Data-Driven Design of a Reference Governor Using Deep Reinforcement Learning

Maria Angelica Arroyo\textsuperscript{1} and Luis Felipe Giraldo\textsuperscript{2}

Abstract—Reference tracking systems involve a plant that is stabilized by a local controller and a command center that indicates the reference set-point the plant should follow. Typically, these systems are subjected to limitations such as poorly designed controllers that do not allow them to achieve the desired performance. In situations where it is not possible to redesign the closed-loop system, it is usual to incorporate a reference governor that instructs the system to follow a modified reference path such that the resultant path is close to the ideal one. Current strategies to design the reference governor need to know a model of the system, which can be an unfeasible task. In this letter, we propose a framework based on deep reinforcement learning that can learn a policy to generate a modified reference that improves the system’s performance in a non-invasive and model-free fashion. To illustrate the effectiveness of our approach, we present two challenging cases: a flight control with a pilot model that includes human reaction delays, and a mean-field control problem for a massive number of space-heating devices. The proposed strategy successfully designs the reference governor that works even in situations that were not seen during the learning process.

I. INTRODUCTION

The main objective in trajectory tracking control is to find the appropriate strategy to make the output of a dynamical system follow a pre-defined trajectory. Even though control design strategies such as robust control \cite{1}, adaptive control \cite{2}, and model predictive control \cite{3} have effectively achieved high performance and stability in this task, scenarios with implementation issues and constraints, unmeasured disturbances, and uncertainties can affect the control action and deteriorate the system’s performance. In situations where it is not possible to redesign or modify the local controller, it is usual to incorporate a reference governor into the system following the scheme shown in Fig. 1. The main idea is to provide a modified reference when the original reference may lead the system to undesired trajectories. This scheme has been studied and implemented for different types of systems and scenarios including linear, nonlinear, and networked and multi-agent systems \cite{4}, \cite{5}. In such approaches to design and analyze a reference governor an important assumption that needs to be considered is that the model of the local controller, plant, and boundaries of the uncertainties have to be known a priori. This assumption is often a complicated or impossible task to conduct. Very little has been done to propose data-driven strategies in a reference governor framework for systems where, besides the infeasibility to change its local controller, there is not an accurate model available that characterizes its dynamics \cite{6}, \cite{7}. We address this problem in this paper. We hypothesize that Reinforcement Learning (RL) can offer powerful tools for designing a reference governor in an online and model-free scheme. RL is a branch of artificial intelligence where an agent learns an optimal control policy by maximizing a reward function while it is interacting with the environment. The incorporation of deep neural networks (DNN) as function approximators into the RL framework is known as deep reinforcement learning (DRL). Complex control problems with discrete and continuous action spaces such as robotic manipulation, bipedal locomotion, and tracking control problems have been successfully solved using DRL \cite{8}–\cite{11}. Inspired by the ideas in RL, in this letter we explore the implementation of a non-invasive, full-online and model-free reference governor using DRL in the context of tracking control problems. As far as we know, this has not been done before. We assume that i) the model of the system is unknown, ii) the local controller cannot be modified, and iii) the control signal given by the local controller cannot be observed. The goal of our DRL agent is to improve the performance of the system only using the information of the output variable and the desired reference. The learned optimal policy is used to generate the reference input of the closed-loop system. In a model-free fashion, this modified reference takes into account the uncertainties, imperfections and constraints of this system to achieve the desired performance. To show the effectiveness of our proposed strategy, we present two complex control problems: human in the loop systems and mean field games. The first study case is a challenging problem due to the human reaction delays as a critical parameter for instabilities. We present numerical simulations for an aircraft pitch angle control with the Neal-Schmidt pilot model, that is a first order lead-lag-type compensator with a gain, time lag with

---

\textsuperscript{1}Maria Angelica Arroyo is with Department of Electrical and Electronic Engineering, University of los Andes, Bogota, Colombia ma.arroyo10@uniandes.edu.co

\textsuperscript{2}Luis Felipe Giraldo is with Department of Electrical and Electronic Engineering, University of los Andes, Bogota, Colombia lf.giraldo404@uniandes.edu.co

---

Fig. 1: Reference governor scheme applied to a closed loop system with constraints.
time-delay. The second study case involves the control of a very large population of individuals using mean field games. We present numerical simulations where the mean temperature of a system composed of a massive number of space heating devices follows a target temperature trajectory using a strategy based on mean field games. Each household is modeled using a stochastic differential equation and has a local linear control law with a poor performance that only requires the information of the target trajectory. We are able to achieve the desired performance with our proposed DRL-based reference governor strategy in a model-free fashion while preserving the local controller of each device in the population.

This paper is organized as follows. In Section II we present our reference governor architecture and the explanation of the DRL elements in a control system context. In Section III, the Human in the Loop Control is introduced and numerical simulations are shown. The case of study Mean Field Games with numerical simulations is presented in Section IV. Finally, in Section V conclusions as well as future research directions are given.

II. REFERENCE GOVERNOR WITH DEEP REINFORCEMENT LEARNING

The solution architecture of the reference governor with DRL is shown by Fig. 2. In this scheme, the command center contains the goal generator and the reference governor. The goal generator block indicates the desired reference \( r_t \) to the reference governor at each time \( t \). The main function of this block is to generate a pre-established trajectory in space to be followed by the controlled system. The reference \( r_t \) can be feedback controlled by the behavior of the system with \( y_t \) as an input. For example, the case of a human operator that interacts with the observed state of the system to indicate the next goal using feedback driving devices such as joystick, flight stick, wheels and touch screens. In the DRL-based reference governor strategy, we assume that the model of the system is unknown, the local controller cannot be modified, and the control signal given by the local controller cannot be observed. Also, we assume that the local controller leads to a poor performance as a consequence of unmeasured disturbances, parametric uncertainties, model approximations and implementation constraints. The DRL agent learns an optimal policy that determines the modified reference \( \tilde{r}_t \) that will improve the the closed-loop system’s performance in an online and model-free fashion. The actual followed trajectory \( y_t \) corresponds to the output variables of the closed-loop system.

In a typical DRL framework, the environment takes the agent’s current state and action as inputs, and returns as outputs the agent’s reward and its next state. By contrast, in our DRL control framework, the environment is the controlled dynamical system which takes the action as an input and returns its next reference state. The reward function is a fundamental component in a DRL framework that measures the success or failure of the agent’s actions in a given state, driving the learning process. We propose an approach where the reward is given by a function that leads the system to minimize the tracking error and to try to satisfy output constraints. Let \( e_t = \| r_t - y_t \| \) be the absolute tracking error. The proposed reward function is given by

\[
R(e_t, y_t, \tilde{r}_t) = G(e_t) + H(\tilde{r}_t) + B(y_t)
\]

where \( G \) is a decreasing function that leads the learning process to an error minimization, and \( H \) and \( B \) are functions that penalize state values and modified reference values (respectively) that do not satisfy constraints that have been previously defined. Functions \( G, B, \) and \( H \) are designed depending on the application where the reference governor scheme is used.

Other important elements are the policy and the action-value function. The policy \( \mu \) is the strategy that the agent employs to determine the next action based on the current state. It maps states to actions that maximize the reward. The action-value function \( Q_\mu(y_t, \tilde{r}_t) \) refers to the long-term return of taking action \( \tilde{r}_t \) under policy \( \mu \) from the current state \( y_t \). In our cases of study, we use deep deterministic policy gradients (DDPG) [12] as our DRL algorithm. DDPG is a popular model-free, actor-critic algorithm that estimates a deterministic policy and that can operate over continuous action spaces. DDPG uses four function approximators: the actor \( \tilde{r}_t = \mu(y_t|\theta^\mu) \), parameterized by \( \theta^\mu \), which specifies the current policy by deterministically mapping states \( y_t \) to actions; the critic \( Q(y_t, \tilde{r}_t|\theta^Q) \), parameterized by \( \theta^Q \), which maps \( y_t \) and \( \tilde{r}_t \) given by the actor into the value function that evaluates the long term return (or future rewards) of applying \( \tilde{r}_t \) in \( y_t \). The other two neural networks in DDPG are used to facilitate the learning process. The critic network is updated with the gradients of the loss function that minimize the difference between the estimated reward and the actual received reward, and the actor network is updated with using deterministic policy gradients. Detailed information on the DDPG algorithm can be found in [12].

III. HUMAN-IN-THE-LOOP: PITCH ANGLE CONTROL

Controlling human in the loop (HIL) systems is a challenging case due to the human reaction delays as a critical parameter for instabilities. We study the pitch angle control problem of an aircraft. This is a problem where the variable to be controlled is the pitch Euler angle (in rad) of the aircraft body axis with respect to the reference axes. We consider the dynamics given by [13] of longitudinal motion of a Boeing 747 airplane linearized at an attitude of 40k ft and velocity of 774 ft/sec. The aircraft is assumed to be operated by a pilot described by a Neal-Schmidt pilot model. In [14], an online approach is used to guarantee stability and to achieve a desired behavior of the closed-loop system, under the assumption that a model of the system is known a priori. In our DRL-based reference governor scheme, we assume that the aircraft is controlled by a local linear controller obtained by a LQR approach, and we explore the application of a reference governor as a reinforcement learning agent trained using DDPG.
Let $x_t = [x_{1t}, x_{2t}, x_{3t}, x_{4t}]^T$ be the state vector of the dynamical system where $x_{1t}$, $x_{2t}$ and $x_{3t}$ are the components of the velocity along the three axes (in crad/sec), and $x_{4t}$ represents the pitch Euler angle of the aircraft body axis with respect to the reference axes (in crad). Let $\xi_t \in \mathbb{R}$ be the internal human state vector, and $c_t$ be command produced by the human. A general class of linear human models with constant time delay is given by [14]

$$
\dot{\xi}_t = A_h \xi_t + B_h \theta_t e^{-\tau} \\
c_t = C_h \xi_t + D_h \theta_1 e^{-\tau} \\
\theta_t = r_t - x_{4t},
$$

where $\theta_t$ is the perceived difference between the reference and the pitch angle of the aircraft, $r_t$ is desired reference for the pitch angle. It is assumed that the human has an internal delayed response $\tau \geq 0$. This is a linear time-invariant model with time delays known as the Neal-Schmidt model [15] widely used to characterize human dynamics. The dynamics of an aircraft that is controlled by a human are given by

$$
\dot{x}_t = A x_t + B (u_t + \delta(x_t)) \\
y_t = x_{4t} \\
\dot{x}_{4t} = x_{4t} - c_t \\
u_t = -K_1 x_t - K_2 x_{4t}.
$$

Note that the aircraft’s elevator control input $u_t$ is the compound action of a proportion of the state vector $x_t$ and a proportion of the integral of the error between the human command and the pitch angle. Constants $K_1$ and $K_2$ are designed by a LQR design approach. Function $\delta(x_t) \in \mathbb{R}$ models a state-dependent error in the control action. In [14], this expression is defined as $\delta(x_t) = W^T [1, x_{1t}, x_{2t}]^T$.

### A. Simulation Results

The simulation parameters are presented in Table I. The learning process is conducted by randomly choosing an initial condition of the state variables of the system and a reference to be reached. After the training stage, we test the reference governor agent in two trajectory tracking scenarios: an aircraft in normal operation and an aircraft with a local control action based on delayed information. We design the reward function in Equation (1) as follows. We define $G(e_t)$ as a linear and decreasing function of the absolute error $G(e_t) = -e_t$, where $\alpha_e = 0.3$. In this problem, since $\bar{r}_t$ is the reference shown to the human, trajectories with quick changes of the modified reference $\bar{r}_t$ are not preferred. Let $\Delta \bar{r}_t = |\bar{r}_t - \bar{r}_{t-1}|$. Then, we define $H$ as

$$
H(\bar{r}_t) = \begin{cases} -\beta_h \Delta \bar{r}_t & \text{if } \Delta \bar{r}_t \geq \rho_h \\ 0 & \text{otherwise} \end{cases}
$$

where $\rho_h = 3.5$ and $\beta_h = 0.1$. Also, since the output of the system is the pitch angle of an aircraft, in this problem, quick changes of the pitch angle variable and values outside a predefined interval are not preferred. Let $\Delta y_t = |y_t - y_{t-1}|$. Then, $B$ is given by

$$
B(y_t) = \begin{cases} -\beta_y \Delta y_t - \delta_y \max(0, |y_t| - L) & \text{if } \Delta y_t \geq \rho_y \\ -\beta_y \max(0, |y_t| - L) & \text{otherwise} \end{cases}
$$

with $\beta_y = 0.1, \delta_y = 1, \rho_y = 3.5$. This function penalizes quick changes of the output of the system and values outside a boundary defined by $L$. We set $L = 18$ crad. Fig. 3a shows the closed-loop response in the context of trajectory tracking without the reference governor. Fig. 3b shows the system response to the modified reference given by the DR-based reference governor. The reference governor is able to find a policy that tries to satisfy the system’s constraints while minimizing the tracking error. The mean absolute error (MAE) for the system without the reference governor is 10.91 crad and for the reference governor scheme it is 3.90 crad. This is an improvement of 64.25% in the MAE.

### Table I: Simulation parameters of the HIL

| $A_h$ | 0.2 |
| $B_h$ | 1 |
| $C_h$ | 0.08 |
| $D_h$ | 0.1 |
| $\tau$ | 0.5 |
| $A$ | $\begin{bmatrix} -0.003 & 0.039 & 0 & -0.322 \\ -0.065 & -0.319 & 7.740 & 0 \\ 0.020 & -0.101 & -0.429 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ |
| $B$ | $[0.010, -0.180, -1.160, 0]^T$ |
| $W$ | $[0.1, 0.3, -0.3]^T$ |
| $Q$ | $\text{diag}(0.0, 0, 1, 2.5)$ |
| $R$ | 10 |

Fig. 2: Interconnection between the reference governor as a DRL agent and the controlled system.
Note that the output of the system in the reference governor scheme remains inside the defined boundaries.

Now, we test the performance of our learned policy in a scenario where there is a delay of 5.7s in the controller of the aircraft, that is, \( u_t = -K_1x_t - 5.7 - K_2x_{ct}(t-0.57) \). Fig. 3c shows the closed-loop response with instabilities due to the delays in the controller, and Fig. 3d shows the system response to the given reference. The reference governor is able to minimize the tracking error, to enforce systems constraints, and to provide a smooth output signal in a scenario that was not directly part of the learning process. The mean absolute error (MAE) for the system without the reference governor is 1.013crad, and 2.46crad for the reference governor scheme. This is an improvement of 75.71%.

IV. MEAN FIELD GAMES: ELECTRIC SPACE HEATERS

The second study case involves populations with many individuals that have common objectives. We study a power system problem where it is required to control the average temperature of a massive number of space heating devices in a way it follows a target temperature trajectory [16]. The dynamics of the space heating devices are described by a stochastic linear system that is controlled by a linear control law based on mean field games (MFGs) [17]. MFGs are commonly used to define strategies to control large-scale complex multi-agent systems [16], [17]. In this control strategy, each space heater is seen as an individual that optimizes its cost function only using local information and a cost coupling term that measures the average effect of the mass formed by all the other individuals. Even though this control law provides some theoretical guarantees and implementation benefits, its ability to lead the system to follow the desired reference needs to be improved. Since this case of study involves a massive number of heaters, it is unfeasible and cost inefficient to modify every single household electric space heater. We show how our DRL-based reference governor scheme is able to improve the performance of the system in a model-free fashion without modifying the local linear control based on MFGs [17].

Let \( x_t^i \) be the air temperature inside the household \( i \) at time \( t \). The thermal dynamics of the controlled household \( i \) are modeled using the stochastic differential equation [16]

\[
\begin{align*}
\frac{dx_t^i}{dt} &= \frac{1}{C_a}[-U_a(x_t^i - x^\text{out}) + Q_h]\delta t + \sigma d\omega_t^i \\
Q_h &= u_t^i + u^\text{free} \\
u^\text{free} &= -\frac{U_a}{C_a} (x^\text{out} - x_t^i),
\end{align*}
\]

where \( x^\text{out} \) is the outside ambient temperature, \( x_0^i \) is the initial temperature of each household, \( C_a \) is the thermal mass of air inside, \( U_a \) is the conductance of air inside, \( Q_h \) is the heat flux from the heater, \( u_t^i \) is a standard Wiener process that characterizes perturbations on the system, \( \sigma \) is a volatility term, and \( N \) is the number of individuals. The heat flux from the heater depends on the control action \( u_t^i \).

The main goal of the control action \( u_t^i \) is to drive household \( i \) to minimize its own cost function while mean field reaches a reference temperature \( x^\text{ref} \). The linear control law that each household locally implements is derived in [17] as the solution of an optimal control problem that tries to minimize the expected value of the error between the control signal and the given reference and the energy spent by the controller subjected to the dynamics of the system modeled by the stochastic differential equation in (3). This solution is given by

\[
\begin{align*}
u_t^i &= -\frac{1}{C_a} \left( \Pi_a x_t^i + s_t^i \right) \\
s_t^i &= s_\infty + \left( \frac{\gamma}{\beta_2 - \lambda_1} \right) (x^\infty_t - x_0^i) e^{\lambda_1 t} \\
s_\infty &= -\frac{\gamma}{\beta_2 - \lambda_1},
\end{align*}
\]

where \( \Pi_a \) is a constant that comes from solving an algebraic Riccati equation, \( s_t^i \) is an offset of the control action, and \( x_0^i \) is the state at \( t = 0 \). The interpretation of the remaining parameters \( \phi, r, \lambda_1, \) and \( \gamma \) is given in detail in [17]. The incorporation of this controlled system into the DRL-based reference governor scheme in Fig. 2 involves the application of a modified reference such that the mean temperature defined as \( y_t = \frac{1}{N} \sum_{i}^N x_t^i \) accurately follows the desired reference. Details on the parameter design can be found in [17].

A. Simulation Results

The learning process is conducted by randomly choosing an initial condition of the state variables of the system and a reference to be reached for four electric space heaters. After training stage, we test the reference governor agent in two trajectory tracking scenarios: a mean field problem with four electric space heaters and a mean field problem with 1000 heaters. Simulation parameters are set as follows: \( C_a = 10kW/°C, U_a = 0.2kW/°C, x^\text{out} = -5°C, \sigma = 0.25, \Pi_a = 0.4, r = 10, \delta = 0.001, \gamma = 0.6, \beta_1 = 0.038, \) and \( \lambda_1 = -0.028. \) The proposed \( G(e_t) \) in Equation (1) is given by \( G(e_t) = -\alpha_g e_t \), where \( \alpha_g = 0.3. \) Function \( G(e_t) \) is decreasing and linear with slope \( -\alpha_g. \) Since it is not convenient to force the system to have sudden changes in the modified reference shown to each electric space heater, function \( H \) is defined as in Equation (2) with \( \beta_h = 0.1 \) and \( \rho_h = 1. \) In this problem, we consider a comfort constraint that dictates that the maximum temperature change per hour on each household is \( 1°C. \) To enforce the comfort constraint, \( B \) is defined as

\[
B(y_t) = \begin{cases} 
-\beta_h \Delta y_t - \delta_h \max(0, \Delta y_t - L) & \text{if } \Delta y_t \geq \rho_h \\
-\delta_h \max(0, \Delta y_t - L) & \text{otherwise}
\end{cases}
\]

with \( \rho_h = \delta_h = L = 1, \beta_h = 0.1 \) and \( \Delta y_t = |y_t - y_{t-1}|. \) This reward function significantly improves tracking error and enforces the comfort constraint to be satisfied. The mean field tracking with no reference governor is shown in Fig. 4a.

As predicted, the linear control action leads the space heaters to an underperforming trajectory. Fig. 4b shows that the
Fig. 3: Human in the Loop - Pitch control. (a) Closed-loop system response with mean absolute error (MAE of 10.91 crad). (b) Closed loop system response with reference governor (MAE of 3.90 crad). (c) Closed-loop system response with instabilities (MAE of 10.13 crad). (d) Closed loop system response with instabilities and reference governor (MAE of 2.46 crad).

The reference governor is able to learn an optimal policy that allows the system to achieve the desired behavior in a model-free fashion while preserving the linear control law. The MAE of the MFG system with no reference governor was 15.62°C, and for the system under the reference governor scheme was 1.17°C, which is an improvement of 92.5% in the MAE. Additionally, using the policy learned by the reference governor agent from the scenario with 4 electric space heaters, we tested a system with 1000 space heaters. In this case, the MAE was improved from 15.07°C with just the local controller to 1.31°C when our reference governor was added. The results are shown in Fig. 4c. In Fig. 4d, we present the temperature change per hour to show that the comfort constraint is satisfied for the mean field. However, some in some households this constraint is violated when the desired reference changes abruptly and the system focuses on minimizing the tracking error as soon as possible. It is important to note that a reward without the $B$ function would lead to a higher violation of the constraint in order to achieve a faster tracking task.

V. CONCLUSIONS

We presented an exploratory study of a data-driven and non-invasive reference governor in trajectory-tracking control systems using deep reinforcement learning. The main objective of this reference governor was to provide a modified reference to the trajectory-tracking system such that the performance of the system is improved. The reference governor was seen as an agent whose learning process is driven by a reward function that depends only on the output of the controlled system and the reference to be reached by the trajectory-tracking system. We proposed a general form of the reward function that leads the system to minimize the tracking error and that penalizes state values and modified reference values that do not satisfy previously defined constraints. We analyzed two challenging study cases where the overall performance of the system was significantly improved: a human in the loop system with time delays, and a control system involving a massive number of electric space heaters where the local control is ruled by a mean field game-based control action. We showed that the DRL-based reference governor is able to improve the performance of the controlled systems.

Future directions include study strategies to have a safe exploration during the learning process to ensure critical constraint satisfaction [18]. Also, recent theoretical analyses of control techniques based on safe reinforcement learning can be applied to study properties of the DRL-based reference governor [19].

REFERENCES

[1] I. Queinnec, S. Tarbouriech, and M. Mazerolles, “Reference tracking controller design for anesthesia,” IFAC-PapersOnLine, vol. 51, pp.
Fig. 4: Mean Field Control - Electric Space Heaters. (a) Closed-loop system response with MAE of 15.62°C. (b) Closed loop system response with reference governor with MAE of 1.17°C. (c) Closed-loop system response with reference governor for 1000 space heaters with MAE improvement of 91.3%. (d) Absolute temperature difference for 1000 heaters.