

The landscape of the optimal control problem: One-shot Optimization versus Dynamic Programming

Jihun Kim, Yuhao Ding, Yingjie Bi, and Javad Lavaei

Abstract—Dynamic programming (DP) has a rich theoretical foundation and a broad range of applications, especially in the classic area of optimal control and the newer area of reinforcement learning (RL). Many optimal control problems can be solved via a single optimization problem, named one-shot optimization, or via a sequence of optimization problems using DP. However, the computation of their global optima often faces the NP-hardness issue, 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. We further incorporate the parameterized policy approach into our problem for the closed-loop system. With a parameterized policy, 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 and RL.

Index Terms—Optimal Control, Landscape, One-shot optimization, Dynamic programming

This work was supported by grants from AFOSR, ARO, ONR, and NSF.

J. Kim, Y. Ding, Y. Bi, and J. Lavaei 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).

A preliminary version of this paper has appeared in 2021 American Control Conference, New Orleans, USA, May 25-28, 2021 [1]. The previous version mainly discussed the deterministic problem with control inputs, while this journal version has significantly extended the results to include both the deterministic and the stochastic problems with a parameterized policy to study a closed-loop system. To address the parameterized problem, our new notion of a local minimizer of the one-shot optimization optimizes the objective function over the parameters modeling the inputs. Furthermore, the notion of a locally minimum control policy of DP is replaced with a local minimizer of DP, which considers an open ball in the parameter space instead of the action space. These new notions enable the investigation of the stochastic dynamics as a finite-dimensional problem.

I. INTRODUCTION

Dynamic programming (DP) is a simple mathematical technique that has been widely used in a variety of fields. Following Bellman's influential work [2] on demonstrating the broad scope of DP and laying the foundation of its theory, many mathematical and algorithmic aspects of DP have been investigated [3]. One main application 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. For example, in the challenging domain of classic Atari 2600 games, the work [9] has demonstrated that the deep Q-learning method based on the generalized policy iteration together 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 to that of a professional human games tester.

Although DP has a rich theoretical foundation and a broad range of applications, 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 sub-problems to global optimality, and the computation of their global optima is NP-hard in general, due to the non-linearity of the dynamics and the non-convexity of the cost function.

Therefore, even though the theory of DP 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 non-global local solution is found in any of the sub-problems. Understanding the performance of local search methods for non-convex 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 and recovery [16]–[19], phase retrieval [15], [20], and online optimization [21], [22].

Recently, there has been an increasing interest in under-

standing 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]–[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 sub-optimal stationary points. However, the literature lacks a rigorous analysis of the spurious solutions of the DP method.

In this paper, we analyze the spurious solutions of the DP method by focusing on the following fundamental question: *What if the globally optimal solution of each sub-problem of DP is replaced with a solution obtained by a local search method?* A challenge in this analysis is that policy optimization even towards 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 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 stationary points or 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 such that the total cost is minimized while the dynamics and input constraints are satisfied. One approach to solving the problem is by formulating it as a one-shot optimization problem, a single whole-period problem, and another approach is using the DP to formulate it as a sequential decision-making problem with multiple single-period sub-problems and solve it in a backward way. Although it is well-known that for the deterministic optimal control problem, the one-shot method and the DP method return the same globally optimal control sequence, it is not yet known what would occur if the global optimizer needed for solving each sub-optimization problem in DP is replaced by a local optimizer. In the remainder of the paper, we will 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. To address this relationship, we first 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 result developed in Section II precisely uncovers the connection between the optimization landscapes of the one-shot and DP optimization problems, showing that the DP method using local search can successfully solve the

TABLE I
THEOREMS AND THE CORRESPONDING ASSUMPTIONS WITH RESULTS.

Theorems Assumptions		Deterministic				Deterministic + Parameterized			Stochastic + Parameterized		
		1	2	3	4	5	6	7	8	9	10
Convex	action space		○								
	parameter space							○			
Policy class	Lipschitz	○									
	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				○			○			○

optimal control problem to global optimality if and only if the one-shot optimization is free of spurious solutions.

However, in RL algorithms, both one-shot method and DP are usually handled by a parameterized policy for the closed-loop control, which necessitates generalizing the results to optimization with respect to the parameters instead of the control inputs themselves. We introduce the problem formulation under a parameterized policy and establish the relationship between the two optimization landscapes. In Section III, we incorporate the parameterized policy in the deterministic problem and prove that each local minimizer of DP corresponds to some local minimizer of the one-shot optimization, whereas its converse may not hold. This implies that the optimization landscape of the one-shot problem is more complex than its DP counterpart, implying that if the one-shot problem has a low complexity, so does the DP problem. 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. This implies that if the DP problem has a very low complexity, the same holds for the one-shot problem. Likewise, in Section IV, we incorporate the parameterized policy in the stochastic problem and derive a similar relationship between the optimization landscapes of the one-shot and DP methods as in the deterministic problem. Finally, concluding remarks are provided in Section V. Table I summarizes the main results of the paper.

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 to the one-shot optimization approach. Hence, it is essential to understand when DP combined with a local search solver works. The results of this paper 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.

Notation: 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 $x \in A - B$ means that x is in the set A but not in the set B . Let $\|\cdot\|$ denote the Euclidean norm and $\nabla_x f(x, y)$ denote the gradient of $f(x, y)$ with respect to x . The notation $\nabla_x^2 f(x) \succ 0$ means that the Hessian of $f(x)$ 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:

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

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 $f_i : \mathbb{R}^N \times \mathbb{R}^M \rightarrow \mathbb{R}^N$. Each time instance i is associated with a stage cost $c_i : \mathbb{R}^N \times \mathbb{R}^M \rightarrow \mathbb{R}$ or the terminal cost $c_n : \mathbb{R}^N \rightarrow \mathbb{R}$. 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 paper, 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

$$\begin{aligned} \min \quad & C(x_0; u_0, \dots, u_{n-1}) \\ \text{s.t.} \quad & u_i \in A, \quad i = 0, \dots, n-1. \end{aligned} \quad (\text{P2})$$

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.,

$$\begin{aligned} J_k(x_k) &= \min_{u_k, \dots, u_{n-1}} C(x_k; u_k, \dots, u_{n-1}) \\ \text{s.t.} \quad & u_i \in A, \quad i = k, \dots, n-1. \end{aligned}$$

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

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

for $k = 0, \dots, n-1$. The optimal cost $J_0(x_0)$ equals the optimal objective value of (P1).

However, due to the non-convexity of the look-ahead value functions, 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 is nonconvex or the dynamics is nonlinear. 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 sub-optimal cost-to-go at time $n-1$ as follows:

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

which is obtained based on the local minimizer $\pi_{n-1}(x)$. Subsequently, it is required to solve the optimal decision-making problem $\min_{u \in A} \{c_{n-2}(x_{n-2}, u) + J_{n-1}^\pi(f_{n-2}(x_{n-2}, u))\}$ 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 sub-optimal cost-to-go functions J_k^π for $k = 0, \dots, n-1$. Given the initial state x_0 , let

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

be the control inputs and the states induced by the policies π_0, \dots, π_{n-1} . Then, $(u_0^*, \dots, u_{n-1}^*)$ is a sub-optimal solution to the original optimal control problem (P1) with the sub-optimal objective value $J_0^\pi(x_0)$. 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 $Q_k^\pi(\cdot, \cdot)$ 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:

$$\begin{aligned} J_n^\pi(x) &= c_n(x), \\ Q_k^\pi(x, u) &= c_k(x, u) + J_{k+1}^\pi(f_k(x, u)), \quad k = 0, \dots, n-1, \\ J_k^\pi(x) &= Q_k^\pi(x, \pi_k(x)), \quad k = 0, \dots, n-1. \end{aligned}$$

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

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

for all $\tilde{u}_i \in B(u_i^*, \epsilon) \cap A$ where $i = 0, \dots, n-1$. It is further called a spurious (non-global) 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 $Q_k^\pi(x, \cdot)$, meaning that there exists $\epsilon_k^*(x) > 0$ such that

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

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

In the following subsections, we 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

Although the input sequence induced by a globally minimal control policy is a global minimizer of the one-shot problem, it turns out that the input sequence induced by a spurious locally minimum control policy of DP does not generally imply a spurious local minimizer of the one-shot problem. However, this implication is indeed the case if some mild conditions are satisfied.

Before presenting the theorem, we first show that the differences between two state sequences starting from the same initial state are closely correlated to the differences between the corresponding control sequences.

Lemma 1: Consider the system under an input sequence (u_0, \dots, u_{n-1}) with its associated state sequence (x_0, \dots, x_n) . Then, there exist continuous and non-decreasing functions $L_k(\delta_0, \dots, \delta_k)$, $k = 0, \dots, n-1$, satisfying $L_k(0, \dots, 0) = 0$ and the following property: for any input sequence $(\tilde{u}_0, \dots, \tilde{u}_{n-1})$ with $\tilde{u}_i \in B(u_i, \delta_i) \cap A$ for all $i \in \{0, \dots, n-1\}$, it holds that

$$\|x_{k+1} - \tilde{x}_{k+1}\| \leq L_k(\delta_0, \dots, \delta_k),$$

where $(\tilde{x}_0, \dots, \tilde{x}_n)$ is the state sequence corresponding to $(\tilde{u}_0, \dots, \tilde{u}_{n-1})$ starting from the initial state x_0 .

Proof: The proof is given in [1] (see Lemma 2). ■

Theorem 1: 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^*) . If π_k is Lipschitz continuous in a neighborhood of x_k^* and ϵ_k^* (see Definition 3) is continuous at x_k^* for $k = 0, \dots, n-1$, then $(u_0^*, \dots, u_{n-1}^*)$ is also a (spurious) local minimizer of the one-shot problem.

Proof: We first show that there exist positive constants $\delta_0, \dots, \delta_{n-1}$ such that for every $L > 0$ and $i = 0, \dots, n-1$, the following relation holds:

$$\delta_i + Ld_i \leq \inf_{x \in \bar{B}(x_i^*, d_i)} \epsilon_i^*(x), \quad (1)$$

where $d_0 = 0$ and $d_i = L_{i-1}(\delta_0, \dots, \delta_{i-1})$ for $i > 0$ with L_{i-1} given in the statement of Lemma 1. For $i > 0$, because L_{i-1} is continuous with $L_{i-1}(0, \dots, 0) = 0$, and ϵ_i^* is continuous at x_i^* , if $\delta_0, \dots, \delta_{n-1}$ are sufficiently small,

$$\inf_{x \in \bar{B}(x_i^*, d_i)} \epsilon_i^*(x) \geq \frac{1}{2} \epsilon_i^*(x_i^*) > 0. \quad (2)$$

Thus, at step $n-1$, there must exist $\delta_0, \dots, \delta_{n-1}$ for which (1) holds. Then, at step $n-2$, if $\delta_0, \dots, \delta_{n-2}$ do not make (1) hold, since L_{n-2} is non-decreasing, we can further reduce $\delta_0, \dots, \delta_{n-2}$ to satisfy (1) for $i = n-2$ without violating (1) for $i = n-1$. By repeating this procedure, one can show that there exist positive constants $\delta_0, \dots, \delta_{n-1}$ such that (1) holds for all $i \in \{0, \dots, n-1\}$. Moreover, we can again reduce $\delta_0, \dots, \delta_{n-1}$ such that each π_i is Lipschitz continuous over the set $\bar{B}(x_i^*, d_i)$.

Now, it is desirable to show that for all $\tilde{u}_i \in B(u_i^*, \delta_i) \cap A$ where $i \in \{0, \dots, n-1\}$,

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

Since (1) implies that $\delta_0 \leq \epsilon_0^*(x_0)$, one can write $J_0^\pi(x_0) \leq c_0(x_0, \tilde{u}_0) + J_1^\pi(\tilde{x}_1)$, where $\tilde{x}_1 = f_0(x_0, \tilde{u}_0)$, $\forall \tilde{u}_0 \in B(u_0^*, \delta_0) \cap A$. For every $i \in \{1, \dots, n-1\}$, by the definition of local optimality of $\pi_i(\tilde{x}_i)$,

$$J_i^\pi(\tilde{x}_i) \leq c_i(\tilde{x}_i, \tilde{u}_i) + J_{i+1}^\pi(\tilde{x}_{i+1}), \quad (3)$$

where $\tilde{x}_{i+1} = f_i(\tilde{x}_i, \tilde{u}_i)$, $\forall \tilde{u}_i \in B(\pi_i(\tilde{x}_i), \epsilon_i^*(\tilde{x}_i)) \cap A$.

Now, we aim to show that (3) also holds for all $\tilde{u}_i \in B(\pi_i(x_i^*), \delta_i) \cap A$; or equivalently, $B(\pi_i(x_i^*), \delta_i) \cap A \subseteq B(\pi_i(\tilde{x}_i), \epsilon_i^*(\tilde{x}_i)) \cap A$. It suffices to prove that

$$\delta_i + \|\pi_i(x_i^*) - \pi_i(\tilde{x}_i)\| \leq \epsilon_i^*(\tilde{x}_i). \quad (4)$$

Because of the Lipschitz continuity of π_i in $\bar{B}(x_i^*, d_i)$, there exists a positive constant L_π such that $\|\pi_i(x_i^*) - \pi_i(\tilde{x}_i)\| \leq L_\pi \|x_i^* - \tilde{x}_i\|$, for all $i = 0, \dots, n-1$. Then, the inequality (4) must hold because of (1).

Now, assume that π is a spurious locally minimum control policy of DP, i.e., $J_0^\pi(x_0) > J_0(x_0)$. Then, since $C(x_0; u_0^*, \dots, u_{n-1}^*) = J_0^\pi(x_0) > J_0(x_0)$ holds, $(u_0^*, \dots, u_{n-1}^*)$ is also a spurious local minimizer of the one-shot problem. ■

Remark 1: The continuity of ϵ_i^* ensures (2), which is required in the proof. One situation where $\epsilon_i^*(x)$ is continuous in a neighborhood of x_i^* is that the local minimizers of $Q_i^\pi(x, \cdot)$ do not bifurcate at x_i^* . As shown later in Example 2 of Section II-E, if a local minimizer of $Q_i^\pi(x, \cdot)$ bifurcates at x_i^* , then $\epsilon_i^*(x)$ is discontinuous at x_i^* and the infimum of $\epsilon_i^*(x)$ may be zero. In this case, there is no guarantee that the induced control input of the locally optimal control policy of DP will also be a local minimizer of the one-shot problem. Thus, the assumption of the continuity of $\epsilon_i^*(x)$ is necessary for the results in Theorem 1 to hold true.

Another situation where the results in Theorem 1 hold true is that the Hessian of $Q_i^\pi(x, \cdot)$ is positive definite. We present this idea below.

Theorem 2: 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^*) . If π_k is twice continuously differentiable in a neighborhood of x_k^* and

$$\nabla_u^2 Q_k^\pi(x_k^*, u_k^*) \succ 0, \quad \forall k \in \{0, \dots, n-1\},$$

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

Proof: The proof is given in [1] (see Theorem 3). ■

Remark 2: 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 either Theorem 1 or Theorem 2 as long as the one-shot problem has no spurious local minima.

C. Stationary points: From DP to one-shot optimization

In Section II-B, we have mentioned that if the assumptions of Theorem 1 or Theorem 2 are not satisfied, an induced controlled input of the locally minimum control policy of DP does not necessarily imply a local minimizer of the one-shot problem. Therefore, it is desirable to discover what property such induced input satisfies for the one-shot optimization problem. In this subsection, we will show that under some conditions the control input induced by a locally minimum control policy of DP is also a stationary point of the one-shot problem.

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^* [31].

Accordingly, we can develop the notion of a stationary point in different settings.

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 $-\nabla_u Q_k^\pi(x, \pi_k(x)) \in \mathcal{N}_A(\pi_k(x))$.

For example, a locally minimum control policy is indeed a stationary control policy of DP. Now, we prove that, under mild assumptions, a stationary control policy implies a stationary point of the one-shot optimization. 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 similarly $\mathbf{D}_k^{f,u}(x, u)$ be the Jacobian matrix of $f_k(x, \cdot)$ at point u while viewing x as a constant.

Theorem 3: 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^*) . If for every $k \in \{0, \dots, n-1\}$:

- 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

$$\nabla_x J_k^\pi(x_k^*) = \nabla_x C(x_k^*; u_k^*, \dots, u_{n-1}^*) \quad (5)$$

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,

$$\begin{aligned} \nabla_x J_k^\pi(x_k^*) &= \nabla_x c_k(x_k^*, u_k^*) + \mathbf{D}_k^{f,x}(x_k^*, u_k^*)^T \nabla_x J_{k+1}^\pi(x_{k+1}^*) \\ &\quad + \mathbf{D}_k^\pi(x_k^*)^T \nabla_u Q_k^\pi(x_k^*, u_k^*). \end{aligned} \quad (6)$$

If u_k^* is in the interior of A , since u_k^* is a stationary point of $Q_k^\pi(x_k^*, \cdot)$, we have $\nabla_u Q_k^\pi(x_k^*, u_k^*) = 0$. Otherwise, by the assumption, it holds that $\mathbf{D}_k^\pi(x_k^*) = 0$. In either case, the last term of (6) is zero. On the other hand,

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

Now, (5) can be obtained by taking $x = x_k^*$ in the above equality and then combining it with the induction hypothesis and (6). Finally, for $k \in \{0, \dots, n-1\}$, one can write

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

in which the second equality is due to (5). Since u_k^* is a stationary point of $Q_k^\pi(x_k^*, \cdot)$, $-\nabla_{u_k} Q_k^\pi(x_k^*, u_k^*) \in \mathcal{N}_A(u_k^*)$. 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 subsection, 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 2: Given a function $g : \mathbb{R}^N \times A \rightarrow \mathbb{R}$, 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 $h : B(x^*, \delta) \rightarrow A$ 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 4: 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: Assume that $(u_0^*, \dots, u_{n-1}^*)$ is a strict local minimizer of the one-shot problem. There exists $\epsilon > 0$ such that

$$C(x_0; u_0^*, \dots, u_{n-1}^*) < C(x_0; u_0, \dots, u_{n-1}), \quad (7)$$

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 π_0, \dots, π_{n-1} , positive numbers $\delta_0, \dots, \delta_n$, and corresponding cost-to-go functions J_0^π, \dots, J_n^π 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 semi-continuous 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

$$f_k(x_k, u_k) \in B(x_{k+1}^*, \delta_{k+1}), \quad \forall (x_k, u_k) \in S_k, \quad (8)$$

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

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

by applying (8) 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 (7) 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 2 to the function Q_k^π with x_k^* and ϵ_k , one can find $0 < \delta_k < \delta'_k$ and a function $h_k : B(x_k^*, \delta_k) \rightarrow A$ 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 (8) 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 semi-continuous 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 is lower semi-continuous, while A is compact. Hence, it follows from the Berge maximum theorem [32] that H_k is also lower semi-continuous on \mathbb{R}^N , which implies that J_k^π is lower semi-continuous on $\mathbb{R}^N - \bar{B}(x_k^*, \delta_k)$. For every point \bar{x}_k on the boundary of $B(x_k^*, \delta_k)$, since H_k is lower semi-continuous 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 semi-continuous at \bar{x}_k .

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

Remark 3: DP can be viewed as a reformulation of the optimal control problem from a single one-shot optimization problem to a sequence of optimization problems. When a non-convex problem is reformulated, its local minimizers could change and for example convexification often serves as a reformulation in a higher-dimensional space that eliminates spurious solutions. However, Theorem 4 shows that, under mild conditions, DP is a reformulation of the one-shot optimization problem that preserves local minimizers. Note that a spurious solution of DP is a set of functions, where a spurious solution of the one-shot optimization is a vector.

Remark 4: By taking the contrapositive of Theorem 4, 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.

Considering Theorems 1, 2, and 4 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 be able to effectively demonstrate the results of this section via visualization, we will provide two low-dimensional examples in this section.

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

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

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

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

This one-shot optimization problem has 3 spurious local minimizers $(-0.523, -0.523)$, $(-0.523, 2.477)$, $(0.938, 0.938)$ and the globally optimal minimizer $(0.938, 3.938)$. The landscape of this objective function is shown in Fig. 1a.

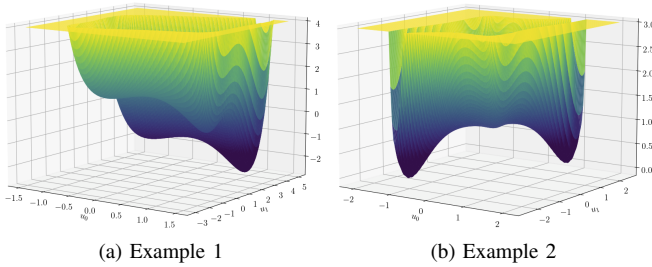


Fig. 1. Landscape of the one-shot optimization: (a) Each (spurious) local minimizer is equivalent to a set of control inputs induced by each (spurious) locally minimum control policy. (b) $(0, 0)$ is a control input induced by a locally minimum control policy but not a local minimizer of the one-shot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

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 the continuous policy

$$\pi_1(x) = \begin{cases} -10, & x < -10, \\ x, & -10 \leq x \leq 10, \\ 10, & x > 10, \end{cases}$$

whose associated cost-to-go function is

$$J_1^\pi(x) = \begin{cases} g(-10), & x < -10, \\ g(x), & -10 \leq x \leq 10, \\ g(10), & x > 10, \end{cases}$$

where $g(x) = -\frac{1}{12}(3x^4 + 16x^3 + 18x^2) + \exp(x^4)$. At the time step $i = 0$ and at the initial state $x_0 = 0$, the Q-function is $Q_1^\pi(0, u_0) = J_1^\pi(u_0)$, which has a spurious local minimizer

at -0.523 and a global minimum at 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.

One can also choose

$$\pi_1(x) = \begin{cases} -10, & x < -13, \\ x + 3, & -13 \leq x \leq 7, \\ 10, & x > 7, \end{cases}$$

whose associated cost-to-go function is

$$J_1^\pi(x) = \begin{cases} g(-13), & x < -13, \\ g(x), & -13 \leq x \leq 7, \\ g(7), & x > 7, \end{cases}$$

where $g(x) = -\frac{1}{12}(3x^4 + 16x^3 + 18x^2 + 27) + \exp(x^4)$. At the time step $i = 0$ and at the initial state $x_0 = 0$, the Q-function is $Q_1^\pi(0, u_0) = J_1^\pi(u_0)$, which has a spurious local minimizer at -0.523 and a global minimum at 0.938 . If we choose $\pi_0(0) = 0.938$, then the locally minimum control policy π is non-spurious 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 the locally minimum control policy 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 4. In addition, it can be noticed that since the minimizer of $Q_1^\pi(x, \cdot)$ does not bifurcate with x , Theorem 1 also 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

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

It has 3 stationary points $(0, 0)$ and $((\log(\frac{8}{3}))^{\frac{1}{4}}, 2(\log(\frac{8}{3}))^{\frac{1}{4}})$ and $(-(\log(\frac{8}{3}))^{\frac{1}{4}}, -2(\log(\frac{8}{3}))^{\frac{1}{4}})$. The later two are the global minimizers of this one-shot problem. To understand why $(0, 0)$ is not a local minimizer of the one-shot problem, 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 locally optimal solution of the one-shot problem. The landscape of this objective function is shown in Fig. 1b. It can also be solved sequentially by DP. For the initial state x_0 , it has 3 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, the Q-function $Q_1^\pi(x, \cdot) = c_1(x, \cdot)$ has 3 stationary points $0, -x, 2x$ and all 3 points will merge to a single point when $x = 0$ and $\nabla_u^2 Q_1(0, 0) = 0$. Therefore,

the assumptions in Theorem 1 and Theorem 2 are violated, and $(0,0)$ is not a local minimizer of the one-shot problem. This clarifies the role of the regularity conditions needed in those theorems. On the other hand, consistent with Theorem 3, $(0,0)$ is a saddle point (which is a stationary point) of the one-shot optimization.

III. DETERMINISTIC PROBLEM WITH 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 the initial state. On the other hand, the dynamic programming 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 [33]. Also, it is reasonable to adopt such parameterized policies for dynamic programming 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, 2, 3, 5, and 6 to incorporate parameterized policies.

Definition 7: Given compact action space A and compact parameter space Θ , let $\mu_\theta(\cdot) : \mathbb{R}^N \rightarrow A$ 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

$$\|\theta - \theta'\| < \delta \Rightarrow \sup_{x \in \mathbb{R}^N} \|\mu_\theta(x) - \mu_{\theta'}(x)\| < \epsilon. \quad (9)$$

Now, we modify the deterministic problems (P1), (P2), and (P3) to a discrete-time finite-horizon deterministic optimal control problem with a parameterized policy as follows:

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

Definition 8: Given a control policy parameter vector $\pi = (\theta_0, \dots, \theta_{n-1})$, 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:

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

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

$$\begin{aligned} \min \quad & J_0^\pi(x_0) \\ \text{s.t.} \quad & \pi = (\theta_0, \dots, \theta_{n-1}) \in \Theta^n, \end{aligned} \quad (\text{PP2})$$

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

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

for $k = 0, \dots, n-1$. Note that π was previously defined as a control policy $(\pi_0, \dots, \pi_{n-1})$, but we use the equivalent definition $(\theta_0, \dots, \theta_{n-1})$ 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 $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$ is said to be a local minimizer of the one-shot optimization if there exists $\epsilon > 0$ such that

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

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

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

$$Q_k^\pi(x, \mu_{\theta_k^*}(x)) \leq Q_k^\pi(x, \mu_{\tilde{\theta}}(x)), \quad \forall \tilde{\theta} \in B(\theta_k^*, \epsilon_k^*) \cap \Theta. \quad (10)$$

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

$$\begin{aligned} \forall k \in \{0, \dots, n-1\}, \forall x \in \mathbb{R}^N, \exists \epsilon_k^*(x) > 0 \text{ such that} \\ Q_k^\pi(x, \mu_{\theta_k^*}(x)) &\leq Q_k^\pi(x, \tilde{\mu}), \quad \forall \tilde{\mu} \in B(\mu_{\theta_k^*}(x), \epsilon_k^*(x)) \cap A, \end{aligned} \quad (11)$$

Definition 10 considers the open ball centered at the policy parameter in the parameter space, while (11) considers the open ball centered at the evaluated action 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 (11) 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 (11), for all $k \in \{0, \dots, n-1\}$ and for all $x \in \mathbb{R}^N$, (10) 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 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 (10). However, it is even possible that $\mu_{\theta_k^*}(x)$ is a strict local maximizer of $Q_k^\pi(x, \cdot)$.

Recall from Remark 1 that (2) was necessary for Theorem 1. Also, 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 $\inf_{x \in \mathbb{R}^N} \epsilon_k^*(x) > 0$ while establishing the relationship from DP to one-shot optimization.

B. From DP to one-shot optimization

In this subsection, 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 5: 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

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

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

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

Theorem 6: 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 6 clearly does not hold since one can generally find a point $x \in \mathbb{R}^N$ such that $\nabla_{\theta_k} Q_k^\pi(x_k^*, \mu_{\theta_k^*}(x_k^*)) \neq \nabla_{\theta_k} Q_k^\pi(x, \mu_{\theta_k^*}(x))$.

C. From one-shot optimization to DP

In this subsection, 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 5 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 by-product, 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: We define our parameterized policy to be a linear combination of arbitrary linearly independent basis functions, while satisfying Definition 7; *i.e.*, Given m functions $f_i : \mathbb{R}^N \rightarrow \mathbb{R}^M$, $i = 1, \dots, m$ and $\theta = [c_1, \dots, c_m]^T \in \Theta$,

$$\mu_\theta(x) = \sum_{i=1}^m c_i f_i(x) \in A \quad (12)$$

where there does not exist $(d_1, \dots, d_m) \neq 0$ such that for all x in any set of non-zero measure, the following equation holds [34]:

$$\sum_{i=1}^m d_i f_i(x) = 0. \quad (13)$$

Remark 8: Since a set of isolated points is a set of measure zero, it is exempt from determining the independence of basis functions. Suppose that x has a continuous distribution. The independence of basis functions implies that if (13) holds for all x in the support of the distribution, $d_1 = \dots = d_m = 0$. On the other hand, suppose that 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 $d_1 = \dots = d_m = 0$ even if (13) 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 details can be found in Appendix A. The usefulness of the parameterized policy also manifests within Representer theorem [35], which shows that a linear combination of kernels fully represents the solution for minimizing empirical risk. It switches the optimization problem in infinite-dimensional function space to finding the finite number of coefficients. This kernel method can also be extended to multi-dimensional vector outputs [36]. 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 remainder 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 5, specifically if the parameterized policy class is a linear combination of basis functions. It turns out that under a mild condition, a local minimizer of the one-shot optimization does not imply a local minimizer of DP in the overparameterized case.

Proposition 2: Consider an overparameterized policy class defined by Definition 13. Let $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$ be a local minimizer of DP. If there exists at least one $k \in \{0, \dots, n-1\}$ 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: The proof is given in Appendix B. ■

Remark 9: Proposition 2 implies that for every $k \in \{0, \dots, n-1\}$, 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 the 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 3: Consider a strict local minimizer of the one-shot optimization $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$. Let (x_0^*, \dots, 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 $(\mu_{\theta_0^*}(x_0^*), \dots, \mu_{\theta_{n-1}^*}(x_{n-1}^*))$.

Proof: The proof is given in Appendix C. ■

Remark 10: If Θ is not assumed to be convex, the conclusion of Lemma 3 does not hold, so the infinite number of control policy parameter vectors can correspond to a single local minimizer of DP as in Proposition 2.

Note that Lemma 3 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 1-step problem

$$\begin{aligned} c_0(x, \mu_\theta(x)) &= \frac{1}{4}\mu_\theta(x)^4 - \frac{1}{3}(x^2 + 2x)\mu_\theta(x)^3 + \\ &\quad \frac{1}{2}(2x^3 + x - 1)\mu_\theta(x)^2 - (x^4 - x^3 + x^2 - x)\mu_\theta(x), \\ c_1(x, \mu_\theta(x)) &= 0, \quad f_0(x, \mu_\theta(x)) = x + \mu_\theta(x). \end{aligned}$$

with the parameterized policy $\mu_\theta(x) = d_1x + d_2$ 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 $(d_1, d_2) \in \Theta$ with the property that $d_1 + d_2 = 0$ or $d_1 + d_2 = 2$ is a local minimizer of the one-shot optimization. Among them, $(1, -1)$ and $(1, 1)$ are strict local minimizers of the one-shot optimization since

$\{(d_1, d_2) : d_1 + d_2 = 0\} \cap \Theta = \{(1, -1)\}$ and $\{(d_1, d_2) : d_1 + d_2 = 2\} \cap \Theta = \{(1, 1)\}$. On the other hand, since $\nabla_{\theta} Q_0^{\pi}(x, \mu_\theta(x)) = \nabla_{\theta} c_0(x, \mu_\theta(x)) = [g(x, \theta)x, g(x, \theta)]^T$ 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.

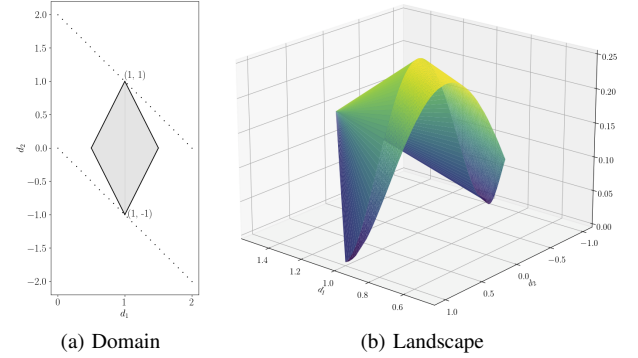


Fig. 2. The domain and the landscape of the one-shot optimization for a deterministic problem: (a) The gray-colored area is the domain of the parameter space. The intersection between the dotted lines and the domain is $\{(1, 1), (1, -1)\}$. (b) Both $(1, 1)$ and $(1, -1)$ are a strict local minimizer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

In light of the above counterexample, one can think of the situation where the parameterized policy contains every locally minimum control policy of DP defined by Definition 3. It turns out that if such a situation is possible, under a convex parameter space, each strict local minimizer of the one-shot optimization is a local minimizer of DP. For this to be true, we need the following assumptions to link between a local minimizer of DP and a locally minimum control policy. Assumption 1 implies that Θ should be large enough to contain relevant parameters to cover the action space A . Also, Assumption 2 can be regarded as the extension of an overparameterized policy.

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

Assumption 2: Given a local minimizer of the one-shot optimization $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$, 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.

Lemma 4: 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 $J_0^\pi(x_0) - J_0^{\pi'}(x_0) = Q_k^\pi(x_k, \mu_{\theta_k^*}(x_k)) - Q_k^\pi(x_k, \mu_{\theta_k'}(x_k))$, 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

$$Q_k^\pi(x_k^*, \mu_{\theta_k^*}(x_k^*)) \leq Q_k^\pi(x_k^*, \mu_{\tilde{\theta}}(x_k^*)), \quad \forall \tilde{\theta} \in B(\theta_k^*, \epsilon_k^*) \cap \Theta. \quad (14)$$

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 1, 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 $\theta_v \in B(\theta_k^*, \epsilon_k^*)$: 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 $0 < \delta < \delta_v$ by the linearity of policy and the convexity of Θ .

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 $0 < \delta < \frac{\epsilon_k^*}{2\|\theta_v - \theta_k^*\|} \delta_v$. Notice that $\|\frac{\epsilon_k^*}{2\|\theta_v - \theta_k^*\|}(\mu_{\theta_v}(x_k^*) - \mu_{\theta_k^*}(x_k^*))\| = \frac{\epsilon_k^*}{2} \cdot \frac{\|F_k^*(\theta_v - \theta_k^*)\|}{\|\theta_v - \theta_k^*\|} \geq \frac{\epsilon_k^*}{2} \sigma_k^* > 0$, where the last inequality is from Assumption 2 and the second last inequality is from the basic property of singular value [37].

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 (14) implies

$$Q_k^\pi(x_k^*, \mu_{\theta_k^*}(x_k^*)) \leq Q_k^\pi(x_k^*, \tilde{u}), \quad \forall \tilde{u} \in B(\mu_{\theta_k^*}(x_k^*), \frac{\epsilon_k^*}{2} \sigma_k^*) \cap A. \quad (15)$$

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 (15). 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 3, $\mu_{\theta_k^*} = \tilde{\pi}_k$ 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 11: Since $\mu_\theta(x)$ is designed to be in A by Definition 7, Assumption 1 is equivalent to saying that $A = \mu_\Theta(x_0^*) = \dots = \mu_\Theta(x_n^*)$. On the other hand, if Θ is not large enough, $\tilde{\pi}$ may not be a locally minimum control policy in that a neighborhood of $\theta_k^* \in \Theta$ may not cover a neighborhood of $\mu_{\theta_k^*}(x_k^*) \in A$. One can think of an extreme situation that Θ is a line but the dimension of A is greater than 1.

Meanwhile, suppose that there exist two different locally minimum control policies in a set of non-zero measure, meaning that at some step k , $\pi_1(x) \neq \pi_2(x)$ for all $x \in I$ where I is a set of non-zero 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 7: 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 there exists only a single locally minimum control policy of DP $\phi = (\phi_0, \dots, \phi_{n-1})$ and the parameterized policy class contains ϕ , then π is a local minimizer of DP.

Proof: Let $\phi' = (\theta'_0, \dots, \theta'_{n-1})$ be the parameters associated with ϕ . For all $k \in \{0, \dots, n-1\}$ 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 4, this completes the proof. ■

Considering both Theorem 5 and 7, one can conclude that under the assumptions of Theorem 7, a local minimizer of DP is *equivalent* to a local minimizer of the one-shot optimization. In other words, if DP has a very low complexity, so does the one-shot problem.

IV. STOCHASTIC PROBLEM WITH A PARAMETERIZED POLICY

A. Problem Formulation

In this Section, we show that the results obtained for the deterministic problem with a parameterized policy also hold for the stochastic problem with a parameterized policy. Since we now take the expectation of the sum of the stage costs and the terminal cost over the trajectories, the issue of strictness, as in Proposition 2, does not take place. Before presenting the related 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 $k \in \{0, \dots, n-1\}$ such that x_0, w_0, \dots, w_{n-1} are mutually independent. The state transition is now governed by the dynamics $f_i : \mathbb{R}^N \times A \times \mathbb{R}^W \rightarrow \mathbb{R}^N$, $i = 0, \dots, n-1$. The dynamics are again defined to be at least twice continuously differentiable.

Now, we modify the deterministic problems with a parameterized policy, *i.e.*, (PP1), (PP2), and (PP3), to a discrete-time finite-horizon stochastic optimal control problem with a parameterized policy:

$$\begin{aligned} \min_{\theta_0, \dots, \theta_{n-1}} \quad & \mathbb{E}_{x_0, w_0, \dots, w_{n-1}} \left[\sum_{i=0}^{n-1} c_i(x_i, \mu_{\theta_i}(x_i)) + c_n(x_n) \right] \\ \text{s.t.} \quad & \theta_i \in \Theta, \quad i = 0, \dots, n-1, \end{aligned} \quad (\text{SP1})$$

where for all $i \in \{0, \dots, n-1\}$, the state x_{i+1} is determined by $f_i(x_i, \mu_{\theta_i}(x_i), w_i)$. Notice that for stochastic problems, x_0 is not given as a point, but has an initial distribution ρ . Afterwards, x_{i+1} is a random variable induced by (x_0, w_0, \dots, w_i) .

Definition 15: Given a control policy parameter vector $\pi = (\theta_0, \dots, \theta_{n-1})$, 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:

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

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

$$\begin{aligned} \min \quad & \mathbb{E}_{x_0} [J_0^\pi(x_0)] \\ \text{s.t.} \quad & \pi = (\theta_0, \dots, \theta_{n-1}) \in \Theta^n, \end{aligned} \quad (\text{SP2})$$

as long as the cost functions $c_i, i = 0, \dots, n-1$, are uniformly bounded, due to the product measure Theorem and Fubini's Theorem [38]. In the remainder of the paper, we assume that the two problems are equivalent.

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

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

Definition 16 (local minimizer of the one-shot optimization): A control policy parameter vector $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$ 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 $\tilde{\pi} = (\tilde{\theta}_0, \dots, \tilde{\theta}_{n-1}) \in (B(\theta_0^*, \epsilon) \cap \Theta) \times \dots \times (B(\theta_{n-1}^*, \epsilon) \cap \Theta)$.

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

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 subsection, 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 the state for each step should be a continuous random variable for the stationary points since the expectation is over all trajectories rather than a single trajectory.

Theorem 8: 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

$$\begin{aligned} \mathbb{E}_{x_0} [J_0^\pi(x_0)] &= \mathbb{E}_{x_0} [Q_0^\pi(x_0, \mu_{\theta_0^*}(x_0))] \leq \mathbb{E}_{x_0} [Q_0^\pi(x_0, \mu_{\tilde{\theta}_0}(x_0))] \\ &= \mathbb{E}_{x_0} [c_0(x_0, \mu_{\tilde{\theta}_0}(x_0)) + \mathbb{E}_{w_0} [Q_1^\pi(\tilde{x}_1, \mu_{\tilde{\theta}_1}(\tilde{x}_1))] \\ &\quad (\tilde{x}_1 = f_0(x_0, \mu_{\tilde{\theta}_0}(x_0), w_0))] \\ &\leq \mathbb{E}_{x_0} [c_0(x_0, \mu_{\tilde{\theta}_0}(x_0)) + \mathbb{E}_{w_0} [Q_1^\pi(\tilde{x}_1, \mu_{\tilde{\theta}_1}(\tilde{x}_1))] \\ &= \mathbb{E}_{x_0} [c_0(x_0, \mu_{\tilde{\theta}_0}(x_0)) + \mathbb{E}_{w_0} [c_1(\tilde{x}_1, \mu_{\tilde{\theta}_1}(\tilde{x}_1)) \\ &\quad + \mathbb{E}_{w_1} [Q_2^\pi(\tilde{x}_2, \mu_{\tilde{\theta}_2}(\tilde{x}_2))] \\ &\quad (\tilde{x}_2 = f_1(\tilde{x}_1, \mu_{\tilde{\theta}_1}(\tilde{x}_1), w_1))] \\ &\leq \dots \leq \mathbb{E}_{x_0, w_0, \dots, w_{n-1}} [J_0^{\tilde{\pi}}(x_0)] \end{aligned}$$

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

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

Now, let $\mathbf{D}_k^\mu(\theta)$ 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 9: Consider a stationary point of DP $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$. If for all $k \in \{0, \dots, n-1\}$,

- 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 $k \in \{1, \dots, n\}$, $J_k^\pi(x)$ is continuously differentiable. For the base step, $J_n^\pi(x) = c_n(x)$ is continuously differentiable. For the induction step, observe that

$$\begin{aligned} \nabla_x J_k^\pi(x) &= \nabla_x [Q_k^\pi(x, \mu_{\theta_k^*}(x))] \\ &= \nabla_x [c_k(x, \mu_{\theta_k^*}(x)) + \int_{\Omega} J_{k+1}^\pi(f_k(x, \mu_{\theta_k^*}(x), w_k)) dp(w_k)] \end{aligned}$$

$$\begin{aligned}
&= \nabla_x [c_k(x, \mu_{\theta_k^*}(x))] + \int_{\Omega} \nabla_x [J_{k+1}^{\pi}(f_k(x, \mu_{\theta_k^*}(x), w_k))] dp(w_k) \\
&= \nabla_x [c_k(x, \mu_{\theta_k^*}(x))] \\
&+ \int_{\Omega} \mathbf{D}_k^{f,x}(x, \mu_{\theta_k^*}(x), w_k)^T \nabla_x J_{k+1}^{\pi}(f_k(x, \mu_{\theta_k^*}(x), w_k)) dp(w_k).
\end{aligned}$$

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 $k \in \{0, \dots, n-1\}$, observe that

$$\begin{aligned}
&\nabla_{\theta_k} Q_k^{\pi}(x_k, \mu_{\theta_k^*}(x_k)) \\
&= \nabla_{\theta_k} [c(x_k, \mu_{\theta_k^*}(x_k)) + \int_{\Omega} J_{k+1}^{\pi}(f_k(x_k, \mu_{\theta_k^*}(x_k), w_k)) dp(w_k)] \\
&= \mathbf{D}_{x_k}^{\mu}(\theta_k^*)^T \nabla_{\mu} c(x_k, \mu_{\theta_k^*}(x_k)) \\
&\quad + \int_{\Omega} \mathbf{D}_k^{f,\theta}(x_k, \mu_{\theta_k^*}(x_k), w_k)^T \nabla_x J_{k+1}^{\pi}(f_k(x_k, \mu_{\theta_k^*}(x_k), w_k)) dp(w_k),
\end{aligned}$$

which is valid because for $k \in \{1, \dots, n\}$, $J_k^{\pi}(x)$ 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 $k \in \{0, \dots, n-1\}$,

$$\begin{aligned}
&\nabla_{\theta_k} \mathbb{E}_{x_0} [J_0^{\pi}(x_0)] = \\
&\int_{\mathbb{R}^N} \int_{\Omega} \dots \int_{\Omega} \nabla_{\theta_k} Q_k^{\pi}(x_k, \mu_{\theta_k^*}(x_k)) dp(w_{k-1}) \dots dp(w_0) dp(x_0),
\end{aligned}$$

Now, note that $\mathcal{N}_{\Theta}(\theta_k^*)$ is nonempty, closed, and convex [31]. 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

$$\begin{aligned}
&-p^T \nabla_{\theta_k} \mathbb{E}_{x_0} [J_0^{\pi}(x_0)] \\
&= -p^T \int_{\mathbb{R}^N} \int_{\Omega} \dots \int_{\Omega} \nabla_{\theta_k} Q_k^{\pi}(x_k, \mu_{\theta_k^*}(x_k)) dp(w_{k-1}) \dots dp(x_0) \\
&= \int_{\mathbb{R}^N} \int_{\Omega} \dots \int_{\Omega} -p^T \nabla_{\theta_k} Q_k^{\pi}(x_k, \mu_{\theta_k^*}(x_k)) dp(w_{k-1}) \dots dp(x_0) \\
&< \int_{\mathbb{R}^N} \int_{\Omega} \dots \int_{\Omega} -p^T \nabla_{\theta_k} \mathbb{E}_{x_0} [J_0^{\pi}(x_0)] dp(w_{k-1}) \dots dp(x_0) \\
&= -p^T \nabla_{\theta_k} \mathbb{E}_{x_0} [J_0^{\pi}(x_0)],
\end{aligned}$$

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 $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$ is a stationary point of the one-shot optimization. ■

Remark 12: Notice that the parameterized policy class defined by Definition 13 satisfies the condition of Theorem 9 as long as all basis functions are continuously differentiable. Therefore, one can apply the result to a linear combination of arbitrary continuously differentiable functions.

C. From one-shot optimization to DP

In this subsection, 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 8 and that of Theorem 9 do not hold. Then, it is clear that 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, among the policy class defined by Definition 13, we use the basic parameterized policy $\mu_{\theta_k}(x) = a_k x + b_k$, where $\theta_k = (a_k, b_k)$. Consider the 2-step problem

$$\begin{aligned}
&x_0 = 0, \quad c_0(x, \mu_{\theta_0}(x)) = 0, \\
&f_0(x, \mu_{\theta_0}(x), w_0) = x + a_0 x + b_0 + w_0, \\
&c_1(x, \mu_{\theta_1}(x)) = \frac{1}{4}(a_1 x + b_1)^4 - \frac{1}{2}(a_1 x + b_1)^2 + x^2, \\
&f_1(x, \mu_{\theta_1}(x), w_1) = x + a_1 x + b_1 + w_1, \\
&c_2(x, \mu_{\theta_2}(x)) = 0 \text{ where } w_0, w_1 \stackrel{\text{iid}}{\sim} \text{Uniform}(-\sqrt{\frac{5}{3}}, \sqrt{\frac{5}{3}}),
\end{aligned}$$

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

$$\begin{aligned}
&\min_{-2 \leq b_0, a_1, b_1 \leq 2} \mathbb{E}_{w_0} \left[\frac{1}{4} \{a_1(b_0 + w_0) + b_1\}^4 \right. \\
&\quad \left. - \frac{1}{2} \{a_1(b_0 + w_0) + b_1\}^2 + (b_0 + w_0)^2 \right]
\end{aligned}$$

It turns out that there are 9 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 4 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 3 stationary points of DP: $(0, 0, \pm 1), (0, 0, 0)$ and 2 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.

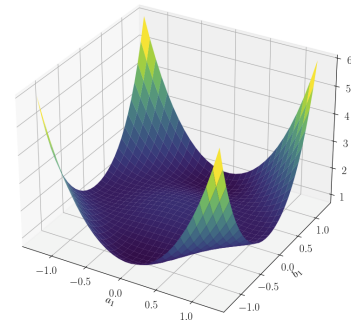


Fig. 3. Landscape of the one-shot optimization for a stochastic problem: b_0 is fixed to 0 in the figure. $(a_1, b_1) = (\pm 1, 0), (0, \pm 1)$ are strict local minimizers of the one-shot optimization but only $(0, \pm 1)$ is a local minimizer of DP.

Now, we present the specific case that a local minimizer of the one-shot optimization implies a local minimizer of DP, similar to Theorem 7. 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 $\phi = (\phi_0, \dots, \phi_{n-1})$ which is also a locally minimum control policy in the interior of A for all $x \in \mathbb{R}^N$. The associated parameters of ϕ are denoted by $\phi' = (\theta'_0, \dots, \theta'_{n-1})$.

Assumption 4: Θ is large enough to contain every parameter $\theta' \in \mathbb{R}^m$ such that $\nabla_{\mu} Q_k^{\phi'}(x, \mu_{\theta'}(x))$ is 0 for all $x \in \mathbb{R}^N$ for any $k \in \{0, \dots, n-1\}$, where ϕ' is defined in Assumption 3.

Theorem 10: Assume that Assumptions 3 and 4 hold. Consider a local minimizer of the one-shot optimization $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$ in the interior of Θ^n . Suppose that x_k^* is a continuous random variable for all $k \in \{0, \dots, n-1\}$, where (x_0^*, \dots, x_n^*) is the random state process associated with π . If the parameterized policy defined by Definition 13 contains a single stationary control policy ϕ , then π is a local minimizer of DP.

Proof: Since ϕ is a single locally minimum control policy, $Q_k^{\phi'}(x, u)$ has no spurious local minima for all $k \in \{0, \dots, n-1\}$. 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 $\pi = (\theta_0^*, \dots, \theta_{n-1}^*)$ 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 $k \in \{0, \dots, n-1\}$.

For the base step, at step $n-1$, since the parameterized policy contains ϕ_{n-1} , it can be expressed as $\phi_{n-1}(x) = \sum_{i=1}^m c_i f_i(x)$, where $f_i : \mathbb{R}^N \rightarrow \mathbb{R}^M$, $i = 1, \dots, m$, $f_i(x) = [f_{i1}(x), \dots, f_{iM}(x)]^T$ and $\theta'_{n-1} = (c_1, \dots, c_m) \in \Theta$. Notice that $Q_{n-1}^{\phi'}(x, \mu_{\theta'_{n-1}}(x)) = Q_{n-1}^{\pi}(x, \mu_{\theta_{n-1}^*}(x))$ since θ_{n-1} is the final parameter of the whole system to determine the control inputs and the state transition. Now, observe that

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

since $\mu_{\theta'_{n-1}}(x)$ is a stationary point of $Q_{n-1}^{\phi'}(x, \cdot)$ in the interior of A . Also, since Θ is large enough, θ'_{n-1} should be a unique solution for $\nabla_{\mu} Q_{n-1}^{\phi'}(x, \mu_{\theta'}(x)) = 0$, which yields the following expression with $\mu_{\theta}(x) = (u_1, \dots, u_M)^T$:

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

where $g_j(x)$, $j = 1, \dots, M$, are nonnegative at $u_j = \sum_{i=1}^m c_i f_{ij}(x)$ and positive at all the other points since ϕ' is the local minimizer of DP.

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

$$\begin{aligned} &\nabla_{\theta_{n-1}} Q_{n-1}^{\pi}(x, \mu_{\theta_{n-1}^*}(x))^T \\ &= \nabla_{\mu} Q_{n-1}^{\pi}(x, \mu_{\theta_{n-1}^*}(x))^T \mathbf{D}_{\theta}^{\mu}(\theta_{n-1}^*) \\ &= \begin{bmatrix} (\sum_{i=1}^m (d_i - c_i) f_{i1}(x)) \cdot g_1(x) \\ \vdots \\ (\sum_{i=1}^m (d_i - c_i) f_{iM}(x)) \cdot g_M(x) \end{bmatrix}^T \begin{bmatrix} f_{11}(x) \cdots f_{m1}(x) \\ \vdots & \ddots & \vdots \\ f_{1M}(x) \cdots f_{mM}(x) \end{bmatrix} \end{aligned} \quad (16)$$

Now, notice that $\nabla_{\theta_{n-1}} \mathbb{E}_{x_0} [J_0^{\pi}(x_0)] = 0$ since π is a stationary point of the one-shot optimization in the interior of Θ^n . Then, observe that

$$\begin{aligned} &\nabla_{\theta_{n-1}} \mathbb{E}_{x_0} [J_0^{\pi}(x_0)] \\ &= \nabla_{\theta_{n-1}} \mathbb{E}_{x_0, w_0, \dots, w_{n-2}} [Q_{n-1}^{\pi}(x_{n-1}^*, \mu_{\theta_{n-1}^*}(x_{n-1}^*))] \quad (17) \\ &= \mathbb{E}_{x_0, w_0, \dots, w_{n-2}} [\nabla_{\theta_{n-1}} Q_{n-1}^{\pi}(x_{n-1}^*, \mu_{\theta_{n-1}^*}(x_{n-1}^*))] = 0. \end{aligned}$$

The second equality comes from $Q_{n-1}^{\pi}(x, \mu_{\theta}(x))$ being differentiable with respect to the parameters due to the linearity of the policy defined by Definition 13. Now, we substitute (16) into (17) to derive an m -dimensional vector equation, and multiply $(d_i - c_i)$ with i^{th} component as follows:

$$\begin{aligned} &\mathbb{E}_{x_0, w_0, \dots, w_{n-2}} \left[\sum_{j=1}^M (d_k - c_k) f_{kj}(x_{n-1}^*) \cdot \right. \\ &\quad \left. \sum_{i=1}^m (d_i - c_i) f_{ij}(x_{n-1}^*) \cdot g_j(x_{n-1}^*) \right] = 0, \end{aligned}$$

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

$$\sum_{j=1}^M \mathbb{E}_{x_0, w_0, \dots, w_{n-2}} \left[\left(\sum_{i=1}^m (d_i - c_i) f_{ij}(x_{n-1}^*) \right)^2 \cdot g_j(x_{n-1}^*) \right] = 0. \quad (18)$$

The term inside the expectation is always nonnegative regardless of the distribution of x_0, w_0, \dots, w_{n-2} . Now, suppose that $d_i \neq c_i$ for some $i \in \{1, \dots, m\}$. For (18) to be satisfied, $\sum_{i=1}^m (d_i - c_i) f_{ij}(x_{n-1}^*)$ should be 0 for every $j \in \{1, \dots, M\}$ for all possible values of x_{n-1} . Recall from Remark 8 to note that it is impossible to satisfy (18) 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. Here, notice that $g_j(x_{n-1}^*)$ is nonnegative only at a single point, and positive elsewhere. Thus, $d_i = c_i$ for all $i \in \{1, \dots, m\}$, which means $\theta_{n-1}^* = \theta'_{n-1}$.

For the induction step, assume that $\theta_k^* = \theta'_k$. Again, $Q_k^{\phi'}(x, \mu_{\theta_k^*}(x)) = Q_k^{\pi}(x, \mu_{\theta_k^*}(x))$ 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 13: Notice that the proof leverages the fact that π is a stationary point of the one-shot optimization. Thus, Theorem 10 can be stated for the stationary points of the one-shot optimization instead of the local minimizers of the one-shot optimization.

Meanwhile, the results of both Theorems 7 and 10 state that a local minimizer of the one-shot optimization is indeed a local minimizer of DP. By taking the contrapositive, one can notice that under the condition that DP has a very low complexity, there is at most one local minimizer of the one-shot optimization since no spurious local minima exist at each single period of DP. This directly implies that there are also no spurious local minima in the entire period one-shot optimization; *i.e.*, if DP has a very low complexity, so does the one-shot problem.

Remark 14: Regarding Assumption 3, it may be difficult to satisfy the precondition that a single stationary control policy should be in the interior of A for all $x \in \mathbb{R}^N$. Thus, instead, we can relax this condition to only the domain of x ; *i.e.*, the set of values that at least one of the states x_0, x_1, \dots, x_n can take. For example, if the state space is finite, satisfying the condition becomes relatively straightforward.

On the other hand, 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 equation (18), which incurs the fact that $\theta_k^* = \theta'_k$ regardless of the distribution of x_0, w_0, \dots, 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, \dots, 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 sub-optimal stationary point with respect to any possible distribution.

Remark 15: To determine the form of $\nabla_{\mu} Q_{n-1}^{\phi'}(x, u)$, it was necessary to argue that $\mu_{\theta'_{n-1}}(x)$ should be a unique solution for $\nabla_{\mu} Q_{n-1}^{\phi'}(x, u) = 0$. For this to be true, there should certainly be only a single stationary control policy. Other than that, this is possible only when Θ is large enough, which necessitates Assumption 4. Interestingly, Assumption 1 also says that the concept of Θ being large enough to cover the action space is crucial even in the deterministic setting to be able to establish a relationship from one-shot to DP. The detailed remarks can be found in Appendix D.

Now, we present the pictorial example of Theorem 10. We use a parameterized policy $\mu_{\theta_k}(x) = a_k x + b_k$ for the k^{th} step, where $\theta_k = (a_k, b_k)$. Consider the 2-step problem

$$\begin{aligned} x_0 &= 0, \quad c_0(x, \mu_{\theta_0}(x)) = 0, \\ f_0(x, \mu_{\theta_0}(x), w_0) &= x + a_0 x + b_0 + w_0, \\ c_1(x, \mu_{\theta_1}(x)) &= \frac{1}{4}(a_1 x + b_1 - x - 0.5)^4 + x^4, \\ f_1(x, \mu_{\theta_1}(x), w_1) &= x + a_1 x + b_1 + w_1, \\ c_2(x, \mu_{\theta_2}(x)) &= 0 \text{ where } w_0, w_1 \stackrel{\text{iid}}{\sim} \text{Uniform}\left(-\sqrt{\frac{5}{3}}, \sqrt{\frac{5}{3}}\right), \end{aligned}$$

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

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

Let the action space be $[-20, 20]$. The action space is large enough to have the locally minimum control policy to be

in the interior of A , which uniquely exists as $(\pi_0, \pi_1) = (0, x+0.5)$. 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.

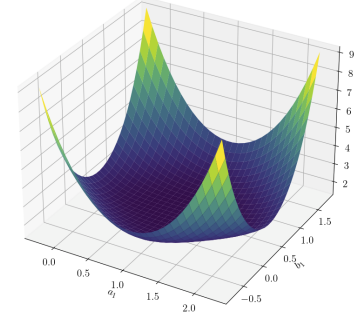


Fig. 4. Landscape of the one-shot optimization under the assumptions of Theorem 10: b_0 is fixed to 0 in the figure. $(a_1, b_1) = (1, 0.5)$ is the only stationary point (local minimizer) of DP and also the only stationary point (local minimizer) of the one-shot optimization.

Considering both Theorem 8 and 10, one can conclude that under the assumptions of 10, a local minimizer of DP is *equivalent* to a local minimizer of the one-shot optimization. In other words, if DP has a very low complexity, the same holds for the one-shot problem.

V. CONCLUSION

In this paper, we studied the (spurious) local solutions of arbitrary optimal control problems through two different formulations: one-shot (single) optimization problem aimed at solving for all input values at the same time, and DP method aimed at finding the input values sequentially. We introduced the notions of spurious (non-global) local minimizers for the one-shot problem and spurious locally minimum control policies for DP. 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. We also proved that the control sequence induced by a stationary control policy of DP corresponds to some stationary point of the one-shot problem.

To help better understand the quality of the local solutions obtained by reinforcement learning algorithms, we incorporated exact parameterized policies into the optimal control problem for both deterministic and stochastic dynamics. We showed that each local minimizer (stationary point) of DP corresponds to a local minimizer (stationary point) of the one-shot optimization, but not the other way around. These discoveries show that the optimization landscape of the one-shot optimization is more complex than its DP counterpart, and thus the one-shot problem having a low complexity implies a low complexity for DP. However, 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 a local minimizer of DP. This implies that DP

having a very low complexity implies a very low complexity for the one-shot problem, and furthermore, a local minimizer of both methods is equivalent.

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 infinite-horizon stationary optimal control problem dealing with Markov-Decision Processes (MDPs) in a typical RL problem, where the policy iteration (DP) is performed at each improvement step. On the other hand, a limitation of this work is our restriction of the policy class. A natural extension would be to extend the parameterized policy class beyond a linear combination of basis functions, such as composite functions widely used in deep neural networks.

APPENDIX

A. Applications of a parameterized policy defined by Definition 13

One example of the parameterized policy satisfying Definition 13 is a piecewise polynomial function. Piecewise polynomial function, also called as splines, is one of the "sufficiently rich" policy to approximate an arbitrary univariate function f . Given the endpoints $a_0, \dots, a_{k+1} \in \mathbb{R}$, also called as knots, and the maximum degree of the polynomial l , one can choose a basis function as

$$f_{ij}(x) = \begin{cases} x^j, & \text{if } a_i \leq x < a_{i+1}, \\ 0, & \text{otherwise,} \end{cases}$$

for $i = 0, \dots, k, j = 0, \dots, l$. The linear combination of these basis functions $\hat{f}(x) = \sum_{i,j} c_{ij} f_{ij}(x)$ is defined in $[a_0, a_{k+1})$ and indeed a piecewise polynomial function parameterized by $\theta = \{c_{ij}\} \in \mathbb{R}^{(k+1) \times (l+1)}$. Such splines or the variations of splines are widely used for curve fitting due to their large capacity to approximate complicated functions [39].

Another application arises when the state and action spaces are finite in Markov-Decision Processes (MDPs). In a direct policy parametrization, the parameters are state-action pair probabilities, where the parameter space is defined as a probability simplex. This setting was covered in the aforementioned works to study the rate of global convergence of policy gradient variants [23]–[27]. For that application, the function whose value is 1 at a single state-action pair and 0 at all the other pairs exemplifies one of the independent basis functions. This indicates that the policy class not only includes deterministic control but also handles stochastic control if the basis functions are properly defined.

B. Proof of Proposition 2

Proof: Consider the state sequence (x_0^*, \dots, x_n^*) induced by a local minimizer of DP π . Let k' be the index for which θ_k^* is in the interior of Θ . Then, $\mu_{\theta_{k'}^*}(x_{k'}^*)$ is the action taken at the step k' , which can be expressed as $\sum_{i=1}^m c_i f_i(x_{k'}^*)$ with $\theta_{k'}^* = (c_0^*, c_1^*, \dots, c_m^*)$ by Definition 13. Since the policy is overparameterized, m is greater than the dimension of the

action M . Then, consider the following matrix equation:

$$\begin{bmatrix} f_0(x_{k'}^*) & f_1(x_{k'}^*) & \dots & f_m(x_{k'}^*) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_m \end{bmatrix} = \mu_{\theta_{k'}^*}(x_{k'}^*) \quad (19)$$

and let F_k^* denote the first matrix in the left hand side. F_k^* is an $M \times m$ constant matrix and $\mu_{\theta_{k'}^*}(x_{k'}^*)$ is an $M \times 1$ constant vector given by $x_{k'}^*$. Since $\theta_{k'}^* = (c_0^*, c_1^*, \dots, c_m^*)$ satisfies (19), it has at least one solution.

On the other hand, since $\theta_{k'}^*$ is in the interior of Θ , one can pick $\epsilon_1 > 0$ such that $B(\theta_{k'}^*, \epsilon_1) \in \Theta$. Since $m > M$, the dimension of the null space of F_k^* is greater than 0. Take any nonzero element v from the null space. Then, for all $\delta \in \mathbb{R}$, $\theta_{k'}^* + \delta v$ satisfies (19). Thus, for $0 \leq \delta \leq \frac{\epsilon_1}{\|v\|}$, $\theta_{k'}^* + \delta v \in B(\theta_{k'}^*, \epsilon_1)$ and satisfies (19), which implies that there exists an infinite number of parameter vectors that preserve the state and action sequence in the neighborhood of $\theta_{k'}^*$. Indeed, the induced cost is also preserved.

By Theorem 5, 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 $\tilde{\pi} = (\theta_0^*, \dots, \tilde{\theta}_k, \dots, \theta_{n-1}^*)$ where $\tilde{\theta}_k \in B(\theta_{k'}^*, \epsilon_2) \cap \Theta$. Let $\epsilon := \min\{\epsilon_1, \epsilon_2\}$.

For $0 \leq \delta \leq \frac{\epsilon}{2\|v\|}$, $B(\theta_{k'}^* + \delta v, \delta\|v\|) \subset B(\theta_{k'}^*, \epsilon)$. Also, recall $\theta_{k'}^* + \delta v$ preserves the induced cost, thus $J_0^{\pi'}(x_0) = J_0^\pi(x_0) \leq J_0^{\tilde{\pi}'}(x_0)$ for all $\tilde{\pi}' = (\theta_0^*, \dots, \tilde{\theta}', \dots, \theta_{n-1}^*)$ where $\pi' = (\theta_0^*, \dots, \theta_{k'}^* + \delta v, \dots, \theta_{n-1}^*)$ and $\tilde{\theta}' \in B(\theta_{k'}^* + \delta v, \delta\|v\|) \cap \Theta$. Thus, for $0 \leq \delta \leq \frac{\epsilon}{2\|v\|}$, $\theta_{k'}^* + \delta v$ is a local minimizer of the one-shot optimization, which completes the proof. ■

C. Proof of Lemma 3

Proof: For every $k \in \{0, \dots, n-1\}$, $\mu_\theta(x_k^*) = \sum_{i=1}^{m_k} c_i f_i(x_k^*)$ where $\theta = (c_1, \dots, c_{m_k})$. Let $\theta_k^* = (c_1^*, \dots, c_{m_k}^*)$. 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

$$\{\theta \in B(\theta_k^*, \epsilon) \cap \Theta : \sum_{i=1}^{m_k} c_i f_i(x_k^*) = \mu_{\theta_k^*}(x_k^*)\} = \{\theta_k^*\}. \quad (20)$$

Assume that there exists $\tilde{\theta} \neq \theta_k^*$ such that $\sum_{i=1}^{m_k} \tilde{c}_i f_i(x_k^*) = \mu_{\theta_k^*}(x_k^*)$ where $\tilde{\theta} = (\tilde{c}_1, \dots, \tilde{c}_{m_k})$. Then, for $\lambda \in [0, 1]$, $\sum_{i=1}^{m_k} (\lambda c_i^* + (1-\lambda)\tilde{c}_i) f_i(x_k^*) = \mu_{\theta_k^*}(x_k^*)$ by linearity and $\lambda\theta_k^* + (1-\lambda)\tilde{\theta} \in \Theta$ by convexity. Letting $\lambda \rightarrow 1$, one can construct $\theta \neq \theta_k^*$ that is an element of the left-hand side of (20). By contradiction, θ_k^* is a unique point that achieves $\mu_{\theta_k^*}(x_k^*)$. ■

D. Remarks on Assumption 4

In the stochastic case, if Θ does not sufficiently cover the action space A , there can exist a stationary point of the one-shot optimization that is not a stationary point of DP. For example, recall Fig. 3, which shows that a stationary point of the one-shot optimization is not necessarily a stationary point of DP. One can truncate the parameter space to $\Theta = [-0.5, 2] \times [0.1, 2]$ to have only a single stationary control

policy $\mu_{\theta_0}(x_0) = 0, \mu_{\theta_1}(x_1) = 1$, which is a locally minimum control policy in the interior of large enough action space since the policy is constant.

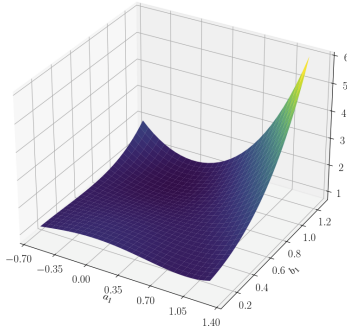


Fig. 5. Truncated landscape of the one-shot optimization for a stochastic problem: b_0 is fixed to 0 in the figure. $(a_1, b_1) = (0, 1)$, $(0.7071, 0.4082)$ are stationary points of the one-shot optimization but only $(0, 1)$ is a stationary point of DP. This demonstrates the necessity of Assumption 4 for Theorem 10.

However, in the truncated parameter space, there exist two stationary points of the one-shot optimization, $(0, 0, 1)$ and $(0, 0.7071, 0.4082)$. The former is a stationary point and a local minimizer of DP, but the latter is not a stationary point of DP. Thus, this example necessitates large enough parameter space for Theorem 10. Fig. 5 shows the truncated landscape of the one-shot optimization when b_0 is fixed to 0.

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 *American Control Conference (ACC)*. IEEE, 2021.
- [2] R. E. Bellman, *Dynamic programming*. Princeton university press, 1957.
- [3] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 1995, vol. 1.
- [4] R. Bellman and R. Kalaba, "On the role of dynamic programming in statistical communication theory," *IRE Transactions on Information Theory*, vol. 3, no. 3, pp. 197–203, 1957.
- [5] E. A. Feinberg, "Optimality conditions for inventory control," in *Optimization Challenges in Complex, Networked and Risky Systems*. INFORMS, 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 *American Control Conference (ACC)*, vol. 2. IEEE, 2002, pp. 1425–1430.
- [7] D. P. Bertsekas, *Reinforcement learning and optimal control*. Athena Scientific Belmont, MA, 2019.
- [8] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [9] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski 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 Transactions on Information Theory*, vol. 65, no. 2, pp. 742–769, 2019.
- [11] C. Liu, L. Zhu, and M. Belkin, "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks," *Applied and Computational Harmonic Analysis*, 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 *International Conference on Machine Learning*. PMLR, 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 Journal on Optimization*, 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," *The Annals of Statistics*, vol. 46, no. 6A, pp. 2747–2774, 2018.
- [15] K. A. Chandrasekher, A. Pananjady, and C. Thrampoulidis, "Sharp global convergence guarantees for iterative nonconvex optimization with random data," *The Annals of Statistics*, 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," *Mathematical Programming*, vol. 193, pp. 75–111, 2022.
- [17] R. Y. Zhang, "Sharp global guarantees for nonconvex low-rank matrix recovery in the overparameterized regime," *arXiv preprint arXiv:2104.10790*, 2021.
- [18] —, "Improved global guarantees for the nonconvex burer-monteiro factorization via rank overparameterization," *arXiv preprint arXiv:2207.01789*, 2022.
- [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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 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," *Mathematical Programming*, 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 Transactions on Automatic Control*, vol. 68, no. 1, pp. 156–171, 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 Transactions on Automatic Control*, vol. 68, no. 1, pp. 80–95, 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," *Journal of Machine Learning Research*, vol. 22, no. 98, pp. 1–76, 2021.
- [24] J. Bhandari and D. Russo, "On the linear convergence of policy gradient methods for finite mdp," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2386–2394.
- [25] L. Xiao, "On the convergence rates of policy gradient methods," *Journal of Machine Learning Research*, vol. 23, pp. 1–36, 2022.
- [26] G. Lan, "Policy mirror descent for reinforcement learning: linear convergence, new sampling complexity, and generalized problem classes," *Mathematical Programming*, 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 Journal on Optimization*, vol. 33, no. 2, pp. 1061–1091, 2023.
- [28] J. Bhandari and D. Russo, "Global optimality guarantees for policy gradient methods," *arXiv preprint arXiv:1906.01786*, 2019.
- [29] D. Silver, G. Lever, N. Manfied, O. Heess, T. Degris, D. Wierstra, and M. A. Riedmiller, "Deterministic policy gradient algorithms," in *International Conference on Machine Learning*. PMLR, 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 *Advances in Neural Information Processing Systems*, 2000, pp. 1057–1063.
- [31] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.
- [32] C. Berge, *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*. Courier Corporation, 1997.
- [33] P. R. Kumar and P. Varaiya, *Stochastic systems: Estimation, identification, and adaptive control*. Prentice Hall, NJ, 1986.
- [34] G. Sansone, *Orthogonal Functions, rev. English ed.* Dover, NY, 1991.
- [35] B. Schölkopf and A. J. Smola, *Learning with kernels*. Academic Press, MA, 1967.
- [36] C. Carmeli, E. De vito, and T. A., "Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem," *Analysis and Applications*, vol. 4, no. 4, pp. 377–408, 2006.
- [37] G. Strang, *Linear algebra and its applications, 3rd ed.* Harcourt, CA, 1988.
- [38] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific Belmont, MA, 1996.
- [39] J. H. Ahlberg, E. N. Nilson, and J. L. Walsh, *The Theory of Splines and Their Applications*. Academic Press, MA, 1967.