Linear Observer Learning by Temporal Difference

Stefano Menchetti, Mario Zanon and Alberto Bemporad

***Abstract*— This paper proposes a method for learning optimal state estimators from input/output data for linear discrete-time stochastic systems. We show that this problem can be expressed in the reinforcement learning framework, suitably adapted to the peculiar problem structure. In particular, we introduce the specific Bellman equation for the state estimation problem and use temporal differences to solve it. We show in simulations that the resulting data-driven method for state estimation converges to the optimal observer.**

I. Introduction

State estimation consists in the problem of estimating and predicting the internal states of a dynamical system based on measured, possibly noisy, output and input data.

From a probabilistic point of view, state estimation methods aim to determine the state estimates from their probability distribution obtained by the Bayes filter [1] using different criteria. For linear dynamical systems with additive Gaussian noise signals, the Kalman Filter (KF) algorithm is the best possible state estimator considering both the Minimum Mean Squared Error (MMSE) [2] and Maximum A Posteriori (MAP) [3] performance indices. For nonlinear dynamical systems, the Particle Filter (PF) [4], which approximates the Bayes filter by using Monte Carlo sampling methods, and the KF-derived algorithms, like the Extended Kalman Filter (EKF) [5] and the Unscented Kalman Filter (UKF) [6], are popular state estimators in practical applications.

The state estimation problem can also be described as the solution of an optimization problem [7], [8]. In this case, if the cost function is defined to include all past information, the algorithm is called a Full Information Estimator (FIE) [9]. The FIE algorithm represents a performance benchmark for the optimization-based estimators, but it becomes computationally intractable as the past information increases. Hence, the Moving Horizon Estimator (MHE) [10], [11], [12] was introduced, which approximates the FIE on a finite-horizon information. Moreover, it is worth noting that both KF and EKF can be formulated as optimization-based estimators [3], [13].

In this paper, our objective is to introduce a data-driven method based on ideas from Reinforcement Learning (RL) [14] to solve the optimal state estimation problem. RL is a set of techniques belonging to the machine learning field that are used to solve the optimal control problem by data sampling [15]. By the duality between the optimal control and optimal estimation problems, RL was proposed as a solution for the optimal state estimation problem in [16], [17]. The first approach proposes a Temporal Difference (TD) method with eligibility trace to compute the state estimates. However, it does not clearly detail how the value function approximation is defined and learned. In particular, the peculiarities of the observer learning problem are not investigated nor exploited. The second approach considers an actor-critic method to introduce a stable state estimator, but it assumes that the state of the system is known exactly during the training phase. In this paper, we assume that the actual state of the system is unknown and we use RL to learn from data a value-function based optimal observer. To that end, we first investigate how a value function can be defined for the optimal estimation problem. We show that this entails some important differences with respect to the control case and we propose an algorithm that can learn the optimal value function. While we focus on the linear case for simplicity, our contribution can be extended to a general setting.

The paper is organized as follows. Section II introduces the state estimation problem and its formulation as a Dynamic Programming (DP) problem. Section III describes how the proposed RL-based observer is built and implemented. Finally, Sections IV and V show, respectively, numerical results and concluding remarks.

II. Problem Statement

Consider the linear time-invariant (LTI) system

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{1a}$$
$$y_t = Cx_t + v_t, \tag{1b}$$

where $t \in \mathbb{N}$ denotes the current time, $x_t \in \mathbb{R}^{n_x}$, $u_t \in \mathbb{R}^{n_u}$, $y_t \in \mathbb{R}^{n_y}$, $w_t \in \mathbb{R}^{n_x}$, and $v_t \in \mathbb{R}^{n_y}$ denote, respectively, the state, input, measured output, process and measurement noise vectors at time t. Our objective is to solve the state estimation problem, i.e., we want to estimate the state x_t at every time t. To do so, we assume to know $(y_t,\ u_t)$ at every time t, as well as the model matrices $A \in \mathbb{R}^{n_x \times n_x}$, $B \in \mathbb{R}^{n_x \times n_u}$ and $C \in \mathbb{R}^{n_y \times n_x}$. Note that we assume no input/output feedthrough for simplicity, although the approach can be straightforwardly extended to cover this case. In the following, we will denote vector time-series as $x_{t_0:t_1} := (x_{t_0}, \ldots, x_{t_1})$.

Given the information $(y_{0:t},\ u_{0:t})$ available at time $t \geq 0$ and the initial state estimate $\bar{x}_0 \in \mathbb{R}^{n_x}$, we formulate the problem of computing the optimal predicted $\hat{x}_{t+1|t} \in \mathbb{R}^{n_x}$ and corrected $\hat{x}_{t|t} \in \mathbb{R}^{n_x}$ state estimates as the FIE problem

$$\hat{x}_{0:t+1|t} := \arg\min_{\chi_{0:t+1}} J_{t+1}(y_{0:t}, u_{0:t}, \bar{x}_0, \chi_{0:t+1}), \tag{2}$$

where $\hat{x}_{t+1|t}$ and $\hat{x}_{t|t}$ are the last two terms of the sequence of optimal state estimates $\hat{x}_{0:t+1|t}$, while $\chi_{0:T+1}$ is the

The authors are with the IMT School for Advanced Studies Lucca, Piazza San Francesco 19, 55100, Lucca, Italy (e-mail: {stefano.menchetti, mario.zanon, alberto.bemporad}@imtlucca.it)

sequence of decision variables with $\chi_t \in \mathbb{R}^{n_x}$ for every t. The cost is

$$J_{t+1}(y_{0:t}, u_{0:t}, \bar{x}_0, \chi_{0:t+1}) = \sum_{i=0}^{t} \gamma^{t-i} \ell(y_i, u_i, \chi_{i+1}, \chi_i) + \gamma^{t+1} \ell_0(\chi_0, \bar{x}_0),$$

with given forgetting factor $0 < \gamma \leq 1$. The stage and initial costs, respectively, ℓ and ℓ_0 are defined as

$$\ell(y_i, u_i, \chi_{i+1}, \chi_i) = \|\chi_{i+1} - A\chi_i - Bu_i\|^2_{Q^{-1}} + \|y_i - C\chi_i\|^2_{R^{-1}},$$

$$\ell_0(\bar{x}_0, \chi_0) = \|\chi_0 - \bar{x}_0\|^2_{P_0^{-1}},$$

with positive definite weighting matrices $Q^{-1} \in \mathbb{R}^{n_x \times n_x}$, $R^{-1} \in \mathbb{R}^{n_y \times n_y}$ and $P_0^{-1} \in \mathbb{R}^{n_x \times n_x}$.

Remark 1: In this paper we assume $\gamma < 1$ for simplicity, although all our results can be extended to the case $\gamma = 1$, with some slight adaptations that will be the subject of future publications. In particular, for $\gamma = 1$, problem (2) is equivalent to the MAP problem of choosing $\hat{x}_{0:t+1}$ that maximizes the conditional probability distribution $p(\chi_{0:t+1} | y_{0:t}, u_{0:t}, \bar{x}_0)$ [1]. The solution of the MAP problem can be proved to yield the well-known KF [3] and the matrices P_0, Q, and R are the covariances of the Gaussian initial state estimation error, the process noise and the output noise, respectively.

A. Dynamic Programming

In order to introduce our RL-based state estimator, it is necessary to formulate the state estimation problem (2) within a DP framework. For this purpose, let the optimal value function be

$$V_{t+1}(\hat{x}_{t+1|t}, \chi_{t+1}) = \|\chi_{t+1} - \hat{x}_{t+1|t}\|^2_{P_{t+1}^{-1}} + c_{t+1}, \quad (4)$$

with

$$\hat{x}_{t+1|t} := A\hat{x}_{t|t-1} + Bu_t + AK_t(y_t - C\hat{x}_{t|t-1}), \quad (5a)$$

$$P_{t+1} := Q + A(\mathbb{I}_{n_x} - K_t C)\frac{P_t}{\gamma}A^\top, \quad (5b)$$

$$S_t := R + C\frac{P_t}{\gamma}C^\top, \quad (5c)$$

$$K_t := \frac{P_t}{\gamma}C^\top S_t^{-1}, \quad (5d)$$

$$c_{t+1} := \|y_t - C\hat{x}_{t|t-1}\|^2_{S_t^{-1}} + \gamma c_t, \quad (5e)$$

where $\mathbb{I}_{n_x}$ is the identity matrix of dimension n_x.

Lemma 1: Let $\hat{x}_{0|-1} = \bar{x}_0$, $c_0 = 0$, and

$$V_0(\hat{x}_{0|-1}, \chi_0) = \ell_0(\hat{x}_{0|-1}, \chi_0) + c_0.$$

For all $t \geq 0$, the optimal value function (4) satisfies

$$V_{t+1}(\hat{x}_{t+1|t}, \chi_{t+1}) = \min_{\chi_{0:t}} J_{t+1}(y_{0:t}, u_{0:t}, \bar{x}_0, \chi_{0:t+1}). \quad (6)$$

Moreover, (5a) yields the last state estimate from the solution of (2).

Proof: We prove that (4) implies (6) by induction. Assuming that (6) holds at time t, we first prove that (6) also holds at time $t+1$. By using

$$J_{t+1}(y_{0:t}, u_{0:t}, \bar{x}_0, \chi_{0:t+1}) = \gamma J_t(y_{0:t-1}, u_{0:t-1}, \bar{x}_0, \chi_{0:t}) + \ell(y_t, u_t, \chi_{t+1}, \chi_t),$$

and (6) at time $t+1$, we have

$$\min_{\chi_{0:t}} \ell(y_t, u_t, \chi_{t+1}, \chi_t) + \gamma J_t(y_{0:t-1}, u_{0:t-1}, \bar{x}_0, \chi_{0:t}) = \min_{\chi_t} \ell(y_t, u_t, \chi_{t+1}, \chi_t) + \gamma V_t(\hat{x}_{t|t-1}, \chi_t), \quad (7)$$

The minimization in the right hand side of (7) is solved in [7] for the undiscounted case, i.e. for $\gamma = 1$. With the minor change of including the forgetting factor $0 < \gamma < 1$, we find

$$\min_{\chi_t} \ell(y_t, u_t, \chi_{t+1}, \chi_t) + \gamma V_t(\hat{x}_{t|t-1}, \chi_t) = \|\chi_{t+1} - \hat{x}_{t+1|t}\|^2_{P_{t+1}^{-1}} + c_{t+1},$$

with $\hat{x}_{t+1|t}$, P_{t+1}, K_{t+1}, c_{t+1} given by (5). Together with (7), this entails that

$$V_{t+1}(\hat{x}_{t+1|t}, \chi_{t+1}) = \min_{\chi_{0:t}} J_{t+1}(y_{0:t}, u_{0:t}, \bar{x}_0, \chi_{0:t+1}). \quad (8)$$

Note that $\hat{x}_{t+1|t}$ minimizes $V_{t+1}(\hat{x}_{t+1|t}, \chi_{t+1})$ and, consequently, also (2).

We are left with proving that (6) holds at time $t = 0$. This is directly obtained by observing that the right hand side of Equation (7) at time $t = 0$ reads

$$J_1(y_0, u_0, \bar{x}_0, \chi_{0:1}) = \ell(y_0, u_0, \chi_1, \chi_0) + \gamma V_0(\hat{x}_{0|-1}, \chi_0)$$

by construction. Then, equation (8) in turn yields

$$V_1(\hat{x}_{1|0}, \chi_1) = \min_{\chi_0} J_1(y_0, u_0, \bar{x}_0, \chi_{0:1}),$$

which is (6) at time $t = 0$. ■

Equation (6) justifies the definition of V as an optimal value function. Moreover, it can be used to show that the optimal value function is recursive, as it satisfies

$$V_{t+1}(\hat{x}_{t+1|t}, \chi_{t+1}) = \min_{\chi_t} \ell(y_t, u_t, \chi_{t+1}, \chi_t) + \gamma V_t(\hat{x}_{t|t-1}, \chi_t).$$

This recursion yields the optimal Bellman equation for state estimation and it formulates the Bellman optimality principle for estimation, i.e., function V_t summarizes all information available up to time t, similarly to the cost-to-go in optimal control, but in reverse time.

It is worth to notice that no expected value is involved in equation (6), since we take a deterministic point of view, by considering the actual realization of the stochastic variable y at every time instant.

Remark 2: The optimal value function V_t, or a suitable approximation, is commonly known in the literature as the arrival cost [7]. It is a fundamental concept for the optimization-based techniques for state estimation. In particular, MHE [10], [12] deeply relies on this concept.

While we proved that the optimal value function V is time varying, we discuss next that, over an infinite horizon and under additional assumptions, it becomes stationary in expectation. In order to prove this fact, we introduce the following common assumption.

Assumption 1: The variables w_t, v_t and x_0 are mutually independent and normally distributed, i.e,

$$\begin{aligned} &w_t \sim \mathcal{N}(0, Q), && Q \succ 0, && Q = Q^\top, \\ &v_t \sim \mathcal{N}(0, R), && R \succ 0, && R = R^\top, \\ &x_0 \sim \mathcal{N}(\bar{x}_0, P_0), && P_0 \succ 0, && P_0 = P_0^\top. \end{aligned}$$

Moreover, w and v are white noise signals. □

Lemma 2: Suppose that Assumption 1 holds, (A, C) is observable and (A, G) is controllable, for some G such that $GG^\top = Q$. Then

$$\lim_{t\to\infty} P_t = P_\star, \qquad P_\star \succ 0, \qquad \lim_{t\to\infty} \mathbb{E}_{\tilde{y}_{0:t}}[\, c_t \,] = c_\star,$$

where $\tilde{y}_t$ is the measurement error and $\mathbb{E}_{\tilde{y}_{0:t}}[\cdot]$ represents the expected value with respect to the probability distributions of the measurement errors from 0 to t. Furthermore, $\lim_{t\to\infty} S_t = S_\star$ and $\lim_{t\to\infty} K_t = K_\star$.

Proof: We observe from definition (5b) that the quantity P_t is computed through the discrete algebraic Riccati recursion

$$P_{t+1} = Q + A\frac{P_t}{\gamma}A^\top - A\frac{P_t}{\gamma}C^\top \left(R + C\frac{P_t}{\gamma}C^\top\right)^{-1} C\frac{P_t}{\gamma}A^\top,$$

that converges to a stationary value $P_\star \succ 0$ as $t \to \infty$, under the previous assumptions made that (A, C), and hence $(A/\sqrt{\gamma}, C/\sqrt{\gamma})$, is observable, (A, G) is controllable, and $Q, R, P_0 \succ 0$ (see [18] for the complete proof). The convergence to $P_\star$ implies from (5d) and (5c) that both K_t and S_t converge to stationary values. Finally, for the sake of brevity, the fact that the constant c_{t+1} converges to $c_\star$ in expectation, is shown in Appendix I. ■

Corollary 3: Suppose that Assumption 1 holds, (A, C) is observable and (A, G) controllable. Assume moreover that $P_0 = P_\star$. Then the value function is stationary and satisfies the Bellman equation

$$\begin{aligned} &V_\star(\hat{x}_{t+1|t}, \chi_{t+1}) \\ &\quad = \min_{\chi_t} \ \ell(y_t, u_t, \chi_{t+1}, \chi_t) + \gamma V_\star(\hat{x}_{t|t-1}, \chi_t) - \tilde{c}_t, \end{aligned} \tag{9}$$

where

$$\begin{aligned} &V_\star(\hat{x}_{t|t-1}, \chi_t) = \|\hat{x}_{t|t-1} - \chi_t\|^2_{P_\star^{-1}} + c_\star, \\ &\tilde{c}_t = \|y_t - C\hat{x}_{t|t-1}\|^2_{\left(R + C\frac{P_\star}{\gamma}C^\top\right)^{-1}} + (\gamma - 1)c_\star, \end{aligned} \tag{10}$$

such that $\mathbb{E}_{\tilde{y}_t}[\, \tilde{c}_t \,] = 0$.

Corollary 3 is of particular interest, since it enables the construction of a stationary observer based on ideas from RL. Furthermore, from the right hand side of equation (9) the state estimates $\hat{x}_{t+1|t}$ and $\hat{x}_{t|t}$ can be computed as

$$\begin{aligned} &(\hat{x}_{t+1|t}, \hat{x}_{t|t}) \\ &\quad = \arg \min_{\chi_{t+1}, \chi_t} \ell(y_t, u_t, \chi_{t+1}, \chi_t) + \gamma V_\star(\hat{x}_{t|t-1}, \chi_t), \end{aligned}$$

without the need to solve problem (2). This implicitly defines the stationary observer policy for the optimal predicted state estimate

$$\mathcal{U}_\star(y_t, u_t, \hat{x}_{t|t-1}) = \hat{x}_{t+1|t}. \tag{11}$$

Consequently, in the next section we will exploit equation (9) to learn policy $\mathcal{U}_\star$ indirectly, i.e., by learning an approximation $\hat{V}_W$ of $V_\star$, where W is a vector of parameters to determine.

III. Observer Learning

Corollary 3 proves that the stationary observer has a corresponding value function that satisfies the Bellman equation (9). This makes it possible to use RL methods to directly learn from data both the value function and the stationary observer itself. However, as we will discuss later in this section, RL algorithms cannot be directly applied to this problem and some modifications are required.

In order to compute the optimal state estimates without prior knowledge of the optimal value function, we introduce a new algorithm that is inspired by the least-squares temporal-difference (LSTD) algorithm [19] for solving the optimal control problem. For this reason, we will denote our method as *Least-Squares Temporal-Difference Observer* or LSTDO.

A. LSTDO Learning Problem Formulation

LSTDO aims at learning the optimal value function $V_\star$, which is equivalent to learning the optimal parameters $P_\star^{-1}$, $c_\star$ that we lump in a parameter vector $W_\star$. Similarly to least-squares methods in RL, by introducing the vector of parameters W and the corresponding approximate value function $\hat{V}_W$, the LSTDO solves for a given batch of data of length N the least-squares problem

$$\min_W \ \sum_{i=0}^{N} \left(V_\star(\hat{x}_+^{(i)}, \chi_+^{(i)}) - \hat{V}_W(\hat{x}_+^{(i)}, \chi_+^{(i)}) \right)^2, \tag{12}$$

which yields the optimality conditions

$$\sum_{i=0}^{N} \nabla_W \hat{V}_W(\hat{x}_+^{(i)}, \chi_+^{(i)}) \Big(V_\star(\hat{x}_+^{(i)}, \chi_+^{(i)}) - \hat{V}_W(\hat{x}_+^{(i)}, \chi_+^{(i)}) \Big) = 0, \tag{13}$$

where $\nabla_W \hat{V}_W$ denotes the gradient vector of $\hat{V}_W$ with respect to the vector of parameters W. Note that the least-squares residuals in (12) correspond to the errors between the (unknown) optimal value function and the approximate one. Focusing on the optimality conditions (13), we observe that, since the optimal value function $V_\star$ is not available, it must be replaced by an estimate. Typical choices in RL are given by Monte-Carlo (MC) or Temporal-Difference (TD) methods [14], but any technique can in principle be applied to solve the observer-learning problem. In this paper, we choose to focus on the simplest TD(0) method [20], where the TD(0) approximation of $V_\star$ is given by

$$\ell(y^{(i)}, u^{(i)}, \chi_+^{(i)}, \chi^{(i)}) + \gamma \hat{V}_W(\hat{x}^{(i)}, \chi^{(i)}). \tag{14}$$

Consequently, the single element $z^{(i)}$ of the batch has to be such that

$$\begin{aligned} z^{(i)} &= (y^{(i)}, u^{(i)}, \hat{x}_+^{(i)}, \hat{x}^{(i)}, \chi_+^{(i)}, \chi^{(i)}) \\ &:= (y_{t^{(i)}}, u_{t^{(i)}}, \hat{x}_{t^{(i)}+1|t^{(i)}}, \hat{x}_{t^{(i)}|t^{(i)}-1}, \chi_{t^{(i)}+1}, \chi_{t^{(i)}}). \end{aligned} \tag{15}$$

where the superscript (i) denotes sample i, collected at time $t^{(i)}$. Note that this notation allows us to use batches of data which are not necessarily obtained from the sequence $t^{(i+1)} = t^{(i)} + 1$.

B. Value Function Approximation

Since $V_\star$ is given by (10), it is a natural choice to select the approximation $\hat{V}_W$ as

$$\hat{V}_W(\hat{x}, \chi) = \|\hat{x} - \chi\|_H^2 + h \tag{16}$$

where the symmetric matrix $H \in \mathbb{R}^{n_x \times n_x}$ and the scalar $h \in \mathbb{R}$ are the parameters which, upon successful learning, should yield $H = P_\star^{-1}$ and $h = c_\star$, and which we lump in vector

$$W = \begin{bmatrix} W_H^\top & h \end{bmatrix}^\top,$$

where $W_H \in \mathbb{R}^{n_x(n_x+1)/2}$ is the vectorization of the symmetric matrix H. The approximation (16) is quadratic with respect to the estimation error $\hat{x} - x$ and linear with respect to the parameter vector W. In the following, we will write $\hat{V}_W$ in the equivalent form

$$\hat{V}_W(\hat{x}, \chi) = W^\top \phi(\hat{x}, \chi),$$

with $\phi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_x(n_x+1)/2+1}$ a suitably defined quadratic function, which we split for convenience of notation as follows:

$$\phi(\hat{x}, x) = \begin{bmatrix} \phi_H^\top(\hat{x}, \chi) & 1 \end{bmatrix}^\top,$$

such that

$$W_H^\top \phi_H(\hat{x}, \chi) = \|\hat{x} - \chi\|_H^2.$$

Note that, by defining $\hat{V}_W$ as linear with respect to the parameters, the TD(0) approximation in (14), yields the optimality condition (13) to compute the optimal vector of parameters W in a single iterate for a given batch.

C. Data Generation: Introducing Exploration

To discuss the algorithm for observer learning, we stress that the observer-learning problem has one particular specificity that distinguishes it from the standard RL problem, i.e., learning a *control* policy. While in the control problem the value function depends only on the state given by the system dynamics, for the observer problem the optimal value function depends not only on the optimal predicted state estimate $\hat{x}_+$ given by $\mathcal{U}_\star(y, u, \hat{x})$, but also on the decision variable χ_+ that is in principle not determined. The specificity outlined above is particularly relevant during the learning phase. As a matter of fact, unlike standard RL, in which it is sufficient to explore by applying off-policy actions, in observer learning one has not only to apply exploration to χ_+, by selecting it off-policy, i.e., as $\chi_+ \neq \mathcal{U}_\star(y, u, \hat{x})$, but also to define a coherent value of χ, such that the Bellman equation (9) holds. This can be understood by observing the TD-error

$$\begin{aligned} &\delta(y, u, \hat{x}_+, \hat{x}, \chi_+, \chi) \\ &\qquad = \ell(y, u, \chi, \chi_+) + \gamma \hat{V}_W(\hat{x}, \chi) - \hat{V}_W(\hat{x}_+, \chi_+), \end{aligned} \tag{17}$$

where the definition of δ entails that, whenever we explore, it is not sufficient to select the future decision variable χ_+, but we also need to select the current one χ. In particular, since our observer is based on optimizing the value function, χ needs to be consistent with χ_+, i.e., χ must also be optimal in the sense that it should minimize the cost, *given* χ_+. We formalize this concept in the optimization problem

$$\mathcal{L}_W(y, u, \hat{x}, \chi_+) = \arg\min_{\chi} \ell(y, u, \chi_+, \chi) + \gamma \hat{V}_W(\hat{x}, \chi), \tag{18}$$

where $\mathcal{L}_W(y, u, \hat{x}, x_+)$ is a smoothing policy which we will use to define our algorithm. Note that $\mathcal{L}_W$ is also used to introduce the observer policy $\mathcal{U}_W$ obtained as

$$\mathcal{U}_W(y, u, \hat{x}) = \arg\min_{\chi_+} \ \ell(y, u, \chi_+, \chi) + \gamma \hat{V}_W(\hat{x}, \chi) \tag{19a}$$

$$\text{s.t.} \ \ \chi = \mathcal{L}_W(y, u, \hat{x}, \chi_+). \tag{19b}$$

It is worth noticing that, upon successful learning, $\mathcal{U}_W$ and $\mathcal{L}_W$ converge to the optimal policies $\mathcal{U}_\star$ and $\mathcal{L}_\star$, where $\mathcal{U}_\star$ is the optimal stationary observer policy in (11) and $\mathcal{L}_\star$ is such that

$$\hat{x}_{t|t} = \mathcal{L}_\star(y_t, u_t, \hat{x}_{t|t-1}, \hat{x}_{t+1|t}),$$

i.e., it is the smoothing policy that allows one to compute the corrected state estimate $\hat{x}_{t|t}$, if the optimal stationary observer policy in (11) is applied.

Finally, we can turn to the definition of a procedure to generate the batch of data. Observing that at time $t^{(i)}$ the quantities $(y^{(i)}, u^{(i)}, \hat{x}^{(i)})$ are known, we need to generate the remaining elements of $z^{(i)}$ defined in (15), i.e., $(\hat{x}_+^{(i)}, \chi_+^{(i)}, \chi^{(i)})$. Using the current observer policy $\mathcal{U}_W$, we can immediately compute the state estimate $\hat{x}_+^{(i)} = \mathcal{U}_W(y^{(i)}, u^{(i)}, \hat{x}^{(i)})$. Afterwards, the predicted decision variable $\chi_+^{(i)}$ is chosen in a neighbourhood of $\hat{x}_+^{(i)}$: though alternatives are possible, we adopt the ϵ-greedy strategy for the sake of simplicity, which yields

$$\chi_+^{(i)} = \hat{x}_+^{(i)} + \xi^{(i)}, \tag{20}$$

where $\xi \sim \mathcal{N}(0, \Xi)$ is a given random vector, with zero mean and $\Xi \succ 0 \in \mathbb{R}^{n_x \times n_x}$. Finally, a consistent current decision variable $\chi^{(i)}$ is obtained as $\chi^{(i)} = \mathcal{L}_W(y^{(i)}, u^{(i)}, \hat{x}^{(i)}, \chi_+^{(i)})$.

D. LSTDO Algorithm

Denoting as N_b the number of batches, Algorithm 1 provides the pseudocode for LSTDO.

IV. Numerical Results

We consider the linear discrete-time dynamical system associated with the three-dimensional Euclidean space described by the axes (μ, η, ζ), with state

$$x_t = \begin{bmatrix} \mu_t & v_t^\mu & \eta_t & v_t^\eta & \zeta_t \end{bmatrix}^\top,$$

```
Algorithm 1: LSTDO Method
Given $\gamma$, $N$, $N_b$, $u_{0:N}$, $\Xi$;
Initialize $W$, $\hat{x}_{0|-1} = \bar{x}_0$;
for $j = 1, 2, \ldots, N_b$ do
    Simulate the real system to have $y_{0:N}$;
    Generate $\hat{x}^{(0:N+1)}$ using $\mathcal{U}_W$;
    Explore as in (20) to define $\chi_+^{(0:N+1)}$;
    for $t^i \in \{0, 1, 2, \ldots, N\}$ do
        Define $\chi^{(i)} = \mathcal{L}_W(y^{(i)}, u^{(i)}, \hat{x}^{(i)}, \chi_+^{(i)})$;
        Compute $\delta^{(i)}$ as in (17);
        Compute $\nabla_W \hat{V}_W(\hat{x}_+^{(i)}, \chi_+^{(i)}) = \phi(\hat{x}_+^{(i)}, \chi_+^{(i)})$;
        Store $\delta^{(i)}$, $\nabla_W \hat{V}_W(\hat{x}_+^{(i)}, \chi_+^{(i)})$
    end
    Solve (13) to update $W$;
    Compute $\mathcal{L}_W$,$\mathcal{U}_W$ as in (18) and (19)
end
```

	batch 0	batch 15	batch 30	batch 49
$n_c(\varepsilon = 10^3)$	0	90	93	100
$n_c(\varepsilon = 10^4)$	0	93	97	100

TABLE I: Number of simulations at convergence

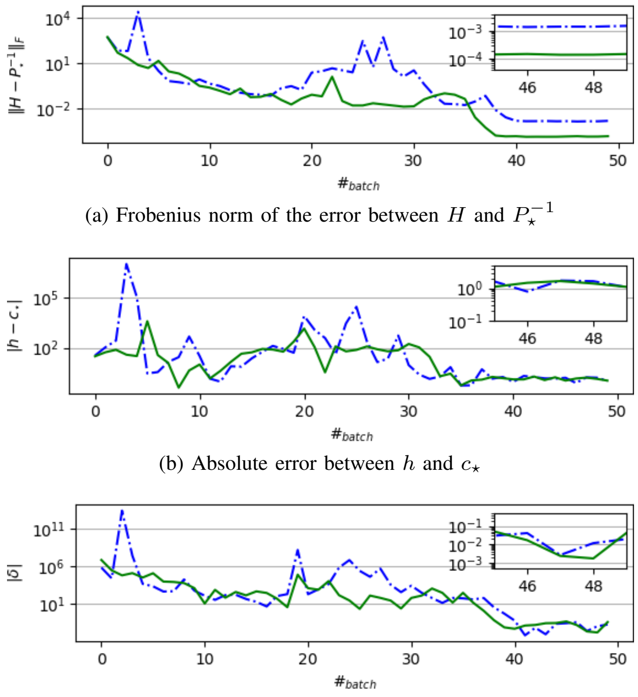
where v_t^μ and v_t^η are the velocities with respect to the axes μ and η. The state-space representation of the dynamical system as in (1) is given by

$$A = \begin{bmatrix} 1 & t_s & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & t_s & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

where $t_s = 0.1$ is the sampling time. We consider an autonomous system, since the presence of a known input does not affect the learning algorithm, hence we assume $B = 0$. This dynamical system describes the discretization of the motion of an aircraft with constant velocity coordinates on a plane described by the coordinates (μ, η) at a constant altitude ζ. The measured output are given by the exact position of the aircraft. The process and the measurement noises are Gaussian white signals with covariance matrices Q and R such that

$$K = \begin{bmatrix} \frac{t_s^3}{3} & \frac{t_s^2}{2} \\ \frac{t_s^2}{2} & t_s \end{bmatrix}, \; Q = \begin{bmatrix} \sigma_\mu^2 K & 0 & 0 \\ 0 & \sigma_\eta^2 K & 0 \\ 0 & 0 & \sigma_\zeta^2 t_s \end{bmatrix}, \; R = 10 \; \mathbb{I}_{n_y}$$

where $\sigma_\mu = \sigma_\eta = \sigma_\zeta = 5$. To apply the LSTDO method we assume $N = 100$, $N_b = 50$, $\gamma = 0.9$. We initialize 100 different simulations with initial H and h such that $H \sim \mathcal{N}(0, 100\mathbb{I}_{n_x})$ and $h \sim \mathcal{N}(0, 100)$. The tuning of LSTDO is done by choosing the amount of exploration, i.e., the parameter ε such that $\xi^{(i)}$ in (20) is sampled according to $\mathcal{N}(0, \varepsilon \; \mathbb{I}_{n_x})$. In our simulation, we use the values $\varepsilon = 10^3$ and $\varepsilon = 10^4$ to tune the LSTDO. In Figure 1 we display for every batch the average value over the different simulations of the Frobenius norm of error $H - P_\star^{-1}$, the absolute value of the error $h - c_\star$ and the absolute value of δ on a logarithmic scale. Table I shows, instead, the number of converged simulations n_c at specific batches. Both Figure 1 and Table I suggest that our method can be used successfully to learn the parameters of the stationary optimal value function $V_\star$. Moreover, the LSTDO method shows fast convergence, since most of the simulations converge in few iterations. Finally, we observe that the amount of exploration affects considerably the performance of the LSTDO and, in particular, that the choice of a larger value for ε is related to quicker convergence and smaller errors on the estimated parameters.



(a) Frobenius norm of the error between H and $P_\star^{-1}$

(b) Absolute error between h and $c_\star$

(c) Absolute value of δ

Fig. 1: Linear case: blue line for $\varepsilon = 10^3$ and green line for $\varepsilon = 10^4$

V. Conclusions

In this paper, we discussed how learning techniques can be also applied to learn optimal observers for linear systems from data. We have highlighted an important difference with learning applied to control, such that we had to introduce some modifications to common RL algorithms. We have demonstrated in simulations that the optimal observer is correctly learned by our algorithm.

Future work will consider extending our framework to cover nonlinear systems and to adapt a wider class of RL algorithms to observer learning.

References

[1] S. Särkkä, *Bayesian Filtering and Smoothing*. Cambridge: Cambridge University Press, 2013.

[2] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME, Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[3] H. Cox, "On the estimation of state variables and parameters for noisy dynamic systems," *IEEE Transaction on Automatic Control*, vol. 9, no. 1, pp. 5–12, 1964.

[4] A. Doucet, S. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and computing*, vol. 10, no. 3, pp. 197–208, 2000.

[5] A. H. Jazwinski, *Stochastic processes and filtering theory*. Courier Corporation, 2007.

[6] S. J. Julier and J. K. Uhlmann, "New extension of the kalman filter to nonlinear systems," in *Signal processing, sensor fusion, and target recognition VI*, vol. 3068. International Society for Optics and Photonics, 1997, pp. 182–193.

[7] J. B. Rawlings, D. Q. Mayne, and M. M. Diehl, *Model Predictive Control: Theory, Computation and Design*. Santa Barbara, California: Nob Hill Publishing, 2019.

[8] D. G. Robertson and J. H. Lee, "A least squares formulation for state estimation," *Journal of process control*, vol. 5, no. 4, pp. 291–299, 1995.

[9] J. B. Rawlings and L. Ji, "Optimization-based state estimation: Current status and some new results," *Journal of Process Control*, vol. 22, no. 8, pp. 1439–1444, 2012.

[10] C. V. Rao, "Moving horizon strategies for the constrained monitoring and control of nonlinear discrete-time systems," Ph.D. dissertation, University of Wisconsin-Madison, 2000.

[11] C. V. Rao, J. B. Rawlings, and J. H. Lee, "Constrained linear state estimation—a moving horizon approach," *Automatica*, vol. 37, no. 10, pp. 1619–1628, 2001.

[12] C. V. Rao, J. B. Rawlings, and D. Q. Mayne, "Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations," *IEEE Transaction on Automatic Control*, vol. 48, no. 2, pp. 246–258, 2003.

[13] J. Humpherys, P. Redd, and J. West, "A fresh look at the kalman filter," *SIAM Review*, vol. 54, no. 4, pp. 801–823, 2012.

[14] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[15] D. Bertsekas, *Reinforcement learning and optimal control*. Athena Scientific, 2019.

[16] J. Morimoto and K. Doya, "Reinforcement learning state estimator," *Neural computation*, vol. 19, no. 3, pp. 730–756, 2007.

[17] L. Hu, C. Wu, and W. Pan, "Lyapunov-based reinforcement learning state estimator," *arXiv preprint arXiv:2010.13529*, 2020.

[18] F. L. Lewis, L. Xie, and D. Popa, *Optimal and robust estimation: with an introduction to stochastic control theory*. CRC press, 2017, ch. 2.

[19] S. J. Bradtke and A. G. Barto, "Linear least-squares algorithms for temporal difference learning," *Machine learning*, vol. 22, no. 1, pp. 33–57, 1996.

[20] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, no. 1, pp. 9–44, 1988.

APPENDIX I
CONVERGENCE OF c_{t+1}

In order to prove that c_{t+1} converges to $c_\star$ in expectation, we initially define the estimation error $\tilde{x}_t := x_t - \hat{x}_{t|t-1}$ and the measurement error $\tilde{y}_t := y_t - C\hat{x}_{t|t-1}$ at time t. Exploiting the dynamical system (1) and the evolution in time of the optimal predicted state estimates (5a), we compute their dynamics

$$\tilde{x}_{t+1} = A(\mathbb{I}_{n_x} - K_t C)\tilde{x}_t + \omega_t - AK_t v_t, \tag{21a}$$

$$\tilde{y}_t = C\tilde{x}_t + v_t. \tag{21b}$$

We observe that from Assumption 1 defining $\hat{x}_{0|-1} = \bar{x}_0$ the estimation error $\tilde{x}_0$ is distributed according to $\mathcal{N}(0, \tilde{P}_0)$, with $\tilde{P}_0 = P_0$. Consequently, it is possible to recursively prove that at every time instant t both $\tilde{x}_t$ and $\tilde{y}_t$ are normally distributed, i.e.,

$$\tilde{x}_{t+1} \sim \mathcal{N}(m_{\tilde{x},t+1}, \tilde{P}_{t+1}), \quad \tilde{y}_t \sim \mathcal{N}(m_{\tilde{y},t}, \tilde{S}_t),$$

since both of them can be described as the sum of normally distributed independent random variables exploiting equations (21a) and (21b) and Assumption 1. Moreover, we compute

$$\begin{aligned} m_{\tilde{x},t+1} &= \mathbb{E}_{\tilde{x}_{t+1}}[\tilde{x}_{t+1}] \\ &= A(\mathbb{I}_{n_x} - K_t C)\mathbb{E}_{\tilde{x}_t}[\tilde{x}_t] + \mathbb{E}_{\omega_t}[\omega_t] - AK_t\mathbb{E}_{v_t}[v_t] \\ &= 0 \\ \tilde{P}_{t+1} &= \mathrm{Var}_{\tilde{x}_{t+1}}[\tilde{x}_{t+1}] \\ &= Q + AK_tRK_t^\top A^\top \\ &\qquad + A(\mathbb{I}_{n_x} - K_t C)\tilde{P}_t(\mathbb{I}_{n_x} - K_t C)^\top A^\top, \end{aligned}$$

and

$$\begin{aligned} m_{\tilde{y},t} &= \mathbb{E}_{\tilde{y}_t}[\tilde{y}_t] = C\mathbb{E}_{\tilde{x}_t}[\tilde{x}_t] + \mathbb{E}_{v_t}[v_t] = 0, \\ \tilde{S}_t &= \mathrm{Var}_{\tilde{y}_t}[\tilde{y}_t] = R + C\tilde{P}_t C^\top. \end{aligned}$$

Knowing from the first part of Lemma 2 that as $t \to \infty$ we have $P_{t+1} = P_\star$, $K_t = K_\star$, $S_t = S_\star$, the discrete Lyapunov equation

$$\begin{aligned} A(\mathbb{I}_{n_x} - K_\star C)\tilde{P}(\mathbb{I}_{n_x} - K_\star C)^\top A^\top - \tilde{P} \\ + \left(Q + AK_\star RK_\star^\top A^\top\right) = 0 \end{aligned}$$

admits a unique positive definite solution $\tilde{P}_\star \succ 0$, such that $\lim_{t\to\infty} \tilde{P}_t = \tilde{P}_\star$, since $Q + AK_\star RK_\star^\top A^\top \succ 0$ and $A(\mathbb{I}_{n_x} - K_\star C)$ is asymptotically stable (see [18]). In turn, this implies that $\tilde{S}_\star = R + C\tilde{P}_\star C^\top \succ 0$.

Finally, we initialize $c_0 = 0$ and exploit (5e) and the definition of $\tilde{y}$ to compute c_{t+1} as

$$c_{t+1} = \sum_{i=0}^{t} \gamma^{t-i} \|\tilde{y}_i\|^2_{S_i^{-1}}.$$

The following relations

$$\lim_{t\to\infty} \sum_{i=0}^{t} \gamma^{t-i} = \frac{1}{1-\gamma}$$

and, for $x \sim \mathcal{N}(m, \Sigma)$,

$$\mathbb{E}_x[x^\top Ax] = \mathit{tr}(A\Sigma) + m^\top Am,$$

hold and we use them to prove that c_{t+1} converge to a costant value in expectation, since at stationarity we observe that

$$\begin{aligned} \lim_{t\to\infty} \mathbb{E}_{\tilde{y}_{0:t}}[\, c_{t+1}\,] &= \lim_{t\to\infty} \sum_{i=0}^{t} \gamma^{t-i} \mathbb{E}_{\tilde{y}_i}\left[\|\tilde{y}_i\|^2_{S_i^{-1}}\right] \\ &= \lim_{t\to\infty} \sum_{i=0}^{t} \gamma^{t-i}\, \mathit{tr}\left(S_i^{-1}\tilde{S}_i\right) \\ &= \frac{1}{1-\gamma}\, \mathit{tr}\left(S_\star^{-1}\tilde{S}_\star\right) \end{aligned}$$

where *tr* is the trace operator. We conclude our proof by defining

$$c_\star = \frac{1}{1-\gamma}\, \mathit{tr}\left(S_\star^{-1}\tilde{S}_\star\right).$$