The Landscape of Deterministic and Stochastic Optimal Control Problems: One-Shot Optimization Versus Dynamic Programming

Jihun Kim, Yuhao Ding, Yingjie Bi, and Javad Lavaei, *Fellow, IEEE*

***Abstract*—Optimal control problems can be solved via a one-shot (single) optimization or a sequence of optimization using dynamic programming (DP). However, the computation of their global optima often faces NP-hardness, and thus only locally optimal solutions may be obtained at best. In this work, we consider the discrete-time finite-horizon optimal control problem in both deterministic and stochastic cases and study the optimization landscapes associated with two different approaches: one-shot and DP. In the deterministic case, we prove that each local minimizer of the one-shot optimization corresponds to some control input induced by a locally minimum control policy of DP, and vice versa. However, with a parameterized policy approach, we prove that deterministic and stochastic cases both exhibit the desirable property that each local minimizer of DP corresponds to some local minimizer of the one-shot optimization, but the converse does not necessarily hold. Nonetheless, under different technical assumptions for deterministic and stochastic cases, if there exists only a single locally minimum control policy, one-shot and DP turn out to capture the same local solution. These results pave the way to understand the performance and stability of local search methods in optimal control.**

***Index Terms*—Dynamic programming (DP), landscape, one-shot optimization, optimal control.**

I. Introduction

DYNAMIC programming (DP) has been widely used in a variety of fields with a rich theoretical foundation and many mathematical and algorithmic aspects [2], [3]. One classic area of DP is to solve optimal control problems, with applications in communication systems [4], inventory control [5], powertrain control [6], and many more. Furthermore, many recent successes in artificial intelligence, especially in reinforcement learning (RL) [7], [8], are also deeply rooted in DP. In the challenging domain of classic Atari 2600 games, the work [9] has demonstrated that the Q-learning method based on the generalized policy iteration along with a deep neural network as the function approximator for the Q-values surpasses the performance of all previous algorithms and achieves a level comparable with that of a professional human games tester.

Despite a strong theoretical framework of DP, the exact solutions of large-scale optimal control problems are often impossible to obtain using DP in practice [7]. Apart from suffering the "curse of dimensionality" when the state space is large, solving DP accurately could also be highly complex. The reason is that DP requires solving optimization subproblems to global optimality, and the computation of their global optima is NP-hard in general, due to the nonlinearity of the dynamics and the nonconvexity of the cost function.

Therefore, although DP theory relies on global optimization solvers, practitioners routinely use local optimization solvers based on first- and second-order numerical algorithms. As a result, the theoretical guarantee of DP could break down as soon as a nonglobal local solution is found in any of the subproblems. Understanding the performance of local search methods for nonconvex problems has been a focal area in machine learning in recent years. This is performed under the notion of spurious solution, which refers to a local minimum that is not a global solution. The specific application areas are neural networks [10], [11], deep learning [12], [13], mixtures of regressions [14], [15], matrix sensing/recovery [16], [17], [18], [19], phase retrieval [15], [20], and online optimization [21], [22].

Recently, there has been an increasing interest in understanding the global convergence of exact or approximate DP algorithms in policy gradient methods for RL, such as projected policy gradient, natural policy gradient, and mirror descent with or without regularizers [23], [24], [25], [26], [27]. Prior to them, the work [28] identified some general algorithm-independent properties of the policy gradient method by establishing a direct connection between policy gradient (one-shot) and policy iteration (DP) objectives. They showed that the global convergence of the policy gradient method is guaranteed if the policy iteration objectives have no suboptimal stationary points. However, the literature lacks a rigorous analysis of the spurious solutions of the DP method.

In this article, we analyze the spurious solutions of the DP method by focusing on the following fundamental question: *What if the globally optimal solution of each subproblem of DP is replaced with a solution obtained by a local search method?* A challenge in this analysis is that policy optimization even toward the spurious solutions can be problematic if the action space is continuous [29]. One can think of the policy iteration with function approximation [30] where the Q-function

Manuscript received 5 April 2024; accepted 9 June 2024. Date of publication 17 June 2024; date of current version 5 December 2024. This work was supported in part by ARO, in part by ONR, in part by AFOSR, in part by NSF, and in part by the UC Noyce Initiative. Recommended by Associate Editor Eric C. Kerrigan. *(Corresponding author: Javad Lavaei.)*

The authors are with the Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720 USA (e-mail: jihun.kim@berkeley.edu; yuhao_ding@berkeley.edu; yb236@cornell.edu; lavaei@berkeley.edu).

Digital Object Identifier 10.1109/TAC.2024.3415459

approximation error is zero. This is a reasonable assumption since a close-to-zero error can be obtained with a sufficiently rich and expressive policy class, such as deep neural networks, which naturally yields the existence of the local minimizer of DP. That motivates our analysis on the comparison between the solutions of one-shot method and DP if they are only solved to the spurious local minimizers, and hence, our algorithm-agnostic study offers a clear understanding on the landscapes for the optimal control problem without considering the secondary issue of the approximation error.

We focus on both deterministic and stochastic discrete-time finite-horizon optimal control problems whose goal is to find an optimal input sequence minimizing the total cost subject to the dynamics. One approach to solving the problem is by formulating it as a one-shot optimization problem, a single entire-period problem, and another approach is using the DP to formulate it as a sequential decision-making problem with multiple single-period subproblems and solve it in a backward way. Although it is known that the one-shot method and the DP method return the same globally optimal control sequence for the deterministic optimal control problem [3], it is not yet known what would occur if the global optimizer needed for solving each suboptimization problem in DP is replaced by a local optimizer. In our work, we compare the two optimization landscapes: one induced by the DP method based on local search algorithms, and the other induced by its corresponding one-shot optimization based on local search methods.

Contribution and Outline: The rest of this article is organized as follows. We investigate the two landscapes holistically for the following three types of control systems.

1) In Section II, we first study *deterministic* systems under a nonparameterized policy. We introduce the notion of locally minimum control policy of DP and prove that under some mild conditions, each (spurious) local minimizer of the one-shot optimization corresponds to the control input induced by a (spurious) locally minimum control policy of DP, and vice versa. This indicates that DP with local search can successfully solve the optimal control problem to global optimality if and only if the one-shot problem is free of spurious solutions.
2) In Section III, we analyze *deterministic* systems under a *parameterized* policy. The necessity to study this problem arises in RL algorithms, where the control policy used by DP is parameterized by neural networks or other means. Thus, we generalize the results of Section II to optimization with respect to the parameters rather than the control inputs themselves. We prove that each local minimizer of DP corresponds to some local minimizer of the one-shot optimization, whereas its converse may not hold. Moreover, we show that if there exists only a single locally minimum control policy with a specific parameterized policy class, namely, a linear combination of independent basis functions, each local minimizer of the one-shot optimization corresponds to a local minimizer of DP.
3) In Section IV, we extend the result to *stochastic* systems under a *parameterized* policy. The stochasticity brings up the challenge to handle an uncountable number of realizations of random variables. We show that surprisingly a similar relationship in the deterministic parameterized problem holds. For both cases, we conclude that the optimization landscape of the one-shot problem is more complex than its DP counterpart in terms of the number of spurious solutions. This implies that if the one-shot problem has a *low* complexity, so does the DP problem. Another result says that if the DP problem has a *very low* complexity, the same holds for the one-shot problem. In this article, our notion of "complexity" of an optimization problem is based on *the number of spurious local minima*. For example, convex optimization problems have very low complexity in light of having no spurious solutions. However, problems with an exponential number of spurious solutions are hard to solve [31]. Note that a reformulation of an optimization problem via a change of variables does not normally change the number of local minima, which justifies why the number of spurious solutions can serve as a complexity measure.

Finally, Section V concludes this article. Table I summarizes the main results of this article.

TABLE I
THEOREMS AND THE CORRESPONDING ASSUMPTIONS WITH RESULTS

Assumptions \ Theorems		Deterministic			Deterministic + Parameterized			Stochastic + Parameterized		
		1	2	3	4	5	6	7	8	9
Convex	action space	○								
	parameter space						○			
Policy class	C^1		○			○			○	
	C^2	○								
	Defined by Definition 13						○			○
	contains a single locally minimum control policy						○			○
Interior policy			○							○
Strict local minimizer				○			○			
Continuous Random state										○
Large parameter space							○			
Result	DP to one-shot	○			○			○		
	DP to one-shot (stationarity)		○			○			○	
	one-shot to DP			○			○			○

In various applications arising in machine learning and model-free approaches for which the model is unknown and simulations are expensive, DP is the only viable choice compared with the one-shot optimization approach. Hence, it is essential to understand when DP combined with a local search solver works. The results of this article explain that the success of DP is closely related to the optimization landscape of a single optimization problem. For instance, the success of the DP method highly depends on the number of spurious solutions of the one-shot optimization problem.

Notations: Let $\mathbb{R}$ denote the set of real numbers. We use $B(c,r)$ to denote the open ball centered at c with radius r, and use $\bar{B}(c,r)$ to denote the closure of $B(c,r)$. The notation $x \in A \setminus B$ means that x is in the set A but not in the set B. Let $\|\cdot\|$ denote the Euclidean norm and $\|\cdot\|_F$ denote the Frobenius norm. Let $\nabla_x f(x,y)$ denote the gradient of $f(x,y)$ with respect to x, and $\nabla_x^2 f(x)$ denote the Hessian of $f(x)$. For the matrix K, $K \succ 0$ means that K is positive definite. The notation C^n means that the function is n-times continuously differentiable. The notation $\mathbb{E}$ denotes the expectation operator.

II. Deterministic Problem

A. Problem Formulation

Consider a general discrete-time finite-horizon deterministic optimal control problem with n time steps

$$\begin{aligned} \min_{u_0,\dots,u_{n-1}\in A} \quad & \sum_{i=0}^{n-1} c_i(x_i,u_i)+c_n(x_n) \\ \text{s.t.} \quad & x_{i+1}=f_i(x_i,u_i), \quad i=0,\dots,n-1 \qquad \text{(P1)} \\ & x_0 \text{ is given} \end{aligned}$$

where $x_i \in \mathbb{R}^N$ is the state at time i, and u_i is the control input at time i that is constrained to be in an action space $A \subseteq \mathbb{R}^M$. The state transition is governed by the dynamics $f_i : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^N$. Each time instance i is associated with a stage cost $c_i : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$ or the terminal cost $c_n : \mathbb{R}^N \to \mathbb{R}$. Given an initial state x_0, the goal of the optimal control problem is to find an optimal control input $(u_0,\dots,u_{n-1})$ minimizing the sum of the stage costs and the terminal cost. In this article, the dynamics f_i and the cost functions c_i are assumed to be at least twice continuously differentiable over $\mathbb{R}^N \times \mathbb{R}^M$, and the action space A is assumed to be compact.

The optimal control problem can be solved by two common approaches. The first approach directly solves (P1) as a one-shot optimization problem that simultaneously solves for all variables. To simplify the analysis, we eliminate the equality constraints in (P1) via the notation $C(x_k;u_k,\dots,u_{n-1})$ defined as the cost-to-go started at the time step k with the initial state x and control inputs $u_k,\dots,u_{n-1}$. In other words

$$\begin{aligned} C(x) &= c_n(x) \\ C(x;u_k,\dots,u_{n-1}) &= c_k(x,u_k) \\ &\quad + C(f_k(x,u_k);u_{k+1},\dots,u_{n-1}) \end{aligned}$$

for $k=0,\dots,n-1$. The one-shot optimization problem (P1) can be equivalently written as

$$\min_{u_0,\dots,u_{n-1}\in A} \quad C(x_0;u_0,\dots,u_{n-1}). \qquad \text{(P2)}$$

The second approach to solving the optimal control problem is based on DP. Let $J_k(x_k)$ denote the optimal cost-to-go at the time step k with the initial state x_k, i.e.,

$$J_k(x_k)=\min_{u_k,\dots,u_{n-1}\in A} \quad C(x_k;u_k,\dots,u_{n-1}).$$

Then, J_k can be computed in a backward fashion from the time step $n-1$ to time 0 through the following recursion:

$$\begin{aligned} J_n(x) &= c_n(x) \\ J_k(x) &= \min_{u\in A}\{c_k(x,u)+J_{k+1}(f_k(x,u))\} \qquad \text{(P3)} \end{aligned}$$

for $k=0,\dots,n-1$. It is worth noting that (P3) yields a set of *functions* that solve the problem for all initial states, whereas (P1) produces a *vector* specific to a given x_0. The optimal cost $J_0(x_0)$ equals the optimal objective value of (P1).

However, due to the nonconvexity of the function, it is generally NP-hard to obtain globally optimal solutions of (P3) for all states and at all times. Specifically, when using the DP to solve the optimal control problem (P1), the first step is to compute $\min_{u\in A}\{c_{n-1}(x_{n-1},u)+c_n(f_{n-1}(x_{n-1},u))\}$ for every $x_{n-1}\in\mathbb{R}^N$, which requires solving nonconvex optimization problems if the cost function or the dynamic is nonconvex. Since these intermediate problems are normally solved via local search methods, the best expectation is to obtain a local minimizer for u_{n-1} as a function of $x\in\mathbb{R}^N$, denoted by the policy $\pi_{n-1}(x)$. As a result, instead of working with truly optimal cost-to-go functions, one may arrive at a suboptimal cost-to-go at time $n-1$ as follows:

$$\begin{aligned} J^\pi_{n-1}(x_{n-1}) &= c_{n-1}(x_{n-1},\pi_{n-1}(x_{n-1})) \\ &\quad + c_n(f_{n-1}(x_{n-1},\pi_{n-1}(x_{n-1}))) \end{aligned}$$

which is obtained based on the local minimizer $\pi_{n-1}(x)$. Subsequently, it is required to solve the optimal decision-making problem $\min_{u\in A}\{c_{n-2}(x_{n-2},u)+J^\pi_{n-1}(f_{n-2}(x_{n-2},u))\}$ for every $x_{n-2}\in\mathbb{R}^N$. By repeating this procedure in a backward fashion toward the time step 0, we obtain a group of policy functions π_k and suboptimal cost-to-go functions J^π_k for $k=0,\dots,n-1$. Given the initial state x_0, let

$$\begin{aligned} & u_0^*=\pi_0(x_0),\ x_1^*=f_0(x_0,u_0^*),\ u_1^*=\pi_1(x_1^*),\ x_2^*=f_1(x_1^*,u_1^*) \\ & \qquad \dots \\ & u_{n-1}^*=\pi_{n-1}(x_{n-1}^*), \quad x_n^*=f_n(x_{n-1}^*,u_{n-1}^*) \end{aligned}$$

be the control inputs and the states induced by the policies $\pi_0,\dots,\pi_{n-1}$. Then, $(u_0^*,\dots,u_{n-1}^*)$ is a suboptimal solution to the original optimal control problem (P1) with the suboptimal objective value $J^\pi_0(x_0)$. This motivates us to define locally minimum control policies based on solving (P3) to local optimality.

Definition 1: Given a control policy $\pi=(\pi_0,\dots,\pi_{n-1})$, the associated Q-functions $Q^\pi_k(\cdot,\cdot)$ and cost-to-go functions $J^\pi_k(\cdot)$ under the policy π are defined in a backward way from the time step $n-1$ to 0 through the following recursion:

$$\begin{aligned} & J^\pi_n(x)=c_n(x) \\ & Q^\pi_k(x,u)=c_k(x,u)+J^\pi_{k+1}(f_k(x,u)), \quad k=0,\dots,n-1 \\ & J^\pi_k(x)=Q^\pi_k(x,\pi_k(x)), \quad k=0,\dots,n-1. \end{aligned}$$

Definition 2 (Local minimizer): A vector $(u_0^*,\dots,u_{n-1}^*)$ is said to be a local minimizer of the one-shot optimization problem (P2) if there exists $\epsilon>0$ such that

$$C(x_0,u_0^*,\dots,u_{n-1}^*)\le C(x_0,\tilde{u}_0,\dots,\tilde{u}_{n-1})$$

for all $\tilde{u}_i\in B(u_i^*,\epsilon)\cap A$, where $i=0,\dots,n-1$. It is further called a spurious (nonglobal) local minimizer of the one-shot optimization problem if $C(x_0,u_0^*,\dots,u_{n-1}^*)>J_0(x_0)$.

Definition 3 (Locally minimum control policy): A control policy $\pi=(\pi_0,\dots,\pi_{n-1})$ is said to be a locally minimum control policy of DP if for all $k\in\{0,\dots,n-1\}$ and for all $x\in\mathbb{R}^N$, the policy $\pi_k(x)$ is a local minimizer of the Q-function $Q^\pi_k(x,\cdot)$, meaning that there exists $\epsilon_k^*(x)>0$ such that

$$Q^\pi_k(x,\pi_k(x))\le Q^\pi_k(x,\tilde{u}) \quad \forall\tilde{u}\in B(\pi_k(x),\epsilon_k^*(x))\cap A.$$

It is further called a spurious locally minimum control policy of DP if $J^\pi_0(x_0)>J_0(x_0)$.

In the following sections, we will show that in the deterministic problem, both approaches capture the same local solutions under mild assumptions.

B. Local Minimizers: From DP to One-Shot Optimization

It is well known that the input sequence induced by a globally minimal control policy is a global minimizer of the one-shot

problem [3]. In this section, we will show that the input sequence induced by a spurious locally minimum control policy of DP also corresponds to a spurious local minimizer of the one-shot problem if some mild conditions are satisfied.

Theorem 1: Assume that A is convex. Consider a (spurious) locally minimum control policy $\pi = (\pi_0, \dots, \pi_{n-1})$, and let the corresponding input and state sequences associated with the initial state x_0 be denoted as $(u_0^*, \dots, u_{n-1}^*)$ and $(x_0^*, \dots, x_n^*)$, respectively. If π_k is twice continuously differentiable in a neighborhood of x_k^* and $\nabla_u^2 Q_k^\pi(x_k^*, u_k^*) \succ 0$ for all $k \in \{0, \dots, n-1\}$, then $(u_0^*, \dots, u_{n-1}^*)$ is also a (spurious) local minimizer of the one-shot problem.

Proof: First, we will use induction to find positive numbers $\delta_0, \dots, \delta_n$ and $\epsilon_0, \dots, \epsilon_{n-1}$ such that

$$\nabla_u^2 Q_k^\pi(x, u) \succ 0 \tag{1}$$

$$\pi_k(x) \in B(u_k^*, \epsilon_k) \tag{2}$$

$$f_k(x, u) \in B(x_{k+1}^*, \delta_{k+1}) \tag{3}$$

for every $x \in B(x_k^*, \delta_k)$, $u \in B(u_k^*, \epsilon_k) \cap A$, and $k \in \{0, \dots, n-1\}$. At the base step $k = n$, we choose an arbitrary $\delta_n > 0$. At the induction step, since f_k is continuous and $\nabla_u^2 Q_k^\pi$ is continuous at (x_k^*, u_k^*), there exist $\delta_k > 0$ and $\epsilon_k > 0$ such that both (1) and (3) are satisfied for all $x \in B(x_k^*, \delta_k)$ and $u \in B(u_k^*, \epsilon_k) \cap A$. Moreover, as π_k is continuous at x_k^*, (2) will be satisfied by further reducing δ_k.

For every $(\tilde{u}_0, \dots, \tilde{u}_{n-1})$ with $\tilde{u}_k \in B(u_k^*, \epsilon_k) \cap A$, let $(\tilde{x}_0, \dots, \tilde{x}_n)$ be its corresponding state sequence (note that $\tilde{x}_0 = x_0$). It follows from (3) that $\tilde{x}_k \in B(x_k^*, \delta_k)$ for all $k \in \{0, \dots, n-1\}$, which together with (2) implies that:

$$\pi_k(\tilde{x}_k) \in B(u_k^*, \epsilon_k) \quad \forall k \in \{0, \dots, n-1\}.$$

In light of (1), $Q_k^\pi(\tilde{x}_k, \cdot)$ is a convex function on the convex set $B(u_k^*, \epsilon_k) \cap A$. Because $\pi_k(\tilde{x}_k) \in B(u_k^*, \epsilon_k) \cap A$ is a local minimizer of the function $Q_k^\pi(\tilde{x}_k, \cdot)$, it must be a global minimizer of this function over $B(u_k^*, \epsilon_k) \cap A$. Thus, for $k \in \{0, \dots, n-1\}$, we have

$$\begin{aligned} c_k(\tilde{x}_k, \tilde{u}_k) + J_{k+1}^\pi(\tilde{x}_{k+1}) = Q_k^\pi(\tilde{x}_k, \tilde{u}_k) &\geq Q_k^\pi(\tilde{x}_k, \pi_k(\tilde{x}_k)) \\ &= J_k^\pi(\tilde{x}_k). \end{aligned}$$

By adding all of the above inequalities, one can obtain

$$C(x_0; \tilde{u}_0, \dots, \tilde{u}_{n-1}) \geq J_0^\pi(x_0) = C(x_0; u_0^*, \dots, u_{n-1}^*)$$

which shows that $(u_0^*, \dots, u_{n-1}^*)$ is a local minimizer of the one-shot problem. Also, if π is a spurious locally minimum control policy of DP, namely, $J_0^\pi(x_0) > J_0(x_0^*)$, then

$$C(x_0; u_0^*, \dots, u_{n-1}^*) = J_0^\pi(x_0) > J_0(x_0).$$

As a result, $(u_0^*, \dots, u_{n-1}^*)$ is also a spurious local minimizer of the one-shot problem. ■

Remark 1: By taking the contrapositive, one can immediately conclude that the DP method cannot produce any spurious locally minimum control policies that satisfy the regularity conditions in Theorem 1 as long as the one-shot problem has no spurious local minima.

C. Stationary Points: From DP to One-Shot Optimization

In this section, we will show that the induced controlled input of a locally minimum control policy of DP corresponds to a stationary point of the one-shot problem, under some conditions milder than the assumptions of Theorem 1.

Definition 4: Given a set S and a continuously differentiable function g, a point $s^* \in S$ is said to be a stationary point of the optimization problem $\min_{s \in S} g(s)$ if

$$-\nabla_s g(s^*) \in \mathcal{N}_S(s^*)$$

where $\mathcal{N}_S(s^*)$ denotes the normal cone of the set S at the point s^* [32].

We branch off into two specific notions of stationarity below.

Definition 5 (Stationary point): A vector of control inputs $(u_0^*, \dots, u_{n-1}^*)$ is said to be a stationary point of the one-shot optimization if for all $k \in \{0, \dots, n-1\}$, it holds that $-\nabla_{u_k} C(x_0; u_0^*, \dots, u_{n-1}^*) \in \mathcal{N}_A(u_k^*)$.

Definition 6 (Stationary control policy): A control policy $\pi = (\pi_0, \dots, \pi_{n-1})$ is said to be a stationary control policy of DP if for all $k \in \{0, \dots, n-1\}$ and for all $x \in \mathbb{R}^N$, it holds that $-\nabla_u Q_k^\pi(x, \pi_k(x)) \in \mathcal{N}_A(\pi_k(x))$.

Now, we will prove that a stationary control policy (which involves a locally minimum control policy) implies a stationary point of the one-shot optimization under mild assumptions. Let $\mathbf{D}_k^\pi(x)$ be the Jacobian matrix of $\pi_k(\cdot)$ at point x, $\mathbf{D}_k^{f,x}(x, u)$ be the Jacobian matrix of the function $f_k(\cdot, u)$ at point x while viewing u as a constant, and $\mathbf{D}_k^{f,u}(x, u)$ be the Jacobian matrix of $f_k(x, \cdot)$ at point u while viewing x as a constant.

Theorem 2: Consider a stationary control policy $\pi = (\pi_0, \dots, \pi_{n-1})$, and let the associated input and state sequences with the initial state x_0 be denoted as $(u_0^*, \dots, u_{n-1}^*)$ and $(x_0^*, \dots, x_n^*)$, respectively. If for every $k \in \{0, \dots, n-1\}$, the following holds:

1) π_k is continuously differentiable in a neighborhood of x_k^*;
2) either $\pi_k(x_k^*)$ is in the interior of A or $\mathbf{D}_k^\pi(x_k^*) = 0$;

then $(u_0^*, \dots, u_{n-1}^*)$ is a stationary point of the one-shot optimization.

Proof: First, we will apply induction to prove that

$$\nabla_x J_k^\pi(x_k^*) = \nabla_x C(x_k^*; u_k^*, \dots, u_{n-1}^*) \tag{4}$$

holds for $k \in \{0, \dots, n\}$. The base step $k = n$ is obvious. For the induction step, observe that

$$\begin{aligned} \nabla_x Q_k^\pi(x, u) &= \nabla_x c_k(x, u) + \mathbf{D}_k^{f,x}(x, u)^T \nabla_x J_{k+1}^\pi(f_k(x, u)) \\ \nabla_x J_k^\pi(x) &= \nabla_x [Q_k^\pi(x, \pi_k(x))] \\ &= \nabla_x Q_k^\pi(x, \pi_k(x)) + \mathbf{D}_k^\pi(x)^T \nabla_u Q_k^\pi(x, \pi_k(x)). \end{aligned}$$

Therefore

$$\begin{aligned} \nabla_x J_k^\pi(x_k^*) = \nabla_x c_k(x_k^*, u_k^*) &+ \mathbf{D}_k^{f,x}(x_k^*, u_k^*)^T \nabla_x J_{k+1}^\pi(x_{k+1}^*) \\ &+ \mathbf{D}_k^\pi(x_k^*)^T \nabla_u Q_k^\pi(x_k^*, u_k^*). \end{aligned} \tag{5}$$

If u_k^* is in the interior of A, we have $\nabla_u Q_k^\pi(x_k^*, u_k^*) = 0$ by stationarity. Otherwise, by the assumption, we have $\mathbf{D}_k^\pi(x_k^*) = 0$. In either case, the last term of (5) is zero. Meanwhile

$$\begin{aligned} &\nabla_x C(x; u_k^*, \dots, u_{n-1}^*) \\ &= \nabla_x c_k(x, u_k^*) + \nabla_x [C(f_k(x, u_k^*); u_{k+1}^*, \dots, u_{n-1}^*)] \\ &= \nabla_x c_k(x, u_k^*) + \mathbf{D}_k^{f,x}(x, u_k^*)^T \\ &\quad \times \nabla_x C(f_k(x, u_k^*); u_{k+1}^*, \dots, u_{n-1}^*). \end{aligned}$$

Now, (4) can be obtained by taking $x = x_k^*$ in the above equality and then combining it with the induction hypothesis and (5). Finally, for $k \in \{0, \dots, n-1\}$, one can write

$$\begin{aligned}
&\nabla_{u_k} C(x_0; u_0^*, \dots, u_{n-1}^*) \\
&= \nabla_u c_k(x_k^*, u_k^*) + \mathbf{D}_k^{f,u}(x_k^*, u_k^*)^T \nabla_x C(x_{k+1}^*; u_{k+1}^*, \dots, u_{n-1}^*) \\
&= \nabla_u c_k(x_k^*, u_k^*) + \mathbf{D}_k^{f,u}(x_k^*, u_k^*)^T \nabla_x J_{k+1}^\pi(x_{k+1}^*) \\
&= \nabla_u Q_k^\pi(x_k^*, u_k^*)
\end{aligned}$$

in which the second equality is due to (4). Since u_k^* is a stationary point of $Q_k^\pi(x_k^*, \cdot)$, $-\nabla_u Q_k^\pi(x_k^*, u_k^*) \in \mathcal{N}_A(u_k^*)$. Thus, $-\nabla_{u_k} C(x_0; u_0^*, \dots, u_{n-1}^*) \in \mathcal{N}_A(u_k^*)$, which proves that $(u_0^*, \dots, u_{n-1}^*)$ is a stationary point of the one-shot optimization. ■

D. Local Minimizers: From One-Shot Optimization to DP

In this section, we will show that each strict local minimizer of the one-shot problem is induced by a locally minimum control policy π of DP. Before proving the theorem, we first provide the following useful lemma.

Lemma 1: Given a function $g : \mathbb{R}^N \times A \to \mathbb{R}$, a point $x^* \in \mathbb{R}^N$, and a number $\epsilon > 0$, if $u^* \in A$ is a strict local minimizer of the function $g(x^*, \cdot)$ and g is continuous in a neighborhood of (x^*, u^*), then there exist $\delta > 0$ and a function $h : B(x^*, \delta) \to A$ such that $h(x^*) = u^*$ and that the following statements hold for all $x \in B(x^*, \delta)$.

1) $h(x)$ is a local minimizer of $g(x, \cdot)$.
2) $h(x) \in B(u^*, \epsilon)$.
3) The function $g(x, h(x))$ is continuous at x.

Proof: The proof is given in [1] (see Lemma 1). ■

Theorem 3: If the one-shot problem has a (spurious) strict local minimizer $(u_0^*, \dots, u_{n-1}^*)$, then there exists a (spurious) locally minimum control policy π of DP with the property that $\pi_k(x_k^*) = u_k^*$ for all $k \in \{0, \dots, n-1\}$, where $(x_0^*, \dots, x_n^*)$ is the state sequence associated with the (spurious) solution of the one-shot problem.

Proof: Let $(u_0^*, \dots, u_{n-1}^*)$ be a strict local minimizer of the one-shot problem. There exists $\epsilon > 0$ such that

$$C(x_0; u_0^*, \dots, u_{n-1}^*) < C(x_0; u_0, \dots, u_{n-1}) \tag{6}$$

for every control sequence $(u_0, \dots, u_{n-1}) \neq (u_0^*, \dots, u_{n-1}^*)$ with the property that $u_i \in B(u_i^*, \epsilon) \cap A$ for $i = 0, \dots, n-1$. In what follows, we will prove by a backward induction that there exist policies $\pi_0, \dots, \pi_{n-1}$, positive numbers $\delta_0, \dots, \delta_n$, and corresponding cost-to-go functions $J_0^\pi, \dots, J_n^\pi$ such that they jointly satisfy the following properties.

1) $\pi_k(x_k)$ is a local minimizer of the function $Q_k^\pi(x_k, \cdot)$ for all $x_k \in \mathbb{R}^N$.
2) $\pi_k(x_k^*) = u_k^*$.
3) For all $x_k \in B(x_k^*, \delta_k)$, it holds that
$$\pi_k(x_k) \in B(u_k^*, \epsilon), \quad f_k(x_k, \pi_k(x_k)) \in B(x_{k+1}^*, \delta_{k+1}).$$
4) J_k^π is lower semicontinuous on $\mathbb{R}^N$ and continuous on $B(x_k^*, \delta_k)$.

For the base step $k = n$, we choose an arbitrary $\delta_n > 0$ and notice that $J_n^\pi(x) = c_n(x)$, implying that J_n^π is always continuous. For $k < n$, assume that $\pi_{k+1}, \dots, \pi_{n-1}$ and $\delta_{k+1}, \dots, \delta_n$ with the above properties have been found.

First, by the continuity of f_k, there exist $\delta_k' > 0$ and $0 < \epsilon_k < \epsilon$ such that

$$f_k(x_k, u_k) \in B(x_{k+1}^*, \delta_{k+1}) \quad \forall (x_k, u_k) \in S_k \tag{7}$$

where $S_k = B(x_k^*, \delta_k') \times (B(u_k^*, \epsilon_k) \cap A)$.

Since $Q_k^\pi(x_k, u_k) = c_k(x_k, u_k) + J_{k+1}^\pi(f_k(x_k, u_k))$ and J_{k+1}^π is continuous on $B(x_{k+1}^*, \delta_{k+1})$, Q_k^π is continuous on S_k. Next, for every $\tilde{u}_k \in B(u_k^*, \epsilon_k) \cap A$, if we define

$$\begin{aligned}
&\tilde{x}_{k+1} = f_k(x_k^*, \tilde{u}_k), \quad \tilde{u}_{k+1} = \pi_{k+1}(\tilde{x}_{k+1}) \\
&\tilde{x}_{k+2} = f_{k+1}(\tilde{x}_{k+1}, \tilde{u}_{k+1}), \quad \tilde{u}_{k+2} = \pi_{k+2}(\tilde{x}_{k+2}) \\
&\qquad \dots \\
&\tilde{x}_{n-1} = f_{n-2}(\tilde{x}_{n-2}, \tilde{u}_{n-2}), \quad \tilde{u}_{n-1} = \pi_{n-1}(\tilde{x}_{n-1})
\end{aligned}$$

by applying (7) and then the third property above repeatedly, we arrive at

$$\tilde{u}_i \in B(u_i^*, \epsilon) \cap A \quad \forall i \in \{k+1, \dots, n-1\}.$$

When $\tilde{u}_k \neq u_k^*$, it follows from (6) and the second property above that:

$$\begin{aligned}
&Q_k^\pi(x_k^*, \tilde{u}_k) = C(x_k^*; \tilde{u}_k, \dots, \tilde{u}_{n-1}) \\
&= C(x_0; u_0^*, \dots, u_{k-1}^*, \tilde{u}_k, \dots, \tilde{u}_{n-1}) - \sum_{i=0}^{k-1} c_i(x_i^*, u_i^*) \\
&> C(x_0; u_0^*, \dots, u_{n-1}^*) - \sum_{i=0}^{k-1} c_i(x_i^*, u_i^*) \\
&= C(x_k^*; u_k^*, \dots, u_{n-1}^*) = Q_k^\pi(x_k^*, u_k^*).
\end{aligned}$$

As a result, u_k^* is a strict local minimizer of $Q_k^\pi(x_k^*, \cdot)$. Applying Lemma 1 to the function Q_k^π with x_k^* and ϵ_k, one can find $0 < \delta_k < \delta_k'$ and a function $h_k : B(x_k^*, \delta_k) \to A$ such that $h_k(x_k^*) = u_k^*$ and that the following statements hold for every $x_k \in B(x_k^*, \delta_k)$.

1) $h_k(x_k)$ is a local minimizer of $Q_k^\pi(x_k, \cdot)$.
2) $h_k(x_k) \in B(u_k^*, \epsilon_k) \subseteq B(u_k^*, \epsilon)$, which together with (7) implies that $f_k(x_k, h_k(x_k)) \in B(x_{k+1}^*, \delta_{k+1})$.
3) The function $Q_k^\pi(x_k, h_k(x_k))$ is continuous at x_k.

Let π_k be the extension of the function h_k by setting $\pi_k(x_k)$ to be any global minimizer of the lower semicontinuous function $Q_k^\pi(x_k, \cdot)$ over the compact set A if $x_k \notin B(x_k^*, \delta_k)$. Obviously, π_k satisfies the first three properties. To verify the last property, observe that

$$J_k^\pi(x_k) = \begin{cases} Q_k^\pi(x_k, h_k(x_k)), & \text{if } x_k \in B(x_k^*, \delta_k) \\ H_k(x_k), & \text{otherwise} \end{cases}$$

in which $H_k(x_k) = \min_{u \in A} Q_k^\pi(x_k, u)$, and therefore, J_k^π is continuous on the set $B(x_k^*, \delta_k)$. In addition, note that J_{k+1}^π and thus Q_k^π are lower semicontinuous, while A is compact. Hence, it follows from the Berge maximum theorem [33] that H_k is also lower semicontinuous on $\mathbb{R}^N$, which implies that J_k^π is lower semicontinuous on $\mathbb{R}^N \setminus \bar{B}(x_k^*, \delta_k)$. For every point $\bar{x}_k$ on the boundary of $B(x_k^*, \delta_k)$, since H_k is lower semicontinuous at $\bar{x}_k$, for every $\bar{\epsilon} > 0$, there exists $\bar{\delta} > 0$ such that

$$J_k^\pi(x_k) \geq H_k(x_k) > H_k(\bar{x}_k) - \bar{\epsilon} = J_k^\pi(\bar{x}_k) - \bar{\epsilon}$$

holds for all $x_k \in B(\bar{x}_k, \bar{\delta})$. Therefore, J_k^π is also lower semicontinuous at $\bar{x}_k$.

By the first and second properties, $\pi = (\pi_0, \ldots, \pi_{n-1})$ is a locally minimum control policy of DP. Also, if $(u_0^*, \ldots, u_{n-1}^*)$ is a spurious local minimizer of the one-shot problem, then $J_0^\pi(x_0) = C(x_0; u_0^*, \ldots, u_{n-1}^*) > J_0(x_0)$, which implies that π is also a spurious locally minimum control policy of DP. ■

Remark 2: Theorem 3 shows that, under mild conditions, DP is a reformulation from a single one-shot optimization problem to a sequence of optimization problems that preserves local minimizers. By taking the contrapositive of Theorem 3, one can immediately obtain the result that the one-shot problem has no spurious strict local minimizers as long as DP has no spurious locally minimum control policies.

Remark 3: Pontryagin's minimum principle implies that a global minimizer of the one-shot problem achieves a global optimality of each DP problem minimizing Hamiltonian. One can restrict the domain to apply the principle to a local minimizer of the one-shot problem; it achieves a local optimality of u_k^* for each DP problem if J_k^π are evaluated at the associated state $x_k^\star$. Theorem 3 is a generalization of Pontryagin's principle in the sense that from each local minimizer of the one-shot problem, we obtain a locally minimum control policy instead of u_k^*; i.e., a set of *functions* that achieves a local optimality of every DP problem for all $x \in \mathbb{R}^N$. We further require a "strict" local minimizer of the one-shot problem to ensure that a local optimality is obtained at all points in the neighborhood of $x_k^\star$. Meanwhile, one can now anticipate that Theorem 1 would correspond to the converse of Pontryagin's principle. The principle provides sufficient conditions for the one-shot problem if we have a convex action space, convex cost functions, and linear dynamics [3], [34]. In contrast, Theorem 1 assumes a convex action space but still has general nonlinear transition dynamics. Theorem 1 instead requires a locally "strictly" convex Q-functions (Hamiltonian) for each DP subproblem. The connection between our results and Pontryagin's principle suggests the possibility of the extension of the above results to the continuous-time setting.

Remark 4: In fact, all results of this article can be naturally generalized to the continuous-time setting, but the analysis is left as future work due to space restrictions. To outline the pathway for generalization, note that the Hamilton–Jacobi–Bellman equation for a given continuous-time system can be obtained from developing a discrete-time model, obtaining the Bellman equation for that model, and then closing the gap between the continuous-time and discrete-time system via taking a limit [3]. Moreover, the infinite-horizon case is also treated in [3] as the stationary limit of a finite-horizon problem, which again allows us to extend our results to the infinite-horizon case.

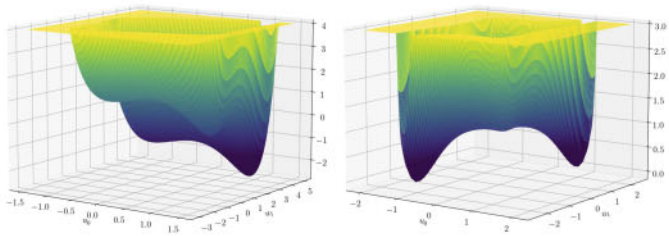
Considering Theorems 1 and 3 altogether, one can conclude that under mild conditions, each local minimizer of the one-shot optimization corresponds to some control input induced by a locally minimum control policy, and vice versa.

E. Numerical Examples

To effectively demonstrate the results of this section via visualization, we will provide two low-dimensional examples.

Example 1: Consider an optimal control problem with the control constraint $A = [-10, 10]$ and

$$c_0(x,u) = 0$$
$$c_1(x,u) = \frac{1}{4}u^4 - \frac{3x+4}{3}u^3 + \frac{3x^2+8x+3}{2}u^2 - x(x+1)(x+3)u + \exp(x^4)$$
$$c_2(x) = 0,\ f_0(x,u) = x+u,\ f_1(x,u) = x+u.$$



(a) (b)

Fig. 1. Landscape of the one-shot optimization. (a) Each local minimizer is equivalent to a set of control inputs induced by each locally minimum control policy—Example 1. (b) (0,0) is a control input induced by a locally minimum control policy but not a local minimizer of the one-shot optimization. However, it is indeed a stationary point of the one-shot optimization—Example 2.

At the initial state $x_0 = 0$, the one-shot problem is written as

$$\min_{u_0 \in A, u_1 \in A} \left\{ \frac{1}{4}u_1^4 - \frac{3u_0+4}{3}u_1^3 + \frac{3u_0^2+8u_0+3}{2}u_1^2 - u_0(u_0+1)(u_0+3)u_1 + \exp(u_0^4) \right\}.$$

This one-shot optimization problem has three spurious local minimizers $(-0.523, -0.523)$, $(-0.523, 2.477)$, and (0.938,0.938), and the globally optimal minimizer (0.938,3.938). The landscape of this objective function is shown in Fig. 1(a).

The optimal control problem can also be solved sequentially by DP. At the time step 1, the Q-function is $Q_1^\pi(x,u_1) = c_1(x,u_1)$, which has the maximum point $x+1$, the spurious local minimizer x, and the global minimizer $x+3$. One can choose between the two different continuous policies

$$\pi_1(x) = \begin{cases} x, & |x| \le 10 \\ 10 \cdot \operatorname{sgn}(x), & \text{otherwise} \end{cases}$$

or

$$\pi_1(x) = \begin{cases} x+3, & -13 \le x \le 7 \\ 10 \cdot \operatorname{sgn}(x), & \text{otherwise} \end{cases}$$

where $\operatorname{sgn}(x)$ denotes the sign of x. The first policy has the cost-to-go function $J_1^\pi(x) = -\frac{1}{12}(3x^4+16x^3+18x^2) + \exp(x^4)$ for $|x| \le 10$ and the second policy has $J_1^\pi(x) = -\frac{1}{12}(3x^4 + 16x^3 + 18x^2 + 27) + \exp(x^4)$ for $-13 \le x \le 7$.

At the time step 0 and the initial state $x_0 = 0$, the Q-function is $Q_0^\pi(0,u_0) = J_1^\pi(u_0)$. For the first policy, the Q-function has a spurious local minimizer at $u_0 = -0.523$ and a global minimum at $u_0 = 0.938$. If we choose $\pi_0(0) = -0.523$, then the induced input under π of DP is $(-0.523, -0.523)$, and if we choose $\pi_0(0) = 0.938$, then the induced input under π of DP is (0.938,0.938). Both of these input sequences are spurious local minimizers of the one-shot problem.

The Q-function of the second policy has a spurious local minimizer at $u_0 = -0.523$ and a global minimum at $u_0 = 0.938$. If we choose $\pi_0(0) = 0.938$, then the locally minimum control policy π is nonspurious and its induced input (0.938,3.938) is the global minimizer of the one-shot problem. However, if we choose $\pi_0(0) = -0.523$, then π is spurious and its induced input $(-0.523, 2.477)$ is the spurious minimizer of the one-shot problem.

In this example, one can observe that each strictly local minimizer of the one-shot problem corresponds to a locally minimum control policy of DP, which validates the result of Theorem 3. In addition, it can be noticed that since $\nabla_u^2 Q_0^\pi(0,-0.523)$ and $\nabla_u^2 Q_0^\pi(0,0.938)$ are both strictly positive for each of the two policies, Theorem 1 clearly holds.

Example 2: Consider the problem in Example 1 but change $c_1(x,u)$ to $\frac{1}{4}u^4 - \frac{x}{3}u^3 - x^2u^2 + \exp(x^4)$. At the initial state $x_0 = 0$, the one-shot problem can be written as

$$\min_{u_0 \in A, u_1 \in A} \left\{ \frac{1}{4}u_1^4 - \frac{u_0}{3}u_1^3 - u_0^2 u_1^2 + \exp(u_0^4) \right\}.$$

It has three stationary points (0,0), $((\log(\frac{8}{3}))^{\frac{1}{4}}, 2(\log(\frac{8}{3}))^{\frac{1}{4}})$, and $(-(\log(\frac{8}{3}))^{\frac{1}{4}}, -2(\log(\frac{8}{3}))^{\frac{1}{4}})$. The latter two are the global minimizers of this one-shot problem. For (0,0), we take $u_0 = u_1 = \epsilon$ and use the Taylor expansion of the exponential function to arrive at $\frac{1}{4}\epsilon^4 - \frac{1}{3}\epsilon^4 - \epsilon^4 + \exp(\epsilon^4) = -\frac{1}{12}\epsilon^4 + 1 + o(\epsilon^4)$, which is strictly less than 1 for sufficiently small values of ϵ. This implies that (0,0) is not a local minimizer of the one-shot problem. The landscape of this objective function is shown in Fig. 1(b). It can also be solved sequentially by DP. For the initial state x_0, it has three different induced input sequences under the locally minimum control policy: $(\log(\frac{8}{3}))^{\frac{1}{4}}, 2(\log(\frac{8}{3}))^{\frac{1}{4}})$, $(-(\log(\frac{8}{3}))^{\frac{1}{4}}, -2(\log(\frac{8}{3}))^{\frac{1}{4}})$, and (0,0). The first two points are the global minimizers of the one-shot problem but (0,0) is not a local minimizer of the one-shot problem.

In this example, $\nabla_u^2 Q_1^\pi(0,0) = \nabla_u^2 c_1(0,0) = 0$ violates the assumptions in Theorem 1, and thus (0,0) is not a local minimizer of the one-shot problem. This clarifies the role of the regularity conditions needed in the theorem. On the other hand, $Q_1^\pi(x,\cdot)$ has three stationary control policies $0, -x, 2x$. Consistent with Theorem 2, (0,0) is a saddle point (which is a stationary point) of the one-shot optimization.

III. Deterministic Problem Under a Parameterized Policy

A. Problem Formulation

In Section II, the one-shot optimization approach is referred to as an open-loop control, in the sense that it determines all the control inputs at once, only given an initial state. On the other hand, the DP approach is referred to as a closed-loop control, in the sense that the control input of each time step is the function of the output of the previous step [3]. In this section, we formulate both approaches to a closed-loop control. To achieve this, we can replace the control inputs of the one-shot optimization with a parameterized policy. We still optimize over a vector at once, which means that it can be solved in a one-shot fashion. However, this method becomes a type of closed-loop control in the sense that a function of both the parameters at each step and the output of the previous step determines the control input [35]. Also, it is reasonable to adopt such parameterized policies for DP as well, which would still be a closed-loop control. Note that both approaches now optimize over a set of parameters so that they can be directly compared in terms of their landscapes. This motivates us to modify Definitions 1–3, 5, and 6 to incorporate parameterized policies.

Definition 7: Given a parameter space Θ and a compact action space A, let $\mu_\theta(\cdot) : \mathbb{R}^N \to A$ be a bounded real-valued function parameterized by $\theta \in \Theta$, which satisfies the continuity assumption that for all $\epsilon > 0$, there exists $\delta > 0$ such that

$$\|\theta - \theta'\| < \delta \Rightarrow \sup_{x \in \mathbb{R}^N} \|\mu_\theta(x) - \mu_{\theta'}(x)\| < \epsilon. \tag{8}$$

Now, we modify the deterministic problems (P1)–(P3) to a discrete-time finite-horizon deterministic optimal control problem under a parameterized policy as follows:

$$\begin{aligned} \min_{\theta_0,\ldots,\theta_{n-1} \in \Theta} \quad & \sum_{i=0}^{n-1} c_i(x_i, \mu_{\theta_i}(x_i)) + c_n(x_n) \\ \text{s.t.} \quad & x_{i+1} = f_i(x_i, \mu_{\theta_i}(x_i)), \qquad i = 0,\ldots,n-1 \\ & x_0 \text{ is given.} \end{aligned} \tag{PP1}$$

Definition 8: Given a control policy parameter vector $\pi = (\theta_0,\ldots,\theta_{n-1})$, the associated Q-functions $Q_k^\pi(\cdot,\cdot)$ and cost-to-go functions $J_k^\pi(\cdot)$ under the policy π are defined in a backward way from the time step $n-1$ to the time step 0 through the following recursion:

$$\begin{aligned} & J_n^\pi(x) = c_n(x) \\ & Q_k^\pi(x, \mu_\theta(x)) = c_k(x, \mu_\theta(x)) + J_{k+1}^\pi(f_k(x, \mu_\theta(x))), \\ & \qquad\qquad k = 0,\ldots,n-1 \\ & J_k^\pi(x) = Q_k^\pi(x, \mu_{\theta_k}(x)), \quad k = 0,\ldots,n-1. \end{aligned}$$

Then, the one-shot optimization problem (PP1) can be equivalently written as

$$\min_{\pi = (\theta_0,\ldots,\theta_{n-1}) \in \Theta^n} J_0^\pi(x_0) \tag{PP2}$$

and DP approach can be written as the following backward recursion:

$$\begin{aligned} & J_n(x) = c_n(x) \\ & J_k(x) = \min_{\theta \in \Theta} \{ c_k(x, \mu_\theta(x)) + J_{k+1}(f_k(x, \mu_\theta(x))) \} \end{aligned} \tag{PP3}$$

for $k = 0,\ldots,n-1$. Note that π was previously defined as a control policy $(\pi_0,..,\pi_{n-1})$, but we use the equivalent definition $(\theta_0,\ldots,\theta_{n-1})$ in the parameterized case. We also call it *control policy parameter vector* alternatively.

Definition 9 (Local minimizer of the one-shot optimization): A control policy parameter vector $\pi = (\theta_0^*,\ldots,\theta_{n-1}^*)$ is said to be a local minimizer of the one-shot optimization if there exists $\epsilon > 0$ such that

$$J_0^\pi(x_0) \le J_0^{\tilde{\pi}}(x_0)$$

for all $\tilde{\pi} = (\tilde{\theta}_0,\ldots,\tilde{\theta}_{n-1}) \in (B(\theta_0^*,\epsilon) \cap \Theta) \times \cdots \times (B(\theta_{n-1}^*,\epsilon) \cap \Theta)$.

Definition 10 (Local minimizer of DP): A control policy parameter vector $\pi = (\theta_0^*,\ldots,\theta_{n-1}^*)$ is said to be a local minimizer of DP if for all $k \in \{0,\ldots,n-1\}$ and for all $x \in \mathbb{R}^N$, the policy parameter θ_k^* is a local minimizer of $Q_k^\pi(x, \mu_{(\cdot)}(x))$, meaning that there exists $\epsilon_k^* > 0$ such that

$$Q_k^\pi(x, \mu_{\theta_k^*}(x)) \le Q_k^\pi(x, \mu_{\tilde{\theta}}(x)) \quad \forall \tilde{\theta} \in B(\theta_k^*, \epsilon_k^*) \cap \Theta. \tag{9}$$

Definition 11 (Stationary point of the one-shot optimization): A control policy parameter vector $\pi = (\theta_0^*,\ldots,\theta_{n-1}^*)$ is said

to be a stationary point of the one-shot optimization if for all $k \in \{0, \dots, n-1\}$, it holds that $-\nabla_{\theta_k} J_0^\pi(x_0) \in \mathcal{N}_\Theta(\theta_k^*)$.

Definition 12 (Stationary point of DP): A control policy parameter vector $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$ is said to be a stationary point of DP if for all $k \in \{0, \dots, n-1\}$ and for all $x \in \mathbb{R}^N$, it holds that $-\nabla_{\theta_k} Q_k^\pi(x, \mu_{\theta_k^*}(x)) \in \mathcal{N}_\Theta(\theta_k^*)$.

Remark 5: By comparing (P2) with (PP2) as well as comparing Definition 2 with Definition 9, notice that one-shot optimization now considers $J_0^\pi(x_0)$ instead of $C(x_0; \theta_0, \dots, \theta_{n-1})$, since the two definitions are equivalent when the parameterized policy is incorporated.

We can compare Definition 10 with the following definition:

$$\forall k \in \{0, \dots, n-1\} \quad \forall x \in \mathbb{R}^N, \ \exists \epsilon_k^*(x) > 0 \text{ such that}$$
$$Q_k^\pi(x, \mu_{\theta_k^*}(x)) \le Q_k^\pi(x, \tilde{u}) \quad \forall \tilde{u} \in B(\mu_{\theta_k^*}(x), \epsilon_k^*(x)) \cap A. \tag{10}$$

Definition 10 considers the open ball centered at the policy parameter in the parameter space, whereas (10) considers the corresponding open ball in the action space. Proposition 1 establishes the relationship between these definitions.

Proposition 1: If an arbitrary control policy parameter vector $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$ satisfies (10) with $\inf_{x \in \mathbb{R}^N} \epsilon_k^*(x) > 0$ for all $k \in \{0, \dots, n-1\}$, then it is a local minimizer of DP.

Proof: Since $\inf_{x \in \mathbb{R}^N} \epsilon_k^*(x) > 0$, by the continuity assumption, for every $k \in \{0, \dots, n-1\}$, there exists $\delta_k > 0$ such that

$$\|\theta - \theta_k^*\| < \delta_k \Rightarrow \sup_{x \in \mathbb{R}^N} \|\mu_\theta(x) - \mu_{\theta_k^*}(x)\| < \inf_{x \in \mathbb{R}^N} \epsilon_k^*(x).$$

That is, for all $\theta \in B(\theta_k^*, \delta_k) \cap \Theta$, $\|\mu_\theta(x) - \mu_{\theta_k^*}(x)\| < \epsilon_k^*(x)$ for all $x \in \mathbb{R}^N$. Notice that Definition 7 implies that $\mu_\theta(x) \in A$ for all $x \in \mathbb{R}^N$. Thus, it holds for all $x \in \mathbb{R}^N$ that

$$\theta \in B(\theta_k^*, \delta_k) \cap \Theta \Rightarrow \mu_\theta(x) \in B(\mu_{\theta_k^*}(x), \epsilon_k^*(x)) \cap A.$$

Thus, given a control policy parameter vector satisfying (10), for all $k \in \{0, \dots, n-1\}$ and for all $x \in \mathbb{R}^N$, (9) holds if one substitutes ϵ_k^* with δ_k. This completes the proof. ■

Remark 6: The converse of Proposition 1 does not hold. For example, suppose that there exists $\epsilon_k^* > 0$ such that $\mu_\theta(x)$ takes the same value for all $\theta \in B(\theta_k^*, \epsilon_k^*) \cap \Theta$. While this control policy satisfies the continuity assumption, θ_k^* is clearly a local minimizer of DP, which satisfies (9). However, it is even possible that $\mu_{\theta_k^*}(x)$ is a strict local maximizer of $Q_k^\pi(x, \cdot)$.

Note that the condition $\inf_{x \in \mathbb{R}^N} \epsilon_k^*(x) > 0$ is necessary for Proposition 1. Thus, the proposition implies that if we use our notion of a local minimizer of DP, we no longer need to assume that $\inf_{x \in \mathbb{R}^N} \epsilon_k^*(x) > 0$ while establishing the relationship from DP to one-shot optimization, which was the case in the (non-parameterized) deterministic case presented in our conference paper (see [1, Th. 2]).

B. From DP to One-Shot Optimization

In this section, we will show that in the deterministic case with a parameterized policy, each local minimizer (stationary point) of DP directly corresponds to some local minimizer (stationary point) of the one-shot optimization.

Theorem 4: Consider a local minimizer of DP $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$. Then, π is also a local minimizer of the one-shot optimization.

Proof: Since $(\theta_0^*, \dots, \theta_{n-1}^*)$ is a local minimizer of DP, there exist $\epsilon_0^*, \dots, \epsilon_{n-1}^* > 0$ such that

$$\begin{aligned} J_0^\pi(x_0) &= Q_0^\pi(x_0, \mu_{\theta_0^*}(x_0)) \le Q_0^\pi(x_0, \mu_{\tilde{\theta}_0}(x_0)) \\ &= c_0(x_0, \mu_{\tilde{\theta}_0}(x_0)) + Q_1^\pi(\tilde{x}_1, \mu_{\theta_1^*}(\tilde{x}_1)) \\ &\qquad (\tilde{x}_1 = f_0(x_0, \mu_{\tilde{\theta}_0}(x_0))) \\ &\le c_0(x_0, \mu_{\tilde{\theta}_0}(x_0)) + Q_1^\pi(\tilde{x}_1, \mu_{\tilde{\theta}_1}(\tilde{x}_1)) \\ &= c_0(x_0, \mu_{\tilde{\theta}_0}(x_0)) + c_1(\tilde{x}_1, \mu_{\tilde{\theta}_1}(\tilde{x}_1)) + Q_2^\pi(\tilde{x}_2, \mu_{\theta_2^*}(\tilde{x}_2)) \\ &\qquad (\tilde{x}_2 = f_1(\tilde{x}_1, \mu_{\tilde{\theta}_1}(\tilde{x}_1))) \\ &\le \cdots \le J_0^{\tilde{\pi}}(x_0) \end{aligned}$$

where $\tilde{\pi} = (\tilde{\theta}_0, \dots, \tilde{\theta}_{n-1}) \in (B(\theta_0^*, \epsilon_0^*) \cap \Theta) \times \cdots \times (B(\theta_{n-1}^*, \epsilon_{n-1}^*) \cap \Theta)$.

Choose $\epsilon = \min\{\epsilon_0^*, \dots \epsilon_{n-1}^*\}$. Then, $J_0^\pi(x_0) \le J_0^{\tilde{\pi}}(x_0)$ for all $\tilde{\pi} = (\tilde{\theta}_0, \dots, \tilde{\theta}_{n-1}) \in (B(\theta_0^*, \epsilon) \cap \Theta) \times \cdots \times (B(\theta_{n-1}^*, \epsilon) \cap \Theta)$. This completes the proof. ■

Theorem 5: Consider a stationary point of DP $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$. Let the corresponding state sequence be $(x_0^*, \dots x_n^*)$. If for every $k \in \{0, \dots, n-1\}$, $\mu_{\theta_k}(x_k)$ is continuously differentiable with respect to θ_k in a neighborhood of (x_k^*, θ_k^*), then π is also a stationary point of the one-shot optimization.

Proof: Notice that $\nabla_{\theta_k} J_0^\pi(x_0) = \nabla_{\theta_k} J_k^\pi(x_k^*) = \nabla_{\theta_k} Q_k^\pi(x_k^*, \mu_{\theta_k^*}(x_k^*))$. Thus, $-\nabla_{\theta_k} J_0^\pi(x_0) \in \mathcal{N}_\Theta(\theta_k^*)$ for all $k \in \{0, \dots, n-1\}$, which means that π is a stationary point of the one-shot optimization. ■

Remark 7: The converse of Theorem 5 clearly does not hold since one can generally find a point $x \in \mathbb{R}^N$ such that $\nabla_{\theta_k} Q_k^\pi(x_k^*, \mu_{\theta_k^*}(x_k^*)) \ne \nabla_{\theta_k} Q_k^\pi(x, \mu_{\theta_k^*}(x))$.

C. From One-Shot Optimization to DP

In this section, we first show that a local minimizer of the one-shot optimization does not necessarily correspond to a local minimizer of DP; i.e., the converse of Theorem 4 does not hold. Then, with Remark 7, it is clear that the optimization landscape of the one-shot optimization is more complex than that of DP. As a byproduct, if the one-shot problem has a *low* complexity, so does the DP problem.

To develop a clear counterexample, we restrict the parameterized policy to a certain class as given below, which automatically satisfies the continuity assumption defined in Definition 7.

Definition 13: Define our parameterized policy to be a linear combination of arbitrary linearly independent basis functions, while satisfying Definition 7; i.e., given m functions $f_i : \mathbb{R}^N \to \mathbb{R}^M$, $i = 1, \dots, m$, and $\theta = [s_1, \dots, s_m]^T \in \Theta$

$$\mu_\theta(x) = \sum_{i=1}^m s_i f_i(x) \in A \tag{11}$$

where there does not exist $(\tilde{s}_1, \ldots, \tilde{s}_m) \neq 0$ such that for all x in any set of nonzero measure, the following equation holds [36]:

$$\sum_{i=1}^{m} \tilde{s}_i f_i(x) = 0. \tag{12}$$

Remark 8: Since a set of isolated points is a set of measure zero, it is exempt from determining the independence of basis functions. When x has a continuous distribution, the independence of basis functions implies that if (12) holds for all x in the support of the distribution, $\tilde{s}_1 = \cdots = \tilde{s}_m = 0$. When x has a discrete distribution, since a set of all the possible values of x is a set of measure zero, the independence of basis functions does not guarantee $\tilde{s}_1 = \cdots = \tilde{s}_m = 0$ even if (12) holds for all possible values of x.

Applications of a parameterized policy defined by Definition 13 arise in a piecewise polynomial function as well as a stochastic control. The usefulness of the parameterized policy also manifests within representer theorem [37]: a linear combination of kernels fully represents the solution of minimizing empirical risk. It switches the optimization problem in infinite-dimensional function space to finding the finite number of coefficients. The minimum number of parameters needed is the number of data points, which is generally much greater than the dimension of the output. Applying this to our parameterized policy, the number of parameters m needs to be greater than the dimension of the action M to cover all data points. For the rest of this section, we call a policy satisfying $m > M$ as an overparameterized policy.

We now provide some evidence to refute the converse of Theorem 4, specifically if the parameterized policy class is a linear combination of basis functions. It turns out that a local minimizer of the one-shot optimization does not necessarily imply a local minimizer of DP in the overparameterized case.

Proposition 2: Consider an overparameterized policy class defined by Definition 13. Let $\pi = (\theta_0^*, \ldots, \theta_{n-1}^*)$ be a local minimizer of DP. If there exists at least one $k \in \{0, \ldots, n-1\}$ such that θ_k^* is in the interior of Θ, then there exists an infinite number of local minimizers of the one-shot optimization corresponding to each local minimizer of DP.

Proof: Consider the state sequence $(x_0^*, \ldots, x_n^*)$ induced by a local minimizer of DP π. Let k be an index for which θ_k^* is in the interior of Θ. Then, one can express the action taken at step k as $\mu_{\theta_k^*}(x_k^*) = \sum_{i=1}^{m} s_i^* f_i(x_k^*)$ with $\theta_k^* = (s_0^*, s_1^*, \ldots, s_m^*)$ by Definition 13. Since the policy is overparameterized, m is greater than the dimension of the action M. Now, consider the matrix equation

$$\begin{bmatrix} f_0(x_k^*) & f_1(x_k^*) & \ldots & f_m(x_k^*) \end{bmatrix} \theta_k = \mu_{\theta_k^*}(x_k^*) \tag{13}$$

where θ_k is an $m \times 1$ vector variable, and let F_k^* denote the first matrix in the left-hand side, which is an $M \times m$ constant matrix given by x_k^*. Equation (13) has at least one solution: θ_k^*.

The dimension of the null space of F_k^* is greater than 0 due to $m > M$. We take any nonzero element v from the null space. Then, for all $\delta \in \mathbb{R}$, $\theta_k^* + \delta v$ satisfies (13). Since θ_k^* is in the interior of Θ, one can pick $\epsilon_1 > 0$ such that $B(\theta_k^*, \epsilon_1) \subset \Theta$. Thus, for $0 \leq \delta \leq \frac{\epsilon_1}{\|v\|}$, $\theta_k^* + \delta v \in B(\theta_k^*, \epsilon_1)$ preserves the state and action sequences associated with θ_k^* due to (13). The induced cost is also indeed preserved.

By Theorem 4, π is a local minimizer of the one-shot optimization. Now, we select $\epsilon_2 > 0$ such that $J_0^{\pi}(x_0) \leq J_0^{\tilde{\pi}}(x_0)$ for all $\tilde{\pi} = (\theta_0^*, \ldots, \tilde{\theta}_k, \ldots, \theta_{n-1}^*)$, where $\tilde{\theta}_k \in B(\theta_k^*, \epsilon_2) \cap \Theta$. Let $\epsilon := \min\{\epsilon_1, \epsilon_2\} > 0$. Then, for $0 \leq \delta \leq \frac{\epsilon}{2\|v\|}$, we have $B(\theta_k^* + \delta v, \delta\|v\|) \subset B(\theta_k^*, \epsilon)$. Since $\theta_k^* + \delta v$ preserves the induced cost, $(\theta_0^*, \ldots, \theta_k^* + \delta v, \ldots, \theta_{n-1}^*)$ is a local minimizer of the one-shot optimization for all $0 \leq \delta \leq \frac{\epsilon}{2\|v\|}$. This completes the proof. ■

Proposition 2 implies that for every $k \in \{0, \ldots, n-1\}$, θ_k of a "strict" local minimizer of the one-shot optimization does not lie in the interior of Θ. Thus, one can think of constructing a strict local minimizer by restricting the area of Θ. It turns out that given a strict local minimizer of the one-shot optimization and the induced input sequence, no other points can retrieve the same input sequence if Θ is convex.

Lemma 2: Consider a strict local minimizer of the one-shot optimization $\pi = (\theta_0^*, \ldots, \theta_{n-1}^*)$. Let $(x_0^*, \ldots, x_n^*)$ be the induced state sequence. Suppose that Θ is convex and the parameterized policy is defined by Definition 13. Then, π is the unique control policy parameter vector that achieves the input sequence $(\mu_{\theta_0^*}(x_0^*), \ldots, \mu_{\theta_{n-1}^*}(x_{n-1}^*))$.

Proof: For every $k \in \{0, \ldots, n-1\}$, $\mu_\theta(x_k^*) = \sum_{i=1}^{m_k} s_i f_i(x_k^*)$, where $\theta = (s_1, \ldots, s_{m_k})$. Let $\theta_k^* = (s_1^*, \ldots, s_{m_k}^*)$. Since π is a strict local minimizer of the one-shot optimization, $\mu_\theta(x_k^*) \neq \mu_{\theta_k^*}(x_k^*)$ in the neighborhood of θ_k^* if $\theta \neq \theta_k^*$; i.e., there exists $\epsilon > 0$ such that

$$\left\{ \theta \in B(\theta_k^*, \epsilon) \cap \Theta : \sum_{i=1}^{m_k} s_i f_i(x_k^*) = \mu_{\theta_k^*}(x_k^*) \right\} = \{\theta_k^*\}. \tag{14}$$

Assume that there exists $\tilde{\theta} \neq \theta_k^*$ such that $\sum_{i=1}^{m_k} \tilde{s}_i f_i(x_k^*) = \mu_{\theta_k^*}(x_k^*)$, where $\tilde{\theta} = (\tilde{s}_1, \ldots, \tilde{s}_{m_k})$. Then, for $\lambda \in [0, 1]$, one can obtain $\sum_{i=1}^{m_k} (\lambda s_i^* + (1-\lambda)\tilde{s}_i) f_i(x_k^*) = \mu_{\theta_k^*}(x_k^*)$ by linearity and $\lambda\theta_k^* + (1-\lambda)\tilde{\theta} \in \Theta$ by convexity. Letting $\lambda \to 1$, one can construct an element of the left-hand side of (14) distinct from θ_k^*. By contradiction, θ_k^* is the unique point that achieves $\mu_{\theta_k^*}(x_k^*)$. ■

Note that Lemma 2 does not necessarily imply that a strict local minimizer of the one-shot optimization is a local minimizer of DP even if Θ is convex. A simple counterexample can be constructed by considering the one-step problem

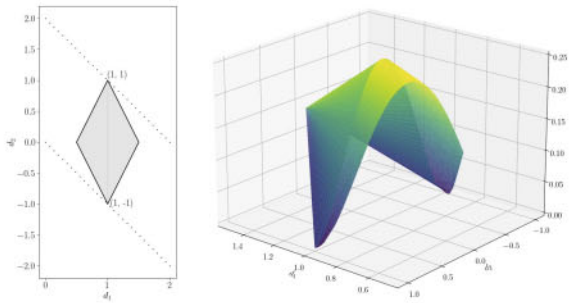
$$c_0(x, \mu_\theta(x)) = \frac{1}{4}\mu_\theta(x)^4 - \frac{1}{3}(x^2 + 2x)\mu_\theta(x)^3 + \frac{1}{2}(2x^3 + x - 1)\mu_\theta(x)^2 - (x^4 - x^3 + x^2 - x)\mu_\theta(x)$$

$$c_1(x, \mu_\theta(x)) = 0, \; f_0(x, \mu_\theta(x)) = x + \mu_\theta(x)$$

with the parameterized policy $\mu_\theta(x) = d_1 x + d_2$, where $\theta = (d_1, d_2)$, and $\Theta = \{(d_1, d_2) : 1 \leq 2d_1 - d_2 \leq 3, 1 \leq 2d_1 + d_2 \leq 3\}$, which is convex. At the initial state $x_0 = 1$, the one-shot problem can be written as

$$\min_{(d_1, d_2) \in \Theta} \left\{ \frac{1}{4}(d_1 + d_2)^4 - (d_1 + d_2)^3 + (d_1 + d_2)^2 \right\}.$$

Each vector $(d_1, d_2) \in \Theta$, which satisfies $d_1 + d_2 = 0$ or $d_1 + d_2 = 2$, is a local minimizer of the one-shot optimization. Since $\{(d_1, d_2) : d_1 + d_2 = 0\} \cap \Theta = \{(1, -1)\}$ and $\{(d_1, d_2) : d_1 + d_2 = 2\} \cap \Theta = \{(1, 1)\}$, we have $(1, -1)$ and (1,1) as strict local minimizers of the one-shot optimization. On the other hand, since $\nabla_\theta Q_0^{\pi}(x, \mu_\theta(x)) =$



(a) (b)

Fig. 2. Domain and the landscape of the one-shot optimization for a deterministic parameterized problem. (a) Gray-colored area is the domain of the parameter space. The intersection between the dotted lines and the domain is $\{(1,1),(1,-1)\}$. (b) Both (1,1) and $(1,-1)$ are a strict local minimizer of the one-shot optimization but only $(1,-1)$ is a local minimizer of DP—Landscape.

$\nabla_\theta c_0(x,\mu_\theta(x)) = [g(x,\theta)x, g(x,\theta)]^T$, where $g(x,\theta) = (\mu_\theta(x) - (x^2+1))(\mu_\theta(x) - x)(\mu_\theta(x) - (x-1))$, a local minimizer of DP should be the parameter that yields $\mu_\theta(x) = x - 1$ or $\mu_\theta(x) = x^2 + 1$ for all $x \in \mathbb{R}^N$. Since a linear policy cannot contain x^2+1, $(1,-1) \in \Theta$ is the only local minimizer of DP. Thus, (1,1) is a strict local minimizer of the one-shot optimization but not a local minimizer of DP. Fig. 2 shows the domain and the landscape of the one-shot optimization.

In light of the above counterexample, one can think of the situation where the parameterized policy contains every locally minimum control policy of DP (see Definition 3). It turns out that if such a situation is possible, given a convex parameter space, each strict local minimizer of the one-shot optimization is a local minimizer of DP under the following assumptions.

Assumption 1: Given a local minimizer of the one-shot optimization π, let $(x_0^*, \dots, x_n^*)$ be the associated state sequence. Then, for all $k \in \{0, \dots, n-1\}$, the $M \times m$ matrix $[f_0(x_k^*)\ f_1(x_k^*)\ \dots\ f_m(x_k^*)]$ has a full row rank.

Assumption 2: Assume that $A \subseteq \cap_{k=1}^n \mu_\Theta(x_k^*)$, where $\mu_\Theta(x_k^*)$ is the image of Θ through $\mu_\theta(x_k^*) : \Theta \to A$.

Lemma 3: Assume that Θ is convex. Consider a strict local minimizer of the one-shot optimization $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$. Suppose that the parameterized policy defined by Definition 13 satisfies Assumptions 1 and 2. If the parameterized policy class contains every locally minimum control policy of DP and at least one of the locally minimum control policies satisfies $\inf_{x\in\mathbb{R}^N} \epsilon_k^*(x) > 0$ for all $k \in \{0, \dots, n-1\}$, then π is a local minimizer of DP.

Proof: Let $(x_0^*, \dots, x_n^*)$ be the state sequence associated with π. Recall that $J_0^\pi(x_0) = \sum_{i=0}^{k-1} c_i(x_i^*, \mu_{\theta_i^*}(x_i^*)) + Q_k^\pi(x_k^*, \mu_{\theta_k^*}(x_k^*))$. One can fix all parameters except θ_k^* to derive that $J_0^\pi(x_0) - J_0^{\pi'}(x_0) = Q_k^\pi(x, \mu_{\theta_k^*}(x)) - Q_k^\pi(x, \mu_{\theta_k'}(x))$, where $\pi' = (\theta_0^*, \dots, \theta_{k-1}^*, \theta_k', \theta_{k+1}^*, \dots, \theta_{n-1}^*)$. Thus, a local minimizer of the one-shot optimization π implies that for all $k \in \{0, \dots, n-1\}$, there exists $\epsilon_k^* > 0$ such that

$$Q_k^\pi(x_k^*, \mu_{\theta_k^*}(x_k^*)) \le Q_k^\pi(x_k^*, \mu_{\tilde\theta}(x_k^*)) \quad \forall \tilde\theta \in B(\theta_k^*, \epsilon_k^*) \cap \Theta. \tag{15}$$

Now, let F_k^* be the $M \times m$ matrix $[f_0(x_k^*)\ f_1(x_k^*)\ \dots\ f_m(x_k^*)]$, where its smallest singular value is denoted by σ_k^*. Given an arbitrary direction $v \in \mathbb{R}^M$, one can take a point u_v that is farthest from $\mu_{\theta_k^*}(x_k^*)$ in the direction of v since the action space A is compact. Let δ_v be the value that achieves $u_v = \mu_{\theta_k^*}(x_k^*) + \delta_v v$. By Assumption 2, there exists $\theta_v \in \Theta$ satisfying $u_v = \mu_{\theta_v}(x_k^*)$, and by Definition 13, $\mu_{\theta_v}(x_k^*)$ is defined by $F_k^* \theta_v$.

Case 1—$\delta_v = 0$: There does not exist $\delta > 0$ such that $\mu_{\theta_k^*}(x_k^*) + \delta v \in A$.

Case 2—$\delta_v > 0$ and $\theta_v \in B(\theta_k^*, \epsilon_k^*)$: Due to the linearity of policy and the convexity of Θ, there exists $\theta_\delta \in B(\theta_k^*, \epsilon_k^*) \cap \Theta$ such that $\mu_{\theta_\delta}(x_k^*) = \mu_{\theta_k^*}(x_k^*) + \delta v$ for all $0 < \delta < \delta_v$.

Case 3—$\delta_v > 0$ and $\theta_v \notin B(\theta_k^*, \epsilon_k^*)$: Consider $\mu_{\theta_k^*}(x_k^*) + \frac{\epsilon_k^*}{2\|\theta_v - \theta_k^*\|}(\mu_{\theta_v}(x_k^*) - \mu_{\theta_k^*}(x_k^*))$. The corresponding parameter is definitely in $B(\theta_k^*, \epsilon_k^*) \cap \Theta$ by the linearity of policy and the convexity of Θ. Then, as in Case 2, there exists $\theta_\delta \in B(\theta_k^*, \epsilon_k^*) \cap \Theta$ such that $\mu_{\theta_\delta}(x_k^*) = \mu_{\theta_k^*}(x_k^*) + \delta v$ for all $0 < \delta < \frac{\epsilon_k^*}{2\|\theta_v - \theta_k^*\|}\delta_v$.

In Case 3, notice that $\|\frac{\epsilon_k^*}{2\|\theta_v - \theta_k^*\|}(\mu_{\theta_v}(x_k^*) - \mu_{\theta_k^*}(x_k^*))\| = \frac{\epsilon_k^*}{2} \cdot \frac{\|F_k^*(\theta_v - \theta_k^*)\|}{\|\theta_v - \theta_k^*\|} \ge \frac{\epsilon_k^*}{2}\sigma_k^* > 0$, where the last inequality is from Assumption 1 and the second last inequality is from the basic property of singular value [38].

Considering all three cases, $\tilde u \in B(\mu_{\theta_k^*}(x_k^*), \frac{\epsilon_k^*}{2}\sigma_k^*) \cap A$ implies that at least one corresponding parameter for each $\tilde u$ is in $B(\theta_k^*, \epsilon_k^*) \cap \Theta$. Thus, one can notice that (15) implies

$$Q_k^\pi(x_k^*, \mu_{\theta_k^*}(x_k^*)) \le Q_k^\pi(x_k^*, \tilde u) \quad \forall \tilde u \in B(\mu_{\theta_k^*}(x_k^*), \frac{\epsilon_k^*}{2}\sigma_k^*) \cap A. \tag{16}$$

We select an arbitrary locally minimum control policy $\phi = (\phi_0, \dots, \phi_{n-1})$ with the property that $\inf_{x\in\mathbb{R}^N} \epsilon_k^*(x) > 0$. Let $\tilde\pi = (\tilde\pi_0, \dots, \tilde\pi_{n-1})$ be the policy such that for all $k \in \{0, \dots, n-1\}$

$$\tilde\pi_k(x_k) = \begin{cases} \mu_{\theta_k^*}(x_k^*), & \text{if } x_k = x_k^* \\ \phi_k(x_k), & \text{otherwise.} \end{cases}$$

Such $\tilde\pi$ is also a locally minimum control policy by (16). This implies that the parameterized policy contains $\tilde\pi$. Also, $\tilde\pi$ achieves the same input sequence $(\mu_{\theta_0^*}(x_0^*), \dots, \mu_{\theta_{n-1}^*}(x_{n-1}^*))$ as the strict local minimizer π. Therefore, by Lemma 2, $\mu_{\theta_k^*} = \tilde\pi_k$ holds. Since $\inf_{x\in\mathbb{R}^N} \epsilon_k^*(x)$ induced by ϕ_k is greater than 0, $\inf_{x\in\mathbb{R}^N} \epsilon_k^*(x)$ induced by $\tilde\pi_k$ is also greater than 0. Then, by Proposition 1, $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$ is a local minimizer of DP. ■

Remark 9: With a given set of parameters $(\theta_0^*, \dots, \theta_{n-1}^*)$, there exists only one associated state sequence for the deterministic parameterized problem. Assumptions 1 and 2 are thus only required for that specific state sequence, where one can readily check the assumptions in advance with known dynamics, parameter space, action space, and policy class. Assumption 1 is a type of regularity condition, which can be regarded as the extension of an overparameterized policy. Assumption 2 implies that Θ should be large enough to contain relevant parameters to cover the action space A. Since $\mu_\theta(x)$ is designed to be in A by Definition 7, Assumption 2 is equivalent to saying that $A = \mu_\Theta(x_0^*) = \cdots = \mu_\Theta(x_n^*)$.

Meanwhile, suppose that there exist two different locally minimum control policies in a set of nonzero measure, meaning that at some step k, $\pi_1(x) \ne \pi_2(x)$ for all $x \in I$, where I is a

set of nonzero measure. Then, there exists an infinite number of locally minimum control policies made up of π_1 and π_2 by alternating between $\pi_1(x)$ and $\pi_2(x)$ along $x \in I$, and the parameterized policy class cannot contain all these policies. We now present the situation that the parameterized policy contains every locally minimum control policy of DP.

Theorem 6: Assume that Θ is convex. Consider a strict local minimizer of the one-shot optimization $\pi = (\theta_0^*, \ldots, \theta_{n-1}^*)$. Suppose that the parameterized policy defined by Definition 13 satisfies Assumptions 1 and 2. If there exists only a single locally minimum control policy of DP $\phi = (\phi_0, \ldots, \phi_{n-1})$ and the parameterized policy class contains ϕ, then π is a local minimizer of DP.

Proof: Let $\phi' = (\theta_0', \ldots, \theta_{n-1}')$ be the parameters associated with ϕ. For all $k \in \{0, \ldots, n-1\}$ and for all $x \in \mathbb{R}^N$, $\phi_k(x)$ is the unique local minimizer of $Q_k^{\phi'}(x, u)$. Having no spurious local minima implies that $\inf_{x \in \mathbb{R}^N} \epsilon_k^*(x) = \infty > 0$. Moreover, the parameterized policy class contains every locally minimum control policy of DP. Since these facts satisfy the preconditions of Lemma 3, this completes the proof. ■

Considering both Theorem 4 and 6, one can conclude that under the assumptions of Theorem 6, a local minimizer of DP is *equivalent* to a local minimizer of the one-shot optimization.

IV. Stochastic Problem Under a Parameterized Policy

A. Problem Formulation

In this section, we will show that the results obtained for the deterministic problem under a parameterized policy also hold for the stochastic problem under a parameterized policy. Since we now take the expectation of the sum of the costs over the trajectories, the issue of strictness, as in Proposition 2, does not take place. Before presenting the theorems, we first define the problem setting in the stochastic case.

Definition 14: Given a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let x_0 be a $\mathcal{F}$-measurable, $\mathbb{R}^N$-valued random variable, which has an initial distribution ρ. Also, let w_k be an $\mathcal{F}$-measurable, $\mathbb{R}^W$-valued random variable for all $k \in \{0, \ldots, n-1\}$ such that $x_0, w_0, \ldots, w_{n-1}$ are mutually independent. The state transition is now governed by the dynamics $f_i : \mathbb{R}^N \times A \times \mathbb{R}^W \to \mathbb{R}^N$, $i = 0, \ldots, n-1$. The dynamics are again defined to be at least twice continuously differentiable.

Now, we modify the deterministic problems under a parameterized policy, i.e., (PP1)–(PP3), to a discrete-time finite-horizon stochastic optimal control problem under a parameterized policy

$$\min_{\theta_0, \ldots, \theta_{n-1} \in \Theta} \quad \mathbb{E}_{x_0, w_0, \ldots, w_{n-1}} \left[\sum_{i=0}^{n-1} c_i(x_i, \mu_{\theta_i}(x_i)) + c_n(x_n) \right]$$
$$\text{where} \quad x_{i+1} = f_i(x_i, \mu_{\theta_i}(x_i), w_i), \quad i = 0, \ldots, n-1. \tag{SP1}$$

Notice that for stochastic problems, x_0 may not be given as a point, but has an initial distribution ρ. Afterward, x_{i+1} is a random variable induced by $(x_0, w_0, \ldots, w_i)$.

Definition 15: Given a control policy parameter vector $\pi = (\theta_0, \ldots, \theta_{n-1})$, the associated Q-functions $Q_k^\pi(\cdot, \cdot)$ and cost-to-go functions $J_k^\pi(\cdot)$ under the policy π are defined in a backward way from the time step $n-1$ to the time step 0 through the following recursion:

$$J_n^\pi(x) = c_n(x)$$
$$Q_k^\pi(x, \mu_\theta(x)) = \mathbb{E}_{w_k}[c_k(x, \mu_\theta(x)) + J_{k+1}^\pi(f_k(x, \mu_\theta(x), w_k))], \quad k = 0, \ldots, n-1$$
$$J_k^\pi(x) = Q_k^\pi(x, \mu_{\theta_k}(x)), \quad k = 0, \ldots, n-1.$$

Then, the one-shot optimization problem (SP1) can be equivalently written as

$$\min_{\pi = (\theta_0, \ldots, \theta_{n-1}) \in \Theta^n} \quad \mathbb{E}_{x_0}[J_0^\pi(x_0)] \tag{SP2}$$

as long as the cost functions c_i, $i = 0, \ldots, n-1$, are uniformly bounded, due to the product measure theorem and Fubini's theorem [39]. In the rest of this article, we assume that the two problems are equivalent.

The DP approach can be written as the following backward recursion:

$$J_n(x) = c_n(x)$$
$$J_k(x) = \min_{\theta \in \Theta} \{\mathbb{E}_{w_k}[c_k(x, \mu_\theta(x)) + J_{k+1}(f_k(x, \mu_\theta(x), w_k))]\}, \quad k = 0, \ldots, n-1. \tag{SP3}$$

Definition 16 (Local minimizer of the one-shot optimization): A control policy parameter vector $\pi = (\theta_0^*, \ldots, \theta_{n-1}^*)$ is said to be a local minimizer of the one-shot optimization if there exists $\epsilon > 0$ such that

$$\mathbb{E}_{x_0}[J_0^\pi(x_0)] \leq \mathbb{E}_{x_0}[J_0^{\tilde{\pi}}(x_0)]$$

for all $\tilde{\pi} = (\tilde{\theta}_0, \ldots, \tilde{\theta}_{n-1}) \in (B(\theta_0^*, \epsilon) \cap \Theta) \times \cdots \times (B(\theta_{n-1}^*, \epsilon) \cap \Theta)$.

Definition 17 (Stationary point of the one-shot optimization): A control policy parameter vector $\pi = (\theta_0^*, \ldots, \theta_{n-1}^*)$ is said to be a stationary point of the one-shot optimization if for all $k \in \{0, \ldots, n-1\}$, it holds that $-\nabla_{\theta_k} \mathbb{E}_{x_0}[J_0^\pi(x_0)] \in \mathcal{N}_\Theta(\theta_k^*)$.

While the one-shot method aims for optimizing the expectation over all steps in the stochastic dynamics, DP studies optimizing Q-function at every step both in the deterministic and stochastic cases. Since we have modified the definition of Q-function to incorporate the expectation, it is natural that the definition of a local minimizer (stationary point) of DP is exactly the same as Definition 10 (12).

B. From DP to One-Shot Optimization

In this section, we will show that, in the stochastic case with a parameterized policy, each local minimizer (stationary point) of DP directly corresponds to some local minimizer (stationary point) of the one-shot optimization, just as in the deterministic case. However, it turns out that for the stationary points, the policy needs to be continuously differentiable with respect to both states and parameters since the expectation is over all trajectories rather than a single trajectory.

Theorem 7: Consider a local minimizer of DP $\pi = (\theta_0^*, \ldots, \theta_{n-1}^*)$. Then, π is also a local minimizer of the one-shot optimization.

Proof: Since $(\theta_0^*, \ldots, \theta_{n-1}^*)$ is a local minimizer of DP, there exist $\epsilon_0^*, \ldots, \epsilon_{n-1}^* > 0$ such that

$$\mathbb{E}_{x_0}[J_0^\pi(x_0)] = \mathbb{E}_{x_0}[Q_0^\pi(x_0, \mu_{\theta_0^*}(x_0))] \le \mathbb{E}_{x_0}[Q_0^\pi(x_0, \mu_{\tilde{\theta}_0}(x_0))]$$
$$= \mathbb{E}_{x_0}[c_0(x_0, \mu_{\tilde{\theta}_0}(x_0)) + \mathbb{E}_{w_0}[Q_1^\pi(\tilde{x}_1, \mu_{\theta_1^*}(\tilde{x}_1))]]$$
$$(\tilde{x}_1 = f_0(x_0, \mu_{\tilde{\theta}_0}(x_0), w_0))$$
$$\le \mathbb{E}_{x_0}[c_0(x_0, \mu_{\tilde{\theta}_0}(x_0)) + \mathbb{E}_{w_0}[Q_1^\pi(\tilde{x}_1, \mu_{\tilde{\theta}_1}(\tilde{x}_1))]]$$
$$= \mathbb{E}_{x_0}[c_0(x_0, \mu_{\tilde{\theta}_0}(x_0)) + \mathbb{E}_{w_0}[c_1(\tilde{x}_1, \mu_{\tilde{\theta}_1}(\tilde{x}_1)) + \mathbb{E}_{w_1}[Q_2^\pi(\tilde{x}_2, \mu_{\theta_2^*}(\tilde{x}_2))]]]$$
$$(\tilde{x}_2 = f_1(\tilde{x}_1, \mu_{\tilde{\theta}_1}(\tilde{x}_1), w_1))$$
$$\le \cdots \le \mathbb{E}_{x_0, w_0, \ldots, w_{n-1}}[J_0^{\tilde{\pi}}(x_0)]$$

where $\tilde{\pi} = (\tilde{\theta}_0, \ldots, \tilde{\theta}_{n-1}) \in (B(\theta_0^*, \epsilon_0^*) \cap \Theta) \times \cdots \times (B(\theta_{n-1}^*, \epsilon_{n-1}^*) \cap \Theta)$. The last inequality is due to the assumption that the two problems (SP1) and (SP2) are equivalent.

Choose $\epsilon = \min\{\epsilon_0^*, \ldots \epsilon_{n-1}^*\}$. Then, $J_0^\pi(x_0) \le J_0^{\tilde{\pi}}(x_0)$ for all $\tilde{\pi} = (\tilde{\theta}_0, \ldots, \tilde{\theta}_{n-1}) \in (B(\theta_0^*, \epsilon) \cap \Theta) \times \cdots \times (B(\theta_{n-1}^*, \epsilon) \cap \Theta)$. This completes the proof. ■

Now, let $\mathbf{D}_x^\mu(\theta)$ be the Jacobian matrix of $\mu_{(\cdot)}(x)$ at point θ, $\mathbf{D}_k^{f,x}(x, \mu_\theta(x), w)$ be the Jacobian matrix of the function $f_k(\cdot, \mu_\theta(\cdot), w)$ at point x while viewing θ as a constant, and similarly $\mathbf{D}_k^{f,\theta}(x, \mu_\theta(x), w)$ be the Jacobian matrix of $f_k(x, \mu_{(\cdot)}(x), w)$ at point θ while viewing x as a constant.

Theorem 8: Consider a stationary point of DP $\pi = (\theta_0^*, \ldots, \theta_{n-1}^*)$. If for all $k \in \{0, \ldots, n-1\}$, the following holds:

1) $\mu_{\theta_k}(x_k^*)$ is continuously differentiable with respect to θ_k in a neighborhood of θ_k^* for all $x_k^* \in \mathbb{R}^N$;
2) $\mu_{\theta_k^*}(x_k)$ is continuously differentiable with respect to x_k everywhere;

then π is a stationary point of the one-shot optimization.

Proof: First, we will apply induction to prove that for every $k \in \{1, \ldots, n\}$, $J_k^\pi(x)$ is continuously differentiable. For the base step, $J_n^\pi(x) = c_n(x)$ is continuously differentiable. For the induction step, observe that

$$\nabla_x J_k^\pi(x) = \nabla_x[Q_k^\pi(x, \mu_{\theta_k^*}(x))]$$
$$= \nabla_x[c_k(x, \mu_{\theta_k^*}(x)) + \int_\Omega J_{k+1}^\pi(f_k(x, \mu_{\theta_k^*}(x), w_k))\mathrm{d}p(w_k)]$$
$$= \nabla_x[c_k(x, \mu_{\theta_k^*}(x))] + \int_\Omega \nabla_x[J_{k+1}^\pi(f_k(x, \mu_{\theta_k^*}(x), w_k))]\mathrm{d}p(w_k)$$
$$= \nabla_x[c_k(x, \mu_{\theta_k^*}(x))] + \int_\Omega \mathbf{D}_k^{f,x}(x, \mu_{\theta_k^*}(x), w_k)^T \times \nabla_x J_{k+1}^\pi(f_k(x, \mu_{\theta_k^*}(x), w_k))\mathrm{d}p(w_k).$$

This observation is based on the existence and continuity of the Jacobian matrix $\mathbf{D}_k^{f,x}(x, \mu_{\theta_k^*}(x), w_k)$ due to Assumption 2, continuity of $\nabla_x J_{k+1}^\pi(f_k(x, \mu_{\theta_k^*}(x), w_k))$ due to the induction step, and therefore the continuity of $\nabla_x[J_{k+1}^\pi(f_k(x, \mu_{\theta_k^*}(x), w_k))]$. This allows us to interchange integration and differentiation in the second equality by Leibniz's integration rule.

Now, for $k \in \{0, \ldots, n-1\}$, observe that

$$\nabla_{\theta_k} Q_k^\pi(x_k, \mu_{\theta_k^*}(x_k))$$
$$= \nabla_{\theta_k}[c(x_k, \mu_{\theta_k^*}(x_k)) + \int_\Omega J_{k+1}^\pi(f_k(x_k, \mu_{\theta_k^*}(x_k), w_k))\mathrm{d}p(w_k)]$$
$$= \mathbf{D}_{x_k}^\mu(\theta_k^*)^T \nabla_\mu c(x_k, \mu_{\theta_k^*}(x_k)) + \int_\Omega \mathbf{D}_k^{f,\theta}(x_k, \mu_{\theta_k^*}(x_k), w_k)^T \times \nabla_x J_{k+1}^\pi(f_k(x_k, \mu_{\theta_k^*}(x_k), w_k))\mathrm{d}p(w_k)$$

which is valid because for $k \in \{1, \ldots, n\}$, $J_k^\pi(x)$ is continuously differentiable and Assumption 1 implies the existence and continuity of $\mathbf{D}_{x_k}^\mu(\theta_k^*)$ and $\mathbf{D}_k^{f,\theta}(x_k, \mu_{\theta_k^*}(x_k), w_k)$. Thus, $\nabla_{\theta_k} Q_k^\pi(x_k, \mu_{\theta_k}(x_k))$ is continuous in a neighborhood of θ_k^* for all $x_k \in \mathbb{R}^N$. Then, for $k \in \{0, \ldots, n-1\}$

$$\nabla_{\theta_k} \mathbb{E}_{x_0}[J_0^\pi(x_0)] = \int_{\mathbb{R}^N} \int_\Omega \cdots \int_\Omega \nabla_{\theta_k} Q_k^\pi(x_k, \mu_{\theta_k^*}(x_k))\mathrm{d}p(w_{k-1}) \ldots \mathrm{d}p(w_0)\mathrm{d}p(x_0).$$

Now, note that $\mathcal{N}_\Theta(\theta_k^*)$ is nonempty, closed, and convex [32]. By the definition of a stationary point of DP, we have $-\nabla_{\theta_k} Q_k^\pi(x_k, \mu_{\theta_k^*}(x_k)) \in \mathcal{N}_\Theta(\theta_k^*)$ for all $x_k \in \mathbb{R}^N$. To prove by contradiction, assume that $-\nabla_{\theta_k} \mathbb{E}_{x_0}[J_0^\pi(x_0)] \notin \mathcal{N}_\Theta(\theta_k^*)$. Let a_k denote the dimension of θ_k. By the separating hyperplane theorem, there exist $p \in \mathbb{R}^{a_k}$ and $\alpha \in \mathbb{R}$ such that

$$-p^T \nabla_{\theta_k} Q_k^\pi(x_k, \mu_{\theta_k^*}(x_k)) < \alpha < -p^T \nabla_{\theta_k} \mathbb{E}_{x_0}[J_0^\pi(x_0)]$$

for all $x_k \in \mathbb{R}^N$. Then, observe that

$$-p^T \nabla_{\theta_k} \mathbb{E}_{x_0}[J_0^\pi(x_0)]$$
$$= -p^T \int_{\mathbb{R}^N} \int_\Omega \cdots \int_\Omega \nabla_{\theta_k} Q_k^\pi(x_k, \mu_{\theta_k^*}(x_k))\mathrm{d}p(w_{k-1}) \ldots \mathrm{d}p(x_0)$$
$$= \int_{\mathbb{R}^N} \int_\Omega \cdots \int_\Omega -p^T \nabla_{\theta_k} Q_k^\pi(x_k, \mu_{\theta_k^*}(x_k))\mathrm{d}p(w_{k-1}) \ldots \mathrm{d}p(x_0)$$
$$< \int_{\mathbb{R}^N} \int_\Omega \cdots \int_\Omega -p^T \nabla_{\theta_k} \mathbb{E}_{x_0}[J_0^\pi(x_0)]\mathrm{d}p(w_{k-1}) \ldots \mathrm{d}p(x_0)$$
$$= -p^T \nabla_{\theta_k} \mathbb{E}_{x_0}[J_0^\pi(x_0)]$$

which is a contradiction. Thus, $-\nabla_{\theta_k} \mathbb{E}_{x_0}[J_0^\pi(x_0)] \in \mathcal{N}_\Theta(\theta_k^*)$, which shows that $\pi = (\theta_0^*, \ldots, \theta_{n-1}^*)$ is a stationary point of the one-shot optimization. ■

C. From One-Shot Optimization to DP

In this section, we first show that a local minimizer (stationary point) of the one-shot optimization does not necessarily correspond to a local minimizer (stationary point) of DP; i.e., the converse of Theorem 7 and that of Theorem 8 do not hold. Then, the optimization landscape of the one-shot optimization is more complex than that of DP. In other words, if the one-shot problem has a *low* complexity, so does the DP problem.

To provide a counterexample, we use the basic parameterized policy that follows Definition 13: $\mu_{\theta_k}(x) = a_k x + b_k$, where $\theta_k = (a_k, b_k)$. Consider the two-step problem

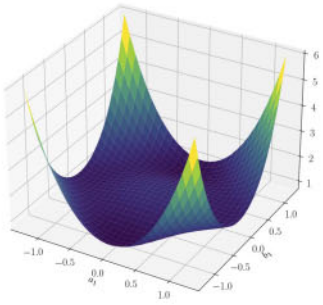


Fig. 3. Landscape of the one-shot optimization for a stochastic parameterized problem: b_0 is fixed to 0 in the figure. $(a_1, b_1) = (\pm 1, 0), (0, \pm 1)$ are strict local minimizers of the one-shot optimization but only $(0, \pm 1)$ is a local minimizer of DP.

$$x_0 = 0,\ c_0(x, \mu_{\theta_0}(x)) = 0$$

$$f_0(x, \mu_{\theta_0}(x), w_0) = x + a_0 x + b_0 + w_0$$

$$c_1(x, \mu_{\theta_1}(x)) = \frac{1}{4}(a_1 x + b_1)^4 - \frac{1}{2}(a_1 x + b_1)^2 + x^2$$

$$f_1(x, \mu_{\theta_1}(x), w_1) = x + a_1 x + b_1 + w_1$$

$$c_2(x, \mu_{\theta_2}(x)) = 0 \text{ where } w_0, w_1 \overset{\text{iid}}{\sim} \text{Uniform}\left(-\sqrt{\frac{5}{3}}, \sqrt{\frac{5}{3}}\right)$$

where $\Theta = [-2, 2] \times [-2, 2]$. The associated one-shot problem can be written as

$$\min_{-2 \le b_0, a_1, b_1 \le 2} \mathbb{E}_{w_0}\left[\frac{1}{4}\{a_1(b_0 + w_0) + b_1\}^4 - \frac{1}{2}\{a_1(b_0 + w_0) + b_1\}^2 + (b_0 + w_0)^2\right].$$

It turns out that there are nine stationary points of the one-shot optimization in the interior of Θ: $(b_0, a_1, b_1) = (0, \pm 0.7071, \pm 0.4082), (0, \pm 1, 0), (0, 0, \pm 1), (0, 0, 0)$. Among them, there are four strict local minimizers of the one-shot optimization: $(0, \pm 1, 0), (0, 0, \pm 1)$. On the other hand, considering $\nabla_{\theta_1} c_1(x, \mu_{\theta_1}(x)) = [g(x, a_1, b_1)x, g(x, a_1, b_1)]$, where $g(x, a_1, b_1) = (a_1 x + b_1)(a_1 x + b_1 - 1)(a_1 x + b_1 + 1)$, there are three stationary points of DP: $(0, 0, \pm 1), (0, 0, 0)$ and two strict local minimizers of DP: $(0, 0, \pm 1)$. This verifies that a local minimizer (stationary point) of DP is indeed a local minimizer (stationary point) of the one-shot optimization but not the other way around. Fig. 3 shows the landscape of the one-shot optimization when b_0 is fixed to 0.

Now, we present the specific case that a local minimizer of the one-shot optimization implies a local minimizer of DP, similar to Theorem 6. The preconditions of theorems are similar in the sense that they both consider the case when DP has a very low complexity in the sense that there is no spurious local minima at each step of DP. The main difference between the theorems comes from whether we consider a single trajectory or the expectation over infinitely many trajectories. We consider this in the view of stationarity. (see Definitions 6, 11, and 12)

Assumption 3: There exists only a single stationary control policy $\phi = (\phi_0, \ldots, \phi_{n-1})$, which is also a locally minimum control policy in the interior of A for all $x \in \mathbb{R}^N$. The parameterized policy defined by Definition 13 contains ϕ, with the associated parameters denoted by $\phi' = (\theta'_0, \ldots, \theta'_{n-1})$.

Theorem 9: Assume that Assumption 3 holds. Consider a local minimizer of the one-shot optimization $\pi = (\theta^*_0, \ldots, \theta^*_{n-1})$ in the interior of Θ^n. If x^*_k is a continuous random variable for all $k \in \{0, \ldots, n-1\}$, where $(x^*_0, \ldots, x^*_n)$ is the random state process associated with π, then π is a local minimizer of DP.

Proof: Since ϕ is a single locally minimum control policy, $Q^{\phi'}_k(x, u)$ has no spurious local minima for all $k \in \{0, \ldots, n-1\}$. Thus, by Proposition 1, the corresponding $\inf_{x \in \mathbb{R}^N} \epsilon^*_k(x) = \infty > 0$ makes ϕ' be a local minimizer of DP.

Consider a stationary point of the one-shot optimization $\pi = (\theta^*_0, \ldots, \theta^*_{n-1})$ in the interior of Θ^n. We will now prove by a backward induction that π should always be ϕ'; i.e., $\theta^*_k = \theta'_k$ for all $k \in \{0, \ldots, n-1\}$.

For the base step, at step $n-1$, since the parameterized policy contains ϕ_{n-1}, it can be expressed as $\phi_{n-1}(x) = \sum_{i=1}^m s_i f_i(x)$, where $f_i : \mathbb{R}^N \to \mathbb{R}^M$, $i = 1, \ldots, m$, $f_i(x) = [f_{i1}(x), \ldots, f_{iM}(x)]^T$, and $\theta'_{n-1} = (s_1, \ldots, s_m) \in \Theta$. Notice that $Q^{\phi'}_{n-1}(x, \mu_{\theta'_{n-1}}(x)) = Q^{\pi}_{n-1}(x, \mu_{\theta'_{n-1}}(x))$ since θ'_{n-1} is the final parameter of the whole system to determine the control inputs and the state transition. Now, observe that

$$\nabla_\mu Q^{\phi'}_{n-1}(x, \mu_{\theta'_{n-1}}(x)) = \nabla_\mu Q^{\pi}_{n-1}(x, \mu_{\theta'_{n-1}}(x)) = 0$$

and $\mu_{\theta'_{n-1}}(x)$ is the unique solution for $\nabla_\mu Q^{\phi'}_{n-1}(x, \cdot) = 0$ since $\mu_{\theta'_{n-1}}(x)$ is the unique stationary point located within the interior of A due to Assumption 3. This yields the following expression with $\mu_\theta(x) = (u_1, \ldots, u_M)^T$:

$$\nabla_\mu Q^{\phi'}_{n-1}(x, \mu_\theta(x)) = \nabla_\mu Q^{\pi}_{n-1}(x, \mu_\theta(x)) = \begin{bmatrix} (u_1 - \sum_{i=1}^m s_i f_{i1}(x)) \cdot g_1(x, \theta) \\ \vdots \\ (u_M - \sum_{i=1}^m s_i f_{iM}(x)) \cdot g_M(x, \theta) \end{bmatrix}$$

where $g_j(x, \theta), j = 1, \ldots, M$, are nonnegative at $\theta = \theta'_{n-1}$ and positive at all the other points since ϕ' is a local minimizer of DP that yields the unique stationary control policy ϕ.

Now, let $\mu_{\theta^*_{n-1}}(x)$ be $\sum_{i=1}^m d_i f_i(x)$, where $\theta^*_{n-1} = (d_1, \ldots, d_m)$. Observe that according to the chain rule, the following expression holds:

$$\begin{aligned} &\nabla_{\theta_{n-1}} Q^{\pi}_{n-1}(x, \mu_{\theta^*_{n-1}}(x))^T \\ &= \nabla_\mu Q^{\pi}_{n-1}(x, \mu_{\theta^*_{n-1}}(x))^T \mathbf{D}^\mu_x(\theta^*_{n-1}) \\ &= \begin{bmatrix} (\sum_{i=1}^m (d_i - s_i) f_{i1}(x)) \cdot g_1(x, \theta^*_{n-1}) \\ \vdots \\ (\sum_{i=1}^m (d_i - s_i) f_{iM}(x)) \cdot g_M(x, \theta^*_{n-1}) \end{bmatrix}^T \\ &\quad \times \begin{bmatrix} f_{11}(x) \cdots f_{m1}(x) \\ \vdots \quad \ddots \quad \vdots \\ f_{1M}(x) \cdots f_{mM}(x) \end{bmatrix}. \end{aligned} \tag{17}$$

Now, notice that $\nabla_{\theta_{n-1}} \mathbb{E}_{x_0}[J^\pi_0(x_0)] = 0$ since π is a stationary point of the one-shot optimization in the interior of Θ^n. Then, observe that

$$\begin{aligned} &\nabla_{\theta_{n-1}} \mathbb{E}_{x_0}[J^\pi_0(x_0)] \\ &= \nabla_{\theta_{n-1}} \mathbb{E}_{x_0, w_0, \ldots, w_{n-2}}[Q^\pi_{n-1}(x^*_{n-1}, \mu_{\theta^*_{n-1}}(x^*_{n-1}))] \\ &= \mathbb{E}_{x_0, w_0, \ldots, w_{n-2}}[\nabla_{\theta_{n-1}} Q^\pi_{n-1}(x^*_{n-1}, \mu_{\theta^*_{n-1}}(x^*_{n-1}))] = 0. \end{aligned} \tag{18}$$

The second equality comes from $Q^{\pi}_{n-1}(x, \mu_\theta(x))$ being differentiable with respect to the parameters due to the linearity of the policy defined by Definition 13. Now, we substitute (17) into (18) to derive an m-dimensional vector equation, and multiply $(d_k - s_k)$ with the kth component as follows:

$$\mathbb{E}_{x_0,w_0,\ldots,w_{n-2}}\Bigg[\sum_{j=1}^{M}(d_k - s_k)f_{kj}(x^*_{n-1}) \times \sum_{i=1}^{m}(d_i - s_i)f_{ij}(x^*_{n-1}) \cdot g_j(x^*_{n-1}, \theta^*_{n-1})\Bigg] = 0$$

for all $k = 1, \ldots, m$. We sum up the m equations and rearrange the terms to derive the following equation:

$$\sum_{j=1}^{M}\mathbb{E}_{x_0,w_0,\ldots,w_{n-2}}\Bigg[\Bigg(\sum_{i=1}^{m}(d_i - s_i)f_{ij}(x^*_{n-1})\Bigg)^2 \times g_j(x^*_{n-1}, \theta^*_{n-1})\Bigg] = 0. \quad (19)$$

The term inside the expectation is always nonnegative regardless of the distribution of $x_0, w_0, \ldots, w_{n-2}$. Now, suppose that $\theta^*_{n-1} \neq \theta'_{n-1}$; i.e., $d_i \neq s_i$ for some $i \in \{1, \ldots, m\}$. Then, we have $g_j(\cdot, \theta^*_{n-1})$ to be strictly positive. As a result, for (19) to be satisfied, $\sum_{i=1}^{m}(d_i - s_i)f_{ij}(x^*_{n-1})$ should be 0 for every $j \in \{1, \ldots, M\}$ for all possible values of x_{n-1}. Recall from Remark 8 to note that it is impossible to satisfy (19) since x^*_{n-1} is a continuous random variable and the policy is defined by Definition 13, i.e., a linear combination of some independent basis functions. Thus, $d_i = s_i$ for all $i \in \{1, \ldots, m\}$, which means $\theta^*_{n-1} = \theta'_{n-1}$.

For the induction step, assume that $\theta^*_k = \theta'_k$. Again, $Q^{\phi'}_k(x, \mu_{\theta_k}(x)) = Q^{\pi}_k(x, \mu_{\theta_k}(x))$ holds, and thus one can apply the same logic as the base step to obtain $\theta^*_{k-1} = \theta'_{k-1}$.

Thus, $\pi = \phi'$ holds, which implies that π is a local minimizer of DP since ϕ' is a local minimizer of DP. ■

Remark 10: The results of both Theorems 6 and 9 state that a local minimizer of the one-shot optimization is indeed a local minimizer of DP under the common assumption that no spurious local minima exist at each step of DP. By taking the contrapositive, one can observe that under such a condition, there is at most one local minimizer of the one-shot optimization, indicating that no spurious local minima exist; i.e., if DP has a *very low* complexity, the same holds for the one-shot problem.

Remark 11: To determine the form of $\nabla_\mu Q^{\phi'}_{n-1}(x, u)$, it was necessary to argue that $\mu_{\theta'_{n-1}}(x)$ should be the unique solution for $\nabla_\mu Q^{\phi'}_{n-1}(x, u) = 0$. For this to be true, there should certainly be only a single stationary control policy, which necessitates Assumption 3. In fact, it may be difficult to satisfy the precondition that a single stationary control policy should be in the interior of A for all $x \in \mathbb{R}^N$. Instead, we can relax this condition to apply only within the domain of x; i.e., the set of values that at least one of the states $x_0, x_1, \ldots, x_n$ can take. For example, if the state space is finite, satisfying the condition becomes relatively straightforward.

Remark 12: The challenging part of a backward induction in the proof arises from the fact that the state at step k is fully determined by the previous steps but one cannot look at the previous steps in the backward induction. Thus, the main idea of the proof leverages (19), which incurs the fact that $\theta^*_k = \theta'_k$ regardless of the distribution of $x_0, w_0, \ldots, w_{k-1}$. Thus, we only need the assumption that there is a single stationary control policy with respect to the given distribution of $x_0, w_0, \ldots, w_{n-1}$. This is a big improvement from the work [28] (see Condition 4 of Section 5.4) in the sense that Condition 4 needs no suboptimal stationary point with respect to any possible distribution.

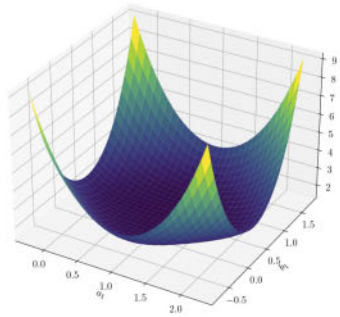


Fig. 4. Landscape of the one-shot optimization under the assumptions of Theorem 9: b_0 is fixed to 0 in the figure. $(a_1, b_1) = (1, 0.5)$ is the only stationary point (local minimizer) of DP and also the only stationary point (local minimizer) of the one-shot optimization.

Now, we present the pictorial example of Theorem 9. Consider the example of two-step problem presented above in Fig. 3, but modify $c_1(x, \mu_{\theta_1}(x))$ to $\frac{1}{4}(a_1 x + b_1 - x - 0.5)^4 + x^4$. The associated one-shot problem can be written as

$$\min_{-2\le b_0,a_1,b_1\le 2} \mathbb{E}_{w_0}\Bigg[\frac{1}{4}\{(a_1 - 1)(b_0 + w_0) + b_1 - 0.5\}^4 + (b_0 + w_0)^4\Bigg].$$

$(\pi_0, \pi_1) = (0, x + 0.5)$ is the only locally minimum control policy, and the parameterized policy class contains this policy as $(b_0, a_1, b_1) = (0, 1, 0.5)$. Clearly, it is a local minimizer of DP. It turns out that the corresponding one-shot problem also has a single stationary point (0,1,0.5), which is also a local minimizer of the one-shot optimization. Fig. 4 shows the landscape of the one-shot optimization when b_0 is fixed to 0.

Considering both Theorems 7 and 9, one can conclude that under the assumptions of Theorem 9, a local minimizer of DP is *equivalent* to a local minimizer of the one-shot optimization.

D. Numerical Experiments

In this section, we will present a high-dimensional experiment on the classical linear quadratic regulator (LQR)

$$f_k(x_k, u_k) = A_k x_k + B_k u_k, \quad k = 0, \ldots, n-1, \quad x_0 \sim \mathcal{D}$$

$$c_k(x_k, u_k) = x_k^\top Q_k x_k + u_k^\top R_k u_k, \quad k = 0, \ldots, n-1$$

$$u_k = K_k x_k, \quad k = 0, \ldots, n-1, \quad c_n(x_n) = x_n^\top Q_n x_n$$

whose goal is to find the optimal parameters: $K_0, \ldots, K_{n-1}$. To solve the problem using DP, we use Xpress Optimizer v9.3.0 [40]. To solve the problem in a one-shot fashion, we use Gurobi Optimizer v11.0.0 [41] with the tolerance of 10^{-4}.

Let $K^*_{k,\mathrm{DP}}$ and $K^*_{k,\mathrm{OS}}$ denote an observed local solution of the kth step parameter K_k obtained by DP and the one-shot problem, respectively. We aim to determine whether each local solution of DP corresponds to some local solutions of the one-shot

TABLE II
RELATIONSHIP BETWEEN DP AND ONE-SHOT SOLUTIONS OF LQR

Numerical difference \ Scenario	(a) Unconstrained	(b) Constrained
$\|K^*_{0,\mathrm{DP}} - K^*_{0,\mathrm{DP}\to\mathrm{OS}}\|_F / \|K^*_{0,\mathrm{DP}}\|_F$	0	0
$\|K^*_{1,\mathrm{DP}} - K^*_{1,\mathrm{DP}\to\mathrm{OS}}\|_F / \|K^*_{1,\mathrm{DP}}\|_F$	0	0
$\|K^*_{0,\mathrm{OS}} - K^*_{0,\mathrm{OS}\to\mathrm{DP}}\|_F / \|K^*_{0,\mathrm{OS}}\|_F$	$2.901\cdot10^{-6}$	$5.399\cdot10^{-4}$
$\|K^*_{1,\mathrm{OS}} - K^*_{1,\mathrm{OS}\to\mathrm{DP}}\|_F / \|K^*_{1,\mathrm{OS}}\|_F$	$2.291\cdot10^{-5}$	**3.156**

The bold value (3.156) being far from 0 supports that the landscape of the one-shot problem is more complex than its DP counterpart. One can find out that the other values in Table II are 0 or close to 0.

problem, and vice versa. One can verify this by first solving DP or the one-shot problem and then providing its solution as the initial parameter values when solving its counterpart. This method is often referred to as "warm start." We expect to observe unchanged values from the initial guess if there is indeed a correspondence. Let $K^*_{k,\mathrm{DP}\to\mathrm{OS}}$ ($K^*_{k,\mathrm{OS}\to\mathrm{DP}}$) denote the solution of K_k obtained by the one-shot problem (DP) using warm start with DP (one-shot) solution as an initial guess. The initial distribution $\mathcal{D}$ introduces the stochasticity to the system and induces the states to be continuous random variables, which obeys the assumption of Theorem 9.

We perform 20 experiments for $x_k \in \mathbb{R}^3$ and $u_k \in \mathbb{R}^4$ with $n = 30$. We randomly generate A_k and B_k, whose entries are all in $[-100, 100]$. We also generate $Q_k = QQ^T$ and $R_k = RR^T + 100I$, where all entries of Q and R are in $[-20, 20]$ and I denotes the identity matrix. $\mathcal{D}$ is the normal distribution with the expectation $20\mathbb{1}$ and the variance VV^T, where $\mathbb{1}$ denotes the vector of ones and all entries in V are in $[-200, 200]$. We consider two scenarios: (a) unconstrained and (b) constrained by the last-step (nonconvex) condition $K^T_{n-1}K_{n-1} \succeq 10\,000I$, where $\succeq$ denotes the Loewner partial ordering (roughly speaking, this condition ensures that the controller has a high gain). Table II gives whether the correspondence holds between the solutions of DP and the one-shot problem using warm start under the two scenarios, presenting the results of the average of 20 experiments.

It turns out that for both scenarios, one can observe that a solution of DP corresponds to each solution of one-shot problem since the one-shot solver directly identifies DP solution as a local solution of the one-shot problem without any numerical update. This implication supports the findings of Theorem 7.

However, whether a one-shot solution implies some DP solutions depends on the problem setting. Our experiment for the unconstrained case shows that a solution of the one-shot problem indeed corresponds to that of DP, implying with the above result that a one-shot solution is equivalent to a DP solution. Previous studies have shown the global convergence of this one-shot problem by proving that LQR satisfies the gradient dominance property, even though the problem is generally nonconvex [42], [43]. Our general approach alternatively observes the DP counterparts. Since every DP subproblem of LQR has no spurious local minima, our experiment implies that the one-shot LQR problem also has none of them and achieves the global convergence, which supports Theorem 9.

On the other hand, the nonconvex constraint on K_{n-1} creates spurious solutions for the $(n-1)^{\text{th}}$ DP, independent of whether the $(n-2)^{\text{th}}, \ldots, 0$th DP steps have any spurious local minima. Our experiment for the constrained case shows that having spurious local minima at the $(n-1)^{\text{th}}$ DP propagates backward to K_1, where the solver fails to guarantee that an observed local minimum of the one-shot optimization corresponds to that of DP. This illustrates that the landscape of the one-shot problem has a higher complexity than its DP counterpart, which was also shown in Figs. 2(b) and 3. This result serves as a counterexample of the converse of Theorems 7 and 8, and it implies that a single high-complexity DP step affects the landscape of the one-shot problem.

V. CONCLUSION

In this article, we studied the optimization landscape of the optimal control problems via two different formulations: one-shot optimization aimed at solving for all input values at the same time, and DP method aimed at finding the input values sequentially. For the deterministic problem, we proved that under some mild conditions, each local minimizer of the one-shot optimization corresponds to an input sequence induced by some locally minimum control policy of DP, and vice versa.

To help better understand the quality of the local solutions obtained by RL algorithms, we incorporated exact parameterized policies into the optimal control problem for both deterministic and stochastic dynamics. We showed that if the one-shot problem has a low complexity, so do the corresponding DP subproblems, indicating the success of DP methods. Moreover, under the condition that there exists only a single locally minimum control policy, with different technical assumptions, both deterministic and stochastic cases yield that a local minimizer of the one-shot optimization is equivalent to a local minimizer of DP.

We focused on the discrete-time finite-horizon optimal control problem in this work. A natural future direction would be to extend this work to the continuous-time and infinite-horizon cases, which was discussed in Remark 4. For safety-critical systems, state constraints may also be enforced, with recursive feasibility being crucial to guarantee the success of DP. One may also want to extend the parameterized policy class beyond a linear combination of basis functions, such as composite functions, widely used in deep neural networks.

REFERENCES

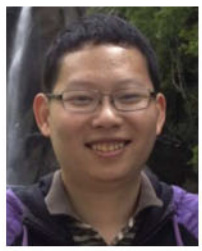
[1] Y. Ding, Y. Bi, and J. Lavaei, "Analysis of spurious local solutions of optimal control problems: One-shot optimization versus dynamic programming," in *Proc. IEEE Amer. Control Conf.*, 2021, pp. 3836–3843.
[2] R. E. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1957.
[3] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Two Volume Set*, 4th ed. Amsterdam, The Netherlands: Athena Scientific, 2012.
[4] R. Bellman and R. Kalaba, "On the role of dynamic programming in statistical communication theory," *IRE Trans. Inf. Theory*, vol. 3, no. 3, pp. 197–203, 1957.
[5] E. A. Feinberg, "Optimality conditions for inventory control," in *Proc. Optim. Challenges Complex, Netw. Risky Syst.*, 2016, pp. 14–45.
[6] I. Kolmanovsky, I. Siverguina, and B. Lygoe, "Optimization of powertrain operating policy for feasibility assessment and calibration: Stochastic dynamic programming approach," in *Proc. IEEE Amer. Control Conf.*, vol. 2, 2002, pp. 1425–1430.
[7] D. P. Bertsekas, *Reinforcement Learning and Optimal Control*. Belmont, MA, USA: Athena Scientific, 2019.
[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA:MIT Press, 2018.
[9] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[10] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 742–769, Feb. 2019.

[11] C. Liu, L. Zhu, and M. Belkin, "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks," *Appl. Comput. Harmon. Anal.*, vol. 59, pp. 85–116, 2022.
[12] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 242–252.
[13] S. Liang, R. Sun, and R. Srikant, "Revisiting landscape analysis in deep neural networks: Eliminating decreasing paths to infinity," *SIAM J. Optim.*, vol. 32, no. 4, pp. 2797–2827, 2022.
[14] S. Mei, Y. Bai, and A. Montanari, "The landscape of empirical risk for non-convex losses," *Ann. Statist.*, vol. 46, no. 6 A, pp. 2747–2774, 2018.
[15] K. A. Chandrasekher, A. Pananjady, and C. Thrampoulidis, "Sharp global convergence guarantees for iterative nonconvex optimization with random data," *Ann. Statist.*, vol. 51, no. 1, pp. 179–210, 2023.
[16] I. Molybog, S. Sojoudi, and J. Lavaei, "Role of sparsity and structure in the optimization landscape of non-convex matrix sensing," *Math. Program.*, vol. 193, pp. 75–111, 2022.
[17] R. Y. Zhang, "Sharp global guarantees for nonconvex low-rank matrix recovery in the overparameterized regime," 2021, *arXiv:2104.10790*.
[18] R. Y. Zhang, "Improved global guarantees for the nonconvex Burer-Monteiro factorization via rank overparameterization," 2022, *arXiv:2207.01789*.
[19] Z. Ma, Y. Bi, J. Lavaei, and S. Sojoudi, "Sharp restricted isometry property bounds for low-rank matrix recovery problems with corrupted measurements," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 7, 2022, pp. 7672–7681.
[20] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval," *Math. Program.*, vol. 176, no. 1-2, pp. 5–37, 2019.
[21] Y. Ding, J. Lavaei, and M. Arcak, "Time-variation in online nonconvex optimization enables escaping from spurious local minima," *IEEE Trans. Autom. Control*, vol. 68, no. 1, pp. 156–171, Jan. 2023.
[22] S. Fattahi, C. Josz, Y. Ding, R. Mohammadi, J. Lavaei, and S. Sojoudi, "On the absence of spurious local trajectories in time-varying nonconvex optimization," *IEEE Trans. Autom. Control*, vol. 68, no. 1, pp. 80–95, Jan. 2023.
[23] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *J. Mach. Learn. Res.*, vol. 22, no. 98, pp. 1–76, 2021.
[24] J. Bhandari and D. Russo, "On the linear convergence of policy gradient methods for finite MDPs," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 2386–2394.
[25] L. Xiao, "On the convergence rates of policy gradient methods," *J. Mach. Learn. Res.*, vol. 23, pp. 1–36, 2022.
[26] G. Lan, "Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes," *Math. Program.*, vol. 198, pp. 1059–1106, 2023.
[27] W. Zhan, S. Cen, B. Huang, Y. Chen, J. D. Lee, and Y. Chi, "Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence," *SIAM J. Optim.*, vol. 33, no. 2, pp. 1061–1091, 2023.
[28] J. Bhandari and D. Russo, "Global optimality guarantees for policy gradient methods," *Operations Res.*, 2024, doi: 10.1287/opre.2021.0014.
[29] D. Silver et al., "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 387–395.
[30] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 1057–1063.
[31] B. Yalçin, H. Zhang, J. Lavaei, and S. Sojoudi, "Factorization approach for low-complexity matrix completion problems: Exponential number of spurious solutions and failure of gradient methods," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2022, pp. 319–341.
[32] R. T. Rockafellar and R J.B. *Wets, Variational Analysis*, vol. 317. Berlin, Germany: Springer, 2009.
[33] C. Berge, *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*. North Chelmsford, MA, USA: Courier Corporation, 1997.
[34] A. Bressan and B. Piccoli, *Introduction to the Mathematical Theory of Control*. San Jose, CA, USA: American Inst. Math. Sci., 2007.
[35] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Englewood Cliffs, NJ, USA: Prentice Hall, 1986.
[36] G. Sansone, *Orthogonal Functions, Rev*. English ed. Mineola, NJ, USA: Dover, 1991.
[37] B. Schölkopf and A. J. Smola, *Learning With Kernels*. Cambridge, MA, USA:Academic Press, 1967.
[38] G. Strang, *Linear Algebra and Its Applications*, 3rd ed. San Diego, CA, USA: Harcourt, 1988.
[39] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete-Time Case*. Belmont, MA, USA: Athena Scientific, 1996.
[40] FICO, "Xpress optimizer reference manual," 2024. [Online]. Available: https://www.fico.com/fico-xpress-optimization/docs
[41] Gurobi Optimization LLC, "Gurobi Optimizer Reference Manual," 2023. [Online]. Available: https://www.gurobi.com/documentation
[42] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1467–1476.
[43] B. Hambly, R. Xu, and H. Yang, "Policy gradient methods for the noisy linear quadratic regulator over a finite horizon," *SIAM J. Control Optim.*, vol. 59, no. 5, pp. 3359–3391, 2021.

Jihun Kim received the B.S. degree in both industrial engineering and statistics from Seoul National University, Seoul, South Korea, in 2022. He is currently working toward the Ph.D. degree in industrial engineering and operations research with the University of California, Berkeley, CA, USA.

His research interests include nonconvex optimization, dynamical systems, and learning-based control, along with their applications to safety-critical systems, such as robotics, aircraft, and autonomous driving.

Yuhao Ding received the B.E. degree in aerospace engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016, the M.S. degree in electrical and computer engineering from the University of Michigan, Ann Arbor, MI, USA, in 2018, and the Ph.D. degree in industrial engineering and operations research with the University of California, Berkeley, CA, USA.



Yingjie Bi received the B.S. degree in microelectronics from Peking University, Beijing, China, in 2014, and the Ph.D. degree in electrical and computer engineering from Cornell University, Ithaca, NY, USA, in 2020.

He is currently a Postdoctoral Scholar with the Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA, USA. His research interests include nonconvex optimization, machine learning, computer networks, and control theory.

Javad Lavaei (Fellow, IEEE) received the Ph.D. degree in control and dynamical systems from the California Institute of Technology, Pasadena, CA, USA, in 2011.

He is currently an Associate Professor with the Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA.

Dr. Lavaei is a Senior Editor of IEEE SYSTEMS JOURNAL and was on the editorial boards of IEEE TRANSACTIONS ON AUTOMATIC CONTROL, IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS, IEEE TRANSACTIONS ON SMART GRID, and IEEE CONTROL SYSTEMS LETTERS.