IEEE Transactions on Automation Science and Engineering (T-ASE) paper, presented at ICRA 2024, Yokohama, Japan.
Cite as T-ASE paper.

Robust Policy Iteration of Uncertain Interconnected Systems with Imperfect Data

Omar Qasem[1], *Member, IEEE* and Weinan Gao[2], *Senior Member, IEEE*

***Abstract*— This paper investigates the robust optimal control problem of a class of continuous-time, partially linear, interconnected systems. In addition to the dynamic uncertainties resulted from the interconnected dynamic system, unknown bounded disturbances and computational errors are taken into account throughout the learning process, wherein the system's dynamics are also assumed unknown. These challenges lead the collected online data to be imperfect. In this scenario, traditional data-driven control techniques, such as adaptive dynamic programming (ADP) and robust ADP, encounter a challenge in learning the optimal control policy precisely due to imperfect data. In this paper, a novel data-driven robust policy iteration method is proposed to solve the robust optimal control problems. Without relying on the knowledge of the system's dynamics, the external disturbances or the complete state, the implementation of the proposed method only needs to access the input and partial state information. Based on the small-gain theorem, the notions of strong unboundedness observability and input-to-output stability, it is guaranteed that the learned robust optimal control gain is stabilizing and that the solution of the closed-loop system is uniformly ultimately bounded despite the existence of dynamic uncertainties and unknown external disturbances. The simulation results reveal the efficiency and practicality of the proposed data-driven control method.**

I. Introduction

In the last bidecade, learning-based control theory has attracted the attention of researchers in control systems [1]–[8]. Since Bellman has first proposed dynamic programming (DP) in the 1950s, DP has quickly evolved and been considered the backbone for solving optimal control problems [9]. Lately, reinforcement learning (RL) has been widely applied in learning-based control methods. Adaptive dynamic programming (ADP) has been considered as an effective RL-based strategy to develop an approximated optimal controller for continuous-time and discrete-time systems. ADP includes policy iteration (PI) and value iteration (VI) [10]–[13]. Moreover, generalized PI [14], [15] and hybrid iteration [16]–[19] techniques have been presented to improve the learning efficiency of existing ADP methods.

However, PI-based algorithms are superior to other ADP methods due to their quadratic convergence rate, since they are based on Newton-Raphson method. Due to its effectiveness and fast convergence rate, PI has been widely deployed in various fields; see [20]–[28] and references therein.

If the exact knowledge of the system's dynamics is available and the computations are precise and accurate, one can guarantee the convergence to the optimal solution. In addition, the learned policy will typically be accurate and exact to the policy obtained by solving the optimization problem analytically. However, the mathematical model is usually simplified in most physical applications, and the mismatch between the actual model and the simplified model is treated as dynamic uncertainties. From the literature of nonlinear control [29], it is known that the presence of dynamic uncertainty makes the feedback control problem extremely challenging in the context of nonlinear systems. In order to overcome this challenge, robust adaptive dynamic programming (RADP) has been developed by the integration of ADP, input-to-state stability theory [30], and nonlinear small-gain techniques [31] to address the presence of dynamic uncertainties in both linear and nonlinear systems. Additionally, a robust optimal control method for nonlinear systems with unknown disturbances has been presented in [32] such that the external disturbances are estimated using nonlinear disturbance observer. Moreover, a data-driven off-policy RL method, based on the Hamilton-Jacobi-Isaac (HJI) equation, is presented in [33] for nonlinear interconnected systems, with uncertain affine input systems under event-triggering. However, the exact measurements of the dynamic uncertainties are required during the learning phase of existing RADP and HJI techniques to learn a robust optimal control policy, which restricts its practicality. Moreover, the traditional techniques do not consider errors encountered by the computations using imperfect data and approximations during the learning phase.

In order to further improve the robustness of ADP methods, robust DP has been studied [26], [34], [35], where the robustness of the learning method due to a small additive external and unknown disturbances is considered, in which such disturbances are induced due to additive input uncertainties [36], [37] or unmodeled effects in the physical systems [34], [38]. Moreover, the design of robust controller using the PI method has been investigated subject to multiplicative and additive noises in [39]–[41], and external disturbances have been considered in [42]–[45]. Nevertheless, the dynamic uncertainties are usually not considered in robust DP methods.

It is worth mentioning that the robustness of DP, ADP and RL techniques to errors induced by the learning processes, studied in [26], [35], are different from the learning techniques that consider developing a controller that is robust to external disturbances and dynamic uncertainties [46]–[50].

Main Contributions: The contributions of this work are summarized as follows: 1) A novel data-driven ADP method based on robust PI is proposed for a class of continuous-time,

[1]O. Qasem is with the Electrical and Computer Engineering Department, School of Engineering and Computing, American International University, Al-Jahra, Kuwait o.qasem@aiu.edu.kw

[2]W. Gao is with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, Lianoning, 110819, China gaown@mail.neu.edu.cn

partially linear, interconnected systems under the influence of external, unknown and bounded disturbances, wherein imperfect data, i.e., partial states and inputs are used to obtain an approximated optimal control policy. 2) The proposed data-driven method is different from the traditional RADP in the sense that both dynamic uncertainties and computational errors induced by the iterative method of PI are considered and investigated. 3) The conditions on the bound of the uncertainties are investigated such that the learned policy is close enough to the actual optimal policy when the external disturbance is small enough. 4) Conditions by means of input-to-output stability, strong unboundedness observability property and small-gain theorem are given, such that once the initial stabilizing control policy is within some neighborhood of the optimal control policy, and the unknown disturbance is bounded, the proposed method will ensure to converge to a small neighborhood of the optimal control policy.

Structure: The rest of this paper is organized as follows. The problem statement and formulation are given in Section II. In Section III, a data-driven implementation of robust PI is given for partially linear interconnected systems with dynamic uncertainties and external disturbances. The simulation results are presented in Section IV with technical discussion. Finally, the conclusion is drawn in Section V.

Notations: Throughout this paper, $\mathbb{R}_+$ and $\mathbb{Z}_+$ represent the sets of non-negative real numbers and non-negative integers, respectively. $|\cdot|$ denotes the Euclidean norm for vectors, or the induced norm for matrices. $|\cdot|_F$ is the Frobenius norm and $|\cdot|_\infty$ is the l^∞-norm for non-negative integers, and the L^∞ norm for non-negative real numbers. For any piecewise continuous function $\mathcal{Q}:\mathbb{R}_+\to\mathbb{R}^q$, $\|\mathcal{Q}\|$ stands for $\sup_{t\geq 0}|\mathcal{Q}(t)|$. The Kronecker product operation is represented by $\otimes$. For any matrices $X, Z\in\mathbb{R}^{m\times n}$ and $U\in\mathbb{R}^{n\times(n+m)}$ we define $\mathcal{H}(U,Z)=\begin{bmatrix}I_n & -Z^{\mathrm{T}}\end{bmatrix}U\begin{bmatrix}I_n\\ -Z\end{bmatrix}$. $\mathrm{vec}(Z)=[z_1^{\mathrm{T}},z_2^{\mathrm{T}},\ldots,z_{n-1}^{\mathrm{T}},z_n^{\mathrm{T}}]^{\mathrm{T}}$, where $z_1,\ldots,z_n$ are the columns of Z. Given $Z^{\mathrm{T}}Z$ is invertible, $Z^\dagger=(Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}$ indicates the pseudo-inverse of Z. Given a symmetric matrix $P\in\mathbb{R}^{n\times n}$ and an arbitrary vector $v\in\mathbb{R}^n$, $\mathrm{vecs}(P)=[p_{11},2p_{12},\ldots,2p_{1n},p_{22},2p_{23},\ldots,2p_{n-1,n},p_{n,n}]^{\mathrm{T}}$ and $\mathrm{vecv}(v)=\mathrm{vecs}(vv^{\mathrm{T}})$. $P\succ(\succeq)0$ and $P\prec(\preceq)0$ indicate P is positive definite (semidefinite) and negative definite (semidefinite), respectively. The notations $\lambda_m(A)$ and $\lambda_M(A)$ denote the minimum and the maximum eigenvalues of the matrix $A\in\mathbb{R}^{n\times n}$, respectively. $\mathrm{I_d}$ denotes the identity function, and I_n denotes the identity matrix of size n. $\circ$ denotes the function composition operator.

II. Problem Statement and Preliminaries

In this paper, a class of continuous-time, partially linear, interconnected system is considered as follows.

$$\dot{x}=Ax+B(u+\Delta_0(z,y)+\Delta), \tag{1}$$

$$\dot{z}=g(z,y),\ y=Cx, \tag{2}$$

where $x\in\mathbb{R}^n$ is the system state vector with its initial condition $x_0\triangleq x(0)$, $u\in\mathbb{R}^m$ is the control input, $y\in\mathbb{R}^p$ is the output of the system, and $A\in\mathbb{R}^{n\times n}$, $B\in\mathbb{R}^{n\times m}$ and $C\in\mathbb{R}^{p\times n}$ are real matrices with A and B unknown. The signal Δ_z is defined by $\Delta_z\triangleq\Delta+\Delta_0(z,y)$. $\Delta_0(\cdot,\cdot):\mathbb{R}^q\times\mathbb{R}^p\to\mathbb{R}^m$ and $g(\cdot,\cdot):\mathbb{R}^q\times\mathbb{R}^p\to\mathbb{R}^q$ are two unknown locally Lipschitz functions satisfying $g(0,0)=0$ and $\Delta_0(0,0)=0$.

The main goal is to develop a robust data-driven PI algorithm to learn a robust optimal control policy for the system described by (1)-(2). By means of input-to-output-stability (IOS), strong unboundedness observability (SUO) property and small-gain theorem, we show that the solution of the interconnected system in closed-loop with the learned control policy is uniformly ultimately bounded (UUB). Throughout this paper, the following assumptions are considered.

Assumption 1: The dynamic uncertainty $\Delta_0(z,y)$ is only measurable during the learning phase, while the external disturbance Δ is unmeasurable anytime and $\|\Delta(t)\|\leq h$.

Assumption 2: The pairs (A,B) and (C,A) are controllable and observable, respectively.

Assumption 3: There exist a class $\mathcal{KL}$ function β_Δ, and a class $\mathcal{K}$ function γ_Δ, such that for any initial state $z(0)$, any measurable locally essentially bounded $y(\cdot)$ on $[0,\infty)$ and any time t in the right maximal interval of definition of the corresponding solution of (2), we have

$$|\Delta_0(t)|\leq\beta_\Delta(|z(0)|,t)+\gamma_\Delta(\|y\|). \tag{3}$$

Assumption 4: There exist a class $\mathcal{KL}$ function β^0, and a class $\mathcal{K}$ function γ^0, such that for any measurable $y(\cdot)$ defined on $[0,T)$ with $(0<T\leq\infty)$, the solution $z(t)$ of (2) right maximally defined on $[0,T')$ $(0<T'\leq T)$ satisfies

$$|z(t)|\leq\beta^0\left(|z(0)|,t\right)+\gamma^0\left(\left\|\left[y_t^{\mathrm{T}},z_t^{\mathrm{T}}\right]^{\mathrm{T}}\right\|\right),\forall t\in[0,T'), \tag{4}$$

where y_t and z_t are the truncated functions of y and z over $[0,t]$, respectively.

In traditional optimal control theory, the main goal is to develop an optimal controller for an undisturbed system described by (1) with $\Delta_z\equiv 0$ while minimizing a cost function described as follows.

$$J(x_0,u)=\int_0^\infty\left(x^{\mathrm{T}}Qx+u^{\mathrm{T}}Ru\right)\mathrm{d}t, \tag{5}$$
$$\text{s.t.}\quad (1),\ \Delta_z\equiv 0,$$

where the weight matrices are $Q\succeq 0$ and $R\succ 0$, and the pair $(A,\sqrt{Q})$ is observable. If the optimal state-feedback control policy was chosen to be

$$u^\star=-K^\star x, \tag{6}$$

then the optimal control gain matrix is computed by

$$K^\star=R^{-1}B^{\mathrm{T}}P^\star,$$

where $P^\star=(P^\star)^{\mathrm{T}}\succ 0$ is the solution to the well-known algebraic Riccati equation (ARE)

$$A^{\mathrm{T}}P^\star+P^\star A+Q-P^\star BR^{-1}B^{\mathrm{T}}P^\star=0. \tag{7}$$

One can notice that (7) is nonlinear in $P^\star$. An efficient model-based PI method was developed by Kleinman in [21] to solve $P^\star$ in (7), summarized by the following lemma.

Lemma 1 ([21]): Let $K_0 \in \mathbb{R}^{m\times n}$ be a stabilizing feedback gain matrix such that $A-BK_0$ is Hurwitz, and $P_k \succ 0$ be the solution of the Lyapunov equation

$$(A-BK_k)^{\mathrm{T}}P_k + P_k(A-BK_k) + Q + K_k^{\mathrm{T}}RK_k = 0, \quad (8)$$

$k \in \mathbb{Z}_+$, and the control gain matrix K_k, with $k = 1, 2, \ldots$, is defined recursively by:

$$K_k = R^{-1}B^{\mathrm{T}}P_{k-1}. \quad (9)$$

Then the following properties hold for any $k \in \mathbb{Z}_+$:

1) $A - BK_k$ is Hurwitz.
2) $P^\star \preceq P_{k+1} \preceq P_k$.
3) $\lim_{k\to\infty} K_k = K^\star$, $\lim_{k\to\infty} P_k = P^\star$.

Based on Lemma 1, given an initial stabilizing control gain matrix K_0, and by repeating (8)-(9), the pair (K_k, P_k) will converge to $(K^\star, P^\star)$ as $k \to \infty$.

It is notable that the method in Lemma 1 requires the exact knowledge of the system matrices (A, B). If the real physical systems' models are linearized, or approximated by the partial knowledge of system's dynamics, the inaccurate system matrices and the estimation errors induced by the iterative learning process and early termination of the PI method, will result in biased control policies after learning. Therefore, we study the conditions, such that one can ensure the convergence of the learned control policy to a small neighborhood of its corresponding actual optimal control policy despite the presence of computational and measurement errors, dynamic uncertainties and disturbances. Additionally, it is assumed that the system dynamics, i.e., (A, B) are unknown.

III. Data-Driven Robust Policy Iteration with Uncertainties and Unknown Disturbances

In this section, we will reach a small-gain condition in Theorem 1 such that the solution of the system (1)-(2) in closed-loop with the controller (6) remains UUB even with the existence of the unknown disturbance. Then, we will propose a data-driven robust PI algorithm, i.e., Algorithm 1, such that the collected data of the underlying dynamic system is corrupted due to the existence of external disturbances $\Delta(t)$, and the uncertainties generated by (2). The design of small-gain condition and the proposed algorithm rely on neither the knowledge of the system's dynamics A, B, g and Δ_0, nor the measurements of the external disturbance Δ.

Theorem 1: Given symmetric matrices $Q \succ \gamma_x I_n$ and $R \succ 0$, such that $\gamma_x \in \mathbb{R}_+$, if the following small-gain condition is satisfied

$$(\mathrm{I_d} + \rho_2) \circ \gamma_1 \circ (\mathrm{I_d} + \rho_1) \circ \gamma_\Delta(s) \leq s, \ \forall s \geq 0 \quad (10)$$

where

$$\gamma_1 = |C|\sqrt{\frac{\lambda_M(P^\star)\lambda_M(R)}{\lambda_m(P^\star)\gamma_x}} \quad (11)$$

with ρ_1, ρ_2 are of class $\mathcal{K}$, then the system (1)-(2) in closed-loop with the control policy (6) is UUB.

Proof: To begin with, we consider a Lyapunov candidate $V = x^{\mathrm{T}}P^\star x$. By taking the time derivative of V along the trajectories of (1), the following is obtained.

$$\begin{aligned}
\dot{V} &= \dot{x}^{\mathrm{T}}P^\star x + x^{\mathrm{T}}P^\star \dot{x} \\
&= x^{\mathrm{T}}\left((A-BK^\star)^{\mathrm{T}}P^\star + P^\star(A-BK^\star)\right)x + 2\Delta_z^{\mathrm{T}}B^{\mathrm{T}}P^\star x \\
&= -x^{\mathrm{T}}\left(Q + P^\star BR^{-1}B^{\mathrm{T}}P^\star\right)x + 2\Delta_z^{\mathrm{T}}B^{\mathrm{T}}P^\star x \\
&\leq -x^{\mathrm{T}}Qx - \left|x^{\mathrm{T}}P^\star BR^{-1/2} - \Delta_z^{\mathrm{T}}R^{1/2}\right|^2 + \Delta_z^{\mathrm{T}}R\Delta_z \\
&\leq -x^{\mathrm{T}}Qx + \Delta_z^{\mathrm{T}}R\Delta_z \quad (12) \\
&\leq -\gamma_x|x|^2 + \lambda_M(R)|\Delta_z|^2. \quad (13)
\end{aligned}$$

Due to the fact that

$$\lambda_m(P^\star)|x|^2 \leq V \leq \lambda_M(P^\star)|x|^2, \quad (14)$$

the following inequality is obtained

$$V \leq \exp\left(-\frac{\gamma_x}{\lambda_M(P^\star)}t\right)V(0) + \frac{\lambda_M(P^\star)\lambda_M(R)}{\gamma_x}\|\Delta_z\|^2. \quad (15)$$

For any initial condition $x_0 \in \mathbb{R}^n$, we have

$$V(0) \leq \lambda_M(P^\star)|x_0|^2 \quad (16)$$

and the following is obtained.

$$\begin{aligned}
V \leq \exp\left(-\frac{\gamma_x}{\lambda_M(P^\star)}t\right)&\lambda_M(P^\star)|x_0|^2 \\
&+ \frac{\lambda_M(P^\star)\lambda_M(R)}{\gamma_x}\|\Delta_z\|^2. \quad (17)
\end{aligned}$$

Using the left side of the inequality (14), comparing it with (17) and taking the square root of both sides, we conclude that

$$|x(t)| \leq \beta_0(|x_0|, t) + \gamma_0(\|\Delta_z\|), \quad (18)$$

with

$$\beta_0(|x_0|, t) = \exp\left(-\frac{\gamma_x}{2\lambda_M(P^\star)}t\right)\sqrt{\frac{\lambda_M(P^\star)}{\lambda_m(P^\star)}}|x_0|, \quad (19)$$

$$\gamma_0(\|\Delta_z\|) = \sqrt{\frac{\lambda_M(P^\star)\lambda_M(R)}{\lambda_m(P^\star)\gamma_x}}\|\Delta_z\|. \quad (20)$$

Since $\beta_0(|x_0|, t)$ is of class $\mathcal{KL}$ and $\gamma_0(\|\Delta_z\|)$ is of class $\mathcal{K}$, the $x-$system described by (1) is input-to-state stable with Δ_z as an input [30]. Furthermore, the following is obtained from (18).

$$|y(t)| \leq \beta_1(|x_0|, t) + \gamma_1(\|\Delta_z\|), \quad (21)$$

where

$$\beta_1(|x_0|, t) = |C|\beta_0(|x_0|, t), \quad (22)$$

$$\gamma_1(\|\Delta_z\|) = |C|\gamma_0(\|\Delta_z\|). \quad (23)$$

Therefore, as depicted in Fig. 1, the x–system has the SUO property with zero-offset and the IOS property [31] with y as its output and $\Delta_z := \Delta_0 + \Delta$ as its input. On the other hand, Assumptions 3-4 imply that the $z-$system described in (2) has the SUO and IOS properties with y as

its input and Δ_0 as its output. Under the small-gain condition (10), one can conclude that the composed system (1)-(2) with the $x-$system in closed-loop with (6) has the ISS property regarding Δ_z as its input; see [31]. Since Δ is bounded, we can always find positive constants b and c, and for every $a \in (0, c)$ there exists a $\mathcal{E} = \mathcal{E}(a, b) > 0$ such that

$$|[x^{\mathrm{T}}(0), z^{\mathrm{T}}(0)]^{\mathrm{T}}| \leq a \rightarrow |[x^{\mathrm{T}}(t), z^{\mathrm{T}}(t)]^{\mathrm{T}}| \leq b, \quad \forall t \geq \mathcal{E},$$

which immediately implies that the solution of the overall system (1)-(2) with (6) is UUB [51, Definition 4.6]. The proof is thus completed. ■

Without loss of generality, we assume $Q \succ I_n$ and $R \succ I_m$ for simplicity. One can always find a $\bar{\boldsymbol{\alpha}} > 1$ such that $\bar{\boldsymbol{\alpha}} Q \succ I_n$ and $\bar{\boldsymbol{\alpha}} R \succ I_m$. $K^\star$ remains the same no matter if we use the scaled weight matrices or the original ones.

From Theorem 1, we see that (6) is a robust optimal control policy if the small-gain condition (10) holds. However, to approximate this robust optimal control policy with imperfect online data and unknown disturbance is by no means easy. To deal with this conundrum, we consider the iterative scheme and analyze its stability in addition to the stability region. Our goal is to develop a robust PI method, that learns the optimal control policy that is robust with the existence of the disturbance $\Delta(t)$ and the uncertainty $\Delta_0(z, y)$.

Given any stabilizing control gain matrix K_k, the system (1) can be written as

$$\dot{x} = A_k x + B(u + K_k x) + B(\Delta_0 + \Delta), \tag{24}$$

where $A_k = A - BK_k$. Let $Q_k = Q + K_k^{\mathrm{T}} R K_k$ and $P_k \succ 0$ solve a Lyapunov equation $A_k^{\mathrm{T}} P_k + P_k A_k = -Q_k$. Taking the time derivative of a Lyapunov function $V_k = x^{\mathrm{T}} P_k x$ along the trajectories of (24) we have

$$\begin{aligned}\dot{V}_k &= \dot{x}^{\mathrm{T}} P_k x + x^{\mathrm{T}} P_k \dot{x} \\ &= x^{\mathrm{T}} \left(A_k^{\mathrm{T}} P_k + P_k A_k\right) x + 2\left(u + K_k x\right)^{\mathrm{T}} B^{\mathrm{T}} P_k x \\ &\quad + 2\Delta_0^{\mathrm{T}} B^{\mathrm{T}} P_k x + 2\Delta^{\mathrm{T}} B^{\mathrm{T}} P_k x.\end{aligned}$$

By letting $L_k = B^{\mathrm{T}} P_k$ we get

$$\begin{aligned}\dot{V}_k &= 2\left(u^{\mathrm{T}} L_k x + x^{\mathrm{T}} K_k^{\mathrm{T}} L_k x + \Delta_0^{\mathrm{T}} L_k x + \Delta^{\mathrm{T}} L_k x\right) - x^{\mathrm{T}} Q_k x \\ &= -\left(x^{\mathrm{T}} \otimes x^{\mathrm{T}}\right) \mathrm{vec}(Q_k) + 2\left(x^{\mathrm{T}} \otimes u^{\mathrm{T}}\right) \mathrm{vec}(L_k) \\ &\quad + 2\left(x^{\mathrm{T}} \otimes x^{\mathrm{T}}\right)\left(I_n \otimes K_k^{\mathrm{T}}\right) \mathrm{vec}(L_k) + 2\left(x^{\mathrm{T}} \otimes \Delta_0^{\mathrm{T}}\right) \mathrm{vec}(L_k) \\ &\quad + 2\left(x^{\mathrm{T}} \otimes \Delta^{\mathrm{T}}\right) \mathrm{vec}(L_k).\end{aligned} \tag{25}$$

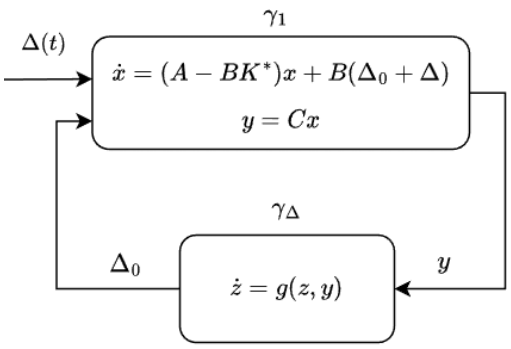

Fig. 1. Illustrative diagram of the interconnected dynamic system described by (1)-(2) in closed-loop with the optimal control policy $u^\star = -K^\star x$

By integrating (25) over the time interval $[t_0, t_s]$, where the sequence $\{t_l\}_{l=0}^{s}$ is an increasing sequence with $t_l = t_{l-1} + \Delta t$ and $\Delta t > 0$, we obtain

$$\Theta(K_k) \begin{bmatrix} \mathrm{vecs}\left(P_k\right) \\ \mathrm{vec}(L_k) \end{bmatrix} = -\mathcal{I}_{xx} \mathrm{vec}(Q_k) + w \tag{26}$$

such that

$$\Theta(K_k) = \left[\delta_{xx}, -2\mathcal{I}_{xu} - 2\mathcal{I}_{xx}\left(I_n \otimes K_k^{\mathrm{T}}\right) - 2\mathcal{I}_{x\Delta_0}\right] \tag{27}$$

where $w \triangleq 2\mathcal{I}_{x\Delta} \mathrm{vec}(L_k)$ is unmeasurable. The following data matrices are defined for any vectors θ and ϕ.

$$\delta_{\theta\theta} = \left[\mathrm{vecv}(\theta)|_{t_0}^{t_1}, \quad \mathrm{vecv}(\theta)|_{t_1}^{t_2}, \quad \ldots \quad, \mathrm{vecv}(\theta)|_{t_{s-1}}^{t_s}\right]^{\mathrm{T}},$$

$$\mathcal{I}_{\theta\phi} = \left[\int_{t_0}^{t_1} (\theta \otimes \phi)\, \mathrm{d}\tau, \int_{t_1}^{t_2} (\theta \otimes \phi)\, \mathrm{d}\tau, \ldots, \int_{t_{s-1}}^{t_s} (\theta \otimes \phi)\, \mathrm{d}\tau\right]^{\mathrm{T}},$$

where the sequence $\{t_l\}_{l=0}^{s}$ is an increasing sequence with $t_l = t_{l-1} + \Delta t$ and $\Delta t > 0$. Since the unknown disturbance and additive noise applied to the system are represented by Δ, and by Assumption 1 we know that $\|\Delta\|$ is bounded, it is concluded that w is also bounded during the learning phase. However, since w is unmeasurable, we can only approximate the solution of (26) by solving the following equation.

$$\hat{\mathcal{Y}}_k \triangleq \begin{bmatrix} \mathrm{vecs}(\bar{P}_k) \\ \mathrm{vec}(\hat{L}_k) \end{bmatrix} = \Theta(\hat{K}_k)^{\dagger} \Xi(\hat{K}_k), \tag{28}$$

where $\Xi(\hat{K}_k) = -\mathcal{I}_{xx} \mathrm{vec}(Q + \hat{K}_k^{\mathrm{T}} R \hat{K}_k)$. In this sense, by defining $\bar{L}_k = B^{\mathrm{T}} \hat{P}_k$, we have

$$G(\hat{P}_k) = \begin{bmatrix} \hat{K}_k^{\mathrm{T}} \bar{L}_k + \bar{L}_k \hat{K}_k - \hat{K}_k^{\mathrm{T}} R \hat{K}_k & \bar{L}_k^{\mathrm{T}} \\ \bar{L}_k & R \end{bmatrix}, \tag{29}$$

$$\hat{G}_k = \begin{bmatrix} \hat{K}_k^{\mathrm{T}} \hat{L}_k + \hat{L}_k \hat{K}_k - \hat{K}_k^{\mathrm{T}} R \hat{K}_k & \hat{L}_k^{\mathrm{T}} \\ \hat{L}_k & R \end{bmatrix}. \tag{30}$$

For any matrix $\mathcal{G}$ we define $\mathcal{G}(\cdot) \triangleq \begin{bmatrix} [\mathcal{G}(\cdot)]_{xx} & [\mathcal{G}(\cdot)]_{ux}^{\mathrm{T}} \\ [\mathcal{G}(\cdot)]_{ux} & [\mathcal{G}(\cdot)]_{uu} \end{bmatrix}$. With that, we are now ready to present the data-driven robust PI algorithm, given in Algorithm 1, with its proof of convergence presented in Theorem 2.

Algorithm 1 Data-Driven Robust PI Algorithm

1: Choose a stabilizing policy $\hat{K}_0 \in \mathbb{R}^{m \times n}$ and a small constant $\hat{\varepsilon} > 0$. $k \leftarrow 0$.
2: Set the maximum number of iterations as $N \in \mathbb{Z}_+$, and choose $\gamma_x > 1$ such that $Q \succ \gamma_x I_n$.
3: Apply an essentially bounded input $u = -\hat{K}_k x + e$, where e is the exploration noise, over $[t_0, t_s]$ and collect the data from the dynamic system (1)-(2) until (31) is satisfied.
4: Compute $\Theta(\hat{K}_k)$ from (27).
5: **repeat**
6: Compute $\hat{\mathcal{Y}}_k$ and $\hat{G}_k$ from (28) and (30), respectively.
7: Compute $\hat{K}_{k+1}$ from $\hat{K}_{k+1} = [\hat{G}_k]_{uu}^{-1}[\hat{G}_k]_{xu}$
8: $k \leftarrow k + 1$
9: **until** $k \geq N$ or $|\bar{P}_k - \bar{P}_{k-1}| \leq \hat{\varepsilon}$
10: Use the approximated control gain matrix $\hat{K}_k$ as a good approximation of the optimal control gain matrix $K^\star$.

Remark 1: To satisfy the persistent excitation, exploration noise is added to the control input during the data collection phase in order to satisfy the following rank condition:

$$\text{rank}\left([\mathcal{I}_{xx}, \mathcal{I}_{xu}, \mathcal{I}_{x\Delta_0}]\right) = n(n+1)/2 + 2nm. \quad (31)$$

Theorem 2: In Algorithm 1, given $Q \succ \max\{\gamma_x, 1\} I_n$, $M \in \mathbb{Z}_+$, $\Delta t \in \mathbb{R}_+$ and $t_0 \in \mathbb{R}_+$, starting from any initial stabilizing control gain $\hat{K}_0$, for any $\epsilon > 0$, there exist a positive constant h such that if $|w|_\infty < h$ and the rank condition (31) is satisfied, then we have the following:

1) $\hat{K}_k$ is stabilizing for all $k = 0, 1, 2, \ldots, N$.
2) $\limsup_{N\to\infty} |\hat{P}_N - P^\star| < \epsilon$, where $\hat{P}_N$ satisfies $\mathcal{H}(G(\hat{P}_N), \hat{K}_N) = 0$.
3) $\limsup_{N\to\infty} |\hat{K}_N - K^\star| < \epsilon$.

Proof: Since $\hat{K}_0$ is stabilizing, then there exists a_k defined as

$$a_k = (m(\sqrt{n} + |\hat{K}_k|)^2 + m(\sqrt{n} + |\hat{K}_{k+1}|)^2)^{-1}$$

such that as long as $|\Delta G_k|_F < a_k$, where $\Delta G_k = \hat{G}_k - G(\hat{P}_k)$, we have each element in $\{\hat{K}_k\}_{k=0}^{\bar{k}}$ is stabilizing, each element in $\{[\hat{G}_k]_{uu}\}_{k=0}^{\bar{k}}$ is nonsingular, and $\{\hat{P}_k\}_{k=0}^{\bar{k}}$ is bounded [35, Lemma 7]. Furthermore,

$$|\hat{K}_{k+1}|_F \le 2|R^{-1}|_F(1 + |B^{\mathrm{T}}\hat{P}_k|_F). \quad (32)$$

We also have

$$|[G(\hat{P}_k)]_{uu}^{-1}([\hat{G}_k]_{uu} - [G(\hat{P}_k)]_{uu})|_F < a_k\sqrt{m} < 0.5. \quad (33)$$

By [52, Lemma 5.8], it is concluded that $[\hat{G}_k]_{uu}$ is invertible. In addition, for any $x \in \mathbb{R}^n$ on the unit ball we have

$$x^{\mathrm{T}}\mathcal{H}(G(\hat{P}_k), \hat{K}_k)x = 0, \quad (34)$$

in which it can be shown that for any $Q \succ I_n$ we obtain

$$x^{\mathrm{T}}((A - B\hat{K}_{k+1})^{\mathrm{T}}\hat{P}_k + \hat{P}_k(A - B\hat{K}_{k+1}))x < 0 \quad (35)$$

for all x in the unit ball. Furthermore, for any $Q \succ \gamma_x I_n$ it has been shown that the overall system will remain UUB at the origin with Δ_z as the input.

Let $\Theta^\star(\cdot)$ and $\Xi^\star(\cdot)$ be the functions defined in (28) associated with $\delta_{xx}^\star$, $\mathcal{I}_{xx}^\star$, $\mathcal{I}_{xu}^\star$ and $\mathcal{I}_{x\Delta_0}^\star$. Then by assuming that (31) is satisfied and by [23, Lemma 6, Theorem 7]

$$\bar{\mathcal{Y}}_k \triangleq \begin{bmatrix}\text{vecs}(\hat{P}_k)\\ \text{vec}(\bar{L}_k)\end{bmatrix} = \Theta^\star\left(\hat{K}_k\right)^\dagger \Xi^\star\left(\hat{K}_k\right) \quad (36)$$

Define $\Delta L_k \triangleq \hat{L}_k - \bar{L}_k$. ΔG_k can be written as follows.

$$\Delta G_k = \begin{bmatrix}(I_n \otimes \hat{K}_k^{\mathrm{T}})\text{vec}(\Delta L_k) + (\hat{K}_k^{\mathrm{T}} \otimes I_n)\text{vec}(\Delta L_k^{\mathrm{T}}) & \Delta L_k^{\mathrm{T}}\\ \Delta L_k & 0\end{bmatrix} \quad (37)$$

By taking the Frobenius norm of (37), using the fact that $|\text{vec}(\Delta L_k)|_F = \left|\text{vec}\left(\Delta L_k^{\mathrm{T}}\right)\right|_F$ and by using matrices norm properties, the following is obtained.

$$\begin{aligned}|\Delta G_k|_F &= \sqrt{\begin{aligned}&2|\text{vec}(\Delta L_k)|_F^2 + |(I_n \otimes \hat{K}_k^{\mathrm{T}})\text{vec}(\Delta L_k)\\ &\qquad + (\hat{K}_k^{\mathrm{T}} \otimes I_n)\text{vec}(\Delta L_k^{\mathrm{T}})|_F^2\end{aligned}}\\ &\le \sqrt{(2 + |(I_n \otimes \hat{K}_k^{\mathrm{T}}) + (\hat{K}_k^{\mathrm{T}} \otimes I_n)|_F^2)|\text{vec}(\Delta L_k)|_F^2}\end{aligned} \quad (38)$$

By using (28), (36) and (38), we can conclude that

$$|\Delta G_k|_F \le \sqrt{2 + 4|\hat{K}_k|_F^2}|\hat{\mathcal{Y}}_k - \bar{\mathcal{Y}}_k|. \quad (39)$$

By [35, Theorem 2], if $|\Delta G_k|_F < \delta_2$, we have $|\hat{P}_k|_F < M_0$, $|\hat{K}_k|_F < M_1$ for some $M_1 > 0$. Let $\bar{M} = \max(M_0, M_1)$ and define a set

$$\begin{aligned}\mathcal{F} = \{K \in \mathbb{R}^{m\times n}, \gamma_1(\|\Delta_z\|) \in \mathcal{K} : |K|_F \le \bar{M}, |P_K|_F \le \bar{M},\\ (A - BK) \text{ is Hurwitz, and (10) is satisfied}\}\end{aligned}$$

where P_K is the solution to (8). To fulfill (39), it is sufficient to prove that for all $K \in \mathcal{F}$ and $|w|_\infty < h'$ we have

$$\begin{aligned}|\Theta(K)^\dagger\Xi(K) - \Theta^\star(K)^\dagger\Xi^\star(K)| < \alpha_1(\alpha_2|\Theta(K) - \Theta^\star(K)|_F\\ + |\Xi(K) - \Xi^\star(K)|_F)\end{aligned} \quad (40)$$

for any $K \in \mathcal{F}$. By linear system theory [53], $|[\mathcal{I}_{xx}, \mathcal{I}_{xu}, \mathcal{I}_{x\Delta_0}] - [\mathcal{I}_{xx}^\star, \mathcal{I}_{xu}^\star, \mathcal{I}_{x\Delta_0}^\star]|_F < \gamma_1(|w|_\infty)$ for γ_1 defined in (23). Thus there always exist a $h < h'$ such that (40) is satisfied. Which implies that if $|P_k - P_{k+1}| < \epsilon$, then $|P_k - P^\star| < \epsilon$. Therefore, items 1) and 2) are thus proved. Item 3) is proved by using the fact that

$$\begin{aligned}|\hat{K}_N - K^\star|_F \le |R^{-1}|_F|\Delta G_{N-1}|_F\\ + |R^{-1}B^{\mathrm{T}}|_F|\hat{P}_{N-1} - P^\star|_F.\end{aligned} \quad (41)$$

By (39) and since it is proved that $\limsup_{N\to\infty} |\hat{P}_N - P^\star| < \epsilon$, then with the given results, item 3) is thus proved. ■

In general, $G(\hat{P}_k)$ is different from $\hat{G}_k$ due to data corruption resulted from computational errors in learning, and noisy measurements from uncertainties . Table I illustrates the main differences between the proposed Algorithm 1 and existing RADP, robust PI and HJI methods.

Methods	This Work	RADP	Robust PI	HJI
Dynamic Uncertainties	Yes	Yes	No	Yes
Computational Errors	Yes	No	Yes	No
External Disturbances	Yes	No	No	Yes
Nonlinearities	Yes	Yes	No	Yes

TABLE I
COMPARISON OF METHODS

IV. ILLUSTRATIVE EXAMPLE

In this section, the efficiency and the practicality of the proposed method is demonstrated through a real-life practical application of a two-machine power system with governor controllers [54]. The system's dynamics are described by

$$\Delta\dot{\boldsymbol{\delta}}_i(t) = \Delta\omega_i(t), \quad (42)$$

$$\Delta\dot{\omega}_i(t) = -\frac{D_i}{2H_i}\Delta\omega_i(t) + \frac{\omega_0}{2H_i}\Delta P_{mi}(t), \quad (43)$$

$$\Delta\dot{P}_{mi}(t) = \frac{1}{T_i}\left(-\Delta P_{mi}(t) + u_i(t) - d_i(t)\right) \quad (44)$$

where for $i, j \in \{1, 2\}$, $\Delta\boldsymbol{\delta}_i(t) = \boldsymbol{\delta}_i(t) - \boldsymbol{\delta}_{i0}$, $\Delta\omega_i(t) = \omega_i(t) - \omega_{i0}$, $\Delta P_{mi}(t) = P_{mi}(t) - P_{ei}(t)$, $u_i(t) = u_{gi}(t) - P_{ei}(t)$, and $d_i(t) =$

Parameter	D_1	D_2	H_1	H_2	T_1	T_2	ω_0
Value	6.4	3	100	6.4	6	6.3	314.159
Unit	p.u.	p.u.	s	s	s	s	rad/s
Parameter	B_{12}	G_{12}	$\boldsymbol{\delta}_{10}$	$\boldsymbol{\delta}_{20}$	E'_{q1}	E'_{q2}	
Value	0.1875	0.05194	1.9	1.7	1.2	1.5	
Unit	$1/\Omega$	$1/\Omega$	rad	rad	p.u.	p.u.	

TABLE II

PARAMETERS VALUES OF THE TWO-MACHINE POWER SYSTEM WITH GOVERNOR CONTROLLERS

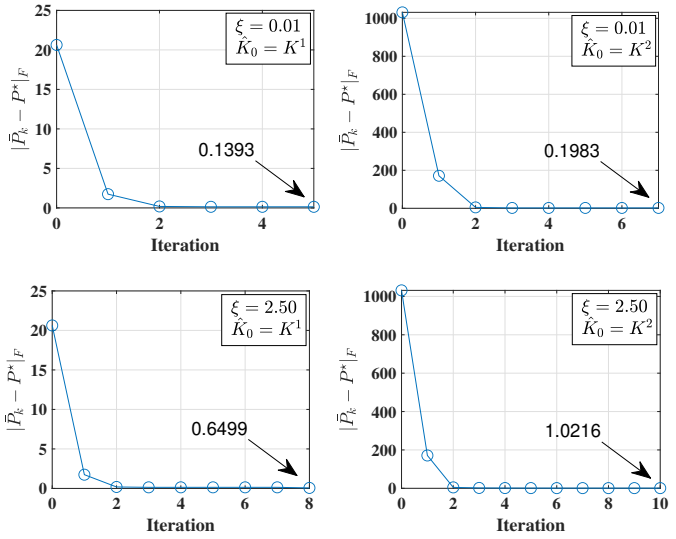

Fig. 2. $|\bar{P}_k - P^\star|_F$ for four different cases under Algorithm 1

$E'_{qi} \sum_{j=1, j\neq i}^{2} E'_{qj} [B_{ij} \cos \boldsymbol{\delta}_{ij} - G_{ij} \sin \boldsymbol{\delta}_{ij}] \times [\Delta\omega_i(t) - \Delta\omega_j(t)]$. $\boldsymbol{\delta}_i(t)$ is the angle of the ith generator; $\boldsymbol{\delta}_{ij} = \boldsymbol{\delta}_i(t) - \boldsymbol{\delta}_j(t)$ represents the angular difference between the ith and the jth generators; $\omega_i(t)$ is the relative rotor speed; $P_{mi}(t)$ and $P_{ei}(t)$ are the mechanical power and the electrical power, respectively; and u_{gi} is the speed governor control signal for the ith generator.

Therefore, the model (42)-(44) has the same structure of (1)-(2) by defining $x = [\boldsymbol{\delta}_1(t) \ \Delta\omega_1(t) \ \Delta P_{e1}(t)]^{\mathrm{T}}$, $z = [\boldsymbol{\delta}_2(t) \ \Delta\omega_2(t) \ \Delta P_{e2}(t)]^{\mathrm{T}}$, $\Delta_0(z, y) = E'_{q1} E'_{q2} [B_{ij} \cos \boldsymbol{\delta}_{ij} - G_{ij} \sin \boldsymbol{\delta}_{ij}] [\Delta\omega_i(t) - \Delta\omega_j(t)]$, and $y_1 = \Delta\omega_1(t)$.

In addition, we consider $\Delta(t)$ to be some additive bounded noise generated by sinusoidal signals applied to the first machine, which represents the x-system. The second machine, which represents the z-system, is considered to be the reference machine.

The values of the parameters used in the simulation are shown in Table II, with their descriptions given in [55]. The weight matrices are chosen to be $Q = 10^6 I_3$ and $R = 1$. Algorithm 1 is applied on the described system with applying exploration noise over the time period $[0, 2.5(s)]$. The state initial condition is set to $x_0 = \begin{bmatrix} 3.5 & 10 & 9 \end{bmatrix}^{\mathrm{T}}$ and that of the uncertainty z is set to $z(0) = \begin{bmatrix} -10 & 6 & -5 \end{bmatrix}^{\mathrm{T}}$. We consider two different initial control gain matrices $\hat{K}_0$, and two different scaling values ξ. The two different initial control gain matrices are considered such that $K^1 = [1010, \ 1508.1, \ 1023.1]$ is a control gain near the optimal one with $|P_{K^1} - P^\star|_F = 20.6293$ and $K^2 = [1500, \ 2239.8, \ 1519.5]$ is a control gain far from the optimal one with $|P_{K^2} - P^\star|_F = 1031.5$. The unknown disturbance $\Delta(t)$ is scaled by $\xi = 0.01$ and $\xi = 2.50$. For simulation purposes we assume $\Delta = \xi \sin(30t)$. The results of the experiments are depicted in Figs. 2-3.

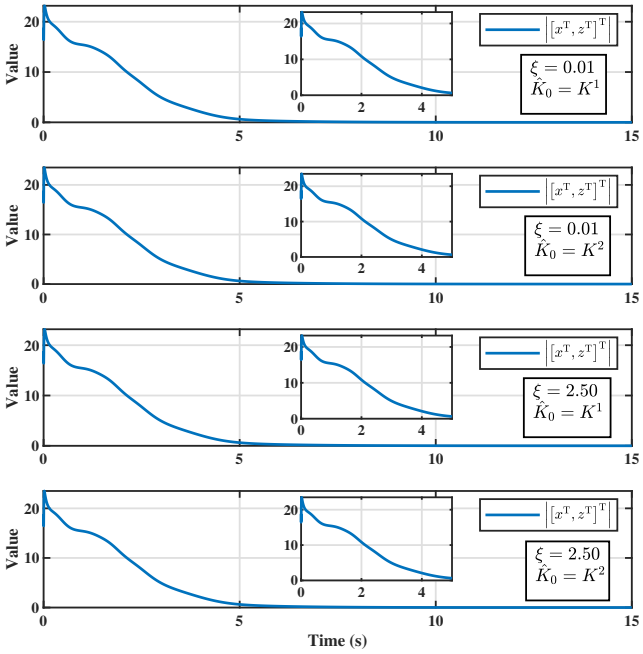

Fig. 3. States trajectories norm under Algorithm 1

It is noticed that even with the existence of the dynamic uncertainties and external disturbance, the algorithm still converges to a near optimal solution for all four cases, considering those cases with K^2 as an initial stabilizing control policy which is far from the optimal control policy. The efficiency of the proposed method is also revealed with increasing the bound of the unknown disturbance, and by starting with a control policy far from the optimal control policy. Notably, the learned policy will still converge as long as the initial value matrix P_0 is within some neighborhood of $P^\star$.

V. CONCLUSIONS

In this paper, a novel data-driven robust policy iteration (PI) method is proposed for solving the robust optimal control problem of continuous-time, partially linear, interconnected systems. The proposed method considers computational errors resulted from the iterative process of PI, the dynamic uncertainties and unknown bounded disturbances. We have reached a new condition on the bound of the external disturbances such that the learned policy is close to the actual robust optimal policy. Through the properties of input-to-output stability and strong unboundedness observability, and the small-gain theory, the solution of the closed-loop system is guaranteed to be uniformly ultimately bounded. Notably, the proposed method does not require the knowledge of the system dynamics or the disturbances, but the imperfect data, i.e., partial states and inputs. An illustrative example is applied, and its simulation results validate the efficacy of the proposed PI method.

REFERENCES

[1] M. Li, P. Zhou, H. Wang, and T. Chai, "Geometric analysis based double closed-loop iterative learning control of output PDF shaping of fiber length distribution in refining process," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 9, pp. 7229–7238, 2019.

[2] Z. P. Jiang, T. Bian, and W. Gao, "Learning-based control: A tutorial and some recent results," *Foundations and Trends in Systems and Control*, vol. 8, no. 3, pp. 176–284, 2020.

[3] K. G. Vamvoudakis and N.-M. T. Kokolakis, "Synchronous reinforcement learning-based control for cognitive autonomy," *Foundations and Trends in Systems and Control*, vol. 8, no. 1-2, pp. 1–175, 2020.

[4] D. Zhao, Z. Xia, and D. Wang, "Model-free optimal control for affine nonlinear systems with convergence analysis," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 4, pp. 1461–1468, 2014.

[5] H. Zhang, C. Qin, and Y. Luo, "Neural-network-based constrained optimal control scheme for discrete-time switched nonlinear system using dual heuristic programming," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 3, pp. 839–849, 2014.

[6] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Systems Magazine*, vol. 32, no. 6, pp. 76–105, 2012.

[7] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, no. 5, pp. 779 – 791, 2005.

[8] O. Qasem, K. Jebari, and W. Gao, "Adaptive dynamic programming and data-driven cooperative optimal output regulation with adaptive observers," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 2538–2543.

[9] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton University Press, 1957.

[10] A. Al-Tamimi, M. Abu-Khalaf, F. Lewis, A. Mellouk, and A. Chebira, "Heuristic dynamic programming nonlinear optimal controller," *Machine Learning*, pp. 361–380, 2009.

[11] Q. Wei, D. Liu, and H. Lin, "Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 840–853, 2015.

[12] T. Bian and Z. P. Jiang, "Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design," *Automatica*, vol. 71, pp. 348 – 360, 2016.

[13] ——, "Reinforcement learning and adaptive optimal control for continuous-time nonlinear systems: A value iteration approach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 7, pp. 2781–2790, 2022.

[14] D. Vrabie and F. L. Lewis, "Generalized policy iteration for continuous-time systems," in *2009 International Joint Conference on Neural Networks*. IEEE, 2009, pp. 3224–3231.

[15] J. Y. Lee, J. B. Park, and Y. H. Choi, "On integral generalized policy iteration for continuous-time linear quadratic regulations," *Automatica*, vol. 50, no. 2, pp. 475–489, 2014.

[16] O. Qasem, W. Gao, and K. G. Vamvoudakis, "Adaptive optimal control of continuous-time nonlinear affine systems via hybrid iteration," *Automatica*, vol. 157, p. 111261, 2023.

[17] O. Qasem, W. Gao, and T. Bian, "Adaptive optimal control of continuous-time linear systems via hybrid iteration," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2021, pp. 01–07.

[18] O. Qasem, M. Davari, W. Gao, D. R. Kirk, and T. Chai, "Hybrid iteration ADP algorithm to solve cooperative, optimal output regulation problem for continuous-time, linear, multi-agent systems: Theory and application in islanded modern microgrids with IBRs," *IEEE Transactions on Industrial Electronics*, pp. 1–12, 2023.

[19] W. Gao, C. Deng, Y. Jiang, and Z. P. Jiang, "Resilient reinforcement learning and robust output regulation under denial-of-service attacks," *Automatica*, vol. 142, p. 110366, 2022.

[20] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, Cambridge, 2018.

[21] D. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, no. 1, pp. 114–115, 1968.

[22] T. Hylla, "Extension of inexact Kleinman-Newton methods to a general monotonicity preserving convergence theory," Ph.D. Dissertation, Universtät Trier, Trier, 2011.

[23] 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, 2012.

[24] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE circuits and systems magazine*, vol. 9, no. 3, pp. 32–50, 2009.

[25] C. Chen, F. L. Lewis, and B. Li, "Homotopic policy iteration-based learning design for unknown linear continuous-time systems," *Automatica*, vol. 138, p. 110153, 2022.

[26] T. Bian and Z. P. Jiang, "Continuous-time robust dynamic programming," *SIAM Journal on Control and Optimization*, vol. 57, no. 6, pp. 4150–4174, 2019.

[27] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477 – 484, 2009.

[28] D. Wang, D. Liu, and H. Li, "Policy iteration algorithm for online design of robust control for a class of continuous-time nonlinear systems," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 627–632, 2014.

[29] M. Krstic, I. Kanellakopoulos, and P. V. Kokotovic, *Nonlinear and Adaptive Control Design*. New York, NY: Wiley, 1995.

[30] E. D. Sontag, "Input to state stability: Basic concepts and results," in *Nonlinear and Optimal Control Theory*, P. Nistri and G. Stefani, Eds. Berlin: Springer-Verlag, 2007, pp. 163–220.

[31] Z. P. Jiang, A. R. Teel, and L. Praly, "Small-gain theorem for ISS systems and applications," *Mathematics of Control, Signals and Systems*, vol. 7, no. 2, pp. 95–120, 1994.

[32] R. Song and F. L. Lewis, "Robust optimal control for a class of nonlinear systems with unknown disturbances based on disturbance observer and policy iteration," *Neurocomputing*, vol. 390, pp. 185–195, 2020.

[33] V. Narayanan, H. Modares, S. Jagannathan, and F. L. Lewis, "Event-driven off-policy reinforcement learning for control of interconnected systems," *IEEE Transactions on Cybernetics*, vol. 52, no. 3, pp. 1936–1946, 2022.

[34] C. De Persis and P. Tesi, "Low-complexity learning of linear quadratic regulators from noisy data," *Automatica*, vol. 128, p. 109548, 2021.

[35] B. Pang, T. Bian, and Z. P. Jiang, "Robust policy iteration for continuous-time linear quadratic regulation," *IEEE Transactions on Automatic Control*, vol. 67, no. 1, pp. 504–511, 2022.

[36] F. Blanchini, "Feedback control for linear time-invariant systems with state and control bounds in the presence of disturbances," *IEEE Transactions on Automatic Control*, vol. 35, no. 11, pp. 1231–1234, 1990.

[37] A. Saberi, Z. Lin, and A. R. Teel, "Control of linear systems with saturating actuators," *IEEE Transactions on Automatic Control*, vol. 41, no. 3, pp. 368–378, 1996.

[38] P. M. Patre, W. MacKunis, K. Dupree, and W. E. Dixon, "Modular adaptive control of uncertain Euler–Lagrange systems with additive disturbances," *IEEE Transactions on Automatic Control*, vol. 56, no. 1, pp. 155–160, 2010.

[39] B. J. Gravell, P. M. Esfahani, and T. H. Summers, "Robust control design for linear systems via multiplicative noise," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 7392–7399, 2020.

[40] B. Gravell, M. Gargiani, J. Lygeros, and T. H. Summers, "Policy Iteration for Multiplicative Noise Output Feedback Control," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 2967–2972.

[41] B. Gravell and T. Summers, "Robust learning-based control via bootstrapped multiplicative noise," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 599–607.

[42] H. Modares, M.-B. N. Sistani, and F. L. Lewis, "A policy iteration approach to online optimal control of continuous-time constrained-input systems," *ISA transactions*, vol. 52, no. 5, pp. 611–621, 2013.

[43] D. Liu, H. Li, and D. Wang, "Neural-network-based zero-sum game for discrete-time nonlinear systems via iterative adaptive dynamic programming algorithm," *Neurocomputing*, vol. 110, pp. 92–100, 2013.

[44] S. Bhasin, M. Johnson, and W. E. Dixon, "A model-free robust policy iteration algorithm for optimal control of nonlinear systems," in *49th IEEE Conference on Decision and Control (CDC)*. IEEE, 2010, pp. 3060–3065.

[45] Q.-Y. Fan, D. Wang, and B. Xu, "H_∞ codesign for uncertain nonlinear control systems based on policy iteration method," *IEEE Transactions on Cybernetics*, 2021.

[46] B. Gravell, P. M. Esfahani, and T. Summers, "Learning optimal controllers for linear systems with multiplicative noise via policy gradient," *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5283–5298, 2020.

[47] K. Zhang, B. Hu, and T. Basar, "On the stability and convergence of robust adversarial reinforcement learning: A case study on linear quadratic systems," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 056–22 068, 2020.

[48] ——, "Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 179–190.

[49] M. Turchetta, A. Krause, and S. Trimpe, "Robust model-free reinforcement learning with multi-objective Bayesian optimization," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 702–10 708.

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

[51] H. K. Khalil, *Nonlinear Systems*. NJ: Prentice Hall PTR, 2002.

[52] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[53] W. M. Wonham, "Controllability subspaces," in *Linear Multivariable Control*. Springer, 1985, pp. 103–130.

[54] P. Kundur, "Power systems stability and control. New York, McGraw-Hill," in *Conference Proceedings*, 1994.

[55] G. Guo, Y. Wang, and D. J. Hill, "Nonlinear output stabilization control for multimachine power systems," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 47, no. 1, pp. 46–53, 2000.