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

Jared Town, Zachary Morrison, and Rushikesh Kamalapurkar

***Abstract*— A key challenge in solving the deterministic inverse reinforcement learning problem online and in real-time is the existence of non-unique solutions. Nonuniqueness necessitates the study of the notion of equivalent solutions 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 is developed to generate solutions that are approximately equivalent. Novel data-richness conditions are developed to facilitate the analysis and simulation results are provided to demonstrate the effectiveness of the developed technique.**

I. Introduction

Inverse reinforcement learning (IRL) is the process of recovering a cost function of an optimal "expert" whose trajectories are consistent with a given dynamic model [1]. This "expert" is assumed to be behaving optimally with respect to some unknown cost function. IRL methods [1]–[14] are often utilized in teaching an autonomous system a specific task in an offline environment. While effective, these implementations 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. These issues are addressed in results such as [15]–[20] where an online model-based IRL method using a single iteration over one continuous trajectory is used to learn the cost function of an expert.

This paper derives inspiration from the history stack observer (HSO) for IRL developed in [15] under the implicit assumption that the IRL problem admits a unique solution up to a scaling factor. Since IRL problems generally admit multiple linearly independent solutions [21], [22], the uniqueness assumption is restrictive. Nonuniqueness is studied in results such as [21], where procedures to determine equivalent performance index are developed. In [22], methods to create and identify inverse optimal control problems that admit multiple solutions are detailed.

The methods recently developed in [23] and [24] guarantee convergence to the set of possible solutions. However, the problem is solved in an offline setting as opposed to the online and real-time problem under consideration in this paper. Furthermore, the result in [23] requires knowledge of the demonstrator's control penalty and a diagonal state penalty matrix.

In this paper, the HSO formulation from [15] is extended to address nonuniqueness of solutions. While the modification made to the observer design resembles ridge regression, the resulting convergence guarantees are surprising and require novel analysis tools and data richness conditions. The analysis shows that if the IRL problem has non-unique solutions, then the developed observer finds an equivalent solution.

The contributions are as follows:

- This article extends the IRL HSO in [15] to problems where the observed trajectories are optimal with respect to multiple cost functions. A learner with access to the state space model, controller input, and measurement data reconstructs an equivalent cost function of an expert.
- A novel data informativity condition is derived for convergence of the observer.
- A novel analysis approach that utilizes the invariance principle is used to provide convergence guarantees.

The paper is structured as follows: Section II contains the problem formulation. Section III contains the modified HSO and stability analysis. Section IV contains simulation Section V concludes the paper.

II. Problem Formulation

Following [15], 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

$$y' = Cx, \tag{2}$$

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 pair (A, B) is assumed stabilizable and the pairs (A, C) and $(A, \sqrt{Q})$ are assumed 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.

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

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

The authors are with the School of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK, USA. {jared.town, zachmor, rushikesh.kamalapurkar}@okstate.edu.
This research was supported, in part, by the National Science Foundation (NSF) under award numbers 1925147 and 2027999. 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.

where $x(\cdot)$ is the system trajectory under the control signal $u(\cdot)$ and starting from the initial condition x_0, and $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$ are unknown positive semi-definite matrices. That is, 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 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, the inputs $u(\cdot)$ and the outputs $y(\cdot)$. Generally, there are multiple cost functionals that are compatible with any set of input-output trajectories and system matrices, A, B, and C. 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 1).

While the HSO in [15] is an effective technique to solve the IRL problem online and in real-time, the update laws rely on inversion of a data matrix which can be invertible only if the IRL problem has a unique solution up to a scaling factor. As such, the method in [15] 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.

III. History Stack Observer for Problems with Nonunique Solutions

To facilitate the discussion, this section provides a brief summary of the HSO developed in [15] and highlights the key problem that is resolved in this paper.

A. History Stack Observer (HSO)

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

$$x^{\mathrm{T}}\left(A^{\mathrm{T}}S + SA - SBR^{-1}B^{\mathrm{T}}S + Q\right)x = 0 \tag{4}$$

for all $x \in \mathbb{R}^n$ and the optimal control equation

$$u(t) = u^*(x(t)) \coloneqq -R^{-1}B^{\mathrm{T}}Sx(t) \tag{5}$$

$\forall t \in \mathrm{R}_{\geq 0}$. The expert's feedback matrix is then given by $K_{Ep} = R^{-1}B^{\mathrm{T}}S$. The HJB equation and the optimal control equation facilitate the definition of an equivalent solution.

Definition 1. A solution $(\hat{Q}, \hat{S}, \hat{R})$ is called an equivalent solution of the IRL problem, corresponding to the set of points $\{(x_i, u_i)\}_{i=1}^N \subset \mathbb{R}^n \times \mathbb{R}^m$, if for all $i = 1, \cdots, N$,

$$x_i^{\mathrm{T}}\left(A^{\mathrm{T}}\hat{S} + \hat{S}A - \hat{S}B\hat{R}^{-1}B^{\mathrm{T}}\hat{S} + \hat{Q}\right)x_i = 0$$

and

$$\hat{K}_p \coloneqq \hat{R}^{-1}B^{\mathrm{T}}\hat{S} = K_{EP}.$$

Remark 1. Note that the idea of an equivalent solution, as defined above, is slightly weaker than equivalent solutions defined in results such as [23] and [18]. In results such as [23] and [18], $(\hat{Q}, \hat{S}, \hat{R})$ is called an equivalent solution if $A^{\mathrm{T}}\hat{S} + \hat{S}A - \hat{S}B\hat{R}^{-1}B^{\mathrm{T}}\hat{S} + \hat{Q} = 0$. Here, we only require that $x_i^{\mathrm{T}}\left(A^{\mathrm{T}}\hat{S} + \hat{S}A - \hat{S}B\hat{R}^{-1}B^{\mathrm{T}}\hat{S} + \hat{Q}\right)x_i = 0$ for all points x_i in our dataset, which renders a larger class of solutions *equivalent*. We concede that obtaining equivalent solutions as defined in results such as [23] and [18] is perhaps more useful in applications. However, when the IRL problem is solved in a model-free setting, we postulate that equivalent solutions in the sense of Definition 1 above (rather, a model-free equivalent thereof, derived using the integral, rather than the differential form of the HJB equation) is the best that can be achieved. Proof and/or further examination of this postulate and extension of the method developed in this paper to the model-free setting are a part of ongoing work.

Given an estimate $\hat{x}$ of the state x, a measurement of the control signal, u, and estimates Q, R, and S of $\hat{Q}$, $\hat{R}$, and $\hat{S}$, respectively, (4) and (5) can be evaluated to develop an observation error that evaluates to zero if the state estimates and estimates of the matrices Q, R, and S are correct. In the following, the observation error is used to improve the estimates by framing the IRL problem as a state estimation problem. To facilitate the observer design, equations (4) and (5) are linearly parameterized as

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

$$\begin{aligned}0 &= \nabla_x\left((W_S^*)^{\mathrm{T}}\sigma_S(x)\right)(Ax + Bu^*(x)) \\ &\quad + (W_Q^*)^{\mathrm{T}}\sigma_Q(x) + (W_R^*)^{\mathrm{T}}\sigma_{R1}(u^*(x)),\end{aligned} \tag{7}$$

where $x^{\mathrm{T}}Sx = (W_S^*)^{\mathrm{T}}\sigma_S(x)$, $x^{\mathrm{T}}Qx = (W_Q^*)^{\mathrm{T}}\sigma_Q(x)$, $u^{\mathrm{T}}Ru = (W_R^*)^{\mathrm{T}}\sigma_{R1}(u)$, and $Ru = \sigma_{R2}(u)W_R^*$, and $W_S^* \in \mathbb{R}^{P_S}$, $W_Q^* \in \mathbb{R}^{P_Q}$, $W_R^* \in \mathbb{R}^M$ are the ideal weights with P_S, P_Q, and M being the number of basis functions in the respective linear parameterizations.

Motivated by (6), and using the estimates $\hat{W}_S$, $\hat{W}_Q$, and $\hat{W}_R$ for W_S^*, W_Q^*, and W_R^* respectively, a control residual error is defined as

$$\Delta_u'\left(x, u, \hat{W}'\right) \coloneqq 2\sigma_{R2}(u)\hat{W}_R + B^{\mathrm{T}}\left(\nabla_x\sigma_S(x)\right)^{\mathrm{T}}\hat{W}_S. \tag{8}$$

Similarly, from (7), the inverse Bellman error is defined as

$$\begin{aligned}\delta'\left(x, u, \hat{W}'\right) &\coloneqq \nabla_x\left((\hat{W}_S)^{\mathrm{T}}\sigma_S(x)\right)(Ax + Bu) \\ &\quad + (\hat{W}_Q)^{\mathrm{T}}\sigma_Q(x) + (\hat{W}_R)^{\mathrm{T}}\sigma_{R1}(u).\end{aligned} \tag{9}$$

Separating out $\hat{W}' = \left[\hat{W}_S, \hat{W}_Q, \hat{W}_R\right]^{\mathrm{T}}$ yields

$$\begin{bmatrix}\delta'\left(x, u, \hat{W}'\right) \\ \Delta_u'\left(x, u, \hat{W}'\right)\end{bmatrix} = \begin{bmatrix}\sigma_{\delta'}(x, u) \\ \sigma_{\Delta_u'}(x, u)\end{bmatrix}\begin{bmatrix}\hat{W}_S \\ \hat{W}_Q \\ \hat{W}_R\end{bmatrix}, \tag{10}$$

where

$$\begin{aligned}&\sigma_{\delta'}(x, u) \coloneqq \\ &\left[(Ax + Bu)^{\mathrm{T}}(\nabla_x\sigma_S(x))^{\mathrm{T}} \quad \sigma_Q(x)^{\mathrm{T}} \quad \sigma_{R1}(u)^{\mathrm{T}}\right]\end{aligned} \tag{11}$$

and

$$\begin{aligned}&\sigma_{\Delta_u'}(x, u) \coloneqq \\ &\left[B^{\mathrm{T}}(\nabla_x\sigma_S(x))^{\mathrm{T}} \quad 0_{m\times P_S+P_Q} \quad 2\sigma_{R2}(u)\right].\end{aligned} \tag{12}$$

The scaling ambiguity inherent in linear quadratic optimal control, which is apparent in the fact that $\hat{W}' = 0$ 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 r_1 arbitrarily and removing it from (10) yields scale-aware definitions of the control residual error and the inverse Bellman error given by

$$\begin{bmatrix} \delta\left(x,u,\hat{W}\right) \\ \Delta_u\left(x,u,\hat{W}\right) \end{bmatrix} \coloneqq \begin{bmatrix} \sigma_\delta(x,u) \\ \sigma_{\Delta_u}(x,u) \end{bmatrix} \begin{bmatrix} \hat{W}_S \\ \hat{W}_Q \\ \hat{W}_R^- \end{bmatrix} + \begin{bmatrix} u_1^2 r_1 \\ 2u_1 r_1 \\ 0_{m-1\times 1} \end{bmatrix}, \quad (13)$$

where $\hat{W}_R^-$ denotes $\hat{W}_R$ with the first element removed.

Pairing the innovation $y - C\hat{x}$ with the inverse bellman error and control residual error from (13) yields the observation error [1]

Using the observation error, the history stack observer is designed to be of the form

$$\begin{bmatrix} \dot{\hat{x}} \\ \dot{\hat{W}} \end{bmatrix} = \underbrace{\begin{bmatrix} A\hat{x} + Bu \\ 0_{P_S+P_Q+M-1} \end{bmatrix}}_{\text{prediction}} + K \underbrace{\left(\begin{bmatrix} Cx \\ \Sigma_u \end{bmatrix} - \begin{bmatrix} C\hat{x} \\ \hat{\Sigma}\hat{W} \end{bmatrix} \right)}_{\text{innovation}} \quad (14)$$

where the gain K is selected to be

$$K \coloneqq \begin{bmatrix} K_3 & 0_{n\times N+Nm} \\ 0_{P_S+P_Q+M-1\times L} & K_4(\hat{\Sigma}^{\mathrm{T}}\hat{\Sigma})^{-1}\hat{\Sigma}^{\mathrm{T}} \end{bmatrix}, \quad (15)$$

and where $\hat{W} = [\hat{W}_S, \hat{W}_Q, \hat{W}_R^-]$ and

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

Remark 2. Motivated similarly to [15], the feedback gain K in (15) can be selected as a Luenberger observer or a Kalman gain.

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 (14), to an equivalent solution (see Theorem 1 below), the history stack is recorded using a minimum singular value maximization 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 (14) 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 selectively added, along with the corresponding control input, to H_2. A new state estimate is selected to replace an existing state estimate in H_2 if the replacement decreases the condition number of $(\hat{\Sigma}^{\mathrm{T}}\hat{\Sigma})$. Once the data in H_2 are such that the condition number of $(\hat{\Sigma}^{\mathrm{T}}\hat{\Sigma})$ 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. The purging process ensures that old and possibly erroneous state estimates are removed from H_1.

The IRL method developed in this paper requires that the expert's behavior is optimal, which implies that $u(t) = K_{EP}x(t)$ for all t. Since true values of the state are not accessible, in general, for the data points stored in the history stack H_1, $K_{EP}\hat{x}(t_i) - u(t_i) \neq 0$, which results in inaccuracy in the estimation of an equivalent solution. Since the state estimates converge to the true state exponentially, the purging process described above ensures that the discrepancy $\max_{i=1,\cdots,N} \|K_{EP}\hat{x}(t_i) - u(t_i)\|$ is monotonically decreasing in time, and so is the resulting inaccuracy in the estimation of an equivalent solution.

Generally, given a system model with output (or state) and control trajectories, there are multiple sets of Q, R, and S matrices that all solve the IRL problem [21], [22]. As such, the IRL problem, as posed in [15], is not well-defined. In fact, the stability theorem in [15] relies on the assumption that $\hat{\Sigma}$ is full rank. Due to purging and improved state estimates, Σ being full rank implies $\hat{\Sigma}$ is eventually full rank, and as a result, $\Sigma W = \Sigma_u$ has a unique solution. Since uniqueness does not generally hold [22], the HSO must be modified to address the non-unique case. In this paper, the full rank condition, and subsequently, the uniqueness assumption is relaxed using an update rule motivated by ridge [25] and lasso [26] regression.

B. Regularized History Stack Observer for Non-Unique Solutions

To avoid the uniqueness assumption, and subsequently, to allow for a rank-deficient $\hat{\Sigma}$, the gain matrix of the HSO is modified in this paper to include a regularization term to yield

$$K \coloneqq \begin{bmatrix} K_3 & 0_{n\times N+Nm} \\ 0_{P_S+P_Q+M-1\times L} & K_4(\hat{\Sigma}^{\mathrm{T}}\hat{\Sigma} + \epsilon I)^{-1}\hat{\Sigma}^{\mathrm{T}} \end{bmatrix}, \quad (16)$$

where $\epsilon \geq 0$ is a small constant selected by the user to ensure invertibility of $\hat{\Sigma}^{\mathrm{T}}\hat{\Sigma} + \epsilon I$. Instead of using the condition number of $(\hat{\Sigma}^{\mathrm{T}}\hat{\Sigma})$ to select data points for storage in the history stack, the condition number of $(\hat{\Sigma}^{\mathrm{T}}\hat{\Sigma} + \epsilon I)$ is used. In addition, since $\hat{\Sigma}$ cannot be full rank, we need a different way to detect whether the recorded data are sufficient for estimation of an equivalent solution.

The following theorems establish that under a novel informativity condition on the recorded data, the modification above leads to an equivalent solution when the IRL problem admits multiple solutions, and the correct solution when the IRL problem admits a unique solution up to a scaling factor. While the modification itself is relatively minor, the above somewhat surprising results are the key contributions of this work.

[1] See [15] for further details.

To facilitate the analysis, let $\Delta(t) \coloneqq \Sigma_u - \hat{\Sigma}\hat{W}(t)$. Using the update law in (14), the time-derivative of Δ can be expressed as

$$\dot{\Delta} = -\hat{\Sigma}K_4(\hat{\Sigma}^{\mathrm{T}}\hat{\Sigma} + \epsilon I)^{-1}\hat{\Sigma}^{\mathrm{T}}\Delta \tag{17}$$

The analysis requires a data informativity condition summarized in Definition 2 below.

Definition 2. The signal $(\hat{x}, u)$ is finitely informative (FI) if there exists a time instance $T > 0$ such that for some $\{t_1, t_2, \ldots, t_N\} \subset [0, T]$,

$$\operatorname{span}\left(\hat{x}(t_i)_{i=1}^{N}\right) = \mathbb{R}^n, \quad \text{and} \quad \Sigma_u \in (\operatorname{Null}(\hat{\Sigma}^{\mathrm{T}}))^{\perp}. \tag{18}$$

The above informativity condition results in a useful relationship between the range space of $\hat{\Sigma}$ and the set of all *feasible* Δ.

Lemma 1. *If $\hat{\Sigma}$ and Σ_u satisfy* (17)*, then*

$$\Omega_\Delta \cap \operatorname{Null}(\hat{\Sigma}^{\mathrm{T}}) = \{0\}, \tag{19}$$

where

$$\Omega_\Delta \coloneqq \Big\{\Delta \in \mathbb{R}^{N(m+1)} \mid \Delta = \Sigma_u - \hat{\Sigma}y, \quad \textit{for some } y \in \mathbb{R}^{P_S+P_Q+M-1}\Big\}. \tag{20}$$

Proof. If $\Delta \in \operatorname{Null}(\hat{\Sigma}^{\mathrm{T}})$. then Δ is given by some linear combination of the basis for the null space of $\hat{\Sigma}^{\mathrm{T}}$. Let Σ_{Null} be a matrix whose columns are the basis vectors of the null space of $\hat{\Sigma}^{\mathrm{T}}$. Then, $\Delta \in \operatorname{Null}(\hat{\Sigma}^{\mathrm{T}})$ implies that $\Delta = \Sigma_{\mathrm{Null}} W_{\mathrm{Null}}$ for some vector W_{Null} whose elements are the coefficients in the linear combination of the basis of the null space of $\hat{\Sigma}^{\mathrm{T}}$ that makes up Δ. This Δ has to also be equal to $\Sigma_u - \hat{\Sigma}\hat{W}$ for some $\hat{W}$. So, there exist weights W_{Null} and $\hat{W}$ such that $\Sigma_{\mathrm{Null}} W_{\mathrm{Null}} = \Sigma_u - \hat{\Sigma}\hat{W}$. Rearranging the terms, there exist weights W_{Null} and $\hat{W}$ such that $\begin{bmatrix}\Sigma_{\mathrm{Null}} & \hat{\Sigma}\end{bmatrix}\begin{bmatrix}W_{\mathrm{Null}} \\ \hat{W}\end{bmatrix} = \Sigma_u$. That is, Σ_u can be written as a linear combination of the columns of $\hat{\Sigma}$ and the columns of Σ_{Null}. However, since $\operatorname{Rank}(\hat{\Sigma}) = \operatorname{Null}(\hat{\Sigma}^{\mathrm{T}})^{\perp}$, every linear combination of columns of $\hat{\Sigma}$ is orthogonal to every linear combination of the columns of Σ_{Null}, we know that Σ_u has two orthogonal components, one that is contained in the range space of $\hat{\Sigma}$ and another that is contained in the null space of $\hat{\Sigma}^{\mathrm{T}}$. If our data are such that $\Sigma_u \in \operatorname{Null}(\hat{\Sigma}^{T})^{\perp}$, then the component that is contained in the null space of $\hat{\Sigma}^{\mathrm{T}}$ is zero. That is, $W_{\mathrm{Null}} = 0$, which implies that $\Delta = 0$. □

Remark 3. Note that if the IRL problem has a unique solution up to a scaling factor, then the condition in Definition 2 is trivially met whenever $N \geq P_S + P_Q + M - 1$ and $\hat{\Sigma}$ is full rank.

Theorem 1 below shows that provided the weights $\hat{W}$ are updated using the update law in (14), and the trajectories are finitely informative as per Definition 2, then Δ converges to the origin.

Theorem 1. *If $\Sigma_u \in \operatorname{Null}(\hat{\Sigma}^{\mathrm{T}})^{\perp}$ and $\epsilon \geq 0$ is selected to ensure invertibility of $\hat{\Sigma}^{\mathrm{T}}\hat{\Sigma} + \epsilon I$, then the solutions of* (17) *with the gain K in* (16) *satisfy $\lim_{t\to\infty}\Delta(t) = \{0\}$.*

Proof. Let $D = \mathbb{R}^{N(m+1)}$ and consider the positive definite and radially unbounded candidate Lyapunov function

$$V(\Delta) = \frac{1}{2}\Delta^{\mathrm{T}}\Delta. \tag{21}$$

The orbital derivative of V along the solutions of (17) is given by

$$\dot{V}(\Delta) = -\Delta^{\mathrm{T}}\hat{\Sigma}K_4(\hat{\Sigma}^{\mathrm{T}}\hat{\Sigma} + \epsilon I)^{-1}\hat{\Sigma}^{\mathrm{T}}\Delta. \tag{22}$$

For any $c > 0$, the sublevel set $\Omega_c \coloneqq \{\Delta \in D | V(\Delta) \leq c\}$ is compact and positively invariant and the set in (20) can be shown to be closed and positively invariant. As such, the intersection $\Omega = \Omega_c \cap \Omega_\Delta$ is compact and positively invariant. By the invariance principle [27, Th 4.4], all trajectories of Δ in (17) 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 $\operatorname{Null}(\hat{\Sigma}^{\mathrm{T}}) \cap \Omega$ as $\hat{\Sigma}^{\mathrm{T}}\Delta = 0$ only when $\Delta \in \operatorname{Null}(\hat{\Sigma}^{\mathrm{T}})$. Furthermore, from Lemma 1, provided $\Sigma_u \in (\operatorname{Null}(\hat{\Sigma}^{\mathrm{T}}))^{\perp}$, the only Δ that can be in $\hat{\Sigma}^{\mathrm{T}} \cap \Omega_\Delta$ is $\Delta = 0$. Since the singleton $\{0\}$ is positively invariant with respect to the dynamics in (17), it is also the largest invariant subset of $\{\Delta \in \Omega | \dot{V}(\Delta) = 0\}$. As a result, by the invariance principle, all trajectories starting 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 trajectories starting in Ω_Δ converge to the origin. □

The theorem above establishes the convergence of Δ to the origin for a given *fixed* $\hat{\Sigma}$ and Σ_u. The following lemma shows that if the state estimates in $\hat{\Sigma}$ are exact, then $\Delta = 0$ generates an equivalent solution.

Lemma 2. *If full state information is available, i.e., $\hat{x} = x$ and as a result, $\hat{\Sigma} = \Sigma$, if $\Delta = \Sigma_u - \Sigma\hat{W} = 0$, and if $\{x_i\}_{i=1}^{N}$ spans $\mathbb{R}^n$, then the matrices $\hat{Q}$, $\hat{S}$, and $\hat{R}$, extracted from $\hat{W}$, constitute an equivalent solution of the IRL problem corresponding to to the history stack H_1.*

Proof. The fact that if $\Delta = 0$ then $x_i^{\mathrm{T}}\left(A^{\mathrm{T}}\hat{S} + \hat{S}A - \hat{S}B\hat{R}^{-1}B^{\mathrm{T}}\hat{S} + \hat{Q}\right)x_i = 0$ holds for all points in H_1 is immediate from the construction of Δ. To prove equivalence, the equality $\hat{R}^{-1}B^{\mathrm{T}}\hat{S} = K_{EP}$ must be established. Indeed, if $\{x_i\}_{i=1}^{N}$ spans $\mathbb{R}^n$ there is a unique matrix K that satisfies $u_i = Kx_i$ for all $i = 1, \ldots, N$. Now letting $\mathbb{U} = [u_1, \ldots, u_N]$ and $\mathbb{X} = [x_1, \ldots, x_N]$, this unique matrix $K = \mathbb{U}\mathbb{X}^{\mathrm{T}}(\mathbb{X}\mathbb{X}^{\mathrm{T}})^{-1}$. It is also known that the observed data satisfies $u_i = K_{EP}x_i$ for all $i = 1, \ldots, N$, because the expert is optimal. Since $\Delta = 0$, the observed data satisfies $u_i = \hat{R}^{-1}B^{\mathrm{T}}\hat{S}x_i$ for all $i = 1, \ldots, N$. Since there is only one matrix K that satisfies $u_i = Kx_i$ for all $i = 1, \ldots, N$, all three of the matrices above must be equal, i.e., $K = K_{EP} = \hat{R}^{-1}B^{\mathrm{T}}\hat{S}$. Therefore, ($\hat{S}$, $\hat{R}$, $\hat{Q}$) constitutes an equivalent solution of the IRL problem corresponding to the history stack H_1. □

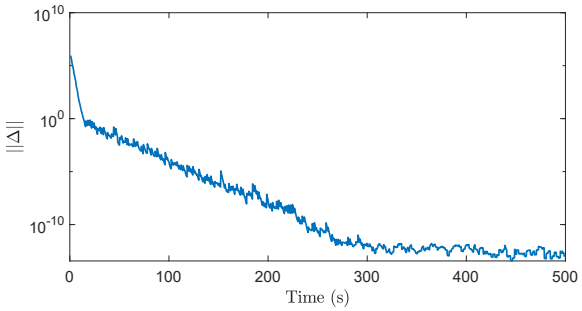

Fig. 1. This plot contains the norm of the error Δ as it is measured throughout the simulation.

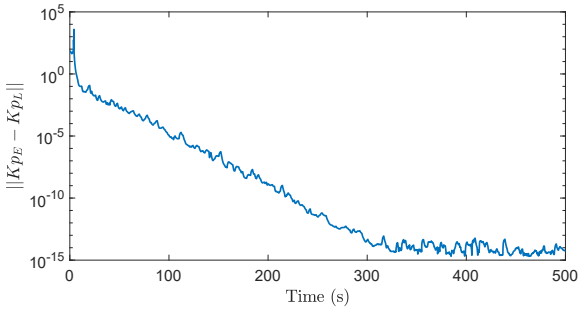

Fig. 2. The feedback matrix of the learner is constructed using the estimated weight trajectories and compared to the expert by the Frobenius norm of the difference.

Theorem 1 and Lemma 2 can be used to obtain the final result summarized into the corollary below.

Corollary 1. *Given $\varpi > 0$, provided t_1 in Definition 2 is sufficiently large, and the history stack is recorded using the purging algorithm described in [15], then* $\left\|\hat{R}^{-1}B^{\mathrm{T}}\hat{S} - K_{Ep}\right\| \leq \varpi$. [2]

Proof. Using similar arguments as the proof of Lemma 2, if $\{\hat{x}(t_i)\}_{i=1}^{N}$ spans $\mathbb{R}^n$ then $\hat{R}^{-1}B^{\mathrm{T}}\hat{S} = \mathbb{U}\hat{\mathbb{X}}^{\mathrm{T}}(\hat{\mathbb{X}}\hat{\mathbb{X}}^{\mathrm{T}})^{-1}$, where $\hat{\mathbb{X}} = [\hat{x}(t_1), \ldots, \hat{x}(t_N)]$ and if $\{x(t_i)\}_{i=1}^{N}$ spans $\mathbb{R}^n$ then $K_{EP} = \mathbb{U}\mathbb{X}^{\mathrm{T}}(\mathbb{X}\mathbb{X}^{\mathrm{T}})^{-1}$, where $\mathbb{X} = [x(t_1), \ldots, x(t_N)]$. The purging process, along with exponential convergence of $\hat{x}$ to x, ensure that the error between $\mathbb{X}$ and $\hat{\mathbb{X}}$ decreases with increasing t_1. As a result, the corollary is established. □

IV. Simulations

In this section, the efficacy of the developed method is demonstrated using an academic example where the IRL problem is known to admit multiple solutions. Each simulation shows the convergence of Δ to zero and of the feedback matrix K_p to the expert's feedback matrix.

A. Academic example

In this section, we construct an academic example that ensures nonuniqueness of IRL solutions using the procedure developed in [22]. The system dynamics are described by

$$\begin{aligned} \dot{x} &= \mathrm{diag}([3,5,7])x + \mathrm{diag}([11,13,17])u \\ y &= \mathrm{diag}([1,1,1])x \end{aligned} \tag{23}$$

where the expert implements a feedback policy that minimizes the cost function in (3) with

$$\begin{aligned} Q &= \mathrm{diag}(1,4,3), \text{ and} \\ R &= \mathrm{diag}(1,1.75,4). \end{aligned} \tag{24}$$

In guaranteeing the invertibility of $\hat{\Sigma}$, $\epsilon = 0.001$. To ensure that the history stack satisfies the condition in (18), we generate an excitation signal comprised of a sum of 30 sinusoids with unit magnitude and randomly selected frequencies and phases ranging from $0.001Hz$ to $0.1Hz$ and $0\mathrm{rad}$ to $\pi\mathrm{rad}$, respectively. This excitation signal is added into the learner system's input (14) and into the expert system's input (1). Data are added to the history stack every 0.08 seconds and is purged when full if the condition number of $\hat{\Sigma}^{\mathrm{T}}\hat{\Sigma} + \epsilon I < 1 \times 10^8$, or 2 seconds since the last purged is reached. A Luenberger observer is utilized for state estimation by selecting the gain K_3 to place the poles of $(A - K_3C)$ at $p_1 = -0.1$, $p_2 = -1.5$ and $p_3 = -2$ using the MATLAB "place" command. As predicted by Theorem 1, Fig. 1 demonstrates Δ convergence to zero and as indicated by Lemma 2, the feedback matrix corresponding to the estimated weights, $\hat{W}$, converges to a neighborhood of the feedback matrix of the expert, as demonstrated in Fig. 2. Finally, Fig. 3 indicates that the cost functional converges to a functional that is different from that of the expert, confirming the existence of multiple equivalent solutions.

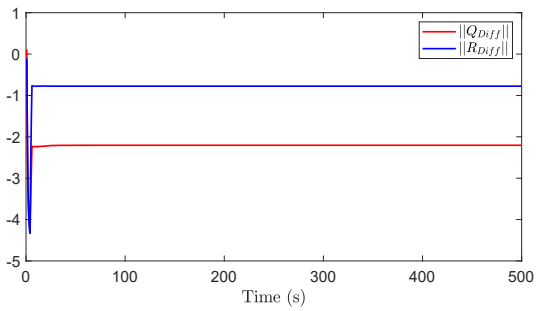

Fig. 3. The trajectories of the Q and R weight estimates are compared to the expert's Q and R values by taking the norm of the difference.

B. Discussion

The simulation requires some tuning effort, where the tuning gains and excitation signals are chosen based on best simulation results. The two primary routes of tuning involve increasing ϵ in small increments, which can be inversely proportional to the time between adding data to the history stack. Increasing time between data allows for more accurate convergence but increases simulation time. Simulation time for the academic example is less than 5 seconds.

[2] $\|\cdot\|$ defines the euclidean norm for a vector when applied to a vector and the Frobenius norm when applied to a matrix.

V. Conclusion

In this paper, a a novel IRL framework is developed for estimation of a cost function, in IRL problems with multiple solutions, through a modification of the HSO introduced in [15]. This modification, while simple, requires an exhaustive and rigorous proof to demonstrate convergence to an equivalent solution when multiple solutions are present. As mentioned previously, most offline IRL methods have disadvantages of being computationally complex and requiring large amounts of data. These issues are resolved though the HSO formulation which is designed to be implemented in an online setting. Future work will involve a deeper theoretical analysis and experimental validation of the developed methods in real-world problems such as learning a quadcopter pilot's cost function through state and input measurements.

References

[1] A. Y. Ng and S. Russell, "Algorithms for inverse reinforcement learning," in *Proc. Int. Conf. Mach. Learn.* Morgan Kaufmann, 2000, pp. 663–670.

[2] S. Russell, "Learning agents for uncertain environments (extended abstract)," in *Proc. Conf. Comput. Learn. Theory*, 1998.

[3] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2004.

[4] P. Abbeel and Y. Ng, Andrew, "Exploration and apprenticeship learning in reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2005.

[5] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich, "Maximum margin planning," in *Proc. Int. Conf. Mach. Learn.*, 2006.

[6] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *Proc. AAAI Conf. Artif. Intel.*, 2008, pp. 1433–1438.

[7] Z. Zhou, M. Bloem, and N. Bambos, "Infinite time horizon maximum causal entropy inverse reinforcement learning," *IEEE Trans. Autom. Control*, vol. 63, no. 9, pp. 2787–2802, 2018.

[8] S. Levine, Z. Popovic, and V. Koltun, "Feature construction for inverse reinforcement learning," in *Adv. Neural Inf. Process. Syst.*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds., vol. 23. Curran Associates, Inc., 2010, pp. 1342–1350.

[9] G. Neu and C. Szepesvari, "Apprenticeship learning using inverse reinforcement learning and gradient methods," in *Proc. Anu. Conf. Uncertain. Artif. Intell.* Corvallis, Oregon: AUAI Press, 2007, pp. 295–302.

[10] U. Syed and R. E. Schapire, "A game-theoretic approach to apprenticeship learning," in *Adv. Neural Inf. Process. Syst.*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 1449–1456.

[11] S. Levine, Z. Popovic, and V. Koltun, "Nonlinear inverse reinforcement learning with Gaussian processes," in *Adv. Neural Inf. Process. Syst.*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 19–27.

[12] K. Mombaur, A. Truong, and J.-P. Laumond, "From human to humanoid locomotion—an inverse optimal control approach," *Auton. Robot.*, vol. 28, no. 3, pp. 369–383, 2010.

[13] B. Lian, V. S. Donge, F. L. Lewis, T. Chai, and A. Davoudi, "Data-driven inverse reinforcement learning control for linear multiplayer games," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[14] R. V. Self, M. Abudia, S. M. N. Mahmud, and R. Kamalapurkar, "Model-based inverse reinforcement learning for deterministic systems," *Automatica*, vol. 140, no. 110242, pp. 1–13, Jun. 2022.

[15] R. V. Self, K. Coleman, H. Bai, and R. Kamalapurkar, "Online observer-based inverse reinforcement learning," *IEEE Control Syst. Lett.*, vol. 5, no. 6, pp. 1922–1927, Dec. 2021.

[16] R. V. Self, "On model-based online inverse reinforcement learning," Ph.D. dissertation, Oklahoma State University, 2020.

[17] N. Rhinehart and K. Kitani, "First-person activity forecasting from video with online inverse reinforcement learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 304–317, 2018.

[18] B. Lian, W. Xue, F. L. Lewis, and T. Chai, "Online inverse reinforcement learning for nonlinear systems with adversarial attacks," *Int. J. Robust Nonlinear Control*, vol. 31, no. 14, pp. 6646–6667, 2021.

[19] M. Herman, V. Fischer, T. Gindele, and W. Burgard, "Inverse reinforcement learning of behavioral models for online-adapting navigation strategies," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 3215–3222.

[20] S. Arora, P. Doshi, and B. Banerjee, "Online inverse reinforcement learning under occlusion," in *Proc. Conf. Auton. Agents MultiAgent Syst.* International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1170–1178.

[21] A. Jameson and E. Kreindler, "Inverse problem of linear optimal control," *SIAM J. Control*, vol. 11, no. 1, pp. 1–19, 1973.

[22] F. Jean and S. Maslovskaya, "Inverse optimal control problem: the linear-quadratic case," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 888–893.

[23] W. Xue, P. Kolaric, J. Fan, B. Lian, T. Chai, and F. L. Lewis, "Inverse reinforcement learning in tracking control based on inverse optimal control," *IEEE Trans. Cybern.*, 2021.

[24] B. Lian, W. Xue, F. L. Lewis, and T. Chai, "Inverse reinforcement learning for adversarial apprentice games," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2021.

[25] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[26] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, pp. 267–288, 1996.

[27] H. K. Khalil, *Nonlinear systems*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2002.