A learning-based optimal tracking controller for continuous linear systems with unknown dynamics: Theory and case study

Jingren Zhang¹, Qingfeng Wang¹ and Tao Wang²

Abstract
In this article, a novel continuous-time optimal tracking controller is proposed for the single-input-single-output linear system with completely unknown dynamics. Unlike those existing solutions to the optimal tracking control problem, the proposed controller introduces an integral compensation to reduce the steady-state error and regulates the feedforward part simultaneously with the feedback part. An augmented system composed of the integral compensation, error dynamics, and desired trajectory is established to formulate the optimal tracking control problem. The input energy and tracking error of the optimal controller are minimized according to the objective function in the infinite horizon. With the application of reinforcement learning techniques, the proposed controller does not require any prior knowledge of the system drift or input dynamics. The integral reinforcement learning method is employed to approximate the $Q$-function and update the critic network on-line. And the actor network is updated with the deterministic learning method. The Lyapunov stability is proved under the persistence of excitation condition. A case study on a hydraulic loading system has shown the effectiveness of the proposed controller by simulation and experiment.

Keywords
Optimal tracking control, reinforcement learning techniques, integral compensation, completely unknown dynamics

Date received: 16 January 2020; accepted: 4 March 2020

Introduction
Accurate tracking control has drawn great research interests in a number of application fields.¹-³ Optimal control deals with problems of minimizing the prescribed objective function in the infinite or finite horizon. Traditional optimal control for linear systems solves the algebraic Riccati equation (ARE) off-line.⁴,⁵ The optimal control policy is regulated as a state feedback according to the gradient of the value function.⁶ However, this kind of controller may suffer from steady-state error because of the disturbance in the system.⁷ And the design of the feedforward part is separate from the optimal regulation.

In this study, the integral compensation and feedforward part are introduced in the optimal controller. The integral term is necessary to maintain the system state around the equilibrium points and reduce the steady-state error. Inspired by the adaptive robust control, a discontinuous projection for the integral compensation is applied to ensure the robustness.⁸,⁹ While the feedforward part can remarkably improve the responses of the control system,¹⁰ which is necessary for high-accuracy control problems. So, an augmented system with the integral compensation, error dynamics, and desired trajectory is established in this study. And the optimal tracking control problem (OTCP) of the augmented system is formulated for minimizing the performance function in the infinite horizon.

However, the introduction of the integral compensation and feedforward brings difficulties to the optimal controller design. The optimal control policy for the augmented system can hardly be obtained by solving the Hamilton–Jacobi–Bellman (HJB) equation directly. And the completely unknown system dynamics are also challenges for the controller design. In this study, a new optimal controller with reinforcement learning techniques is proposed to deal with the problems.

¹State Key Laboratory of Fluid Power and Mechatronic System, Zhejiang University, Hangzhou, China
²Ocean College, Zhejiang University, Hangzhou, China

Corresponding author:
Tao Wang, Ocean College, Zhejiang University, Hangzhou 316000, China.
Email: twang001@126.com

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
mentioned above. Different from many state-of-the-art continuous-time optimal controllers, the proposed controller is built up with a $Q$-function-based actor–critic architecture. The critic is updated by the deterministic learning. Different from those off-line optimal control policy is obtained by the method of the Bellman error and update the critic weights. The reinforcement learning (IRL) method. An integral dynamics, by employing technique. The contribution of this article is as follows.

In this article, we developed an adaptive optimal controller based on the deterministic learning technique. The contribution of this article is as follows. First, the integral compensation and feedforward are added in the control input, so that the control performance can be improved. Second, the OTCP of the system can be solved on-line with completely unknown dynamics, by employing $Q$-function approximation and deterministic learning method. Third, the convergence and Lyapunov stability of the proposed controller are proved. And the effectiveness of the controller is validated by simulation and experiment.

The rest of this article is organized as follows. Section “Optimal tracking problem formulation” presents the OTCP for the augmented system. Section “Optimal controller design” presents the design of the optimal controller. Section “Stability analysis” presents the Lyapunov stability of the proposed controller. Section “Case study” presents a case study on a hydraulic loading system. Section “Conclusion” presents the conclusion.

Optimal tracking problem formulation

**Linear control system with integral compensation**

Consider the single-input-single-output (SISO) affine continuous-time linear system described as

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

$$y(t) = Cx(t)$$

where $x \in \mathbb{R}^n$ is the system state vector, $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^n$ are the unknown system dynamics, $u \in \mathbb{R}$ is the control input, $y \in \mathbb{R}$ is the output system, and $C \in \mathbb{R}^{1 \times n}$ is the output matrix.

**Assumption 1.** The state vector $x$ and the control input $u$ are contained in compact sets.

**Assumption 2.** The system with $A$ and $B$ is controllable. The vector $Ax(t)$ and $B$ are bounded.

**Assumption 3.** Assume that the desired trajectory of the system $x_d(t) \in \mathbb{R}^n$ is bounded and there exists a Lipschitz continuous function $f_d$ such that

$$\dot{x}_d(t) = f_d(x_d(t))$$

The tracking error $e_x(t) \in \mathbb{R}^n$ is defined as

$$e_x(t) = x(t) - x_d(t)$$

The control input $u(t)$ is described as

$$u(t) = u_s(t) + u_f(t)$$

where $u_f(t)$ is the optimal control input composed of the feedforward and feedback parts and $u_s(t)$ is an integral compensation

$$u_s(t) = k_i s(t)$$

where $k_i$ is the integral coefficient, which keeps unchanged in this study. And the integral term $s(t) \in \mathbb{R}$ is described as a discontinuous projection of $e_x(t)$.

$$s(t) = \text{Proj}(e_d(t)) = \begin{cases} 0, & \text{if } e_d \geq e_{\text{max}} \text{ and } C e_d > 0 \\ 0, & \text{if } e_d \leq e_{\text{min}} \text{ and } C e_d < 0 \\ Ce_d, & \text{else} \end{cases}$$

where $e_{\text{max}}$ and $e_{\text{min}}$ are the maximum and minimum of $s(t)$, respectively. It can be seen that $u_s$ is a linear function of $e_d$ when $s(t)$ is in the range $(e_{\text{min}}, e_{\text{max}})$.

**Remark 1.** The traditional optimal regulation method obtains a proportional–derivative (PD)-type controller with feedback only, which may cause steady-state error under uncertain dynamics. In this study, an integral compensation is introduced to eliminate the steady-state error.

**Augmented system and performance function**

Define the augmented system state $X \in \mathbb{R}^{2n}$ as

$$X = [e_x, x_d]$$

The augmented state vector is composed of the tracking error and desired trajectory.

Then, the dynamics of the augmented system can be written as

$$\dot{X}(t) = F(X) + G(X)u(t)$$

where the drift dynamics $F(X)$ and the input dynamics $G(X)$ can be written as

$$F(X) = \begin{bmatrix} A(e_x + x_d) + B k_i s_x - f_d(x_d) & f_d(x_d) \\ f_d(x_d) & 0 \end{bmatrix}, \quad G(X) = \begin{bmatrix} B \\ 0 \end{bmatrix}$$
Remark 2. Because $A$ and $B$ are assumed to be unknown in this study, the dynamics of the augmented system $F(X)$ and $G(X)$ are also unknown.

Remark 3. The coefficient of the integral compensation is set as a constant in this study. Only the feedforward and feedback parts $u_f(t)$ need to be regulated. And it is not necessary for the integral term $s_e(t)$ to be a part of the state vector. The integral compensation is regarded as part of the system drift dynamics instead.

The state vector $X$ is pre-processed by a normalization of the control input can be obtained simultaneously.

The objective of the OTCP is to minimize the performance function of the augmented system

$$V_c(t) = \int^\infty_{t} e^{-\gamma(t-s)} \left[ e_\beta^{T} Q_{f} e_\beta + R_{T} u_f(t) + \gamma V_c(t) \right] ds$$

$V_c(t)$ is a discounted sum of costs in the infinite horizon. The diagonal matrix $Q_{f} \in R^{n \times n}$ and the real number $R_{T} \in R$ are coefficients of the quadratic function.

The optimal control policy for $u_f$ should be obtained on-line under the unknown dynamics

$$u_f = \pi_f'(X)$$

According to Leibniz’s rule, the derivative of $V_c(t)$ can be obtained

$$\dot{V_c}(t) = -e_\beta^{T} Q_{f} e_\beta + R_{T} u_f(t) + \gamma V_c(t)$$

And the tracking HJB equation can be written as

$$H(X, u_f, \nabla V_c) = e_\beta^{T} Q_{f} e_\beta + R_{T} u_f^2 - \gamma V_c(X)$$

$$\nabla V_c(X) \left[ F(X) + G(X)u_f(X) \right] = 0$$

Remark 4. The system dynamics $F(X)$ and $G(X)$ are unknown in this study. So, the traditional linear quadratic regulator (LQR) method is limited in this problem. And the problem can neither be solved by dealing with the HJB equation directly.

Optimal controller design

For systems with completely unknown dynamics, the HJB function can hardly be solved. In this section, an optimal controller is proposed with the actor-critic architecture. The structure of the controller is shown in Figure 1. The $Q$-value approximation is employed to evaluate the performance function. And the optimal control policy is updated on-line by the deterministic learning technique. The feedback and feedforward parts of the control input can be obtained simultaneously.

Critic network and $Q$-function approximation

The state vector is pre-processed by a normalization while considering the difference of scale between the desired trajectory $x_d$ and the tracking error $e_d$

$$\dot{x}_d = \frac{1}{k_{yd}} x_d$$

The state vector is transformed as $\bar{X}$

$$\bar{X} = [x_d, e_d]$$

The value function $V_c(t)$ is the expectation of the performance function, which can be estimated on-line by the $Q$-value

$$V_c(t) = Q_c(\bar{X}, u_f) + e_v$$

where $e_v$ is the approximation error and the $Q$-function $Q_c(\bar{X}, u_f)$ is evaluated based on the actual control input $u_f$.

Remark 5. During the on-line learning process, a probing noise is added on $u_f$ (see section “Experiment results” for more detail). So, the difference between the value function and $Q$-function $e_v$ is mainly caused by the probing noise.

Because the amplitude of probing noise is relatively low, the approximation error $e_v$ can be kept bounded in a compact set according to Assumption 2

$$||e_v|| \leq e_{v_{max}}$$
The $Q$-function can be obtained by a linear approximation

$$Q_c = W_c^T \phi_c(\bar{x},u_f)$$  \hspace{1cm} (18)

where $W_c(t) \in \mathbb{R}^l$ are the weights for ideal approximation and $\phi_c$ is the basis function.

The Bellman equation can be obtained from equation (10), which is written as

$$\int_{t-T}^{t} e^{-\gamma(t-t')} [e_c(t')^T Q \phi_c(t') + R_f u_f(t')^2] dt' + e^{-\gamma T} V_c(t) - V_c(t-T) = 0$$  \hspace{1cm} (19)

According to equations (16) and (19), the tracking Bellman error can be written as

$$\varepsilon_B(t) = \int_{t-T}^{t} e^{-\gamma(t-t')} [e_c(t')^T Q \phi_c(t') + R_f u_f(t')^2] dt' + W_c^T \Delta \phi_c(t)$$  \hspace{1cm} (20)

where

$$\Delta \phi_c(t) = e^{-\gamma T} \phi_c(\bar{x}(t), u_f(t)) - \phi_c(\bar{x}(t-T), u_f(t-T))$$  \hspace{1cm} (21)

According to equations (16) and (18), the tracking Bellman equation error $\varepsilon_B(t)$ can be written as

$$\varepsilon_B(t) = \varepsilon_\delta(t-T) - e^{-\gamma T} \varepsilon_\delta(t)$$  \hspace{1cm} (22)

So, $\varepsilon_B(t)$ is bounded according to equation (17).

The $Q$-value is estimated by the critic network

$$\hat{Q}_c = \hat{W}_c^T \phi_c(\bar{x},u_f)$$  \hspace{1cm} (23)

According to equation (20), the Bellman error with respect to the weights of the critic network $\hat{W}_c$ can be written as

$$\varepsilon_B(t) = \int_{t-T}^{t} e^{-\gamma(t-t')} [e_c(t')^T \hat{Q} \phi_c(t') + R_f u_f(t')^2] dt' + \hat{W}_c^T \Delta \phi_c(t)$$  \hspace{1cm} (24)

In this study, the policy iteration method is employed to minimize the Bellman error. The objective function of the critic network can be written as

$$E_B = \frac{1}{2} \varepsilon_B^2$$  \hspace{1cm} (25)

The update rate of $\hat{W}_c$ can be written as

$$\dot{\hat{W}}_c = -\alpha_c \frac{\partial E_B}{\partial \hat{W}_c} \frac{\partial \hat{W}_c}{\partial \varepsilon_B}$$  \hspace{1cm} (26)

where $\theta_c$ is the regularization coefficient and $k_{nc}$ is the normalization term

$$k_{nc} = 1 + \Delta \phi_c^T \Delta \phi_c$$  \hspace{1cm} (27)

The normalization term is applied to limit the update rate of the network weights.

Define the estimation error of the critic network as

$$\hat{W}_c = W_c - \hat{W}_c$$  \hspace{1cm} (28)

According to equations (20) and (28), the Bellman error $\varepsilon_B(t)$ can be written as

$$\varepsilon_B(t) = -\hat{W}_c^T(t) \Delta \phi_c(t) + \varepsilon_B(t)$$  \hspace{1cm} (29)

So, the critic neural network (NN) estimation error dynamics becomes

$$\dot{\hat{W}}_c = -\alpha_c \left( \Delta \phi_c \Delta \phi_c^T \hat{W}_c(t) - \frac{1}{k_{nc}} \varepsilon_B(t) \Delta \phi_c - \theta_c \hat{W}_c \right)$$  \hspace{1cm} (30)

where $\Delta \phi_c$ is defined as

$$\Delta \phi_c = \frac{\Delta \phi_c}{1 + \Delta \phi_c^T \Delta \phi_c}$$  \hspace{1cm} (31)

**Actor network and deterministic learning technique**

The control policy is improved by the actor network. The deterministic learning technique is applied to update the actor network weights on-line.

The optimal control policy $\pi^*_f$ satisfies

$$\pi^*_f(\bar{x}) = \arg\min Q_a(\bar{x},u_f)$$  \hspace{1cm} (32)

The optimal control input $u^*_f$ can be expressed by a linear approximation

$$u^*_f = \pi^*_f(\bar{x}) = W^T_a \phi_a(\bar{x}) + v_a$$  \hspace{1cm} (33)

where $v_a$ is the approximation error and satisfies

$$||v_a|| \leq v_a_{\text{max}}$$  \hspace{1cm} (34)

$u_f$ is the estimation of $u^*_f$ according to the actor network

$$u_f = \pi_f(\bar{x}) = \hat{W}_a^T \phi_a(\bar{x})$$  \hspace{1cm} (35)

The deterministic learning method is employed to update the weights $\hat{W}_a$ with the $Q$-value in the critic.

$$\dot{\hat{W}}_a = -\alpha_a \left( \frac{\partial Q_a(\bar{x},u_f)}{\partial u_f} \frac{\partial \pi_f(X)}{\partial \hat{W}_a} + \theta_a \hat{W}_a \right)$$  \hspace{1cm} (36)

where $\theta_a$ is a regularization coefficient added to ensure convergence.

The term $\partial Q_a(\bar{x},u_f)/\partial u_f$ can be written as
\[ \frac{\partial Q_c(X_t, u_f)}{\partial u_f} = \nabla_u \phi_c^T \dot{W}_c \] (37)

So, the update rate \( \dot{W}_a \) can be obtained as
\[ \dot{W}_a = -\alpha_a \left( \frac{1}{k_{na}} \phi_a \nabla_u \phi_c^T \dot{W}_c + \theta_a \dot{W}_a \right) \] (38)

where \( k_{na} \) is the normalization term
\[ k_{na} = (1 + \phi_a^T \phi_a)(1 + \nabla_u \dot{V}_c^T \nabla_u \dot{V}_c) \] (39)

Remark 6. The initial weights of the actor network \( W_a(0) \) should be an admissible control policy which is able to stabilize the system.

The estimation error \( \dot{W}_a \) is defined as
\[ \dot{W}_a = W_a - \hat{W}_a \] (40)

Then, the estimation error dynamics of weights \( W_a \) can be written as
\[ \dot{W}_a = \alpha_a \left( \frac{1}{k_{na}} \phi_a \nabla_u \phi_c^T W_c + \theta_a \dot{W}_a \right) \] (41)

Persistently exciting condition

According to equation (26), the convergence of the weights \( W_c \) requires the persistent excitation (PE) condition of \( \Delta \phi_c \). For all \( t \geq 0 \), there exists \( \mu_1 > 0 \) and \( \mu_2 > 0 \) such that
\[ \frac{d}{dt} \int_0^t \Delta \phi_c(\tau) \Delta \phi_c^T(\tau) d\tau \leq \mu_2 I \] (42)

According to equation (38), the term \( \nabla_u \phi_c^T \dot{W}_c \) denotes the gradient of the estimated \( Q \)-function with respect to the control input. And the vector \( \phi_c \) should satisfy the PE condition to ensure the convergence of the weights \( W_a \). For all \( t \geq 0 \), there exist \( \mu_3 > 0 \) and \( \mu_4 > 0 \) such that
\[ \frac{d}{dt} \int_0^t \phi_c(\tau) \phi_c^T(\tau) d\tau \leq \mu_4 I \] (43)

However, the PE condition can hardly be verified on-line.\(^{23,24}\) So, in this study, a probing noise is added on the control input \( u_f \).

Stability analysis

In this section, the stability of the proposed method is proved in the Lyapunov sense.

The Lyapunov function is defined as
\[ J(t) = V_c(t) + \frac{1}{2} \dot{W}_c(t)^T \alpha_c^{-1} \dot{W}_c(t) + \frac{1}{2} \dot{W}_a(t)^T \alpha_a^{-1} \dot{W}_a(t) \] (44)

The derivative of the Lyapunov function is given by
\[ \dot{J}(t) = \dot{V}_c(t) + \dot{W}_c(t)^T \alpha_c^{-1} \dot{W}_c(t) + \dot{W}_a(t)^T \alpha_a^{-1} \dot{W}_a(t) \] (45)

According to equations (12) and (18), \( V_c(t) \) can be written as
\[ V_c(t) = -\varepsilon_a(t)^T Q_c(t) e_a(t) - RT_a \varepsilon_a(t)^2 + \gamma(W_c^T \phi_c + \varepsilon_c) \] (46)

Note that \( u_f \in R \), the term \( u_f(t)^2 \) in equation (46) can be written as
\[ u_f(t)^2 = (W_c^T \phi_c)^2 = W_c^T \phi_c \phi_c^T W_c - 2 \dot{W}_c(t)^T \phi_c \phi_c^T W_c + \dot{W}_c(t)^T \phi_c \phi_c^T \dot{W}_c \] (47)

The first term in equation (46) can be written as
\[ \varepsilon_a^T Q_c(t) e_a(t) \geq \lambda_{\text{min}}(Q_c(t)) \| e_a(t) \|^2 \] (48)

So, the derivative \( V_c(t) \) can be written as
\[ V_c(t) \leq -\lambda_{\text{min}}(Q_c(t)) \| e_a(t) \|^2 - RT_a \varepsilon_a(t)^2 + \gamma(W_c^T \phi_c + \varepsilon_c) \] (49)

According to equation (30), the second term in equation (45) can be written as
\[ \dot{J}_1(t) = \dot{W}_a(t)^T \alpha_a^{-1} \dot{W}_a(t) \]
\[ = -\dot{W}_a(t)^T \Delta \phi_c \Delta \phi_c^T W_c(t) + \frac{1}{k_{na}} \dot{W}_c(t)^T e_a(t) \Delta \phi_c \]
\[ + \dot{W}_c(t)^T \theta_c W_c - \dot{W}_c(t)^T \theta_c \dot{W}_c \] (50)

Using equation (41), the third term in equation (45) can be written as
\[ J_2(t) = \frac{1}{k_{na}} \dot{W}_a(t)^T \phi_a \nabla_u \phi_c^T W_c + \dot{W}_a(t)^T \theta_a \dot{W}_a \] (51)

According to equations (23) and (38), \( J_2(t) \) can be written as
\[ J_2(t) = \frac{1}{k_{na}} \dot{W}_a(t)^T \phi_a \nabla_u \phi_c^T W_c - \frac{1}{k_{na}} \dot{W}_a(t)^T \phi_a \nabla_u \phi_c^T \dot{W}_c + \dot{W}_a(t)^T \theta_a \dot{W}_a \] (52)

According to the basic inequality, the second term in equation (52) can be written as
\[ -\frac{1}{k_{na}} \dot{W}_a(t)^T \phi_a \nabla_u \phi_c \nabla_u \phi_c^T \dot{W}_c \leq \frac{1}{4k_{na}} \dot{W}_a(t)^T \phi_a \phi_a^T \dot{W}_a(t) + \dot{W}_a(t)^T \theta_a \dot{W}_a \] (53)

Therefore
\[ J_0(t) \leq \frac{1}{\kappa_{na}} \tilde{W}_a(t)^T \phi_a \nabla_{\phi_a} \phi_a^T \tilde{W}_a \\
+ \frac{1}{4k_{na}^2} \tilde{W}_a(t)^T \phi_a \phi_a^T \tilde{W}_a(t) \\
+ \tilde{W}_a^T \nabla_{\phi_a} \phi_a \nabla_{\phi_a} \phi_a^T \tilde{W}_a \\
- \tilde{W}_a(t)^T \theta_a \tilde{W}_a \tag{54} \]

Using equations (49), (50), and (54), \( \dot{J}(t) \) becomes
\[ \dot{J}(t) \leq -\lambda_{\min}(Q_T) \| e_d \|^2 - R_T(W_a^T \phi_a)^2 + \gamma_{\phi_e} \\
- \tilde{W}_e(t)^T \nabla_{\phi_e} \tilde{W}_e(t) + \tilde{W}_e(t)^T k_1 - \tilde{W}_a(t)^T N_2 \tilde{W}_a \\
+ \tilde{W}_a^T k_2 + \gamma_{\phi_e} \tilde{W}_e \tag{55} \]

where \( k_1 \) is written as
\[ k_1 = \frac{1}{\kappa_{na}} \delta_d(t) \Delta \phi + \theta_e \tilde{W}_e \tag{56} \]
and \( k_2 \) is written as
\[ k_2 = 2R_T \phi_a \phi_a^T W_a + \theta_a W_a + \frac{1}{\kappa_{na}} \phi_a \nabla_{\phi_a} \phi_a^T W_c \tag{57} \]

According to the range of \( \delta_d \), and the assumption of \( \phi_a \) and \( \phi_e \), it can be concluded that \( k_1 \) and \( k_2 \) are bounded.

And \( N_1 \) and \( N_2 \) are written as
\[ N_1 = \theta_e + \Delta \phi \Delta \phi^T - \nabla_{\phi_a} \phi_e \nabla_{\phi_a} \phi_e^T \tag{58} \]
\[ N_2 = \theta_a + R_T \phi_a \phi_a^T - \frac{1}{4k_{na}^2} \phi_a \phi_a^T \tag{59} \]

The regularization coefficients \( \theta_a \) and \( \theta_e \) should make sure that \( N_1 \) and \( N_2 \) are larger than zero. Then, \( \dot{J}(t) \) becomes negative provided that
\[ \| e_d \| > \sqrt{\frac{\| e_{\phi_e} \|}{\lambda_{\min}(Q_T)}} \tag{60} \]
\[ \| \tilde{W}_e \| > \frac{k_1}{\lambda_{\min}(N_1)} \tag{61} \]
\[ \| \tilde{W}_a \| > \frac{k_2^2 + 4\gamma_{\phi_e} \tilde{W}_e^T \phi_e}{2\lambda_{\min}(N_2)} \tag{62} \]

**Case study**

In this section, the tracking control of a hydraulic loading system for hydraulic motors is taken as a case study.\(^{25}\) The hydraulic loading system utilizes energy regeneration technique to improve efficiency.\(^{26-28}\) The photograph of the experimental setup is shown in Figure 2. Simulation and experiment results are given to verify the effectiveness of the proposed controller. The objective is to achieve high-accuracy pressure control, which can be defined as an OTCP.

**OTCP of the hydraulic loading system**

The simplified schematic of the hydraulic loading system with energy regeneration is shown in Figure 3. The hydraulic loading system is used to test the hydraulic motor (Rexroth A2FM63) mounted on the transmission shaft. The system is driven by a variable frequency induction motor (ABB QABP 355L2A). The variable displacement loading pump (Rexroth A6V2F63) regenerates the mechanical energy and adjusts the system pressure. Two flow meters (KRACHT VC12) and pressure sensors (KELLER PA-33X/600BAR) are mounted at two outlets of the tested motor. A personal computer (PC) receives all the sensor signals and sends the control signal by an I/O card (ADVANTECH USB4716). The objective of the OTCP is to obtain the optimal displacement input of the loading pump so that the performance function (10) can be minimized.
The dynamics of the loading pump can be simplified as a first-order system

\[ T_m \dot{q}_m + q_m = u_m \]  \hspace{1cm} (63)

where \( T_m \) is the time constant of the loading pump, \( q_m \) is the output displacement of the loading pump neglecting the leakage flow, and \( u_m \) (cm\(^3\)/rev) is the displacement input of the loading pump.

The state vector for the optimal controller is written as

\[ x = [p_c, \dot{p}_c]^T \]  \hspace{1cm} (66)

\[ y = p_c \]  \hspace{1cm} (67)

According to equations (63) and (64), the hydraulic loading system can be described as a second-order linear system

\[ \dot{x}(t) = Ax(t) + Bu(t) \]  \hspace{1cm} (68)

\[ A = \begin{bmatrix} 0 & \frac{\beta k_{pl}}{T_m} \\ \frac{1}{T_m} & \frac{1}{T_m} \end{bmatrix} \]  \hspace{1cm} (69)

\[ B = \begin{bmatrix} 0 \\ \frac{\beta \nu_a}{T_m} \end{bmatrix} \]  \hspace{1cm} (70)

The control input \( u(t) \) is written as

\[ u(t) = u_{mp}(t) - q_c \]  \hspace{1cm} (71)

And the range of \( u(t) \) is limited as

\[ -30 \text{cm}^3/\text{rev} \leq u(t) \leq 30 \text{cm}^3/\text{rev} \]  \hspace{1cm} (72)

The state vector for the optimal controller is written as

\[ X = [e_{d1}, e_{d2}, x_{d1}, x_{d2}]^T, e_d = x - x_d \]  \hspace{1cm} (73)

where \( x_{d}(t) \) is the desired trajectory of the system pressure.

The basis function \( \phi_c(X, u_t) \) is chosen as a second-order polynomial

\[ \phi_c(X, u_t) = [X_1^2, X_1 X_2, X_1 X_3, X_1 X_4, X_1 u_{t1}, X_2 X_3, X_2 X_4, X_2 u_{t1}, X_3 X_4, X_3 u_{t1}, X_4 u_{t1}, u_{t1}^2]^T \]  \hspace{1cm} (74)

And \( \phi_c(X) \) is an identity mapping on \( \hat{X} \)

\[ \phi_c(X) = [X_1, X_2, X_3, X_4]^T \]  \hspace{1cm} (75)

The controller is designed with \( k_1 = 1, k_{yd} = 10, Q_T = 10 I_2 \) (\( I_2 \) is the 2\( \times \)2 identity matrix), \( R = 1, \gamma = 0.1, \theta_c = 0.01, \) and \( \theta_a = 0.02. \) The sample time \( T_s = 0.01 \text{s}. \) The learning rates are chosen as

\[ \alpha_c = 0.2 e^{-0.05t} \]  \hspace{1cm} (76)

\[ \alpha_a = 0.1 e^{-0.05t} \]  \hspace{1cm} (77)

Simulation results

The hydraulic loading system is modeled in Matlab/ Simulink with \( T_m = 0.1 \text{s}, \) \( V = 0.008 \text{m}^3, \) \( \beta = 1400 \text{MPa}, \) \( n = 1000 \text{rev/min}, \) \( q_c = 63 \text{cm}^3/\text{rev}, \) and \( k_{pl} = 0.2 \text{L}/(\text{min MPa}). \) And the matrices \( A \) and \( B \) can be written as

\[ A = \begin{bmatrix} 0 & 1 \\ -5.83 & -10.58 \end{bmatrix} \]  \hspace{1cm} (78)

\[ B = \begin{bmatrix} 0 \\ 29.17 \end{bmatrix} \]  \hspace{1cm} (79)

Notice that the system dynamics \( A \) and \( B \) are unknown while designing the optimal controller with our proposed method.

![Figure 4. Simulation results of the proposed controller compared with the PID method: (a) system pressure while tracking a non-periodic signal and (b) comparison of the tracking errors.](image-url)
Figure 4 shows the control performance of the proposed controller while tracking a non-periodic signal compared with a proportional–integral–derivative (PID) controller.\(^{29,30}\)

The feedback gain of the PID controller \(K_{PD}\) is regulated by the LQR method, and the integral coefficient \(K_I\) is the same as that of the proposed method \(K_{PD} = 3, 3, 3, 3\):\(^{3, 4}\).

It can be seen that the tracking error of the proposed controller can be remarkably reduced after the learning process. With the feedforward term in the output, the proposed controller can outperform the PID controller, which is shown in Figure 4(b).

Figure 5 shows the system state \(\bar{X}\). It can be seen that the amplitudes of the four elements are similar after normalization.

The gradient \(\nabla_{\bar{X}} \hat{Q}(\bar{X}, \bar{u})\) is shown in Figure 6. When pressure rises with overshoot, the gradient is positive, and the control input is expected to decrease for improving the control performance.

The convergence during the learning process is shown in Figure 7. The Bellman error keeps bounded and converges to zero gradually. In Figure 7(b), the critic weights vector finally converge to

\[
\hat{W}_c = [0.010, -0.135, -0.009, -0.0367, -0.185, 0.264, -0.032, -0.098, 0.150, 0.070, -0.016, -0.093, -0.081, -0.282, 0.008]^T
\]

The initial weights of actor network are set to be

\[
\hat{W}_a(0) = [2, 0, 0, 0]^T
\]

In Figure 7(c), it can be seen that the weights \(\hat{W}_a\) finally converge to

\[
\hat{W}_a = [1.905, 2.316, 0.081, 0.834]^T
\]

The feedback and feedforward parts are learned simultaneously.
Figure 8. Experiment results of the proposed controller on the hydraulic loading system: (a) system pressure while tracking a non-periodic signal and (b) tracking error compared with the PID method.

Experiment results

Figure 8 shows the experiment results of the proposed optimal controller while tracking a non-periodic signal. The performance of the proposed controller is also compared with a PID controller. It can be seen that the difference of tracking errors between the two controllers is relatively small at the beginning. After several seconds for learning, the tracking error of the proposed controller is remarkably reduced.

Figure 9 shows the convergence of the proposed controller. It can be seen that the Bellman error keeps bounded under the experimental circumstances. The weights of the critic $W_c$ finally converge to

$$ W_c = [0.255, -0.217, 0.007, -0.030, -0.094, 0.163, -0.020, -0.003, 0.084, 0.031, -0.001, -0.015, -0.025, -0.327, 0.035]^T $$

(84)

And $W_a$ finally converges to

$$ W_a = [1.825, 1.154, 0.244, 0.691]^T $$

(85)

So, the convergences of $W_a$ and $W_c$ are also validated in the experiment.

Conclusion

In this article, an SISO continuous-time optimal tracking controller is proposed for linear systems with completely unknown dynamics. The proposed controller is different from those conventional proportional-deviation-type optimal controllers in two aspects. First, the integral compensation and the feedforward part are introduced into the controller, so the control performance can be improved. Second, the reinforcement learning techniques are applied in the controller design, and the optimal control policy can be obtained on-line without the prior knowledge of system dynamics. The Lyapunov stability and the convergence of the system have been proved. A case study on a hydraulic loading system with energy regeneration is given to validate the
control performance. The simulation and experiment results have shown the effectiveness of the proposed controller.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research is supported by the National Natural Science Foundation of China (51475414 and 51875504), as well as the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (51821093).

ORCID iD
Tao Wang https://orcid.org/0000-0002-5121-0599

References
1. Li Z, Liu HHT, Zhu B, et al. Nonlinear robust attitude tracking control of a table-mount experimental helicopter using output feedback. IEEE T Ind Electron 2015; 62(9): 5665–5676.
2. Wang T, Zhang Y, Chen Z, et al. Parameter identification and model-based nonlinear robust control of fluidic soft bending actuators. IEEE/ASME T Mech 2019; 24(3): 1346–1355.
3. Yuan M, Chen Z, Yao B, et al. Fast and accurate motion tracking of a linear motor system under kinematic and dynamic constraints: an integrated planning and control approach. IEEE T Contr Syst T. Epub ahead of print 16 December 2019. DOI: 10.1109/TCST.2019.2955658.
4. Athans M and Falb PL. Optimal control: an introduction to the theory and its applications. IEEE T Automat Contr 2003; 12: 345–347.
5. Weiss M. Robust and optimal control. Automatica 1997; 33: 2095–2095.
6. Molinari BP. The time-invariant linear-quadratic optimal control problem. Automatica 1977; 13: 347–357.
7. Yao B, Bu FP, Reedy J, et al. Adaptive robust motion control of single-rod hydraulic actuators: theory and experiments. IEEE/ASME T Mech 2000; 5: 79–91.
8. Yao B and Tomizuka M. Adaptive robust control of SISO nonlinear systems in a semi-strict feedback form. Automatica 1997; 33: 893–900.
9. Yao B and Tomizuka M. Smooth robust adaptive sliding mode control of robot manipulators with guaranteed transient performance. J Dyn Syst: T ASME 1996; 118: 764–775.
10. Takeshi T. Introduction of feedforward control for an improved optimal regulator system and its application. Int J Control 1983; 37: 191–204.
11. Lewis FL and Vrabie D. Reinforcement learning and adaptive dynamic programming for feedback control. IEEE Circ Syst Mag 2009; 9: 32–50.
12. Li Z, Xue S, Lin W, et al. Training a robust reinforcement learning controller for the uncertain system based on policy gradient method. Neurocomputing 2018; 316: 313–321.
13. Xue S, Li Z and Yang L. Training a model-free reinforcement learning controller for a 3-degree-of-freedom helicopter under multiple constraints. Meas Control 2019; 52(7–8): 844–854.
14. Yang X, Liu D and Wang D. Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints. Int J Control 2014; 87: 553–566.
15. Vamvoudakis KG and Lewis FL. Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem. Automatica 2010; 46: 878–888.
16. Modares H and Lewis FL. Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning. Automatica 2014; 50: 1780–1792.
17. Vamvoudakis KG, Vrabie D and Lewis FL. Online adaptive algorithm for optimal control with integral reinforcement learning. Int J Robust Nonlin 2014; 24: 2686–2710.
18. Xu B, Yang C and Shi Z. Reinforcement learning output feedback NN control using deterministic learning technique. IEEE T Neural Net Learn 2014; 25: 635–641.
19. Liu D and Yang GH. Model-free adaptive control design for nonlinear discrete-time processes with reinforcement learning techniques. Int J Syst Sci 2018; 49: 2298–2308.
20. Schulman J, Levine S, Moritz P, et al. Trust region policy optimization. Int Conf Mach Learn 2015; 37: 1889–1897.
21. Silver D, Lever G, Heess XN, et al. Deterministic policy gradient algorithms. Int Conf Mach Learn 2014; 32: 387–395.
22. Prokhorov DV, Santiago RA and Donald CW. Adaptive critic designs: a case study for neurocontrol. Neural Networks 1995; 8: 1367–1372.
23. Lewis FL and Vamvoudakis KG. Reinforcement learning for partially observable dynamic processes: adaptive dynamic programming using measured output data. IEEE T Syst Man Cy B 2010; 41: 14–25.
24. Chowdhary G, Yucelen T, Maximillian M, et al. Concurrent learning adaptive control of linear systems with exponentially convergent bounds. Int J Adapt Control 2013; 27: 280–301.
25. Zhang J, Wang QF and Wang T. Design, modeling and pressure control of a 300-kW-class hydraulic pump/motor loading system with energy regeneration. Proc IMechE, Part I: J Systems and Control Engineering. Epub ahead of print 4 March 2020. DOI: 10.1177/0959651820903549.
26. Wang T and Wang QF. An energy-saving pressure-compensated hydraulic system with electrical approach. IEEE/ASME T Mech 2014; 19: 570–578.
27. Wang T and Zhou Z. A compact hydrostatic-driven electric generator: design, prototype, and experiment. IEEE/ASME T Mech 2016; 21: 1612–1619.
28. Ho TH and Ahn KK. Design and control of a closed-loop hydraulic energy-regenerative system. Autom Constr 2012; 22: 444–458.
29. Ang KH, Chong G and Li Y. PID control system analysis, design, and technology. IEEE T Contr Syst T 2005; 13: 559–576.
30. Lee YH, Park SW and Lee MY. PID controller tuning for desired closed-loop responses for SI/SO systems. AICHe J 2010; 44: 106–115.