression (Gunasekar et al., 2017). It is natural to analyze the continuous-time limit of an online algorithm for tracking a KKT trajectory of time-varying optimization problem (Simonetto et al., 2016; Tang et al., 2018; Massicot & Marecek, 2019; Fattahi et al., 2019).

1.4. Notations

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

2. Preliminaries and Problem Formulation

In this work, we assume that $f : \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$ is twice continuously differentiable in x and continuously differentiable in $t \geq 0$. Moreover, suppose that f is uniformly bounded from below, meaning that there exists a constant M such that $f(x,t) \geq M$ for all $x \in \mathbb{R}^n$ and $t \geq 0$.

2.1. Time-varying optimization

The first-order stationary condition for (2) is as follows:

$$0 = \nabla_x f(x(t), t) \tag{3}$$

Since the solution is time-varying, we define the notion of stationary trajectories below.

Definition 1. *Given a time interval $I_t \subseteq [0, \infty)$, a continuous trajectory $h(t) : I_t \to \mathbb{R}^n$ is said to be a **stationary trajectory** of the time-varying optimization (2) if $0 = \nabla_x f(h(t), t)$ for all $t \in I_t$.*

In this work, we assume that the real roots of (3) are all isolated at each time t. An isolated stationary trajectory $h(t)$ can theoretically be a mix of local minima, local maxima and saddle points of the function $f(x,t)$ at different times. However, the goal of this work is to study only isolated local minimum trajectories of the time-varying optimization (2).

Definition 2. *A continuous trajectory $h(t) : I_t \to \mathbb{R}^n$ is*

*said to be a **local (or global) I_t-minimum trajectory** of the time-varying optimization (2) if I_t is a maximal interval such that each point of $h(t)$ is a local (or global) minimum of (2) at time $t \in I_t$. In particular, $h(t)$ is called a **local (or global) ∞-minimum trajectory** of the time-varying optimization (2) if $I_t = [t_0, \infty)$.*

After freezing the time t in (2) at a particular value, one may use local search methods to minimize $f(x,t)$. The notion of region of attraction is defined by resorting to the continuous-time model of local search algorithms (for which the step size is not important anymore).

Definition 3. *The **region of attraction** of a local minimum point $h(t)$ of $f(\cdot,t)$ at a given time is defined as:*

$$RA(h(t)) = \Big\{x_0 \in \mathbb{R}^n \;\Big|\; \lim_{\tilde{t}\to\infty} x(\tilde{t}) = h(t), \text{where } \frac{d\tilde{x}(\tilde{t})}{d\tilde{t}} = -\nabla_x f(\tilde{x}(\tilde{t}),t) \quad \text{and} \quad \tilde{x}(0) = x_0\Big\} \tag{4}$$

Definition 4. *Consider arbitrary positive scalars c and r together with an interval $\bar{I}_t$, where $\bar{I}_t \subset I_t$ if I_t is finite and $\bar{I}_t = I_t = [t_0, \infty]$ otherwise. The function $f(x,t)$ is said to be **locally** $(\bar{I}_t, c, r)$**-one-point strongly convex** around the local I_t-minimum trajectory $h(t)$ if*

$$\nabla_x f(e + h(t), t)^\top e \geq c\|e\|^2, \quad \forall e \in D, \quad \forall t \in \bar{I}_t \tag{5}$$

*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** $(\bar{I}_t, c, r)$**-one-point strong convexity** around $h(t)$.*

Note that (5) 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. In this paper, we assume any local I_t-minimum trajectory $h(t)$ satisfies the following assumptions.

Assumption 1. *$h(t)$ is isolated.*

Assumption 2. *The time-varying function $f(x,t)$ is locally one-point $(\bar{I}_t, c, r)$-strongly convex around $h(t)$ for some constants c and r as well as an interval $\bar{I}_t$.*

Assumption 3. *$h(t)$ is continuously differentiable.*

2.2. Derivation of time-varying gradient flow system

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

$$\min_{x\in\mathbb{R}^n} \quad f(x, \tau_0), \tag{6a}$$

$$\min_{x\in\mathbb{R}^n} \quad f(x,\tau_i) + \frac{\alpha\left\|x - x^*_{i-1}\right\|^2}{2(\tau_i - \tau_{i-1})}, i = 1, 2, \ldots \tag{6b}$$

where x^*_{i-1} denotes an arbitrary local minimum of the modified optimization problem (6) 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 (6). 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.

Due to the first-order optimality condition, the local minimum x^*_i of (6) at time step τ_i satisfies the equation:

$$\nabla_x f(x^*_i, \tau_i) + \alpha\frac{x^*_i - x^*_{i-1}}{\tau_i - \tau_{i-1}} = 0 \tag{7}$$

Since the function $f(x,\tau_i)$ is nonconvex in general, the problem (6b) may not have a unique solution x^*_i. In order to cope with this issue, we study the continuous-time limit of (7) as the time step $\tau_{i+1} - \tau_i$ diminishes to zero. This yields the time-varying ordinary differential equation:

$$\alpha\dot{x}(t) = -\nabla_x f(x(t),t), \quad x(0) = x^*_0 \tag{8}$$

When $\alpha = 0$, the differential equation (8) reduces to the algebraic equation (3), which is indeed the first-order stationary condition for the unregularized time-varying optimization (2). When $\alpha > 0$, we will show that (8) has a unique solution defined for all $t \geq 0$ under the assumption that the solutions of (8) lie in a compact set[1].

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

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

$$\bar{x}^*_i = \bar{x}^*_{i-1} - \frac{\tau_i - \tau_{i-1}}{\alpha}\nabla_x f(\bar{x}^*_{i-1}, \tau_{i-1}) \tag{9}$$

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

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

Proposition 1 (Convergence). *Given a local minimum x_0^* of (6a), as the time difference $\Delta_\tau = \tau_{i+1} - \tau_i$ approaches zero, any sequence of discrete local trajectories (x_k^Δ) converges to the (ODE) in the sense that for all fixed $T > 0$:*

$$\lim_{\Delta_\tau \to 0} \max_{0 \le k \le \frac{T}{\Delta_\tau}} \left\| x_k^\Delta - x(\tau_k, \tau_0, x_0^*) \right\| = 0 \qquad (10)$$

and any sequence of $(\bar{x}_k^\Delta)$ updated by (9) converges to the (ODE) in the sense that for all fixed $T > 0$:

$$\lim_{\Delta_\tau \to 0} \max_{0 \le k \le \frac{T}{\Delta_\tau}} \left\| \bar{x}_k^\Delta - x(\tau_k, \tau_0, x_0^*) \right\| = 0 \qquad (11)$$

2.3. Jumping, tracking and escaping

In this section, the objective is to study the case where there are at least two local minima at some time instance of the online optimization problem. Consider a local $I_{t,1}$-minimum trajectory $h_1(t)$ and a local $I_{t,2}$-minimum trajectory $h_2(t)$. Suppose that $f(x,t)$ is locally $(\bar{I}_{t,1}, c_1, r_1)$-one-point strongly convex around $h_1(t)$ and locally $(\bar{I}_{t,2}, c_2, r_2)$-one-point strongly convex around $h_2(t)$. Let $[t_1, t_2] \subset \bar{I}_{t,1} \cap \bar{I}_{t,2}$ be a non-empty interval. We provide the definitions of jumping, tracking and escaping below.

Definition 5. *It is said that the solution of (ODE) **(v,u)-jumps** from $h_1(t)$ to $h_2(t)$ over the time interval $[t_1, t_2]$ if there exist $u > 0$ and $v > 0$ such that*

$$\mathcal{B}_v(h_1(t_1)) \subseteq RA(h_1(t_1)) \qquad (12a)$$
$$\mathcal{B}_u(h_2(t_2)) \subseteq RA(h_2(t_2)) \qquad (12b)$$
$$\forall x_1 \in \mathcal{B}_v(h_1(t_1)) \Longrightarrow x(t_2, t_1, x_1) \in \mathcal{B}_u(h_2(t_2)) \qquad (12c)$$

For now, we assume that $I_{t,2}$ is an infinite time interval and give the following definition of tracking[2].

Definition 6. *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*

$$\begin{aligned} &x(t, t_0, x_0) \in \mathcal{B}_u(h_2(t)), \quad \forall t \ge T \\ &\mathcal{B}_u(h_2(t)) \subseteq RA(h_2(t)), \quad \forall t \ge T \end{aligned} \qquad (13)$$

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*

$$\mathcal{B}_v(h_1(t_0)) \subseteq RA(h_1(t_0)) \qquad (14a)$$
$$\mathcal{B}_u(h_2(t)) \subseteq RA(h_2(t)), \forall t \in \{\tau \in \bar{I}_{t,2} : \tau \ge T\} \qquad (14b)$$
$$\forall x_0 \in \mathcal{B}_v(h_1(t_0)) \Longrightarrow x(t, t_0, x_0) \in \mathcal{B}_u(h_2(t)), \quad \forall t \in \{\tau \in \bar{I}_{t,2} : \tau \ge T\} \qquad (14c)$$

Figure 2 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 the scenario where a solution $x(t, t_0, x_0)$ tracking a poor solution $h_1(t)$ at the beginning ends up jumping to and tracking 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 $h_2(t)$.

[2]The scenario that $I_{t,2}$ is a finite time interval is defined and analyzed in the appendix.

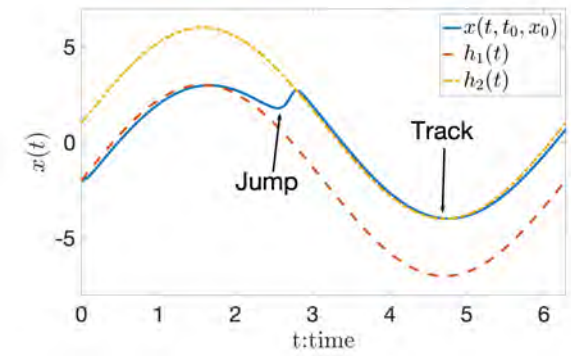


Figure 2. Illustration of jumping and tracking.

3. Optimization landscape after a change of variables

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

$$\dot{e}(t) = -\frac{1}{\alpha} \nabla_x f(e(t) + h_2(t), t) - \dot{h}_2(t), \forall t \ge t_0 \qquad (15)$$

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 $e_0 = x_0 - h_2(t_0)$. Note that $h_1(t)$ and $h_2(t)$ are local solutions of $f(x,t)$ and as long as $f(x,t)$ is time-varying, these functions cannot satisfy (ODE) in general.

3.1. Inertia encouraging the exploration

The first term $\nabla_x f(e + h_2(t), t)$ in (15) can be understood as a time-varying gradient term that encourages the solution of (15) to track the local minimum $h_2(t)$, 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 the appropriate proximal regularization α that balances the

time-varying gradient term and the inertia term, the solution of (ODE) could temporarily track a local minimum trajectory while keeping the potential of exploring other local minimum trajectories.

3.2. Inertia creating a one-point strongly convex landscape

The differential equation (15) can be written as

$$\dot{e}(t) = -\frac{1}{\alpha}\nabla_e\Big(f(e(t)+h_2(t),t) + \alpha\dot{h}_2(t)^\top e(t)\Big) \quad (16)$$

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 (16) starting in a neighborhood of $h_1(t)$ to move towards a neighborhood of $h_2(t)$. We illustrate this concept through Example 1 in the appendix.

From the right-hand side of (16), it can be inferred that if the gradient of $f(\cdot,t)$ is relatively small around 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**.

4. Main results

In this section, we study the jumping, tracking and escaping properties for online optimization.

4.1. Jumping

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

Theorem 2 (Sufficient conditions for jumping from $h_1(t)$ to $h_2(t)$)**.** *Given a local $I_{t,1}$-minimum trajectory $h_1(t)$ and a local $I_{t,2}$-minimum trajectory $h_2(t)$, suppose that the time-varying function $f(x,t)$ is locally $(\bar{I}_{t,1},c_1,r_1)$-one-point strongly convex around $h_1(t)$ and locally $(\bar{I}_{t,2},c_2,r_2)$-one-point strongly convex around $h_2(t)$ in the region $D_1=\{e\in\mathbb{R}^n:\|e\|\le r_2\}$. Assume that there exist a nonempty time interval $[t_1,t_2]\subset\bar{I}_{t,1}\cap\bar{I}_{t,2}$, a regularization parameter α, a connected subset D_4, and a constant $\theta\in(0,1)$ such that five conditions are satisfied:*

1. *Trajectory of the real root: For each fixed $t\in[t_1,t_2]$, $\nabla_x f(e+h_2(t),t)+\alpha\dot{h}_2(t)=0$ has a real root $\bar{e}(t)$ in the region $D_2=\{e\in\mathbb{R}^n:\|e\|\le\rho\}$, where $\rho<r_2$ and $\bar{e}(t)$ is continuously differentiable for all $t\in[t_1,t_2]$.*

2. *Identifying a positively invariant set: $D_2\cup D_3\subset D_4$, where $D_3=\{e_1\in\mathbb{R}^n:e_1+h_2(t_1)\in\mathcal{B}_v(h_1(t_1))\subseteq RA(h_1(t_1))\}$ and D_4 is a compact positively invariant subset with respect to (15), i.e.,*

$$e_1\in D_4\Longrightarrow e(t,t_1,e_1)\in D_4,\quad \forall t\in[t_1,t_2] \quad (17)$$

3. *One-point strong convexification: The time-varying function $f(e+h_2(t),t)+\alpha\dot{h}_2(t)^\top e$ is locally one-point w-strongly convex around $\bar{e}(t)$ for all $e\in D_4$ and for all $t\in[t_1,t_2]$, i.e.,*

$$\begin{aligned}&\Big(\nabla_x f(e+h_2(t),t)+\alpha\dot{h}_2(t)\Big)^\top\Big(e-\bar{e}(t)\Big)\\ &\ge w\left\|e-\bar{e}(t)\right\|^2,\quad \forall e\in D_4,\quad \forall t\in[t_1,t_2]\end{aligned} \quad (18)$$

   where $w>0$ is a constant.

4. *Upper bound on inertia: $\alpha\le\frac{(r_2-\rho)\theta w}{\sup_{t\in[t_1,t_2]}\left(\left\|\dot{\bar{e}}(t)\right\|\right)}$.*

5. *Lower bound for the time interval: The following inequality holds:*

$$t_2-t_1\ge\frac{\alpha}{w(1-\theta)}\ln\Big(\frac{\|e_1-\bar{e}(t_1)\|}{r_2-\rho}\Big),\quad \forall e_1\in D_3 \quad (19)$$

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

To avoid directly solving for the real roots of $\nabla_x f(e+h_2(t),t)+\alpha\dot{h}_2(t)=0$ and checking the condition (18) for all $t\in[t_1,t_2]$, we propose an approach based on the time-averaged dynamics over a small time interval and named it "small interval averaging"[3]. 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 8. *A function $f_{av}^{h_2}(e)$ is said to be the **average function** of $f(e+h_2(t),t)$ over the time interval $[t_1,t_2]$ if*

$$f_{av}^{h_2}(e)=\frac{1}{t_2-t_1}\int_{t_1}^{t_2}f(e+h_2(\tau),\tau)d\tau \quad (20)$$

[3]Our averaging approach distinguishes from classic averaging methods (Hale, 1980; Khalil, 2002; Teel et al., 1999; Aeyels & Peuteman, 1999) and the partial averaging method (Peuteman & Aeyels, 2002) 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 becuase there is no parameter in (15) that can be taken sufficiently small.

Definition 8 yields that $\nabla_x f_{\text{av}}^{h_2}(e)$ is the average function of $\nabla_x f(e + h_2(\tau), \tau)$. The time-invariant system

$$\dot{e} = -\frac{1}{\alpha}\nabla_x f_{\text{av}}^{h_2}(e) - \frac{h_2(t_2) - h_2(t_1)}{t_2 - t_1} \tag{21}$$

is said to be a **partial interval averaged system** of the time-varying system (15) over the time interval $[t_1, t_2]$ with the perturbation term

$$p(\alpha, e, t) = -\frac{1}{\alpha}\left(\nabla_x f(e + h_2(t), t) - \nabla_x f_{\text{av}}^{h_2}(e)\right) - \left(\dot{h}_2(t) - \frac{h_2(t_2) - h_2(t_1)}{t_2 - t_1}\right) \tag{22}$$

Theorem 3 (Sufficient conditions for jumping from $h_1(t)$ to $h_2(t)$ using averaging)**.** *Given a local $I_{t,1}$-minimum trajectory $h_1(t)$ and a local $I_{t,2}$-minimum trajectory $h_2(t)$, suppose that the time-varying function $f(x,t)$ is locally $(\bar{I}_{t,1}, c_1, r_1)$-one-point strongly convex around $h_1(t)$ and locally $(\bar{I}_{t,2}, c_2, r_2)$-one-point strongly convex around $h_2(t)$ in the region $D_1 = \{e \in \mathbb{R}^n : \|e\| \le r_2\}$. Assume that there exist a nonempty time interval $[t_1, t_2] \subset \bar{I}_{t,1} \cap \bar{I}_{t,2}$, a regularization parameter α, and a connected subset D_4 such that the following five conditions are satisfied:*

1. *Equilibrium point of the averaged system: The system (21) has an equilibrium point $\bar{e}$ in the region $D_2 = \{e \in \mathbb{R}^n : \|e\| \le \rho\}$, where $\rho < r_2$.*
2. *Identifying a positively invariant set: $D_2 \cup D_3 \subseteq D_4$, where $D_3 = \{e_1 \in \mathbb{R}^n : e_1 + h_2(t_1) \in \mathcal{B}_v(h_1(t_1)) \subseteq RA(h_1(t_1))\}$ and D_4 is a compact positively invariant subset with respect to (15), i.e.,*
$$e_1 \in D_4 \Rightarrow e(t, t_1, e_1) \in D_4, \quad \forall t \in [t_1, t_2] \tag{23}$$
3. *One-point strong convexification: The time-invariant function $f_{av}^{h_2}(e) + \frac{\alpha(h_2(t_2) - h_2(t_1))^\top}{t_2 - t_1} e$ is locally one-point w-strongly convex around $\bar{e}$ for all $e \in D_4$, i.e.,*
$$\left(\nabla_x f_{av}^{h_2}(e) + \frac{\alpha(h_2(t_2) - h_2(t_1))}{t_2 - t_1}\right)^\top (e - \bar{e}) \ge w\|e - \bar{e}\|^2, \quad \forall e \in D_4 \tag{24}$$
where $w > 0$ is a constant.
4. *Bound on perturbation: Suppose that for all $t \in [t_1, t_2]$ the perturbation $p(\alpha, e, t)$ satisfies the inequality*
$$\|p(\alpha, e, t)\| \le \delta_1(\alpha, t)\|e - \bar{e}\| + \delta_2(\alpha, t), \tag{25}$$
and there exist some positive constants $\eta_1(\alpha)$ and $\eta_2(\alpha)$ such that
$$\int_{t_1}^{t} \delta_1(\alpha, \tau)d\tau \le \eta_1(\alpha)(t - t_1) + \eta_2(\alpha) \tag{26}$$
5. *Guarantee of convergence within $[t_1, t_2]$: The following inequality holds:*
$$r_2 - \rho \ge \beta_2(\alpha)\Big(\|e_1 - \bar{e}\| e^{-\beta_1(\alpha)(t_2 - t_1)} + \int_{t_1}^{t_2} e^{-\beta_1(\alpha)(t_2 - \tau)}\delta_2(\alpha, \tau)d\tau\Big), \quad \forall e_1 \in D_3 \tag{27}$$
where $\beta_1(\alpha) = \frac{w}{\alpha} - 2\eta_1(\alpha)$ and $\beta_2(\alpha) = \exp(\eta_2(\alpha))$.

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

4.2. Tracking

In this subsection, we study the tracking property of the local minimum trajectory $h_2(t)$. First, notice that if $h_2(t)$ is not constant, the right-hand side of (ODE) is nonzero while the left-hand side is zero. Therefore, $h_2(t)$ is not a solution of (ODE) in general. This is because the solution of (ODE) approximates the continuous limit of a discrete local trajectory of the sequential regularized optimization problem (6). However, to preserve the optimality of the solution with regards to the original time-varying optimization problem without any proximal regularization, it is required to guarantee that the solution of (ODE) is close to $h_2(t)$. The next theorem shows that every local ∞-minimum trajectory can be tracked for a relative small α.

Theorem 4 (Sufficient condition for tracking)**.** *Assume that the time-varying function $f(x,t)$ is locally (∞, c_2, r_2)-one-point strongly convex around $h_2(t)$. Then, $h_2(t)$ can be tracked if α is sufficiently small. In particular, given $0 < \theta' < 1$, $\gamma \coloneqq \sup_{t \ge 0}\left\|\dot{h}_2(t)\right\|, u \coloneqq \frac{\alpha\gamma}{\theta' c_2}$, $\|x_0 - h_2(0)\| \le r_2$ and $\alpha < \frac{c_2\theta' r_2}{\gamma}$, the solution $x(t, t_0, x_0)$ will u-track $h_2(t)$ exponentially with the convergence rate $(1 - \theta')\frac{c_2}{\alpha}$, namely,*

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

4.3. Escaping

Combining Theorem 2 or 3 with Theorem 4 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 5 (Sufficient conditions for escaping from $h_1(t)$ to $h_2(t)$)**.** *Given a local $I_{t,1}$-minimum trajectory $h_1(t)$ and a*

local $I_{t,2}$-minimum trajectory $h_2(t)$, suppose that the time-varying function $f(x,t)$ is locally $(\bar{I}_{t,1}, c_1, r_1)$-one-point strongly convex around $h_1(t)$ and locally $(\bar{I}_{t,2}, c_2, r_2)$-one-point strongly convex around $h_2(t)$ in the region $D_1 = \{e \in \mathbb{R}^n : \|e\| \leq r_2\}$. Let $\gamma = \sup_{t\in\bar{I}_{t,2}} \|\dot{h}_2(t)\|$, $0 < \theta' < 1$, $\mathcal{B}_v(h_1(t_1)) \subseteq RA(h_1(t_1))$ and $u = \frac{\alpha\gamma}{\theta' c_2}$. Under the conditions of Theorem 2 or 3 , if $\alpha < \frac{r_2 c_2 \theta'}{\gamma}$, the solution of (ODE) will (v, r_2)-escape from $h_1(t)$ to $h_2(t)$ after $t \geq t_2$.

4.4. Discussions

Adaptive inertia: To leverage the potential of the time-varying perturbation $\alpha\dot{h}_2(t)^\top e$ in re-shaping the landscape of the objective function to become locally one-point strongly convex over a large region, the regularization parameter α should be selected relatively large. On the other hand, to ensure that the solution of (16) will end up tracking a desirable local (or global) minimum trajectory, Theorem 4 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 objective function $f(x,t)$ has many local minimum trajectories, the solution of (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 2 or 3 are satisfied for $h_i(t)$ and $h_{i+1}(t)$ during each time interval. Then, by sequentially deploying Theorem 2 or 3, it can be concluded that the solution of (ODE) will jump from $h_1(t)$ to $h_m(t)$ after $t \geq t_2^m$. Furthermore, if $h_m(t)$ is tractable with the given α, the solution of (ODE) will escape from $h_1(t)$ to $h_m(t)$ after $t \geq t_2^m$.

5. Numerical Example

Example 2. Consider the non-convex function

$$g(x) = 0.5e + 20e^{-d} - 20e^{-\sqrt{0.5(x_1^2+x_2^2)+d^2}} - 0.5e^{(0.5(\cos(2\pi x_1)+\cos(2\pi x_2)))}. \quad (28)$$

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 3. When $d = 0$, this function is called the Ackley function (Ackley, 1987), which is a benchmark function for global optimization algorithms. To make this function twice continuously differentiable, we take $d = 0.01$. Consider the time-varying objective function $f(x,t) = g(x - z(t))$, where $z(t) = [7\sin(t), 7\cos(t)]^\top$. Two local ∞-minimum trajectories of this online optimization problem are $h_1(t) = [1.95, 0.97]^\top + z(t)$ and $h_2(t) = [0,0]^\top + z(t)$. As shown in the appendix, the conditions of Theorem 5 are all met, and therefore the solution of the corresponding (ODE) will escape from $h_1(t)$ to $h_2(t)$. Furthermore, we have verified for 1000 runs of random initialization over $x_0 - z(0) \in [-5,5] \times [-5,5]$ that all solutions of the corresponding (ODE) will sequentially jump over the local minimum trajectories and end up tracking the global trajectory $[0,0]^\top + z(t)$ after $t \geq 10\pi$.

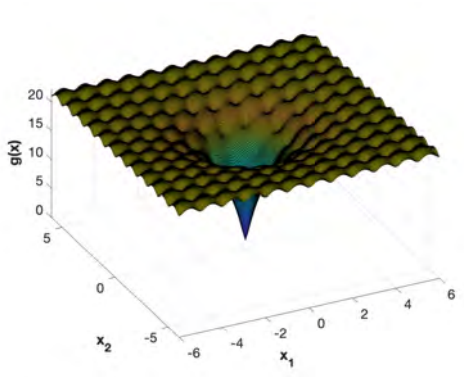


Figure 3. Illustration of Example 2.

6. Conclusion

In this work, we study the landscape of time-varying nonconvex optimization problems. The objective is to understand when simple local search algorithms can find (and track) time-varying global solutions of the problem over time. We introduce a time-varying gradient flow system with controllable inertia as a continuous-time limit of the stationary condition for discretized sequential optimization problems with proximal regularization and online updating scheme. Via a change of variables, the time-varying gradient flow system is regarded as a composition of a time-varying gradient term and a time-varying perturbation term due to the inertia. We show that the time-varying perturbation term due to the inertia re-shapes the landscape by potentially making it one-point strongly convex over a large region during some time interval. We introduce the notions of jumping, tracking and escaping, and use them to develop sufficient conditions under which the time-varying solution jumps 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.

References

Ackley, D. H. *A Connectionist Machine for Genetic Hill-climbing*. Kluwer Academic Publishers, Norwell, MA, USA, 1987. ISBN 0-89838-236-X.

Aeyels, D. and Peuteman, J. On exponential stability of nonlinear time-varying differential equations. *Automatica*, 35(6):1091–1100, 1999.

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

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

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

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

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

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

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

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

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

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

Hale, J. K. Ordinary differential equations. 1980.

Hazan, E., Singh, K., and Zhang, C. Efficient regret minimization in non-convex games. *arXiv preprint arXiv:1708.00075*, 2017.

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

Iserles, A. *A first course in the numerical analysis of differential equations*. Number 44. Cambridge University Press, 2009.

Jadbabaie, A., Rakhlin, A., Shahrampour, S., and Sridharan, K. Online optimization: Competing with dynamic comparators. In *Artificial Intelligence and Statistics*, pp. 398–406, 2015.

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

Josz, C., Ouyang, Y., Zhang, R., Lavaei, J., and Sojoudi, S. A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization. In *Advances in Neural Information Processing Systems*, pp. 2441–2449, 2018.

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

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

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

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

Miller, R. K. and N, M. A. Ordinary differential equations. 1982.

Peuteman, J. and Aeyels, D. Exponential stability of nonlinear time-varying differential equations and partial averaging. *Mathematics of Control, Signals and Systems*, 15(1): 42–70, 2002.

Roy, A., Balasubramanian, K., Ghadimi, S., and Mohapatra, P. Multi-point bandit algorithms for nonstationary online nonconvex optimization. *arXiv preprint arXiv:1907.13616*, 2019.

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

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

Simonetto, A., Mokhtari, A., Koppel, A., Leus, G., and Ribeiro, A. A class of prediction-correction methods for time-varying convex optimization. *IEEE Transactions on Signal Processing*, 64(17):4576–4591, 2016.

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

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016.

Tang, Y., Dvijotham, K., and Low, S. Real-time optimal power flow. *IEEE Transactions on Smart Grid*, 8(6): 2963–2973, 2017.

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

Teel, A. R., Peuteman, J., and Aeyels, D. Semi-global practical asymptotic stability and averaging. *Systems & Control Letters*, 37(5):329–334, 1999.

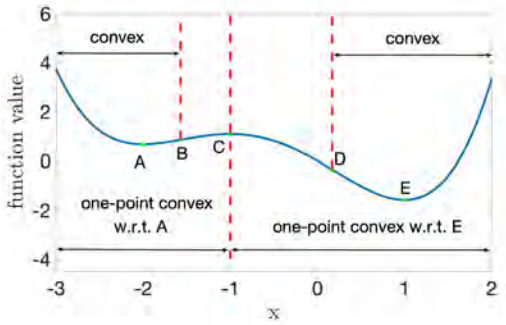
Wibisono, A., Wilson, A. C., and Jordan, M. I. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47): E7351–E7358, 2016.

Zhang, R. Y., Sojoudi, S., and Lavaei, J. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *Journal of Machine Learning Research*, 2019.

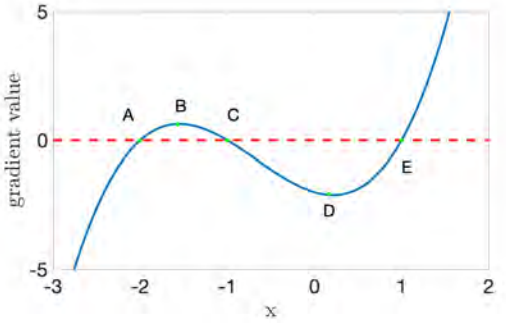
Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pp. 928–936, 2003.

A. One-point strong convexity

Locally one-point strong convexity 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. Figure 4 illustrates the difference between the region of convexity and the region of one-point convexity with respect to a certain point using the time-invariant univariate objective function given in Example 1.



(a) $g(x) = 1/4x^4 + 2/3x^3 - 1/2x^2 - 2x$



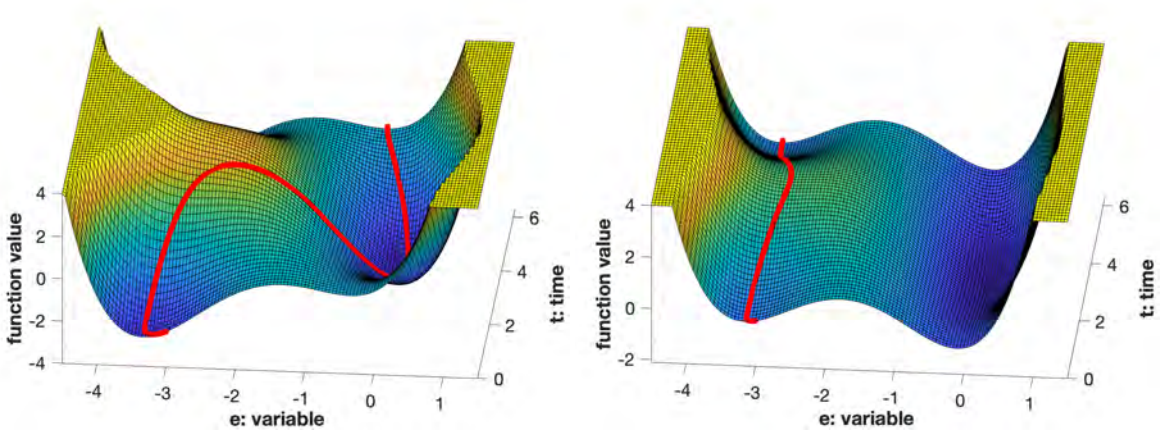
(b) $g'(x) = x^3 + 2x^2 - x - 2$

Figure 4. Illustration of local convexity and one-point convexity for a time-invariant function: Points A and E are the local minimum solutions of $g(x)$; point C is the local maximum solution of $g(x)$; points B and D have a zero second derivative.

B. Inertia creating a one-point strongly convex landscape: A example

Consider again Example 1 and recall that $g(x)$ 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 (16) can be expressed as $\dot{e}(t) = -\frac{1}{\alpha}\nabla_e\Big(g(1 + e(t)) + 5\alpha\cos(t)e(t)\Big)$. The landscape of the new time-varying function $g(1+e)+5\alpha\cos(t)e$ with the variable e is shown for two cases $\alpha = 0.3$ and $\alpha = 0.1$ in Figure 5. The red curves are the solutions of (16) 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 (16) 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, (16) fails to escape the region around $h_1(t)$.

To further inspect the case $\alpha = 0.3$, observe in Figure 6(a) that the landscape of the objective function $g(1 + e) + 1.5\cos(0.9\pi)e$ shows that the region around the spurious



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

Figure 5. Illustration of time-varying landscape after change of variables for Example 1.

local minimum trajectory $h_1(t)$ is one-point strongly convexified with respect to $h_2(t)$ at time $t = 0.9\pi$. This is consistent with the fact that the solution of $\dot{e} = -\frac{1}{0.3}\nabla_x g(1+e) - 5\cos(t)$ starting from $e = -3$ jumps to the neighborhood of 0 around time $t = 0.9\pi$, as demonstrated in Figure 6(c).

Furthermore, if the time interval $[t_1, t_2]$ is relatively large enough to allow transitioning from a neighborhood of $h_1(t)$ to a neighborhood of $h_2(t)$, then the solution of (16) 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 6(b).

C. Proof of existence and uniqueness for (ODE)

Proposition 2. *(Khalil, 2002) Let $f(t,x)$ be piecewise continuous in t and satisfy the Lipschitz condition*

$$\begin{aligned}&\|f(t,x) - f(t,y)\| \le L\|x-y\|,\\ &\forall x, y \in D = \{x \in \mathbb{R}^n : \|x - x_0\| \le r\}, \forall t \in [t_0, t_1]\end{aligned} \tag{29}$$

Then, there exists some $\delta > 0$ such that the state equation $\dot{x} = f(t,x)$ with $x(t_0) = x_0$ has a unique solution over $[t_0, t_0 + \delta]$

Proposition 3. *(Miller & N, 1982) Under the conditions of Proposition 2, there exists a maximal interval $[t_0, T)$ over which the unique solution starting at (t_0, x_0) exists.*

Lemma 1. *Under the conditions of Proposition 2, let $[t_0, T)$ be the maximal interval over which the unique solution starting at (t_0, x_0) exists with $T < \infty$. Let W be any compact subset of D. There exists some $t \in [t_0, T)$ with the property that $x(t) \notin W$.*

Proof. To prove by contradiction, suppose that there is no time t satisfying the stated property. Then, it holds that $x(t) \in W$ for all $t \in [t_0, T)$. It suffices to show that $[t_0, T)$ is not the maximal interval of existence. The solution of $\dot{x} = f(t,x)$ relative to $x(t_1)$ can be written as

$$x(t) = x(t_1) + \int_{t_1}^{t} f(\tau, x(\tau))d\tau, \quad \forall t_1, t \in [t_0, T) \tag{30}$$

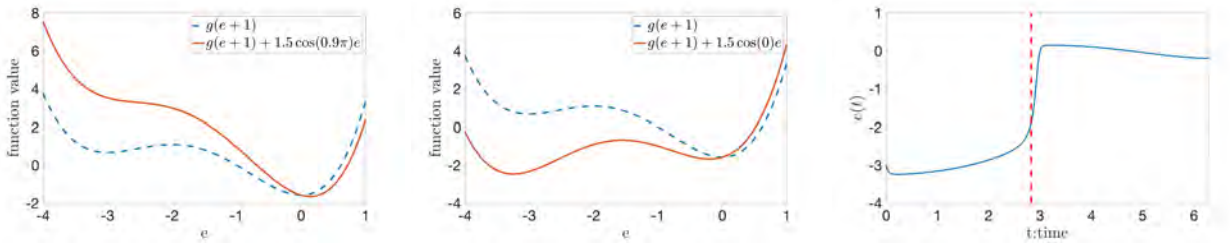
Since $f(t,x)$ is piecewise continuous in t and continuous in x, there exists a constant $M > 0$ such that $\|f(\tau, x(\tau))\| \le M$ for all $\tau \in [t_0, T)$. Thus,

$$\begin{aligned}\|x(t) - x(t_1)\| &= \left\| \int_{t_1}^{t} f(\tau, x(\tau))d\tau \right\| \\ &\le \int_{t_1}^{t} M d\tau \\ &= M(t - t_1)\end{aligned} \tag{31}$$

which implies that $x(t)$ is uniformly continuous on $[t_0, T)$. Then, by the continuous extension theorem, $f(t,x)$ can be defined at the endpoint T in such a way that $f(t,x)$ becomes continuous on $[t_0, T]$. In other words,

$$\begin{aligned}x(T) &= x(t_0) + \lim_{t \to T} \int_{t_0}^{t} f(\tau, x(\tau))d\tau \\ &= x(t_0) + \int_{t_0}^{T} f(\tau, x(\tau))d\tau\end{aligned} \tag{32}$$

Therfore, the solution $x(T)$ is defined and since W is closed, it holds that $x(T) \in W$. Then, it follows from Proposition 2 (applied to the point $(T, x(T))$) that there is a $\delta > 0$ with the property that the solution can be extended to $[t_0, T + \delta]$.



(a) $g(1+e)+1.5\cos(0.9\pi)e$ (b) $g(1+e)+1.5\cos(0)e$ (c) solution of $\dot{e} = -\frac{1}{0.3}\nabla_x g(1+e) - 5\cos(t)$ starting from $e_0 = -3$

Figure 6. Illustration of one-point strong convexification for Example 1.

This contradicts the fact that $[t_0, T)$ is the maximal interval of existence, and completes the proof. □

With the above propositions and lemma, we can show the following result which is similar to the result in (Khalil, 2002).

We now provide the proof for Theorem 1:

Proof. If $f(x,t)$ is piecewise continuous in t and its gradient is locally Lipschitz in x for all $t \geq 0$ and $x \in D \in \mathbb{R}^n$, then $\nabla_x f(t,x)$ satisfies the conditions of Propositions 2-3. It results from Propositions 2-3 that there exists a unique solution for (ODE) over $[t_0, T)$ that is the maximal interval of unique existence. It is enough to show that $T = \infty$. Due to Lemma 1, if the time T is finite, the solution must leave every compact subset of D. However, the solution never leaves the compact set W. This implies that $T = \infty$. □

D. Convergence of (ODE)

Note that (7) can be written in the form of the backward Euler method:

$$x_i^* = x_{i-1}^* - \frac{\tau_i - \tau_{i-1}}{\alpha}\nabla_x f(x_i^*, \tau_i) \tag{33}$$

Furthermore, the online update scheme can be achieved by implementing the forward Euler method to obtain a numerical approximation to the solutions of (ODE):

$$\bar{x}_i^* = \bar{x}_{i-1}^* - \frac{\tau_i - \tau_{i-1}}{\alpha}\nabla_x f(\bar{x}_{i-1}^*, \tau_{i-1}) \tag{34}$$

A direct application of the classical results on convergence of the backward and forward Euler method (Iserles, 2009) immediately shows the following result:

Proposition 4 (Convergence)**.** *Given a local minimum x_0^* of (6a), as the time difference $\Delta_\tau = \tau_{i+1} - \tau_i$ approaches zero, any sequence of discrete local trajectories (x_k^Δ) converges to the (ODE) in the sense that for all fixed $T > 0$:*

$$\lim_{\Delta_\tau \to 0} \max_{0 \leq k \leq \frac{T}{\Delta_\tau}} \left\| x_k^\Delta - x(\tau_k, \tau_0, x_0^*) \right\| = 0 \tag{35}$$

and any sequence of $(\bar{x}_k^\Delta)$ updated by (34) converges to the (ODE) in the sense that for all fixed $T > 0$:

$$\lim_{\Delta_\tau \to 0} \max_{0 \leq k \leq \frac{T}{\Delta_\tau}} \left\| \bar{x}_k^\Delta - x(\tau_k, \tau_0, x_0^*) \right\| = 0 \tag{36}$$

This result shows that the solution of (ODE) starting at a local minimum of (6a) is the continuous limit of the discrete local trajectory of the sequential regularized optimization (6) and the sequence updated by the online scheme (34). For this reason, we only study the continuous-time problem (ODE) in the remainder of this paper.

E. Proof of jumping

Proposition 5 (Comparison lemma)**.** *(Khalil, 2002) Consider the scalar differential equation*

$$\dot{u} = f(t,u), \quad u(t_0) = u_0 \tag{37}$$

where $f(t,u)$ is continuous in t and locally Lipschitz in u for all $t \geq t_0$ and $u \in J \subseteq \mathbb{R}$. Let $[t_0, T)$ (T could be infinity) be the maximal interval of existence of the solution $u(t)$, and suppose that $u(t) \in J$ for all $t \in [t_0, T)$. Let $v(t)$ be a continuous function whose upper right-hand derivative $D^+v(t)$ satisfies the differential inequality

$$D^+v(t) \leq f(t, v(t)), \quad v(t_0) \leq u_0 \tag{38}$$

with $v(t) \in J$ for all $t \in [t_0, T)$. Then, it holds that $v(t) \leq u(t)$ for all $t \in [t_0, T)$.

E.1. Time-varying analysis

We now provide the proof for Theorem 2:

Proof. First, notice that since D_4 is a compact positively invariant set with respect to the dynamics (15), it follows from Theorem 1 that (15) has a unique solution defined for $t \in [t_1, t_2]$ whenever $e_1 \in D_4$. We take a positive semi-definite time-varying function $V(e,t) = \frac{1}{2}\left\|e - \bar{e}(t)\right\|^2$ as

the Lyapunov function for the system (15). The derivative of $V(e)$ along the trajectories of (15) can be expressed as

$$\begin{aligned}\dot{V} =&(e-\bar{e}(t))^\top\Big(-\frac{1}{\alpha}\nabla_x f(e+h_2(t),t)-\dot{h}_2(t)\Big)\\&+(e-\bar{e}(t))^\top\dot{\bar{e}}(t),\quad \forall e\in D_4\\ \le&-\frac{w}{\alpha}\left\|e-\bar{e}(t)\right\|^2+\left\|\dot{\bar{e}}(t)\right\|\left\|e-\bar{e}(t)\right\|,\quad \forall e\in D_4\\ \le&-(1-\theta)\frac{w}{\alpha}\left\|e-\bar{e}(t)\right\|^2-\theta\frac{w}{\alpha}\left\|e-\bar{e}(t)\right\|^2\\&+\sup_{t\in[t_1,t_2]}\Big(\left\|\dot{\bar{e}}(t)\right\|\Big)\left\|e-\bar{e}(t)\right\|,\quad \forall e\in D_4\\ \le&-(1-\theta)\frac{w}{\alpha}\left\|e-\bar{e}(t)\right\|^2,\\ &\forall e\in\left\{e\in D_4:\|e-\bar{e}(t)\|\ge\frac{\alpha\sup_{t\in[t_1,t_2]}\left(\left\|\dot{\bar{e}}(t)\right\|\right)}{\theta w}\right\}\end{aligned}\tag{39}$$

By taking $e_1\in D_3\subset D_4$, since D_4 is a positively invariant set with respect to the dynamics (15) for $t\in[t_1,t_2]$, any trajectory of (15) starting from D_3 will stay in D_4. Thus, the bound in (39) is valid. Let $\delta\coloneqq\sup_{t\in[t_1,t_2]}\left\|\dot{\bar{e}}(t)\right\|$ and $v\coloneqq\frac{\alpha\delta}{\theta w}$. To ensure that the trajectory of (15) enters the time-varying set $\mathcal{B}_{r_2-\rho}=\{e\in\mathbb{R}^n:\|e-\bar{e}(t)\|\le r_2-\rho\}$, it is required to have $\frac{\alpha\delta}{\theta w}\le r_2-\rho$ or $\alpha\le\frac{(r_2-\rho)\theta w}{\delta}$.

Now, it is desirable to show that if the finite time interval $[t_1,t_2]$ is large enough, the solution of (15) will enter the time-varying set $\mathcal{B}_{r_2-\rho}=\{e\in\mathbb{R}^n:\|e-\bar{e}(t)\|\le r_2-\rho\}$ with an exponential convergence rate. Since $\dot{V}$ is negative in $\Gamma=\{e\in D_4:\|e-\bar{e}(t)\|\ge v\}$ and D_4 is a positively invariant set for all $t\in[t_1,t_2]$, a trajectory starting from Γ must stay in D_4 and move in a direction of decreasing $V(e)$. The function $V(e)$ will continue decreasing until the trajectory enters the set $\{e\in D_4:\|e-\bar{e}(t)\|\le v\}$ or until time t_2. Let us show that the trajectory enters $\mathcal{B}_{r_2-\rho}$ 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 $V(e(t),t)=\frac{1}{2}\left\|e-\bar{e}(t)\right\|^2$, (39) can be written as

$$\begin{aligned}&\dot{V}\le-(1-\theta)\frac{2w}{\alpha}V,\\&\forall e\in\Big\{e\in D_4:\|e-\bar{e}(t)\|\ge v\Big\},\forall t\in[t_1,t_2]\end{aligned}\tag{40}$$

By the comparison lemma, $V(\cdot,\cdot)$ satisfies

$$V(e(t),t)\le\exp-(1-\theta)\frac{2w}{\alpha}(t-t_1)V(e_1,t_1)\tag{41}$$

Hence,

$$\left\|e(t)-\bar{e}(t)\right\|\le\exp-(1-\theta)\frac{w}{\alpha}(t-t_1)\left\|e_1-\bar{e}(t_1)\right\|\tag{42}$$

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

E.2. Averaging analysis

To avoid directly solving for the real roots of $\nabla_x f(e+h_2(t),t)+\alpha\dot{h}_2(t)=0$ and checking the condition (18) for all $t\in[t_1,t_2]$, we propose an approach based on the time-averaged dynamics over a small time interval and named it "small interval averaging". This technique guarantees that the solution of the time-varying differential equation (or system) will converge to a residual set of the origin of (15), 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 relatively large enough to allow the transition of points from a neighborhood of $h_1(t)$ to a neighborhood of $h_2(t)$. Therefore, this time interval $[t_1,t_2]$ and its 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 $[t_1,t_2]$.

We now provide the proof for Theorem 3:

Proof. As shown in the proof of Theorem 2, the differential equation (15) has a unique solution defined for $t\in[t_1,t_2]$ whenever $e_1\in D_4$. By using the positive semi-definite function $V(e)=\frac{1}{2}\left\|e-\bar{e}\right\|^2:D_4\to\mathbb{R}$ as the Lyapunov function for the system (15), the derivative of $V(e)$ along the trajectories of (15) can be obtained as

$$\begin{aligned}\dot{V}(e)=&(e-\bar{e})^\top\Big(-\frac{1}{\alpha}\nabla_x f_{\text{av}}^{h_2}(e)-\frac{h_2(t_2)-h_2(t_1)}{t_2-t_1}\\&+p(\alpha,e,t)\Big),\quad\forall e\in D_4\\ \le&-\frac{w}{\alpha}\left\|e-\bar{e}\right\|^2+\delta_1(\alpha,t)\left\|e-\bar{e}\right\|^2\\&+\delta_2(\alpha,t)\left\|e-\bar{e}\right\|,\quad\forall e\in D_4\end{aligned}\tag{43}$$

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

$$\dot{V}(e)\le-\Big[\frac{2w}{\alpha}-2\delta_1(\alpha,t)\Big]V(e)+\delta_2(\alpha,t)\sqrt{2V(e)}\tag{44}$$

To obtain a linear differential inequality, we consider $W(t)=\sqrt{V(e(t))}$. When $V(e(t))\ne0$, it holds that $\dot{W}=\dot{V}/2\sqrt{V}$ and

$$\dot{W}\le-\Big[\frac{w}{\alpha}-\delta_1(\alpha,t)\Big]W+\frac{\delta_2(\alpha,t)}{\sqrt{2}}\tag{45}$$

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

$$\begin{aligned} e(t+\epsilon) =& e(t) + \epsilon\Big(-\frac{1}{\alpha}\nabla_x f_{\text{av}}^{h_2}(e) - \frac{h_2(t_2)-h_2(t_1)}{t_2-t_1} \\ &+ p(\alpha,\bar{e},t)\Big) + o(\epsilon) \\ =& \bar{e} + \epsilon p(\alpha,\bar{e},t) + o(\epsilon) \end{aligned} \tag{46}$$

This implies that

$$\|e(t+\epsilon)-\bar{e}\|^2 = \epsilon^2 \|p(\alpha,\bar{e},t)\|^2 + o(\epsilon^2) \tag{47}$$

Therefore,

$$V(e(t+\epsilon)) = \frac{\epsilon^2}{2}\|p(\alpha,\bar{e},t)\|^2 + o(\epsilon^2) \tag{48}$$

and

$$\begin{aligned} D^+W(t) &= \limsup_{\epsilon\to 0^+} \frac{W(t+\epsilon)-W(t)}{\epsilon} \\ &= \limsup_{\epsilon\to 0^+} \frac{\sqrt{\frac{\epsilon^2}{2}\|p(\alpha,\bar{e},t)\|^2 + o(\epsilon^2)}}{\epsilon} \\ &= \limsup_{\epsilon\to 0^+} \sqrt{\frac{1}{2}\|p(\alpha,\bar{e},t)\|^2 + \frac{o(\epsilon^2)}{\epsilon^2}} \\ &= \frac{1}{\sqrt{2}}\|p(\alpha,\bar{e},t)\| \\ &\le \frac{1}{\sqrt{2}}\delta_2(\alpha,t) \end{aligned} \tag{49}$$

Thus, (45) is also satisfied when $V=0$, and accordingly $D^+W(t)$ satisfies (45) for all values of V. Since W is scalar and the right-hand side of (45) is continuous in t and locally Lipschitz in W for all $t\in[t_1,t_2]$ and $W\ge 0$, the comparison lemma is applicable. In addition, the right-hand side of (45) is linear and a closed-form expression for the solution of the first-order linear differential equation of W can be obtained. Hence, $W(t)$ satisfies

$$W(t) \le \phi(t,t_1)W(t_1) + \frac{1}{\sqrt{2}}\int_{t_1}^{t}\phi(t,\tau)\delta_2(\alpha,\tau)d\tau \tag{50}$$

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

$$\phi(t,t_1) = \exp\Big[-\frac{w}{\alpha}(t-t_1) + \int_{t_1}^{t}\delta_1(\alpha,\tau)d\tau\Big]. \tag{51}$$

Using $W=\sqrt{V}=\frac{1}{\sqrt{2}}\|e-\bar{e}\|$ in (50), we obtain

$$\|e(t)-\bar{e}\| \le \phi(t,t_1)\|e_1-\bar{e}\| + \int_{t_1}^{t}\phi(t,\tau)\delta_2(\alpha,\tau)d\tau \tag{52}$$

Since $\int_{t_1}^{t}\delta_1(\alpha,\tau)d\tau \le \eta_1(\alpha)(t-t_1)+\eta_2(\alpha)$ and using $\beta_1(\alpha)=\frac{w}{\alpha}-\eta_1(\alpha)>0, \beta_2(\alpha)=\exp(\eta_2(\alpha))\ge 1$ in (52), it holds that

$$\begin{aligned} \|e(t)-\bar{e}\| \le& \beta_2(\alpha)\|e_1-\bar{e}\|e^{-\beta_1(\alpha)(t-t_1)} \\ &+ \beta_2(\alpha)\int_{t_1}^{t} e^{-\beta_1(\alpha)(t-\tau)}\delta_2(\alpha,\tau)d\tau \end{aligned} \tag{53}$$

By taking $e_1\in D_3\subset D_4$, since D_4 is a positively invariant set with respect to the dynamics (15) for $t\in[t_1,t_2]$, any trajectory of (15) starting from D_3 will stay in D_4. Thus, the bound in (53) is valid. If t_2 satisfies

$$\begin{aligned} &\beta_2(\alpha)\|e_1-\bar{e}\|e^{-\beta_1(\alpha)(t_2-t_1)} \\ &+ \beta_2(\alpha)\int_{t_1}^{t_2} e^{-\beta_1(\alpha)(t_2-\tau)}\delta_2(\alpha,\tau)d\tau \le r_2-\rho \end{aligned} \tag{54}$$

then $\|e(t_2)-\bar{e}\| \le r_2-\rho$. Since $\bar{e}\in D_2$, we have $\|e(t_2)\|\le r_2$. This shows that the solution of (15) jumps from $h_1(t)$ to $h_2(t)$ during the time interval $[t_1,t_2]$. □

E.3. Discussions

Remark 1. *Consider a special case where $f(x,t)=g(x-z(t))$ such that $z(t):[0,\infty)\to\mathbb{R}^n$ is a continuous differentiable function. Suppose that $g(\cdot)$ has two local minima z_1^* and z_2^*. The online optimization $f(\cdot,t)$ has two isolated minimum trajectories $h_1(t)=z_1^*+z(t)$ and $h_2(t)=z_2^*+z(t)$. The time-varying system after the change of variables $x(t)=e+z_2^*+z(t)$ becomes*

$$\dot{e} = -\frac{1}{\alpha}\nabla_x g(e+z_2^*) - \dot{z}(t) \tag{55}$$

and its partial interval averaged system over the time interval $[t_1,t_2]$ becomes

$$\dot{e} = -\frac{1}{\alpha}\nabla_x g(e+z_2^*) - \frac{z(t_2)-z(t_1)}{t_2-t_1} \tag{56}$$

By selecting $\eta_1(\alpha)=0, \eta_2(\alpha)=0$ and $\delta_2(\alpha,t)=\left\|\dot{z}(t)-\frac{z(t_2)-z(t_1)}{t_2-t_1}\right\| := \delta_2(t)$, the condition (27) reduces to

$$\|e_1-\bar{e}\|e^{-\frac{w}{\alpha}(t_2-t_1)} + \int_{t_1}^{t_2} e^{-\frac{w}{\alpha}(t_2-\tau)}\delta_2(\tau)d\tau \le r_2-\rho \tag{57}$$

which can be relaxed to the simple condition

$$\Big(\|e_1-\bar{e}\| - \frac{\alpha}{w}\Big)e^{-\frac{w}{\alpha}(t_2-t_1)} + \frac{\alpha}{w}\sup_{t\in[t_1,t_2]}\delta_2(t) \le r_2-\rho \tag{58}$$

Remark 2. *In Theorem 3, 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 (15) *has enough time to jump, the time interval* $[t_1, t_2]$ *should not be very small. This trade-off is reflected in* (27) *and* (57)*. In addition, although the estimation of the convergence time in* (27) *and* (57) *may be conservative, the nature of the exponential convergence rate due to the locally one-point strongly convex condition would enable a fast jumping of the solution of* (15) *during* $[t_1, t_2]$.

F. Proof of tracking

First, consider the case when the maximal time interval of $h_2(t)$ is the entire time horizon $[t_0, \infty]$. If the solution of (15) can be shown to be in a small residual set around 0, then it is guaranteed that $x(t, t_0, x_0)$ tracks its nearby local minimum trajectory. Notice that (15) can be regarded as a time-varying perturbation of the system

$$\dot{e} = -\frac{1}{\alpha}\nabla_x f(e + h_2(t), t), \quad \forall t \geq t_0 \tag{59}$$

Since $h_2(t)$ is a local minimum trajectory, it is obvious that $e(t) \equiv 0$ is an equilibrium point of (59). In addition, since $f(e+h_2(t), t)$ is locally (∞, c_2, r_2)-one-point around $h_2(t)$, the stability property of $e = 0$ for (59) can be proved with the help of the proposition below.

Proposition 6. *(Khalil, 2002) Let* $e = 0$ *be an equilibrium point for* $\dot{e} = f(e, t)$ *and* $D = \{e \in \mathbb{R}^n : \|e\| \leq r_2\}$. *Let* $V : [0, \infty) \times D \to \mathbb{R}$ *be a continuously differentiable function such that*

$$k_1 \|e\|^p \leq V(t, e) \leq k_2 \|e\|^p \tag{60}$$

$$\frac{\partial V}{\partial t} + \frac{\partial V}{\partial x} f(t, x) \leq -k_3 \|e\|^p \tag{61}$$

for all $t \geq 0$ *and* $e \in D$, *where* k_1, k_2 *and* k_3 *are positive constants. Then,* $e = 0$ *is exponentially stable.*

Lemma 2. *If* $f(e+h_2(t), t)$ *is locally* (∞, c_2, r_2)*-one-point strongly convex around* $h_2(t)$ *in the region* $D = \{e \in \mathbb{R}^n : \|e\| \leq r_2\}$, *then* $e = 0$ *is a locally exponentially stable equilibrium point of* (59).

Proof. We take a positive semi-definite function $V(e) = \frac{1}{2}\|e\|^2 : D \to \mathbb{R}$ as the Lyapunov function. The derivative of $V(e)$ along the trajectories of (ODE) satisfies

$$\begin{aligned} \dot{V} &= e^\top \Big(-\frac{1}{\alpha}\nabla_x f(e + h_2(t), t)\Big) \\ &\leq -\frac{c}{\alpha}\|e\|^2 \end{aligned} \tag{62}$$

Then, the conditions in Proposition 6 are satisfied for $V = \frac{1}{2}\|e\|^2, p = 2, k_1 = k_2 = \frac{1}{2}$ and $k_3 = \frac{c}{\alpha}$. As a result, $e = 0$ is a locally exponentially stable equilibrium point of (59). □

Since the system (59) has an exponentially stable equilibrium point at $e = 0$, one would expect that the solution of the time-varying perturbed system (15) stays in a small residual set of $e = 0$ if the perturbation $\dot{h}_2(t)$ is relatively small. The perturbation $\dot{h}_2(t)$ being small is equivalent to α being small. The next theorem shows that every local ∞-minimum trajectory can be tracked for a sufficiently small α.

We now provide the proof for Theorem 4:

Proof. Consider the positive semi-definite function $V(e) = \frac{1}{2}\|e\|^2 : D \to \mathbb{R}$ as the Lyapunov function for the system (15), where $D = \{e \in \mathbb{R}^n : \|e\| \leq r_2\}$. The derivative of $V(e)$ along the trajectories of (15) can be written as

$$\begin{aligned} \dot{V} &= e^\top \Big(-\frac{1}{\alpha}\nabla_x f(e + h_2(t), t) - \dot{h}_2(t)\Big), \quad \forall t \geq t_0 \\ &\leq -\frac{c_2}{\alpha}\|e\|^2 + \gamma\|e\|, \quad \forall t \geq t_0 \\ &= -(1-\theta')\frac{c_2}{\alpha}\|e\|^2 - \theta'\frac{c_2}{\alpha}\|e\|^2 + \gamma\|e\|, \quad \forall t \geq t_0 \\ &\leq -(1-\theta')\frac{c_2}{\alpha}\|e\|^2, \quad \forall \|e\| \geq \frac{\alpha\gamma}{\theta' c_2}, \forall t \geq t_0 \end{aligned} \tag{63}$$

We aim to show that if $u \coloneqq \left(\frac{\alpha\gamma}{\theta' c_2}\right) < r_2$ or $\alpha < \frac{c_2\theta' r_2}{\gamma}$, the set D has the property that any trajectory starting in D at t_0 enters the set $\mathcal{B}_u(0) = \{e \in \mathbb{R}^n : \|e\| \leq u\}$ with an exponential convergence rate. Since the derivative $\dot{V}$ is negative on the boundaries ∂D and $\partial\mathcal{B}_u(0)$, (63) implies that the sets D and $\mathcal{B}_u(0)$ are positively invariant. Since D is also a compact set, it follows from Theorem 1 that (15) has a unique solution defined for all $t \geq t_0$ whenever $e_0 \in D$.

Since $\dot{V}$ is negative in $\Gamma = \{e \in \mathbb{R}^n : u \leq \|e\| \leq r_2\}$, any trajectory starting in Γ must move in a direction of decreasing $V(e)$, leading to the property that the function $V(e)$ will continue decreasing until the trajectory enters the set $\mathcal{B}_u(0)$ in finite time and stays therein for all future times. Let us show that the trajectory enters $\mathcal{B}_u(0)$ with an exponential convergence rate. Since $V(e) = \frac{1}{2}\|e\|^2$, (63) can be written as

$$\dot{V} \leq -(1-\theta')\frac{2c_2}{\alpha}V, \quad \|e\| \geq u, \quad \forall t \geq t_0 \tag{64}$$

By the comparison lemma, V satisfies

$$V(e(t)) \leq \exp -(1-\theta')\frac{2c_2}{\alpha}(t - t_0)V(e(0)) \tag{65}$$

Hence,

$$\|e(t)\| \leq \exp -(1-\theta')\frac{c_2}{\alpha}(t - t_0)\|e(0)\| \tag{66}$$

This inequality holds over the interval $[t_0, t_0 + \frac{\alpha}{c_2(1-\theta')}\ln(\frac{r_2}{u})]$ during which $\|e\| \geq u$. Since $\mathcal{B}_u(0)$

is a positively invariant set, $e(t,t_0,e_0)$ will stay in $\mathcal{B}_u(0)$ for all future times. By the change of variables $x(t,t_0,e_0) = e(t,t_0,e_0) + h_2(t)$, we have

$$\begin{aligned} &x(t,t_0,x_0) \in \mathcal{B}_u(h_2(t)) \subseteq RA(h_2(t)), \\ &\forall t \geq t_0 + \frac{\alpha}{c_2(1-\theta')} \ln(\frac{r_2}{u}) \end{aligned} \tag{67}$$

if $x_0 \in \mathcal{B}_{r_2}(h_2(t_0))$. This completes the proof. □

F.1. Concentration of trajectories

Furthermore, if the region $\mathcal{B}_u(h_2(t))$ lies inside the region of strong convexity of $f(x,t)$ around $h_2(t)$, we can show that the solutions of (15) starting from any points in $\mathcal{B}_{r_2}(h_2(t))$ will converge to each other.

Definition 9. *It is said that $f(x,t)$ is **locally** $(\bar{I}_t, \bar{c}, \bar{r})$ **strongly convex** around the local I_t-minimum trajectory $h(t)$ if there exist a constant $\bar{c} > 0$ and a region $\bar{D} = \{e \in \mathbb{R}^n : \|e\| \leq \bar{r}\}$ such that*

$$\nabla_{xx} f(e + h(t), t) \geq \bar{c}, \quad \forall e \in \bar{D}, \quad \forall t \in \bar{I}_t \tag{68}$$

*where $\bar{I}_t \subset I_t$ if I_t is a finite interval and $\bar{I}_t = I_t = [t_0, \infty]$ otherwise. The region $\bar{D} = \{e \in \mathbb{R}^n : \|e\| \leq \bar{r}\}$ is called the **region of locally** $(\bar{I}_t, c, r)$ **strong convexity** around $h(t)$.*

Theorem 6 (Sufficient conditions for contraction)**.** *Assume that the conditions of Theorem 4 are satisfied, and that the time-varying function $f(x,t)$ is locally $(\infty, \bar{c}_2, \bar{r}_2)$ strongly convex around $h_2(t)$. Then, for all points x_0^1 and x_0^2 such that $\|x_0^1 - h_2(0)\| \leq r_2$ and $\|x_0^2 - h_2(0)\| \leq r_2$, the solutions $x(t,t_0,x_0^1)$ and $x(t,t_0,x_0^2)$ will converge to each other in the sense that*

$$\lim_{t\to\infty} \|x(t,t_0,x_0^1) - x(t,t_0,x_0^2)\| = 0 \tag{69}$$

if $\alpha < \frac{c_2\theta'\bar{r}_2}{\gamma}$.

Proof. Under the conditions of Theorem 4 and the condition that $\|x_0^1 - h_2(0)\| \leq r_2$, $\|x_0^2 - h_2(0)\| \leq r_2$, it holds that $\|x(t,t_0,x_0^1) - h_2(t)\| \leq u$ and $\|x(t,t_0,x_0^2) - h_2(t)\| \leq u$ for $t > t' := t_0 + \frac{\alpha}{c_2(1-\theta')} \ln(\frac{r_2}{u})$. If $u \leq \bar{r}_2 \leq r_2$ or $\alpha \leq \frac{c_2\bar{r}_2\theta'}{\gamma}$, we obtain $\nabla_{xx} f(x,t) \leq -\frac{\bar{c}_2}{\alpha} < 0$ for $x \in \mathcal{B}_u(h(t))$ and $t \geq t'$. By denoting $x(t) = x(t,t_0,x_0^1)$ and $z(t) = x(t,t_0,x_0^2)$, the system (ODE) governing these two solutions can be written as

$$\dot{x}(t) = -\frac{1}{\alpha}\nabla_x f(x,t) \tag{70a}$$

$$\dot{z}(t) = -\frac{1}{\alpha}\nabla_z f(z,t) \tag{70b}$$

Applying the mean value theorem to the above equations yields that

$$\dot{x}(t) - \dot{z}(t) = -\frac{1}{\alpha}\Big(\nabla_x f(x,t) - \nabla_z f(z,t)\Big) \tag{71a}$$

$$= -\frac{1}{\alpha}\nabla_{yy} f(y,t)(x(t) - z(t)) \tag{71b}$$

where $y(t) = \lambda(t)x(t) + (1-\lambda(t))z(t)$ for some $0 \leq \lambda(t) \leq 1$. By multiplying $(x(t) - z(t))$ to the both sides of (71b), we arrive at

$$\frac{d}{dt}\|x(t) - z(t)\|^2 = -\frac{2}{\alpha}\nabla_{yy} f(y,t)\|x(t) - z(t)\|^2 \tag{72}$$

Since after $t > t'$, $x(t) \in \mathcal{B}_u(h_2(t))$, $z(t) \in \mathcal{B}_u(h_2(t))$, and $\mathcal{B}_u(h_2(t))$ is a convex set for each fixed t, we have $y(t) \in \mathcal{B}_u(h_2(t))$ for $t > t'$. This implies that $\nabla_{yy} f(y,t) \leq \bar{c}_2$. Then, the solution of (72) satisfies

$$\|x(t) - z(t)\|^2 \leq e^{-\frac{2\bar{c}_2 t}{\alpha}}\|x(t') - z(t')\|^2 \tag{73}$$

Therefore, $\lim_{t\to\infty}\|x(t) - z(t)\| = 0$ □

F.2. Temporary tracking

Next, we consider the case when $h_2(t)$ is define only over a finite maximal time interval $I_{t,2}$. The system after the change of variables $x(t,t_0,x_0) = e(t,t_0,e_0) + h_2(t)$ is defined on a finite time interval:

$$\dot{e} = -\frac{1}{\alpha}\nabla_x f(e + h_2(t), t) - \dot{h}_2(t), \quad \forall t \in I_t \tag{74}$$

In this case, the notion of uniformly asymptotic stability is not well-defined. However, the following result on temporary tracking can be developed.

Theorem 7 (Sufficient conditions for temporary tracking)**.** *Assume that the time-varying function $f(x,t)$ is locally $(\bar{I}_{t,2}, c_2, r_2)$-one-point strongly convex around $h_2(t)$. Let $t_1 = \liminf \bar{I}_{t,2}$ and $t_2 = \limsup \bar{I}_{t,2}$. Given $0 < \theta' < 1$, $\gamma := \sup_{t\in\bar{I}_{t,2}} \|\dot{h}_2(t)\|, u := \frac{\alpha\gamma}{\theta' c_2}$, $\|x(t_1,t_0,x_0) - h_2(t_1)\| \leq r_2$ and $\alpha < \frac{c_2\theta' r_2}{\gamma}$, the solution $x(t,t_0,x_0)$ will temporarily r_2-track $h_2(t)$. In addition, if $t_2 - t_1 > \frac{\alpha}{c_2(1-\theta')}\ln(\frac{r_2}{u})$, the solution $x(t,t_0,x_0)$ will temporarily u-track $h_2(t)$ exponentially with the convergence rate $(1-\theta')\frac{c_2}{\alpha}$, namely,*

$$\begin{aligned} &\forall t_1 \leq t \leq t_1 + \frac{\alpha}{c_2(1-\theta')}\ln(\frac{r_2}{u}): \\ &\|x(t,t_0,x_0) - h_2(t)\| \leq r_2 \exp -(1-\theta')\frac{c_2}{\alpha}(t - t_1), \\ &\forall t_1 + \frac{\alpha}{c_2(1-\theta')}\ln(\frac{r_2}{u}) < t \leq t_2: \\ &\|x(t,t_0,x_0) - h_2(t)\| \leq u. \end{aligned} \tag{75}$$

Proof. Consider the positive semi-definite function $V(e) = \frac{1}{2}\|e\|^2 : D \to \mathbb{R}$ as the Lyapunov function for the system (74), where $D = \{e \in \mathbb{R}^n : \|e\| \leq r_2\}$. Similar to the inequality (63), the derivative of $V(e)$ along the trajectories of (74) satisfies

$$\dot{V} =\leq -(1-\theta')\frac{c_2}{\alpha}\|e\|^2, \quad \forall \|e\| \geq \frac{\alpha\gamma}{\theta' c_2}, \quad \forall t \in \bar{I}_t \tag{76}$$

First, we show that if $u \coloneqq \frac{\alpha\gamma}{\theta' c_2} < r_2$ or $\alpha < \frac{c_2\theta' r_2}{\gamma}$, the set D has the property that any trajectory starting in D at t_1 stays in the set D for all $t \in \bar{I}_{t,2}$. Notice that since the derivative $\dot{V}$ is negative on the boundary ∂D, (63) implies that the set D is positively invariant. Since D is also a compact set, it follows from Theorem 1 that (15) has a unique solution defined for all $t \in \bar{I}_{t,2}$ whenever $e_1 \coloneqq x_1 - h(t_1) \in D$. Then, the set D being positively invariant implies that

$$x(t, t_1, x_1) \in \mathcal{B}_{r_2}(h_2(t)) \subseteq RA(h_2(t)), \quad \forall t \in \bar{I}_{t,2} \tag{77}$$

By choosing $x(t_1, t_0, x_0) = x_1$, one can conclude that $x(t, t_0, x_0)$ will temporarily r_2-track $h_2(t)$. Next, we show that if the finite time interval $\bar{I}_{t,2}$ is large enough, the solution of (74) will enter the set $\mathcal{B}_u(0) = \{e \in \mathbb{R}^n : \|e\| \leq u\}$ with an exponential convergence rate and stays in $\mathcal{B}_u(0)$ for all future times. Since $\dot{V}$ is negative in $\Gamma = \{e \in \mathbb{R}^n : u \leq \|e\| \leq r_2\}$ for all $t \in \bar{I}_{t,2}$, a trajectory starting from Γ must move in a direction of decreasing $V(e)$ and the function $V(e)$ will continue decreasing until the trajectory enters the set $\mathcal{B}_u(0)$ or until time t_2. The fact that the trajectory enters $\mathcal{B}_u(0)$ before t_2 if $t_2 - t_1 > \frac{\alpha}{c_2(1-\theta')}\ln(\frac{r_2}{u})$ is based on the same argument used in the proof of Theorem 4. □

F.3. Discussions

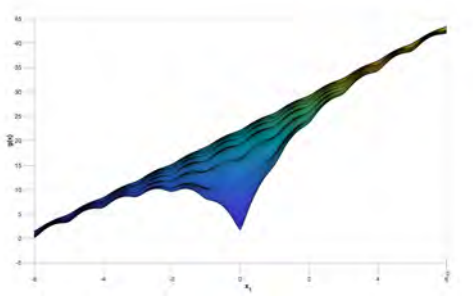
Remark 3. *$u \coloneqq \frac{\alpha\gamma}{\theta' c_2}$ is the ultimate bound of difference between $x(t, t_0, x_0)$ and the local minimum trajectory $h_2(t)$. The smaller the regularization parameter α is, the closer $x(t, t_0, x_0)$ to the local minimum trajectory $h_2(t)$ is.*

Remark 4. *For a fixed ultimate bound u, the convergence rate $(1-\theta')\frac{c_2}{\alpha}$ shows that $x(t, t_0, x_0)$ converges faster to $\mathcal{B}_u(h_2(t))$ as the regularization parameter α reduces.*

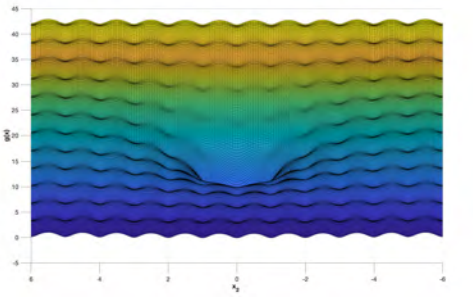
Remark 5. *In the case that the local minimum trajectory $h_2(t)$ is a constant, the upper bound on α simply becomes $\alpha < \infty$. This implies that if the local minimum trajectory $h_2(t)$ is constant, then it will be perfectly tracked with any regularization parameter and can not be escaped by tuning the regularization parameter.*

G. Details on Example 2

Two local ∞-minimum trajectories of this online optimization problem are $h_1(t) = [1.95, 0.97]^\top + z(t)$ and $h_2(t) = [0, 0]^\top + z(t)$. It can be observed in Figures 7(a) and 7(b) that, around time $t = 0$, the time-varying objective function around a neighborhood of $h_1(0)$ is one-point strongly convexified with respect to $h_2(0)$. Thus, one could expect that the solution of (ODE) would jump from $h_1(t)$ to $h_2(t)$. More formally, it can be shown that $g(x)$ is locally $(3.3, 1.1)$-one-point strongly convex with respect to the origin, which implies that $f(x, t)$ is locally $(\infty, 3.3, 1.1)$-one-point strongly convex around $h_2(t)$. To ensure that the solution of (ODE) will track $h_2(t)$, we need to take $\alpha < \frac{c_2 r_2 \theta'}{\sup_{t\geq 0}\|\dot{z}(t)\|}$ for $0 < \theta < 1$. In this case, $\alpha = 0.5$ simply satisfies the tracking condition. Then, by the change of variables $x = e + h_2(t)$, the differential equation (15) can be written as



(a) Landscape of $f(e + h_2(0)) + \alpha\dot{h}_2(0)e$ from the perspective of e_1



(b) Landscape of $f(e + h_2(0)) + \alpha\dot{h}_2(0)e$ from the perspective of e_2

Figure 7. Illustration of Example 2.

$$\dot{e}(t) = \begin{bmatrix} -20e^{-\sqrt{0.5(e_1^2+e_2^2)+d^2}}\frac{e_1}{\sqrt{0.5(e_1^2+e_2^2)+d^2}} \\ -\pi e^{(0.5(\cos(2\pi e_1)+\cos(2\pi e_2)))}\sin(2\pi e_1), \\ -20e^{-\sqrt{0.5(e_1^2+e_2^2)+d^2}}\frac{e_2}{\sqrt{0.5(e_1^2+e_2^2)+d^2}} \\ -\pi e^{(0.5(\cos(2\pi e_1)+\cos(2\pi e_2)))}\sin(2\pi e_2) \end{bmatrix} - \begin{bmatrix} 7\cos(t) \\ -7\sin(t) \end{bmatrix} \tag{78}$$

By selecting the time interval $[0, \frac{\pi}{8}]$, the averaged system

can be obtained as

$$\dot{e}(t) = \begin{bmatrix} -20e^{-\sqrt{0.5(e_1^2+e_2^2)+d^2}} \frac{e_1}{\sqrt{0.5(e_1^2+e_2^2)+d^2}} \\ -\pi e^{(0.5(\cos(2\pi e_1)+\cos(2\pi e_2)))} \sin(2\pi e_1), \\ -20e^{-\sqrt{0.5(e_1^2+e_2^2)+d^2}} \frac{e_2}{\sqrt{0.5(e_1^2+e_2^2)+d^2}} \\ -\pi e^{(0.5(\cos(2\pi e_1)+\cos(2\pi e_2)))} \sin(2\pi e_2) \end{bmatrix} - \begin{bmatrix} \frac{56}{\pi} \sin(\frac{\pi}{8}) \\ \frac{56}{\pi}(\cos(\frac{\pi}{8}) - 1) \end{bmatrix} \tag{79}$$

This system has an equilibrium point at $[-0.0034, 0.0007]^\top$. Then, Condition 1 in Theorem 3 is met with $\rho = 0.01$. Let $D_1 = \mathcal{B}_{1.1}(0)$, $D_2 = \mathcal{B}_{0.01}(0)$, $D_3 = \{e \in \mathbb{R}^n : e_1 + h_2(t_1) \in \mathcal{B}_{0.1}(h_1(t_1))$ and $D_4 = [-0.2, 2.1] \times [-0.1, 1.1]$. It follows that $D_2 \cup D_3 \subseteq D_4$. In addition, on the boundary points $e_1 = 2.1$ and $e_1 = -0.2$, the derivative of e_1 along the trajectory of (78) is negative and positive, respectively, for all $e_2 \in [-0.1, 1.1]$ and $t \in [0, \frac{\pi}{8}]$. Similarly, on the boundary points $e_2 = 1.1$ and $e_2 = -0.1$, the derivative of e_2 along the trajectory of (78) is negative and positive, respectively, for all $e_1 \in [-0.2, 2.1]$ and $t \in [0, \frac{\pi}{8}]$. This implies that D_4 is a positively invariant set with respect to (78) for $t \in [0, \frac{\pi}{8}]$. This shows that Condition 2 in Theorem 3 is also met. Furthermore, (24) and (57) are satisfied for $w = 1.3$. Thus, the conditions of Theorem 5 are all met, and therefore the solution of (78) will $(0.1, 1.1)$-escape from $h_1(t)$ to $h_2(t)$. Furthermore, we have verified for 1000 runs of random initialization over $x_0 - z(0) \in [-5, 5] \times [-5, 5]$ that all solutions of (78) will sequentially jump over the local minimum trajectories and end up tracking the global trajectory $[0, 0]^\top + z(t)$ after $t \geq 10\pi$.