Les Cahiers du GERAD ISSN: 0711–2440

Policy-consistent inverse optimal control for discrete-time nonlinear stochastic systems

Z. Wang, H. Zhang, A. Ringh

G–2026–80

November 2026

La collection *Les Cahiers du GERAD* est constituée des travaux de recherche menés par nos membres. La plupart de ces documents de travail a été soumis à des revues avec comité de révision. Lorsqu'un document est accepté et publié, le pdf original est retiré si c'est nécessaire et un lien vers l'article publié est ajouté.

Citation suggérée : Z. Wang, H. Zhang, A. Ringh (Novembre 2026). Policy-consistent inverse optimal control for discrete-time nonlinear stochastic systems, Rapport technique, Les Cahiers du GERAD G–2026–80, GERAD, HEC Montréal, Canada.

Avant de citer ce rapport technique, veuillez visiter notre site Web (https://www.gerad.ca/fr/papers/G-2026-80) afin de mettre à jour vos données de référence, s'il a été publié dans une revue scientifique.

The series *Les Cahiers du GERAD* consists of working papers carried out by our members. Most of these pre-prints have been submitted to peer-reviewed journals. When accepted and published, if necessary, the original pdf is removed and a link to the published article is added.

Suggested citation: Z. Wang, H. Zhang, A. Ringh (November 2026). Policy-consistent inverse optimal control for discrete-time nonlinear stochastic systems, Technical report, Les Cahiers du GERAD G–2026–80, GERAD, HEC Montréal, Canada.

Before citing this technical report, please visit our website (https://www.gerad.ca/en/papers/G-2026-80) to update your reference data, if it has been published in a scientific journal.

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2026 – Bibliothèque et Archives Canada, 2026

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2026 – Library and Archives Canada, 2026

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Policy-consistent inverse optimal control for discrete-time nonlinear stochastic systems

Ziliang Wang [a, b]

Han Zhang [a, b, c]

Axel Ringh [d]

[a] *School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China*

[b] *Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China*

[c] *GERAD, Montréal (Québec), Canada*

[d] *Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden*

qingyingyese@sjtu.edu.cn
zhanghan_tc@sjtu.edu.cn
axelri@chalmers.se

November 2026
Les Cahiers du GERAD
G–2026–80

Copyright © 2026 GERAD, Wang, Zhang, Ringh

Les textes publiés dans la série des rapports de recherche *Les Cahiers du GERAD* n'engagent que la responsabilité de leurs auteurs. Les auteurs conservent leur droit d'auteur et leurs droits moraux sur leurs publications et les utilisateurs s'engagent à reconnaître et respecter les exigences légales associées à ces droits. Ainsi, les utilisateurs:

- Peuvent télécharger et imprimer une copie de toute publication du portail public aux fins d'étude ou de recherche privée;
- Ne peuvent pas distribuer le matériel ou l'utiliser pour une activité à but lucratif ou pour un gain commercial;
- Peuvent distribuer gratuitement l'URL identifiant la publication.

Si vous pensez que ce document enfreint le droit d'auteur, contactez-nous en fournissant des détails. Nous supprimerons immédiatement l'accès au travail et enquêterons sur votre demande.

The authors are exclusively responsible for the content of their research papers published in the series *Les Cahiers du GERAD*. Copyright and moral rights for the publications are retained by the authors and the users must commit themselves to recognize and abide the legal requirements associated with these rights. Thus, users:

- May download and print one copy of any publication from the public portal for the purpose of private study or research;
- May not further distribute the material or use it for any profit-making activity or commercial gain;
- May freely distribute the URL identifying the publication.

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Abstract : Inverse Optimal Control (IOC) seeks to recover an unknown cost from expert demonstrations, and it provides a systematic way of modeling experts' decision mechanisms while considering the prior information of the dynamics and the cost functions. Nevertheless, existing IOC methods lack consistency guarantees under noisy and nonlinear settings, and hence cannot guarantee policy-consistent recovery. In this paper, we consider a discrete-time nonlinear stochastic system, and it is controlled by an optimal policy that minimizes the expectation of a discounted cumulative cost function across an infinite time-horizon. In particular, the cost function takes the form of a linear combination of a priori known feature functions. In this setting, we first adopt Lasserre's reformulation [12] of the forward problem with occupancy measure. Next, we propose the infinite dimensional IOC algorithm and further approximate it with Lagrange interpolating polynomials, which results in a convex, finite-dimensional sum-of-squares optimization. Moreover, the estimator is shown to be asymptotically and statistically consistent in optimality with respect to the demonstrated expert trajectories. The subsequent policy estimator is also shown to be asymptotically and statistically consistent in policy recovery. Finally, we validate the theoretical results and illustrate the performance of our method with numerical experiments. In addition, the robustness and generalizability performance of the proposed IOC algorithm are also illustrated.

Keywords: Inverse optimal control; nonlinear system; stochastic system; system identification; sum-of-squares optimization

Résumé : La commande optimale inverse (IOC) cherche à retrouver une fonction de coût à partir de démonstrations d'experts, en exploitant les connaissances a priori sur la dynamique et le coût. Nous considérons des systèmes non linéaires stochastiques en temps discret, avec un coût cumulé actualisé sur un horizon infini. Le coût instantané est une combinaison linéaire de fonctions caractéristiques connues. En nous appuyant sur la reformulation de Lasserre du problème direct par des mesures d'occupation, nous construisons un algorithme d'IOC en dimension infinie. Une approximation par des polynômes d'interpolation de Lagrange conduit à un problème convexe d'optimisation par sommes de carrés en dimension finie. Nous établissons la convergence asymptotique et statistique de l'estimateur au sens de l'optimalité par rapport aux trajectoires de l'expert, ainsi que celle de l'estimateur de politique pour la reconstruction de la politique. Des expériences numériques valident les résultats théoriques et illustrent les performances, la robustesse et la capacité de généralisation de la méthode.

Mots clés : Commande optimale inverse; système non linéaire; système stochastique; identification de systèmes; optimisation par sommes de carrés

Acknowledgements: This work was supported by National Natural Science Foundation (NNSF) of China under Grant 62573286, by Natural Science Foundation of Shanghai under Grant 25ZR1401208, by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the Swedish Research Council (VR) under grant 2024–05776.

1 Introduction

Optimal Control (OC) provides a principled framework for designing a control law by specifying it to be the minimizer of a specified performance index. However, the performance of such a controller in the contextual environment critically depends on the choice of the cost function, which in general trades off between the control effort and the feedback system's performance. In practice, accurate specification of such a cost function is often nontrivial; it requires expertise and may not reflect the true objectives of the expert's decisions. Inverse optimal control (IOC), also known as Inverse Reinforcement Learning (IRL), addresses this issue by inferring the underlying cost function of an expert from demonstration data. Therefore, it not only provides valuable insights for controller design, but also further enables the reconstruction of the corresponding closed-loop policy [4, 25].

Compared with Imitation Learning (IL), which directly imitates the observed expert behavior [43], IOC assumes that the expert's cost function follows a known structure, such as a linear combination of predefined feature functions [9, 49, 51]. This assumption restricts the optimal policy's possible structure and provides stronger robustness and generalizability. Motivated by this, IOC methods have been successfully applied to modeling the decision criteria in various domains such as autonomous driving [21], human sensorimotor [36], and animal behaviours [2]. However, for nonlinear stochastic systems, the corresponding IOC problems are challenging due to the absence of explicit expressions for the value functions. This leads to "coupled-grey-black-box" identification problems, compared to the grey-box case in Linear Quadratic (LQ) settings [11, 18, 24, 30, 42, 44–49]. Such a problem still remains even when the cost function structure is known a priori. Moreover, the trajectory observations are usually contaminated by noise in practice. In particular, the noise could be either the observation noise introduced by the sensor measurements, or the process noise that is intrinsically modelled by the Markov transition kernel of the nonlinear stochastic system. Therefore, a key property that an IOC estimator should have to guarantee the robustness is policy consistency. In this paper, policy consistency means that the recovered cost function induces a policy that converges to the demonstrated expert policy, rather than necessarily identifying the unique underlying cost parameters. Nevertheless, if the IOC estimator is an "M-estimator" [40, Chap. 5], consistency is typically only guaranteed at the globally optimal solution of the underlying optimization problem. Therefore, it is also desirable that this optimization is convex, as to avoid issues with local minima and thus achieve consistency in practice.

1.1 Related works

One way to avoid the aforementioned "coupled-grey-black-box" issue is to identify a cost function that satisfies the first-order necessary optimality conditions given by Pontryagin's Maximum Principle (PMP) [20]. However, such algorithms only recover a cost function to which the observed trajectories are only locally optimal, and hence we might not be able to guarantee the feedback system's performance when using the control policy recovered from the estimated cost function. Another straightforward method is to minimize the discrepancy between the expert's policy and the optimal policy induced by the candidate cost functions [1, 13, 51]. Such methods often require repeatedly solving the entire or part of the forward optimal control problem, which makes them computationally demanding and less scalable to complex systems. Nevertheless, by exploiting the specific structures of the system dynamics or the cost function, several studies have avoided the issue of solving the forward problem repeatedly. In particular, bilinear [6], affine [14, 19, 31], and polynomial dynamical systems [27] have been investigated, where certain problem structures enable tractable reformulations. Yet more general nonlinear settings are not considered therein. Furthermore, another line of research incorporates statistical learning principles. Inspired by maximum-entropy regularization, a family of methods recovers the cost function by maximum-likelihood estimation under stochastic expert policies [5, 9, 10, 23, 51]. In addition, [50, 51] discretize the state and control spaces, while others represent the stochastic policies with exponential-family distributions [5, 9, 23] or deep neural networks [10].

Nevertheless, the impact of these approximations on policy recovery accuracy remains theoretically unquantified.

1.2 The contributions

To address the above challenges, we propose a policy-consistent IOC algorithm for discrete-time infinite-horizon nonlinear stochastic systems with a discounted-cost, where the running cost is assumed to be a linear combination of known continuous feature functions. In particular, the method recovers a cost function by solving a convex and well-studied Sum-Of-Squares (SOS) program, and the recovered cost function is asymptotically and statistically consistent in optimality with respect to the demonstrated expert trajectories. In addition, the subsequent policy estimator is also shown to be asymptotically and statistically consistent in policy recovery. Compared to existing approaches, our method (i) handles the nonlinear and stochastic dynamics without restrictive structural assumptions; (ii) avoids solving the forward optimal control problem repeatedly; and (iii) provides theoretical guarantees of asymptotic and statistical consistency in cost function's optimality with respect to the demonstrations as well as in policy recovery, which have not been established simultaneously in existing works. More precisely, our key contributions are as follows:

1. We reformulate the nonlinear, infinite-horizon IOC problem as an infinite dimensional linear programming problem, which minimizes the violation of optimality conditions for the observed trajectories. We further show that samples of finite-time observation suffice to recover a cost function consistent with the demonstrated optimality conditions. By applying a polynomial approximation, we obtain a finite dimensional SOS program that can be solved numerically. Such a method also exempts us from local minima issues.
2. We establish asymptotic and statistical policy consistency of the proposed estimator. Namely, as the polynomial approximation order and the number of demonstration trajectories increase, the optimality gap between the expert policy and the estimated optimal policy vanishes, and the estimated parameters approach the true ones whenever the solution is unique. Moreover, the underlying optimal control policy can be further consistently recovered, e.g., by integrating the estimated optimality condition into "behaviour cloning".
3. Numerical experiments are conducted to validate the theoretical results of the proposed method, demonstrate its estimation accuracy, and evaluate its robustness and generalizability.

The remainder of this paper is organized as follows. Sec. 2 introduces the problem setting: it formulates the forward problem for the discrete-time nonlinear stochastic systems, it formulates the IOC problem, and it states several mild assumptions on the considered system. Next, we present the proposed approach in Sec. 3. In particular, the unknown functions in the IOC formulation are approximated by polynomials, and the problem is converted into a finite-dimensional SOS program — this is the finite-dimensional estimator. Moreover, the theoretical guarantees, including asymptotic and statistical consistency in cost function's optimality with respect to the demonstrations as well as in policy recovery, are established in Sec. 4. In addition, we briefly describe the SOS optimization procedure used in practice in Sec. 5. In Sec. 6, we evaluate the performance of the estimator on three benchmark systems: a LQ regulator, a temperature-control process, and an inverted pendulum. In the inverted pendulum problem, we also compare the controllers obtained via IOC with "behaviour cloning" (see, e.g., [3, 34, 39]) and show that IOC gives controllers that are more robust to distributional shifts and errors in the modelled dynamics. Finally, the paper is concluded in Sec. 7 with some discussion and outlooks.

Notation: $\mathbb{R}_+$ denotes the non-negative real numbers. $C(n,k)$ denotes the binomial coefficient ("n choose k"). We denote $\mathcal{M}(A)$ as the space of all finite signed measures on A, and let $\operatorname{supp}(\mu)$ denote the support of $\mu \in \mathcal{M}(A)$, while $\mathcal{F}(A)$ denotes the space of all bounded measurable functions on A. For any Borel space A, let $\mathcal{B}(A)$ denote the Borel σ-algebra on A. In addition, let $\mathcal{C}(A)$ denote the set of continuous functions whose domain is A. Furthermore, we denote $\|f\|_{p,A} := \left(\int_A |f(x)|^p \,\mathrm{d}x\right)^{\frac{1}{p}}$ as

the p-norm of the bounded function f on the set $A \subset \mathbb{R}^n$. In both this and the previous notation, we sometimes drop the set A when there is no risk of confusion. For vector valued functions $f : \mathbb{R}^n \to \mathbb{R}^n$, $\|f\|_{\infty,A} := \max_{i=1:n} \|f_i\|_{\infty,A}$, where f_i is the i-th element of f. Similarly, $\|v\|_p := (\sum_{i=1}^n |v_i|^p)^{\frac{1}{p}}$ for $v = [v_1, \ldots, v_n]^\top \in \mathbb{R}^n$. On the other hand, we denote $\|\mu\|_{\text{T.V.}} := \sup_{\|v\|_\infty \leq 1} \int v(x)\mu(\mathrm{d}x)$ as the total variation norm. We also denote $\text{Lip}(\beta) \subset \mathcal{C}(A)$ as the set of all Lipschitz continuous functions on A that have the Lipschitz constant β. We further denote $\mathbb{R}[x]$ as the ring of multivariate polynomials of $x \in \mathbb{R}^n$ and Σ^2 as the SOS polynomials. More specifically, $p \in \Sigma^2$ if there exists an integer m and $u_i \in \mathbb{R}[x]$, such that $p = \sum_{i=1}^m u_i^2$. Moreover, we define the Cartesian product as $\times_{i=1}^n A_i := A_1 \times \ldots \times A_n$. $\langle \cdot, \cdot \rangle$ denotes the dual bracket between the dual pair $(\cdot, \cdot)$. Moreover, for linear operator $\mathcal{G}$, we let $\mathcal{G}^*$ denote its adjoint. For any class $B \subset \mathbb{R}^n$, the diameter is defined as $\text{diam}(B) := \sup_{x,y \in B} \|x - y\|_2$. In addition, $\mathbb{I}_\Gamma(x)$ denotes the indicator function of x on set Γ, namely,

$$\mathbb{I}_\Gamma(x) = \begin{cases} 1 & \text{if } x \in \Gamma, \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, we use ***italic bold font*** to denote stochastic elements, and $\boldsymbol{x} \sim \mu$ means that $\boldsymbol{x}$ has distribution μ. Specifically, $\mathcal{N}(b, \sigma^2)$ denotes the normal distribution with mean b and variance σ^2 and $\mathcal{N}_C(b, \sigma^2)$ denotes the truncated normal distribution on the hyper-cube C. $\mathcal{U}(A)$ denotes the uniform distribution on A.

2 Problem formulation

2.1 The "forward" problem

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space that carries the agent's state-action sequence $\{(\boldsymbol{x}_t, \boldsymbol{a}_t)\}_{t=1}^\infty$. We assume that the state-action sequence is governed by a Markov Decision Process $(X, A, \mathbb{K}, Q, \bar{\ell}, \alpha, \mu_1)$, i.e., the agent makes its decision by seeking to find an optimal policy to an infinite-horizon discounted cost optimal control problem. In particular, the optimal policy $\boldsymbol{\pi} = \{\pi_t\}$ is a sequence of stochastic kernels, and the action a_t that the agent taken at each time step t is sampled from the distribution $\pi_t(a|x_1, a_1, \cdots, x_{t-1}, a_{t-1}, x_t)$. It is further assumed that the observed agent's optimal actions a_t for all $t = 1 : N$ are generated from a deterministic stationary policy [12, Def. 2.3.2]. Namely,

Assumption 2.1 (Deterministic stationary policy)**.** It holds for the optimal policy $\bar{\boldsymbol{\pi}} = \{\bar{\pi}_t\}$ that $\bar{\pi}_t(a|x_1, a_1, \cdots, x_{t-1}, a_{t-1}, x_t) = \delta(\bar{\pi}(x_t))$, which is the Dirac measure concentrated at the measurable function $\bar{\pi} : X \to A$.

As a summary, the agent's decision mechanism is governed by solving

$$\min_{\pi} \; J(\pi; \mu_1) := \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1} \bar{\ell}(\boldsymbol{x}_t, \boldsymbol{a}_t)\right], \tag{1a}$$

$$\text{s.t.} \;\; \boldsymbol{x}_{t+1} \sim Q(\cdot|x_t, a_t), \tag{1b}$$

$$\boldsymbol{a}_t \sim \delta(\pi(x_t)) \tag{1c}$$

$$\boldsymbol{x}_1 \sim \mu_1, \tag{1d}$$

$$(\boldsymbol{x}_t, \boldsymbol{a}_t) \in \mathbb{K}, \quad t = 1, 2, \ldots \tag{1e}$$

$$\tag{1f}$$

where

- $X \subseteq \mathbb{R}^{n_x}$, $A \subseteq \mathbb{R}^{n_a}$ are the Borel spaces of states and actions, respectively;
- $\mathbb{K} := \{(x, a) \mid x \in X, a \in A(x)\} \subseteq X \times A$ is the set of feasible state-action pairs, where $\{A(x) \mid x \in X\} \subseteq A$ is the set of feasible control;

- $Q : (\mathcal{B}(X)\times\mathbb{K}) \to [0,1]$ is the transition kernel that characterizes the stochastic forward dynamics (Notably, "additive process noise" is a special case).
- $\bar{\ell} : \mathbb{K} \to \mathbb{R}$ is the running cost;
- $\alpha \in (0,1)$ is the discount factor;
- μ_1 is the initial state's distribution.

2.2 The IOC problem

Assume that as an observer, we can observe M trials of the agent. More precisely, let the trajectories $(\boldsymbol{x}_t^i, \boldsymbol{a}_t^i)_{t=1}^N$ be Independently Identically Distributed (I.I.D) over $i = 1 : M$. The observed trials $\{(x_t^i, a_t^i)\}_{i=1}^M$ are the realizations of the I.I.D state-action pairs $\{(\boldsymbol{x}_t^i, \boldsymbol{a}_t^i)\}_{i=1}^M$. Based on these expert demonstrations, the aim of IOC is to reconstruct a cost function with the prescribed structure and the corresponding policy estimate, which we summarize as follows.

Problem 2.2. Given the discount factor $\alpha \in (0,1)$, the transition kernel $Q(\cdot|x,a)$, and the M observations of the optimal "state-action pair" trajectories $\{(x_t^i, a_t^i)\}_{t=1}^N$ that are generated by (1), construct a consistent estimator of $\bar{\pi} : X \to A$ that takes the structure of $\bar{\ell}$ into account.

Note that different cost functions may lead to identical optimal policies. For instance, $\pi(x) = 0$ is optimal to both $\ell_1(x,a) = a^2$ and $\ell_2(x,a) = a^4$. Hence, $\bar{\theta}_\ell$ is in general not identifiable unless the feature functions vector φ is carefully chosen. Nevertheless, motivated by the practical need of "learning from demonstration", in this paper we choose to not focus on the identifiability of IOC. Instead, the question of identifiability of the cost function is left for future research. Despite this, we can still do consistency analysis in the sense that the collected state-action pairs (x_t^i, a_t^i) are indeed optimal to the reconstructed cost function as the data amount tends to infinity. Moreover, we can also get an estimate $\hat{\pi}$ of the underlying optimal policy $\bar{\pi}$ that corresponds to the estimated cost $\hat{\theta}_\ell^\top \varphi$ (cf. [46]). Such policy estimation intrinsically takes the prior knowledge of the structures in $\bar{\ell}$ into account, in contrast to "behaviour cloning" [3, 34, 39]. More precisely, behaviour cloning parameterizes a family of deterministic stationary policies $\pi(\cdot;\theta)$ (typically a neural network) and solves the following optimization problem

$$\min_\theta \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N} \|\boldsymbol{a}_t^i - \pi(\boldsymbol{x}_t^i;\theta)\|_2^2. \tag{2}$$

Indeed, it also offers a consistent estimator of $\bar{\pi}$, yet such control policy reconstruction is vulnerable to out-of-distribution data samples. Such samples can occur, e.g., due to compounding errors when using the control policy [32], due to distribution shifts in the noise [28], or when trying to generalize the behaviour to new environments [7, 8]. In contrast, by taking the structures in $\bar{\ell}$ into account, we obtain a more robust control policy estimate.

2.3 The technical assumptions

To proceed, we make the following assumptions.

Assumption 2.3 (Markov transition kernel)**.** The Markov transition kernel is a priori known. In addition, the Markov transition kernel $Q(\cdot|x,a)$ is Lipschitz continuous in total-variation distance, i.e., there exists $L_Q \geq 0$ such that for all $\eta_1 := (x_1, a_1) \in \mathbb{K}, \eta_2 := (x_2, a_2) \in \mathbb{K}$,

$$\|Q(\cdot|\eta_1) - Q(\cdot|\eta_2)\|_{\text{T.V.}} \leq L_Q\|\eta_1 - \eta_2\|_2.$$

Assumption 2.4 (State-action set hyper-cube)**.** The state-action set $X \times A$ is a hyper-cube, namely, $X = \times_{i=1}^{n_x}[x_{i,\min}, x_{i,\max}]$ and $A = \times_{i=1}^{n_a}[a_{i,\min}, a_{i,\max}]$. Moreover, for all $x \in X$, it hold for the feasible action set $A(x)$ that $A(x) = A$. Consequently, $\mathbb{K} = X \times A$.

Remark 2.5. Assumption 2.4 is mild from a practical point of view. This is because the hyper-cube can be chosen arbitrarily large to cover the region visited by the expert demonstrations. In this sense, even if the assumption is not satisfied globally in engineering practice, it may still be a reasonable approximation on the part of the state-action space that is most relevant to the expert policy.

Assumption 2.6 (Cost function)**.** The cost function $\bar{\ell}$ has prior structures in the sense that it belongs to a function class that can be parameterized linearly by $\bar{\theta}_\ell \in \mathbb{R}^{n_\ell}$, i.e., $\bar{\ell}(\cdot,\cdot) = \bar{\theta}_\ell^\top \varphi(\cdot,\cdot)$, where $\varphi(\cdot,\cdot) = [\varphi_1(\cdot,\cdot),\ldots,\varphi_{n_\ell}(\cdot,\cdot)]^\top$ is the vector of (known) L_φ-Lipschitz continuous feature functions, $\bar{\ell} \in \mathrm{Lip}(L_\ell)$ where $L_\ell = \|\bar{\theta}\|_1 L_\varphi$.

Assumption 2.7 (Initial value distribution)**.** The support of the initial value distribution μ_1 is X.

3 The algorithm construction

In this section, we construct the estimation algorithm for the cost function and the control policy. This is based on taking the nonlinear optimal control problem and lifing it into a infinite dimensional linear representation. More specifically, to tackle the nonlinearity, we reformulate (1) into a linear program based on occupation measure and analyze its optimality conditions. Based on the aforementioned optimality conditions, we construct the IOC algorithm, formulating an infinite dimensional linear programming problem whose optimal solution is a cost function that corresponds to the observed data's occupation measure. Next, we formulate an equivalent infinite dimensional linear programming problem which eleviates the problem of infinite time-horizon observations. We then approximate the problem of recovering the cost function so that it is numerically tractable, and finally, based on a reconstruction of the cost function, we formulate an estimator for the control policy. Consistency of the estimators are proved in Sec. 4.

3.1 The linear programming reformulation & its optimality conditions

Under the Assumptions 2.3, 2.4 and 2.6, we can reformulate the problem (1) as a linear program in the infinite-dimensional space of measures, as in [12, Chap. 6]. In particular, given a policy $\boldsymbol{\pi}$ and the initial value's distribution μ_1, the state-control occupation measure at time step t is defined as

$$\mu_t^\pi(\Gamma_x \times \Gamma_a) := \mathbb{E}[\mathbb{I}_{\Gamma_x \times \Gamma_a}(\boldsymbol{x}_t, \boldsymbol{a}_t)], \quad \forall (\Gamma_x \times \Gamma_a) \subset \mathbb{K}.$$

In addition, let $\mathcal{P} : \mathcal{M}_+(\mathbb{K}) \to \mathcal{M}_+(X)$ be the linear projection operator that transform any probability measure $\mu \in \mathcal{M}_+(\mathbb{K})$ to its marginal distribution $\tilde{\mu} \in \mathcal{M}_+(X)$ on the state. Moreover, for all $\Gamma_x \subset X$, let $q : \mathcal{M}_+(\mathbb{K}) \to \mathcal{M}_+(X)$ be the linear forward time shift operator defined by

$$(q\mu_t^\pi)(\Gamma_x) = \int_{\mathbb{K}} Q(\Gamma_x | x, a) \mu_t^\pi(\mathrm{d}x, \mathrm{d}a).$$

Therefore, in view of the definition of the transition kernel Q and the state-control occupation measure, it is clear that $\mathcal{P}\mu_{t+1}^\pi = q\mu_t^\pi$. Let us further define the following occupation measure

$$\mu^\pi := \sum_{t=1}^{\infty} \alpha^{t-1} \mu_t^\pi. \tag{3}$$

Equipped with the set-up above, we can reformulate the problem of optimizing (1) as

$$\min_{\mu^\pi \in \mathcal{M}_+(\mathbb{K})} \quad \langle \bar{\ell}, \mu^\pi \rangle \tag{4a}$$

$$\text{s.t.} \quad (\mathcal{P} - \alpha q)\mu^\pi = \mu_1. \tag{4b}$$

In fact, the Assumptions 2.3, 2.4 and 2.6 guarantees that (4) is equivalent to optimizing (1a) [12, Thm. 6.3.7] in the sense that there is an equivalence between the feasible points of the two problems

and that they both attains the same optimal value.[1] The equivalence between (1) and (4) means that we can consider the latter formulation when we consider the IOC problem to come.

To this end, the optimality conditions for the above problem are given in (5), and they can derived as follows. We first relax the constraint (4b) with a multiplier $\bar{V} \in \mathcal{F}(X)$, and get the Lagrangian $\mathcal{L} : \mathcal{M}_+(\mathbb{K}) \times \mathcal{F}(X) \to \mathbb{R}$

$$\begin{aligned}\mathcal{L}(\mu^\pi, \bar{V}) &= \langle \bar{\ell}, \mu^\pi \rangle - \langle \bar{V}, (\mathcal{P} - \alpha q)\mu^\pi - \mu_1 \rangle \\ &= \langle \bar{\ell} + (\alpha q^* - \mathcal{P}^*)\bar{V}, \mu^\pi \rangle + \langle \bar{V}, \mu_1 \rangle.\end{aligned}$$

Considering the first term in the above equation, we note that $\inf_{\mu^\pi \in \mathcal{M}_+(\mathbb{K})} \mathcal{L}(\mu^\pi, \bar{V}) > -\infty$ only if

$$\bar{\ell} + (\alpha q^* - \mathcal{P}^*)\bar{V} \geq 0. \tag{5a}$$

On the other hand, since the above equation holds, the infimum of the Lagrangian is attained for any μ^π such that

$$\langle \bar{\ell} + (\alpha q^* - \mathcal{P}^*)\bar{V}, \mu^\pi \rangle = 0, \tag{5b}$$

for example $\mu^\pi = 0$. However, the latter is not primally feasible, which gives us the last optimality condition, namely

$$(4b), \quad \mu^\pi \in \mathcal{M}_+(\mathbb{K}). \tag{5c}$$

In these equations, the adjoint operators are defined by $q^*\bar{V} : \mathbb{K} \mapsto \mathbb{R}$ and $\mathcal{P}^*\bar{V} : \mathbb{K} \mapsto \mathbb{R}$ via

$$(q^*\bar{V})(x, a) = \int_X \bar{V}(\chi) Q(d\chi | x, a) \tag{6a}$$

$$(\mathcal{P}^*\bar{V})(\boldsymbol{x}, \boldsymbol{u}) = \bar{V}(\boldsymbol{x}) \mathbb{I}_{\mathbb{K}}(\boldsymbol{x}, \boldsymbol{a}) = \bar{V}(\boldsymbol{x}). \tag{6b}$$

Based on the optimality conditions (5), we are ready to design the IOC algorithm.

3.2 The infinite dimensional IOC algorithm

Before we proceed, it is important to note that the IOC problem is ill-posedness if no further regularity conditions posed. Given the transition kernel $Q(\cdot \mid x, a)$, if there is a candidate running cost ℓ and function V that satisfy

$$\ell(x, a) - V(x) + \alpha \int_X V(y) Q(\mathrm{d}y | x, a) = 0, \quad \forall (x, a) \in \mathbb{K},$$

then the Bellman equation

$$V(x) = \min_a \left\{ \ell(x, a) + \alpha \int_X V(y) Q(\mathrm{d}y \mid x, a) \right\}$$

would hold trivially for every admissible action a. Consequently, every control action would appear optimal under such a choice of ℓ. This leads to a trivial and uninformative IOC solution.[2] To rule out such trivial and uninformative solution to the IOC problem, we impose the following regularization assumption.

Assumption 3.1. It holds for the "true" running cost and the "true" value function $(\bar{\ell}, \bar{V})$ that

$$\int_{\mathbb{K}} \left[\bar{\ell}(x, a) - \bar{V}(x) + \alpha \int_X \bar{V}(y) Q(\mathrm{d}y | x, a) \right] \mathrm{d}x\mathrm{d}a \geq 1,$$

where the integration on $\mathbb{K}$ is with respect to the Lebesgue measure.

[1] More specifically, Assumptions 2.3, 2.4 and 2.6 together imply [12, Ass. 4.2.1].

[2] Such pairs (ℓ, V) indeed exist. For example, if $\ell(x, a) \equiv \xi \geq 0$ and $V(x) \equiv \frac{\xi}{1-\alpha}$ for all $(x, a) \in \mathbb{K}$, then the above identity holds for any transition kernel $Q(\cdot \mid x, a)$.

The above assumption rules out the case that the running cost is constant since it requires the admissible state-action set $\mathbb{K}$ to at least contain a non-zero measure set whose components are not optimal. In addition, any choice of constant $\epsilon > 0$ on the right hand side of the above inequality can in principle accomplish this goal. This is because if $(\bar{\ell}, \bar{V})$ satisfies the optimality conditions (5), then any positive scaling of $(\bar{\ell}, \bar{V})$ would also satisfy (5). Thus we choose 1 without loss of generality and removes this degree of freedom.

Next, we construct an IOC algorithm that is based on the optimality conditions (5) and the adjoint operators (6). To this end, we first assume that $\mu^{\bar{\pi}}$ is known (we approximate it with data in Sec. 3.3). Now, for every $(x, a) \in \mathbb{K}$, we define the violation of the complementary slackness condition (5b) as $\psi : \mathbb{K} \to \mathbb{R}$ where

$$\psi(x, a; \theta_\ell, V) := \theta_\ell^\top \varphi(x, a) + \alpha(q^* V)(x, a) - V(x). \tag{7}$$

Then, we construct the IOC algorithm by minimizing the "violation" of (5b). Namely,

$$\min_{\substack{\theta_\ell \in \mathbb{R}^{n_\ell}, V \in \mathcal{F}(X) \\ \psi \in \mathcal{F}(\mathbb{K})}} \ \langle \psi, \mu^{\bar{\pi}} \rangle \tag{8a}$$

$$\text{s.t.} \quad \psi = \theta_\ell^\top \varphi + \alpha(q^* V) - \mathcal{P}^* V \tag{8b}$$

$$\psi(x, a) \geq 0, \forall (x, a) \in \mathbb{K}, \tag{8c}$$

$$\int_{\mathbb{K}} \psi(x, a) \mathrm{d}x \mathrm{d}a \geq 1. \tag{8d}$$

$$\|\theta_\ell\|_1 \leq \beta_\ell, \ \|\psi\|_{\infty, \mathbb{K}} \leq \beta_\psi, \tag{8e}$$

$$\|V\|_{\infty, X} \leq \beta_V. \tag{8f}$$

Note that the norm bounds β_ℓ, β_V and β_ψ in (8e) and (8f) can be chosen to be arbitrarily large. The norm constraint guarantees that the solutions to the optimization problem is always bounded, i.e., that we have a bounded estimator, which is important in the analysis later. Nevertheless, even without the constraint (8e) and (8f), the following Proposition still holds.

Proposition 3.2. The problem (8) attains at least one optimal solution, and the optimal value is 0. Moreover, for any optimal solution $(\theta_\ell^\star, V^\star, \psi^\star)$, $\bar{\pi}$ is an optimal control policy to the optimal control problem (1a) with the running cost $\ell^\star := \theta_\ell^{\star\top} \varphi$.

Proof. First, $\mu^{\bar{\pi}} \in \mathcal{M}_+(\mathbb{K})$ and the constraint (8c) ensure that the objective function (8a) is bounded below by zero. Moreover, the optimal value zero can be attained since the "true" $(\bar{\ell} = \bar{\theta}_\ell^\top \varphi, \bar{V}, \bar{\ell} + \alpha(q^* \bar{V}) - \mathcal{P}^* \bar{V})$ satisfies the optimality condition (5) and thus is both feasible to (8) and have objective function value zero.

Next, let $(\ell^\star, V^\star, \psi^\star)$ be any optimal solution to (8). Then $0 = \langle \psi^\star, \mu^{\bar{\pi}} \rangle = \langle \ell^\star + \alpha(q^* V^\star) - \mathcal{P}^* V^\star, \mu^{\bar{\pi}} \rangle$, and since constraint (8c) is also satisfied, $\bar{\pi}$ satisfies the optimality conditions (5). □

In addition, for bounded $\bar{\ell}$, its corresponding value function $\bar{V}$ is the limit of the well-known Value Iteration (VI), where the corresponding dynamic programming operator is a contraction mapping on $\mathcal{F}(X)$ with respect to supremum norm [12, p. 52]. Thus we have $\left\|\bar{V}\right\|_\infty < \infty$. Furthermore, by Assumption 2.3 and 2.6, $\bar{V}$ is the fixed point of VI, this implies that for all $\eta_1 := (x_1, a_1), \eta_2 := (x_2, a_2) \in \mathbb{K}$, it holds that

$$\begin{aligned}
&\left| \min_{a_1 \in A} (\bar{\ell} + \alpha q^* \bar{V})(x_1, a_1) - \min_{a_2 \in A} (\bar{\ell} + \alpha q^* \bar{V})(x_2, a_2) \right| \\
&\leq \sup_{a \in A} \left| (\bar{\ell} + \alpha q^* \bar{V})(x_1, a) - (\bar{\ell} + \alpha q^* \bar{V})(x_2, a) \right| \\
&\leq \sup_{a \in A} \left[\left| \bar{\ell}(x_1, a) - \bar{\ell}(x_2, a) \right| + \alpha \left| q^* \bar{V}(x_1, a) - q^* \bar{V}(x_2, a) \right| \right] \\
&\leq \sup_{a \in A} \left[(L_\ell + \|\bar{V}\|_\infty L_Q) \|[x_1^\top, a^\top]^\top - [x_2^\top, a^\top]^\top\|_2 \right]
\end{aligned}$$

$$= \underbrace{(L_\ell + \|\bar{V}\|_\infty L_Q)}_{\beta_C} \|x_1 - x_2\|_2$$

and hence $\bar{V} \in \mathrm{Lip}(\beta_C)$. Thus we can restrict ourselves to the set $V \in \mathrm{Lip}(\beta_C)$ and $\psi \in \mathcal{C}(\mathbb{K})$, and the "true" $\bar{V}$ and the corresponding $\bar{\psi}$ are still feasible when solving (8). That is, we instead consider the problem

$$\min_{\substack{\theta_\ell \in \mathbb{R}^{n_\ell}, V \in \mathcal{C}(X) \\ \psi \in \mathcal{C}(\mathbb{K})}} \quad \langle \psi, \mu^{\bar{\pi}} \rangle \tag{9a}$$

$$\text{s.t.} \quad \text{(8b)} - \text{(8f)}, \tag{9b}$$

$$V \in \mathrm{Lip}(\beta_C). \tag{9c}$$

Analogously to the constraints (8e) and (8f), we do not have to know the exact number of β_C in practice since such bound on Lipschitz constant can be taken arbitrarily large.

Nevertheless, we do not have access to the measure $\mu^{\bar{\pi}}$ in practice. One reason is that $\mu^{\bar{\pi}}$ is defined as a sum of infinitely many occupation measures $\mu_t^{\bar{\pi}}$ over time, yet we do not have infinite data length in practice. However, we can show that finite-time observations of the agent are sufficient to achieve the same effect as $\mu^{\bar{\pi}}$.

Proposition 3.3 (Finite-time IOC algorithm)**.** Under Assumption 2.1, for any $\nu \in \mathcal{M}_+(X)$, if $\mathrm{supp}(\nu) = \mathrm{supp}(\mathcal{P}\mu^{\bar{\pi}})$, then solving (9) is equivalent to solving

$$\min_{\substack{\theta_\ell \in \mathbb{R}^{n_\ell}, V \in \mathcal{C}(X) \\ \psi \in \mathcal{C}(\mathbb{K})}} \quad \langle \psi, \nu \otimes \bar{\pi} \rangle \tag{10a}$$

$$\text{s.t.} \quad \text{(8b)} - \text{(8f)}, \text{(9c)}, \tag{10b}$$

where

$$\langle f, \nu \otimes \bar{\pi} \rangle := \int_X \left[\int_A f(x,a) \bar{\pi}(\mathrm{d}a|x) \right] \nu(\mathrm{d}x), \quad \forall f.$$

More specifically, it holds that $\hat{\mu}^{\bar{\pi},N} := \sum_{t=1}^N \mu_t^{\bar{\pi}} = \left(\sum_{t=1}^N \mathcal{P}\mu_t^{\bar{\pi}} \right) \otimes \bar{\pi}$, and under Assumption 2.7, it holds that $\mathrm{supp}\left(\sum_{t=1}^N \mathcal{P}\mu_t^{\bar{\pi}} \right) = \mathrm{supp}(\mathcal{P}\mu^{\bar{\pi}})$.

Proof. Since $\mathrm{supp}(\nu) = \mathrm{supp}(\mathcal{P}\mu^{\bar{\pi}})$, for any $f : X \to \mathbb{R}_+$, it holds that

$$\int_X f(x)(\mathcal{P}\mu^{\bar{\pi}})(\mathrm{d}x) = 0 \Leftrightarrow \int_X f(x)\nu(\mathrm{d}x) = 0.$$

By Proposition 3.2, the optimal value of (8) is zero, and hence the optimal value of (9) is also zero. By following the same argument as that in the proof of Proposition 3.2, it follows from $\nu \otimes \bar{\pi} \in \mathcal{M}_+(\mathbb{K})$ and the constraint (8c) that the optimal value (10) is bounded from below by zero and that the "true" $(\bar{\ell}, \bar{V}, \bar{\ell} + \alpha(q^*\bar{V}) - \mathcal{P}^*\bar{V})$ attains the optimal value zero. Now suppose $(\ell^\star, V^\star, \psi^\star)$ is optimal to (9). Thus we have

$$\begin{aligned}
&(\ell^\star, V^\star, \psi^\star) \text{ minimizes (8)}, \\
\Rightarrow &\langle \psi^\star, \mu^{\bar{\pi}} \rangle = \int_X \left[\int_A \psi^\star(x,a) \bar{\pi}(\mathrm{d}a|x) \right] (\mathcal{P}\mu^{\bar{\pi}})(\mathrm{d}x) = 0, \\
\Leftrightarrow &\langle \psi^\star, \nu \otimes \bar{\pi} \rangle = \int_X \left[\int_A \psi^\star(x,a) \bar{\pi}(\mathrm{d}a|x) \right] \nu(\mathrm{d}x) = 0.
\end{aligned}$$

Since (9) and (10) have the same constraints, $(\ell^\star, V^\star, \psi^\star)$ is also optimal to (10). Conversely, the same argument applies.

□

3.3 The finite dimension approximation of the IOC algorithm

The above proposition implies that, instead of solving (9), we can spare ourselves from infinite data length and solve (10) with $\nu \otimes \bar{\pi} = \hat{\mu}^{\bar{\pi},N}$. Yet we still do not know the distribution $\hat{\mu}^{\bar{\pi},N}$ in practice, and hence we need to further approximate (10) with M I.I.D. samples $\{(x_t^i, a_t^i)\}_{t=1}^N$, $i = 1 : M$. In particular, we approximate the objective function (10a) with empirical average, i.e.,

$$\langle \psi, \hat{\mu}^{\bar{\pi},N} \rangle \approx \frac{1}{M} \sum_{i=1}^{M} \sum_{t=1}^{N} \psi(\boldsymbol{x}_t^i, \boldsymbol{a}_t^i). \tag{11}$$

Nevertheless, the approximated problem still optimizes over $V \in \mathrm{Lip}(\beta_C) \subset \mathcal{C}(X)$ and $\psi \in \mathcal{C}(\mathbb{K})$, which is infinite dimensional and thus difficult to solve. Moreover, the inequality constraint (8c) is challenging to handle since it needs to hold for all $(x, a) \in \mathbb{K}$. To address this issue, we approximate ψ with a polynomial and use Putinar's Positivstellensatz [29][17, Thm. 3.20] to make the polynomial nonnegative. More precisely, we approximate ψ by a $2d_\psi$-th order polynomial on $\mathbb{K}$ (without loss of generality, we consider only even-order polynomials to simplify the analysis to come), and represent the polynomial with Lagrange interpolation basis [35]. Namely, for all $\eta := (x, a) \in \mathbb{K}$, we have

$$\psi(\eta) \approx \hat{\psi}(\eta) := \underbrace{\begin{bmatrix} \hat{\psi}(\eta_{[1]}) & \cdots & \hat{\psi}(\eta_{[D_\psi]}) \end{bmatrix}}_{\theta_\psi^\top} \phi(\eta; \eta_{[\cdot]}), \tag{12}$$

where $D_\psi = C(n_x + n_u + 2d_\psi, 2d_\psi)$ is the dimension of $2d_\psi$-th order polynomial space, $\eta_{[\cdot]} = \{\eta_{[1]}, \ldots, \eta_{[D_\psi]}\}$ are the interpolation nodes, and the Lagrange polynomial basis $\phi(\eta; \eta_{[\cdot]})$ takes the form

$$\phi(\eta; \eta_{[\cdot]}) = \begin{bmatrix} \phi_1(\eta; \eta_{[\cdot]}) & \cdots & \phi_{D_\psi}(\eta; \eta_{[\cdot]}) \end{bmatrix}^\top.$$

Remark 3.4. Note that any polynomial basis would work. Nevertheless, here we choose to work with Lagrange interpolation basis $\phi(\eta; \eta_{[\cdot]})$. The reason is that this leads to computation benefits; more details about this is found in Sec. 5. Moreover, note that polynomial space $\mathcal{P}(X)$ is dense in $\mathcal{C}(X)$, and it holds that $\mathrm{Lip}(\beta_C) \subset \mathcal{C}(X)$. Therefore, polynomials can approximate any function in $\mathrm{Lip}(\beta_C)$ arbitrarily well; more detailed analysis is found in Sec. 4.

Next, to enforce the nonnegativity of $\hat{\psi}$, note that by Assumption 2.4, $\mathbb{K}$ can be characterized by the semi-algebraic set $\{(x, a) | g_i(x, a) \geq 0, i = 1 : n_x + n_a\}$, where $g_i(x, a) = (x_i - x_{i,\min})(x_{i,\max} - x_i)$ and analogously for the bounds on a. The quadratic module [17, Eq. 3.13] generated by $g_1, \ldots, g_{n_x+n_a}$ takes the form

$$\mathcal{Q}(g_{1:n_x+n_a}) := \{u_0 + \sum_{i=1}^{n_x+n_a} u_i g_i \mid u_0, u_i \in \Sigma^2\}.$$

Notably, $\hat{\psi} \in \mathcal{Q}(g_{1:n_x+n_a})$ means that $\hat{\psi}(\eta) \geq 0, \forall \eta \in \mathbb{K}$. Nevertheless, constraining $\hat{\psi}$ to be in $\mathcal{Q}(g_{1:n_x+n_a})$ might, a priori, be too restrictive. To see that it is in fact not the case, we first observe that $\mathbb{K}$ is also Archimedean [17, Def. 3.18]. To see this, we consider the function

$$f(\eta) = \sum_{i=1}^{n_x+n_a} g_i(x, a) = -\|\eta - \zeta\|_2^2 + \text{constant},$$

where $\zeta = [\zeta_1^x, \ldots, \zeta_{n_x}^x, \zeta_1^a, \ldots, \zeta_{n_a}^a]^\top$ and $\zeta_i^x = \frac{x_{i,\min}+x_{i,\max}}{2}$, $\zeta_i^a = \frac{a_{i,\min}+a_{i,\max}}{2}$. It holds that $f \in \mathcal{Q}(h_{1:n_x+n_a}, g_{1:m})$. Moreover, the set $\{\eta | f(\eta) \geq 0\}$ is compact. Hence by [17, Thm. 3.17, Def. 3.18], $\mathbb{K}$ is Archimedean. By Putinar's Positivstellensatz [29][17, Thm. 3.20], all polynomials that are positive on $\mathbb{K}$ must belong to $\mathcal{Q}(g_{1:n_x+n_a})$. In particular, this means that

$$\hat{\psi}(\eta) > 0, \ \forall \eta \in \mathbb{K} \quad \Longrightarrow \quad \hat{\psi} \in \mathcal{Q}(g_{1:n_x+n_a}). \tag{13}$$

Therefore we can approximate (8c) to abitrary accuracy with (13). On the other hand, since (13) says that $\hat{\psi}$ should belong to the quadratic module, it implies that the coefficients θ_ψ must belong to the cone $\left\{\theta \in \mathbb{R}^{D_\psi} | \theta^\top \phi \in \mathcal{Q}(g_{1:n_x+n_a})\right\}$, which, with a slight abuse of notation, we will denote as $\theta_\psi \in \mathcal{Q}(g_{1:n_x+n_a})$.

With the polynomial approximation (12), we can re-write the empirical average (11) as

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\hat{\psi}(\boldsymbol{x}_t^i, \boldsymbol{a}_t^i) = h^\top \theta_\psi,$$

where

$$\boldsymbol{h} = \left[\tfrac{1}{M}\textstyle\sum_{i=1}^{M}\sum_{t=1}^{N}\phi(\boldsymbol{x}_t^i, \boldsymbol{a}_t^i; \eta_{[\cdot]})\right].$$

Similarly, we can further approximate the constraint (8d) as

$$\int_{\mathbb{K}}\hat{\psi}(\eta)\mathrm{d}\eta = \theta_\psi^\top \underbrace{\left[\textstyle\int_{\mathbb{K}}\phi(\eta;\eta_{[\cdot]})\mathrm{d}\eta\right]}_{d} \geq 1.$$

In addition, the structure of the constraint (8b) lends itself to an approximation by linear equations. To see this, noticing that V only depends on x, we hence approximate V by $2d_V$-th order polynomial on X,

$$V(x) \approx \hat{V}(x) := \theta_V^\top r(x),\ \theta_V \in \mathbb{R}^{D_V},$$

where $D_V = C(n_x + 2d_V, 2d_V)$, and $r(x)$ is a basis of $2d_V$-th order polynomials on $\mathbb{R}^{n_x}$. Consequently, the coordinate of $\mathcal{P}^*V$ and q^*V with respect to basis $\phi(\eta;\eta_{[\cdot]})$ can be calculated by their value at the interpolation points, i.e.,

$$(\mathcal{P}^*V)(\eta) = \theta_V^\top \underbrace{\left[(\mathcal{P}^*r)(\eta_{[1]}) \quad \cdots \quad (\mathcal{P}^*r)(\eta_{[D_\psi]})\right]}_{=:G_1^\top}\phi(\eta;\eta_{[\cdot]}),$$
$$(q^*V)(\eta) = \theta_V^\top \underbrace{\left[(q^*r)(\eta_{[1]}) \quad \cdots \quad (q^*r)(\eta_{[D_\psi]})\right]}_{=:G_2^\top}\phi(\eta;\eta_{[\cdot]}).$$

In a similar manner, we further do a coordinate change on ℓ to the same basis $\phi(\eta;\eta_{[\cdot]})$. More specifically, by Assumption 2.6,

$$\ell = \theta_\ell^\top\varphi = \theta_\ell^\top\begin{bmatrix}\varphi_1\\ \vdots \\ \varphi_{n_\ell}\end{bmatrix} \approx \theta_\ell^\top \underbrace{\begin{bmatrix}\gamma_{\varphi_1}^\top\\ \vdots \\ \gamma_{\varphi_{n_\ell}}^\top\end{bmatrix}}_{=:H^\top}\phi,$$

where $\gamma_{\varphi_i}^\top\phi(\eta;\eta_{[\cdot]})$ is an approximation of the feature φ_i in the Lagrange polynomial basis $\phi(\eta;\eta_{[\cdot]})$. Therefore, the constraint (8b) now can be approximated with the following linear constraints on the coefficients of the polynomial approximations

$$\theta_\psi = \underbrace{\begin{bmatrix}H & \alpha G_2 - G_1\end{bmatrix}}_{\Xi_\psi \in \mathbb{R}^{D_\psi\times(n_\ell+D_V)}}\begin{bmatrix}\theta_\ell\\ \theta_V\end{bmatrix} \Rightarrow \underbrace{\begin{bmatrix}I & -H & G_1-\alpha G_2\end{bmatrix}}_{=:G}\begin{bmatrix}\theta_\psi\\ \theta_\ell\\ \theta_V\end{bmatrix} = 0. \tag{14}$$

As a brief summary, we can approximate (8) with the convex optimization problem

$$\min_{\theta_\psi\theta_\ell,\theta_V} \quad \boldsymbol{h}^\top\theta_\psi \tag{15a}$$

$$(14) \tag{15b}$$
$$\theta_\psi \in \mathcal{Q}(g_{1:n_x+n_a}), \tag{15c}$$
$$d^\top \theta_\psi \geq 1, \tag{15d}$$
$$\|\theta_\psi\|_\infty^2 \leq \beta'_\psi, \tag{15e}$$
$$\|\theta_\ell\|_\infty^2 \leq \beta'_\ell,\ \|\theta_V\|_\infty^2 \leq \beta'_V. \tag{15f}$$

Moreover, since all norms on finite dimensional vector spaces are equivalent [16, Theorem 2.4-5], the constraints (8e),(8f) and (9c) approximated in finite dimension is relaxed and simplified to (15e) and (15f), where $\beta'_\psi, \beta'_\ell, \beta'_V$ can be chosen arbitrarily large in practice.

3.4 Expert policy reconstruction

Solving (15) gives the estimates $(\hat{\ell}_{M,\mathbf{d}}, \hat{V}_{M,\mathbf{d}})$ for the "true" running cost and the value function $(\bar{\ell}, \bar{V})$, where $\mathbf{d} = (d_\psi, d_V)$. With the identified $\hat{\ell}_{M,\mathbf{d}}$, any suitable algorithm that solves the forward optimal control problem exactly (if possible) or approximately (e.g. MPC, Adaptive Dynamic Programming, Lasserre's hierarchy, etc.) may then be applied to obtain a policy aligned with the expert demonstrations. As mentioned previously in Sec. 2, such control policy reconstruction method should have better generalizability and robustness compared to behaviour cloning since such method has taken the prior knowledge of the cost structure into account during the policy estimation process.

Another way to obtain a policy which is consistent with the estimated cost function is to, in an approriate way, add this cost function as a regularizer in the behaviour cloning framework. To this end, note that by Assumption 2.1 and (5), it holds for the "true" $(\bar{\ell}, \bar{V})$ and the expert control policy $\bar{\pi}$ that

$$\begin{aligned} 0 &= \langle \bar{\psi} := \bar{\ell} + (\alpha q^* - \mathcal{P}^*)\bar{V}, \mu^{\bar{\pi}} \rangle + \langle \|a - \bar{\pi}(x)\|_2^2], \mu^{\bar{\pi}} \rangle \\ &= \langle \bar{\psi}, (\mathcal{P}\mu^{\bar{\pi}}) \otimes \bar{\pi} \rangle + \langle \|a - \bar{\pi}(x)\|_2^2], \mu^{\bar{\pi}} \rangle \\ &= \int_X \bar{\psi}(x, \bar{\pi}(x))(\mathcal{P}\mu^{\bar{\pi}})(\mathrm{d}x) + \langle \|a - \bar{\pi}(x)\|_2^2, \mu^{\bar{\pi}} \rangle. \end{aligned}$$

On the other hand, due to (8c), for expert policy $\bar{\pi}(x)$ and all deterministic stationary policies π it holds that

$$\int_X \bar{\psi}(x, \pi(x))(\mathcal{P}\mu^{\bar{\pi}})(\mathrm{d}x) + \langle \|\bar{\pi}(x) - \pi(x)\|_2^2, \mu^{\bar{\pi}} \rangle \geq 0.$$

Hence we parameterize π by $\theta_\pi \in \Theta$, where $\Theta \subset \mathbb{R}^{n_{\theta_\pi}}$ is compact and $\pi(x; \cdot) : \Theta \to A$ is continuous for any x, for instance, a neural network $\pi(x; \theta_\pi)$ whose weights are θ_π. And then we estimate $\bar{\pi}$ by solving

$$\min_{\theta_\pi \in \Theta} \langle \bar{\psi}(x, \pi(x; \theta_\pi)) + \|\bar{\pi}(x) - \pi(x; \theta_\pi)\|_2^2, \mathcal{P}\mu^{\bar{\pi}} \rangle.$$

Notably, by following an argument similar to the proof of Proposition 3.3, it is equivalent to solving

$$\min_{\theta_\pi} \langle \bar{\psi}(x, \pi(x; \theta_\pi)) + \|\bar{\pi}(x) - \pi(x; \theta_\pi)\|_2^2, \mathcal{P}\hat{\mu}_t^{\bar{\pi},N} \rangle \tag{16}$$

Though we do not have access to neither the actual distribution $\hat{\mu}_t^{\bar{\pi},N}$, nor $\bar{\psi}$, we can still approximate the above problem. First, the integral can be approximated with the empirical mean of the I.I.D. observations of the optimal "state-action pair" trajectories $\{(x_t^i, a_t^i)\}_{t=1}^N$, $i = 1 : M$. Next, we approximate $\bar{\psi}$ with the approximation $\hat{\psi}_{M,\mathbf{d}}$ that solves (15). Together, this gives the finite-dimensional optimization problem

$$\min_{\theta_\pi \in \Theta} \frac{1}{M} \sum_{t=1}^{N} \sum_{i=1}^{M} \hat{\psi}_{M,\mathbf{d}}(\boldsymbol{x}_t^i, \pi(\boldsymbol{x}_t^i; \theta_\pi)) + \|\boldsymbol{a}_t^i - \pi(\boldsymbol{x}_t^i; \theta_\pi)\|_2^2. \tag{17}$$

As a summary, we have constructed the infinite dimensional IOC algorithm (10) and a finite dimensional approximation (15) that can be solved numerically. More specifically, solving (15) yields an approximation of the cost function as well as the value function that can lead to the "true" policy. As mentioned previously, with the identified cost function, one may apply any standard optimal control method to obtain a policy with good generalization capability. In addition, solving (17) provides another way of estimating the optimal control policy. The complete estimation procedure is described in Algorithm 1.

Algorithm 1 Estimation procedure for cost and policy

Assumptions: Let Assumptions 2.1, 2.3, 2.4, 2.6, 2.7, and 3.1 hold.
Input: Data $\{(x_t^i, a_t^i)\}_{t=1}^N$, $i = 1 : M$, system dynamics $Q(\cdot|x, a)$, state and action space $\mathbb{K} = X \times A$, polynomial degrees (d_ψ, d_V), parametrization $\pi(\cdot; \theta_\pi)$.
Output: $(\hat{\ell}_{M,\mathbf{d}}, \hat{V}_{M,\mathbf{d}})$ and $\pi(\cdot; \hat{\theta}_\pi)$.
1: Design interpolation the points $\eta_{[\cdot]}$ and the polynomial bases $r(x)$.
2: Calculate the matrix G, defined in (14), from system dynamics and sample moments vector h from data.
3: Obtain the estimate $(\hat{\ell}_{M,\mathbf{d}}, \hat{V}_{M,\mathbf{d}})$ by solving (15).
4: Obtain the estimate $\pi(\cdot; \hat{\theta}_\pi)$: either by solving (17) to estimate the expert policy, or by applying any standard optimal control algorithm.

4 Consistency analysis

We now analyze whether the proposed Algorithm 1 can give a cost function estimate that is consistent in optimality with respect to the demonstrated trajectories. Moreover, we also analyze if such algorithm gives a consistent expert policy estimator.

By Proposition 3.3, in theory, the underlying cost function can be estimated by solving (10). Yet such infinite dimensional optimization problem is not numerically tractable and hence the decision variables ψ and V in (10) are approximated by finite-order polynomials so as to yield problem (15). In addition, the integrals are approximated via the empirical averages defined in (11) and (17). To justify these approximations, we present the following theorem, which states that as the data amount M and the polynomial orders $\mathbf{d} = (d_\psi, d_V)$ tend to infinity, the estimate $(\hat{\ell}_{M,\mathbf{d}}, \hat{V}_{M,\mathbf{d}})$ converges to a running cost and a value function that makes the demonstrated expert trajectories optimal.

Remark 4.1. Note that in (9), we optimize (θ_ℓ, V, ψ). To simplify the analysis, in view of (8b), we substitute ψ with the expression of θ_ℓ and V in the remainder of this section.

Theorem 4.2. Under Assumption 2.1, 2.3, 2.4, 2.6, 2.7, and 3.1, as the polynomial degrees d_ψ, d_V and the data sample size M approach infinity, the solution of (15) achieves the optimal value of (10) almost surely. Formally, the following equation holds a.s.

$$\lim_{M\to\infty} \lim_{d_V\to\infty} \lim_{d_\psi\to\infty} \langle \theta_{\ell,M}^{\mathbf{d},\top}\varphi + (\alpha q^* - \mathcal{P}^*)(\theta_{V,M}^{\mathbf{d},\top} r), \hat{\mu}^{\bar{\pi},N}\rangle = 0,$$

where $(\theta_{\ell,M}^{\mathbf{d}}, \theta_{V,M}^{\mathbf{d}})$ is an optimal solution of (15).

Remark 4.3. The above theorem states the fact that, the occupation measure $\hat{\mu}^{\bar{\pi},N}$ that corresponds to the observed optimal state-action pairs $\{(x_t^i, a_t^i)\}_{t=1}^N$, $i = 1, \ldots, M$ is "nearly" optimal to the approximated cost function $\theta_{\ell,M}^{\mathbf{d},\top}\ \varphi$ and the approximated value function $\theta_{V,M}^{\mathbf{d},\top}\ r$. This is because the optimality conditions (5) are "nearly" satisfied.

Remark 4.4. The convergence order in the above theorem: "first increase the polynomial orders, then increase the data sample volume", actually coincide with the practical engineering practice. In particular, since data collection is typically expensive, if we get a bad estimate of the cost function, we would first increase the polynomial order to sufficiently high before we decide to collect more data.

To prove Theorem 4.2, we decompose the overall analysis into Lemmas 4.5, 4.6 and 4.7. First, the following lemma shows that the "uniform law of large numbers" still holds in this infinite dimension function space.

Lemma 4.5. Let $\{\eta_i\}_{i=1}^M$ be independent samples drawn from probability distribution ν. Consider the stochastic optimization problem $\min_{f\in\mathcal{D}} \mathbb{E}_{\boldsymbol{\eta}\sim\nu}[f(\boldsymbol{\eta})]$ and its corresponding empirical approximation problem $\min_{f\in\mathcal{D}} \frac{1}{M}\sum_{i=1}^M f(\boldsymbol{\eta}_i)$, where $\mathcal{D}\subset\mathcal{C}(\mathbb{K})$ is equicontinuous [33, Def 11.27], pointwise bounded and closed. Then it holds that

$$\lim_{M\to\infty}\sup_{f\in\mathcal{D}}\left|\frac{1}{M}\sum_{i=1}^M f(\boldsymbol{\eta}_i)-\mathbb{E}[f]\right|=0, \text{a.s.}$$

Furthermore, let $f^\star$ and $f_M^\star$ denote the optimal solution of the original problem and the approximation problem, respectively. Then it holds that $\mathbb{E}_{\boldsymbol{\eta}\sim\nu}[f_M^\star(\boldsymbol{\eta})] \xrightarrow{a.s.} \mathbb{E}_{\boldsymbol{\eta}\sim\nu}[f^\star(\boldsymbol{\eta})]$.

Proof. By Arzela-Ascoli Theorem [33, Thm 11.28], every sequence in $\mathcal{D}$ has a uniform convergent subsequence on $\mathbb{K}$, and hence $\mathcal{D}$ is compact. Thus $\mathcal{D}$ admits a finite cover number [41, Def 2.1.5] $N(\epsilon,\mathcal{D},\|\cdot\|_\infty)$ for any $\epsilon>0$. Next, pick the centers $\{f_j\}_{j=1}^N$, it holds that $\forall f\in\mathcal{D}$, there is f_j, such that $\|f-f_j\|_\infty\le\epsilon$. Let $E_M[f]:=\frac{1}{M}\sum_{i=1}^M f(\boldsymbol{\eta}_i)$, it holds that

$$\begin{aligned}
&|E_M[f]-\mathbb{E}[f]|\\
&=|E_M[f]-\mathbb{E}[f_j]+\mathbb{E}[f_j]-E_M[f_j]+E_M[f_j]-\mathbb{E}[f_j]|\\
&\le|E_M[f_j]-\mathbb{E}[f_j]|+|\mathbb{E}[f_j]-\mathbb{E}[f]]|+|E_M[f]-E_M[f_j]|\\
&\le|E_M[f_j]-\mathbb{E}[f_j]|+\int\underbrace{|f_j-f|}_{\le\epsilon}d\nu+\frac{1}{M}\sum_{i=1}^M\underbrace{|f(\boldsymbol{\eta}_i)-f_j(\boldsymbol{\eta}_i)|}_{\le\epsilon}\\
&\le|E_M[f_j]-\mathbb{E}[f_j]|+2\epsilon.
\end{aligned}$$

Therefore, it holds that $\sup_{f\in\mathcal{D}}|E_M[f]-\mathbb{E}[f]|\le\max_{j=1:N}|E_M[f_j]-\mathbb{E}[f_j]|+2\epsilon$. By strong laws of large numbers [15, Thm 5.23], as $M\to\infty$, it holds that $|E_M[f_j]-\mathbb{E}[f_j]|\to0$ a.s. for all $j=1:N$. And hence $\max_{j=1:N}|E_M[f_j]-\mathbb{E}[f_j]|\to0$ a.s.. Thus $\limsup_{M\to\infty}\sup_{f\in\mathcal{D}}|E_M[f]-\mathbb{E}[f]|\le2\epsilon$, a.s.. Let $\epsilon\downarrow0$, then we have $\sup_{f\in\mathcal{D}}|E_M[f]-\mathbb{E}[f]|\xrightarrow{a.s.}0$ as $M\to\infty$.

Furthermore, note that the original and the approximated problems share the same feasible domain $\mathcal{D}$. And consequently, by the optimality of $f^\star$ and $f_M^\star$, respectively, it holds that $\mathbb{E}[f^\star]\le\mathbb{E}[f_M^\star], E_M[f_M^\star]\le E_M[f^\star]$. Therefore, it follows that

$$\begin{aligned}
&\mathbb{E}[f^\star]\le\mathbb{E}[f_M^\star]\le E_M[f_M^\star]+|E_M[f_M^\star]-\mathbb{E}[f_M^\star]|\\
&\le E_M[f^\star]+|E_M[f_M^\star]-\mathbb{E}[f_M^\star]|\\
&\le\mathbb{E}[f^\star]+\underbrace{|E_M[f^\star]-\mathbb{E}[f^\star]|}_{\xrightarrow{a.s.}0}+\underbrace{|E_M[f_M^\star]-\mathbb{E}[f_M^\star]|}_{\xrightarrow{a.s.}0}.
\end{aligned}$$

Thus we have $\mathbb{E}_{\boldsymbol{\eta}\sim\mu}[f_M^\star(\boldsymbol{\eta})]\xrightarrow{a.s.}\mathbb{E}_{\boldsymbol{\eta}\sim\mu}[f^\star(\boldsymbol{\eta})]$. □

The following Lemma 4.6 shows that we can approximate V with polynomials, reducing the problem to finite dimensions.

Lemma 4.6. Consider the optimization problem

$$\begin{aligned}
\min_{\substack{\theta_\ell\in\mathbb{R}^{n_\ell}\\ V\in\mathcal{C}(X)}}\quad & \langle\theta_\ell^\top\varphi+T(V),\mu\rangle && \text{(18a)}\\
\text{s.t.}\quad & \theta_\ell^\top\varphi+T(V)\ge0,\forall(x,a)\in\mathbb{K}, && \text{(18b)}\\
& \left\|\theta_\ell^\top\varphi+T(V)\right\|_1\ge1, && \text{(18c)}\\
& \|\theta_\ell\|_1\le\beta'_\ell, && \text{(18d)}\\
& \|V\|_\infty\le\beta'_V, && \text{(18e)}
\end{aligned}$$

$$V \in \mathrm{Lip}(\beta_C), \tag{18f}$$

where $\mu \in \mathcal{M}_+$, $T : \mathcal{C}(X) \to \mathcal{C}(\mathbb{K})$ is a bounded linear operator. On the other hand, its approximated problem takes the form

$$\min_{\substack{\theta_\ell \in \mathbb{R}^{n_\ell} \\ V \in \mathcal{P}_d(X)}} \quad \langle \theta_\ell^\top \varphi + T(V), \mu \rangle \tag{19a}$$

$$\text{s.t.} \quad (18b) - (18f), \tag{19b}$$

where $\{\mathcal{P}_d \subset \mathcal{C}(X)\}_d^\infty$ is the set of d-order polynomials. Let $(\theta_\ell^\star, V^\star)$ and $(\theta_{\ell,d}^\star, V_d^\star)$ be the optimal solution of (18) and (19), respectively. If the feasible set of (18) has a strictly feasible solution, then it holds that

$$\lim_{d\to\infty} \langle \theta_{\ell,d}^{\star\top} \varphi + T(V_d^\star), \mu \rangle = \langle \theta_\ell^{\star\top} \varphi + T(V^\star), \mu \rangle.$$

Proof. Since (18) is convex and its feasible set has a non-empty interior, for all ϵ, there must exist a strictly feasible solution (θ_0, V_0) and ϵ_1 such that $\forall (\theta, V) \in \{(\theta, V) \in \mathbb{R}^{n_\ell} \times \mathcal{C}(X) \mid \|\theta - \theta_0\|_2 + \|V - V_0\|_{\infty,\mathbb{K}} < \epsilon_1\}$ is still feasible and $\langle (\theta_0 - \theta_\ell^\star)^\top \varphi + T(V_0 - V^\star), \mu \rangle \le \epsilon/2$. Moreover, it also holds that $V_0 \in \mathrm{Lip}(\beta_C - \epsilon_C)$ for some $\epsilon_C > 0$. Next, we first extend V_0 from domain $\mathbb{K}$ to $\mathbb{R}^n$ by McShane's theorem [22], namely,

$$\tilde{V}_0(x) = \min_{x \in \mathbb{K}} \{V_0(y) + (\beta_C - \epsilon_C)\|x - y\|\},$$

and notably, $\tilde{V}_0 = V_0$, $\forall x \in \mathbb{K}$, and $\tilde{V}_0 \in \mathrm{Lip}(\beta_C - \epsilon_C)$. Next, let $\rho_\delta(x)$ be a mollifier and we mollify $\tilde{V}_0$ by $V_\delta(x) = (\tilde{V}_0 \circledast \rho_\delta)(x)$, where $\circledast$ denotes the convolution. Hence there exists a δ_ϵ, such that $\|V_{\delta_\epsilon} - \tilde{V}_0\|_{\infty,\mathbb{K}} \le \min\{\epsilon/(4\|T^*\|\|\mu\|_{\text{T.V.}}), \epsilon_1/2\}$, where $\|T^*\|$ is the operator norm of the dual T^*. Furthermore, it also preserves the Lipschitz constant, namely,

$$\begin{aligned}
|V_{\delta_\epsilon}(x) - V_{\delta_\epsilon}(y)| &= \left| \int_{\mathbb{R}^n} [\tilde{V}_0(x - z) - \tilde{V}_0(y - z)] \rho_{\delta_\epsilon}(z) \mathrm{d}z \right| \\
&\le \int_{\mathbb{R}^n} \left| \tilde{V}_0(x - z) - V_0(y - z) \right| \rho_{\delta_\epsilon}(z) \mathrm{d}z \\
&\le \int_{\mathbb{R}^n} (\beta_C - \epsilon_C)\|x - y\|_2 \rho_{\delta_\epsilon}(z) \mathrm{d}z \\
&= (\beta_C - \epsilon_C)\|x - y\|_2 \underbrace{\int_{\mathbb{R}^n} \rho_{\delta_\epsilon}(z) \mathrm{d}z}_{=1}.
\end{aligned}$$

Moreover, it follows from Stone-Weierstrass theorem that there exists a d_ϵ, such that for all $d > d_\epsilon$, $\|\nabla V_d - \nabla V_{\delta_\epsilon}\|_{\infty,X} \le \min\{\epsilon/(4\|T^*\|\|\mu\|_{\text{T.V.}} \cdot \mathrm{diam}(X)), \epsilon_C, \epsilon_1/(2 \cdot \mathrm{diam}(X))\}$, where $V_d \in \mathcal{P}_d$. And hence $\|V_d - V_{\delta_\epsilon}\|_{\infty,X} < \min\{\epsilon/(4\|T^*\|\|\mu\|_{\text{T.V.}}), \epsilon_1/2\}$. Therefore, it holds that $\|V_d - V_0\|_{\infty,X} \le \|V_d - V_{\delta_\epsilon}\|_{\infty,X} + \|V_{\delta_\epsilon} - \tilde{V}_0\|_{\infty,X} \le \min\{\epsilon/(2\|T^*\|\|\mu\|_{\text{T.V.}}), \epsilon_1\}$ and $\|\nabla V_d\|_{\infty,X} \le \|\nabla V_{\delta_\epsilon}\|_{\infty,X} + \|\nabla V_d - \nabla V_{\delta_\epsilon}\|_{\infty,X} \le \beta_C - \epsilon_C + \min\{\epsilon, \epsilon_C\} \le \beta_C$. This implies that (θ_0, V_d) is also feasible to both (18) and (19). Therefore, it holds that

$$\begin{aligned}
&\langle \theta_{\ell,d}^{\star\top} \varphi + T(V_d^\star), \mu \rangle \le \langle \theta_0^\top \varphi + T(V_d), \mu \rangle \\
&= \langle \theta_0^\top \varphi + T(V_0), \mu \rangle + \langle V_d - V_0, T^* \mu \rangle \\
&\le \langle \theta_0^\top \varphi + T(V_0), \mu \rangle + \underbrace{\|V_d - V_0\|_\infty}_{\le \epsilon/(2\|T^*\|\|\mu\|_{\text{T.V.}})} \|T^*\|\|\mu\|_{\text{T.V.}} \\
&\le \langle \theta_\ell^{\star\top} \varphi + T(V^\star), \mu \rangle + \underbrace{\langle (\theta_0 - \theta_\ell^\star)^\top \varphi + T(V_0 - V^\star), \mu \rangle}_{\le \epsilon/2} + \epsilon/2 \\
&\le \langle \theta_\ell^{\star\top} \varphi + T(V^\star), \mu \rangle + \epsilon.
\end{aligned}$$

This implies that for any $\epsilon > 0$, a sufficiently high approximation order d_ϵ guarantees an ϵ-close optimal solution. Consequently, the optimal value to the approximated problem converges to the optimal value of the original problem as the polynomial approximation order tends to infinity. □

Next, Lemma 4.7 shows that polynomial approximations of the basis functions can be introduced without essentially changing the optimization results, in the sense that the feasibility and optimal value are preserved up to an arbitrarily small perturbation.

Lemma 4.7. For $\phi : \mathbb{K} \to \mathbb{R}^n$ that is continuous and $\mu \in \mathcal{M}_+$, consider the optimization problem

$$\min_{\theta\in\mathbb{R}^n} \langle \theta^\top \phi, \mu\rangle \tag{20a}$$
$$\text{s.t. } \theta^\top \phi \geq 0, \forall \eta \in \mathbb{K}, \tag{20b}$$
$$\left\|\theta^\top \phi\right\|_1 \geq 1, \tag{20c}$$
$$\|\theta\|_1 \leq \beta, \tag{20d}$$

as well as the optimization problem obtained by replacing the basis functions ϕ by a polynomial approximation $\tilde{\phi}_d$ of order d, that is, the approximated problem

$$\min_{\theta\in\mathbb{R}^n} \langle \theta^\top \tilde{\phi}_d, \mu\rangle \tag{21a}$$
$$\text{s.t. } \theta^\top \tilde{\phi}_d \geq 0, \forall \eta \in \mathbb{K}, \tag{21b}$$
$$\left\|\theta^\top \tilde{\phi}_d\right\|_1 \geq 1, \tag{21c}$$
$$\|\theta\|_1 \leq \beta/2. \tag{21d}$$

Let $\theta^\star$ and $\theta_d^\star$ be the optimal solution to (20) and (21) respectively. Suppose $\|\theta^\star\|_1 \leq \beta/4$ and the first elements in ϕ and $\tilde{\phi}_d$ are both 1. If $\lim_{d\to\infty} \left\|\phi - \tilde{\phi}_d\right\|_\infty = 0$, then there exists $\theta_d^\star + a_d$ that is feasible for (20) with $\lim_{d\to\infty} \|a_d\|_1 = 0$ and $\lim_{d\to\infty} \langle \theta_d^{\star\top} \phi, \mu\rangle = \langle \theta^{\star\top} \phi, \mu\rangle$.

Proof. Since $\lim_{d\to\infty} \|\phi - \tilde{\phi}_d\|_\infty = 0$, for any small enough $\epsilon < 1$, there exists a d_ϵ such that $\|\phi - \tilde{\phi}_d\|_\infty \leq \epsilon$ for every $d \geq d_\epsilon$. This further implies that for all $\eta \in \mathbb{K}$,

$$|\theta^\top \phi(\eta) - \theta^\top \tilde{\phi}_d(\eta)| \leq \sum_{i=1}^n |\theta_i| \underbrace{|\phi_i(\eta) - \tilde{\phi}_{d,i}(\eta)|}_{\leq\epsilon} \leq \|\theta\|_1 \epsilon, \tag{22}$$

where θ_i, ϕ_i and $\tilde{\phi}_{d,i}$ are the i-th element of the corresponding vectors, respectively. Moreover, let $a_d = [\|\theta_d^\star\|_1 \epsilon, 0, \ldots, 0]^\top$. Since $\theta_d^\star$ minimizes the approximated problem (21), it must satisfy the constraints (21b), (21c) and (21d); and since we assume the first elements of both ϕ and $\tilde{\phi}_d$ are both 1, it follows from (22) that $(\theta_d^\star + a_d)^\top \phi(\eta) = \theta_d^{\star\top} \phi(\eta) + \|\theta_d^\star\|_1 \epsilon \geq \theta_d^{\star\top} \tilde{\phi}_d \geq 0$. Hence it further holds that

$$\begin{aligned}\|(\theta_d^\star + a_d)^\top \phi\|_1 &= \int_{\mathbb{K}} (\theta_d^\star + a_d)^\top \phi(\eta) \mathrm{d}\eta \geq \int_{\mathbb{K}} \theta_d^{\star\top} \tilde{\phi}_d(\eta) \mathrm{d}\eta \\ &= \|\theta_d^\star \tilde{\phi}_d\|_1 \geq 1.\end{aligned}$$

Furthermore, $\theta_d^\star$ also must satisfy (21d) and hence $\|\theta_d^\star + a_d\|_1 \leq \|\theta_d^\star\|_1 + \|a_d\|_1 = \|\theta_d^\star\|_1 + \|\theta_d^\star\|_1 \epsilon \leq \beta$. Thus $\theta_d^\star + a_d$ is also feasible to the original problem (20), and hence it holds that $\langle \theta^{\star\top} \phi, \mu\rangle \leq \langle (\theta_d^\star + a_d)^\top \phi, \mu\rangle$ by optimality.

Similarly, let $b = [\|\theta^\star\|_1 \epsilon, 0, \ldots, 0]^\top$ and in view of we assume $\|\theta^\star\|_1 \leq \beta/4$, it follows that $\theta^\star + b$ is also feasible to the approximated problem (21) and it holds that $\langle \theta_d^{\star\top} \tilde{\phi}_d, \mu\rangle \leq \langle (\theta^\star + b)^\top \tilde{\phi}_d, \mu\rangle$ by optimality. Notably, since μ is a positive measure and in view of (22), we can bound the objective

function values by

$$\begin{aligned}&\langle(\theta_d^\star + a_d)^\top \phi, \mu\rangle = \langle \theta_d^{\star\top}\phi, \mu\rangle + \langle \|\theta_d^\star\|_1\, \epsilon, \mu\rangle \\ &\leq \langle \theta_d^{\star\top}\phi, \mu\rangle + 2\langle \underbrace{\|\theta_d^\star\|_1}_{\leq \beta/2}\, \epsilon, \mu\rangle \leq \langle \theta_d^{\star\top}\tilde{\phi}_d, \mu\rangle + \epsilon\langle \beta, \mu\rangle,\end{aligned}$$

and similarly, it holds that $\langle(\theta^\star + b)^\top \tilde{\phi}_d, \mu\rangle \leq \langle \theta^{\star\top}\phi, \mu\rangle + \epsilon\langle \beta/2, \mu\rangle$. Therefore, it follows that

$$\begin{aligned}&\langle \theta_d^{\star\top}\phi, \mu\rangle = \langle(\theta_d^\star + a_d)^\top \phi, \mu\rangle - \langle \|\theta_d^\star\|_1 \epsilon, \mu\rangle \\ &\geq \langle \theta^{\star\top}\phi, \mu\rangle - \langle \|\theta_d^\star\|_1 \epsilon, \mu\rangle \geq \langle \theta^{\star\top}\phi, \mu\rangle - \epsilon\langle \beta/2, \mu\rangle, \\ &\langle \theta_d^{\star\top}\phi, \mu\rangle \leq \langle(\theta_d^\star + a_d)^\top \phi, \mu\rangle \leq \langle \theta_d^{\star\top}\tilde{\phi}_d, \mu\rangle + \epsilon\langle \beta, \mu\rangle \\ &\leq \langle(\theta^\star + b)^\top \tilde{\phi}_d, \mu\rangle + \epsilon\langle \beta, \mu\rangle \leq \langle \theta^{\star\top}\phi, \mu\rangle + \epsilon\langle 3\beta/2, \mu\rangle.\end{aligned}$$

Since ϵ can be made arbitrarily small, the claim follows. □

Now that we are ready to prove Theorem 4.2.

Proof of Theorem 4.2. Let (θ_ℓ, V, ψ) be any feasible solution to (10). By Assumption 2.3 and 2.6, and since constraints (10b) holds, it holds that for any η_1, η_2

$$\begin{aligned}&|\ell(\eta_1) + \int V(\cdot)Q(\cdot|\eta_1) - \ell(\eta_2) - \int V(\cdot)Q(\cdot|\eta_2)| \\ &\leq |\ell(\eta_1) - \ell(\eta_2)| + |\int V(\cdot)Q(\cdot|\eta_1) - \int V(\cdot)Q(\cdot|\eta_2)| \\ &\leq (\|\theta_\ell\|_1 L_\varphi + \|V\|_\infty L_Q)\, \|\eta_1 - \eta_2\|_2, \\ &\leq (\beta_\ell L_\varphi + \beta_V L_Q)\, \|\eta_1 - \eta_2\|_2.\end{aligned}$$

Hence $\psi \in \mathrm{Lip}(\beta_\ell L_\varphi + \beta_V L_Q)$, therefore ψ lives in an equicontinuous set and Lemma 4.5 applies. Therefore there exists $M_1'(\epsilon)$, such that for all $M \geq M_1'(\epsilon)$ and every feasible ψ,

$$\left|\langle \psi, \hat{\mu}^{\bar{\pi},N}\rangle - \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N} \psi(\boldsymbol{\eta}_t^i)\right| \leq \epsilon, \quad \text{a.s.}, \tag{23}$$

where $\boldsymbol{\eta}_t^i := (\boldsymbol{x}_t^i, \boldsymbol{a}_t^i)$. Moreover, let $(\theta_{\ell,M}, V_M)$ be optimal to the finite-sample problem

$$\begin{aligned}\min_{\theta_\ell, V \in \mathcal{C}(X)} \quad & \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N} \psi(\boldsymbol{\eta}_t^i; \theta_\ell, V) \\ \text{s.t.} \quad & (10b).\end{aligned} \tag{24}$$

By the second conclusion in Lemma 4.5, since the optimal value of (10) is zero (cf. Proposition 3.3), we have

$$\langle \psi(\cdot; \theta_{\ell,M}, V_M), \hat{\mu}^{\bar{\pi},N}\rangle \leq \epsilon, \text{ a.s.} \tag{25}$$

Next, recall that $(\bar{\theta}, \bar{V})$ (and the corresponding $\bar{\psi}$) is an optimal solution of (10), and is therefore feasible for (24). Since the bounds β_ℓ, β_V (β_ψ) can be chosen arbitrarily large, we can construct a strictly feasible point by adding a positive constant to $\bar{V}$. Then by Lemma 4.6, there exists $d_V'(\epsilon, M)$, such that for any $d_V \geq d_V'(\epsilon, M)$,

$$\left|\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\Big(\psi(\boldsymbol{\eta}_t^i; \theta_{\ell,M}^{d_V}, \theta_{V,M}^{d_V,\top} r) - \psi(\boldsymbol{\eta}_t^i; \theta_{\ell,M}, V_M)\Big)\right| \leq \epsilon, \tag{26}$$

where $\theta^{d_V}_{\ell,M}, \theta^{d_V}_{V,M}$ is optimal to

$$\begin{aligned} \min_{\theta_\ell,\theta_V} \ & \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\psi(\boldsymbol{\eta}_t^i;\theta_\ell,\theta_V^\top r) \\ \text{s.t.} \ & \psi(x,a;\theta_\ell,\theta_V^\top r)\geq 0, \forall (x,a)\in\mathbb{K}, \\ & \int \psi(x,a;\theta_\ell,\theta_V^\top r)\mathrm{d}x\mathrm{d}a \geq 1, \quad \text{(15f)}. \end{aligned} \tag{27}$$

Finally, note that the continuous feature function vector φ and cost-to-go q^*V (6a) can be uniformly approximated to arbitrary accuracy by polynomials on $\mathbb{K}$. Moreover, also note that we may, without loss of generality, assume the first element in φ is one (this is the same as adding a constant in the cost function and does not change the optimal control policy), thus Lemma 4.7 applies. Therefore, there exists $d'_\psi(\epsilon, M, d_V)$ such that for all $d_\psi \geq d'_\psi(\epsilon, M, d_V)$ and the optimal solution $(\theta^{\mathbf{d}}_{\ell,M}, \theta^{\mathbf{d}}_{V,M})$ of (15), it holds that

$$\left|\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\Big(\psi(\boldsymbol{\eta}_t^i;\theta^{\mathbf{d}}_{\ell,M},\theta^{\mathbf{d},\top}_{V,M}r)-\psi(\boldsymbol{\eta}_t^i;\theta^{d_V}_{\ell,M},\theta^{d_V,\top}_{V,M}r)\Big)\right|\leq\epsilon. \tag{28}$$

On the other hand, there also exists a_{d_ψ} with $\|a_{d_\psi}\|_1 \leq \epsilon/\|\varphi\|_\infty$ such that $\psi(\cdot;\theta^{\mathbf{d}}_{\ell,M}+a_{d_\psi},\theta^{\mathbf{d},\top}_{V,M}r)$ is feasible for (27). Thus $\psi(\eta;\theta^{\mathbf{d}}_{\ell,M}+a_{d_\psi},\theta^{\mathbf{d},\top}_{V,M}r)\geq 0, \forall \eta\in\mathbb{K}$, and hence

$$\begin{aligned} &\langle\psi(\boldsymbol{\eta}_t^i;\theta^{\mathbf{d}}_{\ell,M},\theta^{\mathbf{d},\top}_{V,M}r),\hat{\mu}^{\bar{\pi},N}\rangle \\ &=\langle\psi(\boldsymbol{\eta}_t^i;\theta^{\mathbf{d}}_{\ell,M}+a_{d_\psi},\theta^{\mathbf{d},\top}_{V,M}r),\hat{\mu}^{\bar{\pi},N}\rangle-\langle a_{d_\psi}^\top\varphi,\hat{\mu}^{\bar{\pi},N}\rangle \\ &\geq -\langle a_{d_\psi}^\top\varphi,\hat{\mu}^{\bar{\pi},N}\rangle\geq -\|a_{d_\psi}\|_1\|\varphi\|_\infty\underbrace{\|\hat{\mu}^{\bar{\pi},N}\|_{\text{T.V.}}}_{=N}\geq -N\epsilon. \end{aligned}$$

Now, for all $M\geq M'_1(\epsilon), d_V\geq d'_V(\epsilon,M), d_\psi\geq d'_\psi(\epsilon,M,d_V)$, the following holds a.s.

$$\begin{aligned} &\langle\psi(\boldsymbol{\eta}_t^i;\theta^{\mathbf{d}}_{\ell,M},\theta^{\mathbf{d},\top}_{V,M}r),\hat{\mu}^{\bar{\pi},N}\rangle \\ &\overset{(23)}{\leq} \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\psi(\boldsymbol{\eta}_t^i;\theta^{\mathbf{d}}_{\ell,M},\theta^{\mathbf{d},\top}_{V,M}r)+\epsilon \\ &\overset{(28)}{\leq} \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\psi(\boldsymbol{\eta}_t^i;\theta^{d_V}_{\ell,M},\theta^{d_V,\top}_{V,M}r)+2\epsilon \\ &\overset{(26)}{\leq} \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\psi(\boldsymbol{\eta}_t^i;\theta_{\ell,M},V_M)+3\epsilon \\ &\overset{(23)}{\leq} \langle\psi(\cdot;\theta_{\ell,M},V_M),\hat{\mu}^{\bar{\pi},N}\rangle+4\epsilon\overset{(25)}{\leq}5\epsilon. \end{aligned}$$

and the claim follows. □

Furthermore, if the optimal solution $\theta^\star_\ell = \bar{\theta}$ is unique up to scalar normalization, namely, all optimal solutions differ only by a positive scaling factor, then the normalized approximated solution also converges to the normalized true optimum. Otherwise, the true cost function parameters $\bar{\theta}$ is not identifiable under the current IOC problem structure. Nevertheless, the expert control policy can still be reconstructed by solving (17). The subsequent corollary establishes that such reconstruction error is also bounded.

Corollary 4.8. Suppose $\pi(x;\cdot):\Theta\to A$ is continuous for any x, and Θ is compact. Let $\hat{\pi}_{M,\mathbf{d}}$ be the solution of (17), and suppose that there exists a $\theta_{\bar{\pi}}$ such that $\bar{\pi}=\pi(x;\theta_{\bar{\pi}})$. It holds that

$$\lim_{M\to\infty}\lim_{d_V\to\infty}\lim_{d_\psi\to\infty}\langle\|\hat{\pi}_{M,\mathbf{d}}-\bar{\pi}\|_2^2,\mathcal{P}\hat{\mu}^{\bar{\pi},N}\rangle=0, \text{ a.s.}$$

Proof. Since $\pi(\cdot;\theta_\pi)$ is continuous with respect to θ_π and θ_π belongs to a compact set, $\|\pi(x;\theta_\pi)-\bar\pi(x)\|_2^2$ is continuous with respect to θ_π for each x. Furthermore, since $\|\pi(\cdot;\theta_\pi)-\bar\pi(\cdot)\|_2^2 \le \mathrm{diam}(A)^2$, the function class $\{\|\pi(\cdot;\theta_\pi)-\bar\pi(\cdot)\|_2^2 \mid \theta_\pi\in\Theta\}$ is Glivenko-Cantelli [40, E.g. 19.8, p. 272]. Thus, there exists $M_2'(\epsilon)$, such that $\forall M\ge M_2'(\epsilon)$ and $\pi\in\{\pi(\cdot;\theta_\pi)|\theta_\pi\in\Theta\}$, the following inequality holds

$$\begin{aligned}&|\langle(\|\pi(\cdot)-\bar\pi(\cdot)\|_2^2, \mathcal{P}\hat\mu^{\bar\pi,N}\rangle - \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\|\pi(\boldsymbol{x}_t^i)-\bar\pi(\boldsymbol{x}_t^i)\|_2^2|\\ &\le \epsilon/4,\ \text{a.s.}\end{aligned} \tag{29}$$

On the other hand, let $(\theta^{\mathbf{d}}_{\ell,M},\theta^{\mathbf{d}}_{V,M})$ be optimal to (15). As shown in Theorem 4.2, for any $\epsilon>0$, there exists $M_3'(\epsilon), d_V'(\epsilon,M), d_{\psi,1}'(\epsilon,M,d_V)$, such that for all $M\ge M_3'(\epsilon), d_V\ge d_V'(\epsilon,M), d_\psi\ge d_{\psi,1}'(\epsilon,M,d_V)$, it holds $\langle\hat\psi_{M,\mathbf{d}},\hat\mu^{\bar\pi,N}\rangle\le\epsilon/4$ a.s, where $\hat\psi_{M,\mathbf{d}}:=\theta^{\mathbf{d},\top}_{\ell,M}\varphi+(\alpha q^*-\mathcal{P}^*)(\theta^{\mathbf{d},\top}_{V,M}r)$.

And applying Lemma 4.7, there also exists $d_{\psi,2}'(\epsilon,M,d_V)$ such that for all $d_\psi\ge d_{\psi,2}'(\epsilon,M,d_V)$, $\tilde\psi_{M,\mathbf{d}}:=(\theta^{\mathbf{d}}_{\ell,M}+a_{d_\psi})^\top\varphi+(\alpha q^*-\mathcal{P}^*)(\theta^{\mathbf{d},\top}_{V,M}r)$ is feasible for (8c) with $\|a_{d_\psi}\|_1\le\epsilon/(4N\|\varphi\|_\infty)$. Thus we have $\hat\psi_{M,\mathbf{d}}(x,a)\ge\tilde\psi_{M,\mathbf{d}}(x,a)-\|a_{d_\psi}\|_1\|\varphi\|_\infty\ge-\epsilon/(4N),\forall(x,a)\in\mathbb{K}$. Therefore, since $\hat\pi_{M,\mathbf{d}}$ is optimal to (17), and in view of (23), it follows that

$$\begin{aligned}&\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\left(-\epsilon/(4N)+\|\hat\pi_{M,\mathbf{d}}(\boldsymbol{x}_t^i)-\bar\pi(\boldsymbol{x}_t^i)\|_2^2\right)\\ &\le\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\hat\psi_{M,\mathbf{d}}(\boldsymbol{x}_t^i,\hat\pi_{M,\mathbf{d}}(\boldsymbol{x}_t^i))+\|\hat\pi_{M,\mathbf{d}}(\boldsymbol{x}_t^i)-\bar\pi(\boldsymbol{x}_t^i)\|_2^2\\ &\le\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\hat\psi_{M,\mathbf{d}}(\boldsymbol{x}_t^i,\bar\pi(\boldsymbol{x}_t^i))+\|\bar\pi(\boldsymbol{x}_t^i)-\bar\pi(\boldsymbol{x}_t^i)\|_2^2\\ &=\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\hat\psi_{M,\mathbf{d}}(\boldsymbol{x}_t^i,\boldsymbol{a}_t^i)\le\underbrace{\langle\hat\psi_{M,\mathbf{d}},\hat\mu^{\bar\pi,N}\rangle}_{\le\epsilon/4}+\epsilon/4\le\epsilon/2,\ \text{a.s.}.\end{aligned}$$

Thus, it follows from (29) that

$$\begin{aligned}&\forall M\ge\max\{M_1'(\epsilon),M_2'(\epsilon),M_3'(\epsilon)\},\qquad d_V\ge d_V'(\epsilon,M),\\ &d_\psi\ge\max\{d_{\psi,1}'(\epsilon,M,d_V),d_{\psi,2}'(\epsilon,M,d_V)\},\end{aligned}$$

it holds that $\langle\|\hat\pi_{M,\mathbf{d}}(\cdot)-\bar\pi(\cdot)\|_2^2,\mathcal{P}\hat\mu^{\bar\pi,N}\rangle\le\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{N}\|\hat\pi_{M,\mathbf{d}}(\boldsymbol{x}_t^i)-\bar\pi(\boldsymbol{x}_t^i)\|_2^2+\epsilon/4\le\epsilon$ a.s. And hence the statement follows. □

5 On practical implementation

As demonstrated in Theorem 4.2, the IOC problem can be solved numerically by solving (15), where $\hat\psi=\theta_\psi^\top\phi\in\mathcal{Q}(g_{1:n_x+n_a})$. This constraint corresponds to a Weighted Sum-of-Squares (WSOS) polynomial condition. Consequently, we solve such problem by following the pipeline proposed in [26], which treats the problem as a non-symmetric conic optimization problem. In particular, it constructs a logarithmically homogeneous self-concordant barrier (LHSCB) function, and applies a primal-dual interior point method. This framework is significantly more efficient and numerically stable than classical semidefinite-programming-based approaches. Next, we will explain the procedure briefly for self-containment.

As mentioned in Sec. 3.3, we represent the polynomials by linear combination of Lagrange interpolation basis. This turns the WSOS polynomial constraints to conic constraints on polynomial

coefficients. Next, we reformulate (15) to the following form

$$\begin{aligned} &\max_{\theta_\psi,\theta_\ell,\theta_V} (-\Xi_\psi^\top h)^\top \begin{bmatrix}\theta_\ell\\ \theta_V\end{bmatrix}, \\ &\text{s.t. } \begin{bmatrix} d^\top(-\Xi_\psi)\\ -\Xi_\psi\end{bmatrix}\begin{bmatrix}\theta_\ell\\ \theta_V\end{bmatrix} + \begin{bmatrix}s_1\\ \theta_\psi\end{bmatrix} = \begin{bmatrix}-1\\ 0\end{bmatrix}, \\ &\begin{bmatrix}s_1\\ \theta_\psi\end{bmatrix} \in \mathbb{R}_+ \times \mathcal{K}, \end{aligned} \tag{30}$$

where Ξ_ψ is defined in (14), $\mathcal{K} := \left\{\theta \in \mathbb{R}^{D_\psi} \mid \theta^\top \phi \in \ \mathcal{Q}(g_{1:n_x+n_a})\right\}$ is the conic constraints on the polynomial coefficients, and $s_1 \in \mathbb{R}_+$ is a slack variable. Moreover, as also mentioned in Sec. 3.3, in practice, we can choose the norm bounds β'_ψ, β'_ℓ, β'_V in (15) to be arbitrarily large to include the "true" cost function $\bar{\ell}$ and the "true" value function $\bar{V}$ as an interior point of the norm constraints (15e), (15f). This means that any "reasonable" estimator should make the norm bound constraints (15e) and (15f) inactive, otherwise it means that the orders of the polynomials or the amount of the data are just not sufficient. Therefore, in practice, we can safely ignore (15e) and (15f) and if the numerical solver fails to solve (30), it implies that we need to add more data or increase the orders of the polynomials.

Next, we view (30) as the dual problem of a non-symmetric conic optimization problem, which takes the form

$$\begin{aligned} &\min_{y\in\mathbb{R}^{D_\psi+1}} -y_1, \\ &\text{s.t. } \begin{bmatrix}\Xi_\psi^\top d & \Xi_\psi^\top\end{bmatrix} y = \Xi_\psi^\top h, \\ &y \in \mathbb{R}_+ \times \mathcal{K}^*, \end{aligned} \tag{31}$$

where $\mathcal{K}^*$ is the dual cone of $\mathcal{K}$. To handle the cone $\mathcal{K}^*$, we adopt the LHSCB function $F : (\mathcal{K}^*)^\circ \to \mathbb{R}$ given by [26, Sec. 6], where $(\mathcal{K}^*)^\circ$ is the interior of $\mathcal{K}^*$. More specifically, let q and p denote the polynomial bases of degree $2d_\psi$ and $(d_\psi - 1)$, respectively. The barrier function F shall take the form

$$F(y) = -\sum_{i=1}^{n_x+n_a} \ln\left(\det\left(\Lambda_i(y)\right)\right),$$

where Λ_i is the linear operator uniquely defined by $\Lambda_i(q) = g_i p p^\top$. Note that the operator Λ_i depends explicitly on the choice of the polynomial basis, and, as mentioned in Remark 3.4, we employ the Lagrange interpolation basis rather than the monomial basis since the latter often yields ill-conditioned $\Lambda_i(y)$ and degrades numerical performance. Moreover, we also use the approximated Fekete points [38] as the Lagrange interpolation nodes to avoid making the resulting polynomial sensitive to the interpolation values.

Finally, we solve (31) with the primal-dual interior-point method proposed in [37]. All subsequent numerical experiments are conducted based on this implementation.

6 Numerical experiments

In this section, we evaluate the performance of the proposed method on IOC problems on three typical systems. In particular,

1. Linear quadratic regulator (LQR), where the optimal policy and value function admit analytically polynomial representations.
2. Temperature control system which is nonlinear and the dynamics is characterized by polynomial.
3. Inverted pendulum system which is nonlinear and the dynamics involves sinusoidal terms.

The experimental results for these three cases are presented below.

6.1 Linear quadratic regulator

Consider the linear system

$$\begin{aligned}
&\boldsymbol{x}_{t+1} = A\boldsymbol{x}_t + B\boldsymbol{a}_t + \boldsymbol{w}_t,\\
&A = \begin{bmatrix} 1 & 0.1 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0.005 \\ 0.1 \end{bmatrix}, \boldsymbol{w}_t \sim \mathcal{N}_{[-1,1]^2}(0, 0.01^2 I),
\end{aligned}$$

where I is identity matrix. The proposed method is evaluated on 400 systems, all of them are with the discount factor $\alpha = 0.99$, and then with the cost function that takes the forms

$$\begin{aligned}
&\ell(x, a) = x^\top Q x + a^\top R a,\\
&Q = \mathrm{diag}(q_1, q_2), R = r\\
&(q_1, q_2, r) = (\bar{q}_1, \bar{q}_2, \bar{r})/\|(\bar{q}_1, \bar{q}_2, \bar{r})\|_2,
\end{aligned}$$

where $(\bar{q}_1, \bar{q}_2, \bar{r})$ is sampled from $\mathcal{U}(0, 1]^3$. Moreover, we set the polynomial degrees to $d_\psi = 3$ and $d_V = 2$, which is higher than needed to estimate the cost function. The reason for doing this is to assess the robustness of the proposed method under an over-complete basis representation. More specifically, a total of 128 systems were random initialized by standard Gauss distribution and simulated over 4 steps to collect data. As shown in Fig.1, the proposed method accurately recovers the coefficients of cost function in the LQR case, with the estimation errors concentrated near zero.

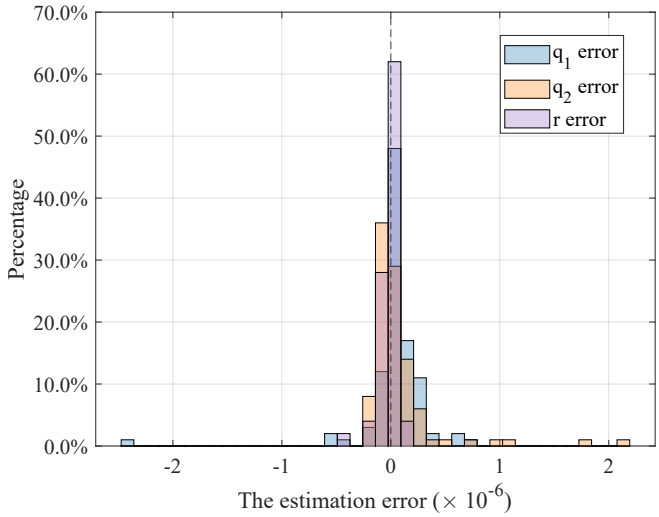


Figure 1: Histogram of Estimation Errors for the LQR Case. Only the central 99.5% of the samples are displayed. The experiment uses redundant polynomial degrees ($d_\psi = 3, d_V = 2$) to assess the robustness of the proposed method under an overcomplete basis representation.

6.2 Temperature control system

Consider the temperature control system

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + 0.5\boldsymbol{a}_t - \mathfrak{a}(\boldsymbol{x}_t - x_{\mathrm{env}}) - \mathfrak{b}(\boldsymbol{x}_t^4 - x_{\mathrm{env}}^4) + \boldsymbol{w}_t,$$

where $\mathfrak{a} = 5 \times 10^{-4}, \mathfrak{b} = 2.268 \times 10^{-8}, x_{\mathrm{env}} = 290, \boldsymbol{w}_t \in \mathcal{N}_{[-200,200]}(0, 2^2)$. In particular, the radiative heat dissipation term $\mathfrak{b}(\boldsymbol{x}_t^4 - x_{\mathrm{env}}^4)$ introduces nonlinearities to the system dynamics. Similar to the LQR case, we construct 100 cost functions with randomly chosen coefficients. In particular,

$$\begin{aligned}
&\ell(x, a) = q(x/200 - 3.25)^2 + r(a/500)^2,\\
&(q, r) = (\bar{q}, \bar{r})/\|(\bar{q}, \bar{r})\|_2,
\end{aligned}$$

where $(\bar{q}, \bar{r})$ is sampled from $\mathcal{U}(0,1]^2$. Moreover, to the best of our knowledge, such nonlinear forward optimal control problems do not have explicit solution for the optimal control policy. Therefore, to numerically generate an expert demonstration dataset, we employ the Model Predictive Control (MPC) algorithm with horizon length 64, and minimize the cost function with discount factor $\alpha = 0.9$. Using this approach, for each cost function, the expert demonstration dataset is generated by simulating 32 temperature control systems over 2 time steps. We estimated the coefficients of these cost functions under different polynomial approximation degrees and evaluated the estimation accuracy by comparing the normalized estimated coefficients with the ground-truth values. As shown in Fig.2, the proposed method performs well for this nonlinear case. Furthermore, higher approximation degrees lead to smaller errors, which is consistent with theoretical expectations.

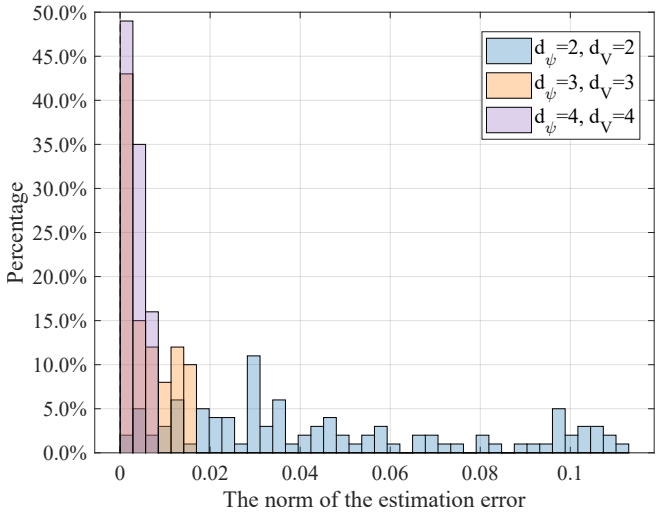


Figure 2: Estimate error histogram for the Temperature control system under different approximation degrees. Only the central 99.5% of data points are plotted. Estimation accuracy improves steadily as the approximation degree increases from 2 to 4, with narrower error distributions for higher-order approximations.

6.3 Inverted pendulum system

Consider the inverted pendulum system

$$\begin{aligned} p_{t+1} &= p_t + 0.01 v_t, \\ v_{t+1} &= 0.999 v_t + 0.01 \sin(p_t) + \cos(p_t) u_t. \end{aligned} \tag{32}$$

Notably, the above dynamics is deterministic and only weakly continuous. Hence it does not satisfy our Assumption 2.3 that the transition kernel is Lipschitz continuous in total variation distance. Yet we choose to do the experiment under this set-up to show that the algorithm is still able to achieve a good performance despite the fact that the stand-alone assumptions are not fulfilled completely. Moreover, analogous to the temperature control problem, this problem does not, to the best of our knowledge, have an explicit solution for the optimal control policy. To numerically generate the expert dataset, we employ an MPC with a 64-step time horizon length and discount factor $\alpha = 0.9$ to minimize the running cost function

$$\ell(p_t, v_t, u_t) = 100 p_t^2 + 10 v_t^2 + u_t^2.$$

Since the amplitude of the coefficient vectors does not influence the optimal control policy, both the ground truth and estimated coefficient vectors are normalized. Then the distances between these normalized coefficient vectors are computed to quantify the relative estimate errors. As shown in Fig. 3, the estimated errors decrease sharply as the data amount increases, and the error stabilizes

around 100 data samples. Further data additions yield no observable improvements, as the inherent precision limit of the approximation itself (rather than data scarcity) becomes the dominant factor restricting error reduction.

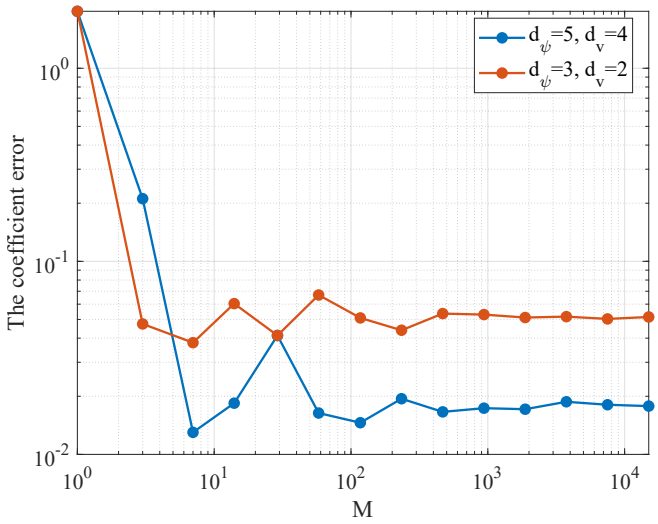


Figure 3: Estimated accuracy versus data volume M. The full dataset contains 16384 one-step trajectories. For each value of M, the training set is obtained by randomly sampling M trajectories from the full dataset. With approximately 100 data samples, the estimated errors reach around 5% for $d_\psi = 3, d_\psi = 2$ and around 2% for $d_\psi = 5, d_\psi = 4$, after that they level off.

Fig. 4 shows further results for different approximation degrees. In these experiments, we also look at the control policy estimation error. To this end, the expert control policy is recovered using the algorithm explained in Section 3.4, where a four-layer neural network with two hidden layers of sizes 128 and 64 is used to fit the control policy. We generate a dataset of 32768 trajectories, with random initial value given by $\mathcal{N}_{[-1,1]^2}(0, \mathrm{diag}(0.314^2, 0))$, envolving 4 time steps. The dataset is randomly divided into training and testing subsets with a ratio of 4:1. The recovery accuracy for the policy is evaluated on the test set using the Symmetric Mean Absolute Percentage Error (SMAPE) between the recovered and the ground truth policies, which is defined as

$$e_{\mathrm{SMAPE}} = \frac{2}{M_{\mathrm{test}}} \sum_{i=1}^{M_{\mathrm{test}}} \frac{|a_i - g(x_i;\theta_\pi)|}{|a_i| + |g(x_i;\theta_\pi)|}.$$

As illustrated in Fig. 4, as the approximation degree increases, both the coefficient estimation error and the policy recovery error decrease, which is consistent with our expectation of enhanced accuracy for higher-order approximations.

Finally, we also present an additional experiment that shows the robustness and generalizability of the proposed method compared to "behaviour cloning". In this setting, the expert data is generated on a new system whose parameters are perturbed, i.e.,

$$\begin{aligned} \boldsymbol{p}_{t+1} &= \boldsymbol{p}_t + 0.01\boldsymbol{v}_t + \boldsymbol{w}_{1,t}, \\ \boldsymbol{v}_{t+1} &= 0.999\boldsymbol{v}_t + 0.0105\sin(\boldsymbol{p}_t) + 1.05\cos(\boldsymbol{p}_t)u_t + \boldsymbol{w}_{2,t}, \end{aligned}$$

where $\boldsymbol{w}_{i,t} \sim \mathcal{N}_{[-0.4,0.4]^2}(0, 0.004^2)$ are independent process noise. In particular, we simulate 16 perturbed systems over 256 steps, and then estimate the coefficients in the cost function by assuming an unchanged system dynamics (32). The initial values are randomly sampled according to $(p_0, v_0) \sim \mathcal{U}[-2.0, 2.0]^2$, and the test system is initialized as $(\boldsymbol{p}_0, v_0) = (3.0, 0.0)$. This simulates the case of out-of-distribution data samples mentioned in Sec. 2. As shown in Fig. 5, the proposed method shows a better performance than behavior cloning method.

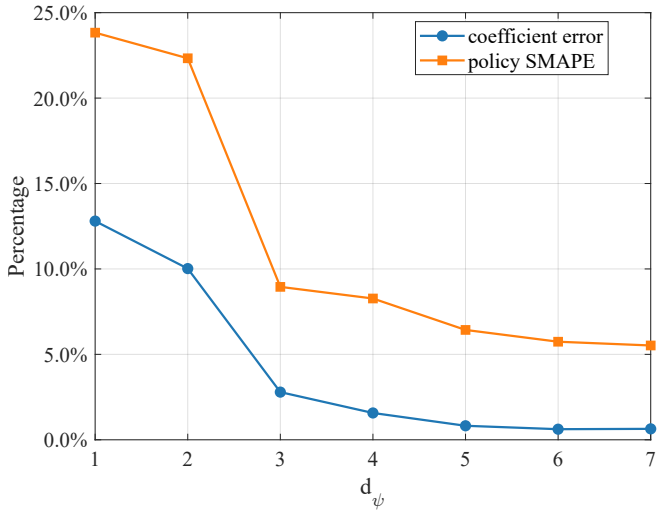


Figure 4: Estimated errors of the cost function coefficient vectors and SMAPE of the recovered control policy across varying approximation degrees. All experiments utilize around 131k data samples. For each d_ψ, best result across different d_V values is reported.

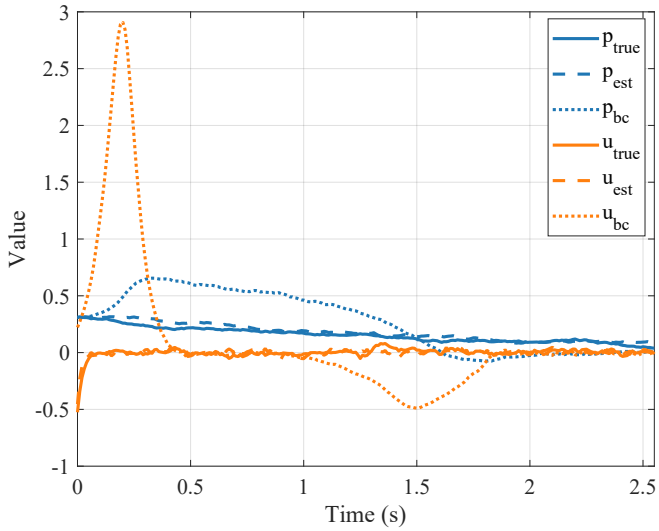


Figure 5: The angle and control trajectories of the test inverted pendulum system. Trajectories are generated by three different controllers: (1) MPC controller using the ground truth coefficients (labeled by "true"); (2) MPC controller using the estimated coefficients (labeled by "est"); (3) Behavior cloning controller (three-layer neural network with 128 hidden units) trained on the same dataset (labeled by "bc").

7 Conclusions

In this paper, we investigate the IOC problem for discrete-time, nonlinear stochastic systems with infinite-horizon discounted cost. In particular, we propose a polynomial-approximation-based approach that reformulates the IOC problem into a Sum-of-Squares optimization problem. This convex formulation avoids local minima and maintains theoretical interpretability.

The proposed estimator is shown to be asymptotically and statistically consistent: as the polynomial approximation order and the number of demonstration trajectories increase, the expert policy becomes nearly optimal under the estimated cost function, and the corresponding optimality gap asymptotically approaches zero. Numerical experiments on three typical optimal control systems val-

idate the theoretical consistency and further demonstrate the estimation accuracy and generalization capability of the proposed method.

Future work will extend the proposed framework to higher-dimensional systems, addressing the scalability challenges of polynomial and SOS approximations. Another promising direction is to handle information-limited settings, including partially observed states, unknown dynamics, or unmeasured control inputs.

References

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In Proceedings of the twenty-first international conference on Machine learning, page 1, 2004.

[2] Zoe Ashwood, Aditi Jha, and Jonathan W Pillow. Dynamic inverse reinforcement learning for characterizing animal behavior. In Advances in Neural Information Processing Systems, volume 35, pages 29663–29676, 2022.

[3] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9329–9338, 2019.

[4] Sourav Dey, Thibault Marzullo, and Gregor Henze. Inverse reinforcement learning control for building energy management. Energy and Buildings, 286:112941, 2023.

[5] Krishnamurthy Dvijotham and Emanuel Todorov. Inverse optimal control with linearly-solvable mdps. In Proceedings of the 27th International conference on machine learning (ICML-10), pages 335–342, 2010.

[6] Victor Nan Fernandez-Ayala, Shankar A Deka, and Dimos V Dimarogonas. Estimating unknown dynamics and cost as a bilinear system with koopman-based inverse optimal control. arXiv preprint arXiv:2501.18318, 2025.

[7] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In Maria Florina Balcan and Kilian Q. Weinberger, editors, Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, pages 49–58, New York, New York, USA, 20–22 Jun 2016. PMLR.

[8] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In International Conference on Learning Representations, 2018.

[9] Emiland Garrabe, Hozefa Jesawada, Carmen Del Vecchio, and Giovanni Russo. On convex data-driven inverse optimal control for nonlinear, non-stationary and stochastic systems. Automatica, 173:112015, 2025.

[10] Hongye Guo, Qixin Chen, Qing Xia, and Chongqing Kang. Deep inverse reinforcement learning for objective function identification in bidding models. IEEE Transactions on Power Systems, 36(6):5684–5696, 2021.

[11] Taosha Guo, Abed AlRahman Al Makdah, Vishaal Krishnan, and Fabio Pasqualetti. Imitation and transfer learning for lqg control. IEEE Control Systems Letters, 7:2149–2154, 2023.

[12] Onésimo Hernández-Lerma and Jean Bernard Lasserre. Discrete-time Markov control processes: basic optimality criteria. Springer Science & Business Media, 1996.

[13] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In Advances in Neural Information Processing Systems, volume 29, 2016.

[14] Frédéric Jean and Sofya Maslovskaya. Injectivity of the inverse optimal control problem for control-affine systems. In 2019 IEEE 58th Conference on Decision and Control (CDC), pages 511–516, 2019.

[15] Olav Kallenberg. Foundations of modern probability. Springer, 1997.

[16] Erwin Kreyszig. Introductory functional analysis with applications. John Wiley & Sons, 1991.

[17] Monique Laurent. Sums of squares, moment matrices and optimization over polynomials. In Emerging applications of algebraic geometry, pages 157–270. Springer, 2009.

[18] Yibei Li, Han Zhang, Yu Yao, and Xiaoming Hu. A convex optimization approach to inverse optimal control. In 2018 37th Chinese Control Conference (CCC), pages 257–262. IEEE, 2018.

[19] Bosen Lian, Wenqian Xue, Frank L Lewis, and Tianyou Chai. Inverse reinforcement learning for multi-player noncooperative apprentice games. Automatica, 145:110524, 2022.

[20] Zihao Liang, Wenjian Hao, and Shaoshuai Mou. A data-driven approach for inverse optimal control. In 2023 62nd IEEE Conference on Decision and Control (CDC), pages 3632–3637, 2023.

[21] Amarildo Likmeta, Alberto Maria Metelli, Giorgia Ramponi, Andrea Tirinzoni, Matteo Giuliani, and Marcello Restelli. Dealing with multiple experts and non-stationarity in inverse reinforcement learning: an application to real-life problems. Machine Learning, 110(9):2541–2576, 2021.

[22] Edward James McShane. Extension of range of functions. 1934.

[23] Negar Mehr, Mingyu Wang, Maulik Bhatt, and Mac Schwager. Maximum-entropy multi-agent dynamic games: Forward and inverse solutions. IEEE Transactions on Robotics, 39(3):1801–1815, 2023.

[24] Marcel Menner and Melanie N. Zeilinger. Convex formulations and algebraic solutions for linear quadratic inverse optimal control problems. In 2018 European Control Conference (ECC), pages 2107–2112, 2018.

[25] Katja Mombaur, Anh Truong, and Jean-Paul Laumond. From human to humanoid locomotion – an inverse optimal control approach. Autonomous robots, 28(3):369–383, 2010.

[26] Dávid Papp and Sercan Yildiz. Sum-of-squares optimization without semidefinite programming. SIAM Journal on Optimization, 29(1):822–851, 2019.

[27] Edouard Pauwels, Didier Henrion, and Jean-Bernard Lasserre. Linear conic optimization for inverse optimal control. SIAM Journal on Control and Optimization, 54(3):1798–1825, 2016.

[28] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 2817–2826. PMLR, 06–11 Aug 2017.

[29] Mihai Putinar. Positive polynomials on compact semi-algebraic sets. Indiana University Mathematics Journal, 42(3):969–984, 1993.

[30] Rahel Rickenbach, Anna Scampicchio, and Melanie N Zeilinger. Inverse optimal control as an errors-in-variables problem. In 6th Annual Learning for Dynamics & Control Conference, pages 375–386. PMLR, 2024.

[31] Luis Rodrigues. Inverse optimal control with discount factor for continuous and discrete-time control-affine systems and reinforcement learning. In 2022 IEEE 61st Conference on Decision and Control (CDC), pages 5783–5788, 2022.

[32] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[33] Walter Rudin. Real and complex analysis. McGraw-Hill, Inc., 1987.

[34] Claude Sammut. Behavioral cloning. In Claude Sammut and Geoffrey I. Webb, editors, Encyclopedia of Machine Learning and Data Mining, pages 120–124. Springer, Boston, MA, USA, 2017.

[35] Thomas Sauer and Yuan Xu. On multivariate Lagrange interpolation. Mathematics of computation, 64(211):1147–1170, 1995.

[36] Matthias Schultheis, Dominik Straub, and Constantin A Rothkopf. Inverse optimal control adapted to the noise characteristics of the human sensorimotor system. In Advances in Neural Information Processing Systems, volume 34, pages 9429–9442, 2021.

[37] Anders Skajaa and Yinyu Ye. A homogeneous interior-point algorithm for nonsymmetric convex conic optimization. Mathematical Programming, 150(2):391–422, 2015.

[38] Alvise Sommariva and Marco Vianello. Computing approximate Fekete points by QR factorizations of Vandermonde matrices. Computers & Mathematics with Applications, 57(8):1324–1336, 2009.

[39] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, page 4950–4957, 2018.

[40] Adrianus W. van der Vaart. Asymptotic statistics. Cambridge university press, Cambridge, United Kingdom, 1998.

[41] Jon Wellner et al. Weak convergence and empirical processes: with applications to statistics. Springer Science & Business Media, 2013.

[42] Chengpu Yu, Yao Li, Hao Fang, and Jie Chen. System identification approach for inverse optimal control of finite-horizon linear quadratic regulators. Automatica, 129:109636, 2021.

[43] Maryam Zare, Parham M Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. IEEE Transactions on Cybernetics, 2024.

[44] Han Zhang, Yibei Li, and Xiaoming Hu. Inverse optimal control for finite-horizon discrete-time linear quadratic regulator under noisy output. In 2019 IEEE 58th Conference on Decision and Control (CDC), pages 6663–6668. IEEE, 2019.

[45] Han Zhang and Axel Ringh. Inverse linear-quadratic discrete-time finite-horizon optimal control for indistinguishable homogeneous agents: A convex optimization approach. Automatica, 148:110758, 2023.

[46] Han Zhang and Axel Ringh. Inverse optimal control for averaged cost per stage linear quadratic regulators. Systems & Control Letters, 183:105658, 2024.

[47] Han Zhang and Axel Ringh. Statistically consistent inverse optimal control for discrete-time indefinite linear–quadratic systems. Automatica, 166:111705, 2024.

[48] Han Zhang, Axel Ringh, Weihan Jiang, Shaoyuan Li, and Xiaoming Hu. Statistically consistent inverse optimal control for linear-quadratic tracking with random time horizon. In 2022 41st Chinese Control Conference (CCC), pages 1515–1522, 2022.

[49] Han Zhang, Jack Umenberger, and Xiaoming Hu. Inverse optimal control for discrete-time finite-horizon linear quadratic regulators. Automatica, 110:108593, 2019.

[50] Zhengyuan Zhou, Michael Bloem, and Nicholas Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. IEEE Transactions on Automatic Control, 63(9):2787–2802, 2018.

[51] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In Aaai, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.