Online Observer-Based Inverse Reinforcement Learning

Ryan Self, Kevin Coleman, He Bai, and Rushikesh Kamalapurkar

***Abstract*— In this paper, a novel approach to the output-feedback inverse reinforcement learning (IRL) problem is developed by casting the IRL problem, for linear systems with quadratic cost functions, as a state estimation problem. Two observer-based techniques for IRL are developed, including a novel observer method that re-uses previous state estimates via history stacks. Theoretical guarantees for convergence and robustness are established under appropriate excitation conditions. Simulations demonstrate the performance of the developed observers and filters under noisy and noise-free measurements.**

I. Introduction

Inverse Reinforcement Learning (IRL) [1]–[3], sometimes referred to as Inverse Optimal Control [4], is a subfield of Learning from Demonstration (LfD) [5] where the goal is to uncover a reward (or cost) function that explains the observed behavior (i.e., input and output trajectories) of an agent. Early results on IRL assumed that the trajectory of the agent under observation is truly optimal with respect to the unknown reward function [2]. Since optimality is in general a strong assumption in a variety of situations, e.g., human operators and trajectories affected by noise or disturbances, IRL is extended to the case of suboptimal demonstrations (i.e., the case where observed behavior does not necessarily reflect the underlying reward function) [6]. While IRL has been an active area of research over the past few decades [7]–[15], most IRL techniques are offline and require a large amount of data in order to uncover the true reward function.

Inspired by recent results in online Reinforcement Learning methods [16]–[18], IRL has been extended to online implementations where the objective is to learn from a single demonstration or trajectory [19]–[22]. In [20], [21], batch IRL techniques are developed to estimate reward functions in the presence of unmeasureable system states and/or uncertain dynamics for both linear and nonlinear systems. The case where the trajectories being monitored are suboptimal due to an external disturbance is addressed in [23], and [22] estimates a feedback policy and generates artificial data using the estimated policy to compensate for the sparsity of data in online implementations. However, results such as [19]–[23], either require full state feedback, or rely on state estimators that require dynamical systems in Brunovsky Canonical form. In addition, none of the aforementioned online IRL methods address uncertainty in the state and control measurements.

This paper builds on the authors' previous work in [22], [23], where concurrent learning (CL) update laws are utilized to estimate reward functions online using output feedback. However, the dynamical systems in [22], [23] are required to be in Brunovsky canonical form, and as such, only the output feedback case where the state is comprised of the output and its derivatives is addressed. In contrast, the IRL observer (IRL-O) technique in this paper generalizes to any observable linear system, since the developed IRL-Os are in a standard observer form where the state estimates are modified based on the innovation (i.e., the error between the actual and the estimated output). As a result, in the case of noisy measurements, they can be implemented as Kalman filters by using the Kalman gain, instead of the developed Lyapunov-based gain design, to select the observer gain. While stability of the filters in the case where the measurements are noisy is not studied in this paper, simulation results demonstrate that the IRL-Os utilizing both the Lyapunov-based gains and the Kalman filter gain are robust to measurement noise.

This paper details two IRL-O formulations. The first method, called the IRL memoryless observer (MLO), is similar to a standard Luenberger observer with a modified observer gain, and guarantees parameter convergence under a persistence of excitation (PE) condition. The second observer implements a novel idea of re-using previous system state estimates and control measurements, along with the Hamilton-Jacobi-Bellman equation, to gain insights into the quality of the current estimate of the reward function. The key advantage of the IRL history stack observer (HSO) over MLO is that it provides an additional guarantee for boundedness of the estimation errors under *finite* (as opposed to *persistent*) excitation [24].

II. Problem formulation

Consider an agent under observation with the following linear dynamics

$$\dot{x} = Ax + Bu, \quad y' = Cx, \tag{1}$$

where $x : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ is the state, $u : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$ is the control, $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are constant system matrices, $y' \in \mathbb{R}^L$ are the outputs, and $C \in \mathbb{R}^{L\times n}$ denotes the output matrix[1].

The authors are with the School of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK, USA. {rself, kevin.coleman10, he.bai, rushikesh.kamalapurkar}@okstate.edu. This research was supported, in part, by the National Science Foundation (NSF) under award number 1925147. 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 agency.

[1]For $a \in \mathbb{R}$, the notation $\mathbb{R}_{\geq a}$ denotes the interval $[a, \infty)$ and the notation $\mathbb{R}_{>a}$ denotes the interval (a, ∞).

The agent under observation is using the policy which minimizes the following performance index

$$J(x_0, u(\cdot)) = \int_0^\infty \left(x(t)^T Q x(t) + u(t)^T R u(t)\right) \mathrm{d}t, \quad (2)$$

where $x(\cdot; x_0, u(\cdot))$ is the trajectory of the agent generated by the optimal control signal $u(\cdot)$ starting from the initial condition x_0. The objective of this paper is to estimate the unknown matrices Q and R by utilizing input-output pairs.

Remark 1. Since Q and R can be selected to be symmetric without loss of generality, the developed IRL method only estimates the elements of Q and R that are on and above the main diagonal.

III. Inverse Reinforcement Learning

Under the premise that the observed agent makes optimal decisions, the state and control trajectories, $x(\cdot)$ and $u(\cdot)$, satisfy the Hamilton-Jacobi-Bellman (HJB) equation [25]

$$H\left(x\left(t\right), \nabla_x \left(V^*\left(x\left(t\right)\right)\right)^T, u\left(t\right)\right) = 0, \forall t \in \mathbb{R}_{\geq 0}, \quad (3)$$

and the optimal control equation $u(x(t)) = -\frac{1}{2} R^{-1} B^T \nabla_x \left(V^*\left(x\left(t\right)\right)\right)^T$, where $V^* : \mathbb{R}^n \to \mathbb{R}$ is the unknown optimal value function and $H : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is the Hamiltonian, defined as $H(x, p, u) := p^T\left(Ax + Bu\right) + x^T Q x + u^T R u$. Given a solution S of the Algebraic Riccati Equation, the optimal value function can be calculated as $V^*(x) = x^T S x$.

To aid in the estimation of the reward function, note that V^*, $x^T Q x$, and $u^T R u$ can be linearly parameterized as $V^*(x) = \left(W_V^*\right)^T \sigma_V(x)$, $x^T Q x = \left(W_Q^*\right)^T \sigma_Q(x)$, and $u^T R u = \left(W_R^*\right)^T \sigma_{R1}(u)$, respectively, where $\sigma_V(x) : \mathbb{R}^n \to \mathbb{R}^P$, $\sigma_Q(x) : \mathbb{R}^n \to \mathbb{R}^P$, and $\sigma_{R1}(u) : \mathbb{R}^m \to \mathbb{R}^M$, are the basis functions, selected as

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

and $W_V^* \in \mathbb{R}^P$, $W_Q^* \in \mathbb{R}^P$, and $W_R^* \in \mathbb{R}^M$, are the ideal weights, given by

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

where, for a given matrix $E \in \mathbb{R}^{n\times n}$, E_{ij} denotes the corresponding element in the ith row and the j-th column of the matrix E, and $E_i^{(-j)}$ denotes the i-th row of the matrix E with the first j elements removed, i.e., $E_3^{(-3)} := \left[E_{34}, E_{35}, \ldots, E_{3(n-1)}, E_{3n}\right]$.

Using $\hat{W}_V$, $\hat{W}_Q$, and $\hat{W}_R$, which are the estimates of W_V^*, W_Q^*, and W_R^*, respectively, in (3), the inverse Bellman error (IBE) $\delta' : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{2P+M} \to \mathbb{R}$ is obtained as $\delta'\left(x, u, \hat{W}'\right) = \hat{W}_V^T \nabla_x \sigma_V(x)\left(Ax + Bu\right) + \hat{W}_Q^T \sigma_Q(x) + \hat{W}_R^T \sigma_{R1}(u)$, where $\hat{W}' := \begin{bmatrix}\hat{W}_V^T & \hat{W}_Q^T & \hat{W}_R^T\end{bmatrix}^T$.

Utilizing $2Ru = -B^T \nabla_x \left(V^*\left(x\right)\right)^T$, Ru can be linearly parameterized as $Ru = \sigma_{R2}(u) W_R^*$, where W_R^* is as previously defined in the IBE and $\sigma_{R2}(u) : \mathbb{R}^m \to \mathbb{R}^{m\times M}$, where the features $\sigma_{R2}(u)$ can be explicitly calculated as

$$\sigma_{R2}(u) = \begin{bmatrix} u^T & 0_{1\times m-1} & \ldots & 0 \\ 0_{1\times m} & \left(u^{(-1)}\right)^T & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0_{1\times m} & 0_{1\times m-1} & \ldots & \left(u^{-(m-1)}\right)^T \end{bmatrix}, \quad (4)$$

where for a given vector $u \in \mathbb{R}^{1\times m}$, $u^{(-j)}$ denotes the vector u with the first j elements removed. Using $\hat{W}_R$ and $\hat{W}_V$ in the optimal controller equation for W_R^* and W_V^*, respectively, after rearranging, a control residual error $\Delta_u' : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{2P+M} \to \mathbb{R}^m$ is obtained as $\Delta_u'(x, u, \hat{W}') = B^T\left(\nabla_x \sigma_V(x)\right)^T \hat{W}_V + 2\sigma_{R2}(u)\hat{W}_R$.

Augmenting the control residual error and the inverse Bellman error yields the error equation

$$\begin{bmatrix}\delta'\left(x, u, \hat{W}'\right) \\ \Delta_u'\left(x, u, \hat{W}'\right)\end{bmatrix} = \begin{bmatrix}\sigma_{\delta'}\left(x, u\right) \\ \sigma_{\Delta_u'}\left(x, u\right)\end{bmatrix} \begin{bmatrix}\hat{W}_V \\ \hat{W}_Q \\ \hat{W}_R\end{bmatrix}, \quad (5)$$

where $\sigma_{\delta'}\left(x, u\right) = \Big[\left(Ax + Bu\right)^T \left(\nabla_x \sigma_V(x)\right)^T, \sigma_Q(x)^T, \sigma_{R1}(u)^T\Big]$, $\sigma_{\Delta_u'}\left(x, u\right) = \Big[B^T(\nabla_x \sigma_V(x))^T, 0_{m\times n}, 2\sigma_{R2}(u)\Big]$.

The IRL problem is then formulated as the need to estimate $\hat{W}_V, \hat{W}_Q$, and $\hat{W}_R$ by minimizing δ' and Δ_u'. However, the IRL problem, as formulated above, is ill-posed, because the minimization problem $\min_{\hat{W}'} |\delta'| + \|\Delta_u'\|$ admits an infinite number of solutions, including the trivial solution $\hat{W}_V = \hat{W}_Q = \hat{W}_R = 0$ and the scaled solutions $\hat{W}_V = \alpha W_V^*, \hat{W}_Q = \alpha W_Q^*$, and $\hat{W}_R = \alpha W_R^*$ $\forall \alpha \in \mathbb{R}_{>0}$. To address the scaling ambiguity and to remove the trivial solution, a single reward weight will be assumed to be known. Since the optimal solution corresponding to a cost function is invariant with respect to arbitrary scaling of the cost function, establishing the scale by assuming that one of the weights as known is without loss of generality. Selecting r_1 as the known weight and removing it from (5) yields

$$\begin{bmatrix}\delta\left(x, u, \hat{W}\right) \\ \Delta_u\left(x, u, \hat{W}\right)\end{bmatrix} = \begin{bmatrix}\sigma_\delta\left(x, u\right) \\ \sigma_{\Delta_u}\left(x, u\right)\end{bmatrix} \begin{bmatrix}\hat{W}_V \\ \hat{W}_Q \\ \hat{W}_R^-\end{bmatrix} + \begin{bmatrix}u_1^2 r_1 \\ 2u_1 r_1 \\ 0_{m-1\times 1}\end{bmatrix}, \quad (6)$$

where $\hat{W}_R^-$ denotes $\hat{W}_R$ with the first element removed, $\hat{W} := \begin{bmatrix}\hat{W}_V^T & \hat{W}_Q^T & \left(\hat{W}_R^-\right)^T\end{bmatrix}^T$, $\sigma_\delta\left(x, u\right) = \Big[\left(Ax + Bu\right)^T \left(\nabla_x \sigma_V(x)\right)^T, \sigma_Q(x)^T, \left(\sigma_{R1}^-(u)\right)^T\Big]$, and $\sigma_{\Delta_u}\left(x, u\right) = \begin{bmatrix}B^T\left(\nabla_x \sigma_V(x)\right)^T & 0_{m\times n} & 2\sigma_{R2}^-(u)\end{bmatrix}$, where $\left(\sigma_{R1}^-(u)\right)^T$ and $\sigma_{R2}^-(u)$ denote $\sigma_{R1}^T(u)$ and $\sigma_{R2}(u)$ with the first columns removed.

We can formulate the IRL problem as a state estimation problem by utilizing the IBE and the controller equation in an

observer framework. Such a formulation allows us to address general output feedback linear systems and to leverage the use of Kalman gains under noisy conditions.

To cast the IRL problem in a state estimation form, the ideal weights are concatenated with the system state to yield the concatenated state vector $z = \begin{bmatrix} x^T & (W^*)^T \end{bmatrix}^T$, where $W^* := \begin{bmatrix} (W_V^*)^T & (W_Q^*)^T & \left((W_R^*)^-\right)^T \end{bmatrix}^T$. Since the ideal weights are constant, the dynamics of the concatenated state is expressed as $\dot{z} = \begin{bmatrix} Ax + Bu \\ 0_{2P+M-1\times 1} \end{bmatrix}$, and $y = h(z)$, where y denotes the measurement vector and $h(z)$ is the corresponding measurement model to be designed in the following.

IV. A Memoryless Observer

The key idea behind MLO is to treat the measurements, y', and the measured/known quantities in (6) as the *output*, $y \in \mathbb{R}^{L+1+m}$, used for estimation of the concatenated state. The output is thus given by $y = \begin{bmatrix} (y')^T & -u_1^2 r_1 & -2u_1 r_1 & 0_{1\times m-1} \end{bmatrix}^T$. The corresponding measurement model is developed by using (6) to express the output as a function of the concatenated state as $h(z) = \begin{bmatrix} Cx \\ \begin{bmatrix} \sigma_\delta(x,u) \\ \sigma_{\Delta_u}(x,u) \end{bmatrix} \begin{bmatrix} W_V^* \\ W_Q^* \\ (W_R^*)^- \end{bmatrix} \end{bmatrix}$. Let $g(\hat{x},u) := \begin{bmatrix} \sigma_\delta(\hat{x},u) \\ \sigma_{\Delta_u}(\hat{x},u) \end{bmatrix}$ and $\sigma_u(u_1) := \begin{bmatrix} -u_1^2 r_1 \\ -2u_1 r_1 \\ 0_{m-1\times 1} \end{bmatrix}$. The observer can then be designed as

$$\begin{bmatrix} \dot{\hat{x}} \\ \dot{\hat{W}} \end{bmatrix} = \begin{bmatrix} A\hat{x} + Bu \\ 0_{2P+M-1\times 1} \end{bmatrix} + K\left(\begin{bmatrix} Cx \\ \sigma_u(u_1) \end{bmatrix} - \begin{bmatrix} C\hat{x} \\ g(\hat{x},u)\hat{W} \end{bmatrix}\right), \tag{7}$$

where $K \in \mathbb{R}^{n+2P+M-1\times L+m+1}$ is the observer gain matrix, designed in the following section.

A. Observer Gain Design and Stability Analysis

In the following analysis, the gain matrix K will be designed in a block diagonal form. In particular, we choose $K_{MLO} := \begin{bmatrix} K_1 & 0_{n\times 1+m} \\ 0_{2P+M-1\times L} & \gamma g(\hat{x},u)^T K_2 \end{bmatrix}$ and $\gamma := {}^1/_{(\nu\|g(\hat{x},u)^T g(\hat{x},u)\|+1)}$ where $\nu \in \mathbb{R}_{\geq 0}$ is a tunable constant.

The following theorem analyzes the stability properties of the resulting MLO using persistence of excitation.

Definition 1. A signal $t \mapsto A(t)$ is called persistently excited, if for all $t \geq 0$ there exists $\alpha, \delta \in \mathbb{R}_{>0}$ such that[2] $\int_{t_0}^{t_0+\delta} A(\tau)^T A(\tau)\, d\tau \geq \alpha I$.

Theorem 1. Provided the gain K_1 is selected such that $(A - K_1 C)$ is Hurwitz, the gain K_2 is selected to be a symmetric positive definite matrix, and $g(\hat{x},u)$ is PE[3], then $\lim_{t\to\infty} \tilde{W}(t) = 0$.

[2]The notation I denotes an identity matrix.

[3]Though the matrix $g(\hat{x},u)^T g(\hat{x},u)$ is singular for all t,

Proof. The dynamics for the system state estimation errors can be described by $\dot{\tilde{x}} = Ax + Bu - A\hat{x} - Bu - K_1 C\tilde{x} = \dot{\tilde{x}} = (A - K_1 C)\tilde{x}$. If $A - K_1 C$ is Hurwitz, then $\tilde{x}$ converges exponentially to the origin.

The dynamics of the weight estimation error can be expressed as $\dot{\tilde{W}} = -\gamma g(\hat{x},u)^T K_2 \sigma_u(u_1) + \gamma g(\hat{x},u)^T K_2 g(\hat{x},u)\hat{W}$. Adding $\pm\gamma g(\hat{x},u)^T K_2 g(\hat{x},u) W^*$ and using the fact that $\sigma_u(u_1) = g(x,u)W^*$, the weight estimation error dynamics can be expressed as a perturbed linear time-varying system

$$\dot{\tilde{W}} = -A(t)\tilde{W} + B(t), \tag{8}$$

where $A(t) := \gamma(t) g(\hat{x}(t),u(t))^T K_2 g(\hat{x}(t),u(t))$ and $B(t) := \gamma(t) g(\hat{x}(t),u(t))^T K_2 (g(\hat{x}(t),u(t)) - g(x(t),u(t)))W^*$. Since $\hat{x}, x, u \in \mathcal{L}_\infty$, Theorem 2.5.1 from [26] implies that the nominal system $\dot{\tilde{W}} = -A(t)\tilde{W}$ is globally exponentially stable (GES) if K_2 is a symmetric positive definite matrix and the signal $(\hat{x},u)$ is PE.

Lemma 4.6 from [27], which states a continuously differentiable and gloablly Lipchitz function is input-to-state stable if the unforced system is GES, can then be invoked with $B(t)$ as the input and $\tilde{W}$ as the state to conclude that (8) is input-to-state stable (ISS). Furthermore, as $t \to \infty$, $\tilde{x}(t) \to 0$, and as a result, $B(t) \to 0$. Exercise 4.58 in [27] can then be invoked to conclude that $\lim_{t\to\infty} \tilde{W}(t) = 0$. □

Remark 2. The condition that $A - K_1 C$ must be Hurwitz can be trivially satisfied if (A, C) is observable.

V. Inclusion of Memory

The observer designed in the previous section relies on *persistent* excitation for stability and convergence. As a result, it suffers from the well-known lack of robustness of PE-based adaptive control methods under loss of excitation. This section develops an observer (called the HSO) that relies on re-use of previously recorded data (henceforth referred to as the history stack) for robustness. If the system trajectories are PE, then the HSO results in convergence of the estimation errors to the origin, similar to the MLO. However, as opposed to the MLO, through the use of a history stack, the HSO guarantees boundedness of the state estimation errors even under loss of excitation.

The output for the HSO is $y(t) = \begin{bmatrix} (y'(t))^T, -u_1^2(t_1)r_1, -2u_1(t_1)r_1, 0_{1\times m-1}, \ldots, -u_1^2(t_N)r_1, -2u_1(t_N)r_1, 0_{1\times m-1} \end{bmatrix}^T$, with the corresponding measurement model, obtained by using past control values and past state estimates in (6), given by

$$h(z) = \begin{bmatrix} Cx \\ \begin{bmatrix} \sigma_\delta(x(t_1),u(t_1)) \\ \sigma_{\Delta_u}(x(t_1),u(t_1)) \\ \vdots \\ \sigma_\delta(x(t_N),u(t_N)) \\ \sigma_{\Delta_u}(x(t_N),u(t_N)) \end{bmatrix} \begin{bmatrix} W_V^* \\ W_Q^* \\ (W_R^*)^- \end{bmatrix} \end{bmatrix}, \tag{9}$$

where $\sigma_\delta\left(x(t_i), u(t_i)\right)$ and $\sigma_{\Delta_u}\left(x(t_i), u(t_i)\right)$ denotes $\sigma_\delta\left(x(t), u(t)\right)$ and $\sigma_{\Delta_u}\left(x(t), u(t)\right)$ evaluated at time t_i, respectively.

It is assumed that at every time instance t, the observer has access to a history stack $\mathcal{H} \coloneqq \left\{\hat{\Sigma}, \Sigma_u\right\}$, defined as

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

where time instances $t_1, \ldots, t_N$ are selected to ensure that the resulting history stack is full rank, as subsequently defined in Def. 2. Denoting the observer gain matrix by $K \in \mathbb{R}^{n+2P+M-1\times L+N(1+m)}$, the HSO is designed as

$$\begin{bmatrix} \dot{\hat{x}} \\ \dot{\hat{W}} \end{bmatrix} = \begin{bmatrix} A\hat{x} + Bu \\ 0_{2P+M-1} \end{bmatrix} + K\left(\begin{bmatrix} Cx \\ \Sigma_u \end{bmatrix} - \begin{bmatrix} C\hat{x} \\ \hat{\Sigma}\hat{W} \end{bmatrix}\right). \tag{10}$$

The error in equation (5) implies that the innovation $\Sigma_u - \hat{\Sigma}\hat{W}$ in (10) corresponds to the weight estimation error $\tilde{W}$ only if $\hat{\Sigma} = \Sigma$. Since $\hat{\Sigma}$ depends continuously on $\hat{x}$ and because $\hat{x}$ exponentially converges to x, $\hat{\Sigma}$ exponentially converges to Σ. As a result, newer and better estimates of x can be leveraged to improve the estimates of W^* by purging and refreshing the history stack $\mathcal{H}$. Due to purging, the time instances $\{t_1, \cdots t_N\}$ and the matrices $\hat{\Sigma}$ and Σ_u are piecewise constant functions of time.

Definition 2. The history stack is called *full rank* if $\operatorname{rank}\left(\hat{\Sigma}\right) = 2P + M - 1$. The signal $(\hat{x}, u)$ is called *finitely informative* (FI) if there exist time instances $0 \leq t_1 < t_2 < \cdots < t_N$ such that the resulting history stack is full rank and *persistently informative* (PI) if for any $T \geq 0$, there exist time instances $T \leq t_1 < t_2 < \cdots < t_N$ such that the resulting history stack is full rank.

A history stack management algorithm similar to [20, Fig. 1] is used to ensure the existence of a time instance t_M such that, if the signal $(\hat{x}, u)$ is FI, then the history stack is full rank for all $t \geq t_M$, and in addition, if it is PI, then $\lim_{t\to\infty} \left\|\Sigma(t) - \hat{\Sigma}(t)\right\| = 0$.

A. Observer Gain Design and Stability Analysis

The HSO gain matrix is designed in the block diagonal form $K_{HSO} \coloneqq \begin{bmatrix} K_3 & 0_{n\times N+Nm} \\ 0_{2P+M-1\times L} & K_4\left(\hat{\Sigma}^T\hat{\Sigma}\right)^{-1}\hat{\Sigma}^T \end{bmatrix}$, where $K_3 \in \mathbb{R}^{n\times L}$ is a constant gain matrix and $K_4 : \mathbb{R}_{\geq 0} \to \mathbb{R}^{2P+M-1\times 2P+M-1}$ is a potentially time-varying gain matrix. Provided the gain matrices are selected to satisfy the hypothesis of Theorem 2 below, the resulting observer in (10) can be shown to be convergent in the presence of PE and bounded under loss of excitation. Finite excitation is needed for the history stack to be full rank so that $(\hat{\Sigma}^T\hat{\Sigma})^{-1}$ is well-defined.

Theorem 2. Provided K_3 is selected such that $(A - K_3C)$ is Hurwitz, $K_4(t)$ is selected such that for $t < t_M$, $K_4(t) = 0$ and for $t \geq t_M$, $K_4(t)$ is symmetric positive definite, $0 < \underline{k} \leq \inf_{t\geq t_M}\{\lambda_{\min} K_4(t)\}$ and $\sup_{t\geq t_M}\{\|K_4(t)\|\} \leq \overline{k} < \infty$, then $\tilde{W}$ is ultimately bounded (UB) if the signal $(\hat{x}, u)$ is FI and $\lim_{t\to\infty}\tilde{W}(t) = 0$ if it is PI.

Proof. Using Theorem 1, if $(A - K_3C)$ is Hurwitz, $\tilde{x}(t) \to 0$ exponentially as $t \to \infty$. Using (10), the dynamics of the weight estimation error can be expressed as $\dot{\tilde{W}} = K_4(t)\hat{W} - K_4(t)\left(\hat{\Sigma}^T(t)\hat{\Sigma}(t)\right)^{-1}\hat{\Sigma}^T(t)\Sigma_u(t)$. Since K_4 is set to 0, the weight estimates are constant over $[0, t_M)$. For $t \geq t_M$, adding $\pm K_4(t)W^*$ to $\dot{\tilde{W}}$, and using the fact that $\Sigma W^* = \Sigma_u$, the weight estimation error dynamics can be treated as the controlled system

$$\dot{\tilde{W}} = -K_4(t)\tilde{W} + K_4(t)w, \tag{11}$$

where $w(t) \coloneqq \left(I - \left(\hat{\Sigma}^T(t)\hat{\Sigma}(t)\right)^{-1}\hat{\Sigma}^T(t)\Sigma(t)\right)W^*$ is treated as the control input. Using the Cauchy-Schwartz Inequality and the Rayleigh-Ritz Theorem [28], the orbital derivative of the positive definite candidate Lyapunov function $V\left(\tilde{W}\right) \coloneqq \frac{1}{2}\tilde{W}^T\tilde{W}$ along the trajectories of (11) can be bounded as

$$\dot{V}\left(t, \tilde{W}\right) \leq -\underline{k}\left\|\tilde{W}\right\|^2 + \overline{k}\left\|\tilde{W}\right\|\|w\|, \quad \forall t \geq t_M, \tag{12}$$

and $\tilde{W} \in \mathbb{R}^{2P+M-1}$. In the domain $\|\tilde{W}\| > \frac{2\overline{k}\|W^*\|}{\underline{k}}\|w\|$, the orbital derivative satisfies the bound $\dot{V}\left(t, \tilde{W}\right) \leq -\frac{\underline{k}}{2}\|\tilde{W}\|^2$. Using Theorem 4.19 from [27], it can be concluded that the controlled system in (11) is input-to-state stable (ISS).

If the signal $(\hat{x}, u)$ is PI, then the history stack can be purged and refreshed infinitely many times such that $w(t) \to 0$ as $t \to \infty$. Utilizing Exercise 4.58 from [27], it can then be concluded that $\tilde{W}(t) \to 0$ as $t \to \infty$.

If the signal $(\hat{x}, u)$ is FI but not PI, then there exists a time instance T such that the history stack remains unchanged for all $t \geq T$. As a result, there exists a constant $\overline{w}$ such that for all $t \geq T$, $\|w(t)\| \leq \overline{w}$. By the definition of ISS, it can then be concluded that $\tilde{W}$ is UB. □

Remark 3. The UB result in the absence of PE is a distinct advantage of HSO over MLO, which provides no such guarantee. Once the system states are no longer exciting, the MLO could potentially become unstable.

Remark 4. The IRL-O formulation is not restricted to the choices of K in Theorems 1 and 2. Different stabilizing or heuristic gain selection methods can be incorporated in the developed framework. For example, motivated by robustness to measurement noise, the use of a Kalman filter for gain selection is explored in Section VI.

VI. Simulations

A key motivation for casting the IRL problem into the observer framework is that the observer can be extended to a Kalman filter in a straightforward fashion to address measurement noise. To implement the developed observers as Kalman filters, all that is needed is to select the gains K_3 and K_4 using the Kalman gain update equations. The following simulation study demonstrates the validity, the robustness, and the performance of the designed observers and their Kalman filter implementation.

While the developed observer IRL methods are applicable to general output feedback linear systems, the concurrent learning (CL) method used for comparison is only applicable to a restricted set of systems (the state estimator in [29] is modified slightly for the non-Brunovsky form of (13)). In the following, to make comparisons feasible, a system that both methods are applicable to is selected.

The agent under observation has linear dynamics

$$\dot{x} = \begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix} x + \begin{bmatrix} 2 \\ 0.5 \end{bmatrix} u, \; y = \begin{bmatrix} 1 & 0 \end{bmatrix} x. \tag{13}$$

The optimal controller, $u^*(x) = -\begin{bmatrix} 4.14 & 5.53 \end{bmatrix} x$, minimizes an LQR problem, $Q = \text{diag}([2, 11])$ and $R = 1.5$, with an optimal value function $V^*(x) = 2.54x_1^2 + 7.59x_2^2 + 4.50x_1x_2$. The ideal weights that are to be estimated are $W_{V1}^* = 2.54, W_{V2}^* = 7.59, W_{V3}^* = 4.50, W_{Q1}^* = 2, W_{Q2}^* = 11$, and $R = 1.5$ is selected as the known value to remove the scaling ambiguity.

Since the system state estimates converge exponentially to the true system states, a time based purging technique similar to [20, Fig. 1] is utilized to reduce the estimation error associated with the system state estimates stored in the history stack. Furthermore, to improve numerical stability of gain computation, the history stack management algorithm also attempts to minimize the condition number of $\hat{\Sigma}^T\hat{\Sigma}$. In the presented simulation studies, the history stacks contain data for five previous time instances and are purged every 0.5 seconds if they can be repopulated.

Two simulation studies are performed. The first shows the performance of the designed observers in a noise-free setting. The second simulation incorporates noise in order to investigate the observers/filter robustness.

The error metric used to compare all of the observers/filters is the summation of the five relative weight estimation errors, defined as

$$\sum \frac{\tilde{W}_i}{W_i^*} := \frac{\|\tilde{W}_{V_1}\|}{W_{V_1}^*} + \frac{\|\tilde{W}_{V_2}\|}{W_{V_2}^*} + \frac{\|\tilde{W}_{V_3}\|}{W_{V_3}^*} + \frac{\|\tilde{W}_{Q_1}\|}{W_{Q_1}^*} + \frac{\|\tilde{W}_{Q_2}\|}{W_{Q_2}^*}. \tag{14}$$

A. Persistently Excited Signal without Noise

1) Two State System: The first simulation study concerns a noise-free environment. The controller that the agent under observation implements is a combination of the optimal controller, u^*, and a known additive excitation signal, i.e., the feedback controller of the agent is $u(t, x(t)) = u^*(x(t)) + u_{exc}(t)$, where $u_{exc}(t) := 5\sin(t) + 18\cos(0.4t) + 36\sin(2t) + 0.5\cos(3t)$ induces excitation in the signal $\hat{x}$.

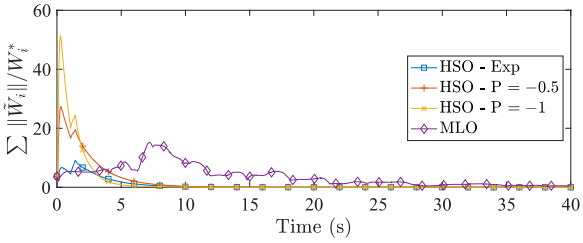

Fig. 1: Weight estimation errors for the developed observers with no noise and PE signal.

The HSO in (10), is implemented using three different K_{HSO} matrices, comprised of the same K_3 matrices, computed using the "place" command in MATLAB for poles $p_1 = -2$ and $p_2 = -4$, and three different K_4 matrices. The first two K_4 matrices are computed using gains $K_4 = -I$ and $K_4 - 0.5I$ (denoted in Fig. 1 as HSO - P $= -1$ and HSO - P $= -0.5$, respectively). The third K_4 matrix is selected to be an exponentially varying gain matrix, $K_4 = (1 - 0.9\exp^{-t})0.5I$ (denoted as HSO - Exp in Fig. 1). The MLO in (7) is implemented using a single K_{MLO} matrix, with K_1 computed using the "place" command for poles $p_1 = -2$ and $p_2 = -4$, and $K_2 = 10000I$.

As seen in Fig. 1, all of the weight estimation errors for the designed observers converge to the origin as expected. Even though there is a larger initial estimation error for the HSO, with constant gains, compared to the MLO, the history stack based observers converge much quicker than the MLO. The initial estimation error can be reduced for the HSO either by moving the poles closer to the origin, or implementing an exponentially varying gain matrix, as in the HSO-Exp case. The exponentially varying gain matrix combines the benefits of initial small gains, when the state estimates are inaccurate, with those of progressively larger gains, leading to fast convergence.

2) Four State System: The second simulation shows a four state system with the exponentially varying HSO and the Kalman filter implementation of the

B. Persistently Excited Signal with Noise

The second simulation is an investigation into noise robustness of the HSO and the Kalman filter implementation of the HSO (called HSO-KF) compared to the CL update law in [22], [23], and the state estimator in [29] (with a slight modification to address the non-Brunovsky form of the dynamics). Zero-mean Gaussian noise is added to y' and u, with three noise variances used to simulate low-noise ($R_1 = \text{diag}([0.1^2, 0.1^2])$), medium noise ($R_2 = \text{diag}(0.5^2, 0.5^2]$) and high noise ($R_3 = \text{diag}([1^2, 1^2])$) scenarios. Fifty Monte-Carlo simulations for each noise level are conducted and compared with the no-noise case. We do not study the behavior of the MLO under noisy measurements due to the added robustness of the HSO due to the use of past data.

The results of the simulation study are shown in Table I. As seen from the data, all three methods perform well in the noise free case, and the performance of all three methods is comparable in the low noise scenario. The advantages of the two HSO methods over the CL method are evident in the medium and high noise scenarios. Both the HSO-Exp and HSO-KF show better robustness to noise when compared to the CL method, especially in high the noise situation (CL steady state (SS) error is almost four times higher than both HSO methods). Comparing the results of HSO-Exp to HSO-KF, HSO-Exp has lower SS errors for the low and medium noise cases, while, HSO-KF has lower SS errors for the no noise and high noise cases. In addition, HSO-KF converges quicker in every case compared to both CL and HSO-Exp, as evidenced by the average over the whole time interval (TT).

TABLE I: Comparison between concurrent learning (CL), KF based implementation of HSO (HSO-KF), and exponential pole selection implementation of HSO (HSO-Exp), with different noise variances. Simulations were ran for 100 seconds over 50 trials with step size $Ts = 0.005$. The standard deviations (SD) simulated are $0.0, 0.1, 0.5$, and 1.0. The metric used for comparison is the average of the average on the trajectories $\sum \tilde{W}_i / W_i^*$, where TT denotes the average over the entire trajectory, and SS denotes the average over the last 30 seconds of the trajectory. The exponential HSO gains are selected similar to Section VI-A.1, except $K_4 = (1 - 0.9\exp^{-t})0.15I$. The Kalman filter gain is selected using the gain matrix $K_{HSO} = \text{diag}([K_3, K_4])$ where K_3 and K_4 are independent Kalman gains.

	CL		HSO-Exp		HSO-KF	
SD	TT	SS	TT	SS	TT	SS
0.0	0.9855	7.43e-05	0.9101	8.41e-05	0.0446	1.86e-14
0.1	0.8977	0.2647	0.8463	0.1652	0.2591	0.2279
0.5	2.1766	2.0336	1.3064	0.6894	0.7291	0.7277
1.0	5.5415	5.5223	1.9055	1.4667	1.5055	1.4111

VII. Conclusion

This paper presents a novel observer-like formulation for performing online estimation of reward functions using input-output observations. Two observers are proposed and their convergence guarantees are established. The Monte-Carlo simulations demonstrate that the developed observer based IRL techniques, utilizing exponentially varying gains and Kalman gains, demonstrate better noise robustness than existing CL based IRL techniques [22], [23].

Future work includes extension of the observer-based IRL methods to nonlinear systems and investigation of data storage algorithms for maintaining informative data in the history stack.

References

[1] S. Russell, "Learning agents for uncertain environments (extended abstract)," in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998.

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

[3] P. Abbeel and A. Y. Ng, "Inverse reinforcement learning," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Springer, Boston, MA, 2010, pp. 554–558.

[4] R. E. Kalman, "When is a linear control system optimal?" *J. Basic Eng.*, vol. 86, no. 1, pp. 51–60, 1964.

[5] S. Schaal, "Learning from demonstration," in *Advances in Neural Information Processing Systems 9*, M. I. Jordan and T. Petsche, Eds. Cambridge, MA: MIT Press, 1997, pp. 1040–1046.

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

[8] M. Wulfmeier, P. Ondruska, and I. Posner, "Maximum entropy deep inverse reinforcement learning," arXiv:1507.04888, 2015.

[9] Z. Li, J. Kiseleva, and M. de Rijke, "Dialogue generation: From imitation learning to inverse reinforcement learning," in *Proc. AAAI Conf. Artif. Intel.*, 2019, pp. 6722–6729.

[10] D. Brown, W. Goo, P. Nagarajan, and S. Niekum, "Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations," arXiv:1904.06387, 2019.

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

[12] ——, "Nonlinear inverse reinforcement learning with Gaussian processes," in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 19–27.

[13] A. Šošić, W. R. KhudaBukhsh, A. M. Zoubir, and H. Koeppl, "Inverse reinforcement learning in swarm systems," in *Proc. Conf. Auton. Agents MultiAgent Syst.* International Foundation for Autonomous Agents and Multiagent Systems, 2017, pp. 1413–1421.

[14] X. Wang and D. Klabjan, "Competitive multi-agent inverse reinforcement learning with sub-optimal demonstrations," in *Proc. Int. Conf. Mach. Learn.*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 5143–5151.

[15] B. Michini and J. P. How, "Bayesian nonparametric inverse reinforcement learning," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, P. A. Flach, T. D. Bie, and N. Cristianini, Eds. Springer Berlin Heidelberg, 2012, vol. 7524, pp. 148–163.

[16] K. G. Vamvoudakis and F. L. Lewis, "Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[17] D. Wang, D. Liu, H. Li, B. Luo, and H. Ma, "An approximate optimal control approach for robust stabilization of a class of discrete-time nonlinear systems with uncertainties," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 46, no. 5, pp. 713–717, 2016.

[18] R. Kamalapurkar, P. Walters, J. A. Rosenfeld, and W. E. Dixon, *Reinforcement learning for optimal feedback control: A Lyapunov-based approach*, ser. Communications and Control Engineering. Springer International Publishing, 2018.

[19] T. Molloy, J. Ford, and T. Perez, "Online inverse optimal control on infinite horizons," in *IEEE Conf. Decis. Control*. IEEE, 2018, pp. 1663–1668.

[20] R. Kamalapurkar, "Linear inverse reinforcement learning in continuous time and space," in *Proc. Am. Control Conf.*, 2018, pp. 1683–1688.

[21] R. V. Self, M. Harlan, and R. Kamalapurkar, "Online inverse reinforcement learning for nonlinear systems," in *Proc. IEEE Conf. Control Technol. Appl.* IEEE, 2019, pp. 296–301.

[22] R. V. Self, S. M. N. Mahmud, K. Hareland, and R. Kamalapurkar, "Online inverse reinforcement learning with limited data," in *Proc. IEEE Conf. Decis. Control*, to appear, see arXiv:2008.08972.

[23] R. V. Self, M. Abudia, and R. Kamalapurkar, "Online inverse reinforcement learning for systems with disturbances," in *Proc. Am. Control Conf.*, to appear, see arXiv:2003.03912.

[24] G. Rotithor, D. Trombetta, R. Kamalapurkar, and A. P. Dani, "Reduced order observer for structure from motion using concurrent learning," in *Proc. IEEE Conf. Decis. Control*, 2019, pp. 6815–6820.

[25] D. Liberzon, *Calculus of variations and optimal control theory: a concise introduction*. Princeton University Press, 2012.

[26] S. S. Sastry and M. Bodson, *Adaptive control: stability, convergence, and robustness*. Upper Saddle River, NJ: Prentice-Hall, 1989.

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

[28] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge: Cambridge University Press., 1993.

[29] R. Kamalapurkar, "Online output-feedback parameter and state estimation for second order linear systems," in *Proc. Am. Control Conf.*, 2017, pp. 5672–5677.