Multistep Look-Ahead Policy Iteration for Optimal Control of Discrete-Time Nonlinear Systems With Isoperimetric Constraints

Tao Li, Qinglai Wei, *Senior Member, IEEE*, and Fei-Yue Wang

***Abstract*—In this article, a novel multistep look-ahead policy iteration with isoperimetric constraints (MLPIIC) method is developed to solve infinite horizon optimal control problems (OCPs) with isoperimetric constraints for discrete-time nonlinear systems. In order to overcome the difficulty that Bellman's principle of optimality does not hold directly in OCPs with isoperimetric constraints, a method to approximate OCPs with isoperimetric constraints by OCPs with new constraints is developed. For the MLPIIC method initialized with an admissible control law, the convergence and optimality of the iterative value function and the feasibility of the iterative control law are proven. Utilizing the function approximator, the implementation of the MLPIIC method is described. Finally, simulation results are provided.**

***Index Terms*—Adaptive dynamic programming (ADP), isoperimetric constraints, nonlinear systems, optimal control, policy iteration.**

I. Introduction

IN RECENT years, there has been a dramatic increase in the demand for system performance. At the same time, modern industrial processes are increasingly complicated [1], [2], [3]. Hence, the optimal control of complex processes is becoming a great challenge. Although dynamic programming is a very useful tool in solving the Bellman equation which is always encountered in optimal control problems (OCPs), it is computationally untenable because of the curse of dimensionality [4]. Nevertheless, adaptive dynamic programming (ADP) algorithms are proposed by [5] and [6] as a way to obtain numerical solutions of the Bellman equation. There are several synonyms for ADP, including adaptive critic designs [7], [8], [9], ADP [10], [11], [12], [13], [14], approximate dynamic programming [15], [16], [17], [18], neuro-dynamic programming [19], [20], neural dynamic programming [21], [22], relaxed dynamic programming [23], [24], and reinforcement learning [25], [26], [27]. ADP methods have been utilized in many different control problems and received more and more attention [28], [29], [30], [31], [32].

The PI algorithm is an important iterative ADP algorithm, which is proposed by [33], and it is verified that the optimal control law is achieved as the iteration index tends to infinity. It is also shown that the system controlled by any iterative control law (ICL) is stable, which is a great merit of the PI algorithm. In [34], a generalized PI algorithm is proposed, and its convergence and optimality are proven. In [35], a local PI algorithm is proposed, which updates the iterative value function (IVF) on a subset of the state space to reduce the computation. Furthermore, the convergence of the multistep look-ahead version of the PI algorithm is analyzed in [36]. For discounted finite Markov decision processes (MDPs), the idea of the multistep look-ahead version of the PI algorithm is first briefly mentioned in [19]. Additionally, the policy improvement through solving a dynamic programming problem that involves feature-based aggregation is discussed in [37] and the first analysis for multistep look-ahead policy improvement is presented by [38] and [39].

Almost all discussions about PI algorithms are interested in optimizing a single performance index function [33], [34], [35], [36]. However, many applications do not require the performance index function to reach the minimum but require it to be less than an upper bound. For example, the objective of guaranteed cost control is to guarantee the boundedness of the performance index function for any uncertainty in the system. Guaranteed cost control problems can be solved by ADP methods [40], [41], [42]. Actually, many OCPs describe

Manuscript received 4 June 2023; accepted 9 October 2023. Date of publication 16 November 2023; date of current version 16 February 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFE0206100; in part by the National Natural Science Foundation of China under Grant 62073321; in part by the National Defense Basic Scientific Research Program under Grant JCKY2019203C029; in part by the Science and Technology Development Fund, Macau, SAR, under Grant FDCT-22-009-MISE, Grant 0060/2021/A2, and Grant 0015/2020/AMJ; and in part by the National Defense Basic Scientific Research Project under Grant JCKY2020130C025. This article was recommended by Associate Editor B. Zhao. *(Corresponding author: Qinglai Wei.)*

Tao Li is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: litao2019@ia.ac.cn).

Qinglai Wei is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Institute of Systems Engineering, Macau University of Science and Technology, Macau, China (e-mail: qinglai.wei@ia.ac.cn).

Fei-Yue Wang is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, also with the Institute of Systems Engineering, Macau University of Science and Technology, Macau, China, and also with the Parallel Intelligence Innovation Center, Qingdao Academy of Intelligent Industries, Qingdao 266109, China (e-mail: feiyue.wang@ia.ac.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSMC.2023.3327492.

Digital Object Identifier 10.1109/TSMC.2023.3327492

goals by two types of performance index functions, where one type of performance index function needs to be optimized, and the other type of performance index function is required to satisfy the upper bound constraint, simultaneously. In order to distinguish between these two types of performance index functions, the constrained performance index functions are referred to as constraint functions. In this research, OCPs minimizing a performance index function with constraint functions are known as OCPs with isoperimetric constraints [43], [44]. OCPs with isoperimetric constraints are also known as OCPs with integral constraints [45]. In [46], the necessary conditions for OCPs with isoperimetric constraints are derived. In [47], the OCP with isoperimetric constraints is transformed into the OCP for an augmented system. In [48], the application of optimal control with isoperimetric constraints in the chemotherapy of tumors is demonstrated. In [49], the OCP with isoperimetric constraints is solved by a novel value iteration method, which is initialized by a feasible control law. For linear OCPs subject to terminal and isoperimetric constraints, Sun [50] derived conditions for guaranteed solvability and addresses terminal constraints by adding penalty terms on terminal states. The results in [45] and [50] depend on the convexity of the OCP with linear systems, quadratic performance index functions, and isoperimetric quadratic constraints, and hence cannot trivially extend to nonlinear systems. OCPs with isoperimetric constraints are also similar to the optimization problems of constrained MDPs (CMDPs) [51] in terms of optimization goals. For finite CMDPs, Chow et al. [52] formulated safe reinforcement learning as CMDPs and presented a Lyapunov method to solve them. For CMDPs with Borel state and action spaces, assumptions that guarantee the solvability of CMDPs are studied considering unbounded reward functions [53]. However, it is not an easy task to extend research results on CMDPs [51], [52], [53] to discrete-time nonlinear dynamical systems in the control theory. For OCPs with isoperimetric constraints, the conclusions in [45] and [50] show that the optimal control is a nonlinear function of the current state and the initial state. These conclusions show that the optimal control law is related to the previous state, which violates the principle of optimality. In addition, according to the conclusion in [54], the value function of OCPs with isoperimetric constraints does not directly satisfy Bellman's principle of optimality. Therefore, the ADP methods cannot be directly applied to solve OCPs with isoperimetric constraints.

In this article, for discrete-time nonlinear OCPs with isoperimetric constraints, a novel multistep look-ahead policy iteration with isoperimetric constraints (MLPIIC) method is developed. The main contributions are summarized as follows.

1) In comparison with studies on linear OCPs with isoperimetric constraints [45], [50], we extend the problem to the case of infinite-horizon performance index functions and nonlinear systems. In comparison with the existing method in [49], the proposed MLPIIC method is initialized by an admissible control law rather than a feasible control law. In comparison with existing results on CMDPs [51], [52], [53], we study the infinite-horizon undiscounted performance index functions with unbounded utility functions and emphasize the stability of the system.
2) To overcome the difficulty that the value function of the OCP with isoperimetric constraints does not directly satisfy Bellman's principle of optimality, by constructing an auxiliary function, the OCP with isoperimetric constraints is approximated as a special OCP, where the principle of optimality holds.
3) Based on the approximation of the OCP with isoperimetric constraints, a novel MLPIIC method is developed to solve them. The MLPIIC method is initialized by an admissible control law, which is the same as traditional PI algorithms. We prove that the constraint function of any ICL is less than the given upper bound, and the system controlled by any ICL is stable. We also analyze the convergence and optimality of the MLPIIC method.
4) By utilizing function approximators, the method to construct an appropriate auxiliary function is developed, and the implementation of the MLPIIC method is described.

This article is organized as follows. In Section II, the OCP with isoperimetric constraints is formulated. In Section III, a method to approximate OCPs with isoperimetric constraints is developed and the MLPIIC method is derived. The convergence of the IVF is shown. In addition, the stability and feasibility of the ICL are demonstrated. In Section IV, the implementation of the MLPIIC method is described. In Section V, the simulation results are presented. Finally, conclusions and future works are given in Section VI.

II. Problem Formulation

The following nonlinear system is studied:

$$x_{k+1} = F(x_k, u_k), \quad k \in \mathbb{N} \tag{1}$$

where $x_k \in \mathbb{R}^n$ is the state vector, $u_k \in \mathbb{R}^m$ is the control vector, $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is the system function, and $\mathbb{N}$ is the set of natural numbers, i.e., $\mathbb{N} = \{0, 1, \ldots\}$.

The control vector is determined by a feedback control law $\mu : \mathbb{R}^n \to \mathbb{R}^m$, i.e., $u_k = \mu(x_k)$. For the system (1) with the initial state x_0 and the feedback control law $\mu(\cdot)$, the performance index function is given by

$$J_\mu(x_0) = \sum_{k=0}^{\infty} U\big(x_k^\mu, \mu(x_k^\mu)\big) \tag{2}$$

where $U(x_k^\mu, \mu(x_k^\mu)) \triangleq U_x(x_k^\mu) + U_u(\mu(x_k^\mu))$ is the utility function, and x_k^μ denotes the kth element in the state trajectory starting from x_0 and controlled by $\mu(\cdot)$, i.e., $x_k^\mu = F(x_{k-1}^\mu, \mu(x_{k-1}^\mu))$ $\forall k \in \mathbb{N}_+$, and $x_0^\mu = x_0$. Note that $\mathbb{N}_+$ is the set of positive natural numbers, i.e., $\mathbb{N}_+ = \{1, 2, \ldots\}$. In (2), $U_x : \mathbb{R}^n \to \mathbb{R}_+$ and $U_u : \mathbb{R}^m \to \mathbb{R}_+$ are continuous positive-definite functions, where $\mathbb{R}_+$ is the set of non-negative real numbers. Apart from the performance index function (2), a constraint function is given by

$$D_\mu(x_0) = \sum_{k=0}^{\infty} d\big(x_k^\mu, \mu(x_k^\mu)\big) \tag{3}$$

where $d(x_k^\mu, \mu(x_k^\mu)) \triangleq d_x(x_k^\mu) + d_u(\mu(x_k^\mu))$ is the constraint utility function. In (3), $d_x(\cdot)$ and $d_u(\cdot)$ are continuous positive-definite functions of the same dimension as $U_x(\cdot)$ and $U_u(\cdot)$, respectively.

Considering a compact set $\Omega \subset \mathbb{R}^n$, including the origin and x_0 as interior points, the following assumptions are standard.

Assumption 1: The system (1) is Lipschitz continuous and stabilizable on Ω. Besides, the system (1) satisfies $F(0, 0) = 0$.

The goal of the OCP with isoperimetric constraints is to solve a feedback control law $\mu(\cdot)$ to move the given initial state x_0 to the origin, minimize the performance index function (2), and simultaneously guarantee the constraint function (3) to satisfy

$$D_\mu(x_0) \le d_0 \tag{4}$$

where $d_0 > 0$ is a given finite upper bound for $D_\mu(x_0)$.

Problem 1 denotes the OCP for the system (1) at x_0 with a performance index (2) and an isoperimetric constraint (4).

Problem 1:

$$\begin{aligned} \min_{\mu(\cdot)} \quad & \left\{J_\mu(x_0) \ : \ D_\mu(x_0) \le d_0\right\} \\ \text{s.t.} \quad & x_{k+1} = F(x_k, \mu(x_k)) \quad \forall k \in \mathbb{N}. \end{aligned}$$

In order to avoid difficulties caused by the difference of the finiteness of $J_\mu(x_0)$ and $D_\mu(x_0)$, the following assumption is made.

Assumption 2: Suppose that if $\mu(\cdot)$ satisfies $J_\mu(x_k) < \infty \ \ \forall x_k$, then $D_\mu(x_k) < \infty \ \ \forall x_k$, and vice versa.

It is worth pointing out that Assumption 2 is a relatively strong assumption since $J_\mu(\cdot)$ and $D_\mu(\cdot)$ are given. We provide an example here to show that Assumption 2 is reasonable. Consider the usual quadratic utility function in (2) and constraint utility function in (3), i.e., $U(x_k, \mu(x_k)) = x_k^\top Q_1 x_k + \mu^\top(x_k) R_1 \mu(x_k)$, $d(x_k, \mu(x_k)) = x_k^\top Q_2 x_k + \mu^\top(x_k) R_2 \mu(x_k)$, where Q_1, R_1, Q_2, and R_2 are symmetric positive-definite matrices with appropriate dimensions. By using Cholesky factorization [55], we can rewrite Q_2 and R_2 as $Q_2 = \mathfrak{Q}_2 \mathfrak{Q}_2^\top$ and $R_2 = \mathfrak{R}_2 \mathfrak{R}_2^\top$, respectively, where $\mathfrak{Q}_2$ and $\mathfrak{R}_2$ are positive-definite lower-triangular matrices. Let $\bar{Q} = \mathfrak{Q}_2^{-1} Q_1 \mathfrak{Q}_2^{-\top}$ and $\bar{R} = \mathfrak{R}_2^{-1} R_1 \mathfrak{R}_2^{-\top}$. Then, according to the properties of the generalized Rayleigh quotient [55], we have

$$\theta_{\min}(\bar{Q}) x_k^\top Q_2 x_k \le x_k^\top Q_1 x_k \le \theta_{\max}(\bar{Q}) x_k^\top Q_2 x_k \tag{5}$$

and

$$\begin{aligned} \theta_{\min}(\bar{R}) \mu^\top(x_k) R_2 \mu(x_k) &\le \mu^\top(x_k) R_1 \mu(x_k) \\ &\le \theta_{\max}(\bar{R}) \mu^\top(x_k) R_2 \mu(x_k) \end{aligned} \tag{6}$$

where $\theta_{\min}(\cdot)$ and $\theta_{\max}(\cdot)$ denote the minimum and maximum eigenvalues of the matrix, respectively. According to (5) and (6), if $\mu(\cdot)$ satisfies $J_\mu(x_k) < \infty \ \ \forall x_k$, then we have

$$\sum_{j=0}^{\infty} x_{k+j}^\top Q_2 x_{k+j} \le \frac{1}{\theta_{\min}(\bar{Q})} \sum_{j=0}^{\infty} x_{k+j}^\top Q_1 x_{k+j} < \infty$$

and

$$\begin{aligned} &\sum_{j=0}^{\infty} \mu^\top(x_{k+j}) R_2 \mu(x_{k+j}) \\ &\quad \le \frac{1}{\theta_{\min}(\bar{R})} \sum_{j=0}^{\infty} \mu^\top(x_{k+j}) R_1 \mu(x_{k+j}) \\ &\quad < \infty. \end{aligned}$$

Thus, we get

$$D_\mu(x_k) = \sum_{j=0}^{\infty} x_{k+j}^\top Q_2 x_{k+j} + \mu^\top(x_{k+j}) R_2 \mu(x_{k+j}) < \infty.$$

If $\mu(\cdot)$ satisfies $D_\mu(x_k) < \infty \ \ \forall x_k$, we can get $\mu(\cdot)$ satisfies $J_\mu(x_k) < \infty \ \ \forall x_k$, in a similar way. Therefore, Assumption 2 holds in this example. We will also discuss the case where Assumption 2 does not hold at the end of Section III.

We introduce the following definition of admissible and feasible control laws.

Definition 1: A control law $\mu(\cdot)$ is defined to be admissible for x_0 if $\mu(\cdot)$ is continuous on Ω, $\mu(0) = 0$, $\mu(\cdot)$ stabilizes (1) starting from x_0, and $J_\mu(x_0)$ and $D_\mu(x_0)$ are finite.

Definition 2: A control law $\mu(\cdot)$ is defined to be feasible if $\mu(\cdot)$ is admissible and $D_\mu(x_0)$ satisfies (4).

Let $\mathfrak{A}(x_0)$ be the set containing all admissible control laws, which is related to the given initial state x_0, i.e.,

$$\begin{aligned} \mathfrak{A}(x_0) = \big\{\mu(\cdot) \ : \ & \mu(\cdot) \text{ stabilizes (1) starting from } x_0 \\ & \mu(0) = 0, J_\mu(x_0) < \infty, D_\mu(x_0) < \infty\big\}. \end{aligned}$$

Let $\mathfrak{F}(x_0)$ denote the set of all feasible control laws, i.e.,

$$\mathfrak{F}(x_0) = \left\{\mu(\cdot) \ : \ \mu(\cdot) \in \mathfrak{A}(x_0), D_\mu(x_0) \le d_0\right\}.$$

For OCPs with isoperimetric constraints, the following assumption is required to guarantee that $\mathfrak{A}(x_0) \neq \emptyset$ and $\mathfrak{F}(x_0) \neq \emptyset$.

Assumption 3: The set of admissible control laws $\mathfrak{A}(x_0)$ is not empty. The control law $\tilde{\mu}(\cdot)$ satisfying

$$\tilde{\mu}(\cdot) \in \arg\min_{\mu(\cdot)} \{D_\mu(x_0) \ : \ \mu(\cdot) \in \mathfrak{A}(x_0)\}$$

is feasible, i.e., $D_{\tilde{\mu}}(x_0) \le d_0$.

For the OCP with isoperimetric constraints, the optimal performance index function is defined as

$$J_c^*(x_0) = \min_{\mu(\cdot)} \{J_\mu(x_0) \ : \ \mu(\cdot) \in \mathfrak{F}(x_0)\} \tag{7}$$

and the optimal control law is defined as

$$\mu_c^*(\cdot) \in \arg\min_{\mu(\cdot)} \{J_\mu(x_0) \ : \ \mu(\cdot) \in \mathfrak{F}(x_0)\}. \tag{8}$$

Note that the solution $\mu_c^*(x_k)$ of minimization in (8) may not be unique. Motivated by [56], the notation "$\in$" is used here to allow the selection of any of the minimizers.

It is worth pointing out that the value function of OCPs with isoperimetric constraints does not directly satisfy the principle of optimality, Hence, the optimal control law $\mu_c^*(\cdot)$ is difficult to achieve via solving the Bellman equation.

III. Multistep Look-Ahead Policy Iteration With Isoperimetric Constraints Method

In this section, the OCP with isoperimetric constraints will be approximated by an OCP with new constraints where Bellman's principle of optimality holds. Then, in order to solve the above OCP with new constraints, the MLPIIC method will be developed. Next, the properties of the MLPIIC method will be proven.

A. Approximation of the OCP With Isoperimetric Constraints

For a control law $\mu(\cdot)$ and functions $Y : \mathbb{R}^n \to \mathbb{R}_+$, $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_+$, the operator $T_{\mu,h}[\,\cdot\,]$ is defined as

$$T_{\mu,h}[Y](x_k) = h(x_k, \mu(x_k)) + Y(F(x_k, \mu(x_k))). \tag{9}$$

Define the operator $T^j_{\mu,h}[\,\cdot\,]$ as the composition of $j \in \mathbb{N}$ operators $T_{\mu,h}[\,\cdot\,]$, i.e.,

$$T^j_{\mu,h}[Y](x_k) = T_{\mu,h}\big[T^{j-1}_{\mu,h}[Y]\big](x_k) \quad \forall j \in \mathbb{N}_+ \tag{10}$$

and

$$T^0_{\mu,h}[Y](x_k) = Y(x_k). \tag{11}$$

According to (9)–(11), we have

$$\begin{aligned} T^j_{\mu,h}[Y](x_k) &= h(x_k, \mu(x_k)) + T^{j-1}_{\mu,h}[Y]\big(x^\mu_{k+1}\big) \\ &= h(x_k, \mu(x_k)) + T_{\mu,h}\Big[T^{j-2}_{\mu,h}[Y]\Big]\big(x^\mu_{k+1}\big) \\ &= h(x_k, \mu(x_k)) + h\big(x^\mu_{k+1}, \mu(x^\mu_{k+1})\big) \\ &\quad + T^{j-2}_{\mu,h}[Y]\big(x^\mu_{k+2}\big) \\ &\;\;\vdots \\ &= \sum_{i=0}^{j-1} h\big(x^\mu_{k+i}, \mu(x^\mu_{k+i})\big) + T^0_{\mu,h}[Y]\Big(x^\mu_{k+j}\Big) \\ &= \sum_{i=0}^{j-1} h\big(x^\mu_{k+i}, \mu(x^\mu_{k+i})\big) + Y\Big(x^\mu_{k+j}\Big) \end{aligned} \tag{12}$$

where $x^\mu_k = x_k$. In the following lemma, it will be shown that $T_{\mu,h}[\,\cdot\,]$ is a monotonic operator.

Lemma 1 [57]: For positive-definite functions $Y(\cdot)$ and $Y'(\cdot)$, which satisfy

$$Y(x_k) \le Y'(x_k) \quad \forall x_k$$

we have

$$T_{\mu,h}[Y](x_k) \le T_{\mu,h}\big[Y'\big](x_k) \quad \forall x_k \quad \forall \mu(x_k).$$

Next, it will be shown that Problem 1 can be approximated by an OCP with a new constraint. Let $\mathfrak{A}(\Omega)$ denote the set of all admissible control laws for Ω, i.e.,

$$\begin{aligned} \mathfrak{A}(\Omega) = \big\{\mu(\cdot) :\ & \mu \text{ stabilizes (1) starting from } x_k \\ & \mu(0) = 0, J_\mu(x_k) < \infty, D_\mu(x_k) < \infty \\ & \forall x_k \in \Omega\big\}. \end{aligned}$$

Let $\mathfrak{F}(\Omega)$ denote the set containing all feasible control laws associated with Ω, i.e.,

$$\mathfrak{F}(\Omega) = \big\{\mu(\cdot) :\ \mu(\cdot) \in \mathfrak{A}(\Omega), D_\mu(x_0) \le d_0\big\}.$$

Since $x_0 \in \Omega$, we know $\mathfrak{A}(\Omega) \subseteq \mathfrak{A}(x_0)$ and $\mathfrak{F}(\Omega) \subseteq \mathfrak{F}(x_0)$ obviously. In the following assumption, it will be supposed that neither $\mathfrak{A}(\Omega)$ nor $\mathfrak{F}(\Omega)$ is an empty set.

Assumption 4: Suppose $\mathfrak{A}(\Omega) \neq \emptyset$ and $\mathfrak{F}(\Omega) \neq \emptyset$.

Suppose there is a known finite positive-definite auxiliary function $\Phi : \mathbb{R}^n \to \mathbb{R}_+$ that satisfies

$$\Phi(x_0) \le d_0. \tag{13}$$

For $p \in \mathbb{N}$, define a set of control laws as

$$S_p \triangleq \{\mu(\cdot) :\ \mu(\cdot) \in \mathfrak{A}(\Omega), T^{2^p}_{\mu,d}[\Phi](x_k) \le \Phi(x_k) \quad \forall x_k \in \Omega\}.$$

Then, the relationship between $S_p, p \in \mathbb{N}$, and $\mathfrak{F}(x_0)$ will be proven.

Theorem 1: If Assumptions 1–4 hold, then

$$\begin{aligned} S_0 \subseteq\ & S_1 \subseteq S_2 \subseteq \ldots \subseteq S_\infty \\ =\ & \{\mu(\cdot) :\ \mu(\cdot) \in \mathfrak{A}(\Omega), D_\mu(x_k) \le \Phi(x_k) \quad \forall x_k \in \Omega\} \\ \subseteq\ & \mathfrak{F}(\Omega) \subseteq \mathfrak{F}(x_0). \end{aligned} \tag{14}$$

Proof: First, we will prove

$$S_\infty = \{\mu(\cdot) :\ \mu(\cdot) \in \mathfrak{A}(\Omega), D_\mu(x_k) \le \Phi(x_k) \quad \forall x_k \in \Omega\}. \tag{15}$$

Let $\mu(\cdot) \in \mathfrak{A}(\Omega)$. According to (12), we have

$$\lim_{j\to\infty} T^{2^j}_{\mu,d}[\Phi](x_k) = \lim_{j\to\infty} \left\{ \sum_{i=0}^{2^j-1} d\big(x^\mu_{k+i}, \mu(x^\mu_{k+i})\big) + \Phi\Big(x^\mu_{k+2^j}\Big) \right\}.$$

Since $\mu(\cdot) \in \mathfrak{A}(\Omega)$, we know

$$\lim_{j\to\infty} x^\mu_{k+2^j} = 0.$$

Then, we have

$$\lim_{j\to\infty} \Phi\Big(x^\mu_{k+2^j}\Big) = 0$$

because $\Phi(\cdot)$ is positive definite. Thus, it can be derived that

$$\begin{aligned} \lim_{j\to\infty} T^{2^j}_{\mu,d}[\Phi](x_k) &= \lim_{j\to\infty} \left\{ \sum_{i=0}^{2^j-1} d\big(x^\mu_{k+i}, \mu(x^\mu_{k+i})\big) + \Phi\Big(x^\mu_{k+2^j}\Big) \right\} \\ &= D_\mu(x_k). \end{aligned}$$

Therefore, (15) is proven.

Next, we will prove

$$S_p \subseteq S_{p+1} \quad \forall p \in \mathbb{N}. \tag{16}$$

Let $\mu(\cdot) \in S_p, p \in \mathbb{N}$. Then, we have

$$T^{2^p}_{\mu,d}[\Phi](x_k) \le \Phi(x_k) \quad \forall x_k \in \Omega. \tag{17}$$

According to Lemma 1, using $T_{\mu,h}[\,\cdot\,]$ to map 2^p times on both sides of (17), we get

$$T^{2^{p+1}}_{\mu,d}[\Phi](x_k) \le T^{2^p}_{\mu,d}[\Phi](x_k) \quad \forall x_k \in \Omega.$$

Thus, we obtain

$$T_{\mu,d}^{2^{p+1}}[\Phi](x_k) \le \Phi(x_k) \quad \forall x_k \in \Omega$$

i.e., $\mu(\cdot) \in S_{p+1}$. Therefore, (16) is proven.

Since $\Phi(x_0) \le d_0$, we have

$$\{\mu(\cdot) : \mu(\cdot) \in \mathfrak{A}(\Omega), D_\mu(x_k) \le \Phi(x_k) \ \forall x_k \in \Omega\} \subseteq \mathfrak{F}(\Omega). \quad (18)$$

According to (15), (16), and (18), (14) can be derived. The proof is completed. ■

Since the value function of Problem 1 does not directly satisfy Bellman's principle of optimality, it is difficult to obtain $\mu_c^*(\cdot)$ by solving the Bellman equation. However, according to Theorem 1, we can use S_0 to approximate $\mathfrak{F}(x_0)$. Then, an OCP subject to $\mu(\cdot) \in S_0$ is constructed as Problem 2.

Problem 2:

$$\min_{\mu(\cdot)} \ \{J_\mu(x_0) : T_{\mu,d}[\Phi](x_k) \le \Phi(x_k) \quad \forall x_k \in \Omega, k \in \mathbb{N}\}$$
$$\text{s.t.} \quad x_{k+1} = F(x_k, \mu(x_k)) \quad \forall k \in \mathbb{N}.$$

Note that $S_p, p \in \mathbb{N}_+$, cannot be used to approximate $\mathfrak{F}(x_0)$ because the optimality principle does not hold in the OCP subject to $\mu(\cdot) \in S_p, p \in \mathbb{N}_+$.

The error resulting from approximating Problem 1 with Problem 2 is analyzed as follows. From (14), $S_\infty \backslash S_0 \ne \emptyset$

$$\{\mu(\cdot) : \mu(\cdot) \in \mathfrak{A}(\Omega), D_\mu(x_k) \le \Phi(x_k) \quad \forall x_k \in \Omega\} \backslash \{\mu(\cdot) : \mu(\cdot) \in \mathfrak{A}(\Omega), D_\mu(x_0) \le d_0\} \ne \emptyset \quad (19)$$

and

$$\mathfrak{F}(\Omega) \backslash \mathfrak{F}(x_0) \ne \emptyset$$

all contribute to the approximation error. According to (19), $\Phi(x_0)$ should be close to d_0 to reduce the approximation error. Since it is challenging to find a function $\Phi(\cdot)$ satisfying $S_\infty = S_0$, it will not be covered in this study. $\Phi(\cdot)$ should also guarantee $S_0 \ne \emptyset$ and the following theorem will provide a sufficient condition for $\Phi(\cdot)$ to guarantee $S_0 \ne \emptyset$.

Theorem 2: Let $\mu_\Phi(\cdot)$ be obtained by

$$\mu_\Phi(x_k) = \arg\min_{\mu(x_k)} T_{\mu,d}[\Phi](x_k) \quad \forall x_k \in \Omega.$$

If Assumptions 1–4 hold, and

$$\Phi(x_k) \ge T_{\mu_\Phi,d}[\Phi](x_k) \quad \forall x_k \in \Omega \quad (20)$$

then $S_0 \ne \emptyset$.

Proof: From (20), it can be derived that

$$\begin{aligned} &\Phi(F(x_k, \mu_\Phi(x_k))) - \Phi(x_k) \\ &\le -d(x_k, \mu_\Phi(x_k)) \\ &\le 0. \end{aligned}$$

According to the Lyapunov stability theory [58], $\Phi(\cdot)$ is a Lyapunov function and the system using $\mu_\Phi(\cdot)$ is asymptotically stable. From (20) and Lemma 1, we get

$$\begin{aligned} \Phi(x_k) &\ge T_{\mu_\Phi,d}[\Phi](x_k) \\ &\ge T_{\mu_\Phi,d}^{\infty}[\Phi](x_k) \\ &= \lim_{j\to\infty} \left\{ \sum_{i=0}^{j-1} d\left(x_{k+i}^{\mu_\Phi}, \mu_\Phi(x_{k+i}^{\mu_\Phi})\right) + \Phi\left(x_{k+j}^{\mu_\Phi}\right) \right\}. \end{aligned}$$

Since $\Phi(\cdot)$ is positive definite and $\mu_\Phi(\cdot)$ is asymptotically stable, we have

$$\begin{aligned} &\lim_{j\to\infty} \left\{ \sum_{i=0}^{j-1} d\left(x_{k+i}^{\mu_\Phi}, \mu_\Phi(x_{k+i}^{\mu_\Phi})\right) + \Phi\left(x_{k+j}^{\mu_\Phi}\right) \right\} \\ &= \lim_{j\to\infty} \sum_{i=0}^{j-1} d\left(x_{k+i}^{\mu_\Phi}, \mu_\Phi(x_{k+i}^{\mu_\Phi})\right) \\ &= D_{\mu_\Phi}(x_k). \end{aligned}$$

Thus, we obtain

$$D_{\mu_\Phi}(x_k) \le \Phi(x_k) < \infty \quad \forall x_k \in \Omega. \quad (21)$$

Then, according to Assumption 2, we know

$$J_{\mu_\Phi}(x_k) < \infty \quad \forall x_k \in \Omega. \quad (22)$$

According to (20)–(22), we know $\mu_\Phi(\cdot) \in S_0$. Therefore, $S_0 \ne \emptyset$. The proof is completed. ■

According to the above analysis, the finite positive-definite function $\Phi(\cdot)$ should satisfy (13) and (20). It is also implied that $\Phi(x_0)$ close to d_0 helps reduce the approximation error. A method to find $\Phi(\cdot)$ is given in Section IV-A.

For Problem 2, assuming that there is a known $\Phi(\cdot)$ satisfying (13) and (20), the optimal performance index function $J^*(\cdot)$ satisfies the Bellman equation

$$J^*(x_k) = \min_{\mu(x_k)} \left\{ T_{\mu,U}\left[J^*\right](x_k) : T_{\mu,d}[\Phi](x_k) \le \Phi(x_k) \right\} \quad (23)$$

and the optimal control law $\mu^*(\cdot)$ satisfies

$$\mu^*(x_k) \in \arg\min_{\mu(x_k)} \left\{ T_{\mu,U}\left[J^*\right](x_k) : T_{\mu,d}[\Phi](x_k) \le \Phi(x_k) \right\}. \quad (24)$$

According to the above analysis, solving Problem 1 is transformed into solving the Bellman equation (23). Next, the MLPIIC method for solving (23) is introduced.

B. Derivation of the MLPIIC Method

In this section, the MLPIIC method to solve (23) is developed. For simplicity of expression, define the control vector sequence of $\mu(\cdot)$ from k to $k+l-1$ as

$$\underline{\mu}_{k \ : \ k+l-1} = \left\{\mu(x_k), \mu\left(x_{k+1}^{\mu}\right), \ldots, \mu\left(x_{k+l-1}^{\mu}\right)\right\} \in \mathbb{R}^{ml}$$

where l is a positive integer.

The MLPIIC method is given as follows. Let $\mu_0(\cdot) \in \mathfrak{A}(\Omega)$ be an arbitrary admissible control law. For $i = 1, 2, \ldots$, the MLPIIC method iterates between the policy evaluation equation

$$V_i^\mu(x_k) = T_{\mu_{i-1},U}\left[V_i^\mu\right](x_k) \quad \forall x_k \in \Omega \quad (25)$$

and the l-step look-ahead policy update equation with isoperimetric constraints

$$\begin{aligned} \mu_i(x_k) \in \arg \min_{\underline{\mu}_{k \ : \ k+l-1}} \ & T_{\mu,U}^l\left[V_i^\mu\right](x_k) \\ \text{s.t. } & T_{\mu,d}[\Phi]\left(x_{k+j}^{\mu}\right) \le \Phi\left(x_{k+j}^{\mu}\right), j = 0, 1, \ldots, l-1 \end{aligned} \quad (26)$$

for all $x_k \in \Omega$. The detailed explanation of (26) is as follows. For a certain x_k and $V_i^\mu(\cdot)$, $T_{\mu,U}^l[V_i^\mu](x_k)$ is a function with respect to $\underline{\mu}_{k\ :\ k+l-1}$ and $T_{\mu,d}[\Phi](x_{k+j}^\mu) \le \Phi(x_{k+j}^\mu), j = 0, 1, \ldots, l-1$, are constraints of $\underline{\mu}_{k\ :\ k+l-1}$. The meaning of (26) is to first solve $\underline{\mu}_{i,k\ :\ k+l-1}^* = \{\mu_i^*(x_k), \mu_i^*(x_{k+1}^{\mu_i^*}), \ldots, \mu_i^*(x_{k+l-1}^{\mu_i^*})\}$ that satisfies

$$\underline{\mu}_{i,k\ :\ k+l-1}^* \in \arg\min_{\underline{\mu}_{k\ :\ k+l-1}} T_{\mu,U}^l\left[V_i^\mu\right](x_k)$$
$$\text{s.t.}\quad T_{\mu,d}[\Phi]\left(x_{k+j}^\mu\right) \le \Phi\left(x_{k+j}^\mu\right), j = 0, 1, \ldots, l-1 \quad (27)$$

and then let $\mu_i(x_k) = \mu_i^*(x_k)$.

The MLPIIC method for discrete-time nonlinear systems is summarized in Algorithm 1.

When (25) cannot be solved directly, an iterative process to solve $V_i^\mu(\cdot)\ \ \forall i \in \mathbb{N}_+$ is given as

$$V_i^\mu(x_k) = \lim_{j\to\infty} T_{\mu_{i-1},U}^j\left[V_{i-1}^\mu\right](x_k)\quad \forall x_k \in \Omega\quad \forall i \in \mathbb{N}_+ \quad (28)$$

where $V_0^\mu(\cdot)$ is an arbitrary positive-definite function. The convergence of (28) is proven in [59].

The developed MLPIIC method possesses inherent differences with the traditional PI algorithms [33], [34], [35]. First, traditional PI methods are generally suitable for optimizing a single performance index function, which cannot work out the OCPs with isoperimetric constraints. However, the MLPIIC method aims to solve OCPs with isoperimetric constraints. Second, for traditional PI algorithms, the ICL is updated with 1 step look-ahead, while it is updated with multistep look-ahead in the MLPIIC method. Third, for traditional PI algorithms, the ICL is obtained through solving unconstrained optimization problems. However, in the MLPIIC method, the ICL is achieved by resolving inequality-constrained optimization problems.

C. Properties of the MLPIIC Method

In this section, the properties of the MLPIIC method will be analyzed. First, Theorem 3 shows that the IVF is nonincreasing and the ICL is feasible. Then, Theorem 4 demonstrates the convergence and optimality of the MLPIIC method.

For simplicity of expression, based on (9)–(11), the operator $T_{\min,h}^l[\,\cdot\,]$ is defined as

$$T_{\min,h}^l[Y](x_k) = \min_{\underline{\mu}_{k\ :\ k+l-1}} T_{\mu,h}^l[Y](x_k)$$
$$\text{s.t.}\quad T_{\mu,d}[\Phi]\left(x_{k+j}^\mu\right) \le \Phi\left(x_{k+j}^\mu\right), j = 0, 1, \ldots, l-1. \quad (29)$$

Let $T_{\min,h}^{jl}[\,\cdot\,]$ be the composition of j operators $T_{\min,h}^l[\,\cdot\,]$, i.e.,

$$T_{\min,h}^{jl}[Y](x_k) = T_{\min,h}^l\left[T_{\min,h}^{(j-1)l}[Y]\right](x_k)\quad \forall j \in \mathbb{N}_+ \quad (30)$$

and

$$T_{\min,h}^{0l}[Y](x_k) = Y(x_k). \quad (31)$$

Theorem 3: For $i = 1, 2, \ldots$, let $V_i^\mu(\cdot)$ and $\mu_i(\cdot)$ be obtained by the MLPIIC method (25)–(26) with $\mu_0(\cdot) \in \mathfrak{A}(\Omega)$. If Assumptions 1–4 hold, then $V_i^\mu(\cdot)$ is positive definite for $i \in \mathbb{N}_+$ and nonincreasing as i increases for $i \in \mathbb{N}_+\backslash\{1\}$, and $\mu_i(\cdot)$ is feasible for $i \in \mathbb{N}_+$.

Algorithm 1 MLPIIC

Initialization: Select a compact set $\Omega \subset \mathbb{R}^n$ including x_0 and the origin.
Select a finite positive definite function $\Phi(\cdot)$ that satisfies (13) and (20).
Select a computation precision ε.
Select the number of steps to look ahead l.
Select an initial admissible control law $\mu_0(\cdot)$.
Let the iteration index i be 0.

Iteration:

1: Let $i = i + 1$. Obtain $V_i^\mu(\cdot)$ by solving
$$V_i^\mu(x_k) = T_{\mu_{i-1},U}\left[V_i^\mu\right](x_k)\quad \forall x_k \in \Omega.$$

2: Obtain $\mu_i(\cdot)$ by solving
$$\mu_i(x_k) \in \arg\min_{\underline{\mu}_{k\ :\ k+l-1}} T_{\mu,U}^l\left[V_i^\mu\right](x_k)$$
$$\text{s.t. } T_{\mu,d}[\Phi]\left(x_{k+j}^\mu\right) \le \Phi\left(x_{k+j}^\mu\right), j = 0, 1, \ldots, l-1$$
for all $x_k \in \Omega$.

3: If $\forall x_k \in \Omega, |V_i^\mu(x_k) - V_{i-1}^\mu(x_k)| \le \varepsilon$, goto next step. Otherwise, goto Step 1.

4: **return** $J^*(\cdot) = V_i^\mu(\cdot)$ and $\mu^*(\cdot) = \mu_i(\cdot)$.

Proof: First, we will prove $V_1^\mu(\cdot)$ is positive definite. Since $V_1^\mu(\cdot)$ satisfies that

$$V_1^\mu(x_k) = T_{\mu_0,U}\left[V_1^\mu\right](x_k)\quad \forall x_k \in \Omega \quad (32)$$

we have

$$V_1^\mu(x_k) = \sum_{j=0}^{\infty} U\left(x_{k+j}^{\mu_0}, \mu_0(x_{k+j}^{\mu_0})\right) \quad (33)$$

because $\mu_0(\cdot)$ is admissible. Thus, we get

$$V_1^\mu(x_k) > 0\quad \forall x_k \ne 0, V_1^\mu(0) = 0. \quad (34)$$

Therefore, $V_1^\mu(\cdot)$ is positive definite.

Second, we will prove $\mu_1(\cdot)$ is feasible. When $i = 1$, we have

$$\mu_1(x_k) \in \arg\min_{\underline{\mu}_{k\ :\ k+l-1}} T_{\mu,U}^l\left[V_1^\mu\right](x_k)$$
$$\text{s.t.}\quad T_{\mu,d}[\Phi]\left(x_{k+j}^\mu\right) \le \Phi\left(x_{k+j}^\mu\right), j = 0, 1, \ldots, l-1 \quad (35)$$

for all $x_k \in \Omega$. Thus, $\mu_1(\cdot)$ satisfies

$$T_{\mu_1,d}[\Phi](x_k) \le \Phi(x_k)\quad \forall x_k \in \Omega. \quad (36)$$

Then, we get

$$\begin{aligned}&\Phi(F(x_k, \mu_1(x_k))) - \Phi(x_k)\\ &\le -d(x_k, \mu_1(x_k))\\ &\le 0.\end{aligned} \quad (37)$$

By the Lyapunov stability criteria [58], $\Phi(\cdot)$ is a Lyapunov function and $\mu_1(\cdot)$ is asymptotically stable. From (36) and Lemma 1, it can be derived that

$$\begin{aligned}\Phi(x_k) &\geq T_{\mu_1,d}[\Phi](x_k)\\ &\geq \lim_{j\to\infty} T^j_{\mu_1,d}[\Phi](x_k)\\ &= \lim_{j\to\infty}\left\{\sum_{i=0}^{j-1} d\left(x^{\mu_1}_{k+i}, \mu_1(x^{\mu_1}_{k+i})\right) + \Phi\left(x^{\mu_1}_{k+j}\right)\right\}\\ &= D_{\mu_1}(x_k) \end{aligned} \tag{38}$$

because $\mu_1(\cdot)$ is asymptotically stable and $\Phi(\cdot)$ is positive definite. Then, $D_{\mu_1}(\cdot)$ is finite. According to Assumption 2, we know $J_{\mu_1}(\cdot)$ is finite. Thus, $\mu_1(\cdot)$ is admissible. Based on (13) and (38), we get

$$D_{\mu_1}(x_0) \leq \Phi(x_0) \leq d_0. \tag{39}$$

Therefore, $\mu_1(\cdot)$ is a feasible control law.

Next, for $i = 2$, we will prove $\mu_2(\cdot)$ is feasible and $V^\mu_3(x_k) \leq V^\mu_2(x_k) \quad \forall x_k \in \Omega$. Because $\mu_1(\cdot)$ is feasible, we can prove $V^\mu_2(\cdot)$ is positive definite similar to (32)–(34). Since $\mu_2(\cdot)$ is obtained by (26), we have

$$\begin{aligned}&\mu_2(x_k) \in \arg \min_{\underline{\mu}_k \,:\, k+l-1} \quad T^l_{\mu,U}\left[V^\mu_2\right](x_k)\\ &\text{s.t.} \quad T_{\mu,d}[\Phi]\left(x^\mu_{k+j}\right) \leq \Phi\left(x^\mu_{k+j}\right), j = 0, 1, \ldots, l-1 \end{aligned} \tag{40}$$

for all $x_k \in \Omega$. Then, similar to (36)–(39), we can prove $\mu_2(\cdot)$ is a feasible control law. According to (40), we have

$$\begin{aligned}T^l_{\mu_2,U}\left[V^\mu_2\right](x_k) &= T^l_{\min,U}\left[V^\mu_2\right](x_k)\\ &\leq T^l_{\mu_1,U}\left[V^\mu_2\right](x_k)\\ &= V^\mu_2(x_k).\end{aligned}$$

Then, according to (12) and Lemma 1, we have

$$\begin{aligned}V^\mu_2(x_k) &\geq T^l_{\mu_2,U}\left[V^\mu_2\right](x_k)\\ &\geq T^{2l}_{\mu_2,U}\left[V^\mu_2\right](x_k)\\ &\vdots\\ &\geq \lim_{j\to\infty} T^j_{\mu_2,U}\left[V^\mu_2\right](x_k)\\ &= V^\mu_3(x_k) \quad \forall x_k \in \Omega.\end{aligned}$$

Suppose that the statement holds for $i = q-1$, $q \in \mathbb{N}_+$ and $q \geq 3$. Because $\mu_{q-1}(\cdot)$ is feasible, we can prove $V^\mu_q(\cdot)$ is positive definite similar to (32)–(34). Since $\mu_q(\cdot)$ is obtained by (26), we have

$$\begin{aligned}&\mu_q(x_k) \in \arg \min_{\underline{\mu}_k \,:\, k+l-1} \quad T^l_{\mu,U}\left[V^\mu_q\right](x_k)\\ &\text{s.t.} \quad T_{\mu,d}[\Phi]\left(x^\mu_{k+j}\right) \leq \Phi\left(x^\mu_{k+j}\right), j = 0, 1, \ldots, l-1 \end{aligned} \tag{41}$$

for all $x_k \in \Omega$. Then, similar to (36)–(39), we can prove $\mu_q(\cdot)$ is a feasible control law. As $\mu_{q-1}(\cdot)$ is a feasible control law, we have

$$\begin{aligned}T^l_{\mu_q,U}\left[V^\mu_q\right](x_k) &= T^l_{\min,U}\left[V^\mu_q\right](x_k)\\ &\leq T^l_{\mu_{q-1},U}\left[V^\mu_q\right](x_k)\\ &= V^\mu_q(x_k).\end{aligned}$$

Then, according to (12) and Lemma 1, we have

$$\begin{aligned}V^\mu_q(x_k) &\geq T^l_{\mu_q,U}\left[V^\mu_q\right](x_k)\\ &\geq T^{2l}_{\mu_q,U}\left[V^\mu_q\right](x_k)\\ &\vdots\\ &\geq \lim_{j\to\infty} T^j_{\mu_q,U}\left[V^\mu_q\right](x_k)\\ &= V^\mu_{q+1}(x_k) \quad \forall x_k \in \Omega.\end{aligned} \tag{42}$$

According to mathematical induction, the proof is completed. ■

Note that the sequence of the IVF $\{V^\mu_i(\cdot)\}_{i=1}^\infty$ is pointwise nonincreasing in the traditional PI algorithm, while $\{V^\mu_i(\cdot)\}_{i=2}^\infty$ is pointwise nonincreasing in the MLPIIC method. Besides, all ICLs are admissible in the traditional PI algorithm, while the initial control law is admissible and subsequent ICLs are feasible in the proposed MLPIIC method. Next, the convergence and optimality of the MLPIIC method are proven.

Theorem 4: For $i = 1, 2, \ldots$, let $V^\mu_i(\cdot)$ and $\mu_i(\cdot)$ be obtained by the MLPIIC method (25)–(26) with $\mu_0(\cdot) \in \mathfrak{A}(\Omega)$. If Assumptions 1–4 hold, then the IVF $V^\mu_i(\cdot)$ converges to the optimal performance index function $J^*(\cdot)$, as $i \to \infty$, where $J^*(\cdot)$ satisfies the Bellman equation (23).

Proof: According to Theorem 3, $V^\mu_i(\cdot)$ is nonincreasing as i increases for $i \in \mathbb{N}_+\backslash\{1\}$ and also lower bounded by zero. Hence, the limit of $V^\mu_i(\cdot)$ exists as $i \to \infty$. The limit of $V^\mu_i(\cdot)$ is defined as

$$V^\mu_\infty(x_k) = \lim_{i\to\infty} V^\mu_i(x_k).$$

The limit of $\mu_i(\cdot)$ is defined as

$$\mu_\infty(x_k) = \lim_{i\to\infty} \mu_i(x_k).$$

First, we will prove $V^\mu_\infty(\cdot)$ satisfies the l-step look-ahead version of the Bellman equation, i.e.,

$$V^\mu_\infty(x_k) = T^l_{\min,U}\left[V^\mu_\infty\right](x_k). \tag{43}$$

According to (42), we get

$$V^\mu_\infty(x_k) \leq V^\mu_{i+1}(x_k) \leq T^l_{\mu_i,U}\left[V^\mu_i\right](x_k) = T^l_{\min,U}\left[V^\mu_i\right](x_k).$$

Let $i \to \infty$, we obtain

$$V^\mu_\infty(x_k) \leq T^l_{\min,U}\left[V^\mu_\infty\right](x_k) \tag{44}$$

Since $\lim_{i\to\infty} V^\mu_i(x_k) = V^\mu_\infty(x_k)$ and $V^\mu_i(\cdot)$ is nonincreasing $\forall \varepsilon > 0$, $\exists q \in \mathbb{N}_+$, such that

$$V^\mu_q(x_k) \geq V^\mu_\infty(x_k) \geq V^\mu_q(x_k) - \varepsilon. \tag{45}$$

Hence, we can get

$$\begin{aligned}V^\mu_\infty(x_k) &\geq T_{\mu_{q-1},U}\left[V^\mu_q\right](x_k) - \varepsilon\\ &= T^l_{\mu_{q-1},U}\left[V^\mu_q\right](x_k) - \varepsilon\\ &\geq T^l_{\mu_{q-1},U}\left[V^\mu_\infty\right](x_k) - \varepsilon\\ &\geq T^l_{\min,U}\left[V^\mu_\infty\right](x_k) - \varepsilon.\end{aligned}$$

Since ε is arbitrary, we obtain

$$V^\mu_\infty(x_k) \geq T^l_{\min,U}\left[V^\mu_\infty\right](x_k). \tag{46}$$

Combining (44) and (46), we can obtain (43).

Second, we will prove $V_\infty^\mu(\cdot)$ is the unique solution of the l-step look-ahead version of the Bellman equation (43) in Ω. Assume that there is another solution $V^*(\cdot)$ of (43), i.e.,

$$V^*(x_k) = T^l_{\min,U}[V^*](x_k)$$

and

$$\mu_V^*(x_k) \in \arg \min_{\underline{\mu}_k \,:\, k+l-1} T^l_{\mu,U}[V^*](x_k)$$
$$\text{s.t. } T_{\mu,d}[\Phi]\left(x^\mu_{k+j}\right) \le \Phi\left(x^\mu_{k+j}\right), j = 0, 1, \ldots, l-1.$$

Next, we will prove $V^*(x_k) = V_\infty^\mu(x_k) \ \ \forall x_k \in \Omega$. Assume that

$$V^*(x_k) < V_\infty^\mu(x_k), \exists x_k \in \Omega. \tag{47}$$

Then, we can prove

$$T^l_{\mu_V^*,U}[V_\infty^\mu](x_k) < V_\infty^\mu(x_k), \exists x_k \in \Omega \tag{48}$$

by contradiction. Assuming that (48) does not hold, we have

$$T^l_{\mu_V^*,U}[V_\infty^\mu](x_k) \ge V_\infty^\mu(x_k) \ \ \forall x_k \in \Omega.$$

Then, we can derive that

$$\sum_{j=0}^{l-1} U\left(x^{\mu_V^*}_{k+j}, \mu_V^*(x^{\mu_V^*}_{k+j})\right) + V_\infty^\mu\left(x^{\mu_V^*}_{k+l}\right) \ge V_\infty^\mu\left(x^{\mu_V^*}_k\right) \tag{49}$$

for all $x_k^{\mu_V^*} \in \Omega$ and

$$\sum_{j=0}^{l-1} U\left(x^{\mu_V^*}_{k+l+j}, \mu_V^*(x^{\mu_V^*}_{k+l+j})\right) + V_\infty^\mu\left(x^{\mu_V^*}_{k+2l}\right) \ge V_\infty^\mu\left(x^{\mu_V^*}_{k+l}\right) \tag{50}$$

for all $x_{k+l}^{\mu_V^*} \in \Omega$. Substituting (50) into (49), we get

$$\sum_{j=0}^{2l-1} U\left(x^{\mu_V^*}_{k+j}, \mu_V^*(x^{\mu_V^*}_{k+j})\right) + V_\infty^\mu\left(x^{\mu_V^*}_{k+2l}\right) \ge V_\infty^\mu\left(x^{\mu_V^*}_k\right) \tag{51}$$

for all $x_k^{\mu_V^*} \in \Omega$. Repeating the process (49)–(51) for $N-2$ times, where $N = 2, 3, \ldots$, we obtain

$$\sum_{j=0}^{Nl-1} U\left(x^{\mu_V^*}_{k+j}, \mu_V^*(x^{\mu_V^*}_{k+j})\right) + V_\infty^\mu\left(x^{\mu_V^*}_{k+Nl}\right) \ge V_\infty^\mu\left(x^{\mu_V^*}_k\right) \ \ \forall x_k^{\mu_V^*} \in \Omega.$$

Let $N \to \infty$. Since $V_\infty^\mu(\cdot)$ is positive definite and $\mu_V^*(\cdot)$ is admissible, we have

$$\lim_{N\to\infty} V_\infty^\mu\left(x^{\mu_V^*}_{k+Nl}\right) = 0.$$

Thus, we obtain

$$V^*(x_k) \ge V_\infty^\mu(x_k) \ \ \forall x_k \in \Omega$$

which contradicts (47). Therefore, if (47) holds, then (48) can be derived. According to (48), we have

$$T^l_{\mu_V^*,U}[V_\infty^\mu](x_k) < V_\infty^\mu(x_k) = T^l_{\mu_\infty,U}[V_\infty^\mu](x_k), \exists x_k \in \Omega$$

which contradicts

$$\mu_\infty(x_k) \in \arg \min_{\underline{\mu}_k \,:\, k+l-1} T^l_{\mu,U}[V_\infty^\mu](x_k)$$
$$\text{s.t. } T_{\mu,d}[\Phi]\left(x^\mu_{k+j}\right) \le \Phi\left(x^\mu_{k+j}\right), j = 0, 1, \ldots, l-1.$$

Therefore, (47) does not hold, i.e.,

$$V^*(x_k) \ge V_\infty^\mu(x_k) \ \ \forall x_k \in \Omega.$$

Similarly, we can prove

$$V^*(x_k) \le V_\infty^\mu(x_k) \ \ \forall x_k \in \Omega$$

by contradiction. Thus, $V^*(\cdot) = V_\infty^\mu(\cdot)$. Therefore, $V_\infty^\mu(\cdot)$ is the unique solution of (43) in Ω.

According to Bellman's principle of optimality, $J^*(\cdot)$ satisfies (43). Therefore, $J^*(\cdot) = V_\infty^\mu(\cdot)$. ■

From Theorem 3 to Theorem 4, we have discussed properties of the MLPCC method when Assumption 2 holds. Furthermore, if we have an initial feasible control law $\mu_f(\cdot)$, then Assumption 2 can be removed and we have the following corollary.

Corollary 1: For $i = 1, 2, \ldots$, let $V_i^\mu(\cdot)$ and $\mu_i(\cdot)$ be obtained by the MLPIIC method (25)–(26) with $\mu_0(\cdot) = \mu_f(\cdot)$ and $\Phi(\cdot) = D_{\mu_f}(\cdot)$. If Assumptions 1, 3, and 4 hold, then $V_i^\mu(\cdot)$ is positive definite and nonincreasing as i increases for $i \in \mathbb{N}_+$. $V_i^\mu(\cdot)$ also converges to the optimal performance index function $J^*(\cdot)$, as $i \to \infty$, where $J^*(\cdot)$ satisfies the Bellman equation (23). Furthermore, for all $i \in \mathbb{N}$, $\mu_i(\cdot)$ is feasible.

Proof: The corollary can be proven similar to Theorems 3 and 4 and the proof is omitted. ■

Remark 1: Optimal control with isoperimetric constraints has been widely applied in boiler–turbine generating systems [60], [61], chemotherapy of tumors [48], and risk-aware control [62]. For boiler–turbine generating systems, by applying the proposed MLPIIC method, an economical control scheme satisfying the power generation demand can be obtained. For the chemotherapy of tumors, by applying the proposed MLPIIC method, we can obtain a treatment plan to minimize cancer cells using a certain amount of drugs. For the risk-aware control of robotics, by applying the proposed MLPIIC method, we can obtain a high-performance and safe control scheme.

IV. Implementation of the MLPIIC Method

In this section, the implementation of the MLPIIC method will be described in detail. First, a method to determine an appropriate auxiliary function $\Phi(\cdot)$ will be introduced. Second, the implementation of the MLPIIC method will be described.

A. Determination of the Auxiliary Function

In this section, the method of determining a finite positive-definite function $\Phi(\cdot)$ that satisfies (13) and (20) will be introduced.

First, by the function approximation theory [63], $\Phi(\cdot)$ is parameterized as

$$\Phi(x_k) = \sum_{j=1}^{N_\Phi} w_j \phi_j(x_k) = W_\Phi^\top \phi(x_k) \tag{52}$$

where N_Φ is the number of basis functions, $\{\phi_j(\cdot)\}_{j=1}^{N_\Phi}$ is a set of continuously differentiable and linearly independent basis functions satisfying $\phi_j(0) = 0$ for $j = 1, 2, \ldots, N_\Phi$, and $W_\Phi = [w_1 \ w_2 \ \cdots \ w_{N_\Phi}]^\top$ is the weight of the basis functions. Then, the problem of determining $\Phi(\cdot)$ is transformed into the problem of determining W_Φ. Second, let an array containing p state vectors randomly sampled from Ω be $\mathfrak{X}$, i.e., $\mathfrak{X} = \{x_k^1, x_k^2, \ldots, x_k^p\}$. Noting that $\Phi(x_0)$ close to d_0 helps reduce the approximation error between Problems 1 and 2, so we set $\rho > 0$ as an upper bound of $d_0 - \Phi(x_0)$. Then, W_Φ can be achieved by solving the following nonlinear inequality equations:

$$\begin{cases} W_\Phi^\top \phi\left(x_k^j\right) > 0, j = 1, 2, \ldots, p \\ W_\Phi^\top \phi\left(x_k^j\right) \geq \min_{\mu(x_k)} T_{\mu,U}\left[W_\Phi^\top \phi\right]\left(x_k^j\right), j = 1, 2, \ldots, p \\ d_0 - \rho \leq W_\Phi^\top \phi(x_0) \leq d_0. \end{cases} \tag{53}$$

The nonlinear inequality (53) can be solved by many methods. In this article, we provide a method to solve (53) as follows. Let one of the solutions of (53) be W_Φ^*. The problem of solving (53) can be transformed into solving the following optimization problem:

$$\begin{aligned} W_\Phi^* \in \arg\min_{W_\Phi} \ & c \\ \text{s.t.} \ & (53) \end{aligned} \tag{54}$$

where c is an arbitrary constant. Then, (54) can be solved by many mature optimization algorithms, such as the interior point method [64] and the genetic algorithm [65].

B. Implementation of the MLPIIC Method

In order to implement the MLPIIC method, similar to (52), we introduce two sets of continuously differentiable and linearly independent basis functions $\{\varphi_j(\cdot)\}_{j=1}^{N_V}$ and $\{\psi_j(\cdot)\}_{j=1}^{N_\mu}$ satisfying $\varphi_j(0) = 0$ and $\psi_j(0) = 0$ for $j = 1, 2, \ldots$, where N_V and N_μ are the numbers of basis functions. Then, $V_i^\mu(\cdot)$ and $\mu_i(\cdot)$ are approximated as

$$\hat{V}_i^\mu(x_k) = \sum_{j=1}^{N_V} w_j^V \varphi_j(x_k) = {W_i^V}^\top \varphi(x_k)$$

and

$$\hat{\mu}_i(x_k) = \sum_{j=1}^{N_\mu} w_j^\mu \psi_j(x_k) = {W_i^\mu}^\top \psi(x_k)$$

where $\varphi(x_k) = [\varphi_1(x_k) \ \varphi_2(x_k) \ \cdots \ \varphi_{N_V}(x_k)]^\top$, $\psi(x_k) = [\psi_1(x_k) \ \psi_2(x_k) \ \cdots \ \psi_{N_\mu}(x_k)]^\top$, $W_i^V = [w_{i,1}^V \ w_{i,2}^V \ \cdots \ w_{i,N_V}^V]^\top$ and $W_i^\mu = [w_{i,1}^\mu \ w_{i,2}^\mu \ \cdots \ w_{i,N_\mu}^\mu]^\top$ are weights of $\hat{V}_i^\mu(\cdot)$ and $\hat{\mu}_i(\cdot)$, respectively. Given input sets and target sets of $\hat{V}_i^\mu(\cdot)$ and $\hat{\mu}_i(\cdot)$, a least-squares solution of W_i^V and W_i^μ can be obtained by the weighted residuals method [66].

The implementation of the MLPIIC method is described as follows. The MLPIIC method is first initialized with an admissible control law $\mu_0(\cdot)$, which can be obtained by feedback linearization [67]. In the first step of the MLPIIC method, we know $V_i^\mu(\cdot)$ can be obtained by (28). However, this process cannot be implemented infinite times. In the implementation of the MLPIIC method, we actually compute the target of $\hat{V}_i^\mu(\cdot)$ by

$$\begin{aligned} &\hat{V}_{i,\text{target}}^\mu(x_k) \\ &= T_{\hat{\mu}_{i-1},U}^K\left[\hat{V}_{i-1}^\mu\right](x_k) \\ &= \sum_{j=0}^{K-1} U\left(x_{k+j}^{\hat{\mu}_{i-1}}, \hat{\mu}_{i-1}(x_{k+j}^{\hat{\mu}_{i-1}})\right) + \hat{V}_{i-1}^\mu\left(x_{k+K}^{\hat{\mu}_{i-1}}\right) \end{aligned} \tag{55}$$

for all $i \in \mathbb{N}_+$, where K is a large positive integer, $\hat{\mu}_0(\cdot) = \mu_0(\cdot)$ and $\hat{V}_0^\mu(\cdot)$ is an arbitrary positive-definite function. In this article, we set K to be a multiple of l, i.e., $K = \kappa l$, where $\kappa \in \mathbb{N}_+$. In the second step, the target of $\hat{\mu}_i(\cdot)$ is calculated by

$$\begin{aligned} \hat{\mu}_i^{\text{target}}(x_k) \in \arg \min_{\underline{\mu}_k \ : \ k+l-1} \ & T_{\mu,U}^l\left[\hat{V}_i^\mu\right](x_k) \\ \text{s.t.} \ \ & T_{\mu,d}[\Phi]\left(x_{k+j}^\mu\right) \leq \Phi\left(x_{k+j}^\mu\right), j = 0, 1, \ldots, l-1. \end{aligned} \tag{56}$$

The implementation of the MLPIIC method is summarized in Algorithm 2.

Algorithm 2 Implementation of the MLPIIC Method

Initialization:

Define the iteration index $i = 0$.
Select the initial admissible control law $\mu_0(\cdot)$.
Select a computation precision ε.
Select a finite positive definite function $\Phi(\cdot)$ that satisfies (13) and (20).
Select an array of p state vectors randomly in Ω, $\mathfrak{X} = \{x_k^1, x_k^2, \ldots, x_k^p\}$.

Iteration:

1: Let $i = i + 1$. Calculate the target of $\hat{V}_i(\cdot)$ by (55) and obtain $\{\hat{V}_{i,\text{target}}^\mu(x_k^1), \hat{V}_{i,\text{target}}^\mu(x_k^2), \ldots, \hat{V}_{i,\text{target}}^\mu(x_k^p)\}$. Obtain the least squares solution of W_i^V.
2: Calculate the target of $\hat{\mu}_i(\cdot)$ by (56) and obtain $\{\hat{\mu}_i^{\text{target}}(x_k^1), \hat{\mu}_i^{\text{target}}(x_k^2), \ldots, \hat{\mu}_i^{\text{target}}(x_k^p)\}$. Obtain the least squares solution of W_i^μ.
3: If

$$|\hat{V}_i^\mu(x_k) - \hat{V}_{i-1}^\mu(x_k)| \leq \varepsilon \quad \forall x_k \in \mathfrak{X},$$

goto next step. Otherwise, goto Step 1.
4: **return** $\hat{V}^*(\cdot) = \hat{V}_i^\mu(\cdot)$ and $\hat{\mu}^*(\cdot) = \hat{\mu}_i(\cdot)$.

V. Simulation Studies

In order to demonstrate the effectiveness of the MLPIIC method, Van der Pol's oscillator [68] is selected as the simulation model. The continuous-time dynamics of Van der Pol's oscillator is

$$\ddot{y} = \left(1 - y^2\right)\dot{y} - y + u.$$

By defining the state vector $x = [y, \dot{y}]^\top$ and discretizing the continuous-time dynamics through the forward Euler method with the sampling time $\Delta t = 0.05$ s, a discrete-time nonlinear model is established as

$$x_{k+1} = x_k + \Delta t f(x_k, u_k)$$

where $x_k = [x_{1k}, x_{2k}]^\top$ is the system state, and

$$f(x_k, u_k) = \begin{bmatrix} x_{2k} \\ (1 - x_{1k}^2)x_{2k} - x_{1k} + u_k \end{bmatrix}.$$

The initial state is $x_0 = [-0.7, 1.437]^\top$.

Let the system state be denoted as $x_k = [x_{1k}, x_{2k}]^\top$. In this simulation, the goal is to maintain the difference between x_{1k} and x_{2k} within a given bound while regulating x_{1k} and x_{2k}. In practical applications, the utility function, constraint utility function, and d_0 are given by experts. In this article, for simulation purposes, the utility function and the constraint utility function are set as $U(x_k, u_k) = 0.25x_k^\top x_k + 0.05u_k^\top u_k$ and $d(x_k, u_k) = 4x_{1k}^2 + 4x_{2k}^2 - 4x_{1k}x_{2k} + 0.05u_k^\top u_k$, respectively. The upper bound d_0 is set as $d_0 = 76$. Note that the constraint utility function indicates the intention to guarantee that x_{1k} and x_{2k} are close.

The implementation details of the MLPIIC method are as follows. Let

$$\Omega = \{x_k : \ -1.5 \le x_{1k} \le 1.5, -1.5 \le x_{2k} \le 1.5\}$$

and sample randomly 500 states from Ω. The auxiliary function $\Phi(\cdot)$ is determined first. The basis functions used to parameterize $\Phi(\cdot)$ are set as

$$\phi(x_k) = \begin{bmatrix} x_{1k}^2 & x_{1k}x_{2k} & x_{2k}^2 & x_{1k}^4 & x_{1k}^3x_{2k} \\ x_{1k}^2x_{2k}^2 & x_{1k}x_{2k}^3 & x_{2k}^4 & x_{1k}^6 & x_{1k}^5x_{2k} \\ x_{1k}^4x_{2k}^2 & x_{1k}^3x_{2k}^3 & x_{1k}^2x_{2k}^4 & x_{1k}x_{2k}^5 & x_{2k}^6 \end{bmatrix}^\top.$$

The weight of $\Phi(\cdot)$ obtained by solving (54) is

$$W_\Phi = [137.10 \quad 23.19 \quad 16.16 \quad -0.69 \quad -0.20 \\ -1.28 \quad -0.11 \quad -0.10 \quad 0.14 \quad -0.01 \\ 0.11 \quad 0 \quad 0 \quad 0.01 \quad 0.03]^\top.$$

From

$$\Phi(x_0) = W_\Phi^\top \phi(x_0) = 75.9775 < d_0$$

we know $\Phi(x_0)$ is close to d_0. The number of steps to look-ahead is set as 2, i.e., $l = 2$. The positive number used to compute the target of $\hat{V}_i^\mu(\cdot)$ in (55) is set as 500, i.e., $K = 500$. In order to approximate the IVF and the ICL by function approximators, both $\varphi(x_k)$ and $\psi(x_k)$ are chosen to be the same as $\phi(x_k)$. An initial control law is achieved via feedback linearization as $\mu_0(x_k) = (1 - x_{1k}^2)x_{2k} - 6x_{1k} - 15x_{2k}$. Note that $\mu_0(\cdot)$ is admissible but not feasible because $D_{\mu_0}(x_0) = 83.65 > d_0$.

Implement the MLPIIC method to achieve the computation precision $\varepsilon = 0.005$. The convergence trajectory of the IVF at x_0, i.e., $\hat{V}_i^\mu(x_0)$, is shown in Fig. 1. In addition, Fig. 2 demonstrates the convergence trajectory of the constraint function of the ICL at x_0, i.e., $D_{\hat{\mu}_i}(x_0)$. The state and control trajectories achieved through applying the final control law to Van der Pol's oscillator system are displayed in Fig. 3(a)–(c), respectively.

To show the effectiveness of the developed MLPIIC method, comparisons with the traditional PI algorithm and the regularization-based PI algorithm are presented. In the regularization-based PI algorithm, the optimal control law with isoperimetric constraints is solved approximately by solving

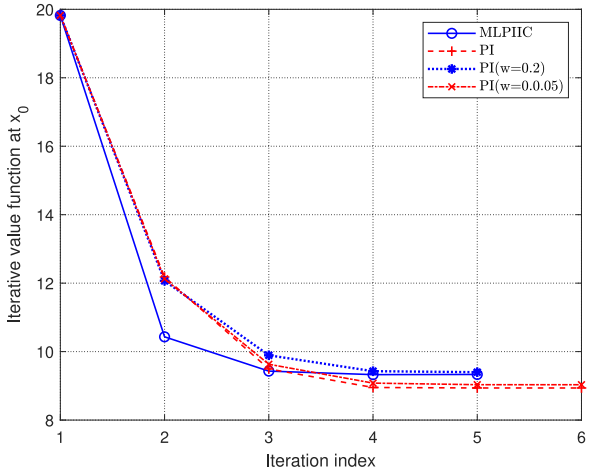

Fig. 1. Convergence trajectory of $\hat{V}_i^\mu(x_0)$.

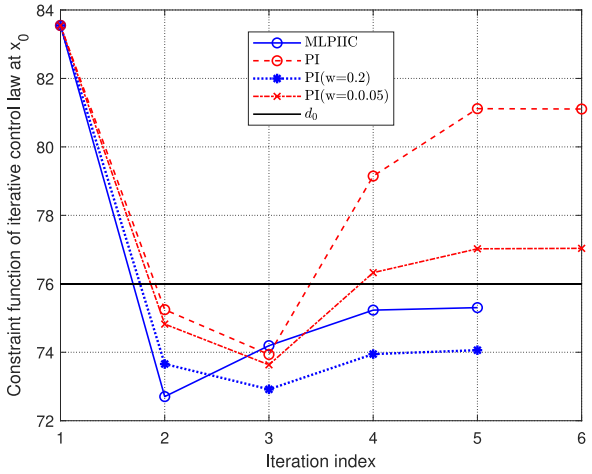

Fig. 2. Convergence trajectory of $D_{\hat{\mu}_i}(x_0)$.

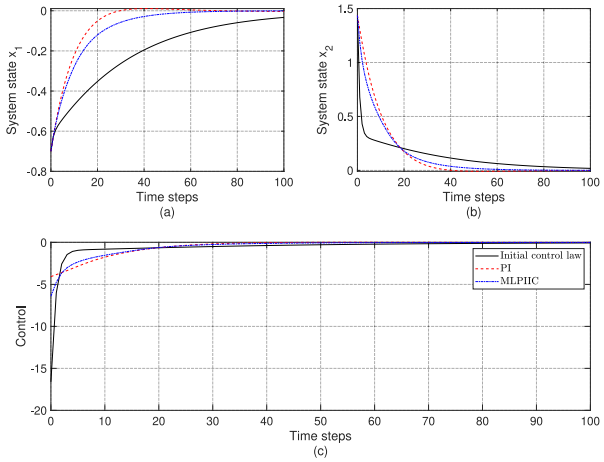

Fig. 3. (a) Trajectory of state x_1. (b) Trajectory of state x_2. (c) Trajectory of control action.

the following regularized OCP using the PI algorithm:

$$\min_{\mu(\cdot)} \ \{J_\mu(x_0) + w(D_\mu(x_0) - d_0)\}$$
$$\text{s.t.} \quad x_{k+1} = F(x_k, \mu(x_k)) \quad \forall x_k, k \in \mathbb{N}$$

where $w \geq 0$ is the regularization parameter. The initial control laws utilized in the comparative experiment and the MLPIIC method are identical. The convergence trajectories of the regularization-based PI algorithm are demonstrated in Figs. 1 and 2, respectively. Applying the final control law in the traditional PI algorithm to the given system, we can obtain the state and control trajectories, which are displayed in Fig. 3(a)–(c), respectively.

The simulation results of the MLPIIC method are analyzed as follows. For the MLPIIC method, the constraint function of the ICL is always less than d_0 when the iteration index is greater than 1 from Fig. 2, which shows that the ICL for $i \in \mathbb{N}_+$ is feasible although $\mu_0(\cdot)$ is not feasible. Furthermore, it is displayed that the IVFs at x_0 are nonincreasing and convergent when the iteration index is greater than 1 from Fig. 1. For the traditional PI algorithm and the regularization-based PI algorithm with $w = 0.05$, Fig. 1 shows that they receive smaller performance index functions than the present MLPIIC method. However, Fig. 2 shows that the constraint functions of the ICL obtained by the traditional PI algorithm and regularization-based PI algorithm with $w = 0.05$ are obviously larger than d_0. Thus, simulation results show that the optimal control laws achieved through these two algorithms are not feasible. By the regularization-based PI algorithm with $w = 0.2$, Figs. 1 and 2 show that a feasible control law whose performance index function is close to that of the MLPIIC method is obtained. However, the regularization-based PI algorithm is time consuming since an appropriate regularization parameter w can only be obtained by trial and error. Through comparative experiments, the convergence, feasibility, and effectiveness of the MLPIIC method are verified.

VI. Conclusion

In this article, a novel MLPIIC method is developed to solve infinite horizon OCPs with isoperimetric constraints. In order to overcome the difficulty that the value function of the OCP with isoperimetric constraints does not directly satisfy Bellman's principle of optimality, by constructing an auxiliary function, the OCP with isoperimetric constraints is approximated as a special OCP, where Bellman's principle of optimality holds. The conditions that the auxiliary function should satisfy are analyzed. Initialized with an admissible control law, the MLPIIC method iterates between the policy evaluation equation and the multistep look-ahead policy update equation. The IVF is proven to converge to the optimal performance index function of the approximated OCP. The feasibility of the ICL is demonstrated. The implementation of the MLPIIC method based on function approximators is described. Numerical results demonstrate the effectiveness of the proposed method.

Our future work is to extend the MLPIIC method to multiple isoperimetric constraints settings and stochastic systems. Furthermore, we will also investigate the application of the MLPIIC method to safety-critical systems.

References

[1] Q. Wei, H. Li, T. Li, and F.-Y. Wang, "A novel data-based fault-tolerant control method for multicontroller linear systems via distributed policy iteration," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 5, pp. 3176–3186, May 2023.

[2] J. Qiu, M. Ma, and T. Wang, "Event-triggered adaptive fuzzy fault-tolerant control for stochastic nonlinear systems via command filtering," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 2, pp. 1145–1155, Feb. 2022.

[3] Q. Yang, W. Cao, W. Meng, and J. Si, "Reinforcement-learning-based tracking control of waste water treatment process under realistic system conditions and control performance requirements," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 8, pp. 5284–5294, Aug. 2022.

[4] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1957.

[5] P. J. Werbos, "Advanced forecasting methods for global crisis warning and models of intelligence," *Gen. Syst. Yearbook*, vol. 23, pp. 25–38, Jan. 1977.

[6] P. J. Werbos, "A menu of designs for reinforcement learning over time," in *Neural Networks Control*, W. T. Miller, R. S. Sutton, and P. J. Werbos, Eds. Cambridge, MA, USA: MIT Press, 1991, pp. 67–95.

[7] D. V. Prokhorov and D. C. Wunsch, "Adaptive critic designs," *IEEE Trans. Neural Netw.*, vol. 8, no. 5, pp. 997–1007, Sep. 1997.

[8] D. V. Prokhorov, R. A. Santiago, and D. C. Wunsch, "Adaptive critic designs: A case study for neurocontrol," *Neural Netw.*, vol. 8, no. 9, pp. 1367–1372, 1995.

[9] M. Wang, K. Wang, L. Huang, and H. Shi, "Observer-based event-triggered tracking control for discrete-time nonlinear systems using adaptive critic design," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 9, pp. 5393–5403, Sep. 2023, doi: 10.1109/TSMC.2023.3269108.

[10] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, "Adaptive dynamic programming," *IEEE Trans. Syst. Man, Cybern. C, Cybern*, vol. 32, no. 2, pp. 140–153, May 2002.

[11] F.-Y. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: An introduction," *IEEE Comput. Intell. Mag.*, vol. 4, no. 2, pp. 39–47, May 2009.

[12] Q. Wei, L. Zhu, T. Li, and D. Liu, "A new approach to finite-horizon optimal control for discrete-time affine nonlinear systems via a pseudo linear method," *IEEE Trans. Autom. Control*, vol. 67, no. 5, pp. 2610–2617, May 2022.

[13] Q. Wei, L. Han, and T. Zhang, "Spiking adaptive dynamic programming based on poisson process for discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 5, pp. 1846–1856, May 2022.

[14] Z. Ming, H. Zhang, Y. Yan, and J. Sun, "Self-triggered adaptive dynamic programming for model-free nonlinear systems via generalized fuzzy hyperbolic model," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 5, pp. 2792–2801, May 2023.

[15] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ, USA: Wiley, 2013.

[16] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 943–949, Aug. 2008.

[17] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Hoboken, NJ, USA: Wiley, 2007.

[18] A. Forootani, R. Iervolino, M. Tipaldi, and S. Dey, "Transmission scheduling for multi-process multi-sensor remote estimation via approximate dynamic programming," *Automatica*, vol. 136, pp. 1–14, Feb. 2022.

[19] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA, USA: Athena Sci., 1996.

[20] D. P. Bertsekas, M. L. Homer, D. A. Logan, S. D. Patek, and N. R. Sandell, "Missile defense and interceptor allocation by neuro-dynamic programming," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 30, no. 1, pp. 42–51, Jan. 2000.

[21] S. Chakraborty and M. G. Simões, "Neural dynamic programming based online controller with a novel trim approach," in *Proc. IEE Control Theory Appl.*, 2005, pp. 95–104.

[22] C. Mu, D. Wang, and H. He, "Novel iterative neural dynamic programming for data-based approximate optimal control design," *Automatica*, vol. 81, pp. 240–252, Jul. 2017.
[23] B. Lincoln and A. Rantzer, "Relaxing dynamic programming," *IEEE Trans. Autom. Control*, vol. 51, no. 8, pp. 1249–1260, Aug. 2006.
[24] A. Rantzer, "Relaxed dynamic programming in switching systems," in *Proc. IEE Proc. Control Theory Appl.*, 2006, pp. 567–574.
[25] G. Wen and B. Li, "Optimized leader–follower consensus control using reinforcement learning for a class of second-order nonlinear multiagent systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 9, pp. 5546–5555, Dec. 2022.
[26] H. Hassani, R. Razavi-Far, M. Saif, and E. Herrera-Viedma, "Reinforcement learning-based feedback and weight-adjustment mechanisms for consensus reaching in group decision making," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 4, pp. 2456–2468, Apr. 2023.
[27] C. Li, Q. Liu, Z. Zhou, M. Buss, and F. Liu, "Off-policy risk-sensitive reinforcement learning-based constrained robust optimal control," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 4, pp. 2478–2491, Apr. 2023.
[28] L. Yuan, T. Li, S. Tong, Y. Xiao, and Q. Shan, "Broad learning system approximation-based adaptive optimal control for unknown discrete-time nonlinear systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 8, pp. 5028–5038, Aug. 2022.
[29] Y. Zhang, B. Zhao, D. Liu, and S. Zhang, "Event-triggered control of discrete-time zero-sum games via deterministic policy gradient adaptive dynamic programming," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 8, pp. 4823–4835, Aug. 2022.
[30] H. Li and Q. Wei, "Optimal synchronization control for multi-agent systems with input saturation: A nonzero-sum game," *Front. Inf. Technol. Electron. Eng.*, vol. 23, no. 7, pp. 1010–1019, May 2022.
[31] Q. Wei, L. Wang, J. Lu, and F.-Y. Wang, "Discrete-time self-learning parallel control," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 1, pp. 192–204, Jan. 2022.
[32] Q. Wei, H. Li, and F.-Y. Wang, "A novel parallel control method for continuous-time linear output regulation with disturbances," *IEEE Trans. Cybern.*, vol. 53, no. 6, pp. 3760–3770, Jun. 2023.
[33] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 3, pp. 621–634, Sep. 2013.
[34] D. Liu, Q. Wei, and P. Yan, "Generalized policy iteration adaptive dynamic programming for discrete-time nonlinear systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 12, pp. 1577–1591, Dec. 2015.
[35] Q. Wei, D. Liu, Q. Lin, and R. Song, "Discrete-time optimal control via local policy iteration adaptive dynamic programming," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3367–3379, Oct. 2016.
[36] A. Heydari, "Convergence analysis of policy iteration," May 2015, *arXiv:1505.05216*.
[37] D. P. Bertsekas, "Feature-based aggregation and deep reinforcement learning: A survey and some new implementations," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 1, pp. 1–31, Jan. 2018.
[38] Y. Efroni, G. Dalal, B. Scherrer, and S. Mannor, "Beyond the one-step greedy approach in reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1387–1396.
[39] Y. Efroni, G. Dalal, B. Scherrer, and S. Mannor, "Multiple-step greedy policies in approximate and online reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5244–5253.
[40] D. Wang and D. Liu, "Learning and guaranteed cost control with event-based adaptive critic implementation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6004–6014, Apr. 2018.
[41] D. Liu, D. Wang, F.-Y. Wang, H. Li, and X. Yang, "Neural-network-based online HJB solution for optimal robust guaranteed cost control of continuous-time uncertain nonlinear systems," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2834–2847, Dec. 2014.
[42] C. Mu and D. Wang, "Neural-network-based adaptive guaranteed cost control of nonlinear dynamical systems with matched uncertainties," *Neurocomputing*, vol. 245, pp. 46–54, Jul. 2017.
[43] J. Banas and A. Vacroux, "On linear systems with a time-varying delay and isoperimetric constraints," *IEEE Trans. Autom. Control*, vol. AC-13, no. 4, pp. 439–440, Aug. 1968.
[44] E. Lee, "Linear optimal control problems with isoperimetric constraints," *IEEE Trans. Autom. Control*, vol. AC-12, no. 1, pp. 87–90, Feb. 1967.
[45] A. Lim, Y. Liu, K. Teo, and J. Moore, "Linear-quadratic optimal control with integral quadratic constraints," *Optim. Control Appl. Methods*, vol. 20, no. 2, pp. 79–92, Apr. 1999.
[46] W. Schmitendorf, "Pontryagin's principle for problems with isoperimetric constraints and for problems with inequality terminal constraints," *J. Optim. Theory Appl.*, vol. 18, no. 4, pp. 561–567, Aug. 1976.
[47] A. Kumar and A. Vladimirsky, "An efficient method for multiobjective optimal control and optimal control subject to integral constraints," *J. Comput. Math.*, vol. 28, no. 4, pp. 517–551, Apr. 2010.
[48] S. Zouhri, M. E. Baroudi, and S. Saadi, "Optimal control with isoperimetric constraint for chemotherapy of Tumors," *Int. J. Appl. Comput. Math.*, vol. 8, no. 4, pp. 1–14, Apr. 2022.
[49] Q. Wei and T. Li, "Constrained-cost adaptive dynamic programming for optimal control of discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 2, 2023, doi: 10.1109/TNNLS.2023.3237586.
[50] J. Sun, "Linear quadratic optimal control problems with fixed terminal states and integral quadratic constraints," *Appl. Math. Optim.*, vol. 83, no. 1, pp. 251–276, Oct. 2021.
[51] E. Altman, *Constrained Markov Decision Processes*. Hoboken, NJ, USA: CRC Press, 1999.
[52] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, "A Lyapunov-based approach to safe reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8092–8101.
[53] E. A. Feinberg, A. Jaśkiewicz, and A. S. Nowak, "Constrained discounted Markov decision processes with Borel state spaces," *Automatica*, vol. 111, Jan. 2020, Art. no. 108582.
[54] P. Soravia, "Viscosity solutions and optimal control problems with integral constraints," *Syst. Control Lett.*, vol. 40, no. 5, pp. 325–335, Aug. 2000.
[55] R. Bhatia, *Matrix Analysis*. New York, NY, USA: Springer, 2013.
[56] D. P. Bertsekas, "Value and policy iterations in optimal control and adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 500–509, Mar. 2015.
[57] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Belmont, MA, USA: Athena Sci., 2012.
[58] R. Kalman and J. Bertram, "Control system analysis and design via the second method of Lyapunov:(I) continuous-time systems (II) discrete time systems," *IRE Trans. Autom. Control*, vol. 4, no. 3, pp. 112–112, Dec. 1959.
[59] D. Liu, H. Li, and D. Wang, "Error bounds of adaptive dynamic programming algorithms for solving undiscounted optimal control problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1323–1334, Jun. 2015.
[60] Y. Zhang, B. Decardi-Nelson, J. Liu, J. Shen, and J. Liu, "Zone economic model predictive control of a coal-fired boiler–turbine generating system," *Chem. Eng. Res. Des.*, vol. 153, pp. 246–256, Jan. 2020.
[61] Q. Wei, J. Lu, T. Zhou, X. Cheng, and F.-Y. Wang, "Event-triggered near-optimal control of discrete-time constrained nonlinear systems with application to a boiler–turbine system," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 3926–3935, Jun. 2022.
[62] A. Tsiamis, D. S. Kalogerias, L. F. Chamon, A. Ribeiro, and G. J. Pappas, "Risk-constrained linear-quadratic regulators," in *Proc. 59th IEEE Conf. Decis. Control (CDC)*, Dec. 2020, pp. 3040–3047.
[63] F. Girosi and T. Poggio, "Networks and the best approximation property," *Biol. Cybern.*, vol. 63, no. 3, pp. 169–176, Aug. 1990.
[64] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[65] W. Banzhaf, P. Nordin, R. E. Keller, and F. D. Francone, *Genetic Programming: An Introduction*. San Francisco, CA, USA: Morgan Kaufmann, 1998.
[66] M. Hatami, *Weighted Residual Methods: Principles, Modifications and Applications*. New York, NY, USA: Academic, 2017.
[67] B. Jakubczyk, "Feedback linearization of discrete-time systems," *Syst. Control. Lett.*, vol. 9, no. 5, pp. 411–416, Nov. 1987.
[68] R. FitzHugh, "Impulses and physiological states in theoretical models of nerve membrane," *Biophys. J.*, vol. 1, no. 6, pp. 445–466, Jul. 1961.

Tao Li received the bachelor's degree in automation from Northeastern University, Shenyang, China, in 2019. He is currently pursuing the Ph.D. degree in control theory and control engineering with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, and the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing.

His current research interests include adaptive dynamic programming, reinforcement learning, optimal control, and neural network-based control.

Qinglai Wei (Senior Member, IEEE) received the B.S. degree in automation and the Ph.D. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2002 and 2009, respectively.

From 2009 to 2011, he was a Postdoctoral Fellow with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently a Professor with the Institute of Automation, Chinese Academy of Sciences, where he is an Associate Director of the State Key Laboratory of Management and Control for Complex Systems. He has authored four books and published over 80 international journal articles. His research interests include adaptive dynamic programming, neural networks-based control, optimal control, nonlinear systems, and their industrial applications.

Prof. Wei was a recipient of the IEEE/CAA JOURNAL OF AUTOMATICA SINICA Best Paper Award, the IEEE System, Man, and Cybernetics Society Andrew P. Sage Best Transactions Paper Award, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS Outstanding Paper Award, the Outstanding Paper Award of *Acta Automatica Sinica*, and the Zhang Siying Outstanding Paper Award of the Chinese Control and Decision Conference. He was also a recipient of the Shuang-Chuang Talents in Jiangsu Province, China, and the Young Researcher Award of Asia–Pacific Neural Network Society. He has been the Vice President of the IEEE Computational Intelligence Society's Beijing Chapter since 2022. He was the General Co-Chair of the 2019 Symposium of Parallel Intelligence, the Invited Session Chair of the IEEE 8th Data Driven Control and Learning Systems Conference, the Special Sessions Chair of the 25th International Conference on Neural Information Processing, the Program Co-Chair of the 24th International Conference on Neural Information Processing, and the Registration Chair of the 12th World Congress on Intelligent Control and Automation 2016 and the 2014 IEEE World Congress on Computational Intelligence 2014. He was a guest editor for several international journals. He is the Associate Editor-in-Chief/the Deputy Editor-in-Chief of the IEEE/CAA JOURNAL OF AUTOMATICA SINICA, *Neurocomputing*, and *Acta Automatica Sinica*. He is also an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON CONSUMER ELECTRONICS, IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, and *Optimal Control Applications and Methods*.

Fei-Yue Wang received the Ph.D. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990.

He joined The University of Arizona, Tucson, AZ, USA, in 1990 and became a Professor and the Director of the Robotics and Automation Laboratory and the Program in Advanced Research for Complex Systems. In 1999, he founded the Intelligent Control and Systems Engineering Center, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Chinese Talents Program from the State Planning Council, and in 2002, he was appointed as the Director of the Key Laboratory of Complex Systems and Intelligence Science, CAS. In 2011, he became the State Specially Appointed Expert and the Director of the State Key Laboratory for Management and Control of Complex Systems. His current research focuses on methods and applications for parallel intelligence, social computing, and knowledge automation.

Prof. Wang received the National Prize in Natural Sciences of China and became an Outstanding Scientist of ACM for his work in intelligent control and social computing in 2007, the IEEE ITS Outstanding Application and Research Awards in 2009 and 2011, respectively, and the IEEE Norbert Wiener Award in 2014. Since 1997, he has been serving as the General or Program Chair of over 30 IEEE, INFORMS, IFAC, ACM, and ASME conferences. He was the President of the IEEE ITS Society from 2005 to 2007, the Chinese Association for Science and Technology, USA, in 2005, and the American Zhu Kezhen Education Foundation from 2007 to 2008, the Vice President of the ACM China Council from 2010 to 2011, and the Vice President and the Secretary General of the Chinese Association of Automation from 2008 to 2018. He was the Founding Editor-in-Chief (EiC) of the *International Journal of Intelligent Control and Systems* from 1995 to 2000, the *IEEE Intelligent Transportation Systems Magazine* from 2006 to 2007, the IEEE/CAA JOURNAL OF AUTOMATICA SINICA from 2014 to 2017, and the *China's Journal of Command and Control* from 2015 to 2020. He was the EiC of the IEEE INTELLIGENT SYSTEMS from 2009 to 2012, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS from 2009 to 2016, and the IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS from 2017 to 2020, and has been the Founding EiC of *Chinese Journal of Intelligent Science and Technology* since 2019 and the EiC of the IEEE TRANSACTIONS ON INTELLIGENT VEHICLES since 2022. He is currently the President of CAA's Supervision Council, IEEE Council on RFID, and the Vice President of the IEEE Systems, Man, and Cybernetics Society. He is a Fellow of INCOSE, IFAC, ASME, and AAAS.