01 Jan 2022

Hamiltonian-Driven Adaptive Dynamic Programming with Efficient Experience Replay

Yongliang Yang

Yongping Pan

Cheng Zhong Xu

Donald C. Wunsch
Missouri University of Science and Technology, dwunsch@mst.edu

Recommended Citation

Y. Yang et al., "Hamiltonian-Driven Adaptive Dynamic Programming with Efficient Experience Replay," *IEEE Transactions on Neural Networks and Learning Systems*, Institute of Electrical and Electronics Engineers, Jan 2022.
The definitive version is available at https://doi.org/10.1109/TNNLS.2022.3213566

Hamiltonian-Driven Adaptive Dynamic Programming With Efficient Experience Replay

Yongliang Yang, *Member, IEEE*, Yongping Pan, *Senior Member, IEEE*,
Cheng-Zhong Xu, *Fellow, IEEE*, and Donald C. Wunsch, II, *Fellow, IEEE*

***Abstract*— This article presents a novel efficient experience-replay-based adaptive dynamic programming (ADP) for the optimal control problem of a class of nonlinear dynamical systems within the Hamiltonian-driven framework. The quasi-Hamiltonian is presented for the policy evaluation problem with an admissible policy. With the quasi-Hamiltonian, a novel composite critic learning mechanism is developed to combine the instantaneous data with the historical data. In addition, the pseudo-Hamiltonian is defined to deal with the performance optimization problem. Based on the pseudo-Hamiltonian, the conventional Hamilton–Jacobi–Bellman (HJB) equation can be represented in a filtered form, which can be implemented online. Theoretical analysis is investigated in terms of the convergence of the adaptive critic design and the stability of the closed-loop systems, where parameter convergence can be achieved under a weakened excitation condition. Simulation studies are investigated to verify the efficacy of the presented design scheme.**

***Index Terms*— Hamilton–Jacobi–Bellman (HJB) equation, Hamiltonian-driven adaptive dynamic programming (ADP), pseudo-Hamiltonian, quasi-Hamiltonian, relaxed excitation condition.**

I. Introduction

RECENT development in control theory and machine learning has promoted reinforcement learning (RL) and adaptive dynamic programming (ADP) for performance optimization of the decision-making problem with long-term cumulative reward. A classical optimal control theory laid the solid theoretical foundation for the dynamical optimization problem, where Pontryagin's maximum principle (PMP) and Bellman's dynamic programming (DP) are the centerpieces of the optimal control theory [1]. Based on PMP, the optimal control policy over finite horizon depends on solving a two-point boundary value problem (TPBVP) [2]. However, solving the TPBVP for general nonlinear dynamical systems remains to be an open problem. On the other hand, the DP suffers from the "curse of dimensionality" [3]. In contrast, RL and ADP are intelligent methods to assist the agent make intelligent decisions to optimize the cumulative reward based on online collected data [4], [5], [6], [7], [8], [9]. Successful application of RL and ADP to control applications can be found in recent literature [10], [11], [12], [13], [14].

A. Related Work

In ADP and RL-based adaptive optimal controller design, the fundamental considerations are twofold. First, the system during the learning process has to be stable, and the closed-loop signals should be bounded to guarantee that the online learning scheme is feasible. As a typical iterative ADP algorithm, in policy iteration, the initial iterative policy should be admissible to ensure that the closed-loop stability in each iteration is stable [15], [16], [17]. However, the initial admissible policy is difficult to be obtained for complex systems with unknown dynamics. Second, the learning convergence of ADP and RL is desired to guarantee that the optimal control policy can be obtained [18], [19], [20]. For both synchronous and iterative ADP algorithms [21], [22], [23], [24], the persistent of excitation (PE) condition on the collected data is critical to the learning convergence to the optimum [25]. However, the PE condition is difficult to be satisfied because it requires the signal to have sufficient rich information over the infinite horizon [26]. In addition, the PE condition requires the signal to be frequency-rich, which might lead to undesired oscillation in the system evolution. Therefore, it is desired to obtain a convergent and stable ADP online learning algorithm without the stringent PE condition.

Experience replay is an efficient method in RL, which repeatedly utilizes the collected historical data for intelligent decision-making [27]. It has been successfully combined with neural networks for autonomous agents to directly learn from the experience based on sequential actions in

Manuscript received 24 March 2022; revised 6 August 2022; accepted 4 October 2022. This work was supported in part by the Natural Science Foundation of Beijing Municipality under Grant J210005; in part by the Guangdong Pearl River Talent Program of China under Grant 2019QN01X154; in part by the National Natural Science Foundation of China under Grant 61903028; in part by the Science and Technology Development Fund, Macao, SAR [Science and Technology Development Fund (FDCT)], under Grant 0015/2019/AKP; and in part by the Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems under Grant 2019B121205007. *(Corresponding author: Yongping Pan.)*

Yongliang Yang is with the Key Laboratory of Knowledge Automation for Industrial Processes Ministry of Education, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: yangyongliang@ieee.org).

Yongping Pan is with the School of Advanced Manufacturing, Sun Yat-sen University, Shenzhen 518107, China (e-mail: panyongp@mail.sysu.edu.cn).

Cheng-Zhong Xu is with the State Key Laboratory of Internet of Things for Smart City, Department of Computer and Information Science, University of Macau, Macau (e-mail: czxu@um.edu.mo).

Donald C. Wunsch, II, is with the Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65409 USA (e-mail: dwunsch@mst.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3213566.

Digital Object Identifier 10.1109/TNNLS.2022.3213566

the environment [28], [29]. Recently, the experience-replay technique has also been employed an adaptive control theory to obviate the stringent PE condition, which has been successfully applied to engineering problems, including adaptive cruise control [30], wastewater treatment process [31], and wind farm control [32]. In concurrent learning for model reference adaptive control [33], the collected data are selected and stored based on a rank condition on the data matrix, which is less restrictive than the PE condition and easy to be verified during the online process. However, the concurrent learning technique requires state derivative to be measurable [34], which might not be feasible in applications. In composite learning, the time-interval integral with filtered signals is designed to construct a novel residual to avoid the measurement of time derivation of plant states [35], where the concept of interval excitation (IE) condition is presented to be weaker than the PE condition and can ensure the parameter estimation convergence [36]. The IE condition is further applied to the adaptive optimal control for linear systems in [37] to provide the adaptive solution with convergence to the solution to the algebraic Riccati equation. However, the optimal control problem of nonlinear system requires solving the Hamilton–Jacobi–Bellman (HJB) equation, which is a nonlinear partial differential equation and difficult to be solved due to the inherent nonlinearity. This article aims to develop an efficient ADP algorithm to solve the HJB equation for nonlinear systems without requiring the PE condition.

The Hamiltonian-driven framework is developed in [38]. Three subproblems for optimal control are categorized as policy evaluation problems for a fixed admissible policy, the performance comparison problem with different admissible policies, and the performance improvement for a given admissible policy. It is shown that the Hamiltonian plays a critical role in the performance optimization problem, and the classical policy iteration algorithm can be viewed as a successive minimization of the iterative Hamiltonian [38]. The Hamiltonian-driven framework is later applied to the performance optimization with intermittent feedback, where the effect of intermittent feedback on the communication bandwidth and the control performance of the iterative ADP algorithms is investigated [39]. In addition, the effect of the residual resulting from the function approximator on the convergence of iterative ADP algorithm is investigated in [40], where a sufficient condition to ensure the closed-loop stability in each iteration and the convergence to the optimum is developed. Recent extensions of the Hamiltonian-driven ADP have been made to solve the differential games [41] and multiobjective optimization [42]. In this article, the Hamiltonian-driven ADP is extended with value function approximation to provide novel defined quasi-Hamiltonian and pseudo-Hamiltonian, which can efficiently combine the instantaneous data with the historical data.

1) Contributions: The contributions of this article are threefold. First, to solve the policy evaluation problem for nonlinear systems with a given admissible policy, the quasi-Hamiltonian is defined to utilize the historical data efficiently. On this basis, an online filtered signal is designed to yield the filtered Bellman equation, which can be solved with a relaxed excitation. Second, for the policy optimization problem for nonlinear systems, the pseudo-Hamiltonian is presented to parameterize the HJB equation, which is named the filtered HJB equation, and efficiently takes the instantaneous data and the online collected data into consideration. Finally, to solve the filtered Bellman equation and filtered HJB equation, a relaxed excitation on the filtered signals is considered to ensure the online learning convergence without the requirement of the stringent PE condition.

2) Structure: The remainder of this article is organized as follows. The optimal control problem of nonlinear dynamical in continuous time with its background knowledge is presented in Section II. In Section III, the quasi-Hamiltonian is presented for efficient data utilization for policy evaluation, where the filtered Bellman equation is developed. A composite critic learning algorithm is presented to solve the filtered Bellman equation with a relaxed excitation condition. Moreover, the filtered HJB equation is derived in Section IV, where the pseudo-Hamiltonian is defined to provide the quadratic parameterization for the filtered HJB equation. On this basis, novel actor–critic learning is developed to solve the filtered HJB equation without the requirement of the PE condition. The simulation study with the presented efficient Hamiltonian-driven ADP algorithm is investigated in Section V. The concluding remarks are made in Section VI.

B. Preliminaries

The following definitions are required for subsequent discussions.

Definition 1 (PE [43]): A vector signal $y(t) \in \mathbb{R}^p$ is PE in $\mathbb{R}^p$ with an excitation level $\beta_1 > 0$, provided that for all $t \in \mathbb{R}^+$, there exist constants $T_{\text{PE}} \in \mathbb{R}^+$ and $\beta_2 > \beta_1$ such that

$$\beta_1 I_{p\times p} \leq \int_t^{t+T_{\text{PE}}} y(\tau)y^{\text{T}}(\tau)\mathrm{d}\tau \leq \beta_2 I_{p\times p}.$$

Definition 2 (Relaxed Excitation Condition): A vector signal $y(t) \in \mathbb{R}^p$ is exciting over interval $[t_a, t_b]$ with an excitation level $\beta_1 > 0$, provided that there exist constants $\beta_1 \in \mathbb{R}^+$ and $\beta_2 \in \mathbb{R}^+$ such that

$$\beta_1 I_{p\times p} \leq \int_{t_a}^{t_b} y(\tau)y^{\text{T}}(\tau)\mathrm{d}\tau \leq \beta_2 I_{p\times p}$$

where $\beta_2 > \beta_1 > 0$.

II. Problem Statement

We consider the following continuous-time nonlinear dynamical system:

$$\dot{x} = f(x) + g(x)u, x(t_0) = x_0 \tag{1}$$

where $x \in \mathbb{R}^n$ denotes the system state, $u \in \mathbb{R}^m$ denotes the control input, and the initial condition x_0 is given. In addition, $f(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ and $g(\cdot) : \mathbb{R}^{n\times m} \to \mathbb{R}^n$ are the system dynamics. On a compact set Ω_x, the functions $f(x)$ and $g(x)$ are locally Lipschitz functions and satisfy $\|f(\cdot)\| \leq \eta_f \|\cdot\|$ and $\|g(\cdot)\| \leq \eta_g$ with positive constants η_f and η_g. Moreover, it is assumed that $f(0) = 0$, which implies that the origin is an equilibrium of the system.

For the optimal control problem, the performance function under consideration takes the following form:

$$V_u(x_0) = \int_{t_0}^{\infty} r(x(\tau), u(\tau)) d\tau \tag{2}$$

with the reward function $r(x, u) = Q(x) + u^{\mathrm{T}} R u$, where $Q(x)$ is a positive semidefinite function and R is a symmetric positive definite matrix.

Definition 3 (Admissible Policy [44]): For the nonlinear dynamical system (1), a feedback policy $u : \mathbb{R}^n \to \mathbb{R}^m$ is said to be admissible on the compact $\Omega \subset \mathbb{R}^n$, denoted as $u \in \mathbb{R}^n$, provided that the following conditions hold.

1) $u(\cdot)$ is continuous on Ω.
2) $u(0) = 0$.
3) The closed-loop system is stable.
4) The performance $V(x_0)$ is finite.

Define the Hamiltonian as

$$H\left(u, x, \frac{\partial V_u(x)}{\partial x}\right) = r(x, u) + \left[\frac{\partial V_u(x)}{\partial x}\right]^{\mathrm{T}} [f(x) + g(x)u]. \tag{3}$$

The value function $V_u(x)$ can be obtained by solving the Bellman equation

$$0 = H\left(u, x, \frac{\partial V_u(x)}{\partial x}\right). \tag{4}$$

Define the optimal value function and optimal control policy as

$$\begin{aligned} V_*(x) &= \min_{u(\cdot)} \int_{t_0}^{\infty} r(x(\tau), u(\tau)) d\tau \\ u_*(x) &= \arg\min_{u(\cdot)} \int_{t_0}^{\infty} r(x(\tau), u(\tau)) d\tau. \end{aligned} \tag{5}$$

According to the optimal control theory [44], the optimal value function and the optimal control policy satisfy the HJB equation

$$\begin{aligned} 0 &= \min_{u(\cdot)} H\left(u, x, \frac{\partial V_*(x)}{\partial x}\right) \\ &= H\left(u_*, x, \frac{\partial V_*(x)}{\partial x}\right). \end{aligned} \tag{6}$$

Applying the stationary condition to the Hamiltoinan, one can obtain the optimal policy with the optimal value gradient as

$$u_*(x) = -\frac{1}{2} R^{-1} g^{\mathrm{T}}(x) \frac{\partial V_*(x)}{\partial x}. \tag{7}$$

Inserting $u_*(x)$ back into (6), the HJB equation can be equivalently expressed as

$$\begin{aligned} 0 = Q(x) &+ \left[\frac{\partial V_*(x)}{\partial x}\right]^{\mathrm{T}} f(x) \\ &- \frac{1}{4}\left[\frac{\partial V_*(x)}{\partial x}\right]^{\mathrm{T}} g(x) R^{-1} g^{\mathrm{T}}(x) \frac{\partial V_*(x)}{\partial x}. \end{aligned} \tag{8}$$

Remark 1: For linear system with the dynamics $\dot{x} = Ax + Bu$ and the reward function $r(x, u) = x^{\mathrm{T}} Q x + u^{\mathrm{T}} R u$, according to the linear quadratic optimal control theory, the value function with a stabilizing feedback policy $u = Kx$ takes the quadratic form $V_u(x) = x^{\mathrm{T}} P_u x$. Accordingly, the Hamiltonian for linear quadratic optimal control is defined as

$$\begin{aligned} H\left(u, x, \frac{\partial V_u(x)}{\partial x}\right) &= x^{\mathrm{T}} Q x + u^{\mathrm{T}} R u + x^{\mathrm{T}} A^{\mathrm{T}} P_u x + u^{\mathrm{T}} B^{\mathrm{T}} P_u x \\ &\quad + x^{\mathrm{T}} P_u A x + x^{\mathrm{T}} P_u B u. \end{aligned} \tag{9}$$

The optimal value function $V_*(x) = x^{\mathrm{T}} P x$ can be obtained by solving the algebraic Riccati equation

$$A^{\mathrm{T}} P + P A - P B R^{-1} B^{\mathrm{T}} P + Q = 0. \tag{10}$$

Applying the stationary condition, the optimal control can be calculated as $u_*(x) = -R^{-1} B^{\mathrm{T}} P x$. □

As investigated by the Hamiltonian-driven ADP framework [38], the fundamental issues in the optimal control problem are the policy evaluation and policy optimization. For the policy evaluation problem, an admissible policy is given and the design object is to learn the corresponding value function satisfying the Bellman equation. For the policy optimization problem, in addition to the adaptive critic network which approximates the optimal value function, the online actor is added to learn the optimal control policy simultaneously. In the following, the traditional Hamiltonian is extended as quasi-Hamiltonian and pseudo-Hamiltonian for policy evaluation and policy optimization problems with efficient experience replay.

III. Hamiltonian-Driven Efficient Policy Evaluation

A. Instantaneous Bellman Equation

Using the function approximators, such as neural network, the value function and value gradient for the fixed policy $u(x)$ can be denoted as the critic network, i.e.,

$$\begin{aligned} V_u(x) &= W_u^{\mathrm{T}} \phi(x) + \sigma_u(x) \\ \frac{\partial V_u(x)}{\partial x} &= \left[\frac{\partial \phi(x)}{\partial x}\right]^{\mathrm{T}} W_u + \frac{\partial \sigma_u(x)}{\partial x} \end{aligned} \tag{11}$$

where $W_u \in \mathbb{R}^{\ell}$ is the critic network weight, $\phi(x) \in \mathbb{R}^{\ell}$ and $((\partial \phi(x))/\partial x) \in \mathbb{R}^{\ell \times n}$ are the basis function and basis gradient, respectively, and $\sigma_u(x) \in \mathbb{R}$ and $((\partial \sigma_u(x))/\partial x) \in \mathbb{R}^{\ell}$ are the value function approximation residual and its gradient, respectively. Based on the universal approximation theorem, the ideal critic weight is defined as

$$W_u = \arg\min_{W} \left\| V_u(x) - W^{\mathrm{T}} \phi(x) \right\|$$

and the ideal value function approximation residual

$$\sigma_u(x) = \min_{W} \left\| V_u(x) - W^{\mathrm{T}} \phi(x) \right\|$$

is bounded in the sense that there exists a positive constant $\bar{\sigma}_u$ such that $\sup_{x \in \Omega} \|\sigma_u(x)\| \le \bar{\sigma}_u$.

Using the value function parameterization (11), the Bellman equation (4) can be rewritten using the critic network as

$$\begin{aligned} 0 = \left[\frac{\partial \sigma_u(x)}{\partial x}\right]^{\mathrm{T}} &[f(x) + g(x)u] \\ &+ W_u^{\mathrm{T}} \frac{\partial \phi(x)}{\partial x} [f(x) + g(x)u] + r(x, u). \end{aligned} \tag{12}$$

Definition 4: Consider the dynamical system (1) with the reward function $r(x, u)$ and the filtered signals in (15). The quasi-Hamiltonian is defined as

$$H_q(W, \varphi, r) = W^{\mathrm{T}}\varphi + r \tag{13}$$

for all $(W, \varphi, r) \in \mathbb{R}^{\ell} \times \mathbb{R}^{\ell} \times \mathbb{R}$.

Consider the following notations:

$$\delta_u(x(t)) = -\left[\frac{\partial \sigma_u(x(t))}{\partial x(t)}\right]^{\mathrm{T}}[f(x(t)) + g(x(t))u(t)]$$
$$\varphi_u(x(t), u(t)) = \frac{\partial \phi(x(t))}{\partial x(t)}[f(x(t)) + g(x(t))u(t)].$$

For notations simplicity, in the following, we denote $\varphi_u(x(t), u(t))$, $r(x(t), u(t))$, and $\delta_u(x(t))$ as $\varphi_u(t)$, $r_u(t)$, and $\delta_u(t)$, respectively. Then, the Bellman equation (12) implies that

$$\begin{aligned} H_q(W_u, \varphi_u(t), r_u(t)) &= W_u^{\mathrm{T}}\varphi_u(t) + r_u(t) \\ &= \delta_u(t). \end{aligned} \tag{14}$$

In (14), the signals $\varphi_u(t)$ and $r_u(t)$ depend on current system state $x(t)$ and control input $u(t)$. Therefore, (14) is referred to as the instantaneous Bellman equation.

B. Filtered Bellman Equation

To begin with, define the following filtered signals:

$$\begin{aligned} \Delta_u(t) &= \int_0^t e^{-\kappa(t-\tau)}\delta_u(x(\tau))d\tau \\ \Phi_u(t) &= \int_0^t e^{-\kappa(t-\tau)}\varphi_u(x(\tau), u(\tau))d\tau \\ Y_u(t) &= \int_0^t e^{-\kappa(t-\tau)}r(x(\tau), u(\tau))d\tau \end{aligned} \tag{15}$$

which can be equivalently obtained using the online filters as follows:

$$\begin{aligned} \dot{\Delta}_u(t) &= -\kappa\Delta_u(t) + \delta_u(x(t)), \quad \Delta_u(0) = 0 \\ \dot{\Phi}_u(t) &= -\kappa\Phi_u(t) + \varphi_u(x(t), u(t)), \quad \Phi_u(0) = 0 \\ \dot{Y}_u(t) &= -\kappa Y_u(t) + r(x(t), u(t)), \quad Y_u(0) = 0. \end{aligned} \tag{16}$$

From the signals definition in (15) and the instantaneous Bellman equation (14), one has

$$\begin{aligned} \Delta_u(t) &= \int_0^t e^{-\kappa(t-\tau)}\left[W_u^{\mathrm{T}}\varphi_u(\tau) + r_u(\tau)\right]d\tau \\ &= W_u^{\mathrm{T}}\int_0^t e^{-\kappa(t-\tau)}\varphi_u(\tau)d\tau + \int_0^t e^{-\kappa(t-\tau)}r_u(\tau)d\tau \\ &= W_u^{\mathrm{T}}\Phi_u(t) + Y_u(t). \end{aligned} \tag{17}$$

In (17), the terms $\Phi(t)$ and $Y_u(t)$ depend on the information of state and control input on the interval $[0, t]$. In addition, for admissible policy u, the boundedness of $\varphi_u(t)$ and $r_u(t)$ further implies the boundedness of $\Phi(t)$ and $Y_u(t)$. In this article, (17) is referred to as the filtered Bellman equation.

With the filtered signals in (17), the quasi-Hamiltonian is defined as

$$H_q(W_u, \Phi_u(t), Y_u(t)) = W_u^{\mathrm{T}}\Phi_u(t) + Y_u(t).$$

Accordingly, the filtered Bellman equation can be expressed as $\Delta_u(t) = H_q(W_u, \Phi_u(t), Y_u(t))$.

Remark 2: The quasi-Hamiltonian can be viewed as a finite-dimensional parameterization of the Hamiltonian $H(u, x, ((\partial V_u(x))/\partial x))$. The advantage of such parameterization can be summarized as follows.

1) The Hamiltonian $H(u, x, ((\partial V_u(x))/\partial x))$ depends on instantaneous data $\{x(t), u(t)\}$, as shown in (3). As investigated in [38], the Hamiltonian (3) can serve as the temporal difference learning error for continuous-time systems.
2) The quasi-Hamiltonian is defined based on the filtered signals $\{\Phi_u(t), Y_u(t)\}$, which are filters of historic data during the interval $[0, t]$. In this article, the quasi-Hamiltonian is viewed as the novel temporal difference error for critic learning.

Therefore, the quasi-Hamiltonian stands for more data collection and is used in the following design.

C. Efficient Critic Learning

For the policy evaluation with a given admissible policy, the critic learning aims to adapt the critic output

$$V_u(x) = \hat{W}_u^{\mathrm{T}}\phi(x) \tag{18}$$

to minimize the critic learning objective

$$J_u\big(\hat{W}_u(t)\big) = \frac{1}{2}\|e_u(t)\|^2 \tag{19}$$

where $e_u(t)$ is the instantaneous critic learning error defined as

$$e_u(t) = \hat{W}_u^{\mathrm{T}}(t)\varphi_u(t) + r_u(t). \tag{20}$$

One can observe that $e_u(t) \to \delta_u(t)$ as $\hat{W}_u(t) \to W_u$ according to the instantaneous Bellman equation (14). The conventional critic learning is designed based on the gradient descent, i.e.,

$$\begin{aligned} \dot{\hat{W}}_u(t) &= -\gamma_u\frac{\partial J_u\big(\hat{W}_u(t)\big)}{\partial \hat{W}_u(t)} \\ &= -\gamma_u\varphi_u(t)\big[\hat{W}_u^{\mathrm{T}}(t)\varphi_u(t) + r_u(t)\big]. \end{aligned} \tag{21}$$

The convergence of the critic learning depends on the PE condition on the signal $\varphi_u(t)$ [21], which is stringent for adaptive systems.

In the following, we consider the efficient critic learning based on the filtered Bellman equation (17). First, we define the composite critic learning error as

$$\varepsilon_u(\tau, t) = \hat{W}_u^{\mathrm{T}}(t)\Phi_u(\tau) + Y_u(\tau) \tag{22}$$

which combines the current critic weight estimation $\hat{W}_u(t)$ and historical information $\{\Phi_u(\tau), Y_u(\tau)\}$. In contrast to the instantaneous learning objective (19), the efficient critic learning aims to minimize the composite learning objective

$$J_u(t) = \frac{1}{2}\int_0^t \frac{\|\varepsilon_u(\tau, t)\|^2}{\left[1 + \Phi_u^{\mathrm{T}}(\tau)\Phi_u(\tau)\right]^2}d\tau. \tag{23}$$

Note that $\hat{W}_u(t)$ is independent of $\{\Phi_u(\tau), Y_u(\tau)\}$. Applying the chain rule, one can obtain the composite learning objective gradient as

$$\begin{aligned}\frac{\partial J_u(t)}{\partial \hat{W}_u(t)} &= \int_0^t \frac{\varepsilon_u(\tau,t)\Phi_u(t)}{\left[1+\Phi_u^{\mathrm{T}}(\tau)\Phi_u(\tau)\right]^2}d\tau\\ &= \int_0^t \frac{\Phi_u(\tau)\Phi_u^{\mathrm{T}}(\tau)}{\left[1+\Phi_u^{\mathrm{T}}(\tau)\Phi_u(\tau)\right]^2}d\tau \cdot \hat{W}_u(t)\\ &\quad+\int_0^t \frac{\Phi_u(\tau)Y_u(\tau)}{\left[1+\Phi_u^{\mathrm{T}}(\tau)\Phi_u(\tau)\right]^2}d\tau. \end{aligned}\tag{24}$$

Denote

$$E_u(t) = \int_0^t \frac{\Phi_u(\tau)\Phi_u^{\mathrm{T}}(\tau)}{\left[1+\Phi_u^{\mathrm{T}}(\tau)\Phi_u(\tau)\right]^2}d\tau \tag{25}$$

$$F_u(t) = \int_0^t \frac{\Phi_u(\tau)Y_u(\tau)}{\left[1+\Phi_u^{\mathrm{T}}(\tau)\Phi_u(\tau)\right]^2}d\tau \tag{26}$$

which can be obtained using the filter design

$$\begin{aligned}\dot{E}_u(t) &= \frac{\Phi_u(t)\Phi_u^{\mathrm{T}}(t)}{\left[1+\Phi_u^{\mathrm{T}}(t)\Phi_u(t)\right]^2}, \quad E_u(0)=0\\ \dot{F}_u(t) &= \frac{\Phi_u(t)Y_u(t)}{\left[1+\Phi_u^{\mathrm{T}}(t)\Phi_u(t)\right]^2}, \quad F_u(0)=0.\end{aligned}$$

Then, the composite learning objective gradient can be rewritten as

$$\frac{\partial J_u(t)}{\partial \hat{W}_u(t)} = E_u(t)\cdot\hat{W}_u(t) + F_u(t). \tag{27}$$

The critic learning is designed as

$$\begin{aligned}\dot{\hat{W}}_u(t) &= -\gamma_u \frac{\partial J_u(t)}{\partial \hat{W}_u(t)}\\ &= -\gamma_u\left[E_u(t)\cdot\hat{W}_u(t) + F_u(t)\right]. \end{aligned}\tag{28}$$

Theorem 1: Denote $\bar{\Phi}_u(t) = ((\Phi_u(t))/(1+\Phi_u^{\mathrm{T}}(t)\Phi_u(t)))$ and suppose that there exists T_u and β_u such that $\bar{\Phi}_u(t)$ satisfies the relaxed excitation condition. Then, for an admissible policy u, the following holds.

1) All the closed-loop signals L_∞-stable on $[0, T_u)$.
2) The critic weight learning error $\tilde{W}_u(t)$ converges to a small neighborhood of the origin exponentially on $[T_u, +\infty)$.

Proof: The first proposition is a standard result in adaptive control and can be referred to existing literature [45].

Next, we consider the second proposition. For the critic learning (28), we consider the Lyapunov candidate $L_u(\tilde{W}_u(t)) = (1/2)\tilde{W}_u^{\mathrm{T}}(t)\gamma_u^{-1}\tilde{W}_u(t)$. Differentiating $L_u(\tilde{W}_u(t))$ yields

$$\begin{aligned}\dot{L}_u\left(\tilde{W}_u(t)\right) &= -\tilde{W}_u^{\mathrm{T}}(t)\gamma_u^{-1}\dot{\hat{W}}_u(t)\\ &= \tilde{W}_u^{\mathrm{T}}(t)\left[E_u(t)\cdot\hat{W}_u(t)+F_u(t)\right]. \end{aligned}\tag{29}$$

From (17), one has

$$E_u(t)\cdot W_u(t) + F_u(t) = G_u(t) \tag{30}$$

where

$$G_u(t) = \int_0^t \frac{\Phi_u(\tau)\Delta_u(\tau)}{\left[1+\Phi_u^{\mathrm{T}}(\tau)\Phi_u(\tau)\right]^2}d\tau. \tag{31}$$

Then,

$$\begin{aligned}&E_u(t)\cdot\hat{W}_u(t)+F_u(t)\\ &\quad= E_u(t)\cdot\hat{W}_u(t)+F_u(t)\\ &\qquad+G_u(t)-E_u(t)\cdot W_u(t)-F_u(t)\\ &\quad= G_u(t)-E_u(t)\cdot\tilde{W}_u(t). \end{aligned}\tag{32}$$

Inserting (32) into (29) yields

$$\begin{aligned}&\dot{L}_u\left(\tilde{W}_u(t)\right)\\ &\quad\le \frac{1}{2}\gamma_G\left\|\tilde{W}_u(t)\right\| - \lambda_{\min}(E_u(t))\left\|\tilde{W}_u(t)\right\|^2\\ &\quad= -(1-\beta_u)\lambda_{\min}(E_u(t))\left\|\tilde{W}_u(t)\right\|^2 + \beta_u\lambda_{\min}(E_u(t))\\ &\qquad\times\left[\frac{\gamma_G}{2\beta_u\lambda_{\min}(E_u(t))}\left\|\tilde{W}_u(t)\right\| - \left\|\tilde{W}_u(t)\right\|^2\right]. \end{aligned}\tag{33}$$

Based on Young's inequality $2ab - a^2 \le b^2$, one has

$$\begin{aligned}&\frac{\gamma_G}{2\beta_u\lambda_{\min}(E_u(t))}\left\|\tilde{W}_u(t)\right\| - \left\|\tilde{W}_u(t)\right\|^2\\ &\qquad\qquad\le \frac{\gamma_G^2}{4\beta_u^2\lambda_{\min}^2(E_u(t))}. \end{aligned}\tag{34}$$

Finally, one has

$$\begin{aligned}\dot{L}_u\left(\tilde{W}_u(t)\right) &\le -cL_u\left(\tilde{W}_u(t)\right) + d\\ c &= 2\gamma_u(1-\beta_u)\lambda_{\min}(E_u(t))\\ d &= \frac{\gamma_G^2}{4\beta_u^2\lambda_{\min}^2(E_u(t))} \end{aligned}\tag{35}$$

which implies that the critic weight learning error $\tilde{W}_u(t)$ converges to a small neighborhood of the origin exponentially according to the Lyapunov stability extension theorem [46]. This completes the proof. ■

Remark 3: As shown in (28), the critic learning depends on the signals $E_u(t)$ and $F_u(t)$ to tune the critic weight $\hat{W}_u(t)$. This novel critic learning is a combination of both instantaneous data and historic data. First, as shown in (25), the signals $E_u(t)$ and $F_u(t)$ are defined based on the filtered signals $Y_u(t)$ and $\Phi_u(t)$. Second, based on (16), $Y_u(t)$ and $\Phi_u(t)$ are filtered signals of $\varphi_u(x(t), u(t))$ and $r(x(t), u(t))$ during the interval $[0, t]$. Therefore, the presented critic learning (28) also takes the historic data into account.

To this end, the Hamiltonian-driven ADP with efficient replay for the policy evaluation problem with a given admissible policy is shown in Fig. 1, where the quasi-Hamiltonian plays an important role in the critic learning.

IV. Hamiltonian-Driven Efficient Policy Optimization

A. Instantaneous HJB Equation

Using the neural network, the value function and value gradient for the fixed policy $u(x)$ can be denoted as the critic network, i.e.,

$$\begin{aligned}V_*(x) &= W_*^{\mathrm{T}}\phi(x) + \sigma_*(x)\\ \frac{\partial V_*(x)}{\partial x} &= \left[\frac{\partial\phi(x)}{\partial x}\right]^{\mathrm{T}}W_* + \frac{\partial\sigma_*(x)}{\partial x} \end{aligned}\tag{36}$$

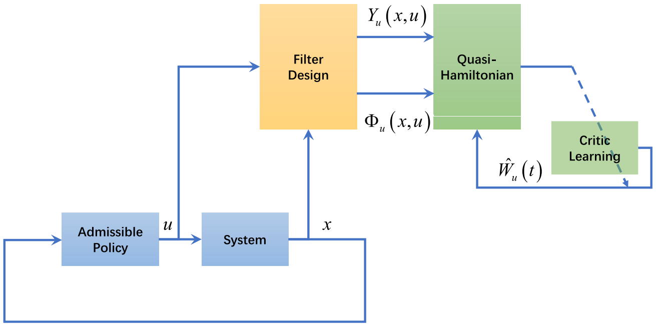


Fig. 1. Hamiltonian-driven ADP with efficient replay for critic learning.

where $W_* \in \mathbb{R}^\ell$ is the critic network weight, $\phi(x)$ and $((\partial\phi(x))/\partial x)$ are the basis function and basis gradient, respectively, and $\sigma_*(x)$ and $((\partial\sigma_*(x))/\partial x)$ are the value function approximation residual and its gradient, respectively. Based on the universal approximation theorem, the ideal critic weight is defined as

$$W_* = \arg\min_W \left\|V_*(x) - W^{\mathrm{T}}\phi(x)\right\|$$

and the ideal value function approximation residual

$$\sigma_*(x) = \min_W \left\|V_*(x) - W^{\mathrm{T}}\phi(x)\right\|$$

is bounded in the sense that there exists a positive constant $\bar{\sigma}_*$ such that $\sup_{x\in\Omega}\|\sigma_*(x)\| \le \bar{\sigma}_*$.

With the value function approximation, the HJB equation can be expressed as

$$\begin{aligned}0 = {} & Q(x) + W_*^{\mathrm{T}}\mu(x) - \frac{1}{4}W_*^{\mathrm{T}}\Theta(x)W_* \\ & - \frac{1}{4}\left[\frac{\partial\sigma_*(x)}{\partial x}\right]^{\mathrm{T}} g(x)R^{-1}g^{\mathrm{T}}(x)\frac{\partial\sigma_*(x)}{\partial x} \\ & - \frac{1}{2}\left[\frac{\partial\sigma_*(x)}{\partial x}\right]^{\mathrm{T}} g(x)R^{-1}g^{\mathrm{T}}(x)\left[\frac{\partial\phi(x)}{\partial x}\right]^{\mathrm{T}} W_* \\ & + \left[\frac{\partial\sigma_*(x)}{\partial x}\right]^{\mathrm{T}} f(x)\end{aligned} \tag{37}$$

with the following notations:

$$\begin{aligned}\mu(x) &= \frac{\partial\phi(x)}{\partial x}f(x) \\ \Theta(x) &= \frac{\partial\phi(x)}{\partial x}g(x)R^{-1}g^{\mathrm{T}}(x)\left[\frac{\partial\phi(x)}{\partial x}\right]^{\mathrm{T}}.\end{aligned} \tag{38}$$

Definition 5: Consider the dynamical system (1) with the signals $\{\mu(x), \Theta(x)\}$ in (38). The pseudo-Hamiltonian is defined as

$$H_p(W, Q, \mu, \Theta) = Q + W^{\mathrm{T}}\mu - \frac{1}{4}W^{\mathrm{T}}\Theta W \tag{39}$$

for all $\{W, Q, \mu, \Theta\} \in \mathbb{R}^\ell \times \mathbb{R} \times \mathbb{R}^\ell \times \mathbb{R}^{\ell\times\ell}$.

Then, the HJB equation (37) can be expressed using the pseudo-Hamiltonian as

$$\begin{aligned}& H_q(W_*, Q(x), \mu(x), \Theta(x)) \\ & \quad = Q(x) + W_*^{\mathrm{T}}\mu(x) - \frac{1}{4}W_*^{\mathrm{T}}\Theta(x)W_* \\ & \quad = \delta_*(x)\end{aligned} \tag{40}$$

with

$$\begin{aligned}\delta_*(x) = {} & \frac{1}{4}\left[\frac{\partial\sigma_*(x)}{\partial x}\right]^{\mathrm{T}} g(x)R^{-1}g^{\mathrm{T}}(x)\frac{\partial\sigma_*(x)}{\partial x} \\ & + \frac{1}{2}\left[\frac{\partial\sigma_*(x)}{\partial x}\right]^{\mathrm{T}} g(x)R^{-1}g^{\mathrm{T}}(x)\left[\frac{\partial\phi(x)}{\partial x}\right]^{\mathrm{T}} W_* \\ & - \left[\frac{\partial\sigma_*(x)}{\partial x}\right]^{\mathrm{T}} f(x).\end{aligned} \tag{41}$$

Note that in (41), the signals $Q(x)$, $\mu(x)$, and $\Theta(x)$ depend on the current state $x(t)$. Therefore, (41) is referred to as the instantaneous HJB equation.

Based on the value function approximation (36), denote the approximate optimal control as

$$\bar{u}_*(x) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\left[\frac{\partial\phi(x)}{\partial x}\right]^{\mathrm{T}} W_* \tag{42}$$

which satisfies

$$\begin{aligned}\delta_*(x) &= Q(x) + W_*^{\mathrm{T}}\mu(x) - \frac{1}{4}W_*^{\mathrm{T}}\Theta(x)W_* \\ &= W_*^{\mathrm{T}}\varphi_*(x, \bar{u}_*) + r(x, \bar{u}_*) \\ &= H_q(W_*, \varphi_*(x, \bar{u}_*), r(x, \bar{u}_*)) \\ &= H_p(W_*, Q(x), \mu(x), \Theta(x))\end{aligned} \tag{43}$$

where $\varphi_*(x, \bar{u}_*)$ is defined as

$$\begin{aligned}\varphi_*(x, \bar{u}_*) &= \frac{\partial\phi(x)}{\partial x}[f(x) + g(x)\bar{u}_*] \\ &= \mu(x) - \frac{1}{2}\Theta(x)W_*.\end{aligned} \tag{44}$$

B. Filtered HJB Equation

Define the following filtered signals:

$$\begin{aligned}Q_f(t) &= \int_0^t e^{-\kappa(t-\tau)}Q(x(\tau))d\tau \\ \Theta_f(t) &= \int_0^t e^{-\kappa(t-\tau)}\Theta(x(\tau))d\tau \\ \mu_f(t) &= \int_0^t e^{-\kappa(t-\tau)}\mu(x(\tau))d\tau.\end{aligned} \tag{45}$$

The pseudo-Hamiltonian with the above filtered signals is

$$\begin{aligned}& H_q\big(W_*, Q_f, \mu_f, \Theta_f\big) \\ & \quad = Q_f(t) + W_*^{\mathrm{T}}\mu_f(t) - \frac{1}{4}W_*^{\mathrm{T}}\Theta_f(t)W_*.\end{aligned} \tag{46}$$

Based on the instantaneous HJB equation (40), one can obtain the following filtered HJB equation as:

$$\begin{aligned}\delta_f(t) &= \int_0^t e^{-\kappa(t-\tau)}\delta_*(x(\tau))d\tau \\ &= H_q\big(W_*, Q_f, \mu_f, \Theta_f\big).\end{aligned} \tag{47}$$

C. Efficient Actor-Critic Learning

Note that in the filtered HJB equation (47), the terms $\{Q_f(t), \mu_f(t), \Theta_f(t)\}$ contain both the instantaneous and historical information of the system. In the following, we design the efficient actor–critic learning to solve the filtered HJB equation (47).

The adaptive critic network is designed as

$$\hat{V}_c(x(t)) = \hat{W}_c^{\mathrm{T}}(t)\phi(x(t)). \tag{48}$$

The actor network is designed as

$$u_a(x(t)) = -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x(t))\left[\frac{\partial\phi(x(t))}{\partial x(t)}\right]^{\mathrm{T}}\hat{W}_a(t). \tag{49}$$

The critic learning is defined as

$$\begin{aligned} e_c(t) &= \hat{W}_c^{\mathrm{T}}(t)\frac{\partial\phi(x(t))}{\partial x(t)}[f(x(t)) + g(x(t))u_a(t)] \\ &\quad + r(x(t), u_a(t)) \\ &= \hat{W}_c^{\mathrm{T}}(t)\mu(x(t)) - \frac{1}{2}\hat{W}_c^{\mathrm{T}}(t)\Theta(x(t))\hat{W}_a(t) \\ &\quad + Q(x(t)) + \frac{1}{4}\hat{W}_a^{\mathrm{T}}(t)\Theta(x(t))\hat{W}_a(t) \end{aligned} \tag{50}$$

which completely depends on the instantaneous system data. In contrast, based on the pseudo-Hamiltonian (39), we present the composite critic learning error as

$$\begin{aligned} \varepsilon_c(\tau, t) &= H_q\big(\hat{W}_c(t), Q_f(\tau), \mu_f(\tau), \Theta_f(\tau)\big) \\ &= \hat{W}_c^{\mathrm{T}}(t)\mu_f(\tau) - \frac{1}{2}\hat{W}_c^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t) \\ &\quad + Q_f(\tau) + \frac{1}{4}\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t). \end{aligned} \tag{51}$$

Denote

$$\begin{aligned} r_a(\tau, t) &= Q_f(\tau) + \frac{1}{4}\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t) \\ \varphi_a(\tau, t) &= \mu_f(\tau) - \frac{1}{2}\Theta_f(\tau)\hat{W}_a(t). \end{aligned} \tag{52}$$

Then,

$$\varepsilon_c(\tau, t) = \hat{W}_c^{\mathrm{T}}(t)\varphi_a(\tau, t) + r_a(\tau, t). \tag{53}$$

The critic learning objective is considered as

$$\bar{J}_c\big(\hat{W}_c(t)\big) = \frac{J_c\big(\hat{W}_c(t)\big)}{B_c(t)} \tag{54}$$

with

$$\begin{aligned} J_c\big(\hat{W}_c(t)\big) &= \frac{1}{2}\int_0^t \|\varepsilon_c(\tau, t)\|^2 d\tau \\ B_c(t) &= 1 + \int_0^t \big[\varphi_a^{\mathrm{T}}(\tau, t)\varphi_a(\tau, t)\big]^2 d\tau. \end{aligned} \tag{55}$$

Using the gradient descent rule, the critic weight update is designed as

$$\begin{aligned} \dot{\hat{W}}_c(t) &= -\gamma_c\frac{\partial\bar{J}_c\big(\hat{W}_c(t)\big)}{\partial\hat{W}_c(t)} \\ &= -\gamma_c\frac{1}{B_c(t)}\frac{\partial J_c\big(\hat{W}_c(t)\big)}{\partial\hat{W}_c(t)}. \end{aligned} \tag{56}$$

The details about calculation of $((\partial J_c(\hat{W}_c(t)))/(\partial\hat{W}_c(t)))$ are provided in Appendix A.

For the actor network (49), we design

$$\dot{\hat{W}}_a(t) = \gamma_a\left\{-K_{\mathrm{aa}}\hat{W}_a(t) + \frac{K_{\mathrm{ca}}F_c(t)}{B_c(t)} + \frac{F_a(t)}{4[B_c(t)]^2}\right\} \tag{57}$$

with

$$\begin{aligned} F_a(t) &= \int_0^t \Theta_f(\tau)\hat{W}_a(t)\varphi_a^{\mathrm{T}}(t, \tau)\hat{W}_c(t)d\tau \\ F_c(t) &= \int_0^t \varphi_a^{\mathrm{T}}(t, \tau)\hat{W}_c(t)d\tau. \end{aligned} \tag{58}$$

The details about the calculation of $F_a(t)$ and $F_c(t)$ are provided in Appendix B.

Theorem 2: Suppose that there exists T_a and β_a such that

$$\int_0^{T_a} \varphi_a(\tau, t)\varphi_a^{\mathrm{T}}(\tau, t)d\tau \succ \beta_a \cdot I_{N\times N}. \tag{59}$$

Then, with the actor–critic learning in (56) and (57), the system state $x(t)$, the critic weight learning error $\tilde{W}_c(t)$, and the actor weight learning error $\tilde{W}_a(t)$ are uniformly ultimately bounded.

Proof: First, the boundedness of closed-loop signals during the interval $[0, T_s]$ can be obtained based on a classical adaptive learning analysis and interested readers can be referred to [21] and [45].

In the following, we investigate the convergence of closed-loop signals on the interval $[T_s, +\infty]$. Consider the Lyapunov candidate

$$\begin{aligned} L\big(x, \tilde{W}_c, \tilde{W}_a\big) &= V_*(x) + L_c\big(\tilde{W}_c\big) + L_a\big(\tilde{W}_a\big) \\ L_c\big(\tilde{W}_c\big) &= \frac{1}{2}\tilde{W}_c^{\mathrm{T}}\gamma_c^{-1}\tilde{W}_c, \quad L_a\big(\tilde{W}_a\big) = \frac{1}{2}\tilde{W}_a^{\mathrm{T}}\gamma_a^{-1}\tilde{W}_a. \end{aligned} \tag{60}$$

With the online actor (49), the system dynamics is

$$\dot{x} = f(x) - \frac{1}{2}g(x)R^{-1}g^{\mathrm{T}}(x(t))\left[\frac{\partial\phi(x(t))}{\partial x(t)}\right]^{\mathrm{T}}\hat{W}_a(t). \tag{61}$$

Considering the value function approximation (36) and applying the chain rule, one has

$$\begin{aligned} \dot{V}_*(x) &= \left[\frac{\partial V_*(x(t))}{\partial x(t)}\right]\dot{x} \\ &= W_*^{\mathrm{T}}\frac{\partial\phi(x(t))}{\partial x(t)}f(x) - \frac{1}{2}W_*^{\mathrm{T}}\frac{\partial\phi(x(t))}{\partial x(t)}g(x)R^{-1} \\ &\quad \times g^{\mathrm{T}}(x(t))\left[\frac{\partial\phi(x(t))}{\partial x(t)}\right]^{\mathrm{T}}\hat{W}_a(t) + \delta_a \\ &= W_*^{\mathrm{T}}\mu(x) - \frac{1}{2}W_*^{\mathrm{T}}\Theta(x)\hat{W}_a(t) + \delta_a \end{aligned} \tag{62}$$

with

$$\delta_a = \left[\frac{\partial\phi(x)}{\partial x}\right]^{\mathrm{T}}\left[f(x) - \frac{1}{2}g(x)R^{-1}g^{\mathrm{T}}(x)\left[\frac{\partial\phi(x)}{\partial x}\right]^{\mathrm{T}}\hat{W}_a(t)\right]$$

satisfying

$$\|\delta_a\| \le L_{d\phi}L_f\|x\| + \frac{1}{2}\lambda_{\min}(R)L_g^2L_{d\phi}^2\big(\big\|\tilde{W}_a(t)\big\| + \|W_*\|\big).$$

Adding and subtracting $(1/2)W_*^{\mathrm{T}}\Theta(x)W_*$ to (62) yields

$$\begin{aligned}\dot{V}_*(x) &= W_*^{\mathrm{T}}\varphi_*(x,\bar{u}_*) + \frac{1}{2}W_*^{\mathrm{T}}\Theta(x)\tilde{W}_a(t) + \delta_a \\ &= \delta_*(x) - Q(x) - \frac{1}{4}W_*^{\mathrm{T}}\Theta(x)W_* \\ &\quad + \frac{1}{2}W_*^{\mathrm{T}}\Theta(x)\tilde{W}_a(t) + \delta_a. \qquad (63)\end{aligned}$$

As shown in (56) and (77), the critic learning can be rewritten as

$$\dot{\hat{W}}_c(t) = -\gamma_c\frac{1}{[B_c(t)]^2}\int_0^t \varphi_a(t,\tau)\varepsilon_c(t,\tau)d\tau. \qquad (64)$$

From (53) and (47), the critic learning error satisfies

$$\begin{aligned}\varepsilon_c(\tau,t) &= Q_f(\tau) + \frac{1}{4}\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t) \\ &\quad + \hat{W}_c^{\mathrm{T}}(t)\varphi_a(\tau,t) \\ &= Q_f(\tau) + \frac{1}{4}\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t) \\ &\quad + \hat{W}_c^{\mathrm{T}}(t)\varphi_a(\tau,t) + \delta_f(\tau) - W_*^{\mathrm{T}}\varphi_*(\tau) \\ &\quad - Q_f(\tau) - \frac{1}{4}W_*^{\mathrm{T}}\Theta_f(\tau)W_* \\ &= \frac{1}{4}\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t) - \frac{1}{4}W_*^{\mathrm{T}}\Theta_f(\tau)W_* \\ &\quad + \hat{W}_c^{\mathrm{T}}(t)\varphi_a(\tau,t) - W_*^{\mathrm{T}}\varphi_*(\tau) + \delta_f(\tau) \\ &= \frac{1}{4}\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t) - \frac{1}{4}W_*^{\mathrm{T}}\Theta_f(\tau)W_* \\ &\quad - \frac{1}{2}\hat{W}_c^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t) + \frac{1}{2}W_*^{\mathrm{T}}\Theta_f(\tau)W_* \\ &\quad - \tilde{W}_c^{\mathrm{T}}(t)\mu_f(\tau) + \delta_f(\tau) \\ &= \frac{1}{2}\tilde{W}_c^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t) + \frac{1}{4}\tilde{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\tilde{W}_a(t) \\ &\quad - \tilde{W}_c^{\mathrm{T}}(t)\mu_f(\tau) + \delta_f(\tau) \\ &= \delta_f(\tau) - \tilde{W}_c^{\mathrm{T}}(t)\varphi_a(\tau,t) + \frac{1}{4}\tilde{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\tilde{W}_a(t).\end{aligned}$$

Inserting the above equation into the critic learning (64) yields

$$\begin{aligned}\dot{\hat{W}}_c(t) &= \gamma_c\frac{\int_0^t \varphi_a(t,\tau)\varphi_a^{\mathrm{T}}(\tau,t)d\tau}{[B_c(t)]^2}\tilde{W}_c(t) \\ &\quad - \gamma_c\frac{\int_0^t \varphi_a(t,\tau)\tilde{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\tilde{W}_a(t)d\tau}{4[B_c(t)]^2} \\ &\quad - \gamma_c\frac{\int_0^t \varphi_a(t,\tau)\delta_f(\tau)d\tau}{[B_c(t)]^2}. \qquad (65)\end{aligned}$$

Then,

$$\begin{aligned}\dot{L}_c\big(\tilde{W}_c\big) &= -\tilde{W}_c^{\mathrm{T}}(t)\gamma_c^{-1}\dot{\hat{W}}_c(t) \\ &= -\tilde{W}_c^{\mathrm{T}}(t)\frac{\int_0^t \varphi_a(t,\tau)\varphi_a^{\mathrm{T}}(\tau,t)d\tau}{[B_c(t)]^2}\tilde{W}_c(t) \\ &\quad + \tilde{W}_c^{\mathrm{T}}(t)\frac{\int_0^t \varphi_a(t,\tau)\delta_f(\tau)d\tau}{[B_c(t)]^2} + \frac{\Lambda(t,\tau)}{4[B_c(t)]^2} \qquad (66)\end{aligned}$$

where

$$\Lambda(t,\tau) = \int_0^t \tilde{W}_c^{\mathrm{T}}(t)\varphi_a(t,\tau)\tilde{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\tilde{W}_a(t)d\tau. \qquad (67)$$

Consider the fact that

$$W_* = \hat{W}_c(t) + \tilde{W}_c(t) = \hat{W}_a(t) + \tilde{W}_a(t). \qquad (68)$$

Then, the term $\Lambda(t,\tau)$ can be further written as

$$\begin{aligned}\Lambda(t,\tau) &= -\int_0^t \tilde{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)W_*\varphi_a^{\mathrm{T}}(t,\tau)W_*d\tau \\ &\quad + \int_0^t \tilde{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)W_*\varphi_a^{\mathrm{T}}(t,\tau)\tilde{W}_c(t)d\tau \\ &\quad + \int_0^t \tilde{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\tilde{W}_c(t)\varphi_a^{\mathrm{T}}(t,\tau)W_*d\tau \\ &\quad + \tilde{W}_a^{\mathrm{T}}(t)F_a(t). \qquad (69)\end{aligned}$$

In addition, for the actor learning (57), one has

$$\begin{aligned}\dot{L}_a\big(\tilde{W}_a\big) &= -\tilde{W}_a^{\mathrm{T}}(t)\gamma_a^{-1}\dot{\hat{W}}_a(t) \\ &= \tilde{W}_a^{\mathrm{T}}(t)K_{\mathrm{aa}}\hat{W}_a(t) - \tilde{W}_a^{\mathrm{T}}(t)\frac{F_a(t)}{4[B_c(t)]^2} \\ &\quad - \tilde{W}_a^{\mathrm{T}}(t)K_{\mathrm{ca}}\frac{\int_0^t \varphi_a^{\mathrm{T}}(t,\tau)\hat{W}_c(t)d\tau}{B_c(t)}. \qquad (70)\end{aligned}$$

Recall the fact in (68), and then,

$$\begin{aligned}\dot{L}_a\big(\tilde{W}_a\big) &= \tilde{W}_a^{\mathrm{T}}(t)K_{\mathrm{aa}}\hat{W}_a(t) - \tilde{W}_a^{\mathrm{T}}(t)\frac{F_a(t)}{4[B_c(t)]^2} \\ &\quad - \tilde{W}_a^{\mathrm{T}}(t)K_{\mathrm{ca}}\frac{\int_0^t \varphi_a^{\mathrm{T}}(t,\tau)d\tau}{B_c(t)}W_* \\ &\quad + \tilde{W}_a^{\mathrm{T}}(t)K_{\mathrm{ca}}\frac{\int_0^t \varphi_a^{\mathrm{T}}(t,\tau)d\tau}{B_c(t)}\tilde{W}_c(t). \qquad (71)\end{aligned}$$

Collecting the results in (63), (66), and (71), one has

$$\begin{aligned}\dot{L}(Z) &\le -Z^{\mathrm{T}}\begin{bmatrix} P_{\mathrm{xx}} & 0 & 0 \\ 0 & P_{\mathrm{cc}} & P_{\mathrm{ca}} \\ 0 & P_{\mathrm{ac}} & P_{\mathrm{aa}} \end{bmatrix}Z + Z^{\mathrm{T}}d + c \\ &= -Z^{\mathrm{T}}P_Z Z + Z^{\mathrm{T}}d + c \qquad (72)\end{aligned}$$

with

$$\begin{aligned}P_{\mathrm{xx}} &= qI_{n\times n}, \quad P_{\mathrm{aa}} = K_{\mathrm{aa}} \\ P_{\mathrm{cc}} &= \frac{\int_0^t \varphi_a(t,\tau)\varphi_a^{\mathrm{T}}(\tau,t)d\tau}{[B_c(t)]^2} \\ P_{\mathrm{ac}} &= P_{\mathrm{ca}}^{\mathrm{T}} = -\frac{K_{\mathrm{ca}}\int_0^t \varphi_a^{\mathrm{T}}(t,\tau)d\tau}{2B_c(t)} \\ &\quad - \frac{\int_0^t \Theta_f(\tau)W_*\varphi_a^{\mathrm{T}}(t,\tau)d\tau}{8[B_c(t)]^2} \\ &\quad - \frac{\int_0^t \Theta_f(\tau)\varphi_a^{\mathrm{T}}(t,\tau)W_*d\tau}{8[B_c(t)]^2} \\ d_x &= L_{d\phi}L_f, \quad d_c = \frac{\int_0^t \varphi_a(t,\tau)\delta_f(\tau)d\tau}{[B_c(t)]^2} \\ d_a &= -\frac{\int_0^t \Theta_f(\tau)W_*\varphi_a^{\mathrm{T}}(t,\tau)W_*d\tau}{4[B_c(t)]^2} + K_{\mathrm{aa}}W_* \\ &\quad - \frac{K_{\mathrm{ca}}\int_0^t \varphi_a^{\mathrm{T}}(t,\tau)d\tau}{B_c(t)}W_* + \frac{1}{2}\Theta(x)W_* \\ &\quad + \frac{1}{2}\lambda_{\min}(R)L_g^2L_{d\phi}^2 \\ c &= \delta_*(x) - \frac{1}{4}W_*^{\mathrm{T}}\Theta(x)W_* \\ &\quad + \frac{1}{2}\lambda_{\min}(R)L_g^2L_{d\phi}^2\|W_*\|. \qquad (73)\end{aligned}$$

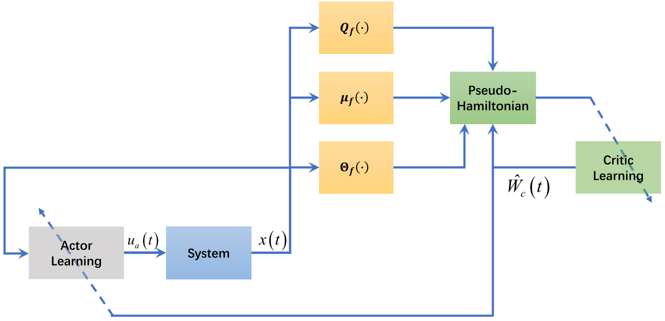


Fig. 2. Hamiltonian-driven ADP with efficient replay for actor–critic learning.

Note that $P_{\rm xx}$ and $P_{\rm aa}$ are positive definite matrices. From condition (59), the matrix $P_{\rm cc}$ is also positive definite. In addition, the Shur complement of $P_{\rm cc}$ is positive definite, i.e., $P_{\rm cc} - P_{\rm ca}P_{\rm aa}^{-1}P_{\rm ac} > 0$, provided that $K_{\rm ca} \ll K_{\rm aa}$. Then, the matrix P_Z can be guaranteed to be positive definite. In addition, with the normalization term $B_c(t)$, the terms d is bounded. Based on the Lyapunov extension theorem [46], the augmented state Z is uniformly ultimately bounded. This completes the proof. ■

To this end, the Hamiltonian-driven ADP with efficient replay for the policy optimization problem is shown in Fig. 2, where the pseudo-Hamiltonian lays the foundation for the actor–critic learning.

Remark 4: In Sections III and IV, we extend the traditional Hamiltonian to quasi-Hamiltonian and pseudo-Hamiltonian, respectively. With the quasi-Hamiltonian, the Bellman equation is parameterized as a linear equation of the unknown critic weight W_u. For the HJB equation, the pseudo-Hamiltonian is presented to produce the filtered HJB equation, which is quadratic in the optimal weight W_*, which is compared to the linear parameterized model [33], [35], [36], [37]. In addition, the presented learning does not require probing noise to guarantee the convergence to the optimum, which can obviate the unnecessary oscillation in the closed-loop signals. □

Remark 5: In existing ADP methods [47], [48], the experience-replay technique is based on concurrent learning [34], of which the historic data are collected and updated based on a matrix rank condition. However, the matrix rank condition has to be checked throughout the implementation to guarantee that the collected data are sufficiently excited, which might increase the computational complexity. In contrast, the experience-replay technique in this article is based on the online filters design (15) and (45), which can be efficiently obtained and has lower computational complexity.

V. Simulation Study

In this section, we investigate the simulation example of optimal stabilization problem with nonlinear dynamical system to demonstrate the efficacy of the presented efficient Hamiltonian-driven ADP. To be specific, the van der pol's oscillator is considered with the dynamics

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -x_1 - \frac{1}{2}(1 - x_1^2)x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ x_1 \end{bmatrix} u \tag{74}$$

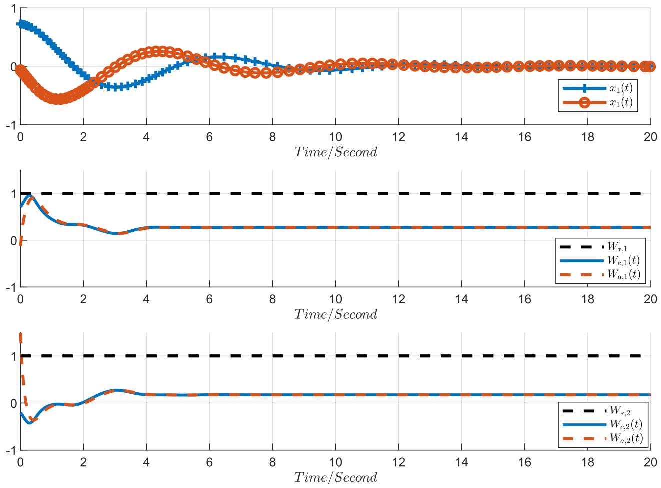


Fig. 3. State evolution and online actor–critic learning for policy optimization problem using conventional approach [21] without probing noise. The optimal critic weight is denoted as $W_* = [W_{*,1} \ W_{*,2}]^{\rm T} = [1 \ 1]^{\rm T}$, as illustrated by the black dashed lines. The adaptive critic learning is denoted as $W_c(t) = [W_{c,1}(t) \ W_{c,2}(t)]^{\rm T}$, as illustrated by the blue solid lines. The online actor learning is denoted as $W_a(t) = [W_{a,1}(t) \ W_{a,2}(t)]^{\rm T}$ and illustrated by the red solid lines.

where $x = [\, x_1 \ \ x_2 \,] \in \mathbb{R}^2$ is the system state and $u \in \mathbb{R}$ is the control input. According to [49], with the performance $V_u(x_0) = \int_{t_0}^{\infty} r(x(\tau), u(\tau))d\tau$ and the reward function $r(x, u) = x_1^2 + u^2$, the optimal value function satisfying the HJB equation

$$0 = x_1^2 + \left[\frac{\partial V_*(x)}{\partial x}\right]^{\rm T} \left[x_2 {-} x_1 - \frac{1}{2}(1 - x_1^2)x_2 \right] - \frac{1}{4}\left[\frac{\partial V_*(x)}{\partial x}\right]^{\rm T} \begin{bmatrix} 0 \\ x_1 \end{bmatrix} \begin{bmatrix} 0 & x_1 \end{bmatrix} \frac{\partial V_*(x)}{\partial x} \tag{75}$$

can be determined as $V_*(x) = x_1^2 + x_2^2$. On this basis, the optimal control policy can be further obtained as $u_* = -x_1 x_2$. In addition, the optimal value function can be parameterized as $V_*(x) = W_*^{\rm T}\phi(x)$ using the basis function $\phi(x) = [\, x_1^2 \ \ x_2^2 \,]^{\rm T}$ and optimal weight $W_* = [\, 1 \ \ 1 \,]^{\rm T}$. Note that with the above value function parameterization, the value function approximation residual $\sigma_*(x)$ is identically equal to zero.

The critic learning rate γ_c should be bigger than the actor learning rate γ_a to guarantee the closed-loop stability and actor–critic learning convergence, as investigated in Theorems 1 and 2. In addition, the filter parameter κ should be small enough to guarantee that the relaxed excitation is satisfied.

Case 1 (Conventional Actor-Critic Learning Without Probing Noise): With the conventional actor–critic learning [21], the evolution of system state and actor–critic weights is shown in Fig. 3. One can observe that in the absence of PE condition, the convergence of the actor–critic learning to the optimal weight cannot be guaranteed.

Case 2 (Conventional Actor-Critic Learning With Probing Noise): To satisfy the PE condition, we add the probing noise $n_u(t)$ into the actor network, i.e.,

$$n_u(t) = e^{-\frac{t}{10}}[\sin(t) + 2\sin(2t)]. \tag{76}$$

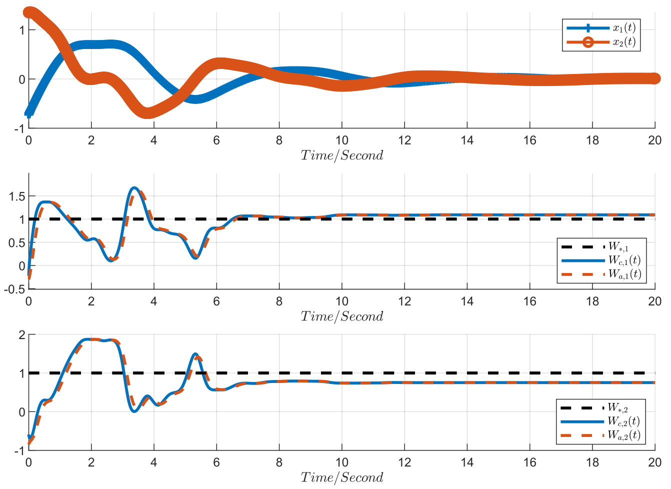


Fig. 4. State evolution and online actor–critic learning for policy optimization problem using conventional approach [21] with probing noise. The optimal critic weight is denoted as $W_* = [W_{*,1}\ W_{*,2}]^{\mathrm{T}} = [1\ 1]^{\mathrm{T}}$, as illustrated by the black dashed lines. The adaptive critic learning is denoted as $W_c(t) = [W_{c,1}(t)\ W_{c,2}(t)]^{\mathrm{T}}$, as illustrated by the blue solid lines. The online actor learning is denoted as $W_a(t) = [W_{a,1}(t)\ W_{a,2}(t)]^{\mathrm{T}}$ and illustrated by the red solid lines.

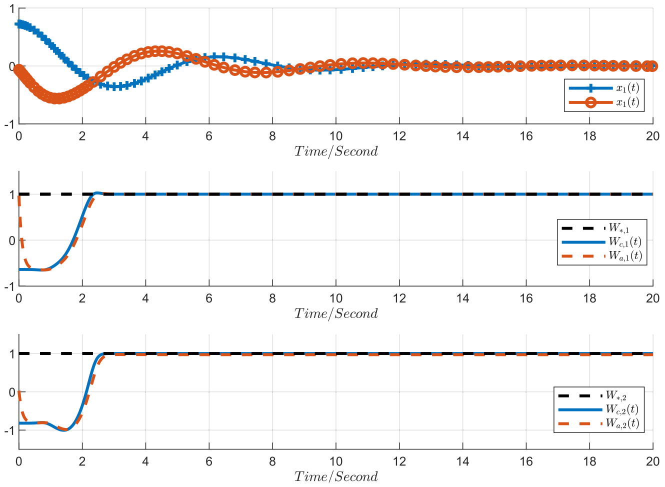


Fig. 5. State evolution and online actor–critic learning for policy optimization problem using presented efficient Hamiltonian-driven ADP. The optimal critic weight is denoted as $W_* = [W_{*,1}\ W_{*,2}]^{\mathrm{T}} = [1\ 1]^{\mathrm{T}}$, as illustrated by the black dashed lines. The adaptive critic learning is denoted as $W_c(t) = [W_{c,1}(t)\ W_{c,2}(t)]^{\mathrm{T}}$, as illustrated by the blue solid lines. The online actor learning is denoted as $W_a(t) = [W_{a,1}(t)\ W_{a,2}(t)]^{\mathrm{T}}$ and illustrated by the red solid lines.

Accordingly, the system response is shown in Fig. 4. One can observe that the satisfaction of PE condition contributes to the actor–critic learning convergence. However, the probing noise also brings undesired oscillation in the evolution of both system state and online actor–critic learning. In addition, the probing noise also results in a bias in the actor–critic learning and the actor–critic weights converges to a neighborhood set around the origin.

Case 3 (Hamiltonian-Driven ADP With Experience-Replay): To relax the stringent PE condition while ensuring the actor–critic learning convergence, we apply the presented Hamiltonian-driven ADP with efficient experience-replay technique to the system, where the results are shown in Fig. 5. Finally, one can observe that the actor–critic weights converge exponentially to the optimal critic weights within 1.5 s. In addition, compared to the case using probing noise, one can observe that there is no undesired oscillation in the closed-loop signals.

VI. Conclusion

In this article, we extend the classical Hamiltonian based on the value function approximation to develop efficient data-based ADP learning. The quasi-Hamiltonian and the pseudo-Hamiltonian are proposed to cope with the policy evaluation and policy optimization problems, respectively. With the quasi-Hamiltonian and pseudo-Hamiltonian, the historical and instantaneous data are simultaneously utilized through additional filters design. Compared with the stringent PE condition, the presented efficient Hamiltonian-driven ADP ensures the learning convergence with a relaxed excitation condition. In addition, the probing noise is also obviated to cancel the undesired oscillation in the closed-loop responses. Simulation examples are investigated to verify the presented design in the end. Future works aim to extend the presented design to distributed optimization and control problems.

Appendix A
Calculation of $((\partial J_c(\hat{W}_c(t)))/(\partial \hat{W}_c(t)))$

Applying the chain rule to the critic learning object (54) and considering (52), one has

$$\begin{aligned}\frac{\partial J_c\big(\hat{W}_c(t)\big)}{\partial \hat{W}_c(t)} &= \int_0^t \varphi_a(t,\tau)\varepsilon_c(t,\tau)d\tau\\ &= \int_0^t \varphi_a(t,\tau)\big[\varphi_a^{\mathrm{T}}(t,\tau)\hat{W}_c(t) + r_a(t,\tau)\big]d\tau.\end{aligned}$$

Define

$$\begin{aligned}M_c(t) &= \int_0^t \varphi_a(t,\tau)\varphi_a^{\mathrm{T}}(t,\tau)d\tau\\ N_c(t) &= \int_0^t \varphi_a(t,\tau)r_a(t,\tau)d\tau. \qquad (77)\end{aligned}$$

Then, the objective gradient can be further expressed as

$$\frac{\partial J_c\big(\hat{W}_c(t)\big)}{\partial \hat{W}_c(t)} = M_c(t)\cdot \hat{W}_c(t) + N_c(t). \qquad (78)$$

From (52), one has

$$\begin{aligned}M_c(t) = &\frac{1}{4}\int_0^t \big[\Theta_f(\tau)\hat{W}_a(t)\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\big]d\tau\\ &+ \int_0^t \big[\mu_f(\tau)\mu_f^{\mathrm{T}}(\tau)\big]d\tau\\ &- \frac{1}{2}\int_0^t \big[\mu_f(\tau)\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\big]d\tau\\ &- \frac{1}{2}\int_0^t \big[\Theta_f(\tau)\hat{W}_a(t)\mu_f^{\mathrm{T}}(\tau)\big]d\tau. \qquad (79)\end{aligned}$$

Using the Kronecker property, one has

$$\begin{aligned}
&\int_0^t \big[\Theta_f(\tau)\hat{W}_a(t)\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\big]d\tau \\
&\quad= \mathrm{vec}_{M,M}^{-1}\big[\psi_{\Theta\Theta}(t)\cdot \mathrm{vec}\big(\hat{W}_a(t)\hat{W}_a^{\mathrm{T}}(t)\big)\big] \\
&\int_0^t \big[\mu_f(\tau)\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\big]d\tau \\
&\quad= \int_0^t \mathrm{vec}_{M,M}^{-1}\big[\psi_{\mu\Theta}(t)\cdot \mathrm{vec}\big(\hat{W}_a^{\mathrm{T}}(t)\big)\big]d\tau \\
&\int_0^t \big[\Theta_f(\tau)\hat{W}_a(t)\mu_f^{\mathrm{T}}(\tau)\big]d\tau \\
&\quad= \mathrm{vec}_{M,M}^{-1}\big[\psi_{\Theta\mu}(t)\cdot \mathrm{vec}\big(\hat{W}_a(t)\big)\big]
\end{aligned} \tag{80}$$

where

$$\begin{aligned}
\psi_{\Theta\Theta}(t) &= \int_0^t \big[\Theta_f(\tau)\otimes\Theta_f(\tau)\big]d\tau \\
\psi_{\mu\Theta}(t) &= \int_0^t \big[\Theta_f(\tau)\otimes\mu_f(\tau)\big]d\tau \\
\psi_{\Theta\mu}(t) &= \int_0^t \big[\mu_f(\tau)\otimes\Theta_f(\tau)\big]d\tau.
\end{aligned} \tag{81}$$

Denote $\psi_{\mu\mu}(t) = \int_0^t \mu_f(\tau)\mu_f^{\mathrm{T}}(\tau)d\tau$. Then,

$$\begin{aligned}
M_c(t) = {}& \frac{1}{4}\mathrm{vec}_{M,M}^{-1}\big[\psi_{\Theta\Theta}(t)\cdot \mathrm{vec}\big(\hat{W}_a(t)\hat{W}_a^{\mathrm{T}}(t)\big)\big] \\
&- \frac{1}{2}\int_0^t \mathrm{vec}_{M,M}^{-1}\big[\psi_{\mu\Theta}(t)\cdot \mathrm{vec}\big(\hat{W}_a^{\mathrm{T}}(t)\big)\big]d\tau \\
&- \frac{1}{2}\mathrm{vec}_{M,M}^{-1}\big[\psi_{\Theta\mu}(t)\cdot \mathrm{vec}\big(\hat{W}_a(t)\big)\big] \\
&+ \psi_{\mu\mu}(t).
\end{aligned} \tag{82}$$

Second, inserting (52) into (77) yields

$$\begin{aligned}
N_c(t) = {}& \int_0^t \big[\mu_f(\tau)Q_f(\tau)\big]d\tau \\
&+\frac{1}{4}\int_0^t \big[\mu_f(\tau)\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t)\big]d\tau \\
&-\frac{1}{2}\int_0^t \big[\Theta_f(\tau)\hat{W}_a(t)Q_f(\tau)\big]d\tau \\
&-\frac{1}{8}\int_0^t \big[\Theta_f(\tau)\hat{W}_a(t)\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t)\big]d\tau.
\end{aligned}$$

Using the Kronecker property, one has

$$\begin{aligned}
&\int_0^t \big[\mu_f(\tau)\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t)\big]d\tau \\
&\quad= \mathrm{vec}_{M,M}^{-1}\big[\psi_{\mu\Theta}(t)\cdot \mathrm{vec}\big(\hat{W}_a^{\mathrm{T}}(t)\big)\big]\cdot\hat{W}_a(t) \\
&\int_0^t \big[\Theta_f(\tau)\hat{W}_a(t)Q_f(\tau)\big]d\tau \\
&\quad= \mathrm{vec}_{M,1}^{-1}\big[\psi_{\Theta Q}(t)\cdot \mathrm{vec}\big(\hat{W}_a(t)\big)\big] \\
&\int_0^t \big[\Theta_f(\tau)\hat{W}_a(t)\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_a(t)\big]d\tau \\
&\quad= \mathrm{vec}_{M,1}^{-1}\big[\psi_{\Theta\Theta}(t)\cdot \mathrm{vec}\big(\hat{W}_a(t)\hat{W}_a^{\mathrm{T}}(t)\big)\big]\cdot\hat{W}_a(t)
\end{aligned}$$

where $\psi_{\Theta Q}(t) = \int_0^t [\Theta_f(\tau)Q_f(\tau)]d\tau$. Denote $\psi_{\mu Q}(t) = \int_0^t [\mu_f(\tau)Q_f(\tau)]d\tau$. Then, $N_c(t)$ can be calculated as

$$\begin{aligned}
N_c(t) = {}& \psi_{\mu Q}(t) - \frac{1}{2}\mathrm{vec}_{M,1}^{-1}\big[\psi_{\Theta Q}(t)\cdot \mathrm{vec}\big(\hat{W}_a(t)\big)\big] \\
&+ \frac{1}{4}\mathrm{vec}_{M,M}^{-1}\big[\psi_{\mu\Theta}(t)\cdot \mathrm{vec}\big(\hat{W}_a^{\mathrm{T}}(t)\big)\big]\cdot\hat{W}_a(t) \\
&- \frac{1}{8}\mathrm{vec}_{M,1}^{-1}\big[\psi_{\Theta\Theta}(t)\cdot \mathrm{vec}\big(\hat{W}_a(t)\hat{W}_a^{\mathrm{T}}(t)\big)\big]\cdot\hat{W}_a(t).
\end{aligned} \tag{83}$$

Appendix B
Calculation of $F_a(t)$ and $F_c(t)$

Based on the definition of $\varphi_a(t,\tau)$ in (52), the term $F_a(t)$ can be rewritten as

$$\begin{aligned}
F_a(t) = {}& \bigg[\int_0^t \Theta_f(\tau)\hat{W}_a(t)\mu_f^{\mathrm{T}}(\tau)d\tau\bigg]\cdot\hat{W}_c(t) \\
&- \frac{1}{2}\bigg[\int_0^t \Theta_f(\tau)\hat{W}_a(t)\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)d\tau\bigg]\cdot\hat{W}_c(t).
\end{aligned}$$

Recalling the facts in (80), one has

$$\begin{aligned}
F_a(t) = {}& \mathrm{vec}_{M,M}^{-1}\big[\psi_{\Theta\mu}(t)\cdot \mathrm{vec}\big(\hat{W}_a(t)\big)\big]\cdot\hat{W}_c(t) \\
&- \frac{1}{2}\mathrm{vec}_{M,M}^{-1}\big[\psi_{\Theta\Theta}(t)\cdot \mathrm{vec}\big(\hat{W}_a(t)\hat{W}_a^{\mathrm{T}}(t)\big)\big]\cdot\hat{W}_c(t).
\end{aligned}$$

Similarity, inserting the definition of $\varphi_a(t,\tau)$ (52) into F_c in (58) yields

$$\begin{aligned}
F_c(t) &= \int_0^t \bigg[\mu_f^{\mathrm{T}}(\tau) - \frac{1}{2}\hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\bigg]\hat{W}_c(t)d\tau \\
&= \int_0^t \mu_f^{\mathrm{T}}(\tau)\hat{W}_c(t)d\tau - \frac{1}{2}\int_0^t \hat{W}_a^{\mathrm{T}}(t)\Theta_f(\tau)\hat{W}_c(t)d\tau \\
&= \bigg[\int_0^t \mu_f^{\mathrm{T}}(\tau)d\tau\bigg]\hat{W}_c(t) - \frac{1}{2}\hat{W}_a^{\mathrm{T}}(t)\bigg[\int_0^t \Theta_f(\tau)d\tau\bigg] \\
&\quad \times \hat{W}_c(t).
\end{aligned}$$

Denote the signals $\Theta_{\mathrm{ff}}(t) = \int_0^t \Theta_f(\tau)d\tau$ and $\mu_{\mathrm{ff}}(t) = \int_0^t \mu_f(\tau)d\tau$.

To this end, $F_c(t)$ can be online calculated as

$$F_c(t) = \mu_{\mathrm{ff}}^{\mathrm{T}}(t)\hat{W}_c(t) - \frac{1}{2}\hat{W}_a^{\mathrm{T}}(t)\Theta_{\mathrm{ff}}(t)\hat{W}_c(t).$$

References

[1] D. Liberzon, *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton, NJ, USA: Princeton Univ. Press, 2012.

[2] A. E. Bryson and Y.-C. Ho, *Applied Optimal Control: Optimization, Estimation and Control*. New York, NY, USA: Taylor & Francis, 1969.

[3] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, vol. 703. Hoboken, NJ, USA: Wiley, 2007.

[4] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, vol. 17. Hoboken, NJ, USA: Wiley, 2013.

[5] L. Busoniu, R. Babuska, B. De Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Boca Raton, FL, USA: CRC Press, 2017.

[6] Y. Jiang and Z.-P. Jiang, *Robust Adaptive Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2017.

[7] D. V. Prokhorov and D. C. Wunsch, "Adaptive critic designs," *IEEE Trans. Neural Netw.*, vol. 8, no. 5, pp. 997–1007, Sep. 1997.

[8] F.-Y. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: An introduction," *IEEE Comput. Intell. Mag.*, vol. 4, no. 2, pp. 39–47, May 2009.

[9] J. Lee and R. S. Sutton, "Policy iterations for reinforcement learning problems in continuous time and space—Fundamental theory and methods," *Automatica*, vol. 126, Apr. 2021, Art. no. 109421.

[10] W. Gao and Z.-P. Jiang, "Adaptive dynamic programming and adaptive optimal output regulation of linear systems," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 4164–4169, Dec. 2016.

[11] Y. Yang, W. Gao, H. Modares, and C.-Z. Xu, "Robust actor-critic learning for continuous-time nonlinear systems with unmodeled dynamics," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 6, pp. 2101–2112, Jun. 2022.

[12] C. Mu, Z. Ni, C. Sun, and H. He, "Air-breathing hypersonic vehicle tracking control based on adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 584–598, Mar. 2017.

[13] Q. Wei, D. Liu, F. L. Lewis, Y. Liu, and J. Zhang, "Mixed iterative adaptive dynamic programming for optimal battery energy control in smart residential microgrids," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 4110–4120, May 2017.

[14] B. Luo, H. N. Wu, and T. Huang, "Optimal output regulation for model-free quanser helicopter with multistep Q-learning," *IEEE Trans. Ind. Electron.*, vol. 65, no. 6, pp. 4953–4961, Jun. 2018.

[15] D. L. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Trans. Autom. Control*, vol. AC-13, no. 1, pp. 114–115, Feb. 1968.

[16] H. Dong, X. Zhao, and B. Luo, "Optimal tracking control for uncertain nonlinear systems with prescribed performance via critic-only ADP," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 1, pp. 561–573, Jan. 2022.

[17] D. Liu, S. Xue, B. Zhao, B. Luo, and Q. Wei, "Adaptive dynamic programming for control: A survey and recent advances," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 1, pp. 142–160, Jan. 2021.

[18] D. Liu, H. Li, and D. Wang, "Error bounds of adaptive dynamic programming algorithms for solving undiscounted optimal control problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1323–1334, Jun. 2015.

[19] D. Liu, Q. Wei, and P. Yan, "Generalized policy iteration adaptive dynamic programming for discrete-time nonlinear systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 12, pp. 1577–1591, Dec. 2015.

[20] C. Mu, D. Wang, and H. He, "Novel iterative neural dynamic programming for data-based approximate optimal control design," *Automatica*, vol. 81, pp. 240–252, Jul. 2017.

[21] K. G. Vamvoudakis and F. L. Lewis, "Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, May 2010.

[22] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, Oct. 2012.

[23] W. Gao and Z.-P. Jiang, "Learning-based adaptive optimal output regulation of linear and nonlinear systems: An overview," *Control Theory Technol.*, vol. 20, no. 1, pp. 1–19, Feb. 2022.

[24] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.

[25] F. Zhao, W. Gao, T. Liu, and Z.-P. Jiang, "Adaptive optimal output regulation of linear discrete-time systems based on event-triggered output-feedback," *Automatica*, vol. 137, Mar. 2022, Art. no. 110103.

[26] Y. Song, K. Zhao, and M. Krstic, "Adaptive control with exponential regulation in the absence of persistent excitation," *IEEE Trans. Autom. Control*, vol. 62, no. 5, pp. 2589–2596, May 2017.

[27] S. Adam, L. Busoniu, and R. Babuska, "Experience replay for real-time reinforcement learning control," *IEEE Trans. Syst., Man, Cybern., C (Appl. Rev.)*, vol. 42, no. 2, pp. 201–212, Mar. 2012.

[28] P. Wawrzynski, "Real-time reinforcement learning by sequential actor–critics and experience replay," *Neural Netw.*, vol. 22, no. 10, pp. 1484–1497, 2009.

[29] S. Kalyanakrishnan and P. Stone, "Batch reinforcement learning in a complex domain," in *Proc. 6th Int. Joint Conf. Auto. Agents Multiagent Syst. (AAMAS)*, 2007, pp. 1–8.

[30] B. Wang, D. Zhao, and J. Cheng, "Adaptive cruise control via adaptive dynamic programming with experience replay," *Soft Comput.*, vol. 23, no. 12, pp. 4131–4144, 2018.

[31] R. Yang, D. Wang, and J. Qiao, "Policy gradient adaptive critic design with dynamic prioritized experience replay for wastewater treatment process control," *IEEE Trans. Ind. Informat.*, vol. 18, no. 5, pp. 3150–3158, May 2022.

[32] H. Dong and X. Zhao, "Composite experience replay-based deep reinforcement learning with application in wind farm control," *IEEE Trans. Control Syst. Technol.*, vol. 30, no. 3, pp. 1281–1295, May 2022.

[33] G. Chowdhary and E. Johnson, "Concurrent learning for convergence in adaptive control without persistency of excitation," in *Proc. 49th IEEE Conf. Decis. Control (CDC)*, Dec. 2010, pp. 3674–3679.

[34] G. Chowdhary, T. Yucelen, M. Mühlegg, and E. N. Johnson, "Concurrent learning adaptive control of linear systems with exponentially convergent bounds," *Int. J. Adapt. Control Signal Process.*, vol. 27, no. 4, pp. 280–301, Apr. 2013.

[35] Y. Pan and H. Yu, "Composite learning robot control with guaranteed parameter convergence," *Automatica*, vol. 89, pp. 398–406, Mar. 2018.

[36] Y. Pan and H. Yu, "Composite learning from adaptive dynamic surface control," *IEEE Trans. Autom. Control*, vol. 61, no. 9, pp. 2603–2609, Sep. 2016.

[37] S. K. Jha, S. B. Roy, and S. Bhasin, "Initial excitation-based iterative algorithm for approximate optimal control of completely unknown LTI systems," *IEEE Trans. Autom. Control*, vol. 64, no. 12, pp. 5230–5237, Dec. 2019.

[38] Y. Yang, D. Wunsch, and Y. Yin, "Hamiltonian-driven adaptive dynamic programming for continuous nonlinear dynamical systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1929–1940, Aug. 2017.

[39] Y. Yang, K. G. Vamvoudakis, H. Modares, Y. Yin, and D. C. Wunsch, "Hamiltonian-driven hybrid adaptive dynamic programming," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 10, pp. 6423–6434, Oct. 2021.

[40] Y. Yang, H. Modares, K. G. Vamvoudakis, W. He, C.-Z. Xu, and D. C. Wunsch, "Hamiltonian-driven adaptive dynamic programming with approximation errors," *IEEE Trans. Cybern.*, early access, Sep. 9, 2021, doi: 10.1109/TCYB.2021.3108034.

[41] Y. Yang, M. Mazouchi, and H. Modares, "Hamiltonian-driven adaptive dynamic programming for mixed $\mathrm{H}_2/\mathrm{H}_\infty$ performance using sum-of-squares," *Int. J. Robust Nonlinear Control*, vol. 31, no. 6, pp. 1941–1963, 2021.

[42] M. Mazouchi, Y. Yang, and H. Modares, "Data-driven dynamic multi-objective optimal control: An aspiration-satisfying reinforcement learning approach," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 22, 2021, doi: 10.1109/TNNLS.2021.3072571.

[43] G. Tao, *Adaptive Control Design and Analysis*, vol. 37. Hoboken, NJ, USA: Wiley, 2003.

[44] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*. Hoboken, NJ, USA: Wiley, 2012.

[45] P. A. Ioannou and J. Sun, *Robust Adaptive Control*. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.

[46] F. L. Lewis, S. Jagannathan, and A. Yesildirak, *Neural Network Control of Robot Manipulators and Nonlinear Systems*. Boca Raton, FL, USA: CRC Press, 1999.

[47] Q. Zhang, D. Zhao, and Y. Zhu, "Event-triggered H_∞ control for continuous-time nonlinear system via concurrent learning," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 7, pp. 1071–1081, Jul. 2017.

[48] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, 2014.

[49] V. Nevistić and J. A. Primbs, "Constrained nonlinear optimal control: A converse HJB approach," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CIT-CDS 96-021, 1996.

Yongliang Yang (Member, IEEE) received the B.S. degree in electrical engineering from Hebei University, Baoding, China, in 2011, and the Ph.D. degree in control theory and control engineering from the University of Science and Technology Beijing (USTB), Beijing, China, in 2018.

From 2015 to 2017, he was a Visiting Scholar with the Missouri University of Science and Technology, Rolla, MO, USA, sponsored by the China Scholarship Council. He was an Assistant Professor with USTB from 2018 to 2020. From 2020 to 2021, he was an independent Post-Doctoral Research Fellow with the State Key Laboratory of Internet of Things for Smart City, Faculty of Science and Technology, University of Macau, Macau. He is currently an Associate Professor with USTB. His research interests include reinforcement learning theory, robotics, distributed optimization, and control for cyber-physical systems.

Dr. Yang was a recipient of the Best Ph.D. Dissertation of the China Association of Artificial Intelligence, the Best Ph.D. Dissertation of USTB, the Chancellor's Scholarship in USTB, the Excellent Graduates Awards in Beijing, and the UM Macao Talent Program in Macau. He is an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Yongping Pan (Senior Member, IEEE) received the Ph.D. degree in control theory and control engineering from the South China University of Technology, Guangzhou, China, in 2011.

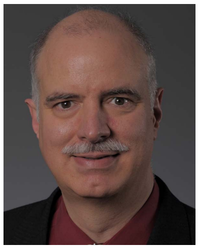
He spent one year in the industry as a Control Systems Engineer in Shenzhen and Guangzhou, China, and conducted research at the Nanyang Technological University, Singapore; the National University of Singapore, Singapore; and The University of Tokyo, Tokyo, Japan, for about eight years. He is currently a Professor with Sun Yat-sen University, Guangzhou. He has authored or coauthored over 130 peer-reviewed academic articles, including more than 100 refereed journal articles. His research interests include automatic control and machine learning with applications to robotics.

Dr. Pan is the Founding Chair of the IEEE Robotics and Automation Society Guangzhou Chapter. He has served as an associate editor for several top-tier journals and has been an organizing committee member for two international conferences.

Cheng-Zhong Xu (Fellow IEEE) received the B.Sc. and M.Sc. degrees from Nanjing University, Nanjing, China, in 1986 and 1989, respectively, and the Ph.D. degree from The University of Hong Kong, Hong Kong, in 1993, all in computer science and engineering.

He is currently the Dean of the Faculty of Science and Technology and the Interim Director of the Institute of Collaborative Innovation, University of Macau (UM), Macau, and a Chair Professor of computer and information science. He was a Professor with Wayne State University, Detroit, MI, USA, and the Director of the Institute of Advanced Computing, Shenzhen Institutes of Advanced Technologies, Chinese Academy of Sciences, Shenzhen, China, before he joined UM in 2019. He is also a Chief Scientist of Key Project on Smart City of MOST, China, and a Principal Instigator of the Key Project on Autonomous Driving of FDCT, Macau. His main research interests lie in parallel and distributed computing and cloud computing.

Dr. Xu was a Best Paper Nominee or Awardee of the 2013 IEEE High Performance Computer Architecture, the 2013 ACM High Performance Distributed Computing, IEEE Cluster 2015, ICPP 2015, GPC 2018, UIC 2018, and AIMS 2019. He was a recipient of the most prestigious "President's Awards for Excellence in Teaching" of Wayne State University in 2002. He was the Chair of the IEEE Technical Committee on Distributed Processing from 2015 to 2020. He serves or served on a number of journal editorial boards, including IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON CLOUD COMPUTING, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, and *Science China: Information Science*.



Donald C. Wunsch, II (Fellow IEEE) received the B.S. degree in applied mathematics from The University of New Mexico, Albuquerque, NM, USA, in 1984, the M.S. degree in applied mathematics and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, WA, USA, in 1987 and 1991, respectively, and the M.B.A. degree from Washington University, St. Louis, MO, USA, in 2006.

He completed the Jesuit Honors Program at Seattle University, Seattle. He is currently the Mary K. Finley Missouri Distinguished Professor with the Missouri University of Science and Technology (Missouri S&T), Rolla, MO, USA, and the Director of the Applied Computational Intelligence Laboratory. His earlier employers were: Texas Tech University, Lubbock, TX, USA; Boeing, Seattle; Rockwell International, Albuquerque; and International Laser Systems, Albuquerque. He has produced 23 Ph.D. recipients in computer engineering, electrical engineering, systems engineering, and computer science. His research interests include real-time learning, unsupervised learning, and reinforcement learning.

Dr. Wunsch, II, is a fellow and the President of the International Neural Networks Society (INNS). He received the NSF CAREER, the INNS Gabor, and the INNS Ada Lovelace. He served as the General Chair for the International Joint Conference on Neural Networks (IJCNN) and served on the St. Patrick's School Board, the IEEE Neural Networks Council, the INNS Board, and the University of Missouri Bioinformatics Consortium. He has chaired the Missouri S&T Information Technology and Computing Committee and the Student Design and Experiential Learning Center Board. He is also serving as the Program Director for the National Science Foundation.