Nonuniqueness and Convergence to Equivalent Solutions in Observer-based Inverse Reinforcement Learning ⋆

Jared Town [a], Zachary Morrison [a], Rushikesh Kamalapurkar [a]

[a] *School of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK, 74078, USA*

Abstract

A key challenge in solving the deterministic inverse reinforcement learning (IRL) problem online and in real-time is the existence of multiple solutions. Nonuniqueness necessitates the study of the notion of equivalent solutions, i.e., solutions that result in a different cost functional but same feedback matrix, and convergence to such solutions. While *offline* algorithms that result in convergence to equivalent solutions have been developed in the literature, online, real-time techniques that address nonuniqueness are not available. In this paper, a regularized history stack observer that converges to approximately equivalent solutions of the IRL problem is developed. Novel data-richness conditions are developed to facilitate the analysis and simulation results are provided to demonstrate the effectiveness of the developed technique.

Key words: inverse reinforcement learning, inverse optimal control, optimal control, adaptive systems, nonlinear observer and filter design

1 Introduction

This paper concerns recovery of the cost functional being optimized by an *expert* through observation of their input-output behavior. The expert is assumed to be controlling a deterministic dynamical system. The controller being implemented by the expert is assumed to be optimal with respect to an unknown cost functional. The objective of the learner is to estimate the cost functional using measurements of the experts inputs and outputs. Cost functional estimation techniques are studied in the literature under the umbrella of inverse reinforcement learning [16]. While IRL typically includes utilization of the estimated cost functionals for behavior imitation using (forward) reinforcement learning, the scope of this paper is limited to cost functional estimation.

IRL methods are often utilized to teach an autonomous system a specific task in an offline environment by observing repeated performance of the same task by the expert [1, 3, 6, 7, 14, 16, 17, 19, 27]. While effective, IRL techniques are generally offline, computationally complex, require multiple trajectories or several iterations over one trajectory, and require a greater amount of data than is readily available in real-time (online) applications. The aforementioned limitations are addressed in results such as [2, 4, 20] where online IRL methods that utilize a single iteration over one continuous trajectory are developed to learn the cost functional of the expert. New techniques to solve the IRL problem up to a scaling factor through non-cooperative linear quadratic differential games are also developed in [7] and [8].

Results such as [2, 4, 8, 20] (implicitly or explicitly) assume that the IRL problem admits a unique solution. Since IRL problems generally admit multiple linearly independent solutions [9, 10], the uniqueness assumption is restrictive. Non-uniqueness is studied in results such as [9], where procedures to determine equivalent cost functionals are developed. It is also shown that IRL problems with multiple solutions arise naturally in state space models that have a product structure (see [10]). Many real-world systems have a product structure, either in the original model or in the linearized model. For example, linearized dynamics of aerospace vehicles have a product structure due to separation of longitudinal and lateral dynamics [10]. The study of IRL prob-

⋆ This research was supported, in part, by the National Science Foundation (NSF) under award numbers 1925147 and 2027999 and the Air Force Office of Scientific Research under award number FA9550-20-1-0127. Any opinions, findings, conclusions, or recommendations detailed in this article are those of the author(s), and do not necessarily reflect the views of the sponsoring agencies.

Email addresses: jared.town@okstate.edu (Jared Town), zachmor@okstate.edu (Zachary Morrison), rushikesh.kamalapurkar@okstate.edu (Rushikesh Kamalapurkar).

lems that admit multiple solutions is thus indispensable in real-world applications.

The IRL methods recently developed in results such as [3, 15, 26] study nonuniqueness of solutions to IRL problems and guarantee convergence to the set of equivalent solutions. In [3, 26] the IRL problem is solved in an offline setting as opposed to the online and real-time problem under consideration in this paper. In results such as [3, 15] equivalent solutions for the state penalty matrix are identified, using measurements of only the control input of the expert. However, these results do not estimate the control penalty of the expert. The technique developed in this paper requires more information than [3, 15] (measurements of the control input *and the output* of the expert), but in contrast with [3, 15], the entire cost functional of the expert, including state *and control* penalties, is estimated.

Motivated by [20], the method developed in this paper identifies an equivalent cost functional for the expert given measurements of the control input and the output of the expert in an observer framework. Specifically, the History Stack Observer (HSO) from [20], originally designed under the uniqueness assumption, is extended to IRL problems that admit multiple solutions. The redesigned HSO is a true extension of the HSO from [20] in the sense that it identifies the true cost functional of the expert, up to a scaling factor, if the IRL problem has a unique solution. While nonuniqueness is studied in the observer context in [25], the definition of equivalence used in this paper is stronger than the one in [25]. As a result, the analysis that proves convergence to equivalent solutions is more involved than the analysis in [25]. In addition, the practically relevant case of convergence to approximately equivalent solutions is studied in this paper.

This article extends the IRL HSO in [20] to problems where the observed trajectories can be optimal with respect to multiple cost functionals. A learner with access to the state space model, controller input, and measurement data reconstructs an equivalent cost functional of an expert. Since recovery of the true cost functional cannot be expected in such problems, analysis of the error between the estimated cost functional and the true cost functional, as done in [20], is no longer useful. In this paper, a novel analysis approach that guarantees convergence of the learned solution to a neighborhood of an equivalent solution is developed. Under sufficient data informativity conditions, a new equivalence metric is designed such that convergence of the equivalence metric to zero implies convergence to an equivalent solution. The developed modification to the HSO is inspired by ridge regression, but has a surprising convergence property. Under ideal conditions (no noise and persistently exciting regressor), the convergence is exact, as opposed to ridge regression, where the solutions are off by a factor proportional to the regularization coefficient.

2 Problem Formulation

The system being controlled by the expert is assumed to be a linear system of the form

$$\dot{x}(t) = Ax + Bu, \tag{1}$$

with output

$$y = Cx(t), \tag{2}$$

where the state is $x \in \mathbb{R}^n$ and the control input is $u \in \mathbb{R}^m$. The system matrices are given as $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$, and the output and output matrix are given as $y \in \mathbb{R}^L$ and $C \in \mathbb{R}^{L\times n}$ respectively.

The expert is assumed to implement an optimal controller that optimizes the cost functional

$$J(x_0, u(\cdot)) = \int_0^\infty \left(x(t)^\top Q x(t) + u(t)^\top R u(t)\right) \mathrm{d}t, \tag{3}$$

where $x(\cdot)$ is the system trajectory under the optimal control signal $u(\cdot)$ and starting from the initial condition x_0, $Q \in \mathbb{R}^{n\times n}$ is an unknown positive semi-definite matrix, and $R \in \mathbb{R}^{m\times m}$ is an unknown positive definite matrix. The following assumption ensures that the IRL problem is well-posed.

Assumption 1 *The pair (A, B) is stabilizable and the pairs (A, C) and $(A, \sqrt{Q})$ are detectable.*

Stabilizability of (A, B) and detectability of $(A, \sqrt{Q})$ is needed for the optimal controller to exist and detectability of (A, C) guarantees the existence of a matrix L such that $A - LC$ is Hurwitz [5, Lemma 21.1]. Under Assumption 1, the policy of the expert is given by $u = K_{Ep}x$, where $K_{Ep} \in \mathbb{R}^{m\times n}$ is obtained by solving the algebraic Riccati equation (ARE) corresponding to the optimal control problem described by the system in (1) and the cost functional in (3).

The learning objective is to estimate, online and in real-time, the unknown matrices in the cost functional using knowledge of the system matrices, A, B, and C, and input-output data. Generally, for a system (A, B, C), a given set of input-output trajectories is optimal with respect to multiple cost functionals. As a result, the true cost functional cannot generally be estimated from data. Instead, an equivalent solution to the IRL problem is sought (see Definition 2 and [26]).

While the HSO in [20] is an effective technique to solve the IRL problem online and in real-time, the analysis focuses on the error between the true cost functional matrices and their estimates, and as such, implicitly assumes uniqueness of solutions. As such, the method in [20] cannot be applied to a large class of IRL problems that admit multiple solutions. In this paper, the HSO is

extended to be applicable to IRL problems that admit multiple solutions. While the extension is similar to the regularization used in ridge regression, the fact that the error between the true cost functional matrices and the obtained estimates can no longer be used as a metric to gauge quality of the estimates necessitates the development of a novel analysis approach.

3 Nonuniqueness and the History Stack Observer

To facilitate the discussion, this section provides a brief summary of the HSO developed in [20] and highlights the key problem that is resolved in this paper.

3.1 Equivalent Solutions and Equivalence Metric

If the state and control trajectories of the system are optimal with respect to the cost functional in (3) and Assumption 1 is met, then there exists a matrix S such that for all $t \geq 0$, the matrices Q, R, A, and B, and the optimal trajectories $x(\cdot)$ and $u(\cdot)$ satisfy the Hamilton-Jacobi-Bellman (HJB) equation

$$x(t)^\top \left(A^\top S + SA - SBR^{-1}B^\top S + Q\right) x(t) = 0, \quad (4)$$

and the optimal control equation

$$u(t) = u^*(x(t)) \coloneqq -R^{-1}B^\top S x(t). \quad (5)$$

The feedback matrix of the expert is then given by $K_{Ep} = R^{-1}B^\top S$. The HJB equation and the optimal control equation facilitate the definition of an equivalent solution.

Definition 2 *A solution $(\hat{Q}, \hat{S}, \hat{R})$ is called an equivalent solution of the IRL problem if it satisfies the ARE $A^\top \hat{S} + \hat{S}A - \hat{S}B\hat{R}^{-1}B^\top \hat{S} + \hat{Q} = 0$ and optimization of the performance index J, with $Q = \hat{Q}$ and $R = \hat{R}$, results in the same feedback matrix as the one utilized by the expert, that is, $\hat{K}_P \coloneqq \hat{R}^{-1}B^\top \hat{S} = K_{Ep}$.*

Given an estimate $\hat{x}$ of the state x, a measurement of the control signal, u, and estimates $\hat{Q}$, $\hat{R}$, and $\hat{S}$ of Q, R, and S, respectively, (4) and (5) can be evaluated to develop an observation error that evaluates to zero if the state estimates are correct and $(\hat{Q}, \hat{R}, \hat{S})$ is an equivalent solution. The observation error is then used to improve the estimates by framing the IRL problem as a state estimation problem. The rest of this subsection is borrowed from [20] and is included here for completeness.

To facilitate the observer design, equations (4) and (5) are linearly parameterized as

$$0 = 2\sigma_{R2}(u)W_R^* + B^\top \left(\nabla_x \sigma_S(x)\right)^\top W_S^*, \quad (6)$$

$$0 = \nabla_x \left((W_S^*)^\top \sigma_S(x)\right)(Ax + Bu) + (W_Q^*)^\top \sigma_Q(x) + (W_R^*)^\top \sigma_{R1}(u), \quad (7)$$

where $x^\top S x = (W_S^*)^\top \sigma_S(x)$, $x^\top Q x = (W_Q^*)^\top \sigma_Q(x)$, $u^\top R u = (W_R^*)^\top \sigma_{R1}(u)$, and $Ru = \sigma_{R2}(u)W_R^*$, where $\left[W_S^{*\top}, W_Q^{*\top}, W_R^{*\top}\right]^\top \in \mathbb{R}^{P_S} \times \mathbb{R}^{P_Q} \times \mathbb{R}^M$ are the ideal weights with P_S, P_Q, and M being the number of basis functions in the respective linear parameterization. For a complete characterization of the weights and the basis functions, see [18].

Using the estimates $\hat{W}_S$, $\hat{W}_Q$, and $\hat{W}_R$ for W_S^*, W_Q^*, and W_R^*, respectively, in (6) and (7), a control residual error and an inverse Bellman error are defined as

$$\Delta_u' \coloneqq 2\sigma_{R2}(u)\hat{W}_R + B^\top \left(\nabla_x \sigma_S(x)\right)^\top \hat{W}_S \text{ and} \quad (8)$$

$$\delta' \coloneqq \nabla_x \left((\hat{W}_S)^\top \sigma_S(x)\right)(Ax + Bu) + (\hat{W}_Q)^\top \sigma_Q(x) + (\hat{W}_R)^\top \sigma_{R1}(u). \quad (9)$$

The scaling ambiguity inherent in linear quadratic optimal control, which is apparent in the fact that $\hat{W}' = [\hat{W}_S^\top, \hat{W}_Q^\top, \hat{W}_R^\top]^\top = 0$ is a solution of (6) and (7), is resolved, without loss of generality, by assigning an arbitrary value to one element of $\hat{W}'$. Selecting the first component of $\hat{W}_R$ to be equal to $r_1 > 0$ and removing it from the weight vector $\hat{W}'$ in (6) and (7) yields scale-aware definitions of the control residual error and the inverse Bellman error, given by

$$\begin{bmatrix} \delta\left(x, u, \hat{W}\right) \\ \Delta_u\left(x, u, \hat{W}\right) \end{bmatrix} = \begin{bmatrix} \sigma_\delta(x, u) \\ \sigma_{\Delta_u}(x, u) \end{bmatrix} \begin{bmatrix} \hat{W}_S \\ \hat{W}_Q \\ \hat{W}_R^- \end{bmatrix} + \begin{bmatrix} u_1^2 r_1 \\ 2u_1 r_1 \\ 0_{m-1\times 1} \end{bmatrix}, \quad (10)$$

where $\hat{W}_R^-$ is a copy of $\hat{W}_R$ with the first element removed, σ_δ is a copy of $\left[(Ax + Bu)^\top (\nabla_x \sigma_S(x))^\top, \sigma_Q(x)^\top, \sigma_{R1}(u)^\top\right]$, with the $(P_S + P_Q + 1)-$th element removed, and σ_{Δ_u} is a copy of $\left[B^\top (\nabla_x \sigma_S(x))^\top, 0_{m\times P_Q}, 2\sigma_{R2}(u)\right]$, with the $(P_S + P_Q + 1)-$th column removed. In this paper, the error system in (10) is used as an *equivalence metric* to develop an observer-based IRL method. The following section provides a brief overview of the observer developed in [20].

3.2 The History Stack Observer

Pairing the innovation $y - C\hat{x}$ with the inverse bellman error and control residual error from (10) yields

the observation error $\omega = \begin{bmatrix} Cx \\ \Sigma_u \end{bmatrix} - \begin{bmatrix} C\hat{x} \\ \hat{\Sigma}\hat{W} \end{bmatrix}$, where $\hat{W} = [\hat{W}_S^\top, \hat{W}_Q^\top, (\hat{W}_R^-)^\top]^\top$,

$$\hat{\Sigma} \coloneqq \begin{bmatrix} \sigma_\delta\left(\hat{x}(t_1), u(t_1)\right) \\ \sigma_{\Delta_u}\left(\hat{x}(t_1), u(t_1)\right) \\ \vdots \\ \sigma_\delta\left(\hat{x}(t_N), u(t_N)\right) \\ \sigma_{\Delta_u}\left(\hat{x}(t_N), u(t_N)\right) \end{bmatrix}, \text{ and } \Sigma_u \coloneqq \begin{bmatrix} -u_1^2(t_1)r_1 \\ -2u_1(t_1)r_1 \\ 0_{m-1\times 1} \\ \vdots \\ -u_1^2(t_N)r_1 \\ -2u_1(t_N)r_1 \\ 0_{m-1\times 1} \end{bmatrix},$$

Using the observation error, the history stack observer is designed in [20] as

$$\begin{bmatrix} \dot{\hat{x}} \\ \dot{\hat{W}} \end{bmatrix} = \begin{bmatrix} A\hat{x} + Bu \\ 0_{P_S+P_Q+M-1} \end{bmatrix} + K\left(\begin{bmatrix} Cx \\ \Sigma_u \end{bmatrix} - \begin{bmatrix} C\hat{x} \\ \hat{\Sigma}\hat{W} \end{bmatrix}\right), \quad (11)$$

where the gain K is selected as

$$K \coloneqq \begin{bmatrix} K_3 & 0_{n\times N+Nm} \\ 0_{P_S+P_Q+M-1\times L} & K_4(\hat{\Sigma}^\top\hat{\Sigma})^{-1}\hat{\Sigma}^\top \end{bmatrix}, \quad (12)$$

where K_3 is selected so that $A-K_3C$ is Hurwitz, and K_4 is scalar multiple of an identity matrix of size P_S+P_Q+M-1. To facilitate the analysis, let Σ be a copy of $\hat{\Sigma}$ where the state estimates are replaced by their true values and let $W^* \coloneqq (r_1/W_R^*(1))[W_S^{*\top}, W_Q^{*\top}, (W_R^{-*})^\top]^\top$, where W_R^{-*} denotes W_R^* with the first element, $W_R^*(1)$, removed.

The matrices $\hat{\Sigma} \in \mathbb{R}^{N(m+1)\times P_S+P_Q+M-1}$ and $\Sigma_u \in \mathbb{R}^{N(m+1)}$ are constructed using the dataset $\{(\hat{x}(t_i), u(t_i))\}_{i=1}^N$, recorded at time instances $\{t_1, \dots t_N\}$, with $N \geq P_S + P_Q + M - 1$. The dataset is referred to hereafter as a *history stack*. To ensure convergence of the weights, updated using (11), to an equivalent solution (see Theorem 7 below), the history stack is recorded using a condition number minimization algorithm. At any time, two separate history stacks, H_1 and H_2 are maintained. The history stack H_1 is used to compute the matrices $\hat{\Sigma}$ and Σ_u in (11) and H_2 is populated with current state estimates and control inputs.

Both history stacks are initialized as zero matrices of the appropriate size. As state estimates become available, they are added, along with the corresponding control input, to H_2, at a predetermined time interval until H_2 is full. After H_2 is full, any newly available state estimates are selected to replace existing state estimates in H_2 if the condition number of $\hat{\Sigma}^\top\hat{\Sigma}$, calculated using the post-replacement history stack, is smaller than the condition number of $\hat{\Sigma}^\top\hat{\Sigma}$ before the replacement. Once the data in H_2 are such that the condition number of $\hat{\Sigma}^\top\hat{\Sigma}$ is lower than a user-selected threshold, and a predetermined amount of time has passed since the last update of H_1, we set $H_1 = H_2$ and purge H_2 by setting it back to a zero matrix. Due to the purging algorithm, the time instances t_i corresponding to the data stored in the history stack H_1 are piecewise constant functions of time.

The IRL method developed in this paper requires that the behavior of the expert is optimal, which implies that $u(t) = K_{Ep}x(t)$ for all t. Since the true values of the state are not accessible, $K_{Ep}\hat{x}(t_i(t)) - u(t_i(t))$ cannot be expected to be equal to 0 for the data points stored in the history stack H_1. This discrepancy between $K_{Ep}\hat{x}(t_i(t))$ and $u(t_i(t))$ results in inaccurate estimates of equivalent solutions. Since the state estimates converge to the true state exponentially, the purging process described above ensures that the discrepancy $\max_{i=1,\cdots,N} \|K_{Ep}\hat{x}(t_i(t)) - u(t_i(t))\|$ is bounded by an exponentially decaying envelope, and so is the resulting inaccuracy in the estimation of an equivalent solution.

4 Regularized History Stack Observer for IRL Problems with Multiple Solutions

Due to purging and improved state estimates, $\hat{\Sigma}$ being full rank implies that Σ is eventually full rank, and as a result, $\Sigma W = \Sigma_u$ has a unique solution. As such, the explicit assumption that $\hat{\Sigma}$ is full rank implies an implicit assumption that the IRL problem admits a unique solution. Lack of uniqueness thus necessitates algorithms that can incorporate a rank-deficient $\hat{\Sigma}$. To that end, a regularized HSO (RHSO) is developed in this paper where the term $K_4(\hat{\Sigma}^\top\hat{\Sigma})^{-1}$ is replaced by a generic positive definite matrix to yield

$$K \coloneqq \begin{bmatrix} K_3 & 0_{n\times N+Nm} \\ 0_{P_S+P_Q+M-1\times L} & K_4\hat{\Sigma}^\top \end{bmatrix}, \quad (13)$$

where K_4 is a positive definite matrix of dimension P_S+P_Q+M-1. In the following lemmas and theorems, it is shown that under a novel informativity condition on the recorded data, the modification above leads to convergence to an equivalent solution when the IRL problem admits multiple solutions and convergence to the true cost functional of the expert, up to a scaling factor, when the IRL problem admits a unique solution. While the modification itself is relatively minor, the above somewhat surprising results are the key contributions of this work. The analysis requires a data informativity condition summarized in Definition 3 below.

Definition 3 *The signal* $(\hat{x}, u)$ *is called* finitely infor-

mative (FI) *if there exists a time instance* $T > 0$ *such that for some* $\{t_1, t_2, \dots, t_N\} \subset [0, T]$,

$$\text{Span}\{\hat{x}(t_i)\}_{i=1}^{N} = \mathbb{R}^n, \quad \Sigma_u \in \text{Range}(\hat{\Sigma}), \quad \textit{and}$$
$$\text{Span}\left\{\hat{x}(t_i)\hat{x}(t_i)^\top\right\}_{i=1}^{N} = \{\mathbb{Z} \in \mathbb{R}^{n\times n} | \mathbb{Z} = \mathbb{Z}^\top\}. \quad (14)$$

In addition, for a given $\epsilon > 0$, *if* $\min\{\text{eig}(XX^\top)\} > \epsilon$ *and* $\min\{\text{eig}(ZZ^\top)\} > \epsilon$, *where* $X := [\hat{x}(t_1), \dots, \hat{x}(t_N)]$, $Z := [\text{uvec}(\hat{x}(t_1)\hat{x}(t_1)^\top), \dots, \text{uvec}(\hat{x}(t_N)\hat{x}(t_N)^\top)] \in \mathbb{R}^{\frac{n(n+1)}{2}\times N}$, *and* $\text{uvec}(\hat{x}(t_i)\hat{x}(t_i)^\top) \in \mathbb{R}^{\frac{n(n+1)}{2}}$ *denotes vectorization of the upper triangular elements of the symmetric matrix* $\hat{x}(t_i)\hat{x}(t_i)^\top \in \mathbb{R}^{n\times n}$, *then* $(\hat{x}, u)$ *is called* $\epsilon-$finitely informative ($\epsilon-$FI).

Remark 4 *The three FI conditions in Definition 3 are utilized in the subsequent analysis to show that as the equivalence metric converges to zero, the corresponding weight estimates converge to an equivalent solution.*

(1) The condition $\text{Span}\{\hat{x}(t_i)\}_{i=1}^{N} = \mathbb{R}^n$ *is an excitation-like condition that requires the state estimates stored in the history stack to be linearly independent. This condition is not restrictive in general, however it can fail if the system has trajectories that are confined to a subspace of dimension less than* n. *This condition can be monitored online by ensuring that the minimum eigenvalue of* $XX^\top$ *is strictly positive, and as shown in Fig. 5, it is met in the simulation study.*

(2) The condition $\text{Span}\left\{\hat{x}(t_i)\hat{x}(t_i)^\top\right\}_{i=1}^{N} = \{\mathbb{Z} \in \mathbb{R}^{n\times n} | \mathbb{Z} = \mathbb{Z}^\top\}$ *is a sufficient condition for* $\hat{x}(t_i)^\top M\hat{x}(t_i) = 0, \forall i = 1, \cdots, N$ *to imply* $M = 0$. *It is not clear how restrictive this condition is, but it can be verified online by ensuring that the minimum eigenvalue of the matrix* $ZZ^\top$ *defined above is strictly positive. As shown in Fig. 6 this condition is met in the simulation study.*

(3) The condition $\Sigma_u \in \text{Range}(\hat{\Sigma})$ *is met provided at least one set of weights* $\hat{W}$ *satisfies* $\Sigma_u = \hat{\Sigma}\hat{W}$, *and as such, is not restrictive. If the IRL problem has a unique solution, then this condition is trivially met whenever* $N \geq P_S + P_Q + M - 1$ *and* $\hat{\Sigma}$ *is full rank. Furthermore, this condition can be verified online using the fact that* $\Sigma_u \in \text{Range}(\hat{\Sigma}) \iff \text{Rank}\left(\begin{bmatrix}\Sigma_u & \hat{\Sigma}\end{bmatrix}\right) = \text{Rank}(\hat{\Sigma})$. *Since the expert is assumed to be optimal,* $\Sigma_u = \Sigma W^*$, *and as a result,* $\Sigma_u \in \text{Range}(\Sigma)$. *Due to improving state estimates and the purging algorithm,* $\hat{\Sigma}$ *converges to* Σ, *and as a result, there exists* $T > 0$ *such that* $\Sigma_u \in \text{Range}(\hat{\Sigma})$ *for all* $t \geq T$. *As shown in Fig. 7 this condition is met in the simulation study.*

If the optimal trajectories of the expert do not meet the excitation conditions, an excitation signal can be added to the control input of the expert. As long as the excitation signal is known to the learner, the learner can infer the optimal control input of the expert needed to implement the developed RHSO.

Remark 5 *In the case of noisy measurements, the feedback gains* K_3 *and* $K_4\hat{\Sigma}^\top$ *in* (13) *can be replaced by Kalman gains. While empirical evidence suggests that the use of the Kalman gain results in improved performance (see [24, Section 2.3.3]), the stability guarantees in this paper are for deterministic systems with* K *selected according to* (12). *Extension of the developed stability guarantees to the case where the measurements are noisy and* K *is the Kalman gain is out of the scope of this paper.*

The following technical lemma is needed to prove convergence of the equivalence metric to zero.

Lemma 6 *If* $\hat{\Sigma}$ *and* Σ_u *satisfy* (14), *then* $\Omega_\Delta \cap \text{Null}(\hat{\Sigma}^\top) = \{0\}$, *where* $\Omega_\Delta := \left\{\Delta \in \mathbb{R}^{N(m+1)} \mid \Delta = \Sigma_u - \hat{\Sigma}\hat{W}, \textit{ for some } \hat{W} \in \mathbb{R}^{P_S+P_Q+M-1}\right\}$.

PROOF. If $\Delta \in \text{Null}(\hat{\Sigma}^\top)$, then $\hat{\Sigma}^\top\Delta = 0$. In addition, if $\Delta \in \Omega_\Delta$, then exists a $\hat{W}$ such that $\hat{\Sigma}^\top\left(\Sigma_u - \hat{\Sigma}\hat{W}\right) = 0$. The FI condition in (14) implies the existence of some W' such that $\Sigma_u = \hat{\Sigma}W'$. Therefore, $\hat{\Sigma}^\top\left(\hat{\Sigma}W' - \hat{\Sigma}\hat{W}\right) = 0$. As a result, $\hat{\Sigma}W' - \hat{\Sigma}\hat{W} \in \text{Null}(\hat{\Sigma}^\top)$. By definition of the range space, $\hat{\Sigma}W' - \hat{\Sigma}\hat{W} \in \text{Range}(\hat{\Sigma})$. Since $\text{Range}(\hat{\Sigma}) = (\text{Null}(\hat{\Sigma}^\top))^\perp$ [22, Section 4.1], $\hat{\Sigma}W' - \hat{\Sigma}\hat{W} \in (\text{Null}(\hat{\Sigma}^\top))^\perp \cap \text{Null}(\hat{\Sigma}^\top)$. Therefore, $\hat{\Sigma}W' - \hat{\Sigma}\hat{W} = 0$, which implies that $\Delta = 0$. □

Theorem 7 below shows that for given *fixed* matrices $\hat{\Sigma}$ and Σ_u that satisfy (14), if the weights $\hat{W}$ are updated using the update law in (11), then the equivalence metric Δ converges to the origin.

Theorem 7 *Let* $\Delta := \Sigma_u - \Sigma\hat{W}$. *If* $\Sigma_u \in \text{Null}(\hat{\Sigma}^\top)^\perp$, *the gain* K *is selected according to* (13), *and the weights* $\hat{W}$ *are updated using the update law in* (11), *then* $\lim_{t\to\infty}\Delta(t) = 0$. *In addition if full state information is available (i.e.,* $\hat{x} = x$ *and as a result,* $\hat{\Sigma} = \Sigma$), $\Delta = 0$, $\text{Span}\{x(t_i)\}_{i=1}^{N} = \mathbb{R}^n$, $\text{Span}\{x(t_i)x(t_i)^\top\}_{i=1}^{N} = \{\mathbb{Z} \in \mathbb{R}^{n\times n} | \mathbb{Z} = \mathbb{Z}^\top\}$, *and if the matrix* $\hat{R}$, *extracted from* $\hat{W}$, *is invertible, then the matrices* $\hat{Q}$, $\hat{S}$, *and* $\hat{R}$, *extracted from* $\hat{W}$, *constitute an equivalent solution of the IRL problem.*

PROOF. Using the update law in (11), the time-derivative of Δ can be expressed as

$$\dot{\Delta} = -\hat{\Sigma}K_4\hat{\Sigma}^\top\Delta. \quad (15)$$

Consider the positive definite and radially unbounded candidate Lyapunov function $V : \mathbb{R}^{N(m+1)} \to \mathbb{R}$ defined as

$$V(\Delta) = \frac{1}{2}\Delta^\top \Delta. \tag{16}$$

The orbital derivative of V along the solutions of (15) is given by

$$\dot{V}(\Delta) = -\Delta^\top \hat{\Sigma} K_4 \hat{\Sigma}^\top \Delta. \tag{17}$$

Note that all points in null space of $\Sigma^\top$ are equilibrium points of (15). Since $\Sigma^\top$ is not assumed to be full rank, $\text{Null}(\Sigma^\top) \neq \{0\}$. As a result, if $\Sigma^\top$ is not full rank, then the origin cannot be an asymptotically stable equilibrium point of (15). The analysis thus requires the invariance principle.

Since $\Omega_\Delta = \{\Sigma_u\} \ominus \text{Range}(\hat{\Sigma})$, where $\ominus$ denotes the Minkowski difference, it is easy to see that provided (14) holds, Ω_Δ is a subspace of $\mathbb{R}^{N(m+1)}$. Indeed, given $\alpha, \beta \in \mathbb{R}$ and $\Delta_1, \Delta_2 \in \Omega_\Delta$, with $\Delta_i = \Sigma_u - \hat{\Sigma}\hat{W}_i$ for $i = 1, 2$, we have $\alpha\Delta_1 + \beta\Delta_2 = \hat{\Sigma}_u + (\alpha + \beta - 1)\Sigma_u - \hat{\Sigma}(\alpha\hat{W}_1 + \beta\hat{W}_2)$. If (14) holds, then $\Sigma_u \in \text{Range}\,\hat{\Sigma}$, and as a result, $\alpha\Delta_1 + \beta\Delta_2 \in \Omega_\Delta$. Since Ω_Δ is a subspace of a finite dimensional topological space, it is also closed.

If $\Delta_0 \in \Omega_\Delta$ then there exists $\hat{W}_0$ such that $\Delta_0 = \Sigma_u - \hat{\Sigma}\hat{W}_0$. Let $t \mapsto \hat{W}_{\Delta_0}(t)$ be a solution of (11) starting from $\hat{W}_0$ with the interval of existence $\mathcal{I}$. For almost all $t \in \mathcal{I}$, we have $\dot{\hat{W}}_{\Delta_0} = K_4\hat{\Sigma}^\top(\Sigma_u - \hat{\Sigma}\hat{W}_{\Delta_0})$, which implies $\frac{\mathrm{d}}{\mathrm{d}t}(\Sigma_u - \hat{W}_{\Delta_0}) = -K_4\hat{\Sigma}^\top(\Sigma_u - \hat{\Sigma}\hat{W}_{\Delta_0})$. Letting $\Delta_{\Delta_0} = \Sigma_u - \hat{\Sigma}\hat{W}_{\Delta_0}$, it can be concluded that for almost all $t \in \mathcal{I}$, $\dot{\Delta}_{\Delta_0}(t) = -K_4\hat{\Sigma}^\top\Delta_{\Delta_0}(t)$. That is, $t \mapsto \Delta_{\Delta_0}(t)$ is a solution of (15) on $\in \mathcal{I}$, starting from Δ_0. Uniqueness of solutions then implies that $t \mapsto \Delta_{\Delta_0}(t)$ is the only solution of (15) on $\in \mathcal{I}$ starting from Δ_0. Using continuity of $t \mapsto \Delta_{\Delta_0}(t)$ along with the facts that Ω_Δ is closed and $\Delta_{\Delta_0}(t) \in \Omega_\Delta$ for almost all $t \in \mathcal{I}$, it can be concluded that $\Delta_{\Delta_0}(t) \in \Omega_\Delta$ for all $t \in \mathcal{I}$. As a result, Ω_Δ is positively invariant with respect to (15).

For any $c > 0$, the sublevel set $\Omega_c := \{\Delta \in \mathbb{R}^{N(m+1)} | V(\Delta) \leq c\}$ is compact. From (17), we conclude that Ω_c is positively invariant with respect to (15). As a result, $\Omega := \Omega_c \cap \Omega_\Delta$ is also positively invariant with respect to (15). Since Ω_c is compact and Ω_Δ is closed, Ω is also compact. The invariance principle [12, Theorem 4.4] can thus be invoked to conclude that all trajectories starting in Ω converge to the largest invariant subset of $\{\Delta \in \Omega \mid \dot{V}(\Delta) = 0\}$.

The set $\{\Delta \in \Omega | \dot{V}(\Delta) = 0\}$, is equal to $\text{Null}(\hat{\Sigma}^\top) \cap \Omega$ as $\hat{\Sigma}^\top\Delta = 0$ only when $\Delta \in \text{Null}(\hat{\Sigma}^\top)$. Furthermore, from Lemma 6, provided $\Sigma_u \in (\text{Null}(\hat{\Sigma}^\top))^\perp$, the only Δ that can be a member of $\text{Null}(\hat{\Sigma}^\top) \cap \Omega_\Delta$ is $\Delta = 0$. Since the set $\{0\}$ is positively invariant with respect to (15), it is also the largest invariant subset of $\{\Delta \in \Omega | \dot{V}(\Delta) = 0\}$.

As a result, by the invariance principle, all trajectories that start in Ω converge to the origin. Since V is radially unbounded, Ω_c can be selected to be large enough to include any initial condition in Ω_Δ. Thus, all solutions of (15) that start in Ω_Δ converge to the origin. In particular, Δ converges to zero along the solutions of the update law in (11).

To prove equivalence when $\Delta = 0$, the equality $\hat{R}^{-1}B^\top\hat{S} = K_{Ep}$ must be established. Indeed, if $\{x(t_i)\}_{i=1}^N$ spans $\mathbb{R}^n$ there is a unique matrix K that satisfies $u(t_i) = Kx(t_i)$ for all $i = 1, \ldots, N$. Letting $U = [u(t_1), \ldots, u(t_N)]$ and $X = [x(t_1), \ldots, x(t_N)]$, this unique matrix is given by $K = UX^\top(XX^\top)^{-1}$. It is also known that because the behavior of the expert is optimal, the observed data satisfy $u(t_i) = -K_{Ep}x(t_i)$ for all $i = 1, \ldots, N$. Since $\Delta = 0$, the observed data points satisfy $u(t_i) = -\hat{R}^{-1}B^\top\hat{S}x(t_i)$ for all $i = 1, \ldots, N$. Since there is only one matrix K that satisfies $u(t_i) = -Kx(t_i)$ for all $i = 1, \ldots, N$, all three of the matrices above must be equal, i.e., $K = K_{Ep} = \hat{R}^{-1}B^\top\hat{S}$.

The fact that if $\Delta = 0$ then $x(t_i)^\top\left(A^\top\hat{S} + \hat{S}A - \hat{S}B\hat{R}^{-1}B^\top\hat{S} + \hat{Q}\right)x(t_i) = 0$ holds for all points in H_1 is immediate from the construction of Δ. Furthermore, with a slight modification of the proof from [21], $(\hat{Q}, \hat{S}, \hat{R})$ can be proven to satisfy the ARE if $\Delta = 0$ and $\{x(t_i)x(t_i)^\top\}_{i=1}^N$ spans all symmetric matrices. To that end, let e_i be the basis vector of zeros with a one in the i^{th} position such that $e_je_k^\top + e_ke_j^\top = \sum_{i=1}^N \alpha_i x(t_i)x(t_i)^\top$ for some $\alpha_1 \cdots \alpha_N \in \mathbb{R}$. Rewriting (4) with $\hat{M} = \left(A^\top\hat{S} + \hat{S}A - \hat{S}B\hat{R}^{-1}B^\top\hat{S} + \hat{Q}\right)$, $\sum_{i=1}^N \alpha_i x(t_i)^\top\hat{M}x(t_i) = \sum_{i=1}^N\sum_{p=1}^n\sum_{q=1}^n \alpha_i x_{i,p}\hat{M}_{p,q}x_{i,q} = \sum_{i=1}^N\sum_{p=1}^n \hat{M}_{p,q}\sum_{q=1}^n \alpha_i x_{i,p}x_{i,q}$. Now, for any fixed j, k, select $\{\alpha_i\}_{i=1}^N$ such that $\sum_{i=1}^N \alpha_i x(t_i)x(t_i)^\top = e_je_k^\top + e_ke_j^\top$, where $\sum_{i=1}^N \alpha_i x(t_i)x(t_i)^\top = \begin{cases} 1 & \text{if } p = j, q = k, \\ 1 & \text{if } p = k, q = j, \\ 0 & \text{otherwise.} \end{cases}$ As a result, $\sum_{i=1}^N\sum_{p=1}^n \hat{M}_{p,q}\sum_{q=1}^n \alpha_i x_{i,p}x_{i,q} = e_k^\top\hat{M}e_j + e_j^\top\hat{M}e_k = \hat{M}_{j,k} + \hat{M}_{k,j} = 2\hat{M}_{j,k} = 0$. Since j and k were arbitrary, $\hat{M} = 0$. That is, the tuple $(\hat{Q}, \hat{S}, \hat{R})$ satisfies the ARE and constitutes an equivalent solution of the IRL problem. □

Remark 8 *The invertibility of $\hat{R}$ is needed for $\hat{K}_P$ to be well-defined. While this is difficult to ensure a priori in general, it can be guaranteed in the specific case where R is diagonal by using a projection operator to ensure that all diagonal elements of $\hat{R}$ remain positive. In this case, the weights are updated using the update law $\dot{\hat{W}} = \text{Proj}\left(K_4\hat{\Sigma}^\top\Delta\right)$, where $\text{Proj}(\cdot)$ denotes smooth projection (see Appendix E of [13]) onto the convex set*

$\mathbb{R}^{P_S} \times \mathbb{R}^{P_Q} \times \mathbb{R}^{m-1}_{\geq \kappa}$, *where* $\mathbb{R}^{m-1}_{\geq \kappa}$ *denotes the set of* $(m-1)-$*dimensional vectors that are element-wise larger than* κ *and* $\kappa > 0$ *is a lower bound for the diagonal entries of* R. *The resulting Lyapunov derivative is* $\dot{V}(\Delta) = -\Delta^\top \hat{\Sigma}\, \mathrm{Proj}\left(K_4 \hat{\Sigma}^\top \Delta\right)$. *Invoking Lemma E.1 from [13], it can be concluded that* $\dot{V}(\Delta) \leq -\Delta^\top \hat{\Sigma} K_4 \hat{\Sigma}^\top \Delta$. *The rest of the analysis then remains unchanged.*

Theorem 7 can be used to obtain the final result summarized in the definition and the theorem below.

Definition 9 *Given* $\varpi \geq 0$ *A solution* $(\hat{Q}, \hat{S}, \hat{R})$ *to the IRL problem is called an* $\varpi-$*equivalent solution of the IRL problem if* $\left\|\hat{M}\right\| \leq \varpi$, *where* $\hat{M} = A^\top \hat{S} + \hat{S}A - \hat{S}B\hat{R}^{-1}B^\top \hat{S} + \hat{Q}$, *and optimization of the performance index* J, *with* $Q = \hat{Q}$ *and* $R = \hat{R}$, *results in a feedback matrix,* $\hat{K}_p := \hat{R}^{-1}B^\top \hat{S}$, *that satisfies* $\left\|\hat{K}_p - K_{Ep}\right\| \leq \varpi$.

Due to the purging algorithm described in Section 3.2, the time instances t_i corresponding to the data stored in the history stack H_1 are piecewise constant functions of time, where $t_1(t)$ denotes the time instance when the oldest datum in the history stack was recorded. The corollary below requires $\liminf_{t\to\infty} t_1(t)$ to be large enough, which translates into the requirement that the excitation in the trajectories of the expert lasts long enough to allow sufficiently many purging events.

The exact lower bound on $\liminf_{t\to\infty} t_1(t)$ needed for convergence to a $\varpi-$equivalent solution is characterized in the proof of Theorem 10 below. The lower bound depends on the value of ϖ, the norm of the feedback gain K_{Ep} of the expert, the user-selected poles of $A-K_3C$, the user-selected gain matrix K_4, the condition numbers of the data matrices X and Z introduced in Definition 3. If $(\hat{x}, u)$ is ϵ-FI, the lower bounds $\min\{\mathrm{eig}(X(t)X(t)^\top)\} > \epsilon$ and $\min\{\mathrm{eig}(Z(t)Z(t)^\top)\} > \epsilon$, for some $\epsilon > 0$ and all $t \geq \underline{T}$, can be easily ensured using a modified history stack management algorithm that maximizes the minimum eigenvalues of $X(t)X(t)^\top$ and $Z(t)Z(t)^\top$.

Theorem 10 *Let* $\underline{T} \geq 0$ *denote the first time instant when* H_1 *is updated. Given* $\varpi > 0$ *if* $\liminf_{t\to\infty} t_1(t)$ *is large enough,* $\Sigma_u(t) \in \mathrm{Null}(\hat{\Sigma}^\top(t))^\perp$ *for all* $t \geq \underline{T}$, K_3 *is selected so that* $A - K_3C$ *is Hurwitz,* $\min\{\mathrm{eig}(X(t)X(t)^\top)\} > \epsilon$ *and* $\min\{\mathrm{eig}(Z(t)Z(t)^\top)\} > \epsilon$, *for some* $\epsilon > 0$ *and all* $t \geq \underline{T}$, *with* X *and* Z *as introduced in Definition 3, and if there exist a constant* $0 \leq \underline{R} < \infty$ *such that the matrix* $\hat{R}(t)$, *extracted from* $\hat{W}(t)$ *is invertible with* $\|\hat{R}^{-1}(t)\| \leq \underline{R}$ *for all* $t \geq \underline{T}$, *then the matrices* $\hat{Q}$, $\hat{S}$, *and* $\hat{R}$, *extracted from* $\hat{W}$, *converge to a* $\varpi-$*equivalent solution of the IRL problem.*

PROOF. The dynamics in (15) ensure that $\Delta(t)$ is bounded for all t. The control residual error established in (8) can be manipulated into the form $\sigma_{\Delta'_u}(\hat{x}(t_i(t)), u(t_i(t)))\,\hat{W}'(t) = \hat{R}(t)\left(\tilde{K}_P(t)\hat{x}(t_i(t)) + K_{Ep}\tilde{x}(t_i(t))\right)$, where $\tilde{K}_P(t) := \hat{R}^{-1}(t)B^\top \hat{S}(t) - K_{Ep}$ and $\tilde{x}(t_i(t)) := x(t_i(t)) - \hat{x}(t_i(t))$. Using the triangle inequality $\left\|\tilde{K}_P(t)\hat{x}(t_i(t))\right\| \leq \left\|\hat{R}^{-1}(t)\sigma_{\Delta'_u}(\hat{x}(t_i(t)), u(t_i(t)))\,\hat{W}'(t)\right\| + \|K_{Ep}\tilde{x}(t_i(t))\|$.

Note that if $\mathrm{Span}\{\hat{x}(t_i(t))_{i=1}^N\} = \mathbb{R}^n$, and in particular, if $\min\{\mathrm{eig}(X(t)X(t)^\top)\} > \epsilon$ then $\exists c > 0$, independent of t, such that $\left\|\tilde{K}_P(t)\hat{x}(t_i(t))\right\| \leq \frac{\varpi}{c}, \forall i$, implies $\left\|\tilde{K}_P(t)\right\| \leq \varpi$. Select $\overline{T}_1$ large enough such that the equivalence metric $\Delta(t)$ satisfies $\left\|\sigma_{\Delta'_u}(\hat{x}(t_i(t)), u(t_i(t)))\,\hat{W}'(t)\right\| \leq \frac{\varpi}{2c\underline{R}}$, for all i and for all $t \geq \overline{T}_1$. Such a $\overline{T}_1$ exists since by Theorem 7, $\lim_{t\to\infty}\Delta(t) = 0$. Select $\overline{T}_2$ large enough so that the state estimation error $\tilde{x}(t_i(t))$ satisfies $\|\tilde{x}(t_i(t))\| \leq \frac{\varpi}{2c\|K_{Ep}\|}$ for all $t \geq \overline{T}_2$. Since $\lim_{t\to\infty}\tilde{x}(t) = 0$, existence of of such a $\overline{T}_2$ follows if $t_1(\overline{T}_2)$ is large enough. Letting $\overline{T} = \max\{\overline{T}_1, \overline{T}_2\}$, it can be concluded that for all $t \geq \overline{T}$, $\left\|\tilde{K}_P(t)\hat{x}(t_i(t))\right\| \leq \frac{\varpi}{c}$, which implies $\left\|\tilde{K}_P(t)\right\| \leq \varpi$.

The inverse Bellman error established in (9) can be manipulated into $\sigma_{\delta'}(\hat{x}(t_i(t)), u(t_i(t)))\,\hat{W}'(t) = \hat{x}(t_i(t))^\top \hat{M}\hat{x}(t_i(t)) + g\left(\hat{K}_P(t), \hat{x}(t_i(t)), K_{Ep}, x(t_i(t))\right)$, where the function g satisfies[1] $g = O\left(\left\|\tilde{K}_P(t)\right\| + \|\tilde{x}(t_i(t))\|\right)$. Using the triangle inequality, $\left|\hat{x}(t_i(t))^\top \hat{M}(t)\hat{x}(t_i(t))\right| \leq \left|\sigma_{\delta'}(\hat{x}(t_i(t)), u(t_i(t)))\,\hat{W}'\right| + \left|g\left(\hat{K}_P(t), \hat{x}(t_i(t)), K_{Ep}, x(t_i(t))\right)\right|$, where $\hat{M}(t) = A^\top \hat{S}(t) + \hat{S}(t)A - \hat{S}(t)B\hat{R}^{-1}(t)B^\top \hat{S}(t) + \hat{Q}(t)$

Since $g = O\left(\left\|\tilde{K}_P(t)\right\| + \|\tilde{x}(t_i(t))\|\right)$ and $\left|\sigma_{\delta'}(\hat{x}(t_i(t)), u(t_i(t)))\,\hat{W}'\right| \leq \|\Delta(t)\|$, a construction similar to the one in the previous paragraph can be used to show that given any $\varepsilon > 0$, that there exists a $\overline{T}$ such that for all $t \geq \overline{T}$ and for all $i = 1, \ldots, N$, $\left|\hat{x}(t_i(t))^\top \hat{M}(t)\hat{x}(t_i(t))\right| \leq \varepsilon$.

Equivalence of matrix norms implies that there exists $c > 0$, independent of t, such that if $\left|\hat{M}_{j,k}(t)\right| \leq \varpi/c$ for all $j, k = 1, \cdots, n$, then $\left\|\hat{M}(t)\right\| \leq \varpi$. As a result, to complete the proof of the theorem, it suffices to construct a $\overline{T}$ such that for all $t \geq \overline{T}$ and for all $j, k = 1, \cdots, n$,

[1] For a positive function g, $f = O(g)$ if there exists a constant M such that $\|f(x)\| \leq Mg(x), \forall x$

$\left|\hat{M}_{j,k}(t)\right| \leq \frac{\varpi}{c}$. To construct such a $\overline{T}$, an ε is constructed such that $\left|\hat{x}(t_i(t))^\top \hat{M}(t)\hat{x}(t_i(t))\right| \leq \varepsilon, i = 1, \ldots, N$ implies $\left|\hat{M}_{j,k}(t)\right| \leq \frac{\varpi}{c}, \forall j, k = 1, \cdots, n$. Existence of the required $\overline{T}$ then follows from the discussion in the previous paragraph.

Let e_i be the basis vector of zeros with a one in the i^{th} position. For a fixed j and k, selecting constants $\alpha_{1,j,k} \cdots \alpha_{N,j,k} \in \mathbb{R}$ and rewriting (4), we have $\sum_{i=1}^N \alpha_{i,j,k}\hat{x}(t_i(t))^\top \hat{M}(t)x(t_i(t)) = \sum_{i=1}^N \sum_{p=1}^n \sum_{q=1}^n \alpha_{i,j,k}\hat{x}_p(t_i(t))\hat{M}_{p,q}(t)\hat{x}_q(t_i(t)) = \sum_{i=1}^N \sum_{p=1}^n \hat{M}_{p,q}(t) \sum_{q=1}^n \alpha_{i,j,k}\hat{x}_p(t_i(t))\hat{x}_q(t_i(t))$. If $\text{Span}\{\hat{x}(t_i(t))\hat{x}(t_i(t))^\top\}_{i=1}^N = \{\mathbb{Z} \in \mathbb{R}^{n\times n} | \mathbb{Z} = \mathbb{Z}^\top\}$, then for any fixed j, k, we can select $\{\alpha_{i,j,k}(t)\}_{i=1}^N$ such that $\sum_{i=1}^N \alpha_{i,j,k}(t)\hat{x}(t_i(t))\hat{x}(t_i(t))^\top = e_j e_k^\top + e_k e_j^\top$, that is, the (p, q) element of $\sum_{i=1}^N \alpha_{i,j,k}(t)\hat{x}(t_i(t))\hat{x}(t_i(t))^\top$ is 1 if $p = j$ and $q = k$, it is also 1 if $p = k$ and $q = j$, and it is zero otherwise. As a result, $\sum_{i=1}^N \sum_{p=1}^n \hat{M}_{p,q}(t) \sum_{q=1}^n \alpha_{i,j,k}(t)\hat{x}_p(t_i(t))\hat{x}_q(t_i(t)) = e_k^\top \hat{M}(t)e_j + e_j^\top \hat{M}(t)e_k = \hat{M}_{j,k}(t) + \hat{M}_{k,j} = 2\hat{M}_{j,k}(t)$. If $\min\{\text{eig}(Z(t)Z(t)^\top)\} > \epsilon$ then the coefficients $\alpha_{i,j,k}$ are bounded such that $\sup_{t\geq T} \max_{i,j,k}(\{|\alpha_{i,j,k}(t)|\}_{i,j,k=1}^{N,n,n}) \leq \alpha < \infty$ for some $\alpha > 0$.

Select $\varepsilon = \frac{2\varpi}{c\alpha N}$ and note that $\left\|\hat{x}(t_i(t))^\top \hat{M}(t)\hat{x}(t_i(t))\right\| \leq \frac{2\varpi}{c\alpha N}, \forall i = 1, \cdots, N$ implies that for all $j, k = 1, \ldots, n$, $\left|2\hat{M}_{j,k}(t)\right| = \left|\sum_{i=1}^N \alpha_{i,j,k}(t)\hat{x}(t_i(t))^\top \hat{M}(t)\hat{x}(t_i(t))\right| \leq \alpha N \max_i \left(\left\{\left\|\hat{x}(t_i(t))^\top \hat{M}(t)\hat{x}(t_i(t))\right\|\right\}_{i=1}^N\right) \leq \frac{2\varpi}{c}$, which implies that for all $j, k = 1, \ldots, n$, $\left|\hat{M}_{j,k}(t)\right| \leq \frac{\varpi}{c}$, which completes the proof of the theorem. □

5 Simulations

5.1 Methods and Results

To demonstrate the ability of the developed method to obtain equivalent solutions to IRL problems that admit multiple solutions, an IRL problem that has a product structure is constructed and linearly transformed. The results in [10] ensure that the resulting transformed IRL problem admits multiple solutions.

The state space model is given by

$$A = \begin{bmatrix} -0.2 & 0.4 & 1.6 \\ 3.7 & 1.6 & -3.1 \\ -3.2 & 0.4 & 4.6 \end{bmatrix}, B = \begin{bmatrix} 1 & 2 & -1 \\ -1 & 3 & 4 \\ 1 & 2 & -3 \end{bmatrix}, C = \begin{bmatrix} 1.7 & -0.4 & -1.1 \\ -0.1 & 0.2 & 0.3 \\ 0.5 & 0 & -0.5 \end{bmatrix}.$$

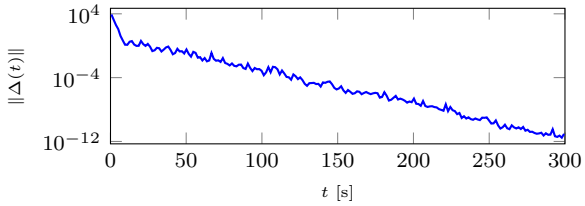

Fig. 1. A log-scale plot of the 2-norm of Δ as a function of time.

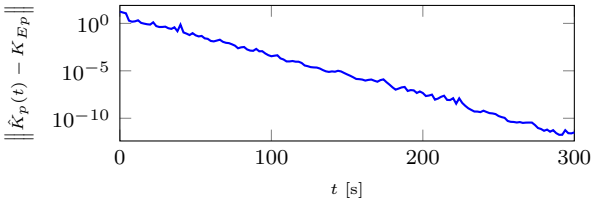

Fig. 2. A log-scale plot of the induced 2-norm of the error between the estimated feedback gain and the feedback gain of the expert as a function of time.

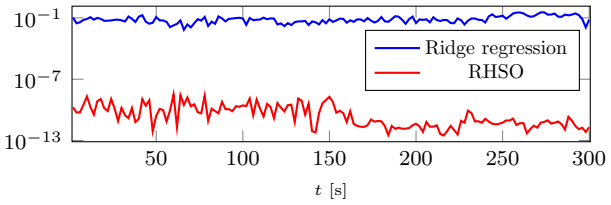

Fig. 3. A log-scale plot of the 2-norm of the error between the state trajectory of the expert and the state trajectory of the learner under the learned feedback gain for a problem that admits multiple solutions. The red trajectory corresponds to the feedback gain learned using the RHSO and the blue trajectory corresponds to the feedback gain computed using offline ridge regression.

The expert implements a feedback policy that minimizes the cost functional in (3) with [2]

$$Q = \begin{bmatrix} 12.32 & -2.74 & -8.26 \\ -2.74 & 0.68 & 1.82 \\ -8.26 & 1.82 & 5.68 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 7 \end{bmatrix}. \tag{18}$$

To ensure that the history stack satisfies the sufficient condition in (14), an excitation signal comprised of a sum of 20 sinusoidal signals is added to the input of the expert in (1). The magnitudes are set to 0.5 and the frequencies and phases are randomly selected from the ranges 0.001 Hz to 1 Hz and 0 rad to π rad, respectively. Since the regressor $\hat{\Sigma}$ is a nonlinear function of $\hat{x}$, a precise characterization of the excitation signal needed to satisfy the finite informativity conditions in Definition 3 is difficult to obtain. Drawing inspiration from persistence

[2] The notation diag(v) represents a diagonal matrix with the elements of the vector v along the diagonal.

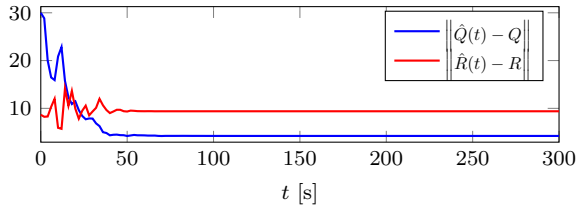

Fig. 4. A plot of the induced 2-norm of the error between the estimated $\hat{Q}$ (red) and $\hat{R}$ (blue) matrices and the Q and R matrices of the expert as a function of time.

of excitation results for linear regressors, the number of frequencies is selected to be higher than the number of unknown parameters, which in this example is 14. The excitation signal is assumed to be known to the learner, so it can be subtracted from the total input of the expert to infer the optimal input of the expert.

To facilitate comparison with ridge regression, the matrix K_4 is selected as $K_4 = (\hat{\Sigma}^\top \hat{\Sigma} + \epsilon I)^{-1}$. Data are added to the history stack every 0.05 seconds and the history stack is purged if it is full and either the condition number of $\hat{\Sigma}^\top \hat{\Sigma} + \epsilon I$ is smaller than 1×10^5, or 2 seconds have elapsed since the last purge. [3] The weights are $\hat{W}$ are randomly sampled from a standard normal distribution.

A Luenberger observer is utilized for state estimation by selecting the gain K_3 to place the poles of $(A - K_3 C)$ at $p_1 = -0.1$, $p_2 = -1.5$ and $p_3 = -2$ using the MATLAB "place" command. These values are selected by trial and error to achieve a sufficiently fast convergence rate for the Luenberger observer. The parameters of the RHSO are held constant for all simulations in this paper unless otherwise stated.

Fig. 1 demonstrates the convergence of Δ to the origin as per Theorem 7 and Fig. 2 demonstrates the convergence of the estimated feedback gain to a neighborhood of the feedback matrix of the expert, as per Theorem 10. Finally, Fig. 4 indicates that the cost functional converges to a functional that is different from that of the expert, confirming that the IRL problem under consideration admits multiple equivalent solutions.

Like most excitation conditions in reinforcement learning, this excitation condition cannot be guaranteed *a priori*. The best practice is to monitor whether it is met online. To examine whether the sufficient conditions detailed in Definition 3 hold, stem plots are generated that equal 1 when the conditions hold and 0 when they do not (see Figs. 5, 6, and 7).

[3] See [11] for further details on condition number minimization.

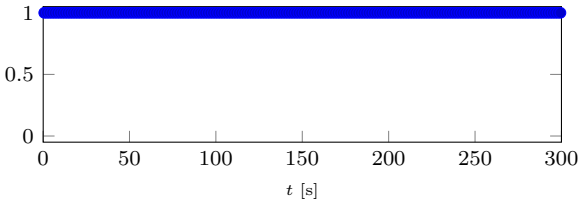

Fig. 5. This plot is equal to 1 if $\text{Span}\{\hat{x}(t_i(t))\}_{i=1}^{N} = \mathbb{R}^n$ and 0 otherwise.

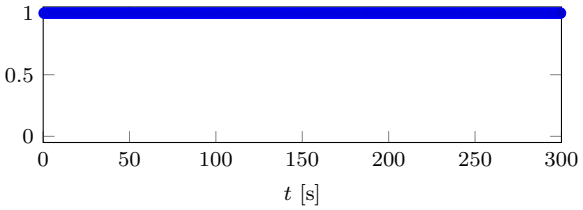

Fig. 6. This plot is equal to 1 if $\text{Span}\{\hat{x}(t_i)\hat{x}(t_i)^\top\}_{i=1}^{N} = \{\mathbb{Z} \in \mathbb{R}^{n \times n} \mid \mathbb{Z} = \mathbb{Z}^\top\}$ and 0 otherwise.

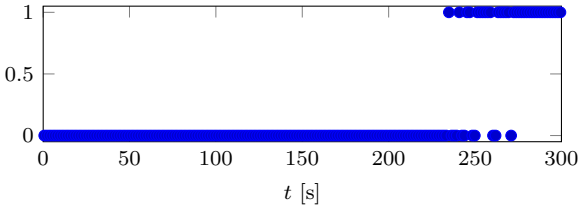

Fig. 7. This plot is equal to 1 if $\Sigma_u(t) \in \text{Range}(\hat{\Sigma}(t))$ and 0 otherwise.

5.2 Discussion

Each simulation shows the convergence of Δ to zero and the convergence of the estimated feedback matrix, $\hat{K}_P$, to the feedback matrix K_{Ep} of the expert. In all simulations, the RHSO converges to either an equivalent solution or the true cost functional of the expert. Therefore, the RHSO is a complete extension to the HSO [20] as it solves IRL problems with unique and non-unique solutions. The particular equivalent solution that the RHSO converges to depends on the initial estimates of the unknown weights $\hat{W}$.

As demonstrated by Fig. 4, convergence to an approximate equivalent solution is achieved in spite of failure to meet the FI condition throughout the simulation. The condition is met, however, at the end of the simulation. Fig. 4 thus indicates that the FI condition is sufficient but not necessary for the RHSO to converge to approximate equivalent solutions. When K_4 is selected as $(\hat{\Sigma}^\top \hat{\Sigma} + \epsilon I)^{-1}$, Δ converges to zero and either a unique or an equivalent solution is obtained, regardless of the magnitude of ϵ. This result is at odds with regularization used in ridge regression, where convergence with an $\epsilon-$dependent bound is obtained. Especially interesting

is the fact that offline ridge regression [23] using matrices Σ_u and $\hat{\Sigma}$ that contain all of the available data fail at finding a $\hat{W}$ that constitutes an equivalent solution to the IRL problem.

6 Conclusion

In this paper, a novel framework for the estimation of a cost functional is developed for IRL problems with multiple solutions. The developed technique is a modification of the HSO in [20]. This modification, while simple, requires a novel analysis approach. The analysis reveals new data-informativity conditions required for convergence of the update laws to an equivalent solution when multiple solutions are present. It is further shown that the RHSO is a proper extension of the HSO, in the sense that it converges to the true cost functional of the expert when the IRL problem has a unique solution.

Simulations demonstrate that the developed adaptive update laws are able to converge to equivalent solutions in IRL problems where offline ridge-regression fails to generate useful solutions. Future research will include applications of the developed method to real-world problems such as learning the cost function of pilots flying unmanned air vehicles using input-output measurements.

References

[1] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proc. Int. Conf. Mach. Learn.*, 2004.

[2] Saurabh Arora, Prashant Doshi, and Bikramjit Banerjee. Online inverse reinforcement learning under occlusion. In *Proc. Conf. Auton. Agents MultiAgent Syst.*, pages 1170–1178. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[3] Vrushabh S. Donge, Bosen Lian, Frank L. Lewis, and Ali Davoudi. Multi-agent graphical games with inverse reinforcement learning. *IEEE Trans. Control Netw. Syst.*, pages 1–12, 2022.

[4] Michael Herman, Volker Fischer, Tobias Gindele, and Wolfram Burgard. Inverse reinforcement learning of behavioral models for online-adapting navigation strategies. In *Proc. IEEE Int. Conf. Robot. Autom.*, pages 3215–3222, 2015.

[5] João P. Hespanha. *Linear systems theory.* Princeton University Press, 2009.

[6] Mahdi Imani and Seyede Fatemeh Ghoreishi. Scalable inverse reinforcement learning through multifidelity bayesian optimization. *IEEE Trans. Neural Netw. Learn. Syst.*, 33(8):4125–4132, 8 2022.

[7] Jairo Inga, Esther Bischoff, Timothy Molloy, Michael Flad, and Soren Hohmann. Solution sets for inverse non-cooperative linear-quadratic differential games. *IEEE Control Syst. Lett.*, 3(4):871–876, 10 2019.

[8] Jairo Inga, Andreas Creutz, and Sören Hohmann. Online inverse linear-quadratic differential games applied to human behavior identification in shared control. In *Proc. Eur. Control Conf.*, pages 323–360, 2021.

[9] Antony Jameson and Eliezer Kreindler. Inverse problem of linear optimal control. *SIAM J. Control*, 11(1):1–19, 1973.

[10] Frédéric Jean and Sofya Maslovskaya. Inverse optimal control problem: the linear-quadratic case. In *Proc. IEEE Conf. Decis. Control*, pages 888–893, 2018.

[11] Rushikesh Kamalapurkar. Linear inverse reinforcement learning in continuous time and space. In *Proc. Am. Control Conf.*, pages 1683–1688, Milwaukee, WI, USA, June 2018.

[12] Hassan K. Khalil. *Nonlinear systems.* Prentice Hall, Upper Saddle River, NJ, third edition, 2002.

[13] Miroslav Krstic, Ioannis Kanellakopoulos, and Peter V. Kokotovic. *Nonlinear and adaptive control design.* John Wiley & Sons, New York, NY, USA, 1995.

[14] Bosen Lian, Vrushabh S Donge, Frank L Lewis, Tianyou Chai, and Ali Davoudi. Data-driven inverse reinforcement learning control for linear multiplayer games. *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[15] Bosen Lian, Wenqian Xue, Frank L. Lewis, and Tianyou Chai. Online inverse reinforcement learning for nonlinear systems with adversarial attacks. *Int. J. Robust Nonlinear Control*, 31(14):6646–6667, 2021.

[16] Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proc. Int. Conf. Mach. Learn.*, pages 663–670. Morgan Kaufmann, 2000.

[17] Stuart Russell. Learning agents for uncertain environments (extended abstract). In *Proc. Conf. Comput. Learn. Theory*, 1998.

[18] Ryan Self, Kevin Coleman, He Bai, and Rushikesh Kamalapurkar. Online observer-based inverse reinforcement learning. arXiv:2011.02057v3, 2023.

[19] Ryan V. Self, Moad Abudia, S M Nahid Mahmud, and Rushikesh Kamalapurkar. Model-based inverse reinforcement learning for deterministic systems. *Automatica*, 140(110242):1–13, June 2022.

[20] Ryan V. Self, Kevin Coleman, He Bai, and Rushikesh Kamalapurkar. Online observer-based inverse reinforcement learning. *IEEE Control Syst. Lett.*, 5(6):1922–1927, December 2021.

[21] Quadratic form vanishing at certain points. https://math.stackexchange.com/q/3230096. accessed: 2019-05-17.

[22] Gilbert Strang. *Introduction to linear algebra.* Wellesley-Cambridge Press, fourth edition, 2009.

[23] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.*, 58:267–288, 1996.

[24] Jared Town. Nonuniqueness and equivalence in online inverse reinforcement learning with applications to pilot performance modeling. Master's thesis, Oklahoma State University, 2023.

[25] Jared Town, Zachary Morrison, and Rushikesh Kamalapurkar. Nonuniqueness and convergence to equivalent solutions in observer-based inverse reinforcement learning. In *Proc. Am. Control Conf.*, pages 3989–3994, July 2023.

[26] Wenqian Xue, Patrik Kolaric, Jialu Fan, Bosen Lian, Tianyou Chai, and Frank L Lewis. Inverse reinforcement learning in tracking control based on inverse optimal control. *IEEE Trans. Cybern.*, 2021.

[27] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proc. AAAI Conf. Artif. Intel.*, pages 1433–1438, 2008.

Detailed Explanation

A preliminary version of the approach developed in this article is published in the proceedings of the 2023 American Control Conference [25]. The definition of equivalence used in this paper is stronger than the one in [25], and as a result, the analysis that proves convergence to equivalent solutions is more involved than the analysis in [25]. In particular, Lemma 6 is new, the data informativity condition in Definition 3 and the proof of Theorem 7 are different, and the analysis of convergence to approximately equivalent solutions (Definition 9 and Theorem 10) is entirely new.

Response to Review Comments

Response to the comments by the associate editor

(1) The extension from paper [20] ([19] in the original manuscript) seems incremental. The paper [23] is not available, so the reviewers, the AE and the senior editor cannot check the extend of the contributions.

Response: The extension from [20] ([19] in the original manuscript) require an entirely new analysis approach. Analysis of weight estimation errors, as done in [20] ([19] in the original manuscript) is no longer appropriate in problems studied here, where the ideal weights do not exist, and as a result, there is no weight estimation error to be analyzed. We therefore respectfully disagree with the assessment that the extension is incremental. We now highlight this contribution in the introduction of the revised manuscript.

We apologize for the mix-up related to the conference version [23]. We intended to cite the conference paper that was uploaded to arXiv, but cited the submitted paper instead. The conference paper is now published and is duly cited in the revised manuscript.

(2) The proof of Theorem 7 can be referred as an application of LaSalle's principle. This can save one page in manuscript if removed. Same for Corollary 10 (Theorem 10 of the revised manuscript).

Response: We cannot use Corollary 4.2 from Khalil's textbook to prove Theorem 7 as suggested by the reviewer. In fact, the system in (15) does not have a globally (or even locally) asymptotically stable equilibrium point at the origin. Any perturbations outside of the subspace Ω_Δ are not guaranteed to return to the origin. As such, that corollary is not applicable, and we are forced to resort to the more general invariance principle. However, in responding to this comment, we realized that uniqueness of solutions to differential equations can be used to conclude that Ω_Δ is forward invariant, which makes the invariance principle applicable to this problem, and we do not need to replicate the proof of the invariance principle. The revised manuscript takes this approach, which simplifies the proof of Theorem 7.

Theorem 10 analyzes asymptotic behavior in the case where Δ is not equal to, but only approaches the origin. We do not see how that corollary can be proved using the invariance principle.

(3) Subsections 3.3, 5.2 and 5.3 can be removed since they do not have theoretical back up especially for 5.3.

Response: In response to this comment, we have removed the mentioned sections from the revised manuscript

(4) As a conclusion, there are several issues in the paper. If the authors decide to resubmit, they have to reduce the paper to brief by following the guidelines above as well as edit the manuscript to respond to the reviewers.

Response: We have edited the manuscript to address the reviewer comments. The revised manuscript is 10 pages, which is 2 pages longer than the typical brief. We respectfully request permission to pay for the two additional pages in the event the manuscript is accepted for publication.

Response to the comments by Reviewer 1

Review 1: This manuscript presents a modified method to determine a quadratic cost function for which the behavior of a given expert system is optimal. Different from a previous work by one of the authors, this method is applicable also when the desired cost function is not unique. The paper is not clearly written and the description of the method is not self-contained. I also have some reserves about the strength of the results. In the following, these issues are discussed in detail.

(1) The authors modified the method published in [20] ([19] in the original manuscript) to make it work for the case in which more than one cost function explains the behavior of the expert. However, in this manuscript the authors only made a superficial review of the results in [20] ([19] in the original manuscript). The consequence is that Section 3 is rather incoherent. For example, the presentation of (7) and (8) is confusing because they do not follow directly from (4) and (5) as claimed. Moreover, the sets of basis functions σ are not defined. This is very important, because (14) makes no sense except for a very specific selection of basis functions.

Response: Since we were asked to resubmit this paper as a brief paper, we were unable to add more details regarding the basis functions used in [20] ([19] in the original manuscript). Instead, we decided to keep the bare minimum detail necessary in Section 3 and we cite the arXiv paper [18], where the basis functions are presented, in detail, and some typographical errors in [20] ([19] in the original manuscript) are also rectified.

(2) Regarding (14), how the authors obtain this expression is incomprehensible unless the reader goes to [20] ([19] in the original manuscript). It is worth mentioning as well that, following the procedure in [20] ([19] in the original manuscript), I did not get exactly the same expression (14). I request the authors to provide the complete, step-by-step procedure to get (14).

Response: An error was found in [20] ([19] in the original manuscript) and has since been rectified, see [18]. Due to page limitations of a brief paper, we are unable to add a derivation in the revised manuscript, but we have included it here for the reviewer's benefit.

Using the assumption of an optimal expert, $x(\cdot)$ and $u(\cdot)$ satisify the HJB equation,

$$H\left(x(t), \nabla_x (V(x(t)))^\top, u(t)\right) = 0, \forall t \in \mathbb{R}_{\geq 0}$$

with the optimal control equation being

$$u^*(x(t)) = -\frac{1}{2} R^{-1} B^\top \nabla_x (V(x(t)))^\top$$

and the Hamiltonian defined as

$$H(x,p,u) \coloneqq p^\top (Ax+Bu) + x^\top Qx + u^\top Ru.$$

We know that the optimal cost is given by

$$V(x) = x^\top Sx$$

where S is a solution to the algebraic Riccati equation

$$A^\top S + SA - SBR^{-1}B^\top S + Q = 0.$$

To aid in the estimation, we do the following linear parameterizations $x^\top Sx = W_S^{*\top}\sigma_S(x)$, $x^\top Qx = W_Q^{*\top}\sigma_Q(x)$, $x^\top Rx = W_R^{*\top}\sigma_{R_1}(x)$, and $Ru = \sigma_{R_2}(u)W_R^*$, where the ideal weights are given by

$$\begin{aligned}
W_S^* &= \left[S_{11}, S_1^{(-1)}, S_{22}, S_2^{(-2)}, \ldots, S_{n-1}^{-(n-1)}, S_{nn}\right]^\top, \\
W_Q^* &= \left[Q_{11}, Q_1^{(-1)}, Q_{22}, Q_2^{(-2)}, \ldots, Q_{n-1}^{-(n-1)}, Q_{nn}\right]^\top, \text{ and} \\
W_R^* &= \left[R_{11}, R_1^{(-1)}, R_{22}, R_2^{(-2)}, \ldots, R_{m-1}^{-(m-1)}, R_{nn}\right]^\top,
\end{aligned}$$

and the basis functions are given by

$$\begin{aligned}
\sigma_S(x) = \sigma_Q(x) &\coloneqq [x_1^2, 2x_1x_2, 2x_1x_3, \ldots, 2x_1x_n, x_2^2, 2x_2x_3, 2x_2x_4, \ldots, x_{n-1}^2, \ldots, 2x_{n-1}x_n, x_n^2]^\top, \\
\sigma_{R1}(u) &\coloneqq [u_1^2, 2u_1u_2, 2u_1u_3, \ldots, 2u_1u_m, u_2^2, 2u_2u_3, 2u_2u_4, \ldots, u_{m-1}^2, \ldots, 2u_{m-1}u_m, u_m^2]^\top,
\end{aligned}$$

and

$$\sigma_{R2}(u) = \begin{bmatrix}
u^\top & 0_{1\times m-1} & 0_{1\times m-2} & \cdots & 0 \\
u_{(1)}e_2^m & \left(u^\top\right)^{(-1)} & 0_{1\times m-2} & \cdots & 0 \\
u_{(1)}e_3^m & u_{(2)}e_2^{m-1} & \left(u^\top\right)^{(-2)} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
u_{(1)}e_m^m & u_{(2)}e_{m-1}^{m-1} & u_{(3)}e_{m-2}^{m-2} & \cdots & \left(u^\top\right)^{-(m-1)}
\end{bmatrix}. \tag{19}$$

In (19), $u^{(-j)}$ denotes the vector u with the first j elements removed, e_i^j denotes a row vector of size j, with a one in the $i-$th position and zeros everywhere else, and $u_{(i)}$ denotes the $i-$th element of u.

The optimal control equation and HJB are then linearly parameterized as

$$0 = 2\sigma_{R2}(u)W_R^* + B^\top \left(\nabla_x \sigma_S(x)\right)^\top W_S^*,$$
$$0 = \nabla_x \left((W_S^*)^\top \sigma_S(x)\right)(Ax + Bu) + (W_Q^*)^\top \sigma_Q(x) + (W_R^*)^\top \sigma_{R1}(u).$$

Using estimates of the ideal weights, a control residual error and an inverse Bellman error are defined as

$$\Delta_u' \coloneqq 2\sigma_{R2}(u)\hat{W}_R + B^\top \left(\nabla_x \sigma_S(x)\right)^\top \hat{W}_S, \text{ and}$$
$$\delta' \coloneqq \nabla_x \left((\hat{W}_S)^\top \sigma_S(x)\right)(Ax + Bu) + (\hat{W}_Q)^\top \sigma_Q(x) + (\hat{W}_R)^\top \sigma_{R1}(u).$$

Separating out the estimated weights $\hat{W}' = \left[\hat{W}_S^\top, \hat{W}_Q^\top, \hat{W}_R^\top\right]^\top$ yields $\begin{bmatrix} \delta'\left(x, u, \hat{W}'\right) \\ \Delta_u'\left(x, u, \hat{W}'\right) \end{bmatrix} = \begin{bmatrix} \sigma_{\delta'}(x, u) \\ \sigma_{\Delta_u'}(x, u) \end{bmatrix} \hat{W}'$, where

$$\sigma_{\delta'}(x, u) = \left[(Ax + Bu)^\top (\nabla_x \sigma_S(x))^\top, \sigma_Q(x)^\top, \sigma_{R1}(u)^\top\right], \text{ and}$$
$$\sigma_{\Delta_u'}(x, u) = \left[B^\top (\nabla_x \sigma_S(x))^\top, 0_{m \times P_Q}, 2\sigma_{R2}(u)\right].$$

Selecting $r_1 = 1$ and removing it from the weight vector $\hat{W}'$ in (6) and (7) yields scale-aware definitions of the control residual error and the inverse Bellman error, given by

$$\begin{bmatrix} \delta\left(x, u, \hat{W}\right) \\ \Delta_u\left(x, u, \hat{W}\right) \end{bmatrix} = \begin{bmatrix} \sigma_\delta(x, u) \\ \sigma_{\Delta_u}(x, u) \end{bmatrix} \begin{bmatrix} \hat{W}_S \\ \hat{W}_Q \\ \hat{W}_R^- \end{bmatrix} + \begin{bmatrix} u_1^2 r_1 \\ 2u_1 r_1 \\ 0_{m-1 \times 1} \end{bmatrix}, \tag{20}$$

where $\hat{W}_R^-$ is a copy of $\hat{W}_R$ with the first element removed, σ_δ is defined as $\sigma_{\delta'}$, with the $(P_S + P_Q + 1)-$th element removed, and σ_{Δ_u} is defined as $\sigma_{\Delta_u'}$, with the $(P_S + P_Q + 1)-$th column removed.

(3) The description of the 'purging' procedure (last paragraph of page 4) is also unclear. The authors write that a new state estimate replaces an existing one in H_2 if the replacement decreases a condition number. However,

they also state that H_2 is set to a zero matrix every several time instances. Thus, there is no state to replace. This procedure must be clarified.

Response: When H_2 is set to a zero matrix, all data vectors in it are zero. In that case, we replace zero data vectors with data vectors computed using a new state estimates. The condition number minimization algorithm takes over once all zero vectors have been replaced. This description has been re-worded in the revised manuscript to clarify the history stack construction process.

(4) Apart from these clarifying issues, there is the problem of the large number of conditions required to make the theoretical results hold. In particular, the conditions required in Theorem 7 and Corollary 10 (Theorem 10 of the revised manuscript) look overwhelming. The authors acknowledge that some of those conditions may be restrictive (e.g. at the end of page 5). Overall, the stated conditions for the proposed method are very technical and make the method look impractical. Could the authors comment, for example, on what could the user do to satisfy the conditions in 3?

Response: The conditions in 3 cannot be enforced a priori due to nonlinear dependence of $\hat{\Sigma}$ on the state estimate $\hat{x}$. However, as shown in Figures 5, 6, and 7, given a history stack, it is relatively straightforward to check whether the conditions are met. 3 in the revised manuscript provides explicit eigenvalue tests for the first and the third condition. The second condition, $\Sigma_u \in \text{Range}\,\hat{\Sigma}$, can be checked using the ranks of $\hat{\Sigma}$ and $[\hat{\Sigma}, \Sigma_u]$. If the rank of the former is equal to the rank of the latter, then $\Sigma_u \in \text{Range}\,\hat{\Sigma}$. The revised manuscript spells this out in Part (3) of Remark 4.

(5) Moreover, Theorem 7 only states that the desired cost function will be obtained if the state estimates are equal to the real states, and if the metric Δ is exactly equal to zero. However, both of those variables only present asymptotic convergence. There is no analysis of the behavior of the approximated parameters in $\hat{W}$ as $\hat{x}$ tends to x, and as Δ tends to zero.

Response: Corollary 10 of the original manuscript (Theorem 10 of the revised manuscript) is where we analyzed exactly what the reviewer is asking for in this comment. To highlight this fact better, we have made Corollary 10 of the original manuscript into a theorem in the revised manuscript.

(6) There is something else I don't understand in Theorem 7. Why is matrix K selected as in (18)? In particular, where does the term $(\hat{\Sigma^\top}\hat{\Sigma} + \epsilon I)^{-1}$ come from? Either I am missing something, or this is never used in the proof. If the intention is simply to say that the right-hand side of (24) is negative semidefinite, then it would be enough to select the lower block entry of (18) as $P\hat{\sigma^\top}$, where P is any positive definite matrix. Then, no matrix inverse is required. The authors should properly justify the design of their method.

Response: The reviewer is correct in saying that instead of $(\hat{\Sigma^\top}\hat{\Sigma} + \epsilon I)^{-1}$, one can select any positive definite symmetric matrix P. We have updated the revised manuscript where we have removed that matrix. Our selection of $(\hat{\Sigma^\top}\hat{\Sigma} + \epsilon I)^{-1}$ was inspired by ridge regression for underdetermined systems, and we say so in the simulation

section of the revised manuscript where we select $(\hat{\Sigma^\top}\hat{\Sigma} + \epsilon I)^{-1}$.

(7) Corollary 10 (Theorem 10 of the revised manuscript) has the same problem as Theorem 7 of requiring a large amount of conditions to hold. At the end, this corollary only shows that the obtained cost function will approximate the real one.

Response: The fact that the obtained cost function will only be an approximation of the real one is a consequence of the excitation only being available on a finite interval. If the excitation is persistent, then asymptotic convergence can be established. Due to the page limits of a brief paper, and since finite excitation is the more common scenario, we analyzed only the finite excitation case.

(8) As I mentioned above, it is unclear how to satisfy the persistence of excitation conditions (19) in 3. This is even more problematic given that the expert system (1) is assumed to use an asymptotically stable linear input that is almost certainly not exciting. Corollary 10 (Theorem 10 of the revised manuscript) even requires that the excitation in the expert's trajectories lasts for a potentially long time. I would like to see a justification about why we should expect that to happen when the trajectories of the expert go to zero in an optimal fashion. In the simulation section, the authors solve this problem by adding an additional exciting signal to the input of the expert system. However, this step was never mentioned in the preceding sections of the manuscript. The problem is that the proofs of the main results in the paper use the fact that the behavior of the expert is optimal, and therefore the proofs do not hold with the additional excitation signal.

Response: The excitation signal is assumed to be known to the learner, and as such, it can be subtracted off to infer the expert's optimal action. As a result, the excitation signal does not affect the analysis. Please note that the state trajectory of the expert does not need to be optimal for the RHSO to work. As long as the learner, at time t, can infer the expert's optimal action in response to a given state $x(t)$, the RHSO can be implemented. In other words, the learner just needs access to sufficiently many state-action pairs of the form $(x, -K_{Ep}x)$, the trajectory $t \mapsto x(t)$ does not need to be optimal (see Remark 4 of the revised manuscript).

(9) As a final important comment, the authors mention in the introduction that there is a conference version of this manuscript. The citation [23], however, seems to refer to this same manuscript, submitted to Automatica. Since I cannot see that conference paper, I cannot judge whether the differences between the content of both versions justifies the publication in Automatica as a regular paper.

Response: We apologize for the mix-up related to the conference version [23]. We intended to cite the conference paper that was uploaded to arXiv, but cited the submitted paper instead. The conference paper is now published and is duly cited in the revised manuscript.

(10) There are typos in the equations throughout the manuscript, particularly when defining column vectors and omitting transposes.

Response: We have made every attempt to ensure that the equations in the revised manuscript are correct.

Response the comments by to Reviewer 2

This paper proposes an algorithm for inverse reinforcement learning, which does not require a uniqueness assumption on the Q, R, and S matrices entailed in the underlying Riccati equation. While this problem is interesting, the presentation needs improvement. In addition, the paper requires quite a few restrictive assumptions, and seems to rely on technically questionable claims. Specific comments below. Regarding the restrictive assumptions:

(1) It is required that $\hat{R}$ is invertible at all times, where $\hat{R}$ is the estimation of the control weighting term in the Riccati equation. But this is simply not possible to verify a priori. Remark 8 claims that a projection operator can be used to enforce this assumption, however, it is known that projection operators negatively interfere with convergence guarantees.

Response: In response to this comment, we have expanded the discussion of the projection operator to show that in the diagonal case, it does not affect stability. Please see Remark 8 of the revised manuscript.

(2) It is required, in (19), that $\Sigma_u \in Range(\hat{\Sigma})$. How can one enforce this assumption? Especially given the reduced degrees of freedom by fixing one of the weights in (14), I cannot see why this assumption should be true

Response: Like most excitation conditions in RL and IRL, this excitation condition cannot be enforced. The best we can do is to monitor whether it is met at each time instant online, which is precisely what is done in Figures 5, 6 and 7. The condition $\Sigma_u \in \text{Range}\,\hat{\Sigma}$, can be checked using the ranks of $\hat{\Sigma}$ and $[\hat{\Sigma}, \Sigma_u]$. If the rank of the former is equal to the rank of the latter, then $\Sigma_u \in \text{Range}\,\hat{\Sigma}$.

Regarding the technical quality:

(1) In the proof of Corollary 10 (Theorem 10 of the revised manuscript), it is claimed that if $\|\tilde{K}_p \hat{x}(t_i(t))\| \leq \frac{\bar{\omega}}{c(t)}$ then $\|\tilde{K}_p\| \leq \bar{\omega}$. I just cannot see why this would be true. For example, suppose that $\hat{x}$ goes to zero while pointwise in time maintaining the full rank status. Then, one might as well have $\tilde{K}_p \to \infty$, and hence $\|\tilde{K}_p\| \to \infty$, while $\|\tilde{K}_p \hat{x}(t_i(t))\| \leq \frac{\bar{\omega}}{c(t)}$ still holds. A similar claim is made later.

We found this omission right after submission of the paper and rectified it, but were unable to update the submission. Please note that in addition to the data matrix being full rank, Theorem 10 in the revised manuscript requires the minimum eigenvalue of the data matrix to be bounded from below. That lower bound on the smallest eigenvalue is what makes the existence of c possible.

(2) The authors focus a lot on the technical issue of $\hat{\Sigma}^\top \hat{\Sigma}$ not being invertible, and thus regularize it by using the matrix $\hat{\Sigma}^\top \hat{\Sigma} + \epsilon I$. One would expect that ϵ would have to be very small here for things to work, yet there is no such condition in the paper. This is because the inverse of $\hat{\Sigma}^\top \hat{\Sigma} + \epsilon I$ is, in fact, not needed at all. To see this, remove the inverse of $\hat{\Sigma}^\top \hat{\Sigma} + \epsilon I$ from (18). Then, (22) will still imply convergence of Δ to 0. So why complicate the presentation and the analysis by using this matrix?

Response: The reviewer is correct in saying that instead of $(\hat{\Sigma^\top}\hat{\Sigma} + \epsilon I)^{-1}$, one can select any positive definite symmetric matrix P. We have updated the revised manuscript where we change that matrix to P. Our selection of $(\hat{\Sigma^\top}\hat{\Sigma} + \epsilon I)^{-1}$ was inspired by ridge regression for underdetermined systems, and we say so in the simulation section of the revised manuscript where we select $P = (\hat{\Sigma^\top}\hat{\Sigma} + \epsilon I)^{-1}$. The discussion on regularization is motivational, and we think that it is still relevant.

(3) Suppose that the data x indeed span R^n, but their excitation decays to 0 exponentially fast (i.e., $x(t) \to 0$ exponentially fast). In this case, it seems that holding on "old data" can be relieving as excitation starts to vanish. But why does the requirement $\lim_{t\to\infty} t_1(t) >> 1$ in Corollary 10 (Theorem 10 of the revised manuscript) suggest otherwise? The authors seem to focus a lot on rank and span properties of matrices pointwise in time, however, in adaptive control these properties are usually insufficient to guarantee convergence.

That requirement would indeed be unnecessary, as the reviewer suggests, if we had access to the full system state, x. Since we only measure the output, y, we can only store estimates of the state, $\hat{x}$. Due to the purging algorithm, the state estimates stored in the history stack get better over time. For a given desired error bound ϖ, we need to wait until the error between $\hat{x}$ and x at points stored in the history stack, becomes sufficiently small for the update law to converge to a $\varpi-$equivalent solution.

Regarding the presentation

(1) The proof of Theorem 7 is a straightforward application of LaSalle's invariance principle. The authors use [13, Theorem 4.4] for the proof of this theorem, but the proof would have been an one-liner if [13, Corollary 4.2] was used instead. The discussion regarding positive limit sets seems redundant.

Response: We cannot use Corollary 4.2 from Khalil to prove Theorem 7 as suggested by the reviewer. In fact, the system in (15) does not have a globally (or even locally) asymptotically stable equilibrium point at the origin. Any perturbations outside of the subspace Ω_Δ are not guaranteed to return to the origin. As such, that corollary is not applicable, and we are forced to resort to the more general invariance principle. However, in responding to this comment, we realized that uniqueness of solutions to ODEs can be used to conclude that Ω_Δ is forward invariant, which makes the invariance principle applicable to this problem, and we do not need to replicate the proof of the invariance principle. The revised manuscript takes this approach, which simplifies the proof of Theorem 7

(2) In (7)-(8), all matrices, such as W_S, σ_Q, etc., should be explicitly defined.

Response: Since we were asked to resubmit as a brief paper, we had to remove detailed descriptions of all basis functions from the revision. We now direct the reader to the arXiv manuscript [18] for the details.

(3) The definition of Σ should appear immediately after Σ is first used, after (15).

Response: It has been defined in the revised manuscript directly after (12), as the non-state estimate version of $\hat{\Sigma}$.

(4) In Remark 5 (Remark 4 of the revised manuscript), it is mentioned "Since the expert is assumed to be optimal, we know that $\Sigma_u \in Range(\Sigma)$. Why is this true?

Note that since the optimal trajectory and the optimal controller satisfy the HJB equation and the control equation, $\Sigma_u = \Sigma W^*$ where W^* is a vector comprised of the expert's W_S^*, W_Q^*, and W_R^*, so we get $\Sigma_u \in Range(\Sigma)$. We have now added this explanation to the revised manuscript.

(5) One of the weights in (14) is fixed arbitrarily beforehand. Why is this without loss of generality? Meanwhile, the authors suggest that they impose this to exclude irregular solutions to the IRL problem. However, the assumption that $\hat{R}$ is invertible seems to imply that arbitrarily fixing this weight is actually insufficient.

Response: Please note that the controller that minimizes the cost $\int_0^\infty \left(x(t)^\top Q x(t) + u(t)^\top R u(t)\right) \mathrm{d}t$ is identical to the optimal controller that minimizes the cost $\int_0^\infty \left(x(t)^\top \alpha Q x(t) + u(t)^\top \alpha R u(t)\right) \mathrm{d}t$ for any $\alpha > 0$, and so are the resulting optimal trajectories. As a result, we can only infer Q and R from measured trajectories up to a scaling factor. Fixing one element of the weights fixes the scaling factor, and hence, is without loss of generality.