2021 60th IEEE Conference on Decision and Control (CDC)
December 13-15, 2021. Austin, Texas

Exploiting the experts: Learning to control unknown SISO feedback linearizable systems from expert demonstrations*

Alimzhan Sultangazin[1], Lucas Fraile[1], and Paulo Tabuada[1]

2021 60th IEEE Conference on Decision and Control (CDC) | 978-1-6654-3659-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDC45484.2021.9683193

***Abstract*— It was shown, in recent work by the authors, that it is possible to learn an asymptotically stabilizing controller from a small number of demonstrations performed by an expert on a feedback linearizable system. These results rely on knowledge of the plant dynamics to assemble the learned controller from the demonstrations. In this paper we show how to leverage recent results on data-driven control to dispense with the need to use the plant model. By bringing these two methodologies — learning from demonstrations and data-driven control — together, this paper provides a technique that enables the control of unknown nonlinear feedback linearizable systems solely based on a small number of expert demonstrations.**

I. Introduction

A. Motivation and related work

The merits of learning from expert demonstrations have been well-reasoned in the literature (see, e.g., [1], [2]). There are many control tasks for which demonstrating the desired behaviour is far simpler than defining such behaviour mathematically (e.g., driving a car in a way that is comfortable to passengers, teaching a robot to manipulate objects or play sports). The ever-growing body of work in learning from demonstrations (LfD) for robot control [2] reflects the need for a well-defined and versatile framework for such tasks. In this work we propose a framework that uses demonstrations of an expert controlling an unknown dynamical system to construct a stabilizing controller for that system.

Our framework relies on combining two methodologies: the LfD methodology described in [3] and the data-driven control methodology described in [4]. The former allows us to construct a controller for a feedback linearizable system from a finite number of expert demonstrations of desired behaviour, provided a small set of sufficiently long demonstrations is collected, whereas the latter provides a method for stabilizing unknown feedback linearizable single-input single-output (SISO) systems requiring only standard linear control techniques and sufficiently fast sampling rates. To set the scene, we briefly discuss the previous work related to both learning from demonstrations component and data-driven control component of this work. This is in no way a comprehensive account of the literature on LfD (see, e.g., the surveys in [2] or [5] for a roadmap of the literature on LfD), but an overview of the approaches most related to ours.

Most LfD methods use the expert demonstrations either to learn the cost function or the policy [2]. The methods learning the former assume that the expert optimizes an unknown cost function (e.g., inverse reinforcement learning methods from [6], [7]). The methodology from [3] used in this work belongs to a class of *policy-learning LfD methods* whose key assumption is that there is a well-defined mapping, called the expert's policy, from states (or observations) to control inputs that determines the expert's response. These methods aim to approximate the expert's policy as best as possible based on a set of expert demonstrations. In the context of machine learning, policy learning is viewed as a supervised-learning problem where states and control inputs are treated as features and labels, respectively (e.g., [8], [9]). These methods, however, require a large number of demonstrations to work well and, if trained solely on data from unmodified expert demonstrations, generate unstable policies that cannot recover from drifts or disturbances [9]. In addition, such methods rarely provide formal stability guarantees and mainly illustrate their performance with experimental results.

To address these and other issues with policy-learning LfD methods, many works use concepts and existing techniques from control theory. In [10], a linear quadratic Gaussian controller is used to show that learning an expert policy is dual to system identification. In fact, there are numerous works where expert demonstrations are assumed to come from an underlying stable dynamical system, and the objective is to learn this dynamical system (see, e.g., [11]–[13]). The work that is conceptually closest to [3], however, is the work by Palan et al. [14] wherein the authors fit a linear policy to expert demonstrations stabilizing a linear system under a set of constraints from [15]. These constraints guarantee that the learned policy also stabilizes the linear system. The methodology of [3] is different from that in [14] in that there is no assumption of linearity of the expert's policy.

The data-driven control methodology from [4] relies on the feedback linearizability assumption and the observation that, by sampling fast enough, the relevant signals can be considered constant in between sampling instants. The latter observation is largely inspired by the work of Fliess and Join in [16], [17]. By leveraging the control techniques developed in [18], [19] for the control of approximate discrete-time models, the work in [4] proposes a technique to asymptotically stabilize unknown feedback linearizable systems. Unlike the data-driven control techniques inspired by behavioural theory (e.g., [20], [21]), no prior data or persistency of excitation is required to use the data-driven control methodology from [4].

The first two authors contributed equally.

*This work was supported in part by the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

[1]A. Sultangazin, L. Fraile, and P. Tabuada are with Department of Electrical and Computer Engineering, University of California - Los Angeles, USA {asultangazin, lfrailev, tabuada}@ucla.edu

B. Contributions

The results of [3] rely on the assumption that we have complete knowledge of the system dynamics. In this work, we use the data-driven control methodology described in [4] to relax this assumption, and propose a method by which we can construct a controller for an unknown feedback linearizable SISO system from a finite number of expert demonstrations of desired behaviour. The number of demonstrations required to construct this controller is $n+1$, where n is the number of states of the system being controlled. Provided the demonstrations are sufficiently long, we can formally prove the learned controller asymptotically stabilizes the system[1].

C. Notations and basic definitions

The notation used in this paper is fairly standard. The integers are denoted by $\mathbb{Z}$, the natural numbers with zero by $\mathbb{N}_0$, the real numbers by $\mathbb{R}$, the non-negative real numbers by $\mathbb{R}_0^+$, and the positive real numbers by $\mathbb{R}^+$. We denote by $\|\cdot\|$ (or by $\|\cdot\|_2$) the Euclidean norm or the induced matrix 2-norm; and by $\|\cdot\|_F$ the matrix Frobenius norm.

Let $\mathcal{X} = \{x_1, \ldots, x_k\}$ be a finite set of points in $\mathbb{R}^n$. A point $x = \sum_{i=1}^k \theta_i x_i$ with $\sum_{i=1}^k \theta_i = 1$ is called *an affine combination* of points in $\mathcal{X}$. A set $\{x_1, \ldots, x_k\}$ in $\mathbb{R}^n$ is *affinely independent* if the set $\{x_2 - x_1, \ldots, x_k - x_1\}$ is linearly independent.

A function $\alpha : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ is of class $\mathcal{K}$ if α is continuous, strictly increasing, and $\alpha(0) = 0$. If α is also unbounded, it is of class $\mathcal{K}_\infty$. A function $\beta : \mathbb{R}_0^+ \times \mathbb{R}_0^+ \to \mathbb{R}_0^+$ is of class $\mathcal{KL}$ if, for fixed $t \geq 0$, $\beta(\cdot, t)$ is of class $\mathcal{K}$ and $\beta(r, \cdot)$ decreases to 0 as $t \to \infty$ for each fixed $r \geq 0$.

The Lie derivative of a function $h : \mathbb{R}^n \to \mathbb{R}$ along a vector field $f : \mathbb{R}^n \to \mathbb{R}^n$, given by $\frac{\partial h}{\partial x} f$, is denoted by $L_f h$. We use the notation $L_f^k h$ for the iterated Lie derivative, i.e., $L_f^k h = L_f(L_f^{k-1} h)$, with $L_f^0 h = h$.

Consider a continuous-time dynamical system of the form:

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

where $x \in \mathbb{R}^n$ is the state and $f : \mathbb{R}_0^+ \times \mathbb{R}^n \to \mathbb{R}^n$ is a smooth function. The origin of (1) is *uniformly asymptotically stable* if there is $\beta \in \mathcal{KL}$ such that:

$$\|x(t)\| \leq \beta(\|x(t_0)\|, t - t_0), \quad \forall t \geq t_0 \geq 0. \tag{2}$$

If β in (2) has the form $\beta(r, t) = Mre^{-\lambda t}$, then the origin of (1) is *uniformly exponentially stable*.

Let $(\mathcal{V}, \|\cdot\|)$ be a normed vector space. Consider a function $f : \mathbb{R}_0^+ \times \mathcal{Q} \to \mathcal{V}$ with $\mathcal{Q} \subset \mathcal{V}$. The notation $f(t, x) = O_x(T)$ denotes existence of constants $M, T \in \mathbb{R}^+$ such that, for all $t \in [0, T]$ and $x \in \mathcal{Q}$, we have $\|f(t, x)\| \leq MT\|x\|$. If we take $T \leq 1$, the following rules apply to this notation:

$$\begin{aligned} O_x(T^2) &= O_x(T) & (O_x(t))^2 &= O_{x^2}(T^2) \\ TO_x(T) &= O_x(T^2) & g(x)O_x(T) &= O_x(T), \end{aligned} \tag{3}$$

[1]The reader is referred to the technical report [22] for the proofs of some results presented here.

for all functions g with a bounded norm, i.e., there is a $b \in \mathbb{R}^+$ such that $\|g(x)\| \leq b$ for all $x \in \mathcal{Q}$. The subscript x^2 in $O_{x^2}(T^2)$ indicates that its upper bound is $MT^2\|x\|^2$.

Let $x : \mathbb{R}_0^+ \to \mathcal{V}$ be a continuous-time signal. We denote the corresponding sampled-data signal by $x_s : \mathbb{N}_0 \to \mathcal{V}$ and define it by $x_s(k) \triangleq x(kT)$.

II. Problem Statement and Preliminaries

A. Problem

Consider an unknown single-input single-output nonlinear system described by:

$$\dot{x} = f(x) + g(x)u, \quad y = h(x), \tag{4}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^n$, and $h : \mathbb{R}^n \to \mathbb{R}$ are smooth functions and we denote by $x \in \mathbb{R}^n$, $u \in \mathbb{R}$, $y \in \mathbb{R}$ the state, the input, and the output of the system, respectively. Assume that the origin is an equilibrium point of (4). We call a triple $(x, u, y) : \mathbb{R}_0^+ \to \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}$ a solution of the system (4) if, for all $t \in \mathbb{R}_0^+$, the equation (4) is satisfied. Furthermore, we refer to the functions x, u, and y as a trajectory, a control input, and an output of the system (4), respectively. Given a sampling period $T > 0$, we assume that the control u is constant over sampling intervals and that the output y is measured at sampling instants kT. We denote the values of the trajectory, the control input, and the output at the sampling instants by $x_s(k) \triangleq x(kT)$, $u_s(k) \triangleq u(kT)$, and $y_s(k) = y(kT)$, respectively.

Definition II.1. A controller $u = \kappa(x)$ is *asymptotically stabilizing* for system (4) if the origin is asymptotically stable for the system (4) with $u = \kappa(x)$.

Suppose there is an unknown asymptotically stabilizing controller $u = \kappa(x)$ for (4), which we refer to as *the expert*. We assume that $\kappa : \mathbb{R}^n \to \mathbb{R}$ is smooth. The expert κ controls (4) and produces $N = n + 1$ finite-length closed-loop solutions of length $\tau_s \in \mathbb{R}$. We denote the set of the expert solutions by $\mathcal{D} = \{(x^i, u^i, y^i)\}_{i=1}^{n+1}$, where, for each i, the trajectory $x^i : [0, \tau_s] \to \mathbb{R}^n$, the output $y^i : [0, \tau_s] \to \mathbb{R}^n$ are continuous, and the control input $u^i : [0, \tau_s] \to \mathbb{R}^m$ is piecewise constant. Each (x^i, u^i, y^i) satisfies $u^i(t) = \kappa(x_s^i(k))$ for all $t \in [kT, (k+1)T)$, $k \in \{0, \ldots, L_s - 2\}$ and $y^i(t) = h(x^i(t))$ for all $t \in [0, \tau_s]$. We also ensure that the "trivial" expert solution, wherein $x(t) = 0$ and $u(t) = 0$ for all $t \in [0, \tau]$, is included in $\mathcal{D}$.

While the expert presents their solutions, we collect the set of measurement samples $\mathcal{D}_s = \{y_s^i\}_{i=1}^{n+1}$ corresponding to the expert solutions in $\mathcal{D}$, where $y_s^i : \{0, 1, \ldots, L_s - 1\} \to \mathbb{R}$ and the length of the measurement sequence $L_s \in \mathbb{N}$ satisfies $(L_s - 1)T = \tau_s$. We want to use $\mathcal{D}_s$ to create a controller for (4) that is guaranteed to asymptotically stabilize the origin.

Remark II.2. The theoretical results to follow apply for any $N \geq n+1$. Having been provided more expert solutions than $n+1$ raises an important question: how can we leverage this extra data to construct the controller that best approximates the expert? An interested reader can refer to Section IV in [3] for a possible answer to this question.

B. Feedback linearization

Assume that the system (4) is feedback linearizable with the output map h having a relative degree n, that is:

$$\begin{aligned} L_g L_f^i h(x) &= 0, \quad \forall i = 0, \dots, n-2, \\ L_g L_f^{n-1} h(x) &\neq 0, \end{aligned}$$

for all $x \in \mathbb{R}^n$. Since $L_g L_f^{n-1} h(x)$ is continuous and is never zero on $\mathbb{R}^n$, it has a constant sign on $\mathbb{R}^n$. We assume that the sign of $L_g L_f^{n-1} h$ is known and, without loss of generality, is taken to be positive.

With the goal of clear exposition in mind, we assume throughout the paper that $n = 2$ — all the results, however, hold for any $n \in \mathbb{N}$. Applying feedback linearization, we can rewrite the unknown dynamics in the coordinates $z = \Phi(x) \triangleq (h(x), L_f h(x))$:

$$\begin{aligned} \dot{z}_1 &= z_2, \\ \dot{z}_2 &= \alpha(z) + \beta(z) u = w, \\ y &= z_1, \end{aligned} \tag{5}$$

where $\alpha = L_f^2 h \circ \Phi^{-1}$, $\beta = L_g L_f h \circ \Phi^{-1}$, and $w \triangleq \alpha(z) + \beta(z) u$.

III. Learning to control unknown systems from expert demonstrations

The proposed methodology consists of the following steps:

1) using the state estimator from [4], estimate the state for each of the measurement sequences in $\mathcal{D}_s$;
2) based on these states, construct a linear time-varying controller $\widehat{v}(k, z)$;
3) use the dynamic controller proposed in [4] with $\widehat{v}(k, \widehat{z})$ as a reference, where $\widehat{z}$ is the state estimate produced by the state estimator in [4].

In what follows, we show that this methodology guarantees asymptotic stability of the origin for the closed-loop system.

A. Approximate model

It was shown in [4] that if we assume that (z, u) always remains within a compact set $\mathcal{R}$ — the proof of Theorem 8.1 in [4] justifies why this is a valid assumption[2] — the dynamics (5) can be approximated with:

$$\begin{aligned} z_{s1}(k+1) &= z_{s1}(k) + z_{s2}(k) T + (\alpha_s(k) + \beta_s(k) u_s(k)) \frac{T^2}{2}, \\ z_{s2}(k+1) &= z_{s2}(k) + (\alpha_s(k) + \beta_s(k) u_s(k)) T, \\ y_s(k) &= z_{s1}(k), \end{aligned} \tag{6}$$

where $\alpha_s(k)$ and $\beta_s(k)$ denote values of $\alpha(z)$ and $\beta(z)$ at time kT, respectively. System (6) can also be written as:

$$\begin{aligned} z_s(k+1) &= A z_s(k) + B(\alpha_s(k) + \beta_s(k) u_s(k)), \\ y_s(k) &= z_{s1}(k), \end{aligned} \tag{7}$$

where the matrices A and B are of the form:

$$A = I + T A_1 \qquad B = T B_1 + T^2 B_2, \tag{8}$$

and (A_1, B_1) is a Brunovsky pair.

[2]The argument is based on the initial conditions residing in a compact contained in some level-set of a Lyapunov function, and then proving that such level-set is forward invariant.

B. State estimator

For state estimation purposes, we follow the example technique provided in [4]. We define the extended state $z_s(k) = (z_{s1}(k), z_{s2}(k), w_s(k))$, where $w_s(k) \triangleq \alpha_s(k) + \beta_s(k) u_s(k)$, for (6) to obtain:

$$\begin{aligned} z_{s1}(k+1) &= z_{s1}(k) + z_{s2}(k) T + w_s(k) \frac{T^2}{2} \\ z_{s2}(k+1) &= z_{s2}(k) + w_s(k) T \\ w_s(k+1) &= w_s(k) \\ y_s(k) &= z_{s1}(k), \end{aligned} \tag{9}$$

which can be written in the form:

$$z_s(k+1) = \tilde{A} z_s(k), \qquad y_s(k) = z_{s1}(k) = \tilde{C} z_s(k).$$

Denoting by $\mathcal{O}$ the observability matrix for the pair $(\tilde{A}^{-1}, \tilde{C})$ allows us to compute the extended state estimate:

$$\widehat{z}_e(k) = (\mathcal{O}^T \mathcal{O})^{-1} \mathcal{O}^T Y(k), \tag{10}$$

where:

$$Y(k) \triangleq \begin{bmatrix} y_s(k) \\ y_s(k-1) \\ \vdots \\ y_s(k-\rho+1) \end{bmatrix}, \tag{11}$$

and $\rho \in \mathbb{N}$, $\rho \geq n+1$, is the number of measurements that will be used for state estimation. Using Proposition 4.2 in [4], we can show that the estimation errors satisfy:

$$z(kT+t) - \widehat{z}_s(k) = O_{(z_s, u_s - u_0)}(T^2) \tag{12}$$

$$w(kT+t) - \widehat{w}_s(k) = O_{(z_s, u_s - u_0)}(T). \tag{13}$$

where $u_0 = -\beta^{-1}(0)\alpha(0)$ and $t \in [0, T]$.

C. Estimating the expert solutions from measurements

One of the uses for the state estimator (9) is to produce the extended state estimates $\{\widehat{z}_e^i\}_{i=1}^{n+1}$ from the set of demonstrations $\mathcal{D}_s$, where each $\widehat{z}_e^i : \{\rho - 1, \dots, L_s - 1\} \to \mathbb{R}^{n+1}$ is the estimated extended state sequence calculated from the measurement sequence y_s^i. Because the state estimator (9) requires ρ measurements to estimate $\widehat{z}_e^i(k)$, the sequence $\widehat{z}_e^i$ is defined only on $\{\rho - 1, \dots, L_s - 1\}$. Without loss of generality, we shift the time index of each extended state sequence z_e^i and its estimate so that it is defined on $\{0, \dots, L-1\}$ instead, where $L \triangleq L_s - \rho + 1$. We also define the length of the demonstration estimate as $\tau \triangleq (L-1)T$.

We define the set of extended expert trajectories $\mathcal{D}_e = \{(z^i, w^i)\}_{i=1}^{n+1}$ and use it to define:

$$Z \triangleq \left[z^2 - z^1 \;\middle|\; \dots \;\middle|\; z^{n+1} - z^1 \right] \tag{14}$$

$$W \triangleq \left[w^2 - w^1 \;\middle|\; \dots \;\middle|\; w^{n+1} - w^1 \right], \tag{15}$$

where $z^i = \Phi(x^i)$ and $w^i = \alpha(z^i) + \beta(z^i) u^i$. We define the estimates of these matrices $\widehat{Z}_s$ and $\widehat{W}_s$ in a similar manner.

In what follows, we also use the matrices defined using the expert solutions from $\mathcal{D}$ given by:

$$U = \left[u^2 - u^1 \;\middle|\; \dots \;\middle|\; u^{n+1} - u^1 \right], \tag{16}$$

$$Z^1 = z^1 \mathbf{1}^T \quad U^1 = u^1 \mathbf{1}^T, \quad U_0 = u_0 \mathbf{1}^T, \tag{17}$$

where $\mathbf{1} \in \mathbb{R}^n$ is a vector of ones, and define:

$$D = (Z_s + Z_s^1, Z_s^1, U_s + U_s^1 - U_0, U_s^1 - U_0). \tag{18}$$

Lemma III.1. *If the estimates $\widehat{z}_s^i$ and $\widehat{w}_s^i$ satisfy* (12) *and* (13) *for a sampling time $T > 0$, then the estimates $\widehat{Z}_s$ and $\widehat{W}_s$ satisfy:*

$$Z(kT + t) - \widehat{Z}_s(k) = O_D(T^2) \tag{19}$$

$$W(kT + t) - \widehat{W}_s(k) = O_D(T), \tag{20}$$

for all $t \in [0, T]$, where D is defined in (18)*.*

Proof. The proof of this lemma can be found in [22]. □

D. Learning control from expert demonstrations

For the discussion that follows, we assume that α and β are known — this assumption will be relaxed in the next subsection. Knowing α and β allows us to apply to the system (5) the following preliminary controller:

$$u(t) = \beta^{-1}(z(t))(-\alpha(z(t)) + v(t)), \tag{21}$$

where $v(t)$ is the new input, resulting in the following:

$$\begin{aligned} \dot{z}(t) &= A_1 z(t) + B_1 v(t), \\ y(t) &= z_1(t). \end{aligned} \tag{22}$$

Below we present a result showing that an affine combination of solutions of (22) is a valid trajectory for (22).

Lemma III.2 ([3])**.** *Suppose we are given a set of finite-length solutions $\{(z^i, w^i, y^i)\}_{i=1}^{n+1}$ of the system* (22)*, where each (z^i, w^i, y^i) is defined for $t \in [0, \tau]$, $\tau \in \mathbb{R}_0^+$. Assume that $\{z^i(0)\}_{i=1}^{n+1}$ is an affinely independent set. Then, under the control law $v(t) = W(t - t_0)\zeta$ with $\zeta = Z^{-1}(0)z(t_0)$, the solution of the system* (22) *with the initial state $z(t_0)$ is:*

$$z(t) = Z(t - t_0)\zeta,$$

for all $t \in [t_0, t_0 + \tau]$, where the matrices $Z(t)$ and $W(t)$ are defined in (14) *and* (15)*, respectively.*

Proof. This lemma can be verified by substitution. □

Remark III.3. The requirement that $\{z^i(0)\}_{i=1}^{n+1}$ is an affinely independent set is a generic property, i.e., this is true for almost all expert demonstrations. In practice, if this condition is violated, a user can eliminate one of the affinely dependent demonstrations and collect additional demonstrations.

We propose using the following control law for the system (22):

$$v(t, z(t)) = W(t - p\tau)Z^{-1}(t - p\tau)z(t) = K(t)z(t), \tag{23}$$

for all $t \in [p\tau, (p + 1)\tau)$ and $p \in \mathbb{N}_0$, where:

$$K(t) \triangleq W(t - p\tau)Z^{-1}(t - p\tau). \tag{24}$$

The following lemma presents sufficient conditions for exponential stability of the system (22) in closed loop with (21)-(23). Its proof is similar to that of Theorem III.4 in [3] and is presented in [22].

Lemma III.4. *Suppose a set of extended trajectories $\mathcal{D}_e = \{(z^i, w^i)\}_{i=1}^{n+1}$ of length $\tau \in \mathbb{R}^+$ is generated by the system* (5) *in closed loop with an asymptotically stabilizing controller $u = \kappa(z)$. Assume that $\{z^i(t)\}_{i=1}^{n+1}$ is affinely independent for $t \in [0, \tau]$, and define $Z(t)$ and $W(t)$ as in* (14) *and* (15)*, respectively. Then, there is $\bar{\tau} \in \mathbb{R}^+$ so that for all $\tau \geq \bar{\tau}$, the origin of the system* (22) *in closed loop with the controller in* (23) *is uniformly exponentially stable.*

Remark III.5. Lemma III.4 shows that there is $\bar{\tau} \in \mathbb{R}^+$ such that $\|\Psi(\tau)\| < 1$ for all $\tau \geq \bar{\tau}$. In practice, a user can determine the upper bound on $\|\Psi(\tau)\|$ by calculating the product of the maximum singular value of $Z(\tau)$ and the minimum singular value of $Z(0)$, and guarantee exponential stability by ensuring that this product is less than 1.

E. Effect of the estimation error

A careful reader must have noticed that we used the exact values of Z and W to construct the controller (23). Recall, however, that we do not have access to Z and W and use the estimates $\widehat{Z}_s = Z + E_Z$ and $\widehat{W}_s = W + E_W$ instead, where, by Lemma III.1, the additive errors satisfy $E_S = O_D(T^2)$ and $E_W = O_D(T)$. Therefore, we, in fact, use[3]:

$$\begin{aligned} \widehat{v}(k, z_s(k)) &= \widehat{W}_s(k - pL')\widehat{Z}_s^{-1}(k - pL')z_s(k) \\ &= \widehat{K}_s(k)z_s(k) \end{aligned} \tag{25}$$

for $k \in \{pL', \ldots, (p + 1)L' - 1\}$ and $p \in \mathbb{N}_0$, where:

$$\widehat{K}_s(k) \triangleq \widehat{W}_s(k - pL')\widehat{Z}_s^{-1}(k - pL'). \tag{26}$$

The following lemma shows how using these estimates merely results in an additive error relative to the ideal controller K in (23). Its proof can be found in [22].

Lemma III.6. *Let $Z : [0, \tau] \to \mathbb{R}^{n \times n}$ and $W : [0, \tau] \to \mathbb{R}^{1 \times n}$ be defined in* (14) *and* (15)*, respectively, and $\widehat{Z}_s : \{0, \ldots, L - 1\} \to \mathbb{R}^{n \times n}$ and $\widehat{W}_s : \{0, \ldots, L - 1\} \to \mathbb{R}^{1 \times n}$ be estimates of Z and W satisfying* (19) *and* (20)*. Assume that $Z(t)$ is non-singular for all $t \in [0, \tau]$ and that $\widehat{Z}_s(k)$ is non-singular for all $k \in \{0, \ldots, L-1\}$. Then, the gains $K(t)$ and $\widehat{K}_s(k)$ defined in* (24) *and* (26)*, respectively, satisfy:*

$$K(t) - \widehat{K}_s(k) = O_D(T), \tag{27}$$

for all $t \in [kT, (k + 1)T]$ and $k \in \{0, \ldots, L - 1\}$.

Remark III.7. The matrix $\widehat{Z}_s(k)$ becomes singular whenever any pair of trajectory estimates $\widehat{z}_s^i(k)$ and $\widehat{z}_s^j(k)$ takes on the same value. If the expert demonstrations are initialized sufficiently far from each other, this usually only happens as the trajectories approach the origin. We can assume that $\widehat{Z}_s(k)$ is non-singular for all $k \in \{0, \ldots, L - 1\}$ since we can remove the tail end of the trajectories whenever the eigenvalues of $\widehat{Z}_s(k)$ become too small.

[3]Even though we will implement the controller with a state estimate $\widehat{z}_s$, in this subsection we assume that the controller has access to the sampled state z_s to simplify the exposition. We relax this assumption in the next subsection.

F. Data-driven dynamic controller

In this subsection, we show how we can control the system (7) *without* knowing the values of α and β. This can be achieved using the dynamic controller:

$$u(k+1) = u(k) + \gamma(\widehat{v}(k, z_s(k)) - w_s(k)), \tag{28}$$

from [4] in closed loop with (4) if we choose $\gamma \in \mathbb{R}^+$ to be sufficiently small. Given that we do not have access to $z_s(k)$ and $w_s(k)$, we resort to using $\widehat{z}_s(k)$ and $\widehat{w}_s(k)$ from (10) instead:

$$u(k+1) = u(k) + \gamma(\widehat{v}(k, \widehat{z}_s(k)) - \widehat{w}_s(k)). \tag{29}$$

According to Lemma 8.2 in [4], to conclude that the controller (29) asymptotically stabilizes the origin of (7), it suffices to show that the system (22) in closed loop with the control law $\widehat{v}(k, z_s(k))$ satisfies the following dissipation inequality:

$$\begin{aligned} V(z_s(k+1)) - V(z_s(k)) \leq -\lambda T \|z_s(k)\|^2 \\ + O_{(z_s, u_s - u_0)^2}(T^2), \end{aligned} \tag{30}$$

where $V : \mathbb{R}^n \to \mathbb{R}$ is a quadratic Lyapunov function, $\lambda \in \mathbb{R}^+$, and $\widehat{v}(k, z_s(k))$ satisfies the relation:

$$\begin{aligned} \widehat{v}(k, z_s(k) + O_{(z_s, u_s - u_0)}(T)) = v(k, z_s(k)) \\ + O_{(z_s, u_s - u_0)}(T). \end{aligned} \tag{31}$$

The following lemma shows that the controller (23) satisfies the requirements (30) and (31) and is proven in [22].

Lemma III.8. *Consider the system* (22) *in closed loop with the controller* (25) *and assume initial conditions reside in a set $\mathcal{R} \subset \mathbb{R}^2$. Then, there exists $\bar{\tau}$, $T^* \in \mathbb{R}^+$ such that for any $\tau \geq \bar{\tau}$ and all $T \in [0, T^*]$ there exist a quadratic time-varying Lyapunov function $V : \mathbb{N}_0 \times \mathbb{R}^n \to \mathbb{R}$ and $\lambda \in \mathbb{R}^+$ such that the dissipation inequality:*

$$\begin{aligned} V(k+1, z_s(k+1)) - V(k, z_s(k)) \leq -\lambda T \|z_s(k)\|^2 \\ + O_{(z_s, u_s - u_0)}(T^2), \end{aligned}$$

holds in the compact $\mathcal{R}$. Furthermore, the controller (25) *satisfies:*

$$\begin{aligned} \widehat{v}(k, z_s(k) + O_{(z_s, u_s - u_0)}(T)) = v(k, z_s(k)) \\ + O_{(z_s, u_s - u_0)}(T). \end{aligned}$$

Remark III.9. In the current work we use time-varying quadratic Lyapunov functions, while in [4] time-invariant quadratic Lyapunov functions are used instead. A detailed analysis of the main proof of [4] shows that due to our proposed Lyapunov functions being quadratic, and the same dissipation inequalities being satisfied, Lemma 8.2 in [4] can be applied with our proposed controller.

In what follows, we present the sufficient conditions for asymptotic stability of the system (4) in closed loop with (29)-(25). The proof uses all the previous results in this paper.

Theorem III.10. *Consider an unknown feedback linearizable SISO system* (4) *where the output function h has a relative degree 2. Let $T \in \mathbb{R}^+$ be the sampling time and $\tau \in \mathbb{R}^+$ be the demonstration length. Suppose we are given a set of finite-length measurement sequences $\mathcal{D}_s = \{y_s^i\}_{i=1}^{n+1}$ generated by the system* (4) *in closed loop with a piecewise constant asymptotically stabilizing controller. Further, suppose the state estimator* (9) *is used to construct a set of finite-length solution estimates $\widehat{\mathcal{D}}_e = \left\{\left(\widehat{z}_s^i, \widehat{w}_s^i\right)\right\}_{i=1}^{n+1}$ of* (7)*. Assume that $\left\{\widehat{z}_s^i(k)\right\}_{i=1}^{n+1}$ is affinely independent for all $k \in \{0, \ldots, L-1\}$, where $L \in \mathbb{N}$ is the solution length, and define $\widehat{Z}_s(k)$ and $\widehat{W}_s(k)$ in a similar way to* (14) *and* (15)*, respectively. For any compact set $\mathcal{S} \subset \mathbb{R}^n$ of initial conditions containing the origin in its interior, there exists a sampling time $\bar{T} \in \mathbb{R}^+$, a constant $b \in \mathbb{R}^+$ (both depending on $\mathcal{S}$) and demonstration length $\bar{\tau} \in \mathbb{R}^+$ so that, for any sampling time $T \in [0, \bar{T}]$ and any demonstration length $\tau \in [0, \bar{\tau}]$, the dynamic controller* (29)*, based on the learned controller* (25)*, using the state estimates provided by an estimation technique satisfying* (12) *renders the closed-loop solutions bounded, i.e., $\|\widehat{z}_s(k)\| \leq b$ for all $k \in \mathbb{N}$ and $\|x(t)\| \leq b$ for all $t \in \mathbb{R}_0^+$. Moreover, $\lim_{t \to \infty} x(t) = 0$.*

Remark III.11. Although we have presented all the results in the context of stabilization to the origin, they easily extend to the problem of trajectory tracking (see Section 4.5 of [23]).

IV. A NUMERICAL EXAMPLE

To demonstrate the performance of the proposed controller, we use it for altitude control of a quadrotor, which is described by the following model:

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = \frac{\sigma_0 - g}{m} + \frac{\sigma_1}{m} u, \tag{32}$$

where $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$ are the position and velocity of quadrotor in the z-axis, respectively, $u \in [0, 1]$ is the PWM signal for thrust, $m \in \mathbb{R}$ is the mass of the quadrotor, and $\sigma_0 \in \mathbb{R}$ and σ_1 describe an affine map from the PWM signal to physical thrust (see [4] for more details).

We define:

$$\alpha \triangleq \frac{\sigma_0 - g}{m}, \quad \beta \triangleq \frac{\sigma_1}{m}, \tag{33}$$

and apply the following expert controller:

$$u = \mathrm{sat}_{[0, 0.9]} \left(\frac{Kx - \alpha}{\beta} \right), \tag{34}$$

with $K = \begin{bmatrix} -5 & -2 \end{bmatrix}$, to the dynamics (32) in simulation. In the resulting expert demonstrations, the expert controller stabilizes the quadrotor state at the origin, starting from different positions and velocities. Since the dimension of the state is equal to 2, we simulate 3 expert demonstrations, record their corresponding measurements $\{y^i\}_{i=1}^3$, and use the state estimator (9) to estimate the expert solutions $\{(\hat{z}^i, \hat{w}^i)\}_{i=1}^3$. In this context, the pairs $(\hat{z}^i, \hat{w}^i)$ are just estimates of evolutions of vertical position, velocity, and acceleration. We check that the recorded demonstrations of length $\tau = 3$ s satisfy the sufficient conditions from Lemma III.4, i.e., the matrix $Z(t)$ is non-singular and $\|Z(\tau)Z^{-1}(0)\| < 1$. and use them to construct a stabilizing controller (25). Currently, we assume there is no measurement noise when collecting expert

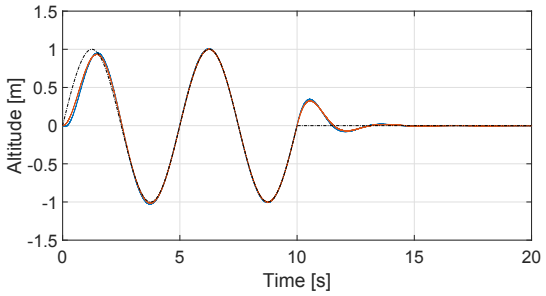

Fig. 1. Reference tracking by the expert controller (red) and the learned controller (blue). Reference is shown with the black dash-dot line.

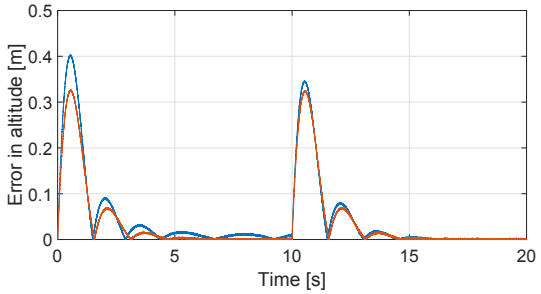

Fig. 2. Comparison of tracking errors of the expert controller (red) and the learned controller (blue).

demonstrations. We intend to address effects of measurement noise on this methodology in future work.

The learned controller is used, together with the state estimator (9) with $\rho = 5$ and the data-driven controller (29) with $\gamma = 0.002$, to track the reference trajectory shown in Figure 1, which consists of two parts: a sinusoid trajectory with period of 5 s from $t = 0$ s to $t = 10$ s and a setpoint to the origin from $t = 10$ s to $t = 20$ s. To elaborate, when tracking a sinusoid $y^R(t) = \sin(2\pi/5)t$, we use the learned controller gain $\hat{K}_s$ from (26) to control the tracking error $e_z = \hat{z}_s(k) - z_s^R(k)$ with:

$$\hat{v}(k, e_z) = \hat{K}_s(k)e_z + \ddot{y}_s^R(k), \tag{35}$$

where $z^R = (y^R, \dot{y}^R)$. When stabilizing to the origin, we use (35) with $y^R \equiv \dot{y}^R \equiv \ddot{y}^R \equiv 0$.

In Figure 1, we present the trajectories for the expert controller and the learned controller tracking the aforementioned reference when the measurement noise is white noise with RMS of 10^{-3}. In Figure 2, we compare the tracking errors of the learned controller with those of the expert controller. After awhile, both the learned controller and the expert controller begin to track the sinusoid well. We can also observe that the expert controller tracks the reference slightly better than the learned controller. Both when tracking the sinusoid and stabilizing to the origin, we can note that the tracking errors exponentially decreases to zero, confirming the theoretical results from the previous sections.

REFERENCES

[1] S. Chernova and A. L. Thomaz, *Robot Learning from Human Teachers*. Morgan & Claypool Publishers, 2014.

[2] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent Advances in Robot Learning from Demonstration," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, no. 1, pp. 297–330, 2020.

[3] A. Sultangazin, L. Fraile, L. Pannocchi, and P. Tabuada, "Watch and Learn: Learning to control feedback linearizable systems from expert demonstrations," UCLA, Tech. Rep., Mar 2021. [Online]. Available: http://www.cyphylab.ee.ucla.edu/Home/publications/UCLA-CyPhyLab-2021-03.pdf

[4] P. Tabuada and L. Fraile, "Data-driven Stabilization of SISO Feedback Linearizable Systems," *arXiv e-prints*, p. arXiv:2003.14240, Mar. 2021. [Online]. Available: https://arxiv.org/abs/2003.14240

[5] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," *Journal of Machine Learning Research*, vol. 22, no. 30, pp. 1–82, 2021.

[6] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the Twenty-First International Conference on Machine Learning*, ser. ICML '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 1.

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

[8] D. A. Pomerleau, *ALVINN: An Autonomous Land Vehicle in a Neural Network*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989, p. 305–313.

[9] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4693–4700.

[10] B. Boots and D. Fox, "Learning dynamic policies from demonstration," in *NIPS Workshop on Advances in Machine Learning for Sensorimotor Control*, vol. 6, no. 1, 2013.

[11] S. M. Khansari-Zadeh and A. Billard, "Learning stable nonlinear dynamical systems with gaussian mixture models," *IEEE Transactions on Robotics*, vol. 27, no. 5, pp. 943–957, 2011.

[12] J. Umlauft and S. Hirche, "Learning stable stochastic nonlinear dynamical systems," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 06–11 Aug 2017, pp. 3502–3510.

[13] M. A. Rana, A. Li, D. Fox, B. Boots, F. Ramos, and N. Ratliff, "Euclideanizing flows: Diffeomorphic reduction for learning stable dynamical systems," in *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, ser. Proceedings of Machine Learning Research, vol. 120. PMLR, 10–11 Jun 2020, pp. 630–639.

[14] M. Palan, S. Barratt, A. McCauley, D. Sadigh, V. Sindhwani, and S. Boyd, "Fitting a Linear Control Policy to Demonstrations with a Kalman Constraint," *arXiv e-prints*, p. arXiv:2001.07572, Jan. 2020.

[15] R. Kálmán, "When is a linear control system optimal," *Journal of Basic Engineering*, vol. 86, pp. 51–60, 1964.

[16] Michel Fliess and Cédric Join, "Model-free control and intelligent pid controllers: Towards a possible trivialization of nonlinear control?" *IFAC Proceedings Volumes*, vol. 42, no. 10, pp. 1531–1550, 2009, 15th IFAC Symposium on System Identification.

[17] M. Fliess and C. Join, "Model-free control," *International Journal of Control*, vol. 86, no. 12, p. 2228–2252, Dec 2013.

[18] D. Nešić and A. R. Teel, "A framework for stabilization of nonlinear sampled-data systems based on their approximate discrete-time models," *IEEE Transactions on Automatic Control*, vol. 49, no. 7, pp. 1103–1122, 2004.

[19] M. Arcak and D. Nešić, "A framework for nonlinear sampled-data observer design via approximate discrete-time models and emulation," *Automatica*, vol. 40, no. 11, pp. 1931–1938, 2004.

[20] J. Coulson, J. Lygeros, and F. Dörfler, "Data-enabled predictive control: In the shallows of the deepc," in *2019 18th European Control Conference (ECC)*, 2019, pp. 307–312.

[21] C. De Persis and P. Tesi, "Formulas for data-driven control: Stabilization, optimality, and robustness," *IEEE Transactions on Automatic Control*, vol. 65, no. 3, pp. 909–924, 2020.

[22] A. Sultangazin, L. Fraile, and P. Tabuada, "Exploiting the experts: Learning to control unknown SISO feedback linearizable systems from expert demonstrations. Technical report," UCLA, Tech. Rep., Aug 2021. [Online]. Available: http://www.cyphylab.ee.ucla.edu/Home/publications/UCLA-CyPhyLab-2021-08.pdf

[23] A. Isidori, *Nonlinear Control Systems*, ser. Communications and Control Engineering. Springer-Verlag London, 1995.