# Online inverse reinforcement learning with limited data

Ryan Self, S M Nahid Mahmud, Katrine Hareland and Rushikesh Kamalapurkar

***Abstract*— This paper addresses the problem of online inverse reinforcement learning for systems with limited data and uncertain dynamics. In the developed approach, the state and control trajectories are recorded online by observing an agent perform a task, and reward function estimation is performed in real-time using a novel inverse reinforcement learning approach. Parameter estimation is performed concurrently to help compensate for uncertainties in the agent's dynamics. Data insufficiency is resolved by developing a data-driven update law to estimate the optimal feedback controller. The estimated controller can then be queried to artificially create additional data to drive reward function estimation.**

## I. Introduction

Based on the premise that the most succinct representation of the behavior of an entity is its reward structure [1], this paper aims to recover the reward (or cost) function of a demonstrator by monitoring its state and control trajectories. Reward function estimation is performed in the presence of modeling uncertainties for situations with limited data via inverse reinforcement learning (IRL) [1], [2].

While IRL in an *offline* setting has a rich history of literature [1]–[11], little work has been done to address IRL in an online setting. One reason for this is the limited data provided by a single demonstration.

Preliminary results on online IRL are available for linear systems, in results such as [12] and [13], and for nonlinear systems, in results such as [14] and [15]. However, [12] and [14] exploit access to demonstrator's feedback policy, [13] requires exact model knowledge, and [15] exploits identical disturbances to provide sufficient excitation. The main contribution of this paper is the development of a novel method for reward function estimation for an agent in situations where estimation of the demonstrator's optimal feedback law is less data-intensive than direct estimation of its reward function.

The novelty in the technique developed in this paper is a recursive model-based IRL approach which facilitates the use of off-trajectory state-action pairs. A majority of IRL methods are trajectory-driven and model-free. As a result, the trajectories need to be sufficiently information-rich for reward function estimation. The technique developed in this paper is model-based, and as a result, once a model is learned, arbitrary state-action pairs can be used for IRL as long as the action is the optimal action for that state. In [12] and [14], the off-trajectory state-action pairs are generated under the assumption that the learner either knows the demonstrator's optimal feedback law or can query the demonstrator to find out what the optimal action would be at a given off-trajectory state. In this paper, we develop a novel IRL approach that relaxes the aforementioned assumption.

The key idea in this paper is to estimate the optimal feedback controller of the agent online, and use that estimate to artificially create off-trajectory data to drive reward function estimation. In the authors' previous work [14], reward function estimation is performed directly using the agents observed trajectories. Instead, in this paper, the trajectory information is used to estimate the optimal feedback controller. This controller is parameterized as a neural network and estimated using a concurrent learning update law. The estimated controller is simultaneously queried to create off-trajectory data which is then used for reward function estimation via IRL. Since the optimal controller is estimated using a neural network, the controller can be estimated independent of the modeling uncertainty. In the developed approach, parameter estimation and two update laws for estimation of the optimal feedback controller and reward function are utilized simultaneously, to achieve uniform ultimate boundedness of the unknown reward function weights.

The paper is organized as follows: Section II explains the notation used throughout the paper. Section III details the problem formulation. Section IV shows how to estimate the optimal controller. Section V explains the IRL algorithm. Section VI shows a simulation example and Section VII concludes the paper.

## II. Notation

The notation $\mathbb{R}^n$ represents the $n-$dimensional Euclidean space, and the elements of $\mathbb{R}^n$ are interpreted as column vectors, where $(\cdot)^T$ denotes the vector transpose operator. The set of positive integers excluding 0 is denoted by $\mathbb{N}$. For $a \in \mathbb{R}$, $\mathbb{R}_{\geq a}$ denotes the interval $[a, \infty)$, and $\mathbb{R}_{>a}$ denotes the interval $(a, \infty)$. If $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^n$, then $[a; b]$ denotes the concatenated vector $\begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{m+n}$. The notations $\mathrm{I}_n$ and $0_n$ denote the $n \times n$ identity matrix and the zero element of $\mathbb{R}^n$, respectively. Whenever it is clear from the context, the subscript $n$ is suppressed.

## III. Problem Formulation

Consider an agent with the following dynamics

$$\dot{x} = f(x, u), \tag{1}$$

The authors are with the School of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK, USA. {rself,nahid.mahmud,katrine.hareland, 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 agencies.

where $x : \mathbb{R}_{\geq T_0} \to \mathbb{R}^n$ is the state, $u : \mathbb{R}_{\geq T_0} \to \mathbb{R}^m$ is the control, and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is a continuously differentiable function.

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

$$J(x_0, u(\cdot)) = \int_{T_0}^{\infty} r(x(t; x_0, u(\cdot)), u(t))\, \mathrm{d}t, \tag{2}$$

where $x(\cdot; x_0, u(\cdot))$ is the trajectory of the agent generated by the optimal control signal $u(\cdot)$ that minimizes the performance index in (2) starting from the initial condition $x_0$ and beginning at time $T_0$. The main objective of the paper is to estimate the unknown reward function, $r$, using input-state pairs.

The following assumptions are used throughout the rest of this paper.

**Assumption 1.** *The unknown reward function $r$ is quadratic in the control, i.e.,*

$$r(x, u) = Q(x) + u^T R u, \tag{3}$$

*where $R \in \mathbb{R}^{m \times m}$ is a positive definite matrix, such that $R = \mathrm{diag}([r_1, \cdots, r_m])$.*

[16], [17] The continuous function $Q$ can be represented using a neural network as $Q(x) = (W_Q^*)^T \sigma_Q(x) + \epsilon_Q(x)$, where $W_Q^* := [q_1, \cdots, q_L]^T$ are ideal reward function weights, $\sigma_Q : \mathbb{R}^n \to \mathbb{R}^L$ are known continuously differentiable features, and $\epsilon_Q : \mathbb{R}^n \to \mathbb{R}$ is the approximation error.

**Assumption 2.** *The dynamics for the agent are affine in control and can be expressed as*

$$\dot{x} = f^o(x, u) + \theta^T \sigma(x, u) + \epsilon(x, u), \tag{4}$$

*where $f^o : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ denotes the continuously differentiable nominal dynamics, $\theta^T \sigma$ is a parameterized estimate of the uncertain part of the dynamics, where $\theta \in \mathbb{R}^{p \times n}$ is a matrix of unknown constant parameters and $\sigma : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$ are known continuously differentiable features, and $\epsilon : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ denotes the function approximation error.*

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) [18] equation

$$H\left(x(t), \left([\nabla_x V^*](x(t))\right)^T, u(t)\right) = 0, \forall t \in \mathbb{R}_{\geq T_0}, \tag{5}$$

where the unknown optimal value function is $V^* : \mathbb{R}^n \to \mathbb{R}$ and $H : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is the Hamiltonian, defined as $H(x, p, u) := p^T f(x, u) + r(x, u)$. The goal of IRL is to estimate the reward function, $r$.

To aid in the estimation of the reward function, let $\hat{V} : \mathbb{R}^n \times \mathbb{R}^P \to \mathbb{R}$, $\left(x, \hat{W}_V\right) \mapsto \hat{W}_V^T \sigma_V(x)$ be a parameterized estimate of the optimal value function $V^*$, where $\hat{W}_V \in \mathbb{R}^P$ are the estimates of the ideal value function weights $W_V^*$ and $\sigma_V : \mathbb{R}^n \to \mathbb{R}^P$ are known continuously differentiable features. Let $\epsilon_V : \mathbb{R}^n \to \mathbb{R}$, defined as $\epsilon_V(x) = V^*(x) - (W_V^*)^T \sigma_V(x)$, be the resulting approximation error. Using $\hat{W}_V$, $\hat{W}_Q$, and $\hat{W}_R$, which are the estimates of $W_V^*$, $W_Q^*$, and $W_R^* := [r_1, \cdots, r_m]^T$, respectively, in (5), the inverse Bellman error $\delta : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{L+P+m} \to \mathbb{R}$ is obtained as

$$\begin{aligned}\delta\left(x, u, \hat{W}\right) =& \hat{W}_V^T \Big([\nabla_x \sigma_V](x)\Big) f(x, u) + \hat{W}_Q^T \sigma_Q(x) \\ &+ \hat{W}_R^T \sigma_u(u), \end{aligned} \tag{6}$$

where $\sigma_u(u) := \left[u_1^2, \cdots, u_m^2\right]$.

For brevity of presentation, it is assumed that a parameter estimator that satisfies the following properties is available. For examples of such parameter estimates, see [14], [19].

**Assumption 3.** *[20, Assumption 2] A compact set $\Theta \subset \mathbb{R}^p$ such that $\theta \in \Theta$ is known a priori. The estimate $\hat{\theta} : \mathbb{R}_{\geq T_0} \to \mathbb{R}^p$ are updated based on a switched update law of the form*

$$\dot{\hat{\theta}} = f_{\theta_s}(\hat{\theta}(t), t),$$

*$\hat{\theta}(T_0) = \hat{\theta}_0 \in \Theta$, where $s \in \mathbb{N}$ denotes the switching index and $\{f_{\theta_s} : \mathbb{R}^p \times \mathbb{R}_{\geq T_0} \to \mathbb{R}^p\}_{s \in \mathbb{N}}$ denotes the family of continuously differentiable functions. The dynamics of the parameter estimation error $\tilde{\theta} : \mathbb{R}_{\geq T_0} \to \mathbb{R}^p$, defined as $\tilde{\theta}(t) := \theta - \hat{\theta}(t)$, can be expressed as $\dot{\tilde{\theta}}(t) = f_{\theta_s}\left(\theta - \tilde{\theta}(t), t\right)$. Furthermore, there exists a continuously differentiable function $V_\theta : \mathbb{R}^p \times \mathbb{R}_{\geq T_0} \to \mathbb{R}_{\geq 0}$ that satisfies*

$$\underline{\nu}_\theta\left(\left\|\tilde{\theta}\right\|\right) \leq V_\theta\left(\tilde{\theta}, t\right) \leq \overline{\nu}_\theta\left(\left\|\tilde{\theta}\right\|\right),$$

*and*

$$\begin{aligned}\left([\nabla_{\tilde{\theta}} V_\theta]\left(\tilde{\theta}, t\right)\right)\left(-f_{\theta_s}\left(\theta - \tilde{\theta}, t\right)\right) + \frac{\partial V_\theta\left(\tilde{\theta}, t\right)}{\partial t} \\ \leq -K\left\|\tilde{\theta}\right\|^2 + D\left\|\tilde{\theta}\right\|,\end{aligned}$$

*for all $s \in \mathbb{N}, t \in \mathbb{R}_{\geq T_0}$, and $\tilde{\theta} \in \mathbb{R}^p$, where $\underline{\nu}_\theta, \overline{\nu}_\theta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ are class $\mathcal{K}$ functions, $K \in \mathbb{R}_{>0}$ is an adjustable parameter, and $D \in \mathbb{R}_{>0}$ is a positive constant.*

Utilizing parameter estimates from Assumption 3, (6) can be updated and expressed as

$$\begin{aligned}\delta'\left(x, u, \hat{W}, \hat{\theta}\right) =& \hat{W}_V^T \Big([\nabla_x \sigma_V](x)\Big) \hat{Y}(x, u, \hat{\theta}) + \hat{W}_Q^T \sigma_Q(x) \\ &+ \hat{W}_R^T \sigma_u(u), \end{aligned} \tag{7}$$

where $\hat{Y}(x, u, \hat{\theta}) := f^o(x, u) + \hat{\theta}^T \sigma(x, u)$ and $\hat{\theta}$ are estimates of unknown parameters. Rearranging, (7) becomes

$$\delta'\left(x, u, \hat{W}', \hat{\theta}\right) = \left(\hat{W}'\right)^T \sigma'\left(x, u, \hat{\theta}\right), \tag{8}$$

where $\hat{W}' := \left[\hat{W}_V; \hat{W}_Q; \hat{W}_R\right]$ and $\sigma'\left(x, u, \hat{\theta}\right) := \left[\Big([\nabla_x \sigma_V](x)\Big) \hat{Y}(x, u, \hat{\theta}); \sigma_Q(x); \sigma_u(u)\right]$.

In the following, the parameter estimator is executed synchronously in with IRL and in real-time.

## IV. Optimal Controller Estimation

Since a large majority of optimal control problems are aimed at driving the state to a set-point or an error signal to zero, information content of the state and control trajectories can quickly decay to zero rendering them unable to provide usable data. More specifically, once the states converge, newer data points from the agent's trajectories will simply provide zero, or near-zero, values for both the states (or errors) and the controls. As a result, the reward function estimate may never converge. Motivated by the observation that knowledge of the optimal controller can be leveraged to artificially create additional data to drive IRL, this section develops a process for finding an estimate of the optimal controller.

### A. Controller Estimation Formulation

Provided Assumptions 1 and 2 are satisfied, the closed-form nonlinear optimal controller corresponding to the reward structure in (2) is

$$u^*(x) = -\frac{1}{2}R^{-1}\Big( \left[\nabla_u f\right](x)\Big)^T \Big( \left[\nabla_x V^*\right](x)\Big)^T, \quad (9)$$

where $u^* := [u_1, u_2, \cdots, u_m]^T$ and $\Big( \left[\nabla_u f\right](x) \Big)$ is found from $f(x,u)$ in (1). To promote estimation, $u^*$ will be represented as

$$u^*(x) = -\left(W_u^*\right)^T \sigma_u(x) + \epsilon_u(x), \quad (10)$$

where $W_u^* \in \mathbb{R}^{K\times m}$ is a matrix of unknown ideal constant parameters, $\sigma_u : \mathbb{R}^n \to \mathbb{R}^K$ are known continuously differentiable features, and $\epsilon_u : \mathbb{R}^n \to \mathbb{R}^m$ is the resulting approximation error.

Collecting state and control signals over time instances, $t_1, t_2, \cdots, t_M$, stored in a history stack, denoted as $\mathcal{H}^u$, (10) can be formulated into the matrix form

$$-\Sigma_u - \Sigma_\sigma \hat{W}_u = \Sigma_\sigma \tilde{W}_u - \Delta_u, \quad (11)$$

where $\Sigma_u := [u^T(t_1); u^T(t_2); \cdots; u^T(t_M)]$, $\Sigma_\sigma := [\sigma_u^T(x(t_1)); \sigma_u^T(x(t_2)); \cdots; \sigma_u^T(x(t_M))]$, and $\Delta_u := [\epsilon_u^T(x(t_1)); \epsilon_u^T(x(t_2)); \cdots; \epsilon_u^T(x(t_M))]$. The weight estimation error is defined as $\tilde{W}_u = W_u^* - \hat{W}_u$, where $\hat{W}_u$ is the estimate of $W_u^*$.

Using (11), a recursive least-squares update law to estimate the unknown weights is designed as

$$\dot{\hat{W}}_u = \alpha_u \Gamma_u \Sigma_\sigma^T \left(-\Sigma_u - \Sigma_\sigma \hat{W}_u\right). \quad (12)$$

where $\alpha_u \in \mathbb{R}_{>0}$ is a constant adaptation gain, and $\Gamma_u : \mathbb{R}_{\geq 0} \to \mathbb{R}^{K\times K}$ is the least-squares gain updated using the update law

$$\dot{\Gamma}_u = \beta_u \Gamma_u - \alpha_u \Gamma_u \Sigma_\sigma^T \Sigma_\sigma \Gamma_u. \quad (13)$$

where $\beta_u \in \mathbb{R}_{>0}$ is the forgetting factor.

### B. Analysis

The time-varying history stack, $\mathcal{H}^u$, is called full rank, uniformly in $t$, if there exists a $\underline{k} > 0$ such that $\forall t \in \mathbb{R}_{\geq T_0}$,

$$0 < \underline{k} < \lambda_{\min}\left\{\Sigma_\sigma^T(t)\Sigma_\sigma(t)\right\}. \quad (14)$$

Using arguments similar to [21, Corollary 4.3.2], it can be shown that if $\lambda_{\min}\left\{\Gamma_u^{-1}(0)\right\} > 0$, and if $\mathcal{H}^u$ is full rank, uniformly in $t$, then the least squares gain matrix satisfies

$$\underline{\Gamma}_u \mathrm{I}_K \leq \Gamma_u(t) \leq \overline{\Gamma}_u \mathrm{I}_K, \quad (15)$$

where $\underline{\Gamma}_u$ and $\overline{\Gamma}_u$ are positive constants.

To facilitate the following analysis, using (11) and (12), the dynamics for the weight estimation error can be described by

$$\dot{\tilde{W}}_u = -\alpha_u \Gamma_u \Sigma_\sigma^T \left(\Sigma_\sigma \tilde{W}_u - \Delta_u\right). \quad (16)$$

**Theorem 1.** *If $\mathcal{H}^u$ is full rank, uniformly in $t$, then $t \mapsto \tilde{W}_u(t)$ is uniformly ultimately bounded.*

*Proof.* Consider the following positive definite candidate Lyapunov function

$$V_u(\tilde{W}_u, t) = \mathrm{tr}(\tilde{W}_u^T \Gamma_u^{-1}(t) \tilde{W}_u), \quad (17)$$

Using the bounds in (15), the candidate Lyapunov function satisfies

$$\frac{1}{\overline{\Gamma}_u}\left\|\tilde{W}_u\right\|^2 \leq V_u\left(\tilde{W}_u, t\right) \leq \frac{1}{\underline{\Gamma}_u}\left\|\tilde{W}_u\right\|^2. \quad (18)$$

Taking the time derivative of (17), and using (13) and (16), along with the identity $\dot{\Gamma}_u^{-1} = -\Gamma_u^{-1}\dot{\Gamma}_u\Gamma_u^{-1}$, after simplifying yields

$$\begin{aligned}\dot{V}_u(\tilde{W}_u, t) = &-\alpha_u \mathrm{tr}(\tilde{W}_u^T \Sigma_\sigma^T \Sigma_\sigma \tilde{W}_u) \\ &+ 2\alpha_u \mathrm{tr}(\tilde{W}_u^T \Sigma_\sigma^T \Delta_u) - \beta_u \mathrm{tr}(\tilde{W}_u^T \Gamma_u^{-1}(t) \tilde{W}_u). \quad (19)\end{aligned}$$

Using the Cauchy-Schwartz inequality, and bounds in (14) and (15), $\dot{V}_u$ can be bounded by

$$\begin{aligned}\dot{V}_u(\tilde{W}_u, t) \leq &-\left(\alpha_u \underline{k} + \frac{\beta_u}{\overline{\Gamma}_u}\right)\left\|\tilde{W}_u\right\|^2 \\ &+ 2\alpha_u \left\|\tilde{W}_u\right\| \|\Sigma_\sigma\| \|\Delta_u\|. \quad (20)\end{aligned}$$

Since the states and controls are both bounded, $\|\Sigma_\sigma\|$ and $\|\Delta_u\|$ are bounded above. The upper bounds are defined as $\overline{\Sigma}_\sigma$ and $\overline{\Delta}_u$. Using these upper bounds and Young's Inequality, $\dot{V}_u$ becomes

$$\dot{V}_u(\tilde{W}_u, t) \leq -A V_u\left(\tilde{W}_u, t\right) + B, \quad (21)$$

where $A$ and $B$ are defined as

$$A := \frac{\underline{\Gamma}_u}{2}\left(\alpha_u \underline{k} + \frac{\beta_u}{\overline{\Gamma}_u}\right), \quad (22)$$

and

$$B := \frac{2(\alpha_u \overline{\Sigma}_\sigma \overline{\Delta}_u)^2}{(\alpha_u \underline{k} + \beta_u/\overline{\Gamma}_u)}. \quad (23)$$

Finding the solution of (21) yields

$$V_u(t) \leq V_{u_0} e^{-A(t-T_0)} + \frac{B}{A}, \quad (24)$$

where $V_{u_0} \geq \left\| V_u\left(\tilde{W}_u\left(T_0\right), T_0\right)\right\|$. It can be concluded that

$$\lim_{t\to\infty} V_u(t) \leq \frac{B}{A}. \tag{25}$$

It can further be concluded that $\tilde{W}_u$ decays exponentially, such that

$$\lim_{t\to\infty} \left\|\tilde{W}_u\left(t\right)\right\| \leq \sqrt{\bar{\Gamma}_u \frac{B}{A}}. \tag{26}$$

□

## V. Inverse Reinforcement Learning

In this section, the optimal feedback estimator developed in this previous section is utilized to create a data-set of estimated near-optimal state-action pairs to drive IRL.

### A. Utilizing Control and Parameter Estimates

Consider a time instance, $t_i$. For each time $t_i$, select an arbitrary state, denoted by $x_i$, and let $\hat{u}_i := \hat{W}_u^T(t_i)\sigma_u(x_i)$ be the estimate of the optimal controller $u_i^*$ at state $x_i$ and $t_i$. The updated inverse Bellman error, when evaluated at the arbitrarily selected state and at time $t_i$ using the estimates of the model and the optimal controller, is given by

$$\delta''\Big(t_i, x_i, \hat{u}_i, \hat{W}'(t_i), \hat{\theta}(t_i)\Big) = \Big(\hat{W}'(t_i)\Big)^T \sigma'\Big(t_i, x_i, \hat{u}_i, \hat{\theta}(t_i)\Big), \tag{27}$$

where

$$\hat{W}'(t_i) := \Big[\hat{W}_V(t_i); \hat{W}_Q(t_i); \hat{W}_R(t_i)\Big]$$

and

$$\sigma'\left(t_i, x_i, \hat{u}_i, \hat{\theta}(t_i)\right) := \Big[\Big(\left[\nabla_x \sigma_V\right](x_i)\Big)(f^o(x_i, \hat{u}_i) + \hat{\theta}^T(t_i)\sigma(x_i, \hat{u}_i)); \sigma_Q\left(x_i\right); \sigma_u\left(\hat{u}_i\right)\Big].$$

Since all positive multiples of a reward function result in the same optimal controller, given state-action pairs, the reward function can only be identified up to a scale. As a result, one of the reward function weights can be arbitrarily assigned.

Since optimal control behaviors are scale-invariant, there is no loss of generality in resolving the scale ambiguity by taking the first element of $\hat{W}_R$ to be known. The inverse BE in (27) can then be expressed as

$$\delta''\Big(t_i, x_i, \hat{u}_i, \hat{W}(t_i), \hat{\theta}(t_i)\Big) = \Big(\hat{W}(t_i)\Big)^T \sigma''\Big(t_i, x_i, \hat{u}_i, \hat{\theta}(t_i)\Big) + r_1 \sigma_{u1}\left(\hat{u}_i\right), \tag{28}$$

where $\hat{W}(t_i) := \Big[\hat{W}_V(t_i); \hat{W}_Q(t_i); \hat{W}_R^-(t_i)\Big]$, the vector $\hat{W}_R^-$ denotes $\hat{W}_R$ with the first element removed, $\sigma_{uj}\left(\hat{u}_i\right)$ denotes the $j$th element of the vector $\sigma_u\left(\hat{u}_i\right)$, the vector $\sigma_u^-$ denotes $\sigma_u$ with the first element removed, and

$$\sigma''\left(t_i, x_i, \hat{u}_i, \hat{\theta}(t_i)\right) := \Big[\Big(\left[\nabla_x \sigma_V\right](x_i)\Big)(f^o(x_i, \hat{u}_i) + \hat{\theta}^T(t_i)\sigma(x_i, \hat{u}_i)); \sigma_Q\left(x_i\right); \sigma_u^-\left(\hat{u}_i\right)\Big]. \tag{29}$$

The closed-form nonlinear optimal controller corresponding to the reward structure in (2) provides the relationship

$$-2Ru^*\left(x_i\right) = \Big(\left[\nabla_u f\right]\left(x_i\right)\Big)^T \Big(\left[\nabla_x \sigma_V\right]\left(x_i\right)\Big)^T W_V^* + \Big(\left[\nabla_u f\right]\left(x_i\right)\Big)^T \Big(\left[\nabla_x \epsilon_V\right]\left(x_i\right)\Big)^T. \tag{30}$$

Utilizing estimates $\hat{\theta}(t_i)$ and data pairs $(x_i, \hat{u}_i)$ in (30), subtracting $H\left(x_i, \Big(\left[\nabla_x V\right]\left(x_i\right)\Big), u^*(x_i)\right)$ from (28), evaluating (28) and (30) at time instances $\{t_i\}_{i=1}^N$, and stacking the results in a matrix form, we get

$$-\hat{\Sigma}\hat{W} - \hat{\Sigma}_{u1} = \hat{\Sigma}\tilde{W} - \Delta, \tag{31}$$

where the weight estimation error is defined as $\tilde{W} = W^* - \hat{W}$, and $\hat{W}$ is the estimate of $W^*$, and

$$\hat{\Sigma} := \Big[\sigma^T\Big(t_1, x_1, \hat{u}_1, \hat{\theta}\left(t_1\right)\Big); \cdots; \sigma^T\Big(t_N, x_N, \hat{u}_N, \hat{\theta}\left(t_N\right)\Big)\Big],$$
$$\hat{\Sigma}_{u1} := \left[\sigma'_{u1}\left(\hat{u}_1\right); \cdots; \sigma'_{u1}\left(\hat{u}_N\right)\right],$$
$$\Delta := \Big[\Delta_\delta(t_1); \Delta_m(t_1); \cdots; \Delta_\delta(t_N); \Delta_m(t_N)\Big],$$

where

$$\sigma'_{u1}(\hat{u}_i) := \left[r_1 \sigma_{u1}\left(\hat{u}_{1i}\right); 2r_1 \hat{u}_{1i}; 0_{(m-1)\times 1}\right],$$
$$\sigma := \left[\sigma''\left[\begin{bmatrix} G \\ 0_{m\times L},\ \begin{bmatrix} 0_{1\times m-1} \\ 2\text{diag}\left([\hat{u}_{2i}, \cdots, \hat{u}_{mi}]\right)\end{bmatrix}\end{bmatrix}^T\right],$$
$$G := \Big(\left[\nabla_x \sigma_V\right]\left(x_i\right)\Big)\Big(\Big(\left[\nabla_u f^o\right]\left(x_i\right)\Big) + \hat{\theta}^T(t_i)\Big(\left[\nabla_u \sigma\right]\left(x_i\right)\Big)\Big),$$

$$\Delta_\delta(t_i) := 2R\tilde{u}_i + \Big(\left[\nabla_u \sigma\right]\left(x_i\right)\Big)^T \tilde{\theta}(t_i)\Big(\left[\nabla_u \sigma_V\right]\left(x_i\right)\Big)^T W_V^* + \Big(\Big(\left[\nabla_u f^o\right]\left(x_i\right)\Big) + \theta^T(t_i)\Big(\left[\nabla_u \sigma\right]\left(x_i\right)\Big)\Big)^T \Big(\left[\nabla_x \epsilon_V\right]\left(x_i\right)\Big)^T + \Big(\left[\nabla_u \epsilon\right]\left(x_i, u_i^*\right)\Big)\Big(\left[\nabla_x \sigma_V\right]\left(x_i\right)\Big) W_V^*,$$

$$\Delta_m(t_i) := \Big(\sigma_u(u_i^*) - \sigma_u(\hat{u}_i)\Big)^T W_R^* + \epsilon_V(x_i) + \epsilon_Q(x_i) + \Big(f^o\left(x_i, u_i^*\right) - f^o\left(x_i, \hat{u}_i\right)\Big)^T \Big(\left[\nabla_x \sigma_V\right]\left(x_i\right)\Big)^T W_V^* + \left(\theta^T\left(\sigma(x_i, u_i^*) - \sigma(x_i, \hat{u}_i)\right)\right)^T \Big(\left[\nabla_x \sigma_V\right]\left(x_i\right)\Big)^T W_V^* + \Big(\tilde{\theta}^T(t_i)\sigma(x_i, \hat{u}_i) + \epsilon(x_i, u_i^*)\Big)^T \Big(\left[\nabla_x \sigma_V\right]\left(x_i\right)\Big)^T W_V^*,$$

and $\hat{u}_{ji}$ is the $j$th element of $\hat{u}_i$.

A history stack, denoted as $\mathcal{H}^{IRL}$, is a set of ordered pairs of parameter estimates, $\hat{\theta}(t_i)$, and data pairs, $(x_i, \hat{u}_i)$, collected over time instance $t_1, t_2, \ldots, t_N$ into matrices $\Big(\hat{\Sigma}, \hat{\Sigma}_{u1}\Big)$.

Due to the fact that $\hat{\Sigma}$ and $\Delta$ depend on the quality of the control and parameter estimates, a purging technique is incorporated in the following to remove poor estimates $\hat{u}$ and $\hat{\theta}$ from $\mathcal{H}^{IRL}$. During the transient phase of the control and parameter estimators, the estimates $\hat{u}$ and $\hat{\theta}$ are likely to be less accurate and the resulting values of $\hat{W}$ are likely to

be poor. Purging facilitates usage of better estimates as they become available.

The recursive update law is then designed as

$$\dot{\hat{W}} = \alpha \Gamma \hat{\Sigma}^T \left( -\hat{\Sigma}\hat{W} - \hat{\Sigma}_{u1} \right). \tag{32}$$

In (32), $\alpha \in \mathbb{R}_{>0}$ is a constant adaptation gain and $\Gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}^{(L+P+m-1)\times(L+P+m-1)}$ is the least-squares gain updated using the update law

$$\dot{\Gamma} = \beta \Gamma - \alpha \Gamma \hat{\Sigma}^T \hat{\Sigma} \Gamma, \tag{33}$$

where $\beta \in \mathbb{R}_{>0}$ is the forgetting factor.

### *B. Analysis*

A Lyapunov based analysis is performed to show convergence for the IRL method in Section V.

The time-varying history stack, $\mathcal{H}^{IRL}$, is called full rank, uniformly in $t$, if there exists a $\underline{\sigma} > 0$ such that $\forall t \in \mathbb{R}_{\geq T_0}$,

$$0 < \underline{\sigma} < \lambda_{\min} \left\{ \hat{\Sigma}^T(t) \hat{\Sigma}(t) \right\}. \tag{34}$$

Using arguments similar to [21, Corollary 4.3.2], it can be shown that if $\lambda_{\min} \left\{ \Gamma^{-1}(T_0) \right\} > 0$, and if $\mathcal{H}^{IRL}$ is full rank, uniformly in $t$, then the least squares gain matrix satisfies

$$\underline{\Gamma} \mathrm{I}_{L+P+m-1} \leq \Gamma(t) \leq \overline{\Gamma} \mathrm{I}_{L+P+m-1}, \tag{35}$$

where $\underline{\Gamma}$ and $\overline{\Gamma}$ are positive constants.

To facilitate the following Lyapunov analysis, using (32), the dynamics for the weight estimation error can be described by

$$\dot{\tilde{W}} = -\alpha \Gamma \hat{\Sigma}^T \left( \hat{\Sigma}\tilde{W} - \Delta \right). \tag{36}$$

The stability result is summarized in the following theorem.

**Theorem 2.** *If $\mathcal{H}^{IRL}$ is full rank, uniformly in $t$, then $t \mapsto \tilde{W}(t)$ is uniformly ultimately bounded.*

*Proof.* Consider the positive definite candidate Lyapunov function

$$V(\tilde{W}, t) = \frac{1}{2} \tilde{W}^T \Gamma^{-1}(t) \tilde{W}. \tag{37}$$

Using the bounds in (35), the candidate Lyapunov function satisfies

$$\frac{1}{2\overline{\Gamma}} \left\| \tilde{W} \right\|^2 \leq V\left( \tilde{W}, t \right) \leq \frac{1}{2\underline{\Gamma}} \left\| \tilde{W} \right\|^2. \tag{38}$$

Taking the time-derivative of (37), and using (33) and (36), along with the identity $\dot{\Gamma}^{-1} = -\Gamma^{-1}\dot{\Gamma}\Gamma^{-1}$, after simplifying the time-derivative can be expressed as

$$\dot{V}(\tilde{W}, t) = -\frac{1}{2}\alpha \tilde{W}^T \hat{\Sigma}^T \hat{\Sigma} \tilde{W} + \alpha \tilde{W}^T \hat{\Sigma}^T \Delta - \frac{1}{2}\beta \tilde{W}^T \Gamma^{-1}(t) \tilde{W}. \tag{39}$$

Substituting in $\hat{\Sigma} = \Sigma - \tilde{\Sigma}$, and using the Cauchy-Schwartz inequality and bounds in (34) and (35), $\dot{V}$ can be bounded by

$$\dot{V}(\tilde{W}, t) \leq -\frac{1}{2} \left( \alpha \underline{\sigma} + \frac{1}{\overline{\Gamma}} \beta \right) \left\| \tilde{W} \right\|^2 + \alpha \|\tilde{W}\| \|\Sigma\| \|\Delta\| + \alpha \|\tilde{W}\| \|\tilde{\Sigma}\| \|\Delta\|. \tag{40}$$

*Remark* 1. Since $(x, u) \mapsto f(x, u)$, $(x, u) \mapsto \sigma(x, u)$, $u \mapsto \sigma_u(u)$, and $x \mapsto \sigma_V(x)$ are continuously differentiable, and since $t \mapsto u(t)$ is bounded, given a compact set $\hat{\chi} \subset \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$, there exist $L_{\sigma 1}, L_{F1}, L_{R1} > 0$ such that

$$\begin{aligned} \sup_{(x,u,\hat{u}) \in \hat{\chi}} \|\tilde{\sigma}(x, u, \hat{u})\| &\leq L_{\sigma 1} \|\tilde{u}\|, \\ \sup_{(x,u,\hat{u}) \in \hat{\chi}} \left\| \tilde{f}^o(x, u, \hat{u}) \right\| &\leq L_{F1} \|\tilde{u}\|, \\ \sup_{(x,u,\hat{u}) \in \hat{\chi}} \|\tilde{\sigma}_u(u, \hat{u})\| &\leq L_{R1} \|\tilde{u}\|. \end{aligned} \tag{41}$$

Using Remark 1, the term $\left\| \tilde{\Sigma} \right\|$ can be expressed in terms of $\tilde{u}$ and $\tilde{\theta}$ as

$$\left\| \tilde{\Sigma} \right\| \leq \left( \|\tilde{u}\| + \left\| \tilde{\theta} \right\| \right) \overline{\Sigma}, \tag{42}$$

where

$$\overline{\Sigma} := N \sup_{(x,u,\hat{u}) \in \hat{\chi}} \Big\{ \|\nabla_x \sigma_V(x)\| \Big( L_{F1} + L_{\sigma 1}(\|\tilde{u}\| + \|\tilde{\theta}\|) + L_{\sigma 1}\|\theta\| + \|\sigma(x, u)\| + \|\nabla_u \sigma(x, u)\| \Big), 2 + L_{R1} \Big\}. \tag{43}$$

The term $\|\Sigma\|$, which contains true values of the unknown states and parameters, is bounded above since it is a function of only true controls and parameters, $u$ and $\theta$, and queried states $x_i$. Let the upper bound on $\|\Sigma\|$ be denoted as

$$\|\Sigma\| \leq \overline{\Sigma}_\sigma, \tag{44}$$

where

$$\overline{\Sigma}_\sigma := N \sup_{\substack{x \in x(\cdot) \\ u \in u(\cdot)}} \Big\{ \|\nabla_x \sigma_V(x)\| \Big( \|f^o(x, u)\| + \|\theta\| \|\sigma(x, u)\| + \|\nabla_u f^o(x, u)\| + \|\theta\| \|\nabla_u \sigma(x, u)\| \Big), \left\| \sigma_u^-(u) \right\|, \left\| \nabla_u \sigma_u^-(u) \right\| \Big\}. \tag{45}$$

The error term $\|\Delta\|$ is bounded above by

$$\|\Delta\| \leq \left( \|\tilde{u}\| + \|\tilde{\theta}\| \right) \overline{\Delta} + \overline{\Delta}_\epsilon, \tag{46}$$

where

$$\overline{\Delta} := N \sup_{(x,u,\hat{u}) \in \hat{\chi}} \Big\{ L_{R1} \|W_R^*\| + 2\|R\| + \|[\nabla_u \sigma](x)\| \|[\nabla_u \sigma_V](x)\| \|W_V^*\| + \|\sigma(x, \hat{u})\| \|[\nabla_x \sigma_V](x)\| \|W_V^*\| + L_{F1} \|[\nabla_x \sigma_V](x)\| \|W_V^*\| + L_{\sigma 1} \|\theta^T\| \|[\nabla_x \sigma_V](x)\| \|W_V^*\| \Big\}, \tag{47}$$

and

$$\overline{\Delta}_\epsilon := N \sup_{(x,u,\hat{u}) \in \hat{\chi}} \Big\{ \|\epsilon_V(x)\| + \|\epsilon_Q(x)\| + \|\epsilon(x, u^*)\| \|[\nabla_x \sigma_V](x)\| \|W_V^*\| + \Big( \|[\nabla_u f^o](x)\| + \|\theta\| \|[\nabla_u \sigma](x)\| \Big) \|[\nabla_x \epsilon_V](x)\| + \|[\nabla_u \epsilon](x, u^*)\| \|[\nabla_x \sigma_V](x)\| \|W_V^*\| \Big\}. \tag{48}$$

Using (42), (44) and (46), $\dot{V}$ becomes

$$\dot{V}(\tilde{W},t) \leq -\frac{1}{2}\left(\alpha\underline{\sigma}+\frac{1}{\overline{\Gamma}}\beta\right)\left\|\tilde{W}\right\|^2+\alpha\overline{\Delta}_\epsilon\overline{\Sigma}_\sigma\left\|\tilde{W}\right\| + \alpha\overline{\Delta}_\epsilon\overline{\Sigma}\left\|\tilde{W}\right\|\left(\|\tilde{u}\|+\left\|\tilde{\theta}\right\|\right)+\alpha\overline{\Sigma}_\sigma\overline{\Delta}\left\|\tilde{W}\right\|\left(\|\tilde{u}\|+\left\|\tilde{\theta}\right\|\right) + \alpha\overline{\Sigma}\ \overline{\Delta}\left\|\tilde{W}\right\|\left(\|\tilde{u}\|+\left\|\tilde{\theta}\right\|\right)^2. \quad (49)$$

Using Young's Inequality $\dot{V}$ then becomes

$$\dot{V}(\tilde{W},t) \leq -\frac{1}{8}\left(\alpha\underline{\sigma}+\frac{1}{\overline{\Gamma}}\beta\right)\left\|\tilde{W}\right\|^2+\frac{2\alpha^2\overline{\Delta}_\epsilon^2\overline{\Sigma}_\sigma^2}{\alpha\underline{\sigma}+\beta/\overline{\Gamma}} + \frac{\alpha^2\overline{\mathcal{E}}^2\left(\overline{\Delta}_\epsilon\overline{\Sigma}+\overline{\Sigma}_\sigma\overline{\Delta}+\overline{\Sigma}\ \overline{\Delta}\ \overline{\mathcal{E}}\right)^2}{\alpha\underline{\sigma}+\beta/\overline{\Gamma}}, \quad (50)$$

where $\overline{\mathcal{E}}=\left(\overline{\tilde{u}}+\overline{\tilde{\theta}}\right)$. The notation, $\overline{\tilde{u}}$ and $\overline{\tilde{\theta}}$, denote bounded $\tilde{u}$ and $\tilde{\theta}$ values stored in the history stack, $\mathcal{H}^{IRL}$. Using the bound in (38), the differential inequality for $\dot{V}$ can be expressed as

$$\dot{V}(\tilde{W},t) \leq -CV\left(\tilde{W},t\right)+D, \quad (51)$$

where

$$C := \frac{\underline{\Gamma}}{4}\left(\alpha\underline{\sigma}+\frac{1}{\overline{\Gamma}}\beta\right), \quad (52)$$

$$D := \frac{\alpha^2\overline{\mathcal{E}}^2\left(\overline{\Delta}_\epsilon\overline{\Sigma}+\overline{\Sigma}_\sigma\overline{\Delta}+\overline{\Sigma}\ \overline{\Delta}\ \overline{\mathcal{E}}\right)^2}{\alpha\underline{\sigma}+\beta/\overline{\Gamma}}+\frac{2\alpha^2\overline{\Delta}_\epsilon^2\overline{\Sigma}_\sigma^2}{\alpha\underline{\sigma}+\beta/\overline{\Gamma}}. \quad (53)$$

Due to purging of $\mathcal{H}^{IRL}$, the estimator is analyzed over discrete time instances. Define the purging instances as $T_1, T_2, \ldots$, and maintain a minimum dwell time, $\mathcal{T}$, such that $T_{s+1}-T_s \geq \mathcal{T} > 0,\ \forall s \in \mathbb{N}$.

Solving equation (51) over any time interval $[T_s, T_{s+1})$, yields

$$\overline{V}_{s+1} \leq \overline{V}_s e^{-C(t-T_s)}+\frac{D_{s+1}}{C}, \quad (54)$$

where $\overline{V}_s \geq \left\|V\left(\tilde{W}\left(T_s\right),T_s\right)\right\|$ and $D_{s+1}$ denotes the value of $D$ over interval $[T_s, T_{s+1})$. Since we know that $\tilde{\theta}$ and $\tilde{u}$ decay exponentially to a bound, we know that $\overline{\mathcal{E}}$ is decreasing exponentially. Therefore, due to the decreasing error term $\overline{\mathcal{E}}$, it can be seen that

$$D_s > D_{s+1}, \forall s = 1, 2, \ldots \quad (55)$$

and

$$\overline{D} := \lim_{s\to\infty} D_s = \frac{2\alpha^2\overline{\Delta}_\epsilon^2\overline{\Sigma}_\sigma^2}{\alpha\underline{\sigma}+\beta/\overline{\Gamma}} + \frac{\alpha^2\overline{\mathcal{E}}_N^2\left(\overline{\Delta}_\epsilon\overline{\Sigma}+\overline{\Sigma}_\sigma\overline{\Delta}+\overline{\Sigma}\ \overline{\Delta}\ \overline{\mathcal{E}}_N\right)^2}{\alpha\underline{\sigma}+\beta/\overline{\Gamma}}, \quad (56)$$

where $\overline{\mathcal{E}}_N := \left(\sqrt{\overline{\Gamma}_u\frac{B}{A}}+\overline{\theta}_\infty\right)$, and $\overline{\theta}_\infty$ denotes the ultimate bound of the parameter estimation error $\tilde{\theta}$. Furthermore, the dwell time condition results in the bound

$$\overline{V}_{s+1} \leq \overline{V}_s e^{-C\mathcal{T}}+\frac{D_{s+1}}{C}, \forall s = 0, 1, 2, \ldots \quad (57)$$

If the bounds $D_{s+1}$ are selected so that

$$D_{s+1} > 2D_s e^{-C\mathcal{T}}, \forall s = 0, 1, 2, \ldots, \quad (58)$$

then

$$\overline{V}_{s+1} \leq \frac{2D_{s+1}}{C}, \forall s = 0, 1, 2, \ldots, \quad (59)$$

where $D_0 := \frac{C\overline{V}_0}{2}$. As a result, it can be concluded that

$$\lim_{s\to\infty}\sup\overline{V}_s \leq \frac{2\overline{D}}{C}, \quad (60)$$

and as a result $\lim_{s\to\infty}\sup\left\|\tilde{W}\left(T_s\right)\right\| \leq 2\sqrt{\overline{\Gamma}\frac{\overline{D}}{C}}$.

□

## VI. Simulation

To demonstrate the performance of the developed method, a linear optimal trajectory tracking problem, using the method developed in [22], [23], is utilized in order to have a known value function for comparison.

Consider an agent with the following linear dynamics

$$\dot{x} = \begin{bmatrix}0 & 1\\ \theta_1 & \theta_2\end{bmatrix}+\begin{bmatrix}0\\ \theta_3\end{bmatrix}u, \quad (61)$$

where the unknown parameters are $\theta_1 = -0.5, \theta_2 = -0.5$, and $\theta_3 = 1$. The parameter estimation technique utilized is developed in [19].

The trajectory the agent is attempting to follow is generated from the linear system

$$\dot{x}_d = \begin{bmatrix}0 & 1\\ -2 & 0\end{bmatrix}x_d. \quad (62)$$

The optimal control problem designed on the error dynamics is

$$J(e_0,\mu(\cdot)) = \int_{T_0}^{\infty} e(t)^T\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}e(t)+10\mu(t)^2\,\mathrm{d}t, \quad (63)$$

resulting in the ideal reward function weights $Q = \text{diag}([W_{Q_1},\ W_{Q_2}]) = \text{diag}([1,\ 1])$ and $R = 10$ where the error dynamics are

$$\dot{e} = \begin{bmatrix}0 & 1\\ -0.5 & -0.5\end{bmatrix}e+\begin{bmatrix}0\\ 1\end{bmatrix}\mu, \quad (64)$$

where $e = x - x_d$, $\mu = u - u_d$, and $u_d = [-1.5,\ 0.5]\ x_d$. The optimal value function to be estimated is

$$V^* = W_{V_1}e_1^2+W_{V_2}e_2^2+W_{V_3}e_1e_2, \quad (65)$$

where the ideal values are $W_{V_1} = 1.82, W_{V_2} = 2.30$, and $W_{V_3} = 1.83$. The optimal controller is $\mu = -[0.092,\ 0.230]e$.

Fig. 1 shows the tracking error and Fig. 2 shows the parameter estimation error. The parameters used for the two simulations are: $\beta = 0.5, \alpha = {}^{0.01}/_{50}, \beta_u = 2, \alpha_u = 1, M = 50, N = 50$ and a step size of $0.005s$.

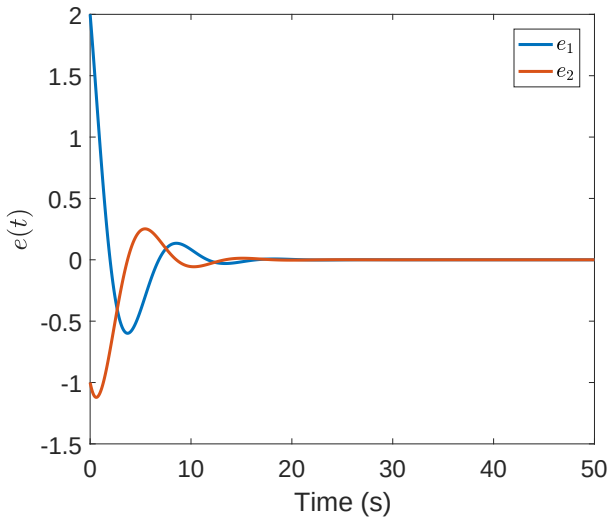


Fig. 1. Trajectory tracking error.

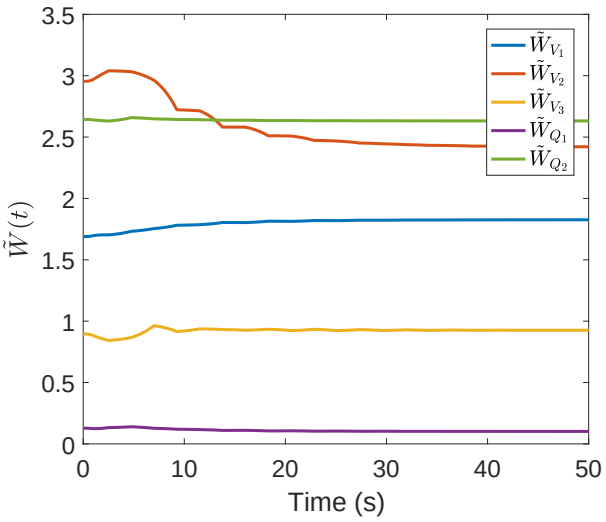


Fig. 3. Reward and value function estimation error without data querying.

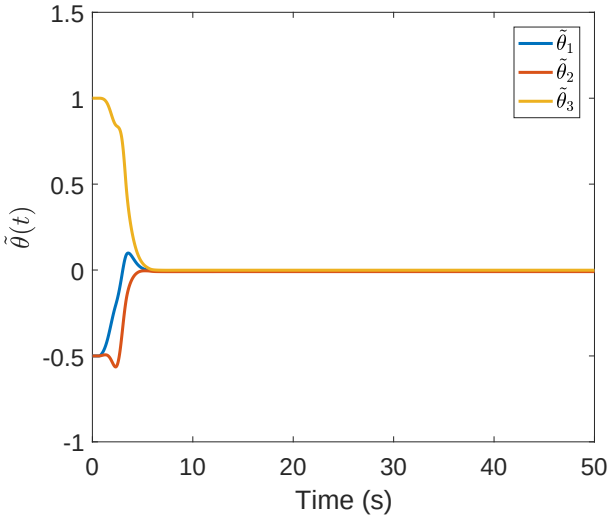


Fig. 2. Parameter estimation error.

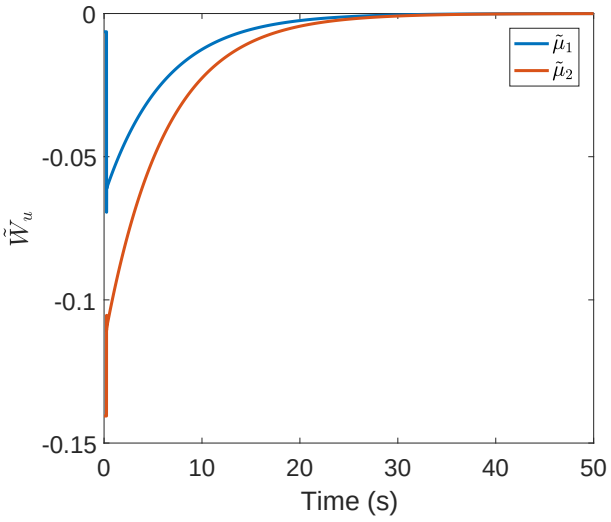


Fig. 4. Optimal feedback controller estimation error.

### A. IRL without Data Querying

The first simulation utilizes the state and control trajectories directly for IRL, and does not estimate the optimal controller for additional data. Fig. 3 shows reward and value function estimation errors without queried data.

As seen in Fig. 3, the reward and value function estimates do not converge to the ideal values. Looking closer, the estimates do not change much at all. The reason for this is once the history stacks are purged to remove poor parameter estimates, $\hat{\theta}$, the tracking errors, $e$, have decreased near the origin. Meaning, the data that IRL is utilizing, both $e$ and $\mu$, are at or near zero. This data does not provide sufficient information in order to accurately estimate the reward function.

### B. IRL Formulation with Data Querying

The second simulation shows the results of the novel control-estimation-based technique developed in this paper, with queried data points. Utilizing the estimate of the optimal controller, the estimate is queried with random states $x_i$ in the set $[-1, 1]$, which produce estimates of the optimal controller, $\hat{u}_i$. The pairs $(x_i, \hat{u}_i)$ are then iteratively collected in $\mathcal{H}^{IRL}$ and IRL is performed utilizing the update law in (32).

Fig. 4 shows the estimation error for the optimal feedback controller, and Fig. 5 shows the reward and value function estimation errors.

As seen in Fig. 5, the new IRL approach estimates the ideal values of the reward and value functions online. Though the tracking errors of the system dynamics have already converged, due to the non-zero queried state and control

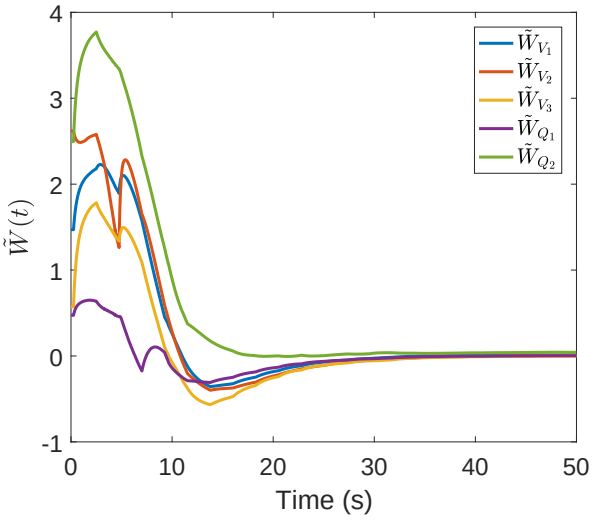


Fig. 5. Reward and value function estimation error with data querying.

values available through feedback estimation, IRL is able to converge.

## VII. Conclusion

This paper presents a new approach to performing reward function estimation online for situations with limited data. The approach utilizes a concurrent learning update law to estimate the optimal feedback controller of the agent online. This estimate is then utilized to artificially create additional data to promote reward function estimation. Theoretical guarantees are provided showing uniform ultimate boundedness of the unknown reward and value functions estimation errors using Lyapunov theory. A simulation example is performed that clearly shows the benefit of the method and how this additional queried data helps promote reward function estimation.

Future work will include analyzing the performance of this approach for systems with unmeasurable states and the affect of noise on optimal control estimation.

## 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 *Proceedings of the eleventh annual conference on Computational learning 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 *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.

[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 *Advances in Neural Information Processing Systems 20*, 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 *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.

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

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

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

[15] R. V. Self, M. Abudia, and R. Kamalapurkar, "Online inverse reinforcement learning for systems with disturbances," in *Proc. Am. Control Conf.*, Jul. 2020, to appear.

[16] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, pp. 359–366, 1985.

[17] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Netw.*, vol. 4, pp. 251–257, 1991.

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

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

[20] R. Kamalapurkar, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for approximate optimal regulation," *Automatica*, vol. 64, pp. 94–104, Feb. 2016.

[21] P. Ioannou and J. Sun, *Robust adaptive control*. Prentice Hall, 1996.

[22] R. Kamalapurkar, H. T. Dinh, S. Bhasin, and W. E. Dixon, "Approximate optimal trajectory tracking for continuous-time nonlinear systems," *Automatica*, vol. 51, pp. 40–48, Jan. 2015.

[23] R. Kamalapurkar, L. Andrews, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for infinite-horizon approximate optimal tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 753–758, Mar. 2017.