Output regulation of unknown linear systems using average cost reinforcement learning

Farnaz Adib Yaghmaie, Svante Gunnarsson and Frank L. Lewis

Adib Yaghmaie, F., Gunnarsson, S., Lewis, F. L., (2019), Output regulation of unknown linear systems using average cost reinforcement learning, *Automatica*, 110, 108549.
https://doi.org/10.1016/j.automatica.2019.108549

Output Regulation of Unknown Linear Systems using Average Cost Reinforcement Learning ⋆

Farnaz Adib Yaghmaie [a], Svante Gunnarsson [a], Frank L. Lewis [b]

[a] *Department of Electrical Engineering, Linköping University, Linköping, Sweden*

[b] *University of Texas at Arlington, Texas, USA and Northeastern University, Shenyang, China*

Abstract

In this paper, we introduce an optimal average cost learning framework to solve output regulation problem for linear systems with unknown dynamics. Our optimal framework aims to design the controller to achieve output tracking and disturbance rejection while minimizing the average cost. We derive the Hamilton-Jacobi-Bellman (HJB) equation for the optimal average cost problem and develop a reinforcement algorithm to solve it. Our proposed algorithm is an off-policy routine which learns the optimal average cost solution completely model-free. We rigorously analyze the convergence of the proposed algorithm. Compared to previous approaches for optimal tracking controller design, we elevate the need for judicious selection of the discounting factor and the proposed algorithm can be implemented completely model-free. We support our theoretical results with a simulation example.

Key words: Output Regulation, Reinforcement Learning , Linear Systems, Optimal Control

1 Introduction

Output regulation is one of the central concepts in control theory which aims at tracking and disturbance rejection [9,16,21,19,20]. Usually, the dynamics of the reference and the disturbance are combined into a single dynamical system called *exo-system* in the literature [9]. There are two general approaches to design the controller for the output regulation problem. The first approach is a feedforward method where a feedback controller is designed to stabilize the system and a feedforward controller is obtained by solving the output regulation equation to keep track of the exo-system. The second approach is to include an internal model of the exo-system in the dynamic controller and then, the control signal is a feedback from the internal state of the controller and the state of the system. Both approaches need a full knowledge of the system and the exo-system dynamics. Proportional–Integral–Derivative (PID) controllers are also widely used for practical (but not generally exact) tracking [12].

Over the past decades, Reinforcement Learning (RL) techniques are used to design adaptive/optimal schemes for control of systems with unknown dynamics [17,4]. Recently, RL techniques have been further extended to solve tracking-type problems [6,13–15,23,8]. In [22,23,8,10], suboptimal approaches are suggested where the feedforward part of the controller is obtained by dynamic inversion assuming that the dynamics is known and the feedback part is obtained by solving optimal control problems using RL techniques. These methods are suboptimal because only the feedback part of the controller is considered in the cost function. The main challenge in considering the feedforward part in the optimal controller design is that the feedforward part results in an infinite cost; because it contains a term from the exo-system (or an internal model of the exo-system) which is non-dissipating. One possible way to fix this issue is to optimize a discounted cost [13–15]. The discounting factor needs to satisfy an upper bound to ensure local asymptotic stability of the tracking error while it cannot be selected near to zero to avoid an infinite cost and a long transient response. Note that the discounted cost may not be a correct measure of the original cost, for example when the cost is the consumed energy of the system [3]. Alternatively, [6,5] suggest

⋆ This paper was not presented at any IFAC meeting. Corresponding author Farnaz Adib Yaghmaie. Tel. +46-762909978.

Email addresses: farnaz.adib.yaghmaie@liu.se (Farnaz Adib Yaghmaie), svante.gunnarsson@liu.se (Svante Gunnarsson), lewis@uta.edu (Frank L. Lewis).

a non-discounted framework for the output regulation controller design where it is required to build and evaluate tracking-type errors using information of the output matrices and such, they are not completely model-free.

To elevate the need for judicious selection of the discounting factor and to design the output regulation controller without any knowledge about the dynamics, we bring together optimal average cost, output regulation theory, and RL techniques. There are two main contributions in this paper. Our first contribution is to introduce an average cost optimization to solve the output regulation problem. This makes our formulation and results independent of the discounting factor and its selection [13–15]. Our second contribution is to propose a completely model-free online RL algorithm to solve the output regulation problem. In comparison, the RL approaches in [6,5,13] need a knowledge of the input or output matrices.

Notations: Let I and $\mathbf{0}$ denote an identity and a zero matrices with appropriate dimensions respectively. Let $\otimes$ denote the Kronecker product. Let ∇f denote the gradient of function $f(x)$ with respect to x. The (semi) positive definiteness constraint on the matrix Q is formulated as $(Q \geq 0)$, $Q > 0$. The set of all eigenvalues of a square matrix A is denoted by $Spec(A)$ and the minimum eigenvalue is denoted by $\underline{\mu}(A)$. The transpose of a matrix is denoted by the superscript $\dagger$. Consider matrix $A = [a_1, ..., a_m] \in \mathbb{R}^{n\times m}$. Then, $vec(A) = [a_1^\dagger, a_2^\dagger, ..., a_m^\dagger]^\dagger \in \mathbb{R}^{nm}$. Consider a symmetric matrix $P = [p_{ij}] \in \mathbb{R}^{n\times n}$. Then, $vecs(P) = [p_{11}, p_{12}, .., p_{1n}, p_{22}, .., p_{2n}, ..., p_{nn}]^\dagger \in \mathbb{R}^{n(n+1)/2}$. Consider a vector $x = [x_i] \in \mathbb{R}^n$. Then $vecv(x) = [x_1^2, 2x_1x_2, ..., 2x_1x_n, x_2^2, ..., 2x_2x_n, ..., x_n^2]^\dagger \in \mathbb{R}^{n(n+1)/2}$.

2 Output Regulation Problem

Consider the following dynamical system

$$\begin{aligned} \dot{x} &= Ax + Bu + Dv, && (1)\\ y &= Cx, && (2)\\ \dot{v} &= Sv, && (3)\\ w &= Fv, && (4)\\ e &= y - w = Cx - Fv, && (5) \end{aligned}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$ denote the state, the control and the output of the system, and $v \in \mathbb{R}^q$, $w \in \mathbb{R}^p$ denote the state and the output of the exo-system. Here, the exo-system (3) contains the dynamics of the reference signal to be tracked and the disturbance to be rejected [9]. It is desired to design the controller u such that the output regulation error, denoted by $e \in \mathbb{R}^p$, converges to zero. Let $v(t, v_0)$ denote the solution of (3) at time t initiated at v_0 and let $x(t, x_0, v, u)$ denote the solution of (1) at time t by the control u initiated at x_0. For simplicity, we define the augmented state of the system and the exo-system as $X(t) = \begin{bmatrix} x(t)^\dagger & v(t)^\dagger \end{bmatrix}^\dagger$. Stacking (1)-(3), the augmented system is defined as

$$\begin{aligned} \dot{X} &= \begin{bmatrix} A & D \\ \mathbf{0} & S \end{bmatrix} X + \begin{bmatrix} B \\ \mathbf{0} \end{bmatrix} u = A_a X + B_a u, \\ e &= \begin{bmatrix} C & -F \end{bmatrix} X = C_a X. \end{aligned} \tag{6}$$

Problem 1 *Consider* (1)-(5). *Design*

$$u = K_{fb}x + K_{ff}v \tag{7}$$

such that the output regulation error e *converges to an arbitrary small vicinity of zero.*

We make the following assumption regarding the dynamical systems in (1)-(6).

Assumption 1 *The pair* (A, B) *is stabilizable and* (A_a, C_a) *is detectable.*

Assumption 2 $Re(\lambda) \leq 0$, $\forall \lambda \in Spec(S)$.

Assumption 3 *The linear matrix equations*

$$\Pi S = A\Pi + B\Gamma + D, \quad C\Pi - F = \boldsymbol{0}, \tag{8}$$

are solved by some $\Pi \in \mathbb{R}^{n\times q}$ *and* $\Gamma \in \mathbb{R}^{m\times q}$.

Theorem 1 ([9]) *Let Assumptions 1-2 hold and assume that* K_{fb} *is selected such that* $A + BK_{fb}$ *is strictly stable. Then, the controller*

$$u = K_{fb}x + (\Gamma - K_{fb}\Pi)v \tag{9}$$

solves Problem 1 if and only if Assumption 3 holds.

Introduce $\xi = x - \Pi v$. Using (8), we have

$$\begin{aligned} \dot{\xi} =& (A + BK_{fb})\xi, \; u = K_{fb}\xi + \Gamma v, \\ e =& C\xi. \end{aligned} \tag{10}$$

Hence, if the controller (9) solves the output regulation problem, it also results $\xi \to \mathbf{0}$ and vice versa.

3 Optimal Output Regulation Problem

Consider the following quadratic performance index

$$V(X(t), u(t)) = \int_t^{+\infty} r(X(\tau), u(\tau))\, d\tau \tag{11}$$

where $r(X, u) = e^\dagger Qe + u^\dagger Ru$ is the *running cost* with $Q > 0$ and $R > 0$. The performance index in (11) is

called the *total (not-discounted) value function* and it can be optimized only if the exo-system (3) is asymptotically stable; otherwise, the total value function becomes infinite because of the feedforward controller from the exo-system in (3). The finiteness of the value function is a key in Bellman principle of optimality [2], and a problem with an infinite value function is not well defined. Alternatively, one can consider optimizing of the average value function instead of (11). Define the *average value function* as

$$V_a(X(t),u(t)) = \lim_{T\to\infty} \frac{1}{T}\int_t^{t+T} r(X(\tau),u(\tau))\,d\tau \quad (12)$$

which is bounded if Assumption 2 holds. If we set $t=0$ in (12) then, we obtain the *average cost*

$$J_a(X_0,u(t)) = \lim_{T\to\infty} \frac{1}{T}\int_0^{T} r(X(\tau),u(\tau))\,d\tau. \quad (13)$$

In this paper, we are interested in finding control policies of the form $u(t)=u(x(t),v(t))$ from the set of *average cost admissible policies.*

Definition 1 ***(Average cost admissible policy)*** *Consider dynamical system* (1)-(5)*. The policy (feedback/feedforward controller) $u(x,v)$ is Average Cost Admissible (ACA), denoted by $u(x,v)\in\mathcal{U}_{ACA}$, if it satisfies the following: (a) it is continuous, (b) $u(\mathbf{0})=\mathbf{0}$, (c) $\dot{x}=Ax+Bu(x,\mathbf{0})$ is asymptotically stable, and (d) the average cost* (13) *is bounded.*

Now, we compare the ACA with the Classical definition of Admissible policy (CA) in [1]. Properties (a)-(b) are the same for both the ACA and CA. Property (c) concerns the stabilizability of the dynamical system. Since the exo-system in (3) is not stabilizable, only stabilizability of (1) is considered in ACA. Regarding (d), a CA policy makes the value function (11) finite while an ACA policy makes the average value function (12) finite. Finally, we note that a CA policy is always an ACA policy but the converse is not true. For example, the output regulation controller (9) is ACA but not CA because the average value function (12) is finite but the total value function (11) becomes infinite by (9) because of the term from the exo-system, i.e. $(\Gamma - K_{fb}\Pi)v$, which is non-dissipating.

Problem 2 *Consider* (1)-(5)*. Find an optimal ACA policy $u^*=K^*X(t)$ to bring the output regulation error e to an arbitrary small vicinity of zero by minimizing* (12)*.*

The *optimal average value function* associated with the optimal policy $u^*=K^*X(t)$ is denoted by

$$V_a^*(X(t)) = \lim_{T\to\infty} \frac{1}{T}\int_t^{t+T} r(X(\tau),u^*)\,d\tau, \quad (14)$$

and the *optimal average cost* is denoted by λ^*

$$\lambda^* = J_a(X_0,u^*) = \lim_{T\to\infty} \frac{1}{T}\int_0^{T} r(X(\tau),u^*)\,d\tau. \quad (15)$$

For an output regulation problem satisfying Assumptions 1-3, one can show that the optimal average cost depends on the initial state of the exo-system.

Theorem 2 *Consider* (1)-(5)*. Let Assumptions 1-3 hold. Assume that K_{fb} is selected such that $A+BK_{fb}$ is strictly stable such that the output regulation problem is solvable by the controller $u=K_{fb}x+(\Gamma-K_{fb}\Pi)v = K_{fb}\xi+\Gamma v$. Then, the average cost $\lambda=J_a(x_0,v_0,u)$ is expressed in a quadratic form*

$$\lambda(\nu)=\nu^\dagger M\nu,\ M=\lim_{T\to\infty}\frac{1}{T}\int_0^T (e^{S\tau})^\dagger\Gamma^\dagger R\Gamma e^{S\tau}\,d\tau, \quad (16)$$

where ν can be v_0 or any other point on the trajectory of v initiated at v_0; i.e. $v(t,v_0),\ 0\le t<\infty$.

PROOF. Let $\lambda=J_a(x_0,v_0,u)$. Since $A+BK_{fb}$ is stable $\xi\to\mathbf{0}$ and $e\to\mathbf{0}$. Hence, the terms containing e and ξ in the running cost, produce a bounded total cost and a zero average cost in the view of $T\to\infty$. As a result, the average cost is given by $\lambda=\lim_{T\to\infty}\frac{1}{T}\int_0^T v^\dagger\Gamma^\dagger R\Gamma v\,d\tau$. The solution for v in (3) initiated at v_0 is $v(t)=e^{St}v_0$. Consider the time $0\le t<\infty$. Then,

$$\lambda=\lim_{T\to\infty}\frac{1}{T}[\int_0^t v^\dagger\Gamma^\dagger R\Gamma v\,d\tau+\int_t^T v^\dagger\Gamma^\dagger R\Gamma v\,d\tau].$$

Since the time t is finite, $\lim_{T\to\infty}\frac{1}{T}\int_0^t v^\dagger\Gamma^\dagger R\Gamma v d\tau=0$ and the average cost reads

$$\lambda=\lim_{T\to\infty}\frac{1}{T}\int_t^T v_0^\dagger(e^{S\tau})^\dagger\Gamma^\dagger R\Gamma e^{S\tau}v_0\,d\tau.$$

Changing the integral variable $\rho=\tau-t$, the average cost reads $\lambda=v_0^\dagger(e^{St})^\dagger Me^{St}v_0$. Hence, one can consider $\nu\equiv v(t,v_0),\ 0\le t<\infty$.

3.1 Solution to the Optimal Average Cost Problem

Lemma 1 *Minimizing the average value function* (12) *is equivalent to minimizing the following infinite-horizon optimal control problem*

$$V_\infty(X(t),u(t)) = \int_t^{+\infty}(r(X(\tau),u(\tau))-\lambda^*)\,d\tau \quad (17)$$

in the sense that an optimal control to (17) *is also optimal for* (12).

PROOF.

By (15), for any u, we have $\int_0^T (r(X(\tau),u(\tau))-\lambda^*)d\tau \geq 0$ and the equality holds for u^*. Hence, (17) achieves its minimum value with u^* which is also optimal for (12).

Using Lemma 1, the Hamilton-Jacobi-Bellman (HJB) equation associated with the dynamics (1)-(3) and the value function (17) is given by

$$\lambda^* = \min_{u\in\mathcal{U}_{ACA}} [r(X(t),u(t)) + \nabla V_\infty^{*\dagger}(A_a X + B_a u)]. \tag{18}$$

Then, the following policy

$$u^* = -\frac{1}{2}R^{-1}B_a^\dagger \nabla V_\infty^*(x) \tag{19}$$

optimizes the value function (17) and by Lemma 1, it also minimizes the average value function (12). Next, we prove that V_∞^* is quadratic.

Theorem 3 *Consider* (1)-(5). *Let Assumptions 1-3 hold. Assume that the output regulation problem is solvable by the controller* $u^* = K_{fb}x + (\Gamma - K_{fb}\Pi)v$. *Then, the infinite horizon value function* V_∞^* *is quadratic*

$$V_\infty^* = X^\dagger(t)P^*X(t) = \begin{bmatrix} x^\dagger & v^\dagger \end{bmatrix} \begin{bmatrix} P_x^* & P_{xv}^* \\ P_{xv}^{*\dagger} & P_v^* \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix}, \tag{20}$$

where $P_x^* > 0$ *and* P_{xv}^* *uniquely satisfy*

$$\begin{aligned} &A^\dagger P_x^* + P_x^* A - P_x^* B R^{-1} B^\dagger P_x^* + C^\dagger Q C = \boldsymbol{0}, \\ &P_{xv}^* S + (A - BR^{-1}B^\dagger P_x^*)^\dagger P_{xv}^* + P_x^* D - C^\dagger Q F = \boldsymbol{0}, \end{aligned} \tag{21}$$

and P_v^* *satisfies*

$$\begin{aligned} &S^\dagger P_v^* + P_v^* S - P_{xv}^{*\dagger} B R^{-1} B^\dagger P_{xv}^* \\ &+ P_{xv}^{*\dagger} D + D^\dagger P_{xv}^* + F^\dagger Q F - M = \boldsymbol{0}. \end{aligned} \tag{22}$$

PROOF. Let $K_{ff} = \Gamma - K_{fb}\Pi$. Using $u^* = K_{fb}x + K_{ff}v$, the average value function reads

$$V_\infty^*(X(t),u(t)) = \int_t^\infty (X^\dagger \bar{Q} X - \lambda^*)\, d\tau, \tag{23}$$

$$\bar{Q} = \begin{bmatrix} C^\dagger Q C + K_{fb}^\dagger R K_{fb} & -C^\dagger Q F + K_{fb}^\dagger R K_{ff} \\ -F^\dagger Q C + K_{ff}^\dagger R K_{fb} & F^\dagger Q F + K_{ff}^\dagger R K_{ff} \end{bmatrix}.$$

The solutions for the differential equations (1) and (3) using the policy $u^* = K_{fb}x + K_{ff}v$ are

$$\begin{aligned} x(t+T,x(t),v(t),u^*) &= e^{(A+BK_{fb})T}x(t) \\ &+ (\int_0^T e^{(A+BK_{fb})(T-\tau)}(BK_{ff}+D)e^{S\tau}\, d\tau)v(t) \\ &= L_x(T)x(t) + L_{xv}(T)v(t), \\ v(t+T,v(t)) &= e^{ST}v(t) = L_v(T)v(t). \end{aligned}$$

Using the above solutions in (23) results in (20) with

$$\begin{aligned} P_x^* &= \int_t^\infty L_x^\dagger (C^\dagger Q C + K_{fb}^\dagger R K_{fb}) L_x\, d\tau, \\ P_{xv}^* &= \int_t^\infty \{L_x^\dagger (C^\dagger Q C + K_{fb}^\dagger R K_{fb}) L_{xv} \\ &+ L_x^\dagger (-C^\dagger Q F + K_{fb}^\dagger R K_{ff}) L_v\}\, d\tau, \\ P_v^* &= \int_t^\infty \{L_{xv}^\dagger (C^\dagger Q C + K_{fb}^\dagger R K_{fb}) L_{xv} + L_v^\dagger (F^\dagger Q F \\ &- \Gamma^\dagger R K_{fb}\Pi - \Pi^\dagger K_{fb}^\dagger R\Gamma + \Pi^\dagger K_{fb}^\dagger R K_{fb}\Pi) L_v \\ &+ 2L_{xv}^\dagger (-C^\dagger Q F + K_{fb}^\dagger R K_{ff}) L_v\}\, d\tau, \end{aligned} \tag{24}$$

where we have used (16) to derive the last equation. Till now, we have proved that V_∞^* is quadratic. Next, we show that the relations in (21)-(22) hold. Using the quadratic form (20), the HJB (18) reads

$$\begin{aligned} \begin{bmatrix} x^\dagger & v^\dagger \end{bmatrix} &(A_a^\dagger P^* + P^* A_a - P^* B_a R^{-1} B_a^\dagger P^* + \\ &\begin{bmatrix} C^\dagger Q C & -C^\dagger Q F \\ -F^\dagger Q C & F^\dagger Q F \end{bmatrix}) \begin{bmatrix} x \\ v \end{bmatrix} - v^\dagger M v = \boldsymbol{0}. \end{aligned}$$

Substituting (6) in the above, (21)-(22) are concluded. Note that (21) has a unique positive definite solution P_x^* based on Assumption 1. Moreover, the solution P_{xv}^* is unique because $A - BR^{-1}B^\dagger P_x^*$ and $-S$ do not share any eigenvalues (see Theorem 4.4.6 of [7]).

3.2 Main result

In this subsection, we prove that the controller (19) which is obtained by minimizing the average cost, solves the output regulation problem.

Theorem 4 *Consider* (1)-(5) *and let Assumptions 1-3 hold. Let* $(V_\infty^*, u^*, \lambda^*)$ *form a solution to* (18). *Then, the control* u^* *in* (19) *solves Problem 2 and the output regulation error* e *is Uniformly Ultimately Bounded (UUB) with bound* $\|e\| \leq \sqrt{\lambda^*/\underline{\mu}(Q)}$.

PROOF. Since $(V_\infty^*, u^*, \lambda^*)$ form a solution to (18), by Lemma 1, u^* also minimizes (12). It remains to show

Algorithm 1 Model-based routine for average cost Learning

1: **Initialize:** $u^{(0)} = K^{(0)}X \in \mathcal{U}_{ACA}$, $k = 0$.
2: **repeat**
3: Apply $u^{(k)}$ and collect the required information.
4: Given $u^{(k)}$, find $P^{(k)}$, $\lambda^{(k)}$ from

$$\begin{aligned}\lambda^{(k)} =& e^\dagger Qe + u^{(k)\dagger} Ru^{(k)} \\ &+ 2X^\dagger P^{(k)\dagger}(A_a X + B_a u^{(k)}).\end{aligned} \tag{26}$$

5: Improve the policy by

$$u^{(k+1)} = -R^{-1}B_a^\dagger P^{(k)} X. \tag{27}$$

6: **until** $\|P^{(k)} - P^{(k-1)}\| + |\lambda^{(k)} - \lambda^{(k-1)}| < \epsilon_1$

that u^* solves the output regulation problem. We consider V_∞^* as the candidate Lyapunov function to prove stability of $\xi = x - \Pi v$ in (10) with u^*. Note that V_∞^* is a valid Lyapunov function for ξ because $V_\infty^*(\xi \equiv \mathbf{0}) = \int_t^\infty (v^\dagger \Gamma^\dagger R \Gamma v - \lambda^*)\, d\tau = 0$ and $V_\infty^*(\xi) \geq 0$. By (18), the time derivative of V_∞^* reads

$$\dot{V}_\infty^* = -(e^\dagger Qe + u^{*\dagger} Ru^* - \lambda^*). \tag{25}$$

Hence, $\dot{V}_\infty^* < 0$ if $e^\dagger Qe > \lambda^*$ which is guaranteed by $\|e\|^2 \underline{\mu}(Q) > \lambda^*$. As a result, the error is UUB [11] and $\|e\| \leq \sqrt{\lambda^*/\underline{\mu}(Q)}$. One can make $\|e\|$ arbitrary small by selecting Q sufficiently large.

4 Reinforcement Learning Frameworks for Optimal Average Cost

In this section, we first present a model-based approach to solve the optimal average cost problem and then, we propose an off-policy IRL routine which learns the optimal solution without any knowledge about the dynamics.

4.1 Model-based Algorithm

Here, we give Algorithm 1 which is an iterative model-based routine to solve the optimal control problem in (18)-(19). Algorithm 1 is essentially a Netwon's iteration to solve (18) which is discussed [15] and it is modified according to the average cost formulation. In Algorithm 1, ϵ_1 is the convergence threshold.

4.2 Model-free Off-Policy Integral Reinforcement Algorithm

Now, we present an off-policy algorithm to learn the optimal average cost problem completely model-free. The

Algorithm 2 Off-policy IRL for average cost Learning

1: **Initialize:** $u^{(0)} = K^{(0)}X \in \mathcal{U}_{ACA}$, $k = 0$.
2: **repeat**
3: Apply $u = K^{(k)}X + n$ and collect the required information at N sample times.
4: Find $P^{(k)}$, $\lambda^{(k)}$, $K^{(k+1)}$ from

$$\begin{aligned}X^\dagger(t)P^{(k)}X(t) =& \int_t^{t+\delta t} (e^\dagger Qe + u^{(k)\dagger} Ru^{(k)})\, d\tau \\ &+ X^\dagger(t+\delta t)P^{(k)}X(t+\delta t)) - \delta t \lambda^{(k)} \\ &+ 2\int_t^{t+\delta t} (u - u^{(k)})^\dagger R K^{(k+1)} X\, d\tau.\end{aligned} \tag{29}$$

5: **until** $\|P^{(k)} - P^{(k-1)}\| + |\lambda^{(k)} - \lambda^{(k-1)}| + \|K^{(k+1)} - K^{(k)}\| < \epsilon_2$

idea is to apply a behavioral policy $u = u^{(k)} + n$ while learning a sequence of controllers $u^{(k)}$ which converge to the optimal control u^*. Note the behavioral policy u differs from $u^{(k)}$ by an exponentially decreasing probing noise n. Examples are given in [18]. Using the behavioral policy, the system in (6) can be written as

$$\dot{X} = A_a X + B_a u^{(k)} + B_a(u - u^{(k)}). \tag{28}$$

Differentiating $V_\infty^{(k)}$ along with (28) reads

$$\dot{V}_\infty^{(k)} = \nabla V_\infty^{(k)\dagger}(A_a X + B_a u^{(k)}) + \nabla V_\infty^{(k)\dagger} B_a(u - u^{(k)}).$$

Integrating $\dot{V}_\infty^{(k)}$ along $[t, t+\delta t]$, using (26)-(27), we have

$$\begin{aligned}&V_\infty^{(k)}(X(t+\delta t)) - V_\infty^{(k)}(X(t)) \\ &= \int_t^{t+\delta t} \nabla V_\infty^{(k)\dagger}(A_a X + B_a u^{(k)} + B_a(u - u^{(k)}))d\tau \\ &= \int_t^{t+\delta t} (\lambda^{(k)} - r(X, u^{(k)}))d\tau \\ &\quad - 2\int_t^{t+\delta t} X^\dagger K^{(k+1)\dagger} R(u - u^{(k)})d\tau.\end{aligned}$$

By considering the quadratic form $V_\infty^{(k)}(X(t)) = X^\dagger(t)P^{(k)}X(t)$, (29) is concluded. This equation can be used to find $P^{(k)}$, $K^{(k+1)}$ and $\lambda^{(k)}$ simultaneously. The off-policy IRL routine is summarized in Algorithm 2, where $\epsilon_2 > 0$ is the convergence threshold. The following theorem concerns monotonicity and convergence of Algorithm 2

Theorem 5 *Consider* (1)-(5). *Let Assumptions 1-3 hold. Let* $\{V_\infty^{(k)}, \lambda^{(k)}, u^{(k)}\}_{k=1}^\infty$ *satisfy* (29) *in Algorithm 2. Then, (i) The improved policy* (29) *is ACA. (ii) Let* $V_a^{(k)} = V_a(x(t), u^{(k)}(t))$. *Then* $V_a^{(k+1)} \leq V_a^{(k)}$. *(iii) The*

sequence of $\{V_\infty^{(k)}, \lambda^{(k)}, u^{(k)}\}_{k=1}^\infty$ *uniformly converges to* $\{V_\infty^*, \lambda^*, u^*\}$.

PROOF.

(i) We show that properties (a-d) in Definition 1 hold for $u^{(k+1)} = K^{(k+1)}X$. (a-b) are immediately concluded from $u^{(k+1)} = K^{(k+1)}X$. (c) Set $v \equiv \mathbf{0}$ and $n \equiv \mathbf{0}$. Since (1) is stabilizable, then $\lambda^{(k)} = 0$. Let $\bar{V}_\infty = V_\infty^{(k)}(v \equiv \mathbf{0})$. Since $u^{(k)}$ and $V_\infty^{(k)}$ satisfy (29), one has

$$\nabla \bar{V}_\infty^{(k)\dagger} Ax = -r(x, u^{(k)}) - \nabla \bar{V}_\infty^{(k)\dagger} Bu^{(k)}.$$

Consider the Lyapunov function $\bar{V}_\infty^{(k)}$ for the system (1) driven by the policy $u^{(k+1)} = u^{(k+1)}(x, \mathbf{0})$. Using the above equation, the time derivative of $\bar{V}_\infty^{(k)}$ reads

$$\begin{aligned}\dot{\bar{V}}_\infty^{(k)}(x, u^{(k+1)}) &= \nabla \bar{V}_\infty^{(k)\dagger} Ax + \nabla \bar{V}_\infty^{(k)\dagger} Bu^{(k+1)} \\ &= -r(x, u^{(k)}) - \nabla \bar{V}_\infty^{(k)\dagger} B(u^{(k)} - u^{(k+1)}). \quad (30)\end{aligned}$$

It has been shown in Lemma 1 of [15] that the off-policy algorithm produces the same sequence of value functions and improved policies as Algorithm 1 for learning a discounted cost. A similar proof can be brought also for the average cost learning. As a result, by using $K^{(k+1)} = -R^{-1}B_a^\dagger P^{(k)}$ in (27) and completing the squares, we have

$$\begin{aligned}&\dot{\bar{V}}_\infty^{(k)}(x, u^{(k+1)}) \\ &\quad = -r(x, u^{(k)}) + 2u^{(k+1)\dagger} R(u^{(k)} - u^{(k+1)}) \\ &\quad = -x^\dagger Q x - u^{(k+1)\dagger} R u^{(k+1)} \\ &\qquad - (u^{(k)} - u^{(k+1)})^\dagger R(u^{(k)} - u^{(k+1)}) < 0\end{aligned}$$

which shows that $\dot{x} = Ax + Bu^{(k+1)}(x, \mathbf{0})$ is asymptotically stable. (d) Since both x and v are bounded $u^{(k+1)}$ is also bounded. Hence, the average cost in (13) is finite.

(ii) Evaluate two Lyapunov candidate functions $V_\infty^{(k+1)}$ and $V_\infty^{(k)}$ along the system trajectory by policy $u^{(k+1)}$

$$\begin{aligned}&V_\infty^{(k)} - V_\infty^{(k+1)} \\ &= \int_t^{+\infty} \{\dot{V}_\infty^{(k+1)}(X, u^{(k+1)}) - \dot{V}_\infty^{(k)}(X, u^{(k+1)})\}\, d\tau \\ &= \int_t^{+\infty} \{\nabla V_\infty^{(k+1)\dagger}(A_a + B_a u^{(k+1)}) \\ &\qquad - \nabla V_\infty^{(k)\dagger}(A_a + B_a u^{(k+1)})\}\, d\tau.\end{aligned}$$

Because we evaluate the system trajectory by policy $u^{(k+1)}$, we set $n \equiv \mathbf{0}$. Then, using the fact that $\{V_\infty^{(k)}, \lambda^{(k)}, u^{(k)}\}_{k=1}^\infty$ satisfy (29)

$$\begin{aligned}&V_\infty^{(k)} - V_\infty^{(k+1)} = \int_t^{+\infty} \{-u^{(k+1)\dagger} R u^{(k+1)} + \lambda^{(k+1)} \\ &+ u^{(k)\dagger} R u^{(k)} - \lambda^{(k)} + \nabla V_\infty^{(k)\dagger} B_a (u^{(k)} - u^{(k+1)})\}\, d\tau.\end{aligned}$$

Using (27), the above equation reads

$$\begin{aligned}&V_\infty^{(k)} - V_\infty^{(k+1)} + \int_t^{+\infty} (\lambda^{(k)} - \lambda^{(k+1)})\, d\tau \qquad (31) \\ &\quad = \int_t^{+\infty} (u^{(k)} - u^{(k+1)})^\dagger R(u^{(k)} - u^{(k+1)})\, d\tau.\end{aligned}$$

By (12) and (17) we have

$$\lim_{T\to\infty} V_\infty^{(k)} + \int_t^{t+T} \lambda^{(k)}\, d\tau = \lim_{T\to\infty} T V_a^{(k)}. \quad (32)$$

Using the aforementioned expression, (31) reads

$$\begin{aligned}&V_a^{(k)} - V_a^{(k+1)} = \\ &\lim_{T\to\infty} \frac{1}{T} \int_t^{+\infty} (u^{(k)} - u^{(k+1)})^\dagger R(u^{(k)} - u^{(k+1)})\, d\tau \\ &\qquad \geq 0. \qquad (33)\end{aligned}$$

Hence, $\{V_a^{(k)}\}_{k=1}^\infty$ is a decreasing sequence and it is lower bounded by V_a^*.

(iii) Since $V_a^{(k)}$ is continuous in X and by the result in part (ii), $\nabla V_a^{(k)} \to \nabla V_a^*$ and by (32), $\nabla V_\infty^{(k)} \to \nabla V_\infty^*$. As a result $u^{(k)} \to u^*$. The triple $\{V_\infty^{(k)}, \lambda^{(k)}, u^{(k)}\}_{k=1}^\infty$ satisfies (29). Because of pointwise convergence of $\nabla V_\infty^{(k)}$ and $u^{(k)}$ to their optimal values, the pointwise convergence of $\lambda^{(k)}$ and $V_\infty^{(k)}$ to λ^* and V_∞^* is also concluded.

4.3 Least-Square Implementation of Algorithm 2

Here, we discuss the Least-Squares (LS) implementation of Algorithm 2. Equation (29) reads

$$\theta(t)^\dagger W^{(k)} = \phi(t), \quad (34)$$

$$W^{(k)} = \begin{bmatrix} vecs(P^{(k)}) \\ vec(K^{(k+1)}) \\ \lambda^{(k)} \end{bmatrix}, \ \phi(t) = \int_t^{t+\delta t} r(X, u^{(k)})\, d\tau,$$

$$\theta(t) = \begin{bmatrix} vecv(X(t)) - vecv(X(t+\delta t)) \\ -2\int_t^{t+\delta t} X \otimes R(u - u^{(k)})\, d\tau \\ \delta t \end{bmatrix}.$$

Hence, a least square estimation of $W^{(k)}$ is given by

$$W^{(k)} = (\Theta\Theta^\dagger)^{-1}\Theta\Phi, \tag{35}$$
$$\Theta = [\theta(t_1), ..., \theta(t_N)],\ \Phi = [\phi(t_1), ..., \phi(t_N)]^\dagger,$$

with $N \geq (n+q)(n+q+1)/2 + m(n+q) + 1$. After convergence of Algorithm 2, one can estimate the kernel of the average cost M by selecting random initial conditions v_{0i} for the exo-system and measuring the associated average costs $\lambda(v_{0i})$ for $N_M \geq q(q+1)/2$ times. Then, a least square estimation of M can be made.

5 Simulation Results

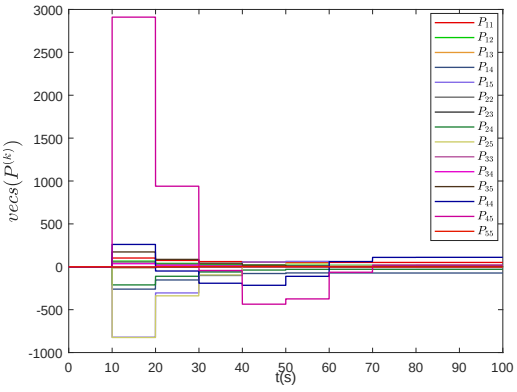
Consider an output regulation problem for the F16 aircraft system with the following dynamics [15]

$$\dot{x} = \begin{bmatrix} -1.019 & 0.905 & -0.002 \\ 0.822 & -1.077 & -0.176 \\ 0 & 0 & -1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ 5 \end{bmatrix} u + \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} v,$$
$$y = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} x, \tag{36}$$
$$\dot{v} = \begin{bmatrix} 0 & 1 \\ -0.01 & 0 \end{bmatrix} v, \quad w = \begin{bmatrix} 1 & 0 \end{bmatrix} v. \tag{37}$$

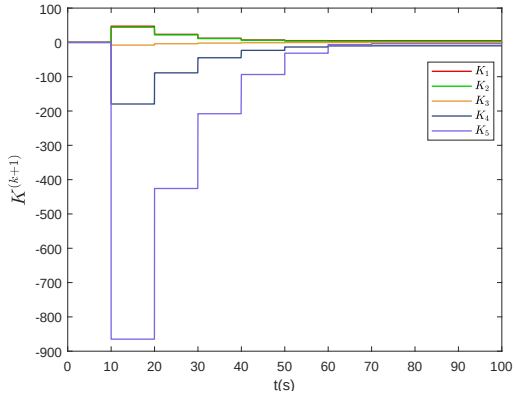
Let $Q = 100$, $R = 1$. The analytical solution to the output regulation problem using full model information is summarized in Table 1.

We use Algorithm 2 to solve this output regulation problem completely model-free. We set $\delta t = 0.1$, $N = 100$, $K^{(0)} = \mathbf{0}$. We randomize the state of the system after each 5 samples to ensure that we have enough independent samples for the least square estimation in (35). We select the probing noise as $n = 0.01e^{-0.1t}\sin(t)$. From a practical point of view, the behavioral policy is applied to (36) and the information of the system ($X(t)$, $X(t+\delta t)$ and the integral in (29)) is recorded at N sampled times. This information is then used to obtain the improved controller gain by Algorithm 2.

Algorithm 2 converges after 9 iterations and after convergence, we estimate M using 100 different initial conditions for the exo-system. Figure 1 shows the evolution of $vecs(P^{(k)})$ and $K^{(k+1)}$, and the converged values are reported in Table 1. From Table 1, we can see that $P_x^{(9)}$, $P_{xv}^{(9)}$ and the feedback gain converge to the same value obtained analytically but the feedforward gain and $\hat{M}$ differ slightly from the analytical solutions. We use the converged optimal controller $u^{(10)} = K^{(10)}X$ to control the system in (36)-(37) for $t = 100\ s$. From Fig. 2, we can see that using the converged optimal controller, the F16 aircraft (36) successfully tracks the exo-system



(a) The weight $vecs(P^{(k)})$



(b) The controller gain $K^{(k+1)}$

Fig. 1. The evolution of $vecs(P^{(k)})$ and $K^{(k+1)}$ during learning

(37) without knowing the dynamics models and the average cost of the tracking is given by $\lambda = v_0^\dagger M v_0$ where v_0 is initial state of the exo-system.

Table 1. Solutions by the analytical method and Algorithm 2. $\diamond$ indicates an unspecified value.

Analytical method						
$vecs(P_x^*) =$	[50.61	17.26	-1.01	9.12	-0.68	0.07]
$vec(P_{xv}^{*\dagger}) =$	[-72.61	-7.59	-29.41	-5.71	2.00	0.75]
$vecs(P_v^*) =$	[$\diamond$	$\diamond$	$\diamond$]			
$K^* =$	[5.03	3.41	-0.33	-10.0	-3.74]	
$vecs(M) =$	[0.11	0	10.25]			
Algorithm 2						
$vecs(P_x^{(9)}) =$	[50.61	17.26	-1.01	9.12	-0.68	0.07]
$vec(P_{xv}^{(9)\dagger}) =$	[-72.61	-7.59	-29.41	-5.71	2.00	0.75]
$vecs(P_v^{(9)}) =$	[110.24	19.96	-1.10]			
$K^{(10)} =$	[5.03	3.41	-0.33	-9.98	-3.77]	
$vecs(\hat{M}) =$	[0.18	0.02	10.53]			

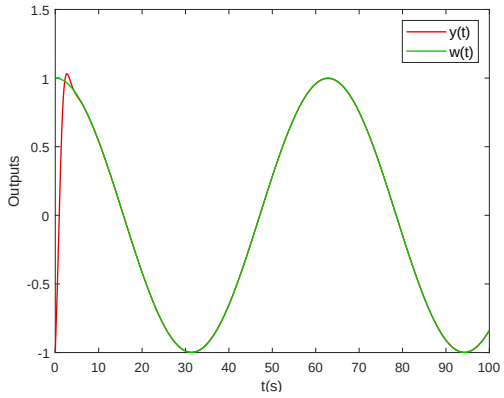


Fig. 2. y and w using $K^{(9)}$

6 Conclusion and Future Works

In this paper, we have suggested the theory of optimal average cost learning for output regulation controller design of linear systems. We have developed a completely model-free online integral reinforcement learning algorithm to solve this problem. Important features of our RL algorithm for output regulation controller design are that a discounting factor is not needed, and the off-policy algorithm is completely model-free. Our future works will focus on extending these results when the control constraints appear and when the system is time-variant.

ACKNOWLEDGMENT

Farnaz Adib Yaghmaie is supported by the Vinnova Competence Center LINK-SIC and by the Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP). Svante Gunnarsson is supported by the Vinnova Competence Center LINK-SIC

References

[1] Randal W Beard, George N Saridis, and John T Wen. Approximate solutions to the time-invariant Hamilton-Jacobi-Bellman equation. *Journal of Optimization theory and Applications*, 96(3):589–626, 1998.

[2] Richard Bellman. *Dynamic programming*. Courier Corporation, 1958.

[3] Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

[4] Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming: an overview. In *Proceedings of the 34th IEEE Conference on Decision and Control*, volume 1, pages 560–564. IEEE Publ. Piscataway, NJ, 1995.

[5] Weinan Gao and Zhong-Ping Jiang. Global optimal output regulation of partially linear systems via robust adaptive dynamic programming. *IFAC-PapersOnLine*, 48(11):742–747, 2015.

[6] Weinan Gao and Zhong-Ping Jiang. Adaptive dynamic programming and adaptive optimal output regulation of linear systems. *IEEE Transactions on Automatic Control*, 61(12):4164–4169, 2016.

[7] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 1990.

[8] Yuzhu Huang and Derong Liu. Neural-network-based optimal tracking control scheme for a class of unknown discrete-time nonlinear systems using iterative ADP algorithm. *Neurocomputing*, 125:46–56, 2014.

[9] Alberto Isidori. *Nonlinear control systems*. Springer Science & Business Media, 2013.

[10] Rushikesh Kamalapurkar, Huyen Dinh, Shubhendu Bhasin, and Warren E Dixon. Approximate optimal trajectory tracking for continuous-time nonlinear systems. *Automatica*, 51:40–48, 2015.

[11] Hassan K Khalil. *Noninear Systems*, volume 2. 1996.

[12] Bin Li, Kok Lay Teo, Cheng-Chew Lim, and Guang Ren Duan. An optimal PID controller design for nonlinear constrained optimal control problems. *Discrete and Continuous Dynamical Systems–Series B*, 16(4):1101–1117, 2011.

[13] Hamidreza Modares and Frank L Lewis. Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning. *IEEE Transactions on Automatic Control*, 59(11):3051–3056, 2014.

[14] Hamidreza Modares and Frank L Lewis. Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning. *Automatica*, 50(7):1780–1792, 2014.

[15] Hamidreza Modares, Frank L Lewis, and Zhong-Ping Jiang. H_∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning. *IEEE transactions on neural networks and learning systems*, 26(10):2550–2562, 2015.

[16] Eduardo D Sontag. Adaptation and regulation with signal detection implies internal model. *Systems & control letters*, 50(2):119–126, 2003.

[17] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[18] Farnaz Adib Yaghmaie and David J Braun. Reinforcement learning for a class of continuous-time input constrained optimal control problems. *Automatica*, 99:221–227, 2019.

[19] Farnaz Adib Yaghmaie, Frank L Lewis, and Rong Su. Output regulation of linear heterogeneous multi-agent systems via output and state feedback. *Automatica*, 67:157–164, 2016.

[20] Farnaz Adib Yaghmaie, Rong Su, Frank L Lewis, and Sorin Olaru. Bipartite and cooperative output synchronizations of linear heterogeneous agents: A unified framework. *Automatica*, 80:172–176, 2017.

[21] Farnaz Adib Yaghmaie, Rong Su, Frank L Lewis, and Lihua Xie. Multi-party consensus of linear heterogeneous multi-agent systems. *IEEE Transactions on Automatic Control*, 60(11):5578–5589, 2017.

[22] Huaguang Zhang, Lili Cui, Xin Zhang, and Yanhong Luo. Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method. *IEEE Transactions on Neural Networks*, 22(12):2226–2236, 2011.

[23] Huaguang Zhang, Qinglai Wei, and Yanhong Luo. A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(4):937–942, 2008.