Statistically consistent inverse optimal control for discrete-time indefinite linear–quadratic systems

Automatica 166 (2024) 111705

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Statistically consistent inverse optimal control for discrete-time indefinite linear–quadratic systems☆

Han Zhang [a,b], Axel Ringh [c,*]

[a] Department of Automation, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
[b] The Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China
[c] Department of Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, 41296 Gothenburg, Sweden

ARTICLE INFO

Article history:
Received 16 December 2022
Received in revised form 20 December 2023
Accepted 27 March 2024
Available online 16 May 2024

Keywords:
Inverse optimal control
Inverse reinforcement learning
Indefinite linear quadratic regulator
System identification
Convex optimization
Semidefinite programming
Time-varying system matrices

ABSTRACT

The Inverse Optimal Control (IOC) problem is a structured system identification problem that aims to identify the underlying objective function based on observed optimal trajectories. This provides a data-driven way to model experts' behavior. In this paper, we consider the case of discrete-time finite-horizon linear–quadratic problems where: the quadratic cost term in the objective is not necessarily positive semi-definite; the planning horizon is a random variable; we have both process noise and observation noise; the dynamics can have a drift term; and where we can have a linear cost term in the objective. In this setting, we first formulate the necessary and sufficient conditions for when the forward optimal control problem is solvable. Next, we show that the corresponding IOC problem is identifiable. Using the conditions for existence of an optimum of the forward problem, we then formulate an estimator for the parameters in the objective function of the forward problem as the globally optimal solution to a convex optimization problem, and prove that the estimator is statistical consistent. Finally, the performance of the algorithm is demonstrated on two numerical examples.

1. Introduction

Optimal control is a powerful framework in which control decisions are performed in order to minimize some given objective function; see, e.g., one of the monographs (Anderson & Moore, 2007; Bertsekas, 2000). In fact, many processes in nature can be modeled as optimal control problems with respect to some criteria (Alexander, 1996). However, in applications of optimal control, a fundamental problem is to design an appropriate objective function. In order to induce an appropriate control response, the object function needs to be adapted to the contextual environment in which the system is operating. This is a difficult task, which relies heavily on the designers' experience and imagination.

Instead of designing the cost criteria, one way to overcome this difficulty would be to identify the cost function from the observations of an expert system that behaves "optimally" in the environment and thus "imitating" the expert behavior. The latter is known as Inverse Optimal Control (IOC) (Kalman, 1964), and has received considerable attention. In particular, IOC reconstructs the objective function of the expert system and hence predicts the closed-loop system's behavior using observed data as well as the knowledge of underlying system dynamics.

As one of the most classical optimal controller designs, linear–quadratic optimal regulators has been widely used in engineering. Though most of the literature considers the case when Q (the penalty parameter for the states) is positive semi-definite, indefinite linear–quadratic optimal control (Chen, Li, & Zhou, 1998; Ferrante & Ntogramatzidis, 2015, 2016; Rami, Chen, & Zhou, 2002; Ran & Trentelman, 1993) has found applications in, e.g., mathematical finance (Zhou & Li, 2000), crowd evacuation (Toumi, Malhamé, & Le Ny, 2020, 2021), and controller design for automatic steering of ships (Reid, Tugcu, & Mears, 1983). We are thus motivated to develop an IOC framework for general indefinite linear–quadratic optimal control. The linear–quadratic IOC problem has been studied under many different settings, including the infinite-horizon case in both continuous time (Anderson & Moore, 2007; Boyd, El Ghaoui, Feron, & Balakrishnan, 1994) and discrete time (Priess, Conway, Choi, Popovich, & Radcliffe, 2014), respectively, as well as the finite-horizon case in both continuous time (Li, Yao, & Hu, 2020; Li, Zhang, Yao, & Hu, 2018) and discrete time (Keshavarz, Wang, & Boyd, 2011;

☆ The work of Han Zhang was supported by National Natural Science Foundation (NNSF) of China under Grant 62103276, and the work of Axel Ringh was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, Sweden. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Simone Formentin under the direction of Editor Alessandro Chiuso.

* Corresponding author.
E-mail addresses: zhanghan_tc@sjtu.edu.cn (H. Zhang), axelri@chalmers.se (A. Ringh).

https://doi.org/10.1016/j.automatica.2024.111705

Yu, Li, Fang, & Chen, 2021; Zhang & Ringh, 2023; Zhang, Ringh, Jiang, Li, & Hu, 2022; Zhang, Umenberger, & Hu, 2019), respectively. IOC is also closely connected to inverse reinforcement learning (Ng & Russell, 2000), and this perspective has been used in Lian, Xue, Lewis, and Chai (2021), Xue, Kolaric et al. (2021) and Xue, Lian et al. (2021) to consider infinite horizon discrete-time and continuous-time linear–quadratic set-ups for regulation, tracking, and adversary scenarios, respectively. However, to the best of our knowledge, IOC frameworks for general indefinite linear–quadratic optimal control has not been considered. There are also other important aspects that have not been fully investigated in the aforementioned literature. More precisely, any real-world data would inevitably contain noise: it can either be process noise, observation noise, or both. Therefore, from robustness and accuracy perspectives, it is important to have a statistically consistent estimator, i.e., that converges to the true underlying parameter values as the number of observation grows. Moreover, many of the aforementioned IOC algorithms that are based on optimization either suffer from the fact that the estimation problems are nonconvex (Keshavarz et al., 2011; Yu et al., 2021; Zhang et al., 2019), and can therefore have issues with local minima, or suffer from the fact the estimation procedure needs to know the control gain a priori (Anderson & Moore, 2007; Boyd et al., 1994; Li et al., 2020, 2018; Priess et al., 2014). In the latter case, the estimation normally needs to be done in a two-stage procedure and the information is thus not used in the most efficient way. Furthermore, most of the literature on linear–quadratic IOC consider the regulation problem. However, in many experimental set-ups, an expert agent may have more complicated tasks than regulation, e.g., tracking a reference signal. Finally, real-world data can be of different time lengths, and this needs to be handled in a systematic way in order not to deteriorate the estimates.

In this work, we address these issues. More specifically, we extend our previous work (Zhang & Ringh, 2023; Zhang et al., 2022), and consider the generalized linear–quadratic, indefinite, discrete-time IOC problem with both process noise and measurement noise. The contribution of this work is three-fold:

(1) We give necessary and sufficient conditions for the well-posedness of the generalized indefinite discrete-time finite-horizon linear–quadratic optimal control problem.
(2) We prove the identifiability of the parameters in the objective function.
(3) We construct an IOC algorithm that works for both the positive semi-definite and the indefinite case. The algorithm is based on convex optimization, and we show that the estimator is statistically consistent. In addition, the convex optimization formulation guarantees that the statistically consistent estimate that corresponds to the global optimum can actually be attained in practice.

Notation: For a matrix G, $G^\dagger$ denotes the Moore–Penrose pseudo-inverse. For a square matrix $G = \left[\begin{smallmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{smallmatrix}\right]$, where G_{11} and G_{22} are square, we define the Schur complements $G\backslash G_{22} = G_{11} - G_{21}G_{22}^\dagger G_{12}$ and $G\backslash G_{11} = G_{22} - G_{12}G_{11}^\dagger G_{21}$ (see, e.g., Horn and Zhang (2005, Sec. 1.6)). $\mathbb{S}^n$ denotes the set of $n \times n$ symmetric matrices, while $\mathbb{S}^n_+$ denotes the set of $n \times n$ positive semi-definite matrices. For $G \in \mathbb{S}^n_+$, $G^{\frac{1}{2}}$ is the (unique) positive semi-definite matrix square root. $\|\cdot\|$ denotes l_2-norm and $\|\cdot\|_F$ denotes Frobenius norm. $\mathscr{B}^n_\varphi(x) := \{y \in \mathbb{R}^n \mid \|x - y\| < \varphi\}$, and $\bar{\mathscr{B}}^n_\varphi(x)$ denotes its closure. Moreover, $\mathbb{S}^n_+(\varphi) := \{G \in \mathbb{S}^n_+ \mid \|G\|_F \le \varphi\}$. For $G_1, G_2 \in \mathbb{S}^n$, $G_1 \succeq G_2$ means that $G_1 - G_2 \in \mathbb{S}^n_+$. For integers ν_1, ν_2, we let $\nu_1 : \nu_2$ denote $\nu_1, \nu_1 + 1, \dots, \nu_2$, and define the expression as empty if $\nu_2 < \nu_1$. By Im and ker we denote the image space and kernel, respectively, and by $^\perp$ we denote the orthogonal complement. Finally, we use ***italic bold font*** to denote stochastic elements, and use $\xrightarrow{p}$ to denote convergence in probability, i.e., for random elements $\{\boldsymbol{a}^i\}_{i=1}^\infty$ and $\boldsymbol{a}$, $\boldsymbol{a}^i \xrightarrow{p} \boldsymbol{a}$ means that for all $\varepsilon > 0$, $\lim_{i\to\infty} \mathbb{P}(\|\boldsymbol{a}^i - \boldsymbol{a}\| > \varepsilon) = 0$.

2. Problem formulation

In this section, we introduce the forward problem as well as the inverse problem. To solve the inverse problem, we use measured (noisy) optimal trajectories from an expert that performs the given task multiple times. This gives multiple demonstration trajectories that can be used to learn the cost.

We start by introducing the mathematical formulation of the forward optimal control problem. To this end, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space that carries a random vector $\bar{\boldsymbol{x}} \in \mathbb{R}^n$, stochastic processes $\{\boldsymbol{w}_t \in \mathbb{R}^n\}_{t=1}^\infty$, $\{\boldsymbol{v}_t \in \mathbb{R}^n\}_{t=1}^\infty$ (the measurement noise to appear in (2)), and a random variable $\boldsymbol{N} \in \{2, 3, \dots, \nu\} \subset \mathbb{Z}_+$. It is assumed that for each realization $(\bar{x}, N)$ of the random element $(\bar{\boldsymbol{x}}, \boldsymbol{N})$ (corresponding to the initial position and planning horizon length), the agent's control decision $\boldsymbol{u}_t$ is determined by a stochastic generalized linear–quadratic control problem, namely,

$$\min_{\substack{\boldsymbol{x}_{1:\nu},\\ \boldsymbol{u}_{1:\nu}}} J_N := \mathop{\mathbb{E}}_{\boldsymbol{w}_{\nu-N+1:\nu-1}} \Big[\frac{1}{2}\boldsymbol{x}_\nu^T \bar{Q} \boldsymbol{x}_\nu + \bar{q}^T \boldsymbol{x}_\nu + \sum_{t=\nu-N+1}^{\nu-1} \big[\frac{1}{2}\boldsymbol{x}_t^T \bar{Q} \boldsymbol{x}_t + \bar{q}^T \boldsymbol{x}_t + \frac{1}{2}\boldsymbol{u}_t^T \bar{R} \boldsymbol{u}_t \big] \Big] \tag{1a}$$

$$\text{s.t. } \boldsymbol{x}_{t+1} = A\boldsymbol{x}_t + B\boldsymbol{u}_t + d + \boldsymbol{w}_t, \quad t = \nu - N + 1 : \nu - 1, \tag{1b}$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t, \quad t = 1 : \nu - N, \tag{1c}$$

$$\boldsymbol{x}_1 = \bar{x}, \tag{1d}$$

$$\boldsymbol{u}_1 = \cdots = \boldsymbol{u}_{\nu-N} = 0, \tag{1e}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $\bar{Q} \in \mathbb{S}^n$, $\bar{R} \in \mathbb{S}^m$, and $\bar{q}, d \in \mathbb{R}^n$. More specifically, the minimization in (1) is over admissible control strategies with complete state information, i.e., $\boldsymbol{u}_t$ is a function that maps from $\mathbb{R}^n$ to $\mathbb{R}^m$, $\boldsymbol{u}_t : \boldsymbol{x}_t \mapsto \boldsymbol{u}_t(\boldsymbol{x}_t)$ (see, e.g., Åström (2006, Chp. 8)).

The formulation (1) is motivated by the fact that an agent can have different time horizon lengths (and initial values) to complete different tasks, while the underlying decision principle (i.e., running cost) remains unchanged since the principle is connected to the agent's characteristics. In particular, given a realization of the time-horizon length $\boldsymbol{N} = N$ and initial value $\bar{\boldsymbol{x}} = \bar{x}$, the agent starts to apply its control from the initial value $\bar{x}$ at the time instant $t = \nu - N + 1$ and the agent maintains *the same running cost* for each control execution. This formulation gives a systematic way to handle real-world data with different time lengths (see Zhang et al. (2022)). Moreover, note that since the dynamics (1b) and the running cost in (1a) are time-invariant, by Bellman's principle of optimality, (1) can be reformulated to optimal control problems with planning horizon length N and that start to control from an initial state at time point $t = 1$. However, it turns out to be convenient to align the optimal demonstration trajectories with different lengths at the end time point, and view the demonstration trajectories as if the expert starts to control the system at different time instants, i.e., to formulate the problem as in (1).

Remark 2.1. The reason why we only consider a time-invariant external forcing term d in (1b) is to simplify the presentation. The results of the paper also hold for time-varying forcing terms d_t, provided that the agent knows all the future forcing terms for $t = \nu - N + 1 : \nu - 1$. Similarly, this formulation can be

extended to tracking-problems by letting the linear cost term be time-varying, $q_t = -Qx_t^r$ where x_t^r is the reference signal, and the results in the paper follows analogously (cf. Zhang et al. (2022)). Notably, with time-varying d_t or q_t, the arguments in the paragraph just before (using time-invariance of running cost and dynamics) does not hold. Nevertheless, the formulation (1) is still of interest in application such as the tracking in rehabilitation trainings, see Zhang et al. (2022).

In the remainder of the paper, we make the following assumptions.

Assumption 2.2 (*Controlability and Full Rank*)**.** The system (A, B) is controllable, A is invertible, and B is of full-column rank.

Assumption 2.3 (*Independent White Random Noise*)**.** The discrete time stochastic processes $\{\boldsymbol{w}_t\}_{t=1}^{\infty}$ and $\{\boldsymbol{v}_t\}_{t=1}^{\infty}$ are independent zero-mean white-noise processes. More specifically, this means that $\mathbb{E}[\boldsymbol{w}_t] = 0$, $\mathbb{E}[\boldsymbol{v}_t] = 0$, $\forall t$, and $\operatorname{cov}(\boldsymbol{w}_t, \boldsymbol{w}_s) = \Sigma_w \delta(t-s)$, $\operatorname{cov}(\boldsymbol{v}_t, \boldsymbol{v}_t) = \Sigma_v \delta(t-s)$, where $\delta(t)$ is the Dirac-delta function, and where $\Sigma_w \succeq 0$ and $\Sigma_v \succeq 0$ are a priori known. Moreover, the random element $(\bar{\boldsymbol{x}}, \boldsymbol{N})$ is independent of the two stochastic processes.

Assumption 2.4 (*Support of the Planning Horizon*)**.** The constant $\nu \in \mathbb{Z}_+$ is known, and $\nu \geq n+1$. Moreover, the probability distribution for $\boldsymbol{N}$ satisfies $\mathbb{P}(\boldsymbol{N} \in [2, \nu]) = 1$, and $\mathbb{P}(\boldsymbol{N} = \nu) > 0$.

The rationale behind the first assumption is that we are considering a controllable discrete-time system that is not over-actuated,[1] and that is sampled from a continuous-time system. The third assumption means that the longest possible planning horizon is known, that it is sufficiently long, and that the longest horizon ν can be realized, i.e., it has a nonzero probability.

Next, note that since $\bar{Q}$ might not be positive semi-definite, (1) might not admit an optimal solution (see, e.g. Ferrante and Ntogramatzidis (2015, 2016)). We analyze the well-posedness of the forward problem (1) in depth in Section 3. However, before that, we have the following proposition which illustrates the reason why we emphasize the longest time-horizon length in Assumption 2.4. The proof can be found in the Appendix.

Proposition 2.5. *Under Assumptions 2.2 and 2.3, if the optimal control problem (1) with the objective function given by $(\bar{Q}, \bar{q}, \bar{R})$ admits a solution for planning horizon $N = \nu$ for any $\bar{x} \in \mathbb{R}^n$, then it admits a solution for all $N = 2 : \nu$ for any $\bar{x} \in \mathbb{R}^n$.*

With Assumption 2.4 and Proposition 2.5 in mind, we are thus interested in parameters that belong to the following set:

$$\begin{aligned}\mathscr{F}(\bar{R}) = \{&(Q, q) \in \mathbb{S}^n \times \mathbb{R}^n \mid \text{the optimal control problem} \\ &(1) \text{ with the objective function given by } (Q, q, \bar{R}) \text{ admits} \\ &\text{solutions a.s. under the distribution of } \bar{\boldsymbol{x}} \text{ and for all} \\ &N \in \{2, 3, \dots, \nu\}\}.\end{aligned}$$

More specifically, if $(\bar{Q}, \bar{q}) \in \mathscr{F}(\bar{R})$, then the forward problem is well-posed for any $N \in \{2, \dots, \nu\}$ under the distribution of $\bar{\boldsymbol{x}}$ almost surely. In addition, for the inverse problem we assume that the observation of the optimal states $\boldsymbol{x}_{1:\nu}$ are contaminated by observation noise:

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{v}_t, \quad t = 1 : \nu, \tag{2}$$

and that we observe M trials of the agent. More precisely, let $\boldsymbol{y}_t^i$ have the same distribution as $\boldsymbol{y}_t$, for all $i = 1 : M$ and all $t = 1 : \nu$. Then the observed M trajectories of the agent's trials, $\{y_t^i\}_{i=1}^M$, are just realizations of the I.I.D. random vectors $\{\boldsymbol{y}_t^i\}_{i=1}^M$.

Before we formulate the IOC problem, we also make the following assumption.

Assumption 2.6 (*Initial Value Distribution*)**.** The random element $(\bar{\boldsymbol{x}}, \boldsymbol{N})$ is such that $\mathbb{E}[\|\bar{\boldsymbol{x}}\|^2] < \infty$. Moreover, for all $N \in \{2, \dots, \nu\}$ such that $\mathbb{P}(\boldsymbol{N} = N) > 0$, it holds that for all $\chi \in \mathbb{R}^n$, there exists a $\rho > 0$ such that $\mathbb{P}(\bar{\boldsymbol{x}} \in \mathscr{B}_\epsilon^n(\rho\chi) \mid \boldsymbol{N} = N) > 0$ for all $\epsilon > 0$.

Intuitively speaking, the above assumption states that, for each planning horizon length of interest, the initial value for the forward problem can be in any "direction" from the origin. This turns out to be important for both the forward and the inverse problem. The latter will be discussed in Section 4.

With the setup presented in this section, we summarize the IOC problem to be considered in this paper. For the sake of simplicity, we consider the case $\bar{R} = I$ when designing the IOC algorithm.

Problem 2.7 (*General Stochastic Linear–Quadratic IOC*)**.** Suppose the unknown $(\bar{Q}, \bar{q}) \in \mathscr{F}(I)$. Given the optimal state trajectory observations $\{y_t^i\}_{t=1}^{\nu}$ of the agent's trials $i = 1 : M$ that are governed by (1), estimate the corresponding $(\bar{Q}, \bar{q})$ in the objective function (1a) that governs the agents' motion.

3. Forward problem analysis

Before we present the IOC algorithm set-up, we first need to analyze the forward problem. More precisely, we need to characterize the set $\mathscr{F}(\bar{R})$ and find the necessary and sufficient optimality conditions for the existence of such generalized indefinite linear–quadratic optimal control. This is not only because we want to ensure that the forward-problem is well-behaved, but also to construct the IOC algorithm based on the optimality conditions. Moreover, we also analyze the properties of the time-varying closed-loop system matrices that will be useful in developing the IOC algorithm. For the theoretical development in this section, we do not assume that $\bar{R} = I$.

3.1. Necessary and sufficient conditions for existence of optimal control

To this end, we first derive the necessary and sufficient conditions for existence of optimal control to (1). The results build on Ferrante and Ntogramatzidis (2015); in particular, some of the proof ideas are inspired by the proof in Ferrante and Ntogramatzidis (2015, Thm. 2.1). However, we not only extend the result to a more general setting of stochastic linear–quadratic problems, but also show that solvability of the forward problem, i.e., that $(\bar{Q}, \bar{q}) \in \mathscr{F}(\bar{R})$, can be characterized in different, but equivalent, ways. The main result of this section is the following.

Theorem 3.1 (*Boundedness of Forward Problem*)**.** *Under Assumptions 2.3 and 2.6, the following statements are equivalent:*

(1) $(\bar{Q}, \bar{q}) \in \mathscr{F}(\bar{R})$.

(2) Let $\bar{P}_{1:\nu}$ and $\bar{\eta}_{1:\nu}$ be generated by the Riccati iterations

$$\bar{P}_\nu = \bar{Q}, \tag{3a}$$

$$\bar{P}_t = A^T\bar{P}_{t+1}A + \bar{Q} - A^T\bar{P}_{t+1}B(B^T\bar{P}_{t+1}B + \bar{R})^\dagger \times B^T\bar{P}_{t+1}A, \quad t = 1 : \nu - 1; \tag{3b}$$

$$\bar{\eta}_\nu = \bar{q}, \tag{3c}$$

$$\bar{\eta}_t = \left(A - B(B^T\bar{P}_{t+1}B + \bar{R})^\dagger B^T\bar{P}_{t+1}A\right)^T \times (\bar{\eta}_{t+1} + \bar{P}_{t+1}d) + \bar{q}, \quad t = 1 : \nu - 1. \tag{3d}$$

[1] It means that, given the system state's evolution from time step t to $t+1$, there is only one possible control input to realize that.

Denote

$$\bar{\mathfrak{S}}_t := B^T \bar{P}_{t+1} A, \tag{4a}$$

$$\bar{\mathfrak{R}}_t := B^T \bar{P}_{t+1} B + \bar{R}, \tag{4b}$$

$$\bar{g}_t := B^T \bar{\eta}_{t+1} + B^T \bar{P}_{t+1} d. \tag{4c}$$

For all $t = 1 : \nu - 1$, it holds that

$$\bar{\mathfrak{R}}_t \succeq 0, \tag{4d}$$

$$\ker(\bar{\mathfrak{R}}_t) \subset \left[\ker(\bar{\mathfrak{S}}_t^T) \cap \ker(\bar{g}_t^T)\right]. \tag{4e}$$

(3) *There exists $\{\bar{P}_t \in \mathbb{S}^n\}_{t=1:\nu}$, $\{\bar{\eta}_t \in \mathbb{R}^n\}_{t=1:\nu}$, and $\{\bar{\xi}_t \in \mathbb{R}\}_{t=1:\nu}$ such that*

$$\bar{P}_\nu = \bar{Q}, \quad \bar{\eta}_\nu = \bar{q}, \tag{5a}$$

$$\bar{H}_t := \begin{bmatrix} \bar{\mathfrak{R}}_t & \bar{\mathfrak{S}}_t & \bar{g}_t \\ \bar{\mathfrak{S}}_t^T & A^T \bar{P}_{t+1} A + \bar{Q} - \bar{P}_t & \bar{\beta}_t \\ \bar{g}_t^T & \bar{\beta}_t^T & \bar{\xi}_t \end{bmatrix} \succeq 0 \tag{5b}$$

$$\operatorname{rank}(\bar{H}_t) = \operatorname{rank}(\bar{\mathfrak{R}}_t) \tag{5c}$$

where $\bar{\beta}_t := \bar{q} + A^T \bar{P}_{t+1} d + A^T \bar{\eta}_{t+1} - \bar{\eta}_t$ and where $\bar{\mathfrak{S}}_t$, $\bar{\mathfrak{R}}_t$, and $\bar{g}_t$ are as in (4a), (4b), and (4c), respectively, for all $t = 1 : \nu - 1$.

(4) *The Hamilton–Jacobi–Bellman equation (HJBE)*

$$V_\nu(\chi_\nu) := \frac{1}{2} \chi_\nu^T \bar{Q} \chi_\nu + \bar{q}^T \chi_\nu, \tag{6a}$$

$$V_t(\chi_t) = \min_{\mu_t} \Big\{ \frac{1}{2} \chi_t^T \bar{Q} \chi_t + \bar{q}^T \chi_t + \frac{1}{2} \mu_t^T \bar{R} \mu_t \tag{6b}$$

$$+ \mathbb{E}_{\boldsymbol{w}_t} \left[V_{t+1}(A\chi_t + B\mu_t + d + \boldsymbol{w}_t) \right] \Big\}, \ t = 1 : \nu - 1$$

has a solution. More precisely, this means that $V_t(\chi_t)$ is bounded from below for any $\chi_t \in \mathbb{R}^n$, for $t = 1 : \nu - 1$. Moreover, the solution has the form

$$V_t(\chi_t) = \frac{1}{2} \chi_t^T \bar{P}_t \chi_t + \bar{\eta}_t^T \chi_t + \bar{\gamma}_t, \quad t = 1 : \nu, \tag{7}$$

where $\bar{P}_{1:\nu}$ and $\bar{\eta}_{1:\nu}$ are generated by (3) and

$$\bar{\gamma}_\nu = 0 \tag{8a}$$

$$\bar{\gamma}_t = \bar{\gamma}_{t+1} - \frac{1}{2} \bar{g}_t^T \bar{\mathfrak{R}}_t^\dagger \bar{g}_t + \frac{1}{2} d^T \bar{P}_{t+1} d + \bar{\eta}_{t+1}^T d + \frac{1}{2} \operatorname{tr}(\bar{P}_{t+1} \Sigma_w), \quad t = 1 : \nu - 1, \tag{8b}$$

with $\bar{\mathfrak{R}}_{1:\nu-1}$ as in (4b) and $\bar{g}_{1:\nu-1}$ as in (4c).

In addition, if any of the four above conditions hold, then the optimal control signal $\boldsymbol{u}_t$ for (1) is parametrized by any arbitrary vector $\lambda_t \in \mathbb{R}^n$, and is given by

$$\boldsymbol{u}_t = -\bar{\mathfrak{R}}_t^\dagger (\bar{\mathfrak{S}}_t \boldsymbol{x}_t + \bar{g}_t) + \mathcal{P}_t^{\ker(\bar{\mathfrak{R}})} \lambda_t, \quad t = \nu - N + 1 : \nu - 1, \tag{9a}$$

$$\boldsymbol{u}_t = 0, \quad t = 1 : \nu - N, \tag{9b}$$

where $\mathcal{P}_t^{\ker(\bar{\mathfrak{R}})} = I - \bar{\mathfrak{R}}_t^\dagger \bar{\mathfrak{R}}_t$ is the projection operator onto the kernel space of $\bar{\mathfrak{R}}_t$.

Proof. First, assume that (4) holds. Note that for an agent with a planning horizon realization $N = \nu$, by the principle of optimality in dynamic programming (see, e.g., Bertsekas (2000, p. 18)), we can easily show that (4) $\Longrightarrow$ (1) (cf. Bertsekas (2000, Prop. 1.3.1)). For the case $N < \nu$, note that the HJBE is still valid for $t = \nu - N + 1 : \nu$, and thus the agents behavior is still optimal in $t = \nu - N + 1 : \nu$. Since the systems behavior for $t = 1 : \nu - N$ is completely determined by (1c), and (1e), by Bellman's principle of optimality, the solution to the HJBE still gives the optimal control. This implies (1), and hence shows that (4) $\Longrightarrow$ (1).

Before we proceed, first note that for any $N \in \{2, \dots, \nu\}$,

$$0 = \sum_{t=\nu-N+1}^{\nu-1} \Big\{ \frac{1}{2} \boldsymbol{x}_{t+1}^T \bar{P}_{t+1} \boldsymbol{x}_{t+1} + \bar{\eta}_{t+1}^T \boldsymbol{x}_{t+1} - \frac{1}{2} \boldsymbol{x}_t^T \bar{P}_t \boldsymbol{x}_t - \bar{\eta}_t^T \boldsymbol{x}_t \Big\} + \frac{1}{2} \boldsymbol{x}_{\nu-N+1}^T \bar{P}_{\nu-N+1} \boldsymbol{x}_{\nu-N+1} + \bar{\eta}_{\nu-N+1}^T \boldsymbol{x}_{\nu-N+1} - \frac{1}{2} \boldsymbol{x}_\nu^T \bar{P}_\nu \boldsymbol{x}_\nu - \bar{\eta}_\nu^T \boldsymbol{x}_\nu.$$

Taking expectation with respect to $\boldsymbol{w}_{\nu-N+1:\nu-1}$ on both sides of the above equation, adding the latter expression to (1a), and using (1b), (3a), (3c), and Assumption 2.3, we can write the objective function as

$$\begin{aligned} J_N &= \mathop{\mathbb{E}}_{\boldsymbol{w}_{\nu-N+1:\nu-1}} \Big[\frac{1}{2} \boldsymbol{x}_\nu^T \underbrace{(\bar{Q} - \bar{P}_\nu)}_{=0} \boldsymbol{x}_\nu + \underbrace{(\bar{q} - \bar{\eta}_\nu)^T}_{=0} \boldsymbol{x}_\nu \\ &+ \sum_{t=\nu-N+1}^{\nu-1} \Big\{ \frac{1}{2} (A\boldsymbol{x}_t + B\boldsymbol{u}_t + d + \boldsymbol{w}_t)^T \bar{P}_{t+1} (A\boldsymbol{x}_t + B\boldsymbol{u}_t + d + \boldsymbol{w}_t) \\ &+ \bar{\eta}_{t+1}^T (A\boldsymbol{x}_t + B\boldsymbol{u}_t + d + \boldsymbol{w}_t) - \frac{1}{2} \boldsymbol{x}_t^T \bar{P}_t \boldsymbol{x}_t - \bar{\eta}_t^T \boldsymbol{x}_t + \frac{1}{2} \boldsymbol{x}_t^T \bar{Q} \boldsymbol{x}_t + \bar{q}^T \boldsymbol{x}_t \\ &+ \frac{1}{2} \boldsymbol{u}_t^T \bar{R} \boldsymbol{u}_t \Big\} + \frac{1}{2} \boldsymbol{x}_{\nu-N+1}^T \bar{P}_{\nu-N+1} \boldsymbol{x}_{\nu-N+1} + \bar{\eta}_{\nu-N+1}^T \boldsymbol{x}_{\nu-N+1} \Big] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{w}_{\nu-N+1:\nu-1}} \Big[\sum_{t=\nu-N+1}^{\nu-1} \Big\{ \frac{1}{2} \begin{bmatrix} \boldsymbol{u}_t^T & \boldsymbol{x}_t^T & 1 \end{bmatrix} \bar{H}_t \begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{x}_t \\ 1 \end{bmatrix} \Big\} \\ &+ \frac{1}{2} \boldsymbol{x}_{\nu-N+1}^T \bar{P}_{\nu-N+1} \boldsymbol{x}_{\nu-N+1} + \bar{\eta}_{\nu-N+1}^T \boldsymbol{x}_{\nu-N+1} \Big] \\ &+ \sum_{t=\nu-N+1}^{\nu-1} \frac{1}{2} d^T \bar{P}_{t+1} d + \frac{1}{2} \operatorname{tr}(\bar{P}_{t+1} \Sigma_w) + \bar{\eta}_{t+1}^T d - \frac{1}{2} \bar{\xi}_t \end{aligned} \tag{10}$$

with $\bar{H}_t$ in the form of (5b) and $\bar{\xi}_t = \bar{g}_t^T \bar{\mathfrak{R}}_t^\dagger \bar{g}_t$ in $\bar{H}_t$. Note that the last row in the above equation is constant with respect to the state and the control and hence can be discarded in the optimization problem. On the other hand, since $\{\bar{P}_t\}_{t=1}^\nu$ and $\{\bar{\eta}_t\}_{t=1}^\nu$ are generated by (3), $\bar{H}_t$ can also be written as

$$\bar{H}_t = \begin{bmatrix} \bar{\mathfrak{R}}_t & \bar{\mathfrak{S}}_t & \bar{g}_t \\ \bar{\mathfrak{S}}_t^T & \bar{\mathfrak{S}}_t^T \bar{\mathfrak{R}}_t^\dagger \bar{\mathfrak{S}}_t & \bar{\mathfrak{S}}_t^T \bar{\mathfrak{R}}_t^\dagger \bar{g}_t \\ \bar{g}_t^T & \bar{g}_t^T \bar{\mathfrak{R}}_t^\dagger \bar{\mathfrak{S}}_t & \bar{g}_t^T \bar{\mathfrak{R}}_t^\dagger \bar{g}_t \end{bmatrix} = \begin{bmatrix} \bar{\mathfrak{R}}_t \\ \bar{\mathfrak{S}}_t^T \\ \bar{g}_t^T \end{bmatrix} \bar{\mathfrak{R}}_t^\dagger \begin{bmatrix} \bar{\mathfrak{R}}_t & \bar{\mathfrak{S}}_t & \bar{g}_t \end{bmatrix}. \tag{11}$$

Next, we use the above trick to prove that (1) $\Longrightarrow$ (2). This is done by proving the contraposition of the statement. To this end, suppose (4d) and (4e) cease to hold at the Nth backward iteration (3), where $N \in \{2, \dots, \nu\}$. Namely, $\bar{\mathfrak{R}}_{\nu-N+t} \succeq 0$, $\ker(\bar{\mathfrak{R}}_{\nu-N+t}) \subset \left[\ker(\bar{\mathfrak{S}}_{\nu-N+t}^T) \cap \ker(\bar{g}_{\nu-N+t}^T)\right]$ still holds for $t = 2 : N$ but not for $t = 1$. We proceed by showing that this implies that for this planning horizon length realization N, (1) is not bounded from below. In particular, by (11) and $\ker(\bar{\mathfrak{R}}_{\nu-N+t}) \subset \left[\ker(\bar{\mathfrak{S}}_{\nu-N+t}^T) \cap \ker(\bar{g}_{\nu-N+t}^T)\right]$, $\forall t = 2 : N$, it follows that $\bar{H}_{\nu-N+t}$ can be written as in (11), which also implies that $\bar{H}_{\nu-N+t} \succeq 0$ for $t = 2 : N$. Hence, in view of (1d), (1c) and the fact that $\{\boldsymbol{w}_t\}_{t=1}^\infty$ is independent of other random elements from Assumption 2.3, for the given "initial state" realization $\boldsymbol{x}_{\nu-N+1} = \bar{x} \in \mathbb{R}^n$ from which the agent starts tracking, the objective function can be written as

$$\begin{aligned} J_N = \mathop{\mathbb{E}}_{\boldsymbol{w}_{\nu-N+1:\nu-1}} \Big[\frac{1}{2} \begin{bmatrix} \boldsymbol{u}_{\nu-N+1}^T & \boldsymbol{x}_{\nu-N+1}^T & 1 \end{bmatrix} \bar{H}_{\nu-N+1} \\ \times \begin{bmatrix} \boldsymbol{u}_{\nu-N+1} \\ \boldsymbol{x}_{\nu-N+1} \\ 1 \end{bmatrix} + \sum_{t=2}^{N-1} \Big\{ \frac{1}{2} \Big\| (\bar{\mathfrak{R}}_{\nu-N+t}^\dagger)^{\frac{1}{2}} \Big(\bar{\mathfrak{R}}_{\nu-N+t} \boldsymbol{u}_{\nu-N+t} \\ + \bar{\mathfrak{S}}_{\nu-N+t} \boldsymbol{x}_{\nu-N+t} + \bar{g}_{\nu-N+t} \Big) \Big\|^2 \Big\} \end{aligned}$$

$$+\frac{1}{2}\underbrace{\boldsymbol{x}_{\nu-N+1}^T}_{\bar{x}^T}\bar{P}_{\nu-N+1}\underbrace{\boldsymbol{x}_{\nu-N+1}}_{\bar{x}}+\bar{\eta}_{\nu-N+1}^T\underbrace{\boldsymbol{x}_{\nu-N+1}}_{\bar{x}}\Big]. \tag{12}$$

Notably, it holds that $\frac{1}{2}\bar{x}^T\bar{P}_{\nu-N+1}\bar{x}+\bar{\eta}_{\nu-N+1}^T\bar{x} \le \frac{1}{2}\sigma_{\max}(\bar{P}_{\nu-N+1})\|\bar{x}\|^2 + \|\bar{\eta}_{\nu-N+1}\|\cdot\|\bar{x}\| := \tau(\bar{x})$, where $\sigma_{\max}(\cdot)$ is the largest eigenvalue of a matrix. Also note that, for any $\boldsymbol{u}_{\nu-N+1}$ and $\boldsymbol{x}_{\nu-N+1} = \bar{x}$, by the dynamics (1b) we get an $\boldsymbol{x}_{\nu-N+2}$. Selecting $\boldsymbol{u}_{\nu-N+2} = -\bar{\mathfrak{R}}_{\nu-N+2}^\dagger(\bar{\mathfrak{S}}_{\nu-N+2}\boldsymbol{x}_{\nu-N+2}+\bar{g}_{\nu-N+2}) + \mathcal{P}_{\nu-N+2}^{\ker(\bar{\mathfrak{R}})}\lambda_{\nu-N+2}$, for some arbitrary $\lambda_{\nu-N+2}\in\mathbb{R}^n$, this would give the next state $\boldsymbol{x}_{\nu-N+3}$. Recursively selecting the other control signals for $t=\nu-N+3:\nu-1$ accordingly, i.e., as $\boldsymbol{u}_t = -\bar{\mathfrak{R}}_t^\dagger(\bar{\mathfrak{S}}_t\boldsymbol{x}_t+\bar{g}_t)+\mathcal{P}_t^{\ker(\bar{\mathfrak{R}})}\lambda_t$, for some arbitrary $\lambda_t\in\mathbb{R}^n$, then all terms in the summation in (12) would become zero. Therefore, for the objective function J_N it holds that

$$\begin{aligned}J_N \le \mathop{\mathbb{E}}_{\boldsymbol{w}_{\nu-N+1:\nu-1}}\Big[&\frac{1}{2}\boldsymbol{u}_{\nu-N+1}^T\bar{\mathfrak{R}}_{\nu-N+1}\boldsymbol{u}_{\nu-N+1}+\bar{x}^T\bar{\mathfrak{S}}_{\nu-N+1}^T\\ &\times\boldsymbol{u}_{\nu-N+1}+\bar{g}_{\nu-N+1}^T\boldsymbol{u}_{\nu-N+1}\Big]+\bar{x}^T\bar{\mathfrak{S}}_{\nu-N+1}^T\bar{\mathfrak{R}}_{\nu-N+1}^\dagger\\ &\times\bar{\mathfrak{S}}_{\nu-N+1}\bar{x}+\bar{g}_{\nu-N+1}^T\bar{\mathfrak{R}}_{\nu-N+1}^\dagger\bar{\mathfrak{S}}_{\nu-N+1}\bar{x}+\tau(\bar{x}).\end{aligned} \tag{13}$$

If $\bar{\mathfrak{R}}_{\nu-N+1}\not\succeq 0$, it is clear that we can choose $\boldsymbol{u}_{\nu-N+1} = \alpha v_-(\bar{\mathfrak{R}}_{\nu-N+1})$, where $v_-(\bar{\mathfrak{R}}_{\nu-N+1})$ is an eigenvector that corresponds to a negative eigenvalue of $\bar{\mathfrak{R}}_{\nu-N+1}$. In this case, since such choice of $\boldsymbol{u}_{\nu-N+1}$ does not depend on the random vectors $\boldsymbol{w}_{\nu-N+1:\nu-1}$, the expectation would be marginalized out. Letting $\alpha\to+\infty$ would make J_N tend to minus infinity. Thus, unless (4d) holds, irrespective of the initial condition $\bar{x}$ there is a planning horizon length realization $N\in\{2,\dots,\nu\}$ such that the cost function is not bounded from below.

On the other hand, if (4d) holds but $\ker(\bar{\mathfrak{R}}_{\nu-N+1}) \not\subset \big[\ker(\bar{\mathfrak{S}}_{\nu-N+1}^T)\cap\ker(\bar{g}_{\nu-N+1}^T)\big]$, then there exists a vector $v\in\mathbb{R}^m$ with norm one, such that $\bar{\mathfrak{R}}_{\nu-N+1}v = 0$, and such that $\bar{\mathfrak{S}}_{\nu-N+1}^T v\neq 0$ or $\bar{g}_{\nu-N+1}^T v\neq 0$. Without loss of generality, assume that $\bar{g}_{\nu-N+1}^T v\ge 0$; otherwise we instead consider $-v$. By Assumption 2.6, it follows that we can find $\rho>0$ such that $\mathbb{P}\left(\bar{\boldsymbol{x}}\in\mathscr{B}_\epsilon^n(\rho\bar{\mathfrak{S}}_{\nu-N+1}^T v)\right)>0, \forall\epsilon>0$. Now, consider $\boldsymbol{u}_{\nu-N+1}=\alpha v$ and $\boldsymbol{x}_{\nu-N+1}=\bar{x}=\rho\bar{\mathfrak{S}}_{\nu-N+1}^T v+\tilde{v}$, where $\tilde{v}\in\mathscr{B}_{\epsilon_1}^n(0)$ and where $\epsilon_1>0$ will be determined shortly. Note that with such choice, the expectation in (13) would be again marginalized out. Then it holds for the objective function that

$$\begin{aligned}J_N \le{}&\alpha[\rho v^T\bar{\mathfrak{S}}_{\nu-N+1}\bar{\mathfrak{S}}_{\nu-N+1}^T v+\bar{g}_{\nu-N+1}^T v+\tilde{v}^T\bar{\mathfrak{S}}_{\nu-N+1}^T v]\\ &+(\rho v^T\bar{\mathfrak{S}}_{\nu-N+1}+\tilde{v}^T)\bar{\mathfrak{S}}_{\nu-N+1}^T\bar{\mathfrak{R}}_{\nu-N+1}^\dagger\bar{\mathfrak{S}}_{\nu-N+1}\\ &\times(\rho\bar{\mathfrak{S}}_{\nu-N+1}^T v+\tilde{v})+\bar{g}_{\nu-N+1}^T\bar{\mathfrak{R}}_{\nu-N+1}^\dagger\bar{\mathfrak{S}}_{\nu-N+1}\\ &\times(\rho\bar{\mathfrak{S}}_{\nu-N+1}^T v+\tilde{v})+\tau(\rho\bar{\mathfrak{S}}_{\nu-N+1}^T v+\tilde{v}).\end{aligned}$$

Since $\rho v^T\bar{\mathfrak{S}}_{\nu-N+1}\bar{\mathfrak{S}}_{\nu-N+1}^T v+\bar{g}_{\nu-N+1}^T v = \rho\|\bar{\mathfrak{S}}_{\nu-N+1}^T v\|^2+\bar{g}_{\nu-N+1}^T v > 0$, we can always make $\epsilon_1>0$ small enough so that $\rho\|\bar{\mathfrak{S}}_{\nu-N+1}^T v\|^2+\bar{g}_{\nu-N+1}^T v+\tilde{v}^T\bar{\mathfrak{S}}_{\nu-N+1}^T v>0, \forall\tilde{v}\in\mathscr{B}_{\epsilon_1}^n(0)$. Letting $\alpha\to-\infty$ would make J_N tend to minus infinity. Recalling that $\mathbb{P}(\bar{\boldsymbol{x}}\in\mathscr{B}_\epsilon^n(\rho\bar{\mathfrak{S}}_{\nu-N+1}^T v))>0, \forall\epsilon\in(0,\epsilon_1)$, this shows that there exists a set of initial value realizations $\bar{x}$ with non-zero probability such that the forward problem is ill-posed. This proves that (1) $\Longrightarrow$ (2).

To prove that (2) $\Longrightarrow$ (4), we make an ansatz that the solution to the HJBE (6) is $V_t(\chi_t) = \frac{1}{2}\chi_t^T\bar{P}_t\chi_t+\bar{\eta}_t^T\chi_t+\bar{\gamma}_t$, with terms generated by (3) and (8), respectively. This ansatz fulfills (6a) and (8), and plugging it into the left hand side of (6b) we get

$$\begin{aligned}\min_{\mu_t}\Big\{&\frac{1}{2}\chi_t^T\bar{Q}\chi_t+\bar{q}^T\chi_t+\frac{1}{2}\mu_t^T\bar{R}\mu_t+\mathop{\mathbb{E}}_{\boldsymbol{w}_t}\Big[\frac{1}{2}(A\chi_t+B\mu_t\\ &+d+\boldsymbol{w}_t)^T\bar{P}_{t+1}(A\chi_t+B\mu_t+d+\boldsymbol{w}_t)\\ &+\bar{\eta}_{t+1}^T(A\chi_t+B\mu_t+d+\boldsymbol{w}_t)+\bar{\gamma}_{t+1}\Big]\Big\},\end{aligned}$$

for $t=1:\nu-1$. By Assumption 2.3, we can expand the expectation regarding $\boldsymbol{w}_t$, and removing constant terms that are irrelevant to the optimization this gives

$$\begin{aligned}\min_{\mu_t}\Big\{&\frac{1}{2}\mu_t^T\bar{\mathfrak{R}}_t\mu_t+(\bar{\mathfrak{S}}_t\chi_t+\bar{g}_t)^T\mu_t+(A\chi_t+d)^T\bar{P}_{t+1}\\ &\times(A\chi_t+d)+\eta_{t+1}^T(A\chi_t+d)+\frac{1}{2}\chi_t^T\bar{Q}\chi_t+\bar{q}^T\chi_t\Big\},\end{aligned} \tag{14}$$

which is an unconstrained quadratic optimization problem with respect to μ_t. By (4d) we know that it is a convex problem, and hence it has an optimal solution if and only if the gradient is zero in some point. To verify that it has a solution, and to work out the optimal control, we take the derivative of (14) with respect to μ_t and equate it to zero, which gives $\bar{\mathfrak{R}}_t\mu_t = -\bar{\mathfrak{S}}_t\chi_t-\bar{g}_t$, $t=1:\nu-1$. Since (4e) holds, we have that $\bar{\mathfrak{S}}_t\chi_t+\bar{g}_t\in\mathrm{Im}(\bar{\mathfrak{S}}_t)\oplus\mathrm{Im}(\bar{g}_t) = \left[\ker(\bar{\mathfrak{S}}_t^T)\cap\ker(\bar{g}_t^T)\right]^\perp \subset \ker(\bar{\mathfrak{R}}_t)^\perp = \mathrm{Im}(\bar{\mathfrak{R}}_t)$, where $\oplus$ denotes the direct sum of the subspaces. Hence, the above equation has a solution, and the control signal takes the form of $\mu_t = -\bar{\mathfrak{R}}_t^\dagger(\bar{\mathfrak{S}}_t\chi_t+\bar{g}_t)+\mathcal{P}_t^{\ker(\bar{\mathfrak{R}})}\lambda_t, \forall\lambda_t\in\mathbb{R}^n, t=1:\nu-1$, which minimizes the right hand side of (14). Plug the aforementioned equation into (14), use the property of (4e) and in view of (3), (8), we have that the quadratic, first order and constant terms regarding χ_t equates between the left and right hand sides. Thus the ansatz is indeed a solution to the HJBE.

To complete the proof, we now show the equivalence between (2) and (3). First, we prove (2) $\Longrightarrow$ (3), and start with noting that if (3) holds, then (5a) holds trivially. Next, with $\bar{\xi}_t = \bar{g}_t^T\bar{\mathfrak{R}}_t^\dagger\bar{g}_t$, we know from the above argument that due to the kernel containment (4e), $\bar{H}_t$ can be expressed as in (11), for $t=1:\nu-1$. By (4d), $\bar{\mathfrak{R}}_t^\dagger\succeq 0$ and hence $\bar{H}_t\succeq 0$, i.e., (5b) holds. On the other hand, by the rank property of Schur complement (Horn & Zhang, 2005, p. 43), it holds that $\mathrm{rank}(\bar{H}_t) = \mathrm{rank}(\bar{\mathfrak{R}}_t)+\mathrm{rank}(\bar{H}_t\backslash\bar{\mathfrak{R}}_t)$. Now, observe that

$$\bar{H}_t\backslash\bar{\mathfrak{R}}_t = \begin{bmatrix}\bar{\mathfrak{S}}_t^T\bar{\mathfrak{R}}_t^\dagger\bar{\mathfrak{S}}_t & \bar{\mathfrak{S}}_t^T\bar{\mathfrak{R}}_t^\dagger\bar{g}_t\\ \bar{g}_t^T\bar{\mathfrak{R}}_t^\dagger\bar{\mathfrak{S}}_t & \bar{g}_t^T\bar{\mathfrak{R}}_t^\dagger\bar{g}_t\end{bmatrix}-\begin{bmatrix}\bar{\mathfrak{S}}_t^T\\ \bar{g}_t^T\end{bmatrix}\bar{\mathfrak{R}}_t^\dagger\begin{bmatrix}\bar{\mathfrak{S}}_t & \bar{g}_t\end{bmatrix}=0,$$

and hence $\mathrm{rank}(\bar{H}_t)=\mathrm{rank}(\bar{\mathfrak{R}}_t)$, i.e., (5c) holds.

Now we prove (3) $\Longrightarrow$ (2). If (5) holds, it follows that $\bar{\mathfrak{R}}_t\succeq 0$, i.e., (4d) holds. By properties of the generalized Schur complement (Horn & Zhang, 2005, Thm. 1.20 and p. 43), it follows that $(I-\bar{\mathfrak{R}}_t\bar{\mathfrak{R}}_t^\dagger)\begin{bmatrix}\bar{\mathfrak{S}}_t & \bar{g}_t\end{bmatrix}=0$, which implies (4e), that

$$\bar{H}_t\backslash\bar{\mathfrak{R}}_t := \begin{bmatrix}A^T\bar{P}_{t+1}A+\bar{Q}-\bar{P}_t & \bar{\beta}_t\\ \beta_t & \bar{\xi}_t\end{bmatrix}-\begin{bmatrix}\bar{\mathfrak{S}}_t^T\\ \bar{g}_t^T\end{bmatrix}\bar{\mathfrak{R}}_t^\dagger\begin{bmatrix}\bar{\mathfrak{S}}_t & \bar{g}_t\end{bmatrix}\succeq 0,$$

and that $\mathrm{rank}(\bar{H}_t)=\mathrm{rank}(\bar{\mathfrak{R}}_t)+\mathrm{rank}(\bar{H}_t\backslash\bar{\mathfrak{R}}_t)$. Since $\mathrm{rank}(\bar{H}_t)=\mathrm{rank}(\bar{\mathfrak{R}}_t)$, we must have that $\bar{H}_t\backslash\bar{\mathfrak{R}}_t=0$, and hence (3) follows. This finishes the proof of the entire theorem. □

Remark 3.2. Note that in the equivalence between point (2) and (3) in Theorem 3.1, we have that (5a) and (5c) are equivalent to that the Riccati recursions in (3) are satisfied, and that (5b), i.e., that $\bar{H}_t\succeq 0$ holds, is equivalent to (4d) and (4e).

Remark 3.3. By extending the state space to $\tilde{\boldsymbol{x}}_t = [\boldsymbol{x}_t^T, 1]^T$ and with state space matrices given by

$$\tilde{A}=\begin{bmatrix}A & d\\ \boldsymbol{0}^T & 1\end{bmatrix},\ \tilde{B}=\begin{bmatrix}B\\ 0\end{bmatrix},\ \tilde{\bar{Q}}=\begin{bmatrix}\bar{Q} & \bar{q}\\ \bar{q}^T & 0\end{bmatrix},\ \tilde{\bar{R}}=\bar{R}, \tag{15}$$

we can indeed rewrite the forward problem (1) as a "standard" LQ problem (albeit possibly indefinite). However, note that this does not lead to a classic LQR in the sense that $(\tilde{A},\tilde{B})$ is not controllable. Moreover, $\tilde{\bar{\boldsymbol{x}}}=[\bar{x}^T,1]^T$ does not satisfy Assumption 2.6. Furthermore, extending such a rewriting to time-varying d_t or the

tracking problem, where $q_t = Qx_t^r$ and x_t^r is the reference signal, is not possible without specifying the problem structure. Therefore, existing IOC results cannot be directly applied to the rewritten problem. Nevertheless, the form (15) is still useful in some of the analysis.

3.2. Analysis of the closed-loop system matrices

In view of (9), it seems like there might be infinitely many choices of control signals that are optimal. However, as we shall see, under the assumptions we impose the optimal control signal is *unique* for the considered forward problem (1).

Proposition 3.4. *Let $\bar{R} \succ 0$. Under Assumptions 2.2, 2.3 and 2.6, $(\bar{Q}, \bar{q}) \in \mathscr{F}(\bar{R})$ is equivalent to $\bar{\mathfrak{R}}_t \succ 0$, $t = 1 : \nu - 1$.*

Proof. The implication "$\Longleftarrow$" follows from Theorem 3.1. To prove the implications "$\Longrightarrow$": by Theorem 3.1, since $(\bar{Q}, \bar{q}) \in \mathscr{F}(\bar{R})$, there exist matrices and vectors that fulfill (3) and (4). In particular, since A is invertible, by (4a) and (4b) we have that $\bar{\mathfrak{R}}_t = B^T\bar{P}_{t+1}B + \bar{R} = B^T\bar{P}_{t+1}AA^{-1}B + \bar{R} = \bar{\mathfrak{S}}_tA^{-1}B + \bar{R}$, for $t = 1 : \nu - 1$. Now, by (4e) we have that $\ker(\bar{\mathfrak{R}}_t) \subset \ker(\bar{\mathfrak{S}}_t^T)$ holds for $t = 1 : \nu - 1$. In particular, this means that

$$z \in \ker(\bar{\mathfrak{R}}_t) \implies \begin{cases} \bar{\mathfrak{R}}_t z = 0 \\ \bar{\mathfrak{S}}_t^T z = 0 \end{cases} \implies \begin{cases} -\bar{\mathfrak{S}}_tA^{-1}Bz = \bar{R}z \\ z^T\bar{\mathfrak{S}}_t = 0. \end{cases}$$

This in turn means that for any $z \in \ker(\bar{\mathfrak{R}}_t)$, it holds that $z^T\bar{R}z = -z^T\bar{\mathfrak{S}}_tA^{-1}Bz = 0$, and since $\bar{R} \succ 0$ this means that $z = 0$. Therefore, the only vector in $\ker(\bar{\mathfrak{R}}_t)$ is the zero-vector, and since $\bar{\mathfrak{R}}_t$ is positive semi-definite (see (4d)), this implies that $\bar{\mathfrak{R}}_t$ is in fact strictly positive definite for $t = 1 : \nu - 1$. □

Corollary 3.5. *Under the assumptions in Proposition 3.4, the optimal control signal for problem (1) takes the form*

$$\boldsymbol{u}_t = -\bar{\mathfrak{R}}_t^{-1}(\bar{\mathfrak{S}}_t\boldsymbol{x}_t + \bar{\mathfrak{g}}_t), \quad t = \nu - N + 1 : \nu - 1 \tag{16a}$$

$$\boldsymbol{u}_t = 0, \quad t = 1 : \nu - N. \tag{16b}$$

Remark 3.6. Even if $\bar{R}$ is not strictly positive definite, by the proof of Proposition 3.4 we can still partially characterize $\ker(\bar{\mathfrak{R}}_t)$. In particular, $z \in \ker(\bar{\mathfrak{R}}_t)$ implies that $z^T\bar{R}z = 0$. If $\bar{R}$ is full rank, this is the neutral subspace in the indefinite inner product space defined by $\bar{R}$; see, e.g., Gohberg, Lancaster, and Rodman (2005, Chp. 2).

By Corollary 3.5, under the conditions in Proposition 3.4 there is a unique solution to the forward optimal control problem (1). In this case, the system's behavior (after it starts applying control) is determined by $\tilde{\boldsymbol{x}}_{t+1} = \tilde{A}_{cl}(t; \bar{Q}, \bar{q})\tilde{\boldsymbol{x}}_t + \tilde{\boldsymbol{w}}_t$, for $t = \nu - N + 1 : \nu$, where $\tilde{A}_{cl}(t; \bar{Q}, \bar{q})$ is the closed-loop system matrix at time t for the extended state-space model (see Remark 3.3). In particular,

$$\tilde{A}_{cl}(t; \bar{Q}, \bar{q}) = \begin{bmatrix} A - B\bar{\mathfrak{R}}_t^{-1}\bar{\mathfrak{S}}_t & d - B\bar{\mathfrak{R}}_t^{-1}\bar{\mathfrak{g}}_t \\ \mathbf{0}^T & 1 \end{bmatrix}, \tag{17}$$

with $\bar{\mathfrak{R}}_t$, $\bar{\mathfrak{S}}_t$, and $\bar{\mathfrak{g}}_t$ as in (4), and hence it implicitly depends on $(\bar{Q}, \bar{q})$ in the objective function (1a). This means that the (conditional) distribution of the agent's optimal trajectory $\mathbb{P}(\boldsymbol{x}_{1:\nu} \mid \boldsymbol{N} = N, \bar{\boldsymbol{x}} = \bar{x})$ and optimal control $\mathbb{P}(\boldsymbol{u}_{1:\nu-1} \mid \boldsymbol{N} = N, \bar{\boldsymbol{x}} = \bar{x})$ are implicitly given by solving (1). Moreover, under mild regularity conditions on the probability distribution of $(\bar{\boldsymbol{x}}, \boldsymbol{N})$, the formulation in (1) then defines joint probability distributions for $(\boldsymbol{x}_{1:\nu}, \boldsymbol{N}, \bar{\boldsymbol{x}})$ and $(\boldsymbol{u}_{1:\nu-1}, \boldsymbol{N}, \bar{\boldsymbol{x}})$ (cf. Kallenberg (1997, Thm. 5.3)). Before we continue analyzing the identifiability, we present the following corollary that is useful in the analysis to come.

Corollary 3.7. *Under the assumptions in Proposition 3.4, given any $R \succ 0$, for any $(Q, q) \in \mathscr{F}(R)$, let $\{P_t\}_{t=1}^{\nu}$ be the solution to (3) that corresponds to Q. Accordingly, let $\mathfrak{R}_t$, $\mathfrak{S}_t$ and $\mathfrak{g}_t$ be defined as in Theorem 3.1, for $t = 1 : \nu - 1$. Then the matrix $A_{cl}(t; Q) := A - B\mathfrak{R}_t^{-1}\mathfrak{S}_t$, as well as the matrix $\tilde{A}_{cl}(t; Q, q)$ in (17), are invertible for all $t = 1 : \nu - 1$.*

Proof. The fact that the matrix $A_{cl}(t; Q)$ is invertible for $t = 1 : \nu - 1$ follows by an argument similar to the proof of Zhang et al. (2019, Thm. 2.1), since $R \succ 0$ and $\mathfrak{R}_t \succ 0$ holds for $t = 1 : \nu - 1$. To show that $\tilde{A}_{cl}(t; Q, q)$ has full rank, we simply note that it has the upper block-triangular form (17), and since $A_{cl}(t; Q)$ has full rank so does $\tilde{A}_{cl}(t; Q, q)$. □

4. Identifiability analysis and persistent excitation

Next, we investigate the inverse problem of recovering $(\bar{Q}, \bar{q})$ from observations of optimal trajectories to problem (1). Throughout the rest we will therefore, unless explicitly stated otherwise, assume that $\bar{R} = I$ and $(\bar{Q}, \bar{q}) \in \mathscr{F}(\bar{R} = I)$. We start by considering the identifiability of the problem.

To this end, first note that from the analysis in Section 3, for any parameters $(Q, q) \in \mathscr{F}(I)$, the agent's behavior is completely determined by the time-varying closed-loop system matrices $\tilde{A}_{cl}(t; Q, q)$ in (17). In the spirit of Ljung and Glad (1994, Sec. 5.1), we can thus see the sequence of closed-loop system matrices $\{\tilde{A}_{cl}(t; Q, q)\}_{t=1}^{\nu-1}$ as the model structure. Therefore, the fundamental question for identifiability is if there exist two different sets of parameters (Q, q) and (Q', q') such that $\tilde{A}_{cl}(t; Q, q) = \tilde{A}_{cl}(t; Q', q')$ for all $t = 1 : \nu - 1$.

Proposition 4.1 (*Identifiability*)**.** *Under Assumptions 2.2–2.4 and 2.6, given $(Q, q), (Q', q') \in \mathscr{F}(I)$, if $\tilde{A}_{cl}(t; Q, q) = \tilde{A}_{cl}(t; Q', q')$ for all $t = 1 : \nu - 1$, then $(Q, q) = (Q', q')$.*

Proof. Assume that $\tilde{A}_{cl}(t; Q, q) = \tilde{A}_{cl}(t; Q', q')$ for $t = 1 : \nu - 1$, and let $(P_k, \eta_k, \mathfrak{S}_k, \mathfrak{R}_k, g_k)_{k=\nu-N+1}^{\nu-1}$ and $(P_k', \eta_k', \mathfrak{S}_k', \mathfrak{R}_k', g_k')_{k=\nu-N+1}^{\nu-1}$ be the solutions to (3) and (4) for (Q, q) and (Q', q'), respectively. Moreover, let $Q' = Q + \Delta Q$, $q' = q + \Delta q$, $P_t' = P_t + \Delta P_t$, $\eta_t' = \eta_t + \Delta\eta_t$, $\mathfrak{S}_t' = \mathfrak{S}_t + \Delta\mathfrak{S}_t$, $\mathfrak{R}_t' = \mathfrak{R}_t + \Delta\mathfrak{R}_t$, and $g_t' = g_t + \Delta g_t$. Since $\tilde{A}_{cl}(t; Q, q) = \tilde{A}_{cl}(t; Q', q')$, for $t = 1 : \nu - 1$, it follows that $A_{cl}(t; Q) = A_{cl}(t; Q')$, for $t = 1 : \nu - 1$. Then following the line of arguments in Zhang et al. (2019, Thm. 2.1), we can conclude that $\Delta Q = 0$. This in turn implies that ΔP_t, $\Delta\mathfrak{R}_t$ and $\Delta\mathfrak{S}_t$ are all zero (cf. (3) and (4)).

Next, since $\tilde{A}_{cl}(t; Q, q) = \tilde{A}_{cl}(t; Q', q')$ and $\Delta\mathfrak{R}_t = 0$, for $t = 1 : \nu - 1$, it follows that $d - B\mathfrak{R}_t^{-1}g_t = d - B\mathfrak{R}_t^{-1}g_t'$ for $t = 1 : \nu - 1$. Therefore, it holds that $B\mathfrak{R}_t^{-1}\Delta g_t = 0$ for $t = 1 : \nu - 1$. Since B is full column rank by assumption, and since $\mathfrak{R}_t^{-1} \succ 0$ by Proposition 3.4, we must have that $\Delta g_t = 0$ for $t = 1 : \nu - 1$. In view of (4c), we therefore have that $B^T\Delta\eta_{t+1} = \Delta g_t = 0$, $t = 1 : \nu - 1$. Then, in view of (3c) and (3d), this in turn implies that $\Delta\eta_\nu = \Delta q$, $\Delta\eta_t = A^T\Delta\eta_{t+1} + \Delta q$, $t = 1 : \nu - 1$. Thus, we have that $\Delta\eta_\nu = \Delta q$, which means that $B^T\Delta\eta_\nu = B^T\Delta q = 0$. Continuing, we have that $B^T\Delta\eta_{\nu-1} = B^T(A^T\Delta\eta_\nu + \Delta q) = B^TA^T\Delta q + B^T\Delta q = B^TA^T\Delta q = 0$. Then following the line of arguments in Zhang et al. (2019, Thm. 2.1), we can conclude that $\Delta q = 0$. The fact that $\Delta Q = 0$, $\Delta q = 0$ implies that $Q = Q'$, $q = q'$. □

This means that the parameters (Q, q) that characterizes the closed-loop system matrices are identifiable. Moreover, in view of (1), we can see $(\bar{\boldsymbol{x}}, \boldsymbol{N})$ as the "input" of the model and $\boldsymbol{y}_{1:\nu}$ as the "output". To this end, in order to uniquely identify the parameters $(\bar{Q}, \bar{q})$, the "input" $(\bar{\boldsymbol{x}}, \boldsymbol{N})$ needs to be "persistently exciting" (Ljung & Glad, 1994, Sec. 5.1). Notably, Assumption 2.4

gives a persistent excitation condition regarding $\boldsymbol{N}$. Moreover, Assumption 2.6 turns out to give a persistent excitation condition for the initial value $\bar{\boldsymbol{x}}$. In fact, we have the following result, the proof of which we defer to the Appendix.

Lemma 4.2. *Let $(\bar{\boldsymbol{x}}, \boldsymbol{N})$ be as in Assumption 2.6. Then, for all $N \in \{2, \dots, \nu\}$ such that $\mathbb{P}(\boldsymbol{N} = N) > 0$, $\operatorname{cov}_{\bar{\boldsymbol{x}}|\boldsymbol{N}=N}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}) \succ 0$.*

This result can now be used to prove the following Lemma, which is useful in the IOC algorithm construction to come. Similarly, the proof of this Lemma is also deferred to the Appendix.

Lemma 4.3 (*Persistent Excitation*)**.** *Suppose that $(\bar{Q}, \bar{q}) \in \mathscr{F}(I)$ and let $\tilde{\boldsymbol{x}}_t := [\boldsymbol{x}_t^T, 1]^T$. Under Assumptions 2.2–2.4 and 2.6, it holds that $\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=\nu}[\tilde{\boldsymbol{x}}_t \tilde{\boldsymbol{x}}_t^T] \succ 0$, and $\mathbb{E}[\|\tilde{\boldsymbol{x}}_t\|^2] < \infty$ for all $t = 1 : \nu$.*

5. The IOC algorithm

In this section, we construct the IOC algorithm for general linear–quadratic systems with different time-horizon lengths. In particular, we show that the algorithm is statistically consistent, i.e., that it converges in probability to the true underlying parameter. For the sake of brevity, in some of the following we sometimes use the notation $(\cdot)$ for the arguments of some functions.

In order to construct the IOC algorithm, we further make the following assumption.

Assumption 5.1 (*Bounded Parameters*)**.** The parameter tuple $(\bar{Q}, \bar{q})$ that governs the agents tracking behavior lies in the compact set

$$\mathbb{G}(\varphi) := \left\{ (\bar{Q} \in \mathbb{S}^n, \bar{q} \in \mathbb{R}^n) \mid \| \begin{bmatrix} \bar{Q} & \bar{q} \\ \bar{q} & 0 \end{bmatrix} \|_F \leq \varphi \right\},$$

for some (potentially unknown) $0 < \varphi < \infty$.

This assumption is mild, since when we solve the corresponding inverse problem in practice, we can always set φ arbitrary large if we have no prior knowledge on the norm bound of the parameters.

5.1. Construction and empirical approximation

To this end, the algorithm is constructed based on the necessary and sufficient optimality conditions in Theorem 3.1. More precisely, the IOC algorithm will be built upon an optimization problem which is constructed so that it has a unique optimal solution $(Q^\star, q^\star)$ which is the "true" $(\bar{Q}, \bar{q}) \in \mathscr{F}(I)$.

First, we construct the objective function for the optimization problem. Let $(Q, q) \in \mathscr{F}(I)$, and let $\bar{\mu}_t$ be the optimal control signal to $(\bar{Q}, \bar{q})$. Given a realization of the planning horizon N, it holds for all state $\chi_t \in \mathbb{R}^n$ that

$$\begin{aligned} 0 &= \min_{\mu_t} \Big\{ \frac{1}{2}\chi_t^T Q \chi_t + q^T \chi_t + \frac{1}{2}\|\mu_t\|^2 \\ &+ \mathop{\mathbb{E}}_{\boldsymbol{w}_t}[V_{t+1}(A\chi_t + B\mu_t + d + \boldsymbol{w}_t)] \Big\} - V_t(\chi_t) \\ &\leq \frac{1}{2}\chi_t^T Q \chi_t + q^T \chi_t + \frac{1}{2}\|\bar{\mu}_t\|^2 + \mathop{\mathbb{E}}_{\boldsymbol{w}_t}[V_{t+1}(A\chi_t + B\bar{\mu}_t + d \\ &+ \boldsymbol{w}_t)] - V_t(\chi_t), \end{aligned}$$

where the inequality follows since $\bar{\mu}_t$ is not necessarily optimal to $(Q, q, \bar{R} = I)$, and where $V_t(\cdot)$ has the form (7), and $P_{t:t+1}$, $\eta_{t:t+1}$, and $\gamma_{t:t+1}$ in $V_t(\cdot)$ are determined by (Q, q) via (3). Moreover, seen intuitively from the other perspective, for given χ_t and $\bar{\mu}_t$, we expect the inequality to hold unless we plug in $(Q, q, \bar{R} = I)$ which renders the state χ_t and control $\bar{\mu}_t$ optimal. We hence define the "violation" of HJBE at each time step $t = \nu - N + 1 : \nu - 1$ by

$$\begin{aligned} \psi_{t,N}(Q, q; \chi_t, \mu_t) &:= \mathop{\mathbb{E}}_{\boldsymbol{w}_t}[V_{t+1}(A\chi_t + B\mu_t + d + \boldsymbol{w}_t)] \\ &+ \frac{1}{2}\chi_t^T Q \chi_t + q^T \chi_t + \frac{1}{2}\|\mu_t\|_2^2 - V_t(\chi_t), \end{aligned}$$

since we expect that $\psi_{t,N}(Q, q; \chi_t, \mu_t) \geq 0$. The latter will be formally proved in Theorem 5.2.

Plugging (7) and (8) in to the above equation, we have

$$\begin{aligned} \psi_{t,N}(Q, q; \chi_t, \mu_t) &= \mathop{\mathbb{E}}_{\boldsymbol{w}_t}\Big[\frac{1}{2}(A\chi_t + B\mu_t + d + \boldsymbol{w}_t)^T P_{t+1} \\ &\times (A\chi_t + B\mu_t + d + \boldsymbol{w}_t) + \eta_{t+1}^T (A\chi_t + B\mu_t + d + \boldsymbol{w}_t)\Big] \\ &+ \frac{1}{2}\chi_t^T Q \chi_t + q^T \chi_t + \frac{1}{2}\|\mu_t\|^2 - \frac{1}{2}\chi_t^T P_t \chi_t - \eta_t^T \chi_t \\ &+ \frac{1}{2} g_t^T \mathfrak{R}_t^\dagger g_t - \frac{1}{2} d^T P_{t+1} d - \eta_t^T d - \frac{1}{2}\operatorname{tr}(P_{t+1}\Sigma_w). \end{aligned} \tag{18}$$

Given a realization of the planning horizon N, let $\boldsymbol{x}_{\nu-N+1:\nu}$ and $\boldsymbol{u}_{\nu-N+1:\nu-1}$ be the optimal trajectory and control. We let $\chi_t = \boldsymbol{x}_t$ and $\mu_t = \boldsymbol{u}_t$, and take the expectation of $\psi_{t,N}(Q, q; \boldsymbol{x}_t, \boldsymbol{u}_t)$ with respect to $\boldsymbol{x}_t|\boldsymbol{N} = N$. In view of (1b), this gives

$$\begin{aligned} &\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\psi_{t,N}(Q, q; \boldsymbol{x}_t, \boldsymbol{u}_t)] = \\ &\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\mathop{\mathbb{E}}_{\boldsymbol{w}_t}\Big[\frac{1}{2}\underbrace{(A\boldsymbol{x}_t + B\boldsymbol{u}_t + d + \boldsymbol{w}_t)^T}_{\boldsymbol{x}_{t+1}} P_{t+1} \underbrace{(A\boldsymbol{x}_t + B\boldsymbol{u}_t + d + \boldsymbol{w}_t)}_{\boldsymbol{x}_{t+1}} \\ &+ \eta_{t+1}^T \underbrace{(A\boldsymbol{x}_t + B\boldsymbol{u}_t + d + \boldsymbol{w}_t)}_{\boldsymbol{x}_{t+1}}\Big] \\ &+ \frac{1}{2}\boldsymbol{x}_t^T Q \boldsymbol{x}_t + q^T \boldsymbol{x}_t + \frac{1}{2}\|\boldsymbol{u}_t\|^2 - \frac{1}{2}\boldsymbol{x}_t^T P_t \boldsymbol{x}_t - \eta_t^T \boldsymbol{x}_t \\ &+ \frac{1}{2} g_t^T \mathfrak{R}_t^\dagger g_t - \frac{1}{2} d^T P_{t+1} d - \eta_t^T d - \frac{1}{2}\operatorname{tr}(P_{t+1}\Sigma_w)\Big] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{x}_{t+1}|\boldsymbol{N}=N}\Big[\frac{1}{2}\boldsymbol{x}_{t+1}^T P_{t+1} \boldsymbol{x}_{t+1} + \eta_{t+1}^T \boldsymbol{x}_{t+1}\Big] \\ &+ \mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\frac{1}{2}\boldsymbol{x}_t^T Q \boldsymbol{x}_t + q^T \boldsymbol{x}_t + \frac{1}{2}\|\boldsymbol{u}_t\|^2 - \frac{1}{2}\boldsymbol{x}_t^T P_t \boldsymbol{x}_t - \eta_t^T \boldsymbol{x}_t\Big] \\ &+ \frac{1}{2} g_t^T \mathfrak{R}_t^\dagger g_t - \frac{1}{2} d^T P_{t+1} d - \eta_t^T d - \frac{1}{2}\operatorname{tr}(P_{t+1}\Sigma_w). \end{aligned} \tag{19}$$

However, the above expression is constructed based on $\boldsymbol{x}_t$, while the observations are in terms of $\boldsymbol{y}_t$. To rewrite it in terms of $\boldsymbol{y}_t$, first we simply add and subtract some terms in the expression above:

$$\begin{aligned} \mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\psi_{t,N}(Q, q; \boldsymbol{x}_t, \boldsymbol{u}_t)] &= \mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\psi_{t,N}(Q, q; \boldsymbol{x}_t, \boldsymbol{u}_t)] \\ &+ \frac{1}{2}\operatorname{tr}(P_{t+1}\Sigma_v) - \frac{1}{2}\operatorname{tr}(P_t\Sigma_v) + \frac{1}{2}\operatorname{tr}(Q\Sigma_v) \\ &- \frac{1}{2}\operatorname{tr}(P_{t+1}\Sigma_v) + \frac{1}{2}\operatorname{tr}(P_t\Sigma_v) - \frac{1}{2}\operatorname{tr}(Q\Sigma_v). \end{aligned} \tag{20}$$

On the other hand, by Assumption 2.3, $\{\boldsymbol{v}_t\}_{t=1}^\infty$ are independent of any other stochastic elements. Using the cyclic permutation property of the matrix trace operator, we know that $\mathbb{E}_{\boldsymbol{v}_t}[\boldsymbol{v}_t^T P_{t+1} \boldsymbol{v}_t] = \mathbb{E}_{\boldsymbol{v}_t}[\operatorname{tr}(\boldsymbol{v}_t^T P_{t+1} \boldsymbol{v}_t)] = \mathbb{E}_{\boldsymbol{v}_t}[\operatorname{tr}(P_{t+1}\boldsymbol{v}_t\boldsymbol{v}_t^T)] = \operatorname{tr}(P_{t+1}\Sigma_v)$. Similarly, we also have $\mathbb{E}_{\boldsymbol{v}_t}[\boldsymbol{v}_t^T P_t \boldsymbol{v}_t] = \operatorname{tr}(P_t\Sigma_v)$, $\mathbb{E}_{\boldsymbol{v}_{t+1}}[\boldsymbol{v}_t^T Q \boldsymbol{v}_t] = \operatorname{tr}(Q\Sigma_v)$, $d^T P_{t+1} d = \operatorname{tr}(P_{t+1} d d^T)$. In view of (8), (2) and the fact that $\mathbb{E}_{\boldsymbol{w}_t}[\boldsymbol{w}_t] = 0$, $\mathbb{E}_{\boldsymbol{v}_t}[\boldsymbol{v}_t] = 0$, using (19) and (20) we can rewrite

$\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\psi_{t,N}(Q,q;\boldsymbol{x}_t,\boldsymbol{u}_t)]$ as

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{v}_{t:t+1}}\Big[\mathop{\mathbb{E}}_{\boldsymbol{x}_{t+1}|\boldsymbol{N}=N}\Big[\frac{1}{2}\underbrace{(\boldsymbol{x}_{t+1}+\boldsymbol{v}_{t+1})^T}_{\boldsymbol{y}_{t+1}^T}\\
&\times P_{t+1}\underbrace{(\boldsymbol{x}_{t+1}+\boldsymbol{v}_{t+1})}_{\boldsymbol{y}_{t+1}}+\eta_{t+1}^T\underbrace{(\boldsymbol{x}_{t+1}+\boldsymbol{v}_{t+1})}_{\boldsymbol{y}_{t+1}}\Big]\\
&+\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\frac{1}{2}\underbrace{(\boldsymbol{x}_t+\boldsymbol{v}_t)^T}_{\boldsymbol{y}_t^T}Q\underbrace{(\boldsymbol{x}_t+\boldsymbol{v}_t)}_{\boldsymbol{y}_t}+q^T\underbrace{(\boldsymbol{x}_t+\boldsymbol{v}_t)}_{\boldsymbol{y}_t}\\
&+\frac{1}{2}\|\boldsymbol{u}_t\|^2-\frac{1}{2}\underbrace{(\boldsymbol{x}_t+\boldsymbol{v}_t)^T}_{\boldsymbol{y}_t^T}P_t\underbrace{(\boldsymbol{x}_t+\boldsymbol{v}_t)}_{\boldsymbol{y}_t}-\eta_t^T\underbrace{(\boldsymbol{x}_t+\boldsymbol{v}_t)}_{\boldsymbol{y}_t}\Big]\\
&+\frac{1}{2}g_t^T\mathfrak{R}_t^\dagger g_t-\frac{1}{2}\operatorname{tr}(P_{t+1}dd^T)-\eta_t^Td-\frac{1}{2}\operatorname{tr}(P_{t+1}\Sigma_w)\\
&-\frac{1}{2}\operatorname{tr}(P_{t+1}\Sigma_v)+\frac{1}{2}\operatorname{tr}(P_t\Sigma_v)-\frac{1}{2}\operatorname{tr}(Q\Sigma_v)\Big]\\
&=\mathop{\mathbb{E}}_{\boldsymbol{y}_{t+1}|\boldsymbol{N}=N}\Big[\frac{1}{2}\boldsymbol{y}_{t+1}^TP_{t+1}\boldsymbol{y}_{t+1}+\eta_{t+1}^T\boldsymbol{y}_{t+1}\Big]+\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\frac{1}{2}\|\boldsymbol{u}_t\|^2\Big]\\
&+\mathop{\mathbb{E}}_{\boldsymbol{y}_t|\boldsymbol{N}=N}\Big[\frac{1}{2}\boldsymbol{y}_t^TQ\boldsymbol{y}_t+q^T\boldsymbol{y}_t-\frac{1}{2}\boldsymbol{y}_t^TP_t\boldsymbol{y}_t-\eta_t^T\boldsymbol{y}_t\Big]\\
&+\frac{1}{2}g_t^T\mathfrak{R}_t^\dagger g_t-\frac{1}{2}\operatorname{tr}(P_{t+1}dd^T)-\eta_t^Td-\frac{1}{2}\operatorname{tr}(P_{t+1}\Sigma_w)\\
&-\frac{1}{2}\operatorname{tr}(P_{t+1}\Sigma_v)+\frac{1}{2}\operatorname{tr}(P_t\Sigma_v)-\frac{1}{2}\operatorname{tr}(Q\Sigma_v)\\
&=:\mathop{\mathbb{E}}_{\boldsymbol{y}_{t:t+1}|\boldsymbol{N}=N}[\tilde{\psi}_{t,N}(Q,q,P_{t:t+1},\eta_t,\xi_t;\boldsymbol{y}_{t:t+1})]+\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\frac{1}{2}\|\boldsymbol{u}_t\|^2],
\end{aligned}
$$

where we introduce $\xi_t := g_t^T\mathfrak{R}_t^\dagger g_t$. We construct the objective function $\Psi(Q,q,P_{1:\nu},\eta_{1:\nu},\xi_{1:\nu-1})$ by summing up the above equation from $t=\nu-N+1$ to $\nu-1$, but excluding the terms $\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\frac{1}{2}\|\boldsymbol{u}_t\|^2]$ which are constants, and taking the expectation over $\boldsymbol{N}$. In particular,

$$
\Psi(\cdot) := \sum_{N=2}^{\nu}\mathbb{P}(\boldsymbol{N}=N)\mathop{\mathbb{E}}_{\boldsymbol{y}_{\nu-N+1:\nu}|\boldsymbol{N}=N}\left[\tilde{\psi}_N(\cdot)\right], \tag{21}
$$

where

$$
\begin{aligned}
\tilde{\psi}_N(\cdot)&=\sum_{t=\nu-N+1}^{\nu-1}\tilde{\psi}_{t,N}(\cdot)\\
&=\frac{1}{2}\boldsymbol{y}_\nu^TP_\nu\boldsymbol{y}_\nu+\eta_\nu^T\boldsymbol{y}_\nu-\frac{1}{2}\boldsymbol{y}_{\nu-N+1}^TP_{\nu-N+1}\boldsymbol{y}_{\nu-N+1}\\
&-\eta_{\nu-N+1}^T\boldsymbol{y}_{\nu-N+1}-\frac{1}{2}\operatorname{tr}(P_\nu\Sigma_v)+\frac{1}{2}\operatorname{tr}(P_{\nu-N+1}\Sigma_v)\\
&+\sum_{t=\nu-N+1}^{\nu-1}\Big(\frac{1}{2}\xi_t-\frac{1}{2}\operatorname{tr}(P_{t+1}dd^T)-\eta_{t+1}^Td+\frac{1}{2}\boldsymbol{y}_t^TQ\boldsymbol{y}_t\\
&+q^T\boldsymbol{y}_t-\frac{1}{2}\operatorname{tr}(P_{t+1}\Sigma_w)-\frac{1}{2}\operatorname{tr}(Q\Sigma_v)\Big).
\end{aligned}
\tag{22}
$$

The objective function (21) can be rewritten as a joint expectation over $\boldsymbol{y}_{1:\nu}$ and $\boldsymbol{N}$. However, we find the form in (21) more useful both in analysis and in implementation.

The objective function (21) is constructed with the idea that $\Psi(\cdot)+\sum_{t=1}^{\nu-1}\mathbb{E}_{\boldsymbol{x}_t,N}\left[\frac{1}{2}\|\boldsymbol{u}_t\|^2\right]$, where the latter is the discarded constant, should be bounded from below by 0 for all $(Q,q)\in\mathscr{F}(I)$. Therefore, ideally we would consider the problem of finding the point $(Q^\star,q^\star)$ that minimizes (21) subject to (3) and (4). From Theorem 3.1, we know that (3) and (4) are equivalent to (5) and $\xi_t = g_t^T\mathfrak{R}_t^\dagger g_t$ for $t=1:\nu-1$. However, while (21) is linear and hence a convex function, neither the constraint (5) nor the constraint $\xi_t = g_t^T\mathfrak{R}_t^\dagger g_t$ for $t=1:\nu-1$ are convex. Even so, note that the only nonconvex part in (5) is (5c). We therefore consider the relaxed convex problem obtained by removing the constraints (5c) and $\xi_t = g_t^T\mathfrak{R}_t^\dagger g_t$ for $t=1:\nu-1$.[2] In particular, let

$$
\begin{aligned}
\mathscr{D} := \Big\{\,(Q,q,P_{1:\nu},\eta_{1:\nu},\xi_{1:\nu-1})\mid(Q,q)\in\mathbb{G}(\varphi),\{P_t\in\mathbb{S}_+^n(\varphi)\}_{t=1:\nu},\\
\{\eta_t\in\mathscr{B}_\varphi^n(0)\}_{t=1:\nu},\{\xi_t\in\mathscr{B}_\varphi^1(0)\}_{t=1:\nu-1}\,\Big\}.
\end{aligned}
$$

The optimization problem for IOC reads

$$
\begin{aligned}
\min_{(Q,q,P_{1:\nu},\eta_{1:\nu},\xi_{1:\nu-1})\in\mathscr{D}}\quad&\Psi(Q,q,P_{1:\nu},\eta_{1:\nu},\xi_{1:\nu-1})\\
\text{s.t.}\quad&P_\nu=Q,\quad\eta_\nu=q &\text{(23a)}\\
&H_t\succeq 0,\quad t=1:\nu-1, &\text{(23b)}
\end{aligned}
$$

where H_t has the same form as (5b). In Section 5.2, we prove that the unique optimal solution to this optimization problem is indeed $(\bar{Q},\bar{q})$.

Nevertheless, the distribution of $\tilde{\boldsymbol{x}}$, $\{\boldsymbol{w}_t\}$, $\{\boldsymbol{v}_t\}$ and $\boldsymbol{N}$ are usually not a priori known in practice, and hence the distribution of $\boldsymbol{y}_t$ and $\boldsymbol{N}$ are not known. Therefore, it is not possible to calculate the objective function (21) explicitly, and hence we cannot solve (23) directly. But since we have the optimal state trajectory observations $\{y_t^i\}_{t=1}^{\nu}$ of the agent's trials, i.e., realizations of I.I.D. random processes $\{\boldsymbol{y}_t^i\}_{t=1}^{\nu}$ for $i=1:M$, we can instead empirically estimate the objective function. To this end, let M_N denote the number of observations which has a planning horizon of N time steps. Clearly, $\sum_{N=2}^{\nu}M_N = M$. Then, for each value of N, the expectation in (21) is approximated by the empirical mean as $\mathbb{E}_{\boldsymbol{y}_{\nu-N+1:\nu}|\boldsymbol{N}=N}[\tilde{\psi}_N(\cdot)]\approx\frac{1}{M_N}\sum_{i_N=1}^{M_N}\{$ the expression in (22) with all elements $\boldsymbol{y}_t$ replaced by $\boldsymbol{y}_t^{i_N}\}$. On the other hand, approximating the expectation over $\boldsymbol{N}$ is the same as estimating the probabilities $\mathbb{P}(\boldsymbol{N}=N)$ using the empirical estimates M_N/M. This, together with the above expression, gives an approximation of (21) as

$$
\begin{aligned}
&\Psi(Q,q,P_{1:\nu},\eta_{1:\nu},\xi_{1:\nu-1})\approx\Psi_E^{\boldsymbol{y}}(Q,q,P_{1:\nu},\eta_{1:\nu},\xi_{1:\nu-1})\\
&=\frac{1}{M}\sum_{N=2}^{\nu}\sum_{i_N=1}^{M_N}\Big[\frac{1}{2}\boldsymbol{y}_\nu^{i_N T}P_\nu\boldsymbol{y}_\nu^{i_N}+\eta_\nu^T\boldsymbol{y}_\nu^{i_N}-\frac{1}{2}\boldsymbol{y}_{\nu-N+1}^{i_N T}P_{\nu-N+1}\boldsymbol{y}_{\nu-N+1}^{i_N}\\
&-\eta_{\nu-N+1}^T\boldsymbol{y}_{\nu-N+1}^{i_N}-\frac{1}{2}\operatorname{tr}(P_\nu\Sigma_v)+\frac{1}{2}\operatorname{tr}(P_{\nu-N+1}\Sigma_v)\\
&+\sum_{t=\nu-N+1}^{\nu-1}\Big(\frac{1}{2}\xi_t-\frac{1}{2}\operatorname{tr}(P_{t+1}dd^T)-\eta_{t+1}^Td+\frac{1}{2}\boldsymbol{y}_t^{i_N T}Q\boldsymbol{y}_t^{i_N}+q^T\boldsymbol{y}_t^{i_N}\\
&-\frac{1}{2}\operatorname{tr}(P_{t+1}\Sigma_w)-\frac{1}{2}\operatorname{tr}(Q\Sigma_v)\Big)\Big].
\end{aligned}
\tag{24}
$$

We therefore consider the estimator

$$
\begin{aligned}
\min_{(Q,q,P_{1:\nu},\eta_{1:\nu},\xi_{1:\nu-1})\in\mathscr{D}}\quad&\Psi_E^{\boldsymbol{y}}(Q,q,P_{1:\nu},\eta_{1:\nu},\xi_{1:\nu-1})\\
\text{s.t.}\quad&\text{(23a)–(23b) hold.}
\end{aligned}
\tag{25}
$$

In practice, an estimate is obtained by solving (25) for a given realization $\{y_{1:\nu}^i\}_{i=1}^M$ of $\{\boldsymbol{y}_{1:\nu}^i\}_{i=1}^M$. We will use the notation $\Psi_E^{\boldsymbol{y}}(\cdot)|_{\boldsymbol{y}=y}$ to denote the objective function at the given realization.

5.2. Statistical consistency analysis

In this section, we analyze the statistical consistency of the IOC algorithm. To proceed, we first show that the optimization problem (23) is well-posed, i.e., the objective function (21) is bounded from below on its feasible domain (23a)–(23b). In addition, we show the "true" $(\bar{Q},\bar{q})$ is actually the unique global minimizer. The proof of the theorem is deferred to the Appendix.

[2] Note that substituting $\xi_t = g_t^T\mathfrak{R}_t^\dagger g_t$ into the matrix (5b), also gives a convex problem; see Nordström (2011, 2018).

Theorem 5.2. *Let $(\bar{Q}, \bar{q}) \in \mathscr{F}(I)$ be the "true" parameters of the stochastic linear–quadratic control problem (1) that governs the agent, and let $\boldsymbol{x}_{1:\nu}$, $\boldsymbol{u}_{1:\nu}$, and $\boldsymbol{y}_{1:\nu}$ be distributed accordingly. Under Assumptions 2.2–2.4, 2.6 and 5.1, for any feasible solution $(Q, q, P_{1:\nu}, \eta_{1:\nu}, \xi_{1:\nu-1})$ of the optimization problem (23), the objective function (21) is bounded from below by $-\sum_{t=1}^{\nu-1} \mathbb{E}_{\boldsymbol{x}_t,\boldsymbol{N}} \left[\frac{1}{2}\|\boldsymbol{u}_t\|^2\right]$. Moreover, for φ that is large enough, $(\bar{Q}, \bar{q}, \bar{P}_{1:\nu}, \bar{\eta}_{1:\nu}, \bar{\xi}_{1:\nu-1})$ is the unique globally optimal solution achieving the lower bound, where $\bar{P}_{1:\nu}, \bar{\eta}_{1:\nu}$ are generated by (3) and $\bar{\xi}_t = \bar{g}_t^T \bar{\mathfrak{R}}_t^\dagger \bar{g}_t$, $t = 1:\nu-1$.*

Having shown that the optimization problem (23) has $(\bar{Q}, \bar{q})$ as unique globally optimal solution, next we turn to the estimator (25). We show that it is statistically consistent, but to this end we first have the following Lemmas.

Lemma 5.3 (*Boundedness of Estimator*)**.** *The feasible region in problem (25) is compact. Moreover, for any realization, the cost function $\Psi_E^{\boldsymbol{y}}(Q, q, P_{1:\nu}, \eta_{1:\nu}, \xi_{1:\nu-1})|_{\boldsymbol{y}=y}$ is bounded on the feasible region, and the optimization problem (25) is convex and admits an optimal solution.*

Proof. The feasible region is convex and compact, and for any realization, $\Psi_E^{\boldsymbol{y}}(\cdot)|_{\boldsymbol{y}=y}$ is a linear function. This means that (25) is a convex problem and by Weierstrass' theorem, it admits an optimal solution. □

Lemma 5.4 (*Uniform Law of Large Numbers*)**.** *For large enough φ and under Assumptions 2.2–2.4, 2.6 and 5.1, the optimal value of the problem $\sup |\Psi_E^{\boldsymbol{y}}(\cdot) - \Psi(\cdot)|$ subject to $(Q, q, P_{1:\nu}, \eta_{1:\nu}, \xi_{1:\nu-1}) \in \mathscr{D}$ and (23a)–(23b), converges to 0 almost surely as $M \to \infty$.*

Proof. The proof follows along the lines of Zhang and Ringh (2023, Proof of Lem. 4.2), using bounds from Lemmas 4.3 and 5.3, and is omitted for brevity. □

Theorem 5.5 (*Statistical Consistency*)**.** *For large enough φ and under Assumptions 2.2–2.4, 2.6 and 5.1, given a realization of M trajectories, let $(Q^M, q^M, P_{1:\nu}^M, \eta_{1:\nu}^M, \xi_{1:\nu-1}^M)$ be a corresponding optimal solution to (25). Then $Q^M \xrightarrow{p} \bar{Q}$ and $q^M \xrightarrow{p} \bar{q}$ as $M \to \infty$.*

Proof. The result follows by verifying the conditions in van der Vaart (1998, Thm. 5.7). In particular, since (25) is convex, $(Q^M, q_{1:\nu}^M, P_{1:\nu}^M, \eta_{1:\nu}^M, \xi_{1:\nu-1}^M)$ is a globally optimal solution. This means that $\Psi_E^{\boldsymbol{y}}(Q^M, q^M, P_{1:\nu}^M, \eta_{1:\nu}^M, \xi_{1:\nu-1}^M)|_{\boldsymbol{y}=y} \leq \Psi_E^{\boldsymbol{y}}(\bar{Q}, \bar{q}, \bar{P}_{1:\nu}, \bar{\eta}_{1:\nu}, \bar{\xi}_{1:\nu-1})|_{\boldsymbol{y}=y}$. Moreover, since convergence almost surely implies convergence in probability (Kallenberg, 1997, Lem. 3.2), Lemma 5.4 implies that the optimal value in the statement of the Lemma converges to 0 in probability as $M \to \infty$. Finally, the fact that the feasible region to (23) and (25) is compact (see Lemma 5.3), and that (23) has a unique optimal solution (see Theorem 5.2), by van der Vaart (1998, p. 46) the last condition also holds. Hence, the result follows. □

5.3. On implementation and the computational complexity of the estimator

To get a point estimate from the estimator (25), the data (i.e., the observed trajectories) is used in the optimization problem (25). This problem can be solved using any appropriate method for solving the convex optimization problem, in the form of almost exactly as stated, by disciplined convex programming, e.g. YALMIP (Löfberg, 2004). Nevertheless, the cost function can be rewritten to make the solving process more efficient. To this end, observe that for any $Z \in \mathbb{S}^n$ and any $a \in \mathbb{R}^n$, $a^T Z a = \operatorname{tr}(Zaa^T)$. This means that by defining $\boldsymbol{y}_t^{(N)} = \sum_{i_N=1}^{M_N} \boldsymbol{y}_t^{i_N}$ and $\boldsymbol{Y}_t^{(N)} = \sum_{i_N=1}^{M_N} \boldsymbol{y}_t^{i_N} (\boldsymbol{y}_t^{i_N})^T$, the objective function (24) can be rewritten as in terms of expressions of the form

$$\sum_{i_N=1}^{M_N} \frac{1}{2} \boldsymbol{y}_\nu^{i_N T} P_\nu \boldsymbol{y}_\nu^{i_N} = \frac{1}{2} \operatorname{tr}(P_\nu \boldsymbol{Y}_\nu^{(N)}),$$

$$\sum_{i_N=1}^{M_N} \eta_\nu^T \boldsymbol{y}_\nu^{i_N} = \eta_\nu^T \boldsymbol{y}_\nu^{(N)}, \quad \sum_{i_N=1}^{M_N} \xi_t = \frac{M_N}{2} \xi_t,$$

with analogous expressions of all other terms. Note that $\boldsymbol{y}_t^{(N)}$ and $\boldsymbol{Y}_t^{(N)}$ are collecting all the samples at time-point t from trajectories with the same planning horizon length N, and that these can be pre-computed from the data before assembling the optimization problem (25). Moreover, the sizes of $\boldsymbol{y}_t^{(N)}$ and $\boldsymbol{Y}_t^{(N)}$ only depend on the dimension of the state space, n. It means that the size of the optimization problem does not grow with the amount of data collected. More specifically, since $Q \in \mathbb{S}^n$, $q \in \mathbb{R}^n$, $\{P_t \in \mathbb{S}_+^n\}_{t=1:\nu}$, $\{\eta_t \in \mathbb{R}^n\}_{t=1:\nu}$, and $\{\xi_t \in \mathbb{R}\}_{t=1:\nu-1}$, the number of variables in the problem is $n(n+1)/2 + n + \nu n(n+1)/2 + \nu n + \nu$. Moreover, the LMI constraints in (23b) are ν symmetric matrices of size $(m+n+1)\times(m+n+1)$. This means that, e.g., $n = 12$, $m = 4$, and $\nu = 80$ gives a problem with a total of 7370 scalar variables and 80 LMI constraints of size 17×17. As we demonstrate in Section 6.1, this can be handled by standard off-the-shelf convex optimization solvers.

6. Numerical examples

In this section, we present two numerical examples. The first example, in Section 6.1, illustrates that the problem (25) can be solved efficiently with off-the-shelf convex optimization solvers. The second example, in Section 6.2, applies the developed methodology to a non-zero sum pursuit-evasion game, where the pursuer models the evaders objective function using collected data. In both examples, the problem is solved on a MacBook Pro with Apple M1 eight-core CPU and 16 GB of RAM, and the implementation is done using YALMIP (Löfberg, 2004) in Matlab and solved by MOSEK (MOSEK ApS, 2019).

6.1. Demonstration of performance for a system with both moderate size and moderate planning horizon

To illustrate the performance of the method, we generate a system with moderate size and planning horizon length. In particular, to ensure that Assumption 2.2 holds, we generate continuous-time matrices $\hat{A} \in \mathbb{R}^{12\times12}$ and $\hat{B} \in \mathbb{R}^{12\times4}$ in controllable canonical form

$$\hat{A} = \begin{bmatrix} & I_4 & & \\ & & I_4 & \\ & & & I_4 \\ a_1 I_4 & a_2 I_4 & a_3 I_4 & a_4 I_4 \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} \mathbf{0}_4 \\ \mathbf{0}_4 \\ \mathbf{0}_4 \\ I_4 \end{bmatrix}.$$

We sample the coefficients a_i, $i = 1:4$ from a standard normal distribution $\mathcal{N}(0, 1)$. Next, we discretize the system by letting $A = e^{\hat{A}\Delta t}$ and $B = \int_0^{\Delta t} e^{\hat{A}t} dt \hat{B}$, using the sampling period $\Delta t = 0.1$. We choose $\bar{Q}$ to be the Hermitian part of a randomly drawn matrix with shifted eigenvalues so that the smallest eigenvalue is -0.1.[3] We set $\nu = 80$, and verify that the conditions in point (2) in Theorem 3.1 hold, i.e., that $\bar{\mathfrak{R}}_t \succ 0$ for $t = 1:79$. The process noise $\boldsymbol{w}_t$ and measurement noise $\boldsymbol{v}_t$ are drawn from multi-variate

[3] Namely, we let $G' = (G+G^T)/2$ and $\bar{Q} = G' - (\sigma_{\min}(G') + 0.1)I$, where $\sigma_{\min}(\cdot)$ is the smallest eigenvalue of a matrix and where $G \in \mathbb{R}^{12\times12}$ and elements are randomly drawn from $\mathcal{N}(0, 1)$. We shift the eigenvalues in order to make sure that we get $(\bar{Q}, \bar{q}) \in \mathscr{F}(I)$.

normal distribution $\mathcal{N}(0, \Sigma_w)$ and $\mathcal{N}(0, \Sigma_v)$, respectively, with covariance matrices that are randomly generated from a Wishart distribution of degree 12, i.e., with the same degrees of freedom as the dimension of the state. Moreover, the Wishart distribution used to draw the covariance matrices has itself a random covariance of $0.01GG^T$, where each element in $G \in \mathbb{R}^{12\times 12}$ was drawn from a standard normal distribution. Finally, we generate $M = 5 \times 10^4$ optimal trajectories, with the planning horizon lengths N drawn uniformly from the integers in the interval $[2, 80]$ and with initial value $\bar{x}$ drawn from $\mathcal{N}(0, 100I_{12})$.

The time to solve the optimization problem (25), as reported by MOSEK, is 4.85 s. Moreover, the relative error of the estimate , defined as $\frac{\|\tilde{Q}_{est}-\tilde{\bar{Q}}\|_F}{\|\tilde{\bar{Q}}\|_F}$ where $\tilde{\bar{Q}}$ is defined in (15) and $\tilde{Q}_{est}$ is defined analogously, is 0.0347. This shows that the IOC problem for systems of "moderate" size and planning horizon length can be efficiently solved with off-the-shelf solvers.

6.2. Identification of cost in non-zero sum pursuit-evasion game

In this section, we demonstrate the performance of the proposed IOC algorithm on a non-zero sum two-dimensional finite-horizon linear–quadratic pursuit-evasion game, cf. Starr and Ho (1969). For a more extensive treatment of pursuit-evasion games, see, e.g., Başar and Olsder (1982). To this end, let $\boldsymbol{x}_t \in \mathbb{R}^2$ be the distance between the pursuer and the evader, and let $\boldsymbol{u}_t^p, \boldsymbol{u}_t^e \in \mathbb{R}^2$ be the control signal of the pursuer and the evader, respectively. In particular, for each realization $(\bar{x}, N)$ of $(\bar{\boldsymbol{x}}, \boldsymbol{N})$, we assume that the evader solves the following problem

$$\min_{\substack{\boldsymbol{x}_{1:\nu},\\ \boldsymbol{u}^e_{1:\nu}}} \ \mathbb{E}_{\boldsymbol{w}_{\nu-N+1:\nu-1}}\Big[\frac{1}{2}\boldsymbol{x}_\nu^T Q^e \boldsymbol{x}_\nu + \sum_{t=\nu-N+1}^{\nu-1}[\frac{1}{2}\boldsymbol{x}_t^T Q^e \boldsymbol{x}_t + \frac{1}{2}\|\boldsymbol{u}_t^e\|^2]\Big] \tag{26a}$$

$$\text{s.t. } \boldsymbol{x}_{t+1} = A\boldsymbol{x}_t + B\boldsymbol{u}_t^e + B\boldsymbol{u}_t^p + \boldsymbol{w}_t,$$

$$t = \nu - N + 1 : \nu - 1, \tag{26b}$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t, \quad t = 1 : \nu - N \tag{26c}$$

$$\boldsymbol{x}_1 = \bar{x}, \tag{26d}$$

$$\boldsymbol{u}_1^e = \cdots = \boldsymbol{u}_{\nu-N}^e = 0, \tag{26e}$$

where (A, B) is discretized in the same way as in Section 6.1 from a continuous-time dynamics $\dot{\boldsymbol{x}} = \hat{A}\boldsymbol{x} + \hat{B}\boldsymbol{u}^e + \hat{B}\boldsymbol{u}^p$ using the sampling period $\Delta t = 0.1$, and where $\nu = 20$. Notably, $Q^e \prec 0$. In practice, as a pursuer, Q^e is unknown. In order to gain advantages over the evader and predict its future movements, a pursuer can first use some "trivial dummy movements" $\boldsymbol{u}^p_{\nu-N+1:\nu}$ that are easy for the evader to predict (i.e., known by the evader) in the first few rounds of the game. During these rounds, the pursuer collects the evader's behavior data and use the proposed IOC algorithm to estimate Q^e. In particular, here we assume the pursuer choose control $\boldsymbol{u}_t^p$ to be a constant during the data collection phase for convenience. Consequently, the forcing term $d = B\boldsymbol{u}_t^p$ would be constant for the evader (cf. (1b)).[4] The pursuer observes the noisy distance (see (2)) between the pursuer and evader, which is the optimal solution to (26).

To simulate this, we choose $\hat{A} = 0$, $\hat{B} = I_2$, $Q^e = -0.1I_2$, and for each time step in the trajectories process noise $\boldsymbol{w}_t$ and measurement noise $\boldsymbol{v}_t$ are drawn from multi-variate normal distribution $\mathcal{N}(0, \Sigma_w)$ and $\mathcal{N}(0, \Sigma_v)$, respectively, with covariance matrices

$$\Sigma_w \approx 10^{-2}\begin{bmatrix}1.04 & 0.68\\ 0.68 & 1.00\end{bmatrix} \text{ and } \Sigma_v \approx 10^{-2}\begin{bmatrix}2.33 & -2.25\\ -2.25 & 2.18\end{bmatrix}.$$

[4] In fact, it does not matter what kinds of strategy the pursuer uses in the data collection phase, as long as the evader can foresee it, since, as mentioned in Remark 2.1, the results still hold for time-varying d_t.

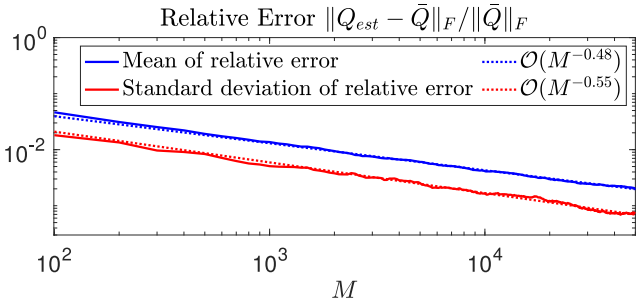


Fig. 1. Log–log plot of the mean and standard deviation of the relative error of Q_{est} as a function of the number of trajectories.

The latter matrices were randomly generated by drawing two elements from a Wishart distribution of degree 2, i.e., with the same degrees of freedom as the dimension of the state. The Wishart distribution is itself generated analogously to the distribution in Section 6.1. As "dummy" movements for the pursuer, we choose $\boldsymbol{u}_t^p = [-1, -1]^T$, $t = \nu - N + 1 : \nu$, and hence the constant forcing term in the dynamics is given by $d = B[-1, -1]^T$. Finally, the random variable $\boldsymbol{N}$ is taken to be uniformly distributed on the integers between 2 and $\nu = 20$.

To test the performance of the algorithm, we generate 100 batches of trajectories, where each batch consists of 50 000 trajectories. For each batch, we divide the trajectories into groups of size $M = 100 + 100(k-1)$, for $k = 1, \ldots, 500$, where each larger group contains all the trajectories of a smaller group. For each such group of trajectories, we solve the IOC problem (with φ set to 10^6), and this procedure is repeated for all the 100 batches. This means that we obtain estimates $Q_{est}^{\ell,M}$, for $\ell = 1, \ldots, 100$ and $M = 100, 200, \ldots, 50\,000$. For each value of M, the relative error $\|Q_{est}^{\ell,M} - Q^e\|_F/\|Q^e\|_F$ is averaged over the batches, and the resulting empirical mean and empirical standard deviation (as a function of M) is shown in Fig. 1. From the figure we see that, in line with the statistical consistency proved in Theorem 5.5, both the mean and the standard deviation decreases with increasing M. Moreover, in Fig. 1 the logarithm of the mean and the logarithm of the standard deviation appears to be (approximately) affine in $\log(M)$. The figure also shows the corresponding lines obtained by fitting an affine model to each of the two sets of logarithmic data. From this fit, we see that $\texttt{Mean of relative error} \approx \mathcal{O}(M^{-0.48})$ and $\texttt{Standard deviation of relative error} \approx \mathcal{O}(M^{-0.55})$. We hence suspect that the convergence rate is $\mathcal{O}(M^{-0.5})$, and that $\sqrt{M}(Q_M - \bar{Q})$ is asymptotically normal, just like most M-estimators such as maximum log-likelihood (van der Vaart, 1998, p. 51). However, a theoretical analysis of this is left for future work.

7. Conclusion

In this work, we have considered the inverse optimal control problem for discrete-time finite-horizon general indefinite linear–quadratic problems with stochastic planning horizons. We first investigate the necessary and sufficient conditions for when the forward problem is solvable. The identifiability of the corresponding inverse optimal control problem is analyzed and proved. Furthermore, based on the underlying necessary and sufficient condition, we construct the estimator of the inverse optimal control problem as the solution to a convex optimization problem, and prove that the estimator is statistically consistent. The performance of the estimator is illustrated on a numerical example of identifying the evaders cost in non-zero sum pursuit-evasion game.

Appendix. Deferred proofs

Proof of Proposition 2.5. We show the contraposition of the statement, i.e., that if $(\bar{Q}, \bar{q}, \bar{R})$ is such that there exists an initial value $\bar{x} \in \mathbb{R}^n$ and a time-horizon length $N \in \{2, \dots, \nu\}$ so that the optimal control problem (1) is unbounded from below, then there exists an initial value $\bar{x}'$ so that (1) is unbounded from below for planning horizon length ν. To this end, consider planning horizon $\boldsymbol{N} = \nu$. By Theorem 3.1, if $(\bar{Q}, \bar{q}) \notin \mathscr{F}(\bar{R})$ then there exists an N such that (4d) or (4e) does not hold for $t = \nu - N + 1$. Splitting the summation in the objective function as $J_\nu = \mathbb{E}_{\boldsymbol{w}_{1:\nu-1}}\Big[\frac{1}{2}\boldsymbol{x}_\nu^T\bar{Q}\boldsymbol{x}_\nu + \bar{q}^T\boldsymbol{x}_\nu + \sum_{t=\nu-N+1}^{\nu-1}[\frac{1}{2}\boldsymbol{x}_t^T\bar{Q}\boldsymbol{x}_t + \bar{q}^T\boldsymbol{x}_t + \frac{1}{2}\boldsymbol{u}_t^T\bar{R}\boldsymbol{u}_t] + \sum_{t=1}^{\nu-N}[\frac{1}{2}\boldsymbol{x}_t^T\bar{Q}\boldsymbol{x}_t + \bar{q}^T\boldsymbol{x}_t + \frac{1}{2}\boldsymbol{u}_t^T\bar{R}\boldsymbol{u}_t]\Big]$, and following along the lines of the proof of "(1) $\implies$ (2)" in Theorem 3.1, similar to (13) we get the following inequality $J_\nu \leq \mathbb{E}_{\boldsymbol{w}_{1:\nu-1}}\Big[\frac{1}{2}\boldsymbol{u}_{\nu-N+1}^T\bar{\mathfrak{R}}_{\nu-N+1}\boldsymbol{u}_{\nu-N+1} + \boldsymbol{x}_{\nu-N+1}^T\bar{\mathfrak{S}}_{\nu-N+1}^T\boldsymbol{u}_{\nu-N+1} + \bar{g}_{\nu-N+1}^T\boldsymbol{u}_{\nu-N+1} + \boldsymbol{x}_{\nu-N+1}^T\bar{\mathfrak{S}}_{\nu-N+1}^T\bar{\mathfrak{R}}_{\nu-N+1}^\dagger\bar{\mathfrak{S}}_{\nu-N+1}\boldsymbol{x}_{\nu-N+1} + \bar{g}_{\nu-N+1}^T\bar{\mathfrak{R}}_{\nu-N+1}^\dagger\bar{\mathfrak{S}}_{\nu-N+1}\boldsymbol{x}_{\nu-N+1} + \tau(\boldsymbol{x}_{\nu-N+1}) + \sum_{t=1}^{\nu-N}[\frac{1}{2}\boldsymbol{x}_t^T\bar{Q}\boldsymbol{x}_t + \bar{q}^T\boldsymbol{x}_t + \frac{1}{2}\boldsymbol{u}_t^T\bar{R}\boldsymbol{u}_t]\Big]$. Moreover, from the same proof we know that we can select $\boldsymbol{x}_{\nu-N+1}$ in order to make the terms outside of the last summation unbounded from below. Now, by invertibility of A, we can select $\boldsymbol{u}_t = 0$ and $\boldsymbol{x}_t = A^{-1}(\boldsymbol{x}_{t+1} - d - \boldsymbol{w}_t)$ for $t = 1 : \nu - N$. For any value of $\boldsymbol{x}_{\nu-N+1}$, this gives an initial condition and a sequence of states and controls that fulfill the constraints (1b)–(1e) (note that (1c) and (1e) are vacuous since $\boldsymbol{N} = \nu$). Moreover it is easily seen that J_ν is bounded from above by an expression similar to the one in the previous proof, but containing an additional constant $\tilde{\tau}(\boldsymbol{x}_{\nu-N+1})$. Following a logic similar to the reminder of the proof of "(1) $\implies$ (2)" in Theorem 3.1, the result follows. □

Proof of Lemma 4.2. Let $N \in \{2, \dots, \nu\}$ be such that $\mathbb{P}(\boldsymbol{N} = N) > 0$, and let $\operatorname{cov}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}) = \sum_{i=1}^n \lambda_i v_i$ be an orthonormal eigendecomposition of the symmetric matrix. This means that we can write $\tilde{\boldsymbol{x}} = \sum_{i=1}^n \boldsymbol{\alpha}_i v_i$ for some real-valued random variables $\boldsymbol{\alpha}_i$, $i = 1, \dots, n$. Assume that $\operatorname{cov}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}})$ is not positive definite. Then at least one eigenvalue is zero; without loss of generality, let $\lambda_1 = 0$. Then we have that $0 = v_1^T \operatorname{cov}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}})v_1 = \mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(v_1^T\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T v_1) - \mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(v_1^T\tilde{\boldsymbol{x}})\mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\tilde{\boldsymbol{x}}^T v_1) = \mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\boldsymbol{\alpha}_1^2) - \mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\boldsymbol{\alpha}_1)\mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\boldsymbol{\alpha}_1)$, and thus that $(\mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\boldsymbol{\alpha}_1))^2 = \mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\boldsymbol{\alpha}_1^2)$. By Jensen's inequality (Bauschke & Combettes, 2017, Prop. 9.24), we know that $(\mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\boldsymbol{\alpha}_1))^2 \leq \mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\boldsymbol{\alpha}_1^2)$, and by following the proof of Bauschke and Combettes (2017, Prop. 9.24), we see that equality holds if and only if $\boldsymbol{\alpha}_1$ is constant a.s. To this end, let $\boldsymbol{\alpha}_1 = c$ a.s. for some constant c. If $c = 0$, then for $\chi = v_1$, Assumption 2.6 does not hold. If $c \neq 0$, then the probability mass of $\tilde{\boldsymbol{x}}$ is located on a hyperplane defined by $\boldsymbol{\alpha}_1 = c$, which does not pass through the origin. In this case, let $\chi = v_2$ and note that for $\epsilon < c$ we have that $\mathbb{P}(\tilde{\boldsymbol{x}} \in \mathscr{B}_\epsilon^n(\rho v_2)) = 0$ for all ρ, hence violating Assumption 2.6. Therefore, we must have $\operatorname{cov}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}) \succ 0$. □

Proof of Lemma 4.3. Note that

$$\operatorname*{cov}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}) = \underbrace{\mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}\left[\begin{bmatrix}\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T & \tilde{\boldsymbol{x}} \\ \tilde{\boldsymbol{x}}^T & 1\end{bmatrix}\right]}_{\mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}[\tilde{\tilde{\boldsymbol{x}}}\tilde{\tilde{\boldsymbol{x}}}^T]} \setminus 1.$$

Hence by Lemma 4.2 and Horn and Zhang (2005, Thm. 1.12), we know that for all N such that $\mathbb{P}(\boldsymbol{N} = N) > 0$, $\mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}[\tilde{\tilde{\boldsymbol{x}}}\tilde{\tilde{\boldsymbol{x}}}^T] \succ 0$. On the other hand, by Assumption 2.3, $\boldsymbol{w}_t$ is uncorrelated with the noiseless $\boldsymbol{z}_t := A\boldsymbol{x}_t + B\boldsymbol{u}_t + d$, for $t = \nu - N + 1 : \nu - 1$, and for such t it thus holds that

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{t+1}|\boldsymbol{N}=N}\left[\tilde{\boldsymbol{x}}_{t+1}\tilde{\boldsymbol{x}}_{t+1}^T\right] = \mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\tilde{\boldsymbol{z}}_t\tilde{\boldsymbol{z}}_t^T] + \begin{bmatrix}\Sigma_w & 0 \\ 0 & 0\end{bmatrix}$$

$$= \tilde{A}_{cl}(t; \bar{Q}, \bar{q}) \mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\tilde{\boldsymbol{x}}_t\tilde{\boldsymbol{x}}_t^T]\tilde{A}_{cl}(t; \bar{Q}, \bar{q})^T + \begin{bmatrix}\Sigma_w & 0 \\ 0 & 0\end{bmatrix}, \tag{A.1}$$

where $\tilde{\boldsymbol{z}}_t = \begin{bmatrix}\boldsymbol{z}_t^T & 1\end{bmatrix}^T$. In particular, note that this holds for $t = \nu - N + 1$. By induction, since $\mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}[\tilde{\tilde{\boldsymbol{x}}}\tilde{\tilde{\boldsymbol{x}}}^T] \succ 0$, since $\tilde{A}_{cl}(t; \bar{Q}, \bar{q})$ is invertible for all $t = 1 : \nu - 1$, and since positive definiteness is invariant under congruence, we thus have $\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\tilde{\boldsymbol{x}}_t\tilde{\boldsymbol{x}}_t^T] \succ 0$ for all N such that $\mathbb{P}(\boldsymbol{N} = N) > 0$, and in particular thus for $\boldsymbol{N} = \nu$.

Now we show $\mathbb{E}[\|\tilde{\boldsymbol{x}}_t\|^2] < \infty$. First, note that $\mathbb{E}[\|\tilde{\tilde{\boldsymbol{x}}}\|^2] = \mathbb{E}[\|\tilde{\boldsymbol{x}}\|^2] + 1 < \infty$ by Assumption 2.6. Next, taking trace on both sides of (A.1), moving the trace inside the expectation, rearranging terms, and using Cauchy–Schwarz inequality, we have that $\mathbb{E}_{\boldsymbol{x}_{\nu-N+2}|\boldsymbol{N}=N}\left[\|\tilde{\boldsymbol{x}}_{\nu-N+2}\|^2\right] \leq \mathbb{E}_{\tilde{\boldsymbol{x}}|\boldsymbol{N}=N}[\|\tilde{\tilde{\boldsymbol{x}}}\|^2] \cdot \|\tilde{A}_{cl}(\nu-N+1; \bar{Q}, \bar{q})\|_F^2 + \operatorname{tr}(\Sigma_w)$. Using a similar induction argument, we thus have that $\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\|\tilde{\boldsymbol{x}}_t\|^2] < \infty$ for all $t = \nu - N + 1 : \nu$ and all N such that $\mathbb{P}(\boldsymbol{N} = N) > 0$. Finally, $\mathbb{E}[\|\tilde{\boldsymbol{x}}_t\|^2] = \sum_{N=1}^{\nu} \mathbb{P}(\boldsymbol{N} = N)\mathbb{E}_{\tilde{\boldsymbol{x}}_t|\boldsymbol{N}=N}[\|\tilde{\boldsymbol{x}}_t\|^2] < \infty$, which proves the lemma. □

Proof of Theorem 5.2. From the construction of the objective function in Section 5.1, and in view of (1e), $\Psi(\cdot) + \sum_{t=1}^{\nu-1} \mathbb{E}_{\boldsymbol{x}_t,\boldsymbol{N}}\left[\frac{1}{2}\|\boldsymbol{u}_t\|^2\right] = \sum_{N=2}^{\nu} \mathbb{P}(\boldsymbol{N} = N) \sum_{t=\nu-N+1}^{\nu-1} (\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\psi_{t,N}(Q, q; \boldsymbol{x}_t, \boldsymbol{u}_t) + \frac{1}{2}\|\boldsymbol{u}_t\|^2])$. Now, by the definition of $\psi_{t,N}(Q, q; \boldsymbol{x}_t, \boldsymbol{u}_t)$ in (18), using Assumption 2.3 and computations analogous to those in (19) we have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\psi_{t,N}(Q, q; \boldsymbol{x}_t, \boldsymbol{u}_t)] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\frac{1}{2}(A\boldsymbol{x}_t + B\boldsymbol{u}_t + d)^T P_{t+1}(A\boldsymbol{x}_t + B\boldsymbol{u}_t + d) \\
&\quad + \eta_{t+1}^T(A\boldsymbol{x}_t + B\boldsymbol{u}_t + d) - \frac{1}{2}\boldsymbol{x}_t^T P_t\boldsymbol{x}_t - \eta_t^T\boldsymbol{x}_t + \frac{1}{2}\xi_t \\
&\quad + \frac{1}{2}\boldsymbol{x}_t^T Q\boldsymbol{x}_t + q^T\boldsymbol{x}_t\Big] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\frac{1}{2}\begin{bmatrix}\boldsymbol{u}_t^T & \boldsymbol{x}_t^T & 1\end{bmatrix} H_t \begin{bmatrix}\boldsymbol{u}_t \\ \boldsymbol{x}_t \\ 1\end{bmatrix} - \frac{1}{2}\|\boldsymbol{u}_t\|^2\Big],
\end{aligned} \tag{A.2}$$

where H_t has the form (5b). On the other hand, since $(Q, q, P_{1:\nu}, \eta_{1:\nu}, \xi_{1:\nu-1})$ is feasible, by the constraint (23b) we have that $H_t \succeq 0$. Therefore, it holds that

$$\begin{aligned}
\Psi(\cdot) &= \sum_{N=2}^{\nu} \mathbb{P}(\boldsymbol{N} = N) \sum_{t=\nu-N+1}^{\nu-1} \mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\\
&\quad \frac{1}{2}\begin{bmatrix}\boldsymbol{u}_t^T & \boldsymbol{x}_t^T & 1\end{bmatrix} H_t \begin{bmatrix}\boldsymbol{u}_t \\ \boldsymbol{x}_t \\ 1\end{bmatrix} - \frac{1}{2}\|\boldsymbol{u}_t\|^2\Big] \\
&\geq -\sum_{N=2}^{\nu} \mathbb{P}(\boldsymbol{N} = N) \sum_{t=\nu-N+1}^{\nu-1} \mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\left[\frac{1}{2}\|\boldsymbol{u}_t\|^2\right] \\
&= -\sum_{t=1}^{\nu-1} \mathop{\mathbb{E}}_{\boldsymbol{x}_t,\boldsymbol{N}}\left[\frac{1}{2}\|\boldsymbol{u}_t\|^2\right].
\end{aligned} \tag{A.3}$$

This proves the first part of the theorem.

Next, we show that the lower bound is actually attained by $(\bar{Q}, \bar{q}, \bar{P}_{1:\nu}, \bar{\eta}_{1:\nu}, \bar{\xi}_{1:\nu-1})$. By using Theorem 3.1 we have that the true underlying Q and q, together with corresponding solution $\{\bar{P}_t \in \mathbb{S}^n\}_{t=1:\nu}$ and $\{\bar{\eta}_t \in \mathbb{R}^n\}_{t=1:\nu}$ to the Riccati recursions (3), and with $\bar{\xi}_t = \bar{g}_t^T\bar{\mathfrak{R}}_t^\dagger\bar{g}_t$ for $t = 1 : \nu - 1$, is a feasible solution to the optimization problem, if φ is large enough. For this feasible solution $(\bar{Q}, \bar{q}, \bar{P}_{1:\nu}, \bar{\eta}_{1:\nu}, \bar{\xi}_{1:\nu-1})$, we can decompose the corresponding $\bar{H}_t$ as in (11) and in this case, together with (16), and (A.2), $\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\psi_{t,N}(Q, q; \boldsymbol{x}_t, \boldsymbol{u}_t)]$ can be written as

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\frac{1}{2}\begin{bmatrix}\boldsymbol{u}_t^T & \boldsymbol{x}_t^T & 1\end{bmatrix}\begin{bmatrix}\bar{\mathfrak{R}}_t \\ \bar{\mathfrak{S}}_t^T \\ \bar{g}_t^T\end{bmatrix}\bar{\mathfrak{R}}_t^\dagger\begin{bmatrix}\bar{\mathfrak{R}}_t & \bar{\mathfrak{S}}_t^T & \bar{g}_t^T\end{bmatrix}\begin{bmatrix}\boldsymbol{u}_t \\ \boldsymbol{x}_t \\ 1\end{bmatrix} - \frac{1}{2}\|\boldsymbol{u}_t\|^2\Big]$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\underbrace{\frac{1}{2}\|(\bar{\mathfrak{R}}_t)^{\frac{1}{2}}(\bar{\mathfrak{R}}_t\boldsymbol{u}_t+\bar{\mathfrak{S}}_t\boldsymbol{x}_t+\bar{g}_t)\|^2}_{=0}-\frac{1}{2}\|\boldsymbol{u}_t\|^2\Big],$$

This shows that the lower bound for the objective function $\Psi(\cdot)$ is attained by $(\bar{Q},\bar{q},\bar{P}_{1:\nu},\bar{\eta}_{1:\nu},\bar{\xi}_{1:\nu-1})$.

Finally, we show that the "true" $(\bar{Q},\bar{q},\bar{P}_{1:\nu},\bar{\eta}_{1:\nu},\bar{\xi}_{1:\nu-1})$ is actually the unique global optimizer to (23). To this end, let $(Q^\star,q^\star,P^\star_{1:N},\eta^\star_{1:N},\xi^\star_{1:N-1})$ be an optimal solution to (23), and let us also use * to denote other vectors and matrices obtained using this optimal solution. Since the solution is optimal, it must be feasible, which implies that $H_t^\star \succeq 0, \forall t = 1:\nu-1$. Hence it follows that $\mathfrak{R}_t^\star \succeq 0$, $\ker(\mathfrak{R}_t^\star) \subset \big[\ker(\mathfrak{S}_t^{\star T})\cap\ker(g_t^{\star T})\big]$ and $H_t^\star\backslash\mathfrak{R}_t^\star \succeq 0$, see Horn and Zhang (2005, Thm. 1.20, p. 43). In view of the above "kernel containment", (A.2), and (A.3), the optimal value of $\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=N}[\psi_{t,N}(Q^\star,q^\star;\boldsymbol{x}_t,\boldsymbol{u}_t)]$ can be further rewritten as

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\frac{1}{2}\begin{bmatrix}\boldsymbol{u}_t^T & \boldsymbol{x}_t^T & 1\end{bmatrix}\begin{bmatrix} I & & \\ \mathfrak{S}_t^{\star T}\mathfrak{R}_t^{\star\dagger} & I & \\ g_t^{\star T}\mathfrak{R}_t^{\star\dagger} & & I\end{bmatrix}$$
$$\times\begin{bmatrix}\mathfrak{R}_t^\star & & \\ & A^TP_{t+1}^\star A+Q^\star-P_t^\star-\mathfrak{S}_t^{\star T}\mathfrak{R}_t^{\star\dagger}\mathfrak{S}_t^\star & \beta_t^\star-\mathfrak{S}_t^{\star T}\mathfrak{R}_t^{\star\dagger}g_t^\star\\ & \beta_t^{\star T}-g_t^{\star T}\mathfrak{R}_t^{\star\dagger}\mathfrak{S}_t^\star & \xi_t^\star-g_t^\star\mathfrak{R}_t^{\star\dagger}g_t^\star\end{bmatrix}$$
$$\times\begin{bmatrix} I & \mathfrak{R}_t^{\star\dagger}\mathfrak{S}_t^{\star\dagger} & \mathfrak{R}_t^{\star\dagger}g_t^{\star\dagger}\\ & I & \\ & & I\end{bmatrix}\begin{bmatrix}\boldsymbol{u}_t\\ \boldsymbol{x}_t\\ 1\end{bmatrix}-\frac{1}{2}\|\boldsymbol{u}_t\|^2\Big].$$

Recalling the notation $\tilde{\boldsymbol{x}}_t=[\boldsymbol{x}_t^T,1]^T$ and the fact that $\mathfrak{R}_t^\star\succeq 0$, we in turn get

$$\Psi(\cdot)=\sum_{N=2}^{\nu}\mathbb{P}(\boldsymbol{N}=N)\sum_{t=\nu-N+1}^{\nu-1}\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[-\frac{1}{2}\|\boldsymbol{u}_t\|^2$$
$$+\frac{1}{2}\begin{bmatrix}\boldsymbol{u}_t^T+\boldsymbol{x}_t^T\mathfrak{S}_t^{\star T}\mathfrak{R}_t^{\star\dagger}+g_t^{\star T}\mathfrak{R}_t^{\star\dagger} & \boldsymbol{x}_t^T & 1\end{bmatrix}$$
$$\times\begin{bmatrix}\mathfrak{R}_t^\star & \\ & H_t^\star\backslash\mathfrak{R}_t^\star\end{bmatrix}\begin{bmatrix}\boldsymbol{u}_t+\mathfrak{R}_t^{\star\dagger}\mathfrak{S}_t^\star\boldsymbol{x}_t+\mathfrak{R}_t^{\star\dagger}g_t^\star\\ \boldsymbol{x}_t\\ 1\end{bmatrix}\Big]$$
$$=\sum_{t=1}^{\nu-1}\mathop{\mathbb{E}}_{\boldsymbol{x}_t,\boldsymbol{N}}\Big[-\frac{1}{2}\|\boldsymbol{u}_t\|^2\Big]+\sum_{N=2}^{\nu}\mathbb{P}(\boldsymbol{N}=N)\sum_{t=\nu-N+1}^{\nu-1}\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[$$
$$\frac{1}{2}\,\|\,(\mathfrak{R}_t^\star)^{\frac{1}{2}}(\boldsymbol{u}_t+\mathfrak{R}_t^{\star\dagger}\begin{bmatrix}\mathfrak{S}_t^\star & g_t^\star\end{bmatrix}\tilde{\boldsymbol{x}}_t)\,\|^2+\frac{1}{2}\operatorname{tr}\Big((H_t^\star\backslash\mathfrak{R}_t^\star)\tilde{\boldsymbol{x}}_t\tilde{\boldsymbol{x}}_t^T\Big)\Big],$$

Note that since $H_t^\star\backslash\mathfrak{R}_t^\star\succeq 0$, all terms except $\sum_{t=1}^{\nu-1}\mathbb{E}_{\boldsymbol{x}_t,\boldsymbol{N}}[-\frac{1}{2}\|\boldsymbol{u}_t\|^2]$ are non-negative. Hence, in order for the lower bound $\sum_{t=1}^{\nu-1}\mathbb{E}_{\boldsymbol{x}_t,N}[-\frac{1}{2}\|\boldsymbol{u}_t\|^2]$ to be attained, we must have that for $t=\nu-N+1:\nu-1$

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\|\,(\mathfrak{R}_t^\star)^{\frac{1}{2}}(\boldsymbol{u}_t+\mathfrak{R}_t^{\star\dagger}\begin{bmatrix}\mathfrak{S}_t^\star & g_t^\star\end{bmatrix}\tilde{\boldsymbol{x}}_t)\,\|^2\Big]=0, \tag{A.4a}$$
$$\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\operatorname{tr}\Big((H_t^\star\backslash\mathfrak{R}_t^\star)\tilde{\boldsymbol{x}}_t\tilde{\boldsymbol{x}}_t^T\Big)\Big]=\operatorname{tr}\Big(H_t^\star\backslash\mathfrak{R}_t^\star\Big)\mathop{\mathbb{E}}_{\boldsymbol{x}_t|\boldsymbol{N}=N}\Big[\tilde{\boldsymbol{x}}_t\tilde{\boldsymbol{x}}_t^T\Big]=0, \tag{A.4b}$$

for all N such that $\mathbb{P}(\boldsymbol{N}=N)>0$. In particular, by Assumption 2.4 it must be true for $\boldsymbol{N}=\nu$. From Lemma 4.3, we know that $\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=\nu}[\tilde{\boldsymbol{x}}_t\tilde{\boldsymbol{x}}_t^T]\succ 0$. Thus it follows that, for $N=\nu$, (A.4b) implies that $H_t^\star\backslash\mathfrak{R}_t^\star=0$ holds for $t=1:\nu-1$. By using the observation in Remark 3.2, we therefore have that $(Q^\star,q^\star,P^\star_{1:\nu},\eta^\star_{1:\nu})$ satisfies the generalized Riccati iterations (3).

Now, to show that the optimal solution to (23) is unique, first assume that $(Q^\star,q^\star,P^\star_{1:\nu},\eta^\star_{1:\nu})$ is an optimal solution such that $\mathfrak{R}_t^\star\succ 0$, for $t=1:\nu-1$. In this case, also $(\mathfrak{R}_t^\star)^{\frac{1}{2}}\succ 0$ for $t=1:\nu-1$. By (A.4a), this means that, conditioned on $\boldsymbol{N}=\nu$, we have $\boldsymbol{u}_t=-\mathfrak{R}_t^{\star\dagger}\begin{bmatrix}\mathfrak{S}_t^\star & g_t^\star\end{bmatrix}\tilde{\boldsymbol{x}}_t$ a.s. for $t=1:\nu-1$. But conditioned on $\boldsymbol{N}=\nu$, we also have that $\boldsymbol{u}_t=-\bar{\mathfrak{R}}_t^\dagger\begin{bmatrix}\bar{\mathfrak{S}}_t & \bar{g}_t\end{bmatrix}\tilde{\boldsymbol{x}}_t$. Therefore $\mathfrak{R}_t^{\star\dagger}\begin{bmatrix}\mathfrak{S}_t^\star & g_t^\star\end{bmatrix}\tilde{\boldsymbol{x}}_t=\bar{\mathfrak{R}}_t^\dagger\begin{bmatrix}\bar{\mathfrak{S}}_t & \bar{g}_t\end{bmatrix}\tilde{\boldsymbol{x}}_t$, a.s., for $t=1:\nu-1$. Multiplying from the right with $\tilde{\boldsymbol{x}}_t^T$ and taking expectation $\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=\nu}$ on both sides, we have that $\mathfrak{R}_t^{\star\dagger}\begin{bmatrix}\mathfrak{S}_t^\star & g_t^\star\end{bmatrix}\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=\nu}[\tilde{\boldsymbol{x}}_t\tilde{\boldsymbol{x}}_t^T]=\bar{\mathfrak{R}}_t^\dagger\begin{bmatrix}\bar{\mathfrak{S}}_t & \bar{g}_t\end{bmatrix}\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=\nu}[\tilde{\boldsymbol{x}}_t\tilde{\boldsymbol{x}}_t^T]$, for $t=1:\nu-1$. Using Lemma 4.3, we know that $\mathbb{E}_{\boldsymbol{x}_t|\boldsymbol{N}=\nu}[\tilde{\boldsymbol{x}}_t\tilde{\boldsymbol{x}}_t^T]\succ 0$ and hence the matrix is full rank. Therefore, it must hold that $\mathfrak{R}_t^{\star\dagger}\begin{bmatrix}\mathfrak{S}_t^\star & g_t^\star\end{bmatrix}=\bar{\mathfrak{R}}_t^\dagger\begin{bmatrix}\bar{\mathfrak{S}}_t & \bar{g}_t\end{bmatrix}$, $t=1:\nu-1$, and thus, by (17), that $\bar{A}_{cl}(t;Q^\star,q^\star)=\bar{A}_{cl}(t;\bar{Q},\bar{q})$, $t=1:\nu-1$. By Proposition 4.1, we therefore have that $\bar{Q}=Q^\star$ and that $\bar{q}=q^\star$. This implies that in the subset of the feasible region (23a)–(23b) where $\mathfrak{R}_t\succ 0$, for $t=1:\nu-1$, it holds that $(\bar{Q},\bar{q},\bar{P}_{1:\nu},\bar{\eta}_{1:\nu},\bar{\xi}_{1:\nu-1})$ is the unique globally optimal solution to (23). Next, suppose that there exists a minimizer $(Q^\star,q^\star,P^\star_{1:\nu},\eta^\star_{1:\nu},\xi^\star_{1:\nu-1})$ such that $\mathfrak{R}_t^\star\succeq 0$ but not strictly positive definite for some $t\in\{1,\dots,\nu-1\}$. In particular, by Proposition 3.4 this means that $(Q^\star,q^\star,P^\star_{1:\nu},\eta^\star_{1:\nu},\xi^\star_{1:\nu-1})\neq(\bar{Q},\bar{q},\bar{P}_{1:\nu},\bar{\eta}_{1:\nu},\bar{\xi}_{1:\nu-1})$. Since (23) is a convex optimization problem that attains an optimal solution, the set of all optimal solutions is a nonempty convex set (Rockafellar, 1970, Thm. 27.2). This means that all points $(Q^\alpha,q^\alpha,P^\alpha_{1:\nu},\eta^\alpha_{1:\nu},\xi^\alpha_{1:\nu-1}):=(\alpha\bar{Q}+(1-\alpha)Q^\star,\alpha\bar{q}+(1-\alpha)q^\star,\{\alpha\bar{P}_t+(1-\alpha)P_t^\star\}_{t=1}^{\nu},\{\alpha\bar{\eta}_t+(1-\alpha)\eta_t^\star\}_{t=1}^{\nu},\{\alpha\bar{\xi}_t+(1-\alpha)\xi_t^\star\}_{t=1}^{\nu-1})$ are optimal, for all $\alpha\in[0,1]$. Since the eigenvalues of $\mathfrak{R}_t$, $t=1:\nu-1$, depends smoothly on P_t (see (4)), we can select α close enough to 1 so that $(Q^\alpha,q^\alpha,P^\alpha_{1:\nu},\eta^\alpha_{1:\nu},\xi^\alpha_{1:\nu-1})$ will be such that $\mathfrak{R}_t^\alpha\succ 0$ for all $t=1:\nu-1$. However, this contradicts the fact that $(\bar{Q},\bar{q},\bar{P}_{1:\nu},\bar{\eta}_{1:\nu},\bar{\xi}_{1:\nu-1})$ is the unique globally optimal solution to (23) with $\mathfrak{R}_t\succ 0$, $t=1:\nu-1$. Therefore, there can be no optimal solution such that $\mathfrak{R}_t^\star$ is not (strictly) positive definite for all $t\in\{1,\dots,\nu-1\}$, and hence $(\bar{Q},\bar{q},\bar{P}_{1:\nu},\bar{\eta}_{1:\nu},\bar{\xi}_{1:\nu-1})$ is the unique globally optimal solution to (23). □

References

Alexander, R. McNeill (1996). *Optima for animals*. Princeton, NJ: Princeton University Press.

Anderson, Brian D. O., & Moore, John B. (2007). *Optimal control: linear quadratic methods*. Mineola, NY: Dover Publications.

Başar, Tamer, & Olsder, Geert Jan (1982). *Dynamic noncooperative game theory*. London, UK: Academic Press.

Bauschke, H. H., & Combettes, P. L. (2017). *Convex analysis and monotone operator theory in Hilbert spaces* (2nd ed.). Cham: Springer.

Bertsekas, Dimitri P. (2000). *Dynamic programming and optimal control: Volume 1* (2nd ed.). Belmont, MA: Athena scientific.

Boyd, Stephen, El Ghaoui, Laurent, Feron, Eric, & Balakrishnan, Venkataramanan (1994). *Linear matrix inequalities in system and control theory*. Philadelphia, PA: SIAM.

Chen, Shuping, Li, Xunjing, & Zhou, Xun Yu (1998). Stochastic linear quadratic regulators with indefinite control weight costs. *SIAM Journal on Control and Optimization*, *36*(5), 1685–1702.

Ferrante, Augusto, & Ntogramatzidis, Lorenzo (2015). A note on finite-horizon LQ problems with indefinite cost. *Automatica*, *52*, 290–293.

Ferrante, Augusto, & Ntogramatzidis, Lorenzo (2016). A discussion on the discrete-time finite-horizon indefinite LQ problem. In *2016 IEEE 55th conference on decision and control* (pp. 216–220). IEEE.

Gohberg, Israel, Lancaster, Peter, & Rodman, Leiba (2005). *Indefinite linear algebra and applications*. Basel: Birkhäuser Verlag.

Horn, Roger A., & Zhang, Fuzhen (2005). Basic properties of the Schur complement. In Fuzhen Zhang (Ed.), *The Schur complement and its applications* (pp. 17–46). Boston, MA: Springer.

Kallenberg, Olav (1997). *Foundations of modern probability*. New York, NY: Springer.

Kalman, Rudolf E. (1964). When is a linear control system optimal? *Journal of Basic Engineering*, *86*(1), 51–60.

Keshavarz, Arezou, Wang, Yang, & Boyd, Stephen (2011). Imputing a convex objective function. In *2011 IEEE international symposium on intelligent control* (pp. 613–619). IEEE.

Li, Yibei, Yao, Yu, & Hu, Xiaoming (2020). Continuous-time inverse quadratic optimal control problem. *Automatica*, *117*, Article 108977.

Li, Yibei, Zhang, Han, Yao, Yu, & Hu, Xiaoming (2018). A convex optimization approach to inverse optimal control. In *2018 37th Chinese control conference* (pp. 257–262). IEEE.

Lian, Bosen, Xue, Wenqian, Lewis, Frank L., & Chai, Tianyou (2021). Robust inverse Q-learning for continuous-time linear systems in adversarial environments. *IEEE Transactions on Cybernetics*.

Ljung, Lennart, & Glad, Torkel (1994). On global identifiability for arbitrary model parametrizations. *Automatica*, *30*(2), 265–276.

Löfberg, J. (2004). YALMIP : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD conference*. Taipei, Taiwan.

MOSEK ApS (2019). The MOSEK optimization toolbox for MATLAB manual. Version 9.0.

Ng, Andrew Y., & Russell, Stuart (2000). Algorithms for inverse reinforcement learning.. In *Proceeding of the 17th international conference on machine learning* (pp. 663–670).

Nordström, Kenneth (2011). Convexity of the inverse and Moore–Penrose inverse. *Linear Algebra and its Applications*, *434*(6), 1489–1512.

Nordström, Kenneth (2018). A note on the convexity of the Moore–Penrose inverse. *Linear Algebra and its Applications*, *538*, 143–148.

Priess, M Cody, Conway, Richard, Choi, Jongeun, Popovich, John M, & Radcliffe, Clark (2014). Solutions to the inverse LQR problem with application to biological systems analysis. *IEEE Transactions on Control Systems Technology*, *23*(2), 770–777.

Rami, Mustapha Ait, Chen, X., & Zhou, Xun Yu (2002). Discrete-time indefinite LQ control with state and control dependent noises. *Journal of Global Optimization*, *23*(3), 245–265.

Ran, André C. M., & Trentelman, Harry L. (1993). Linear quadratic problems with indefinite cost for discrete time systems. *SIAM Journal on Matrix Analysis and Applications*, *14*(3), 776–797.

Åström, Karl J. (2006). *Introduction to stochastic control theory*. Mineola, NY: Dover, Unabridged republication of original published by Academic Press, 1970.

Reid, Robert E., Tugcu, A. Kemal, & Mears, Barry C. (1983). An optimal controller arising from minimization of a quadratic performance criterion of indefinite form. *IEEE Transactions on Automatic Control*, *28*(10), 985–987.

Rockafellar, R. Tyrrell (1970). *Convex Analysis*. Princeton, NJ: Princeton University Press.

Starr, Alan Wilbor, & Ho, Yu-Chi (1969). Nonzero-sum differential games. *Journal of Optimization Theory and Applications*, *3*(3), 184–206.

Toumi, Noureddine, Malhamé, Roland, & Le Ny, Jerome (2020). A tractable mean field game model for the analysis of crowd evacuation dynamics. In *2020 59th IEEE conference on decision and control* (pp. 1020–1025). IEEE.

Toumi, Noureddine, Malhamé, Roland, & Le Ny, Jerome (2021). A spatial partitioning based crowd evacuation model. In *2021 60th IEEE conference on decision and control* (pp. 5247–5252). IEEE.

van der Vaart, Adrianus W. (1998). *Asymptotic statistics*. Cambridge, United Kingdom: Cambridge University Press.

Xue, Wenqian, Kolaric, Patrik, Fan, Jialu, Lian, Bosen, Chai, Tianyou, & Lewis, Frank L (2021). Inverse reinforcement learning in tracking control based on inverse optimal control. *IEEE Transactions on Cybernetics*.

Xue, Wenqian, Lian, Bosen, Fan, Jialu, Kolaric, Patrik, Chai, Tianyou, & Lewis, Frank L (2021). Inverse reinforcement Q-learning through expert imitation for discrete-time systems. *IEEE Transactions on Neural Networks and Learning Systems*.

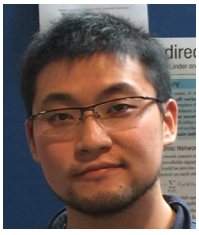
Yu, Chengpu, Li, Yao, Fang, Hao, & Chen, Jie (2021). System identification approach for inverse optimal control of finite-horizon linear quadratic regulators. *Automatica*, *129*, Article 109636.

Zhang, Han, & Ringh, Axel (2023). Inverse linear-quadratic discrete-time finite-horizon optimal control for indistinguishable homogeneous agents: A convex optimization approach. *Automatica*, *148*, Article 110758.

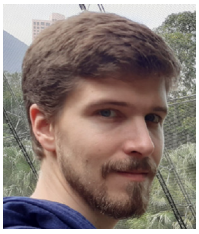
Zhang, Han, Ringh, Axel, Jiang, Weihan, Li, Shaoyuan, & Hu, Xiaoming (2022). Statistically consistent inverse optimal control for linear-quadratic tracking with random time horizon. In *2022 41st Chinese control conference* (pp. 1515–1522). http://dx.doi.org/10.23919/CCC55666.2022.9902327.

Zhang, Han, Umenberger, Jack, & Hu, Xiaoming (2019). Inverse optimal control for discrete-time finite-horizon linear quadratic regulators. *Automatica*, *110*, Article 108593.

Zhou, Xun Yu, & Li, Duan (2000). Continuous-time mean-variance portfolio selection: A stochastic LQ framework. *Applied Mathematics and Optimization*, *42*(1), 19–33.



Han Zhang received his Ph.D. from Dept. of Mathematics, KTH Royal Institute of Technology, Sweden in 2019. He obtained both this B.S. and M.S. degree from Dept. of Automation, Shanghai Jiao Tong University in 2011 and 2014 respectively. He is now an associate professor in Dept. of Automation, Shanghai Jiao Tong University. His main research interests are control of inverse optimal control, game theory and their application in robotics.



Axel Ringh received a M.Sc. degree in Engineering Physics in 2014, and a Ph.D. degree in Applied and Computational Mathematics in 2019, both from KTH Royal Institute of Technology, Stockholm, Sweden. From 2019 to 2021 he was a postdoctoral researcher with the Department of Electronic and Computer Engineering, the Hong Kong University of Science and Technology, Hong Kong, China, and since 2021 he is assistant professor at the Department of Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, Gothenburg, Sweden. He is the recipient of the European Control Conference 2015 Best Student Paper Award, and the SIAM Activity Group on Control and Systems Theory Best SICON Paper Prize 2023. His research interests are in the field of applied mathematics, specifically in optimization and systems theory, and the intersection with areas such as control theory, signal processing, inverse problems, and machine learning.