A Model Free Learning Algorithm to Control Autonomous Streams over IoT

Alireza Farahmandi
Naval Air Warfare Center
China Lake, CA 93555
alireza.farahmandi@navy.mil

Gary A. Hewer
Naval Air Warfare Center
China Lake, CA 93555
gary.hewer@navy.mil

Brian C. Reitz
Naval Air Warfare Center
China Lake, CA 93555
brian.reitz@navy.mil

Katia Estabridis
Naval Air Warfare Center
China Lake, CA 93555
katia.estabridis@navy.mil

Kyriakos G. Vamvoudakis
Georgia Tech
Atlanta, GA 30332
kyriakos@gatech.edu

ABSTRACT
This paper presents the application and effectiveness of a recent novel model-free Q-learning algorithm to control linear systems with unknown dynamics in support of the growing Internet of Things (IoT) ecosystem.

ACM Classification Keywords
I.2.6. Artificial Intelligence: Learning; I.2.8. Problem Solving, Control Methods, and Search: Control Theory

Author Keywords
Q-learning, Optimal control, Uncertain systems

INTRODUCTION
In recent years, Internet-connected devices have been extensively utilized in various applications and, with the proliferation of the IoT and cloud based services, it is easy to envision a diverse set of devices becoming an integral component of the IoT ecosystem. Massive deployment of these devices is expected as their technology matures and cost lowers. IoT devices can be configured with a diversity of sensors to collect data in coordination with other IoT devices. As the number of IoT devices and their utility increases, it becomes essential to design and implement efficient mechanisms to reconfigure, program, and control them. Traditional controllers are based on an underlying dynamical model of the device which is designed to a particular configuration of the platform and usually requires retuning when tasks and performance requirements are changed. Reinforcement Learning (RL), as an online-learning approach, can be implemented in real time to control IoT devices without any knowledge of the system dynamics or the environment. In this paper, we present the application and effectiveness of an online model-free Q-learning algorithm to derive controllers for different types of systems.

IOT'18, October 15–18, 2018, Santa Barbara, CA, USA

ACM ISBN 978-1-4503-6564-2/18/10.
DOI: https://doi.org/10.1145/3277593.3277640

Q-LEARNING APPROACH
Our model-free Q-learning approach is based on the recent work presented in [7]. This online algorithm was proposed to solve the infinite-horizon optimal control problem of an LTI system with completely unknown dynamics.
Considering an LTI continuous-time system,

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1}$$

with $x(0) = x_0$ and $t \geqslant 0$, where $x(t) \in \mathbb{R}^n$ is the state vector and, $u(t) \in \mathbb{R}^m$ is the control input, and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ are the plant and input matrices respectively and assuming controllable dynamics, the goal is to find the optimal value function V^* defined by,

$$V^*(x(t)) := \min_u \int_t^\infty \frac{1}{2}(x^T Mx + u^T Ru)\, d\tau,\ \forall t \tag{2}$$

but without any information of the system dynamics (A,B). Note that $M \succeq 0$ and $R \succ 0$ are user defined matrices and the pair $(\sqrt{M},A)$ is detectable. The assumption that (A,B) is controllable and $(\sqrt{M},A)$ is detectable, will guarantee that the algebraic Riccati equation will have a unique non-negative solution [7] when the control contains full-state feedback.
Traditionally, with the full knowledge of the dynamics, the Hamiltonian associated with (1) and (2) can be considered as follows,

$$H(x,u,\frac{\partial V^*}{\partial x}) = \frac{\partial V^*}{\partial x}^T (Ax+Bu) + \frac{1}{2}x^T Mx + \frac{1}{2}u^T Ru \tag{3}$$

and the optimal control can be found to be,

$$u^*(x) = \arg\min_u H(x,u,\frac{\partial V^*}{\partial x}) = -R^{-1}B^T \frac{\partial V^*}{\partial x} \tag{4}$$

For the linear system (1), the value function can be considered as quadratic in the state,

$$V^*(x) = \frac{1}{2}x^T Px,\ \forall x \tag{5}$$

where $P \in \mathbb{R}^{n\times n}$ is the unique symmetric positive definite matrix that solves this Riccati equation,

$$A^T P + PA - PBR^{-1}B^T P + M = 0 \tag{6}$$

Therefore, the optimal control (4) can be written as follows,

$$u^*(x) = -R^{-1}B^T Px,\ \forall x \tag{7}$$

Note that (6) and (7) require full knowledge of the system dynamics.
This new Q-learning approach solves the optimal control problem without any information of the system dynamics, by adjusting parameters in an adaptive way.
The Q-function is defined by adding the Hamiltonian (3) to the value function (5),

$$\begin{aligned} Q(x,u) &:= V^*(x) + H(x,u,\frac{\partial V^*}{\partial x}) \\ &= V^*(x) + \frac{1}{2}x^T P(Ax+Bu) + \frac{1}{2}(Ax+Bu)^T Px \\ &+ \frac{1}{2}x^T Mx + \frac{1}{2}u^T Ru,\ \forall x,u \end{aligned} \tag{8}$$

By defining $U := [x^T u^T]^T$ a compact quadratic Q-function can be written as follows,

$$\begin{aligned} Q(x,u) &:= \frac{1}{2}U^T \begin{bmatrix} P+M+PA+A^T P & PB \\ B^T P & R \end{bmatrix} U \\ &= \frac{1}{2}U^T \begin{bmatrix} Q_{xx} & Q_{xu} \\ Q_{ux} & Q_{uu} \end{bmatrix} U,\ \forall x,u \end{aligned} \tag{9}$$

In this framework, by solving $\frac{\partial Q(x,u)}{\partial u} = 0$ the optimal control can be found as follows,

$$u^*(x) = \arg\min_u Q(x,u) = -Q_{uu}^{-1}Q_{ux}x \tag{10}$$

Note that (10) is the model-free formulation of (4).
By introducing an actor/critic structure as follows, we can tune the parameters of the Q-function to solve the infinite-horizon optimal control problem of a LTI system with completely unknown dynamics.

$$\begin{aligned} Q(x,u) &= W_{critic}^T (U \otimes U) \\ u(x) &= W_{actor}^T x \end{aligned} \tag{11}$$

In (11), the critic approximator W_{critic} will approximate the Q-function and the actor approximator W_{actor} will approximate the optimal controller as derived in [7].

NUMERICAL EXAMPLES

To demonstrate the range of effectiveness of this proposed Q-learning algorithm we applied it to five classic LTI examples that were taken from the literature in [5, 4, 1, 6, 2]. For the examples we present, the initial state vector and the initial weights for the critic were generated randomly between 0 and 1 before being scaled by a constant. The initial weights for the actor are constant multiplicities of the identity matrix. The actor and critic gradient descent parameters were selected as $\alpha_a = 0.01$ and $\alpha_c = 35$ respectively. The delay time is the iteration time step and the number of iterations in each simulation can be calculated by dividing the total simulation time in the figures by this delay time. Also, in all of the simulations, we do not add any exploration noise.

Example 1: This 5th order unstable system is from Kailath [4][5] with

$$A = \begin{bmatrix} -0.0297 & 0.331 & -1.13 & 0 & 0 \\ -1 & -0.0042 & 0.128 & 0 & 1 \\ 0 & -0.0461 & -0.803 & 1 & 0 \\ 0.0438 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1000 & 0 \\ 0 & 1000 \end{bmatrix}$$

$M = I$, $R = 0.1I$ and delay time $T = 0.001$ sec. This system is controllable and observable. The evolution of the states (dashed lines) in this system along with the optimal LQR solution (solid lines) are shown in Figure 1.

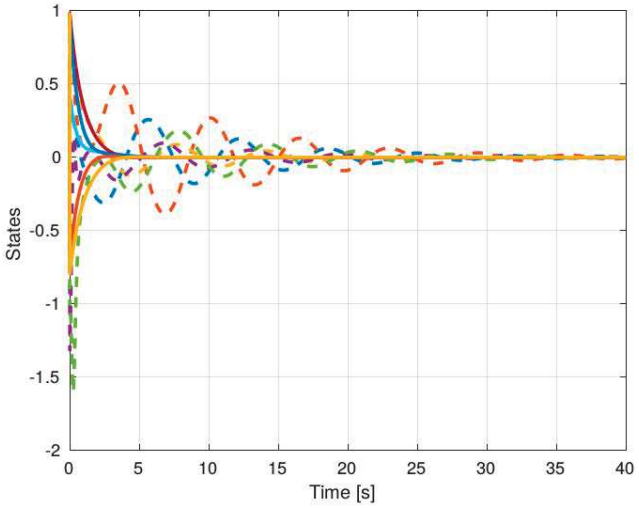

Figure 1. Evolution of system states in example 1

Example 2: This 6th order system is the lateral-directional, rigid-body model of the L-1011 aircraft in cruise flight and is taken from Andry-Shapiro-Chung [1][5]. The plant matrix of the system A is

$$\begin{bmatrix} -20 & 0 & 0 & 0 & 0 & 0 \\ 0 & -25 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ -0.744 & -0.032 & 0 & -0.154 & -0.0042 & 1.54 \\ 0.337 & -1.12 & 0 & 0.249 & -1 & -5.2 \\ 0.02 & 0 & 0.0386 & -0.996 & -0.000295 & -0.117 \end{bmatrix}$$

The input and state weighting matrices are

$$B = \begin{bmatrix} 20 & 0 \\ 0 & 25 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Also, we selected $R = I$ and delay time $T = 0.001$ sec. This system is controllable and observable. The evolution of the states (dashed lines) in this system along with the optimal LQR solution (solid lines) are shown in Figure 2.

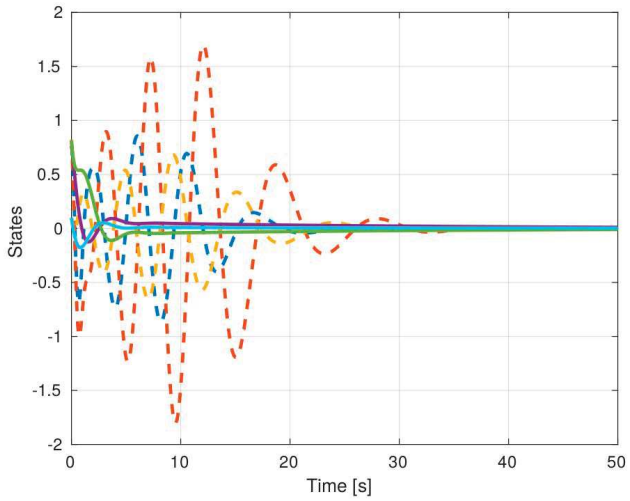

Figure 2. Evolution of system states in example 2

Example 3: This 10th order example is a stable system taken from Lainiotis [6][5] with

$$A = \begin{bmatrix} 0 & I \\ A_0 & A_0 \end{bmatrix}$$

$$B = [1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1]^T$$

where

$$A_0 = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -2 \end{bmatrix}, \qquad M = I$$

In this case, we selected $R = 0.1I$ and a delay time of $T = 0.001$ seconds with the system being observable and controllable. The evolution of the states (dashed lines) in this system along with the optimal LQR solution (solid lines) are shown in Figure 3.

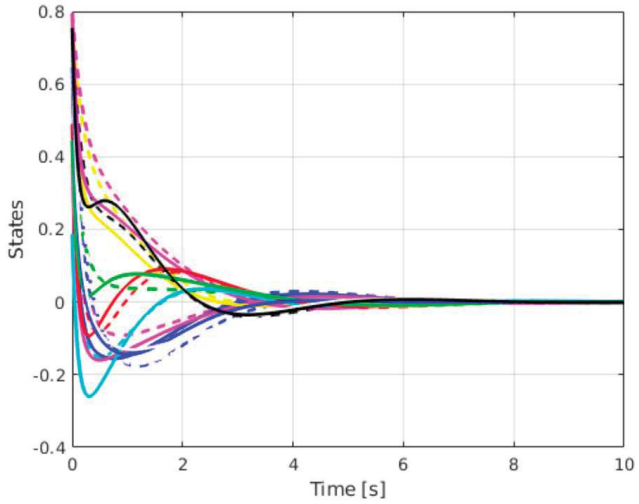

Figure 3. Evolution of system states in example 3

Example 4: This example is a 9th order unstable system taken from Davison-Maki [2]. For this system we have the A matrix,

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.2165 & -0.0356 & 0 & -0.0299 & 0 & -0.027 & 0 \\ -0.0458 & 1 & -0.0133 & 0.0004 & 0 & 0.0006 & 0 & 0.0007 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -29.81 & -0.0546 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -169 & -0.13 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -334.3 & -0.1828 \end{bmatrix}$$

and we have,

$$B = [0 \quad -1.138 \quad -0.0348 \quad 0 \quad 29.56 \quad 0 \quad 47.25 \quad 0 \quad 16.4]^T$$

$$M = diag(0.1, 0.05, 0.5, 10^{-4}, 10^{-4}, 10^{-4}, 10^{-4}, 10^{-4}, 10^{-4})$$

Also, $R = 0.1I$ and the delay time was selected as $T = 0.001$ sec. This system is controllable and observable. The evolution of the states (dashed lines) in this system along with the optimal LQR solution (solid lines) are shown in Figure 4.

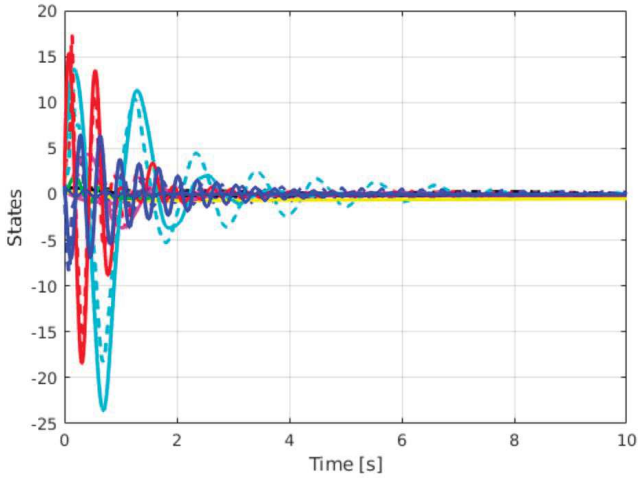

Figure 4. Evolution of system states in example 4

Example 5: This 7th order system is based on the F-16 linearized lateral dynamics and is taken from Lewis [3]. The plant matrix of this stable system A, is

$$\begin{bmatrix} -0.3220 & 0.0640 & 0.0364 & -0.9917 & 0.0003 & 0.0008 & 0 \\ 0 & 0 & 1 & 0.0037 & 0 & 0 & 0 \\ -30.6492 & 0 & -3.6784 & 0.6646 & -0.7333 & 0.1315 & 0 \\ 8.5396 & 0 & -0.0254 & -0.4764 & -0.0319 & -0.0620 & 0 \\ 0 & 0 & 0 & 0 & -20.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -20.2 & 0 \\ 0 & 0 & 0 & 57.2958 & 0 & 0 & -1 \end{bmatrix}$$

We have

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 20.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 20.2 & 0 \end{bmatrix}$$

$$M = diag\,(50 \quad 100 \quad 100 \quad 50 \quad 0 \quad 0 \quad 1)$$

and we selected $R = 0.5I$ with delay time $T = 0.1$ sec. This system is controllable and observable. The evolution of the states (dashed lines) in this system along with the optimal LQR solution (solid lines) are shown in Figure 5.

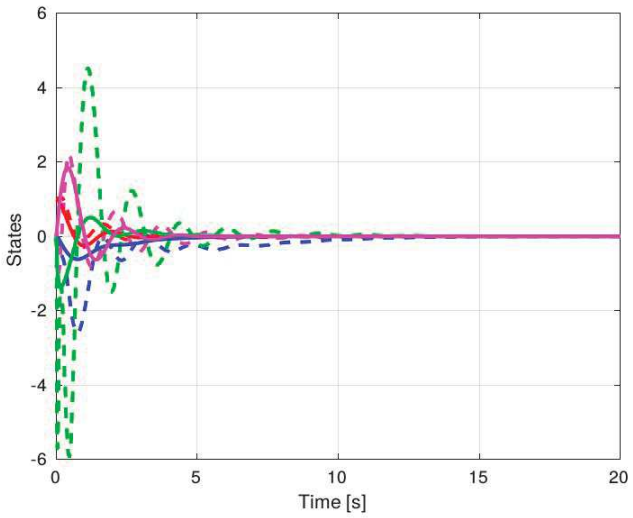

Figure 5. Evolution of system states in example 5

CONCLUSION

This paper demonstrates the application of a model-free Q-learning algorithm to estimate optimal controller settings for time-invariant linear systems without any knowledge of the system dynamics. This algorithm, in its present form, is very sensitive to initialization parameters and the fine tuning of these parameters can drive the solution close to the optimal solution. Future research efforts will include initialization techniques that can find an initial solution in the neighborhood of the optimal solution similar to graduated optimization methodologies. At the same time our future work will also focus on higher-order quadrotor models which are envisioned to become part of the IoT ecosystem. Quadrotors are always exposed to new dynamics, sensor loads and / or environments, therefore, this Q-learning algorithm is a potential solution to address changing environments and configurations of IoT devices.

ACKNOWLEDGMENTS

This work was supported in part by an NSF CAREER CPS-1750789, by NATO under grant No. SPS G5176, by ONR Minerva Initiative under grant N00014-18-1-2160 and by ONR grant N00014-18WX01381.

REFERENCES

1. A. N. Jr. Andry, E. Y. Shapiro, and J. C. Chung. 1983. Eigenstructure assignment for linear systems. *IEEE Trans. Aerospace Electron. Systems* 19 (1983).
2. E. Davison and M. Maki. 1973. The numerical solution of the matrix Riccati differential equation. *IEEE Trans. Autom. Control* 18, 1 (1973), 71–73.
3. D. Vrabie F. L. Lewis and V. L. Syrmos. 2012. *Optimal Control, 3rd Edition*. John Wiley and Sons, Hoboken, New Jersey.
4. T. Kailath. Some new algorithms for recursive estimation in constant linear systems. *IEEE Trans. Inform. Theory* (????).
5. C. Kenney and R. Leipnik. 1985. Numerical integration of the differential matrix Riccati equation. *IEEE Trans. Automat. Control* 30, 10 (1985), 962–970.
6. D. G. Lainiotis. 1976. Partitioned Riccati solutions and integration-free doubling algorithms. *IEEE Trans. Automat. Control* 21 (1976), 677–689.
7. K. G. Vamvoudakis. 2017. Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach. *System and Control Letters* 100 (2017), 14–20.