ROBUSTNESS OF ONLINE IDENTIFICATION-BASED POLICY ITERATION TO NOISY DATA

Bowen Song, Andrea Iannelli
Institute for Systems Theory and Automatic Control
University of Stuttgart
Stuttgart
{bowen.song, andrea.iannelli}@ist.uni-stuttgart.de

ABSTRACT

This article investigates the core mechanisms of indirect data-driven control for unknown systems, focusing on the application of policy iteration (PI) within the context of the linear quadratic regulator (LQR) optimal control problem. Specifically, we consider a setting where data is collected sequentially from a linear system subject to exogenous process noise, and is then used to refine estimates of the optimal control policy. We integrate recursive least squares (RLS) for online model estimation within a certainty-equivalent framework, and employ PI to iteratively update the control policy. In this work, we investigate first the convergence behavior of RLS under two different models of adversarial noise, namely point-wise and energy bounded noise, and then we provide a closed-loop analysis of the combined model identification and control design process. This iterative scheme is formulated as an algorithmic dynamical system consisting of the feedback interconnection between two algorithms expressed as discrete-time systems. This system theoretic viewpoint on indirect data-driven control allows us to establish convergence guarantees to the optimal controller in the face of uncertainty caused by noisy data. Simulations illustrate the theoretical results.

***K*eywords** Data-driven Control · Policy Iteration · Recursive Least Squares · Robustness · Nonlinear Systems

1 Introduction

Data-driven control is a very active area of research aimed at developing control strategies for systems where a precise mathematical model is unavailable, a scenario increasingly common in complex modern applications. This field encompasses a wide range of approaches with different problem settings, techniques, and objectives. It is beyond the scope of this Introduction to review them all, and we refer the reader to the following works and references therein [1, 2, 3, 4, 5, 6]. One direction closely related to this work is indirect data-driven control, where data is collected first to estimate a system model, which is then used inside model-based control methods [7, 8, 9, 10]. This approach makes use of system identification [11] and, in cases where the controller is updated during operation, is related to indirect adaptive control [12] or model-based reinforcement learning [13]. By blending data-driven insights with established model-based strategies, indirect data-driven control offers a flexible framework for tackling control problems in complex, dynamic environments.

In this article, we focus on a classic problem of increasing importance in the optimal control and reinforcement learning communities: policy iteration (PI) for solving the linear quadratic regulator (LQR) problem. PI is a dynamic programming algorithm for optimal control [14, 15] that plays a foundational role in approximate dynamic programming and reinforcement learning algorithms [16, 17, 18, 19, 20]. The PI algorithm consists of two main steps—policy evaluation and policy improvement—both of which traditionally rely on an accurate model of the plant. In the standard formulation, convergence to the optimal policy is guaranteed under certain assumptions about the cost and system's dynamics [21].

The LQR problem is a foundational optimal control problem frequently used as a benchmark to compare data-driven control approaches and, owing to its tractability, analytically understand their fundamental properties [7, 18, 20, 22,

23, 24]. For example, [18] investigates the impact of additive uncertainties in model-based PI for continuous-time LQR, while [20] examines the robustness of PI in the presence of parameter uncertainties. The regret analysis of system identification and LQR algorithms has been investigated from a statistical learning perspective in [25, 26, 27]. In [23], a data-driven policy gradient method that integrates recursive least squares (RLS) with a model-based policy gradient approach is proposed, with convergence analyzed using averaging theory, and in [28], an adaptive control framework is proposed for LQR problem. Additionally, the authors' previous work [24] compared indirect and direct data-driven PI for LQR and provided their advantages and disadvantages via theoretical analysis. System-theoretic tools [29] were employed for analysis in [23, 24, 28]. The works discussed earlier [23, 24, 25, 26, 27] adopt an indirect data-driven control framework, which involves system identification followed by controller design based on the estimated system dynamics. In contrast, the direct data-driven control framework bypasses the system identification step and directly optimizes the control policy. In [30, 31], direct data-driven policy gradient methods leveraging zeroth-order optimization were proposed for the noise-free discrete-time and continuous-time LQR problem. Similarly, [32] introduced a data-driven policy gradient method incorporating a novel zeroth-order gradient estimation technique for the noise-free LQR problem. More recently, [33] proposed a direct adaptive data-driven policy gradient method to handle LQR with noise.

In this study, we develop an indirect data-driven policy iteration approach to solve the LQR problem for an unknown system subject to additive adversarial process noise. Specifically, we consider the twofold scenario where the noise is point-wise bounded and energy bounded. We begin by examining the convergence properties of RLS identification, providing a finite-sample analysis that extends existing asymptotic convergence results for RLS [34]. Our analysis is meaningful for providing guarantees in indirect data-driven control that employ RLS for online system identification. Then, we consider the feedback interconnection between the RLS algorithm and the PI scheme, where the gain matrix is refined iteratively through PI steps that use model estimates generated by RLS from online noisy data. By leveraging an algorithmic dynamical systems viewpoint on this interconnection, we frame this iterative process as a nonlinear feedback interconnection and carry out a system theoretic closed-loop analysis. With these results, we establish the conditions under which the algorithmic system converges to the desired values (i.e., the optimal controller and the true system model) and, if convergence is not achieved, we provide a guaranteed upper bound on the suboptimal solution. Our analysis captures the noise in the online collected data as a source of disturbance, enabling an input-to-state stability result with an intuitive, practical interpretation. In contrast to previous studies, such as [23, 24, 28, 30, 31, 32], which assume noise-free data, our approach accommodates adversarial noise and relaxes assumptions necessary for closed-loop analysis compared to our previous work [24]. The analysis in our work provides insights into the impact of noise within the indirect data-driven policy iteration framework. This work serves as an example for analyzing online concurrent learning and controller design algorithms, highlighting how noise influences convergence and control performance.

The main contributions of this work are summarized as follows:

- Convergence analysis of RLS under pointwise bounded noise and energy-bounded noise.
- A system-theoretic analysis of the concurrent learning and controller design algorithm using noisy data.

The paper is organized as follows: Section 2 introduces the problem setting and provides essential preliminaries. Section 3 investigates the convergence properties of recursive least squares with adversarial noisy data. Section 4 details the methodologies of the indirect data-driven policy iteration and analyzes the convergence properties of the coupled RLS and PI system. Section 5 illustrates the theoretical findings. Finally, Section 6 provides a concluding summary of the work.

Notations:

We denote by $A \succeq 0$ and $A \succ 0$ a positive semidefinite and positive definite matrix A, respectively. The symbol $\mathbb{S}_+^n$ represents the set of real $n \times n$ symmetric and positive semidefinite matrices. The sets of non-negative and positive integers are denoted by $\mathbb{Z}_+$ and $\mathbb{Z}_{++}$, respectively. The Kronecker product is represented as $\otimes$, and $vec(A) = [a_1^\top, a_2^\top, ..., a_n^\top]^\top$ stacks the columns a_i of matrix A into a vector. The symbols $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the floor function, which returns the greatest integer smaller or equal than $x \in \mathbb{R}$, and ceil function, which returns the smallest integer greater or equal than $x \in \mathbb{R}$, respectively. For matrices and vectors, $|\cdot|$ denotes their Frobenius and Euclidean norm, respectively. A function belongs to class $\mathcal{K}$ if it is continuous, strictly increasing, and vanishing at the origin. A function $\beta(x,t)$ is called $\mathcal{KL}$ function if $\beta(x,t)$ decreases to 0 as $t \to 0$ for every $x \geq 0$ and $\beta(\cdot,t) \in \mathcal{K}$ for all $t \geq 0$. For $Y \in \mathbb{R}^{m\times n}$ and $r > 0$, we define $\mathcal{B}_r(Y) := \{X \in \mathbb{R}^{m\times n} \big| \, |X - Y| < r\}$. We consider generic sequences $\{Y_t\}$ as maps $\mathbb{Z}_+ \to \mathbb{R}^{m\times n}$, and we denote by $\|Y\|_\infty := \sup_{t\in\mathbb{Z}_+} |Y_t|$ and $\|Y\|_2 := \sum_{t=0}^{\infty} |Y_t|$.

2 Problem Setting and Preliminaries

We consider discrete-time linear time-invariant (LTI) systems of the form

$$x_{t+1} = Ax_t + Bu_t + w_t, \tag{1}$$

where $x_t \in \mathbb{R}^{n_x}$ is the system state, $u_t \in \mathbb{R}^{n_u}$ is the control input and t denotes the timestep. The system matrices (A, B) are unknown but assumed to be stabilizable, as is standard in data-driven control approaches [22, 35]; $w_t \in \mathbb{R}^{n_x}$ represents the adversarial process noise acting on the system.

In this work, we consider two models for the noise:

- point-wise bounded noise [36], where the noise sequence $\{w_t\}$ satisfies:

$$\|w\|_\infty \leq L_\infty, \quad L_\infty \in (0, \infty), \tag{2}$$

  where L_∞ is an upper bound on the noise magnitude;
- energy bounded noise [37], where the noise sequence $\{w_t\}$ satisfies:

$$\|w\|_2 \leq L_2, \quad L_2 \in (0, \infty), \tag{3}$$

  where L_2 represents the noise energy. This type of noise is a specific form of case of point-wise bounded noise with the additional property, the magnitude of the noise converges to zero quickly enough to be summable, implying $\lim_{t\to\infty} |w_t| = 0$, which offers advantages in certain control applications.

The objective is to design a state-feedback controller $u_t = Kx_t$ that minimizes the following infinite horizon cost for the noise-free plant:

$$J(x_t, K) = \sum_{k=t}^{+\infty} r(x_k, u_k) = \sum_{k=t}^{+\infty} x_k^\top Q x_k + u_k^\top R u_k, \tag{4}$$

where $R \succ 0$ and $Q \succeq 0$. When a stabilizing gain K is applied, ensuring that $A+BK$ is Schur stable, the corresponding cost $J(x_t, K)$ can be expressed in terms of the quadratic form $x_t^\top P x_t$. Here, $P \succ 0$ represents the quadratic kernel of the cost function associated with K [22], which is determined by the model-based Bellman equation:

$$P = Q + K^\top R K + (A + BK)^\top P (A + BK). \tag{5}$$

In optimal control theory [14], it is well established that the solution to the linear quadratic regulator (LQR) problem is a linear state-feedback control. The optimal feedback gain K^* is determined by:

$$K^* = -(R + B^\top P^* B)^{-1} B^\top P^* A, \tag{6a}$$

$$P^* = Q + A^\top P^* A - A^\top P^* B (R + B^\top P^* B)^{-1} B^\top P^* A, \tag{6b}$$

where P^* is the quadratic kernel of the optimal cost (value function) and is the unique solution of the discrete algebraic Riccati equation (DARE) in (6b). The system of equations in (6) provides a way to compute the optimal feedback gain K^* that minimizes cost (4).

2.1 Policy Iteration

Even when the system model is known, directly solving DARE (6b) can become computationally challenging for high-dimensional systems. Policy iteration (PI) offers an efficient, iterative method to compute the optimal gain K^* by-passing this calculation. The fundamental model-based version of the PI algorithm [38], which requires knowledge of the system matrices A and B, is summarized in Algorithm 1.

The key properties of Algorithm 1 are presented in the following theorem.

Theorem 1 *Properties of model-based PI [38][24, Theorem 4]*
If the system dynamics (A, B) are stabilizable, and K_0 is stabilizing, then

1. $P_0 \succeq P_1 \succeq ... \succeq P^*$;
2. K_i stabilizes the system (A, B), $\forall i \in \mathbb{Z}_+$;
3. $\lim_{i\to\infty} P_i = P^*$, $\lim_{i\to\infty} K_i = K^*$;

Algorithm 1 Model-based policy iteration.

Require: A, B, a stabilizing policy gain K_0
for $i = 0, 1, ..., +\infty$ **do**
Policy Evaluation: find P_i

$$P_i = Q + K_i^\top R K_i + (A + BK_i)^\top P_i (A + BK_i) \tag{7}$$

Policy Improvement: update gain K_{i+1}

$$K_{i+1} = -(R + B^\top P_i B)^{-1} B^\top P_i A \tag{8}$$

end for

4. $|P_{i+1} - P^*| \leq c|P_i - P^*|$ with $c < 1, \forall i \in \mathbb{Z}_+$.

This theorem establishes that, under stabilizability of (1) and appropriate initialization of K_0, the sequence $\{P_i\}$ generated by policy iteration converges exponentially to the optimal solution P^*, with K_i stabilizing the system at each iteration. Theorem 1 is a standard result on PI. However, leveraging a dynamical system viewpoint, we can obtain an additional results.

2.1.1 PI System Analysis

In [20], we investigated the convergence of PI algorithm with nominal system (A, B) by equivalently reformulating it as a dynamical system. The main steps are as follows. Define the functions $\alpha(P_i) := B^\top P_i A$ and $\beta(P_i) := R + B^\top P_i B$, where $\beta(P_i)$ is a positive definite matrix and thus always invertible. By substituting the policy improvement step (8) into the policy evaluation step (7), the relationship between P_i and P_{i+1} is given by:

$$\begin{aligned} P_{i+1} = Q &+ A^\top P_{i+1} A \\ &+ \alpha(P_i)^\top \beta(P_i)^{-1} \beta(P_{i+1}) \beta(P_i)^{-1} \alpha(P_i) \\ &- \alpha(P_{i+1})^\top \beta(P_i)^{-1} \alpha(P_i) \\ &- \alpha(P_i)^\top \beta(P_i)^{-1} \alpha(P_{i+1}). \end{aligned} \tag{9}$$

Using the identity $vec(EFG) = (F^\top \otimes E) vec(G)$ from [39] and defining

$$\Gamma(P_i) := Q + \alpha(P_i)^\top \beta(P_i)^{-1} R \beta(P_i)^{-1} \alpha(P_i), \tag{10}$$

we can rewrite (9) as:

$$\mathcal{A}(P_i) vec(P_{i+1}) = vec\left(\Gamma(P_i)\right), \tag{11}$$

where $\mathcal{A}(P_i) := I_{n_x} \otimes I_{n_x} - \Omega(P_i) \otimes \Omega(P_i)$ and $\Omega(P_i) := A^\top - \alpha(P_i)^\top \beta(P_i)^{-1} B^\top$.
If $\mathcal{A}(P_i)$ is invertible, we have:

$$vec(P_{i+1}) = \mathcal{A}(P_i)^{-1} vec\left(\Gamma(P_i)\right). \tag{12}$$

The transformation from (9) to (12) involves reshaping the vectorized terms back into a square matrix, thereby establishing the iterative relationship between P_{i+1} and P_i. This process can be formalized as:

$$P_{i+1} = \mathcal{L}^{-1}_{(A,B,P_i)}\left(\Gamma(P_i)\right). \tag{13}$$

where $\mathcal{L}^{-1}_{(\cdot)}(\cdot)$ is an operator that reconstructs the matrix P_{i+1} using (A, B) and P_i.

This formulation allows the sequence $\{P_i\}$ obtained from Algorithm 1 to be interpreted as a discrete-time dynamical system, abstracting the PI algorithm into an algorithmic dynamic and enabling the analysis of its convergence properties, which serves as the foundation for the subsequent analysis. To this aim, the invertibility of $\mathcal{A}(P_i)$ must be ensured. According to Theorem 1, when $P_i \succeq P^*$, the invertibility of $\mathcal{A}(P_i)$ is guaranteed. This condition yields convergence of (13) to P^*, as established in Theorem 1.

Additionally, in [20, Theorem 4], we explored an alternative condition that guarantees the invertibility of $\mathcal{A}(P_i)$ and ensures exponential convergence, without relying on the well-known condition $P_i \succeq P^*$, as discussed in Theorem 1.

Theorem 2 (Exponential convergence of PI [20]) *There exists a constant $\delta_1 > 0$, such that for any $P_i \in \mathcal{B}_{\delta_1}(P^*)$, $\mathcal{A}(P_i)$ is invertible and the following inequality holds:*

$$|P_{i+1} - P^*| \leq \sigma |P_i - P^*|, \qquad \forall i \in \mathbb{Z}_+, \tag{14}$$

where $\sigma \in (0, 1)$.

The advantage of Theorem 2 is to guarantee the existence of a region around the optimal P^* such that if P_0 is initialized there, the sequence $\{P_i\}$ generated by PI guarantees the invertibility of $\mathcal{A}(P_i)$. Figure 1 illustrates the region where $P \succeq P^*$ as the shaded area, indicating where convergence is guaranteed by Theorem 1. The remaining region, depicted within the circle, represents the area where convergence is ensured by Theorem 2.

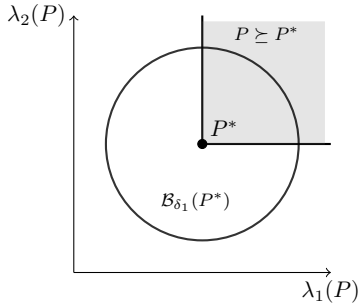

Figure 1: 2-dimensional Graphic Representation

2.2 Recursive Least Squares

When the system dynamics are unknown, least squares identification is a possible strategy to identify the model parameters. We can rewrite system (1) as:

$$x_{t+1} = Ax_t + Bu_t + w_t = \underbrace{[A\ B]}_{=:\theta} \underbrace{\begin{bmatrix} x_t \\ u_t \end{bmatrix}}_{=:d_t} + w_t. \tag{15}$$

Given a dataset $\{d_k, x_{k+1}\}_{k=1}^T$ collected over a trajectory of length T, an estimate $\hat{\theta}$ of system matrix θ can be obtained by minimizing the least-squares loss function [11]:

$$\theta \in \arg\min_{\hat{\theta}} \sum_{k=1}^{T} (x_{k+1} - \hat{\theta} d_k)^\top (x_{k+1} - \hat{\theta} d_k). \tag{16}$$

When the matrix $H_T := \left(\sum_{k=1}^{T} d_k d_k^\top \right)$ is invertible, $\hat{\theta}$ has a closed-form solution:

$$\hat{\theta} = \left(\sum_{k=1}^{T} x_{k+1} d_k^\top \right) H_T^{-1}. \tag{17}$$

This (batch) least squares approach estimates the parameters in a single step, utilizing all data points at once. In contrast, the recursive least squares (RLS) algorithm is particularly valuable for online estimation scenarios [12], whereby estimates are incrementally updated as new data becomes available. Defining the estimated system matrix at time t as $\hat{\theta}_t := [\hat{A}_t, \hat{B}_t]$, the RLS algorithm update equations, are given as follows and summarized in Algorithm 2.

$$H_t = H_{t-1} + d_t d_t^\top, \tag{18a}$$

$$\hat{\theta}_t = \hat{\theta}_{t-1} + (x_{t+1} - \hat{\theta}_{t-1} d_t) d_t^\top H_t^{-1}. \tag{18b}$$

Algorithm 2 Recursive least squares.

Require: An initial estimate of the system dynamic $\hat{\theta}_0$ and $H_0 \succ 0$
  for $t = 1, ..., \infty$ **do**
    Given $\{x_{t+1}, d_t\}$
    $H_t = H_{t-1} + d_t d_t^\top$
    $\hat{\theta}_t = \hat{\theta}_{t-1} + (x_{t+1} - \hat{\theta}_{t-1} d_t) d_t^\top H_t^{-1}$
  end for

In the context of RLS, it is essential to quantify the time-varying estimation error, denoted as $\Delta\theta_t := \hat{\theta}_t - \theta$, which evolves as data are collected over time. This term arises due to the initial estimation error and the effect of the noise models (2) and (3). Using Algorithm 2, we can derive the recursive expression for the estimation error $\Delta\theta_t$ as follows:

$$\begin{aligned}
\Delta\theta_t &= \hat{\theta}_{t-1} H_{t-1} H_t^{-1} + (\theta d_t + w_t) d_t^\top H_t^{-1} - \theta H_t H_t^{-1} \\
&= (\hat{\theta}_{t-1} - \theta) H_{t-1} H_t^{-1} + w_t d_t^\top H_t^{-1} \\
&= (\hat{\theta}_{t-2} - \theta) H_{t-2} H_t^{-1} + (w_t d_t^\top + w_{t-1} d_{t-1}^\top) H_t^{-1} \\
&= (\hat{\theta}_0 - \theta) H_0 H_t^{-1} + \left(\sum_{k=1}^{t} w_k d_k^\top \right) H_t^{-1}.
\end{aligned} \tag{19}$$

In the derivations above, the first equality uses the RLS update equation in (18) and the last equality is obtained by the recursively applying (19). In the next section, we will analyze how the estimation error behaves under the presence of adversarial noisy data.

3 Recursive Least Squares with Adversarial Noise Data

Before analyzing the property of RLS with noisy data, we first recall a property of data sequence $\{d_t\}$, where d_t is defined in (15), which plays a crucial role in ensuring the convergence of the RLS estimator. This property, named local persistency, captures the excitation level of the data sequence.

Definition 1 *Local persistency [24, Definition 2]*
A sequence $\{Y_t\} \in \mathbb{S}_+^n$ is locally persistent if there exist $N \geq 1, M \geq 1$ and $\alpha > 0$ such that , for all $j = Mk + 1$ with $k \in \mathbb{Z}_+$,

$$\sum_{t=0}^{N-1} Y_{t+j} \succeq \alpha I_n. \tag{20}$$

The numbers α and N are respectively, the lower bound and persistency window of $\{Y_t\}$. M is the persistency interval. A sequence $\{Y_t\} \in \mathbb{R}^{n \times m}$ is locally persistent if $\{Y_t Y_t^\top\}$ is locally persistent.

The concept of local persistency was first introduced in our previous work [24] as a relaxed condition of the persistency condition from in [34]. Local persistency provides a sufficient condition for the convergence of the RLS algorithm with noise-free data, as demonstrated in [24, Theorem 2]. We introduce the following assumption which holds throughout the work.

Assumption 1 *The data sequence $\{d_t\}$ is locally persistent with parameters $N = N_d,\ M = M_d$ and $\alpha = \alpha_d$.*

Assumption 1 can be met by appropriately selecting the excitation signal u_t. In a later section, we will address how to design u_t to satisfy this assumption. Building on this assumption, we now extend the analysis to include the convergence of RLS under the influence of adversarial noise. To facilitate this, we introduce an additional assumption regarding the data sequence:

Assumption 2 (Boundedness of data sequence $\{d_t\}$) *The data sequence $\{d_t\}$ satisfies:*

$$\|d\|_\infty \leq \bar{d}, \tag{21}$$

where $\bar{d} \in (0, \infty)$ is a constant.

Because of the boundedness of the noise sequence $\{w_t\}$, Assumption 2 is guaranteed if we apply a stabilizing gain K. Having established these preliminaries, we now present the following theorem that analyzes the convergence properties of the RLS estimation error in the presence of bounded noisy data.

Theorem 3 *If Assumptions 1 and 2 are satisfied and the noise satisfies* (2)*, then the estimation error of RLS initialized with $\hat{\theta}_0$ and $H_0 = aI (a > 0)$ is bounded by:*

$$|\hat{\theta}_t - \theta| \leq \beta_\theta(|\hat{\theta}_0 - \theta|, t) + \gamma_\theta(\|w\|_\infty), \quad \forall t \in \mathbb{Z}_{++} \tag{22}$$

where $\beta_\theta(|\hat{\theta}_0 - \theta|, t) := \frac{a(M_d+N_d)|\hat{\theta}_0-\theta|}{\min(a,\alpha_d)t}$; $\gamma_\theta(x) := \bar{d}\eta x$; $\bar{d}$ is defined in (21)*; $\eta := \frac{(n_x+n_u)(M_d+N_d)}{\min(a,\alpha_d)}$.*

The proof of Theorem 3 is provided in Appendix A.1. The result of Theorem 3 can be interpreted as an input-to-state stability (ISS) result [40, 41]. The function $\beta_\theta(\cdot,\cdot)$ is a $\mathcal{KL}$ function, representing the error due to initialization $\hat{\theta}_0$, which decreases to zero as t approaches infinity. The function $\gamma_\theta(\cdot)$ is a $\mathcal{K}$ function, capturing the error introduced by the noise term w_t. This function is non-zero unless $\|w\|_\infty = 0$. Based on Theorem 3, we can derive the following corollary, which is a standard corollary of ISS results.

Corollary 1 *Using the assumptions and notations of Theorem 3, if $\lim_{t\to\infty} |w_t| = 0$, then we have $\lim_{t\to\infty} |\hat{\theta}_t - \theta| = 0$.*

The proof of Corollary 1 closely follows the steps outlined in [24, Appendix D3] and is omitted here. As discussed in Section 2, the energy-bounded noise condition in (3) represents a particular case of (2), where $\lim_{t\to\infty} |w_t| = 0$. Thus, Theorem 3 and Corollary 1 are applicable. However, by directly using (3), a stronger result than those provided in (22) and Corollary 1 can be achieved.

Corollary 2 (RLS with energy bounded noisy data) *Using the assumptions and notations of Theorem 3, if the noise is energy bounded, i.e. satisfying* (3)*, the estimation error of RLS is bounded by:*

$$|\hat{\theta}_t - \theta| \leq \beta_\theta(|\hat{\theta}_0 - \theta|, t) + \beta_e(\|w\|_2, \sqrt{t}),\ \forall t \in \mathbb{Z}_{++}, \tag{23}$$

with $\beta_e(\|w\|_2, \sqrt{t}) := \bar{d}\eta \frac{\|w\|_2}{\sqrt{t}}$.

The proof of Corollary 2 is provided in Appendix A.2. According to the corollary, the estimation error is bounded by two $\mathcal{KL}$-function. As t approaches infinity, the estimation error converges to zero, which recovers with the result in Corollary 1.

The analysis in this section provides analytical insight into the role of noise in RLS, illustrating how noise affects estimation accuracy and convergence. These results can be integrated with robust control techniques to guarantee the performance of indirect data-driven control employing online RLS algorithms.

Before concluding our discussion on RLS, we quantify the maximum estimation error of RLS for point-wise bounded noise, which can be derived from Theorem 3 as:

$$\overline{\Delta\theta}(\hat{\theta}_0, \bar{d}) := \max\{|\hat{\theta}_0 - \theta|, \beta_\theta(|\hat{\theta}_0 - \theta|, 1) + \bar{d}\eta\|w\|_\infty\}. \tag{24}$$

The first term in (24) represents the estimation error determined by the initialization at $t = 0$, and the second term is the upper bound provided by Theorem 3 for $t \geq 1$. This quantity can be interpreted as the largest estimation error for all $t \in \mathbb{Z}_+$, i.e. $|\Delta\theta_t| \leq \overline{\Delta\theta}(\hat{\theta}_0, \bar{d})$, and it is determined by the initialization $\hat{\theta}_0$ and the upper bound on the data sequence $\bar{d}$ defined in (21). Similarly, for the energy bounded noise satisfying (3), we can derive the maximum estimation error from Corollary 1 as:

$$\overline{\Delta\theta}_e(\hat{\theta}_0, \bar{d}) := \max\{|\hat{\theta}_0 - \theta|, \beta_\theta(|\hat{\theta}_0 - \theta|, 1) + \bar{d}\eta\|w\|_2\}. \tag{25}$$

4 Online Identification-based Policy Iteration

In this section, we analyze the online identification-based policy iteration (ORLS+PI), which integrates the model-based PI from Algorithm 1 with the RLS algorithm presented in Algorithm 2. This approach offers a practical solution for performing policy iteration in scenarios where the system dynamics are unknown. By concurrently optimizing the policy and conducting online system identification, the algorithm aims to improve the control performance iteratively. Our primary focus is to investigate the convergence properties and limitations of this combined approach from a system-theoretic perspective and its robustness to noise.

4.1 Algorithm Definition

For the ORLS+PI algorithm, we collect the data sequence $\{d_t\}$ online with the control input u_t given as:

$$u_t = \hat{K}_t x_t + e_t, \tag{26}$$

where $\hat{K}_t$ is the feedback gain and e_t is a potentially non-zero feedforward term.

Remark 1 (Remark on $\hat{K}_t$) *The gain $\hat{K}_t$ in* (26) *is referred to the on-policy gain [23], meaning that data are generated using the policy currently being updated. In this case, the $\hat{K}_t$ is generated by ORLS+PI algorithm. However, as discussed in [24, Section 5.4], one advantage of indirect data-driven policy iteration is that the excitation can be also performed off-policy. i.e. the data can be generated using a different stabilizing policy K, that is not updated by the algorithmic dynamics.*

Remark 2 (Remark on e_t) *The term e_t represents an additional degree of freedom of the online policy, which can be used, for example, as an exploratory signal that explores the system in a random or targeted way [8, 42]. The purpose of including e_t is to ensure the local persistency of the data sequence $\{d_t\}$, i.e. Assumption 1. However, it is important to note that the subsequent analysis is agnostic to the specific choice of e_t. In this work, we assume that the sequence of the signal $\{e_t\}$ is bounded, i.e.*

$$\|e\|_\infty \leq \bar{e}. \tag{27}$$

where $\bar{e} \in (0, \infty)$ is a constant that represents the upper bound of the signal magnitude at each timestep.

The ORLS+PI algorithm involves at each iteration t the following steps:

- Given a policy gain $\hat{K}_t$, which either originates from the initialization $(t = 1)$ or the previous timestep, the cost function kernel estimate $\hat{P}_t$ is computed by solving the model-based Bellman equation (5) using the current system estimates $\left(\hat{A}_{t-1}, \hat{B}_{t-1}\right)$:

$$\begin{aligned} \hat{P}_t = Q + \hat{K}_t^\top R \hat{K}_t + \\ \left(\hat{A}_{t-1} + \hat{B}_{t-1}\hat{K}_t\right)^\top \hat{P}_t \left(\hat{A}_{t-1} + \hat{B}_{t-1}\hat{K}_t\right). \end{aligned} \tag{28}$$

- The physical system is excited with the control input u_t introduced in (26). The state-input data $\{x_t, u_t, x_{t+1}\}$ is then used to recursively update the system dynamics estimates $\left(\hat{A}_t, \hat{B}_t\right)$ using the RLS Algorithm:

$$H_t = H_{t-1} + d_t d_t^\top, \tag{29a}$$

$$\hat{\theta}_t = \left(\hat{\theta}_{t-1} H_{t-1} + x_{t+1} d_t^\top\right) H_t^{-1}. \tag{29b}$$

- Using the updated estimates $\left(\hat{A}_t, \hat{B}_t\right)$, the policy is improved by solving for the new feedback gain $\hat{K}_{t+1}$:

$$\hat{K}_{t+1} = -\left(R + \hat{B}_t^\top \hat{P}_t \hat{B}_t\right)^{-1} \hat{B}_t^\top \hat{P}_t \hat{A}_t. \tag{30}$$

To ensure the feasibility of the ORLS+PI algorithm, particularly regarding equations (28) and (30), we will provide a detailed discussion on this topic in a later section.

Remark 3 (Timestep t) *In this work, we use a single index t for both the RLS estimate update and the PI policy update. While, in principle, each update could be tracked by its own independent index. The analysis in this section can be extended to handle different timescales for each update, following the approach outlined in [24].*

The ORLS+PI algorithm is summarized in Algorithm 3 and is depicted in Figure 2 through a block diagram that emphasizes the dynamic viewpoint leveraged in this work. The closed-loop system, consisting of the physical system and the controller, is connected by the solid black lines in the figure and is subject to the exogenous noise term w_t. The algorithmic dynamics, formed by the PI and RLS algorithms, is placed inside the bottom shaded area and its interconnections are depicted by the dashed black lines.

Algorithm 3 ORLS+PI Algorithm

Require: $\hat{A}_0, \hat{B}_0, H_0$, the initial optimal policy gain $\hat{K}_1$ for system $(\hat{A}_0, \hat{B}_0)$
for $t = 1, ..., \infty$ **do**
Policy Evaluation: find $\hat{P}_t$ **by** (28)
Excite the system with $u_t = \hat{K}_t x_t + e_t$
Collect the data $\leftarrow (x_t, u_t, x_{t+1})$
Use RLS in Algorithm 2 to update $\hat{A}_t, \hat{B}_t$
Policy Improvement: update gain $\hat{K}_t$ **by** (30)
end for

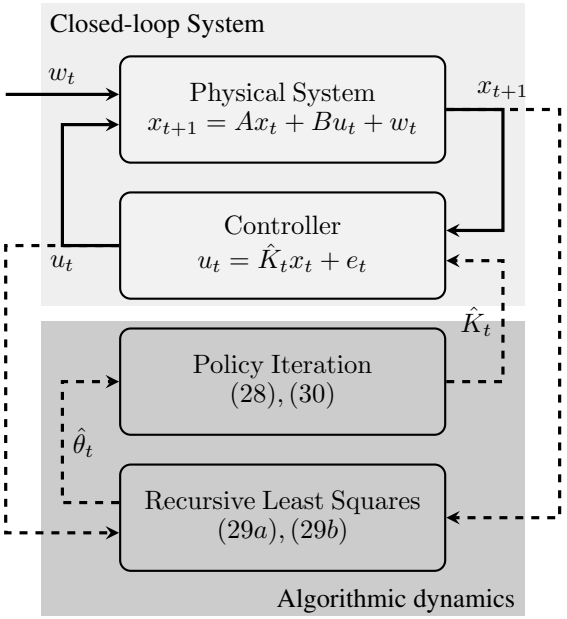

Figure 2: Concurrent identification and policy iteration scheme

4.2 Convergence Analysis of ORLS+PI Algorithm

As illustrated in Figure 2, the dynamics of the policy iteration (PI) and recursive least-squares (RLS) can be analyzed as a feedback interconnection of two coupled dynamical systems. In the "system PI", the inputs are the estimates $\left(\hat{A}_t, \hat{B}_t\right)$ obtained from the RLS, and the dynamics are described by (30) and (28). In the "system RLS", the inputs are the data $\{d_t\}$ and $\{x_{t+1}\}$ collected online from the physical system and perturbed by the noise, with the dynamics described by (29a) and (29b).

The properties of "system PI" were recalled in Section 2.1 and the properties of "system RLS" were investigated in Section 3, which provides insight into the behavior of the RLS algorithm under adversarial noise conditions. To facilitate our analysis, we introduce the following notations:

$$\hat{\alpha}_t := \hat{B}_t^\top \hat{P}_t \hat{A}_t \tag{31a}$$

$$\hat{\beta}_t = \hat{\beta}_t^\top := R + \hat{B}_t^\top \hat{P}_t \hat{B}_t. \tag{31b}$$

Before stating the main result, we introduce the following assumption.

Assumption 3 *The estimates* $\left(\hat{A}_t, \hat{B}_t\right)$ *obtained from RLS are stabilizable* $\forall t \in \mathbb{Z}_+$*. Given a stabilizable estimate* $\left(\hat{A}_t, \hat{B}_t\right)$*, we assume that* $\hat{P}_t \succeq P^*_{\left(\hat{A}_t, \hat{B}_t\right)}$ $\forall t \in \mathbb{Z}_+$*, where* $\hat{P}_t$ *is obtained via* (28) *and* $P^*_{\left(\hat{A}_t, \hat{B}_t\right)}$ *is the quadratic kernel of the value function associated with* $\left(\hat{A}_t, \hat{B}_t\right)$ *and is calculated by solving* (6b)*.*

Assumption 3 is a direct translation in the online identification-based setting of the standard requirement for the formulation of policy iteration, (cf. Theorem 1). For further discussions and details on how to realize this assumption, we refer to [24, Assumption 1, Assumption 2]. We are finally ready to state the main convergence and robustness result of Algorithm 3.

Theorem 4 (ORLS+PI Analysis 1**)** *If Assumption 3 is satisfied, then the ORLS+PI system formulated by* (28)-(30) *admits the following equivalent dynamical system representation:*

$$\begin{aligned}\hat{\theta}_{t+1} = &\left(\hat{\theta}_t \left(H_0 + \sum_{k=1}^{t} d_k d_k^\top\right)\right. \\ &\left. + \sum_{k=1}^{t} x_{t+1} d_t^\top\right)\left(H_0 + \sum_{k=1}^{t} d_k d_k^\top\right)^{-1},\end{aligned} \tag{32a}$$

$$\hat{P}_{t+1} = \mathcal{L}^{-1}_{\left(\hat{A}_t, \hat{B}_t, \hat{P}_t\right)}\left(Q + \hat{\alpha}_t^\top \hat{\beta}_t^{-1} R \hat{\beta}_t^{-1} \hat{\alpha}_t\right). \tag{32b}$$

Additionally, if Assumptions 1 and 2 are satisfied and the noise satisfies (2)*, then with the initialization* $H_0 = aI (a > 0)$ *and arbitrary* $\hat{\theta}_0$*, the estimates* $\hat{P}_t$ *and* $\hat{\theta}_t$ *satisfy the following relationships for all* $t \in \mathbb{Z}_{++}$*:*

$$\left|\hat{P}_t - P^*\right| \leq \beta_c\left(\left|\hat{P}_0 - P^*\right|, t\right) + \gamma_c\left(\|\Delta\theta\|_\infty\right), \tag{33a}$$

$$|\hat{\theta}_t - \theta| \leq \beta_\theta(|\hat{\theta}_0 - \theta|, t) + \gamma_\theta(\|w\|_\infty), \tag{33b}$$

where

- $\beta_c(\cdot,\cdot) := c^t \left|\hat{P}_0 - P^*\right|$ *is a* $\mathcal{KL}$*-function with* $c \in (0,1)$ *defined in Theorem 1;*
- $\gamma_c\left(\|\cdot\|_\infty\right) := \frac{\bar{C}}{1-c}\|\cdot\|_\infty$ *is a* $\mathcal{K}$*-function with constant* $\bar{C}$ *given in the proof* (50)*;*
- $\beta_\theta(\cdot,\cdot)$ *and* $\gamma_\theta(\cdot)$ *are defined in Theorem 3.*

The proof of Theorem 4 is provided in Appendix A.3 and is the result of combining Theorem 3 with [24, Theorem 6].

We observe here that, regarding Assumption 2, there is no guarantee that the stabilizing property of $\hat{K}_t$ will hold for the true system (A, B). In the on-policy setting (see Remark 1), we cannot ensure the boundedness of the data sequence. However, as discussed in Remark 1, the excitation can be performed off-policy. Specifically, all the analyses still hold if the system is excited using a fixed pre-stabilizing gain K. In this off-policy case, Assumption 2 can be guaranteed.

Theorem 4 describes the convergence properties of the ORLS+PI algorithm for arbitrary initial $\hat{\theta}_0$. If an assumption on the maximum estimation error (24), which also depends on $\hat{\theta}_0$ is made, then Assumptions 2 and 3 are not anymore required.

Assumption 4 *The maximum estimation error of RLS satisfies the following condition:*

$$\overline{\Delta\theta}(\hat{\theta}_0, \bar{D}) \leq \min\{\bar{a}_p, \bar{b}_p\}, \tag{34}$$

where $\bar{a}_p$ *and* $\bar{b}_p$ *are constants defined in* (52) *(see Theorem 6 in Appendix A.4) and* $\bar{D}$ *is defined in* (58) *(see Lemma 21 in Appendix A.4).*

The value of $\bar{D}$ is quantitatively determined by both the upper bound of the noise and the sequence $\{\hat{K}_t\}$ applied to the system. Assumption 4 requires that the maximum estimation error from RLS remains within acceptable limits. This can be used in conjunction with recent findings on the inherent robustness of PI with inexact models [20] to show that Algorithm 2 converges under different assumptions than Theorem 4. Under Assumption 4, we can derive the following theorem.

Theorem 5 (ORLS+PI Analysis 2) *If Assumption 4 is satisfied and the initial $\hat{K}_0$ is selected as the optimal gain calculated by solving* (6) *using $(\hat{A}_0, \hat{B}_0, Q, R)$, then the ORLS+PI algorithm formulated by* (28)-(30) *admits the equivalent dynamical system representation in* (32)*. Additionally, if Assumption 1 is satisfied and the noise satisfies* (2)*, then with the initialization $H_0 = aI (a > 0)$ and an initial $\hat{\theta}_0$ satisfying Assumption 4, the estimates $\hat{P}_t$ and $\hat{\theta}_t$ satisfy the following relationships for all $t \in \mathbb{Z}_{++}$:*

$$\left|\hat{P}_t - P^*\right| \leq \beta_\sigma \left(\left|\hat{P}_0 - P^*\right|, t\right) + \gamma_\sigma \left(\|\Delta\theta\|_\infty\right), \tag{35a}$$

$$|\hat{\theta}_t - \theta| \leq \beta_\theta(|\hat{\theta}_0 - \theta|, t) + \gamma_D(\|w\|_\infty), \tag{35b}$$

where:

- $\beta_\sigma(\cdot,\cdot) := \sigma^t \left|\hat{P}_0 - P^*\right|$ *is a $\mathcal{KL}$-function with $\sigma \in (0,1)$ defined in Theorem 2;*
- $\gamma_\sigma(\|\cdot\|_\infty) := \frac{\bar{p}_a + \bar{p}_b}{1-\sigma} \|\cdot\|_\infty$ *is a $\mathcal{K}$-function with $\bar{p}_a$ and $\bar{p}_b$ given in the proof* (52)*;*
- $\beta_\theta(\cdot,\cdot)$ *is defined in Theorem 3;*
- $\gamma_D(\|\cdot\|_\infty) := c_D\|\cdot\|_\infty$ *with $c_D := \bar{D}\eta$; η is defined in Theorem 3 and $\bar{D}$ is defined in* (58)*.*

The proof of Theorem 5 is provided in Appendix A.4. Here, we outline the main steps involved in the proof. The proof relies primarily on Theorem 3, which establishes the convergence of the RLS under a bounded data sequence and point-wise bounded noise, and on [20, Theorem 6], which describes the inherent robustness of PI. The proof proceeds as follows:

1. **Condition on Initialization** $\hat{\theta}_0$: The inherent robustness of PI guarantees that $\hat{K}_t$ stabilizes the system for all $t \in \mathbb{Z}_+$. Because in addition we have point-wise bounded noise and control inputs, we determine the upper bounded of the sequence $\{d_t\}$ denoted by $\bar{D}$. Then we determine the necessary condition (Assumption 4) to sure that Theorem 6 holds;
2. **PI inherent robustness**: Leveraging the robustness properties of PI from [20, Theorem 6], we can directly establish inequality (35a);
3. **System stabilization and bounded data sequence**: We have shown that the data sequence $\{d_t\}$ is upper bounded by $\bar{D}$. This allows us to prove inequality (35b);

Remark 4 (Comparison between Theorems 4 and 5) *Theorem 4 extends the results of [24, Theorem 6] to case studies involving bounded noisy data. Theorem 4 relies on Assumptions 2 and 3 to derive ISS results* (32)*. These assumptions provide a result by imposing no restrictions on the initial condition $\hat{\theta}_0$ of RLS.*

In contrast, Theorem 5 removes the Assumptions 2 and 3 by introducing a specific condition on initialization and the upper bound of the data sequence, which is partially influenced by the noise level, as defined in (34)*. This requirement ensures that the maximum estimation error stays within the level of inherent robustness of PI. Therefore, the results under Theorem 5 only hold when the estimation error is sufficiently small.*

As discussed earlier, for Theorem 4, we can only perform off-policy excitation during the online data collection to ensure the boundedness of the data sequence. However, in the case of Theorem 5, the closed-loop stability of the physical system is guaranteed. Therefore, we can directly employ excitation with the on-policy gain $\hat{K}_t$.

Remark 5 (Remark to Assumption 4) *Assumption 4 cannot be directly verified, as we only know the existence of $\bar{a}_p$ and $\bar{b}_p$. However, from a system-theoretical perspective provided by Theorem 5, we know that if the initial condition is close to the true system and the upper bound of the noise is small, the coupled system is input-to-state stable with respect to the upper bound of the noise and the estimation error of system matrices. Moreover, the on-policy gain ensures stability as stated in Remark 4. In other words, compared to Theorem 5, with better prior knowledge of the system matrices, fewer assumptions are required to guarantee the performance of concurrent learning and controller design procedure.*

Based on Theorem 4 and Theorem 5, we can now derive the following corollaries that help interpret the two theorems in a more intuitive and practical manner.

Corollary 3 (Finite sample analysis of $|\hat{P}_t - P^*|$) *Using the notations and assumptions of Theorem 4 and given an iteration $t_{\mathrm{re}} > 1$, the distance between $|\hat{P}_t - P^*|$ can be quantified as:*

$$\begin{aligned}\left|\hat{P}_t - P^*\right| \leq \beta_c\left(\left|\hat{P}_{t_{\mathrm{re}}} - P^*\right|, t - t_{\mathrm{re}}\right) \\ + \gamma_c\left(\sup_{k\geq t_{\mathrm{re}}} |\Delta\theta_k|\right), \quad \forall t \geq t_{\mathrm{re}};\end{aligned} \tag{36}$$

Similarly, using similar notations and assumptions of Theorem 5, we have:

$$\begin{aligned}\left|\hat{P}_t - P^*\right| \leq \beta_\sigma\left(\left|\hat{P}_{t_{\mathrm{re}}} - P^*\right|, t - t_{\mathrm{re}}\right) \\ + \gamma_\sigma\left(\sup_{k\geq t_{\mathrm{re}}} |\Delta\theta_k|\right), \quad \forall t \geq t_{\mathrm{re}}.\end{aligned} \tag{37}$$

The proof of Corollary 3 follows directly by reformulating the equations (33a) and (35a).

Corollary 4 *Under the conditions of Theorem 4 and Theorem 5, if $\lim_{t\to\infty}|w_t| = 0$, then $\lim_{t\to\infty}|\hat{\theta}_t - \theta| = 0$, $\lim_{t\to\infty}|\hat{P}_t - P^*| = 0$ and $\lim_{t\to\infty}|\hat{K}_t - K^*| = 0$.*

Corollary 4 is a standard corollary of ISS results and it can be proved for example by following the steps outlined in [24, Appendix D3]. From this corollary, if the data sequence is locally persistent and noise term w_t vanished at infinity, $\{\hat{P}_t\}$ obtained from the concurrent learning and controller design algorithm converges asymptotically to the optimal P^*.

Corollary 5 (Energy bounded noise) *Using the notations of Theorem 5, for the energy bounded noise satisfying* (3)*, if*

$$\overline{\Delta\theta}_e(\hat{\theta}_0, \bar{D}) \leq \min\{\bar{a}_p, \bar{b}_p\}, \tag{38}$$

where $\overline{\Delta\theta}_e(\cdot,\cdot)$ is defined in (25)*, then the ORLS+PI algorithm formulated by* (28)-(30) *admits the equivalent dynamical system representation in* (32)*. If Assumption 1 is satisfied, the estimates $\hat{P}_t$ and $\hat{\theta}_t$ satisfy the following relationships for all $t \in \mathbb{Z}_{++}$:*

$$\left|\hat{P}_t - P^*\right| \leq \beta_\sigma\left(\left|\hat{P}_0 - P^*\right|, t\right) + \gamma_\sigma\left(\|\Delta\theta\|_\infty\right), \tag{39a}$$

$$|\hat{\theta}_t - \theta| \leq \beta_\theta(|\hat{\theta}_0 - \theta|, t) + \beta_D(\|w\|_2, \sqrt{t}), \tag{39b}$$

where $\beta_D(\|w\|_2, \sqrt{t}) := \bar{D}\eta\frac{\|w\|_2}{\sqrt{t}}$ and η is defined in Theorem 3.

Corollary 5 can be proved by integrating the results of Corollary 2 with Theorem 5. Based on (39), we can recover the asymptotic results stated in Corollary 3. A similar corollary for Theorem 4 can also be derived by combining Corollary 2 is omitted here.

In this work and previous [24], we analyze the ORLS+PI algorithm as a dynamical system and provide input-to-state stability (ISS) results to characterize the closed-loop behavior. In [24], we focused on noise-free data, considering the persistency level of the data sequence. In contrast, this work accounts for bounded noisy data and assumes that the sequence is locally persistent. The analysis in this section provides a mathematical description of how adversarial noise impacts the performance of the ORLS+PI algorithm. This insight enables us to characterize the conditions under which noise affects estimation accuracy and convergence, informing guidelines for robust algorithm initialization and parameter tuning in noisy environments.

5 Simulations

In this section, we present simulation results[1] to illustrate some of the properties of online identification-based policy iteration discussed in the previous sections.

[1]The Matlab codes used to generate these results are accessible from the repository: https://github.com/col-tasas/2024-SysIDbasedPIwithNoisyData

5.1 Comparison between different types of noise

We consider the following system which was already used in prior studies [7, 22, 24]:

$$x_{t+1} = \underbrace{\begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}}_{A} x_t + \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{B} u_t + w_t. \tag{40}$$

The weight matrices Q and R are set to $0.001I_3$ and I_3, respectively. The initial estimates for system matrices A and B are set as:

$$\begin{aligned} \hat{A}_0 &= A + 0.5I_3, \\ \hat{B}_0 &= B + 0.5I_3. \end{aligned} \tag{41}$$

The matrix H_0 for RLS is initialized as $0.1I_6$. The initial stabilizing policy gain $\hat{K}_0$ is set to the optimal LQR gain associated with $(\hat{A}_0, \hat{B}_0, Q, R)$. The dithering signal e_t of the policy (26) is distributed uniformly with each entry sampled independently from the interval $[-10, 10]$. Figure 3 illustrates the convergence of the quadratic kernel of the value function $\hat{P}_t$, representing the closed-loop evaluation of the cost function with the feedback gain $\hat{K}_t$ under different noise conditions, which are set as:

$$\text{PB1} : |w_t| = \frac{0.5}{t} + 0.5; \tag{42a}$$

$$\text{PB2} : |w_t| = \frac{0.5}{t}; \tag{42b}$$

$$\text{EB} : |w_t| = \frac{0.5}{t^2}. \tag{42c}$$

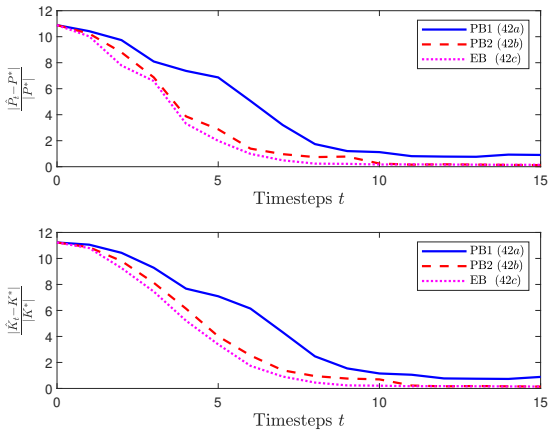

Figure 3: Comparisons of convergence behaviors of ORLS+PI with different types of noise

The blue solid line shows the convergence under point-wise bounded noise. This setup results in a non-vanishing error between $\hat{P}_t(\hat{K}_t)$ and $P^*(K^*)$ due to the persistent noise component, as discussed in Corollary 3. The red dashed line uses noise vanishing as $t \to \infty$ but is not energy bounded. This condition yields convergence to the optimal values, as detailed in Corollary 4. The magenta dotted line shows energy bounded noise. This configuration, in line with Corollary 5, achieves convergence to the optimal values.

5.2 Comparison between Policy Iteration and Policy Gradient

We compare our OLRS+PI algorithm with a recently proposed method that combines online RLS with a model-based policy gradient approach [23], referred to here as ORLS+PG. The system dynamics (A, B) and the weight matrices Q

and R are set according to the example proposed in [23]:

$$A = \begin{bmatrix} -0.53 & 0.42 & -0.44 \\ 0.42 & -0.56 & -0.65 \\ -0.44 & -0.65 & 0.35 \end{bmatrix}, \quad B = \begin{bmatrix} 0.43 & -0.82 \\ 0.53 & -0.78 \\ 0.26 & -0.40 \end{bmatrix},$$
$$Q = \begin{bmatrix} 6.12 & 1.72 & 0.53 \\ 1.72 & 6.86 & 1.72 \\ 0.53 & 1.72 & 5.73 \end{bmatrix}, \quad R = \begin{bmatrix} 1.15 & -0.23 \\ -0.23 & 3.62 \end{bmatrix}.$$

The initial estimates $\hat{A}_0$, $\hat{B}_0$, and the matrix H_0 required for both ORLS+PI and OLRS+PG are set to $1.3A$, $0.7B$, and $H_0 = 0.01I_5$, respectively. The initial feedback gain $\hat{K}_0$ is set to the optimal gain for the LQR problem associated with $(\hat{A}_0, \hat{B}_0, Q, R)$. The OLRS+PG method uses the same online policy employed in Algorithm 2, with a feedback term $\hat{K}_t x_t$ plus a dithering signal $e_t \in [-10, +10]$ to ensure sufficiently informative data. The stepsize γ of ORLS+PG is empirically set to 0.005. Figure 4 investigates the convergence of kernel of closed-loop evaluation $\hat{P}_t$ by considering three different types of noise, set as (42).

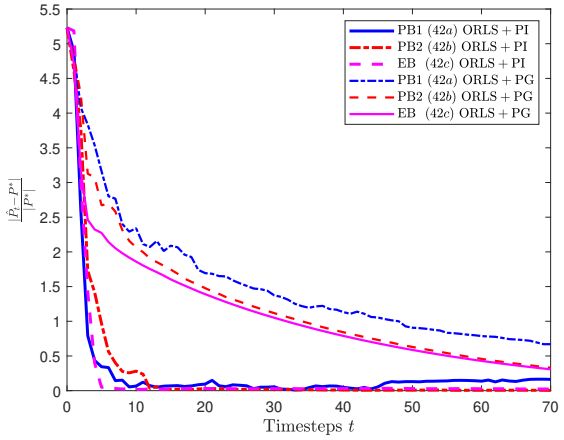

Figure 4: Comparison of ORLS+PI with ORLS+PG

As seen in Figure 4, the ORLS+PI method exhibits faster convergence of $\hat{P}_t$ compared to the ORLS+PG methods. This is due to the nature of the PI method, which can be viewed as a Newton method. For the ORLS+PG methods, the stepsize can only be tuned empirically, and selecting an optimal stepsize to ensure convergence remains an open question. Instead, for ORLS+PI, owing to the analyses carried out in this work, there are systematic guidelines for choosing the initialization based on the bounds of the data sequence. Examining (58) reveals that the upper bound also grows as the noise magnitude increases. Consequently, when the noise is larger, the initialization must be chosen closer to the true system to ensure convergence.

6 Conclusion

In this work, we studied the application of indirect data-driven policy iteration to the LQR problems when data are subject to adversarial bounded noise. First we analyzed the convergence properties of RLS, establishing an upper bound on the estimation error. This result is meaningful for the indirect data-driven control method, as it provides guarantees on control performance by quantifying the accuracy of model estimates obtained from noisy data. Subsequently, we conceptualized the algorithm as a feedback interconnection between an identification scheme and the PI algorithm, both framed as algorithmic dynamical systems that realize concurrent learning and control. We analyzed the convergence properties of such a nonlinear closed-loop under different noise and parameters initialization scenarios to provide a comprehensive picture on the robustness of such data-driven schemes. In future work, it will be important to explore unbounded stochastic noise and investigate its impact on the performance of RLS and the coupled RLS+PI system. Additionally, we aim to explore direct data-driven policy iteration, which bypasses the system identification step and directly utilizes data to formulate the PI procedure, with a particular focus on its performance in noisy data and the relative strengths and weaknesses with respect to indirect schemes.

Acknowledgement

Bowen Song acknowledges the support of the International Max Planck Research School for Intelligent Systems (IMPRS-IS). Andrea Iannelli acknowledges the German Research Foundation (DFG) for support of this work under Germany's Excellence Strategy - EXC 2075 – 390740016.

A Technical Proof

A.1 Proof of Theorem 3

Proof 1 (Proof of Theorem 3) *From* (19)*, we have:*

$$\begin{aligned} |\Delta\theta_t| &\leq a|\Delta\theta_0||H_t^{-1}| + \left(\sum_{k=1}^{t}|w_k||d_k|\right)|H_t^{-1}| \\ &\leq a|\Delta\theta_0||H_t^{-1}| + \bar{d}\left(\sum_{k=1}^{t}|w_k|\right)|H_t^{-1}|. \end{aligned} \tag{43}$$

With the definition of local persistency, we have:

$$\lambda_{\min}(H_t) \geq a + \lfloor \frac{t}{\lceil \frac{N_d}{M_d} \rceil M_d} \rfloor \alpha_d \geq a + \lfloor \frac{t}{M_d + N_d} \rfloor \alpha_d.$$

Then we have:

$$\begin{aligned} |H_t^{-1}| &\leq \frac{n_x + n_u}{a + \lfloor \frac{t}{M_d+N_d} \rfloor \alpha_d} \\ &\leq \frac{(n_x+n_u)(M_d+N_d)}{\min(a,\alpha_d)t},\ \forall t \in \mathbb{Z}_{++}. \end{aligned} \tag{44}$$

Substituting (44) *into* (43)*, we obtain:*

$$|\hat{\theta}_t - \theta| \leq \beta(|\hat{\theta}_0 - \theta|, t) + c\frac{\sum_{k=0}^{t}|w_k|}{t},\ \forall t \in \mathbb{Z}_{++}. \tag{45}$$

Based on the bound defined in (2)*, we obtain:*

$$\sum_{k=1}^{t}|w_k| \leq t \sup_t \sqrt{w_t^\top w_t} \leq t\|w\|_\infty. \tag{46}$$

Substituting (46) *into* (45)*, we conclude the proof of Theorem 3.*

A.2 Proof of Corollary 2

Proof 2 *For the proof of Corollary 2, we use the AM–GM inequality,*

$$\sum_{k=1}^{t}|w_k| \leq \sqrt{t}\sqrt{\sum_{k=1}^{t} w_k^\top w_k} \leq \sqrt{t}\sqrt{\sum_{k=1}^{\infty} w_k^\top w_k} \leq \sqrt{t}\|w\|_2. \tag{47}$$

Substituting (47) *into* (45)*, we conclude the proof.*

A.3 Proof of Theorem 4

Proof 3 (Proof of Theorem 4) *Based on the Assumptions 1 and 2,* (33b) *is directly proved. Now we turn to* (33a)*, Assumption 3 guarantees the formulation of standard PI procedure. Following the same step in [24, Appendix D6]*

$$\hat{P}_{t+1} = \mathcal{L}^{-1}_{\left(A,B,\hat{P}_t\right)}\left(\Gamma(\hat{P}_t)\right) + \varepsilon\left(\Delta A_t, \Delta B_t\right), \tag{48}$$

where

$$\begin{aligned}\varepsilon\left(\Delta A_t, \Delta B_t\right) := &-\mathcal{L}^{-1}_{\left(A,B,\hat{P}_t\right)}\left(\Gamma(\hat{P}_t)\right)\\ &+\mathcal{L}^{-1}_{\left(\hat{A}_t,\hat{B}_t,\hat{P}_t\right)}\left(Q+\hat{\alpha}_t^\top\hat{\beta}_t^{-1}R\hat{\beta}_t^{-1}\hat{\alpha}_t\right),\end{aligned} \tag{49}$$

and $\Gamma(\hat{P}_t)$ *is defined in* (10)*. Using the same arguments in [24], we can prove that:*

$$|\varepsilon\left(\Delta A_t, \Delta B_t\right)| \leq \bar{C}|\Delta\theta_t|, \tag{50}$$

where $\bar{C}$ *is polynomial of* (A, B, Q, R)*. For the detailed computation steps and derivation of* $\bar{C}$*, we refer to [24, Appendix D6]. Then we can prove:*

$$\begin{aligned}|\hat{P}_t - P^*| &\leq c|\hat{P}_{t-1} - P^*| + \bar{C}|\Delta\theta_t|\\ &\leq c^t|\hat{P}_0 - P^*| + \bar{C}\left(1 + c + ... + c^{t-1}\right)\|\Delta\theta\|_\infty\\ &\leq c^t|\hat{P}_0 - P^*| + \frac{\bar{C}}{1-c}\|\Delta\theta\|_\infty.\end{aligned} \tag{51}$$

Then we conclude the proof of (33a)*.*

A.4 Proof of Theorem 5

Proof 4 (Proof of Theorem 5) *In this proof, the robustness of PI algorithms plays a central role, as outlined in our previous work [20, Theorem 7]. For clarity and completeness, we recall this theorem here:*

Theorem 6 (Robustness of PI [20]) *Given* σ *and* δ_1 *defined in Theorem 2, there always exist constants* $\bar{a}_p(\delta_1, \sigma) \geq 0$ *and* $\bar{b}_p(\delta_1, \sigma) \geq 0$ *such that if* $\|a\|_\infty \leq \bar{a}_p$*,* $\|b\|_\infty \leq \bar{b}_p$ *and* $\hat{P}_0 \in \mathcal{B}_{\delta_1}(P^*)$*, where sequences* $\{a_t\}$ *and* $\{b_t\}$ *are defined as*

$$a_t := |\Delta A_t|,\ b_t := |\Delta B_t|, \tag{52}$$

with $\Delta A_t := \hat{A}_t - A,\ \Delta B_t := \hat{B}_t - B$*, then*

1. $\hat{K}_t$ *is stabilizing,* $\forall t \in \mathbb{Z}_+$*;*
2. *the following holds,:*

$$\begin{aligned}|\hat{P}_t - P^*| \leq \beta_p(|\hat{P}_0 - P^*|, t) + \gamma_1(\|a\|_\infty)\\ + \gamma_2(\|b\|_\infty) \leq \delta_1,\ \forall t \in \mathbb{Z}_+,\end{aligned} \tag{53}$$

   where $\beta_p(x,t) := \sigma^t x$*;* $\gamma_1(x) := \frac{\bar{p}_a}{1-\sigma}x$*;* $\gamma_2(x) := \frac{\bar{p}_b}{1-\sigma}x$ *with constants* $\bar{p}_a, \bar{p}_b > 0$*;*

3. *if* $\lim_{t\to\infty}|\Delta A_t| = 0$ *and* $\lim_{t\to\infty}|\Delta B_t| = 0$*, then* $\lim_{t\to\infty}|\hat{P}_t - P^*| = 0$*.*

To proceed with the proof, we first verify the conditions under which Theorem 6 holds.

From Theorem 6, if $\|a\|_\infty \leq \bar{a}_p$*,* $\|b\|_\infty \leq \bar{b}_p$ *and* $\hat{P}_0 = P^*$*, ensuring that the conditions of Theorem 6 hold, then we have* $|\hat{P}_t - P^*| \leq \delta_1,\ \forall t \in \mathbb{Z}_+$*. Moreover, this guarantees that* $\lim_{t\to\infty}|\hat{P}_t - P^*| \leq \delta_1$*. Now, consider fixed matrices* $\tilde{A}$ *and* $\tilde{B}$ *satisfying* $|\tilde{A} - A| \leq \bar{a}_p$ *and* $|\tilde{B} - B| \leq \bar{b}_p$*, then given an initial condition* $\hat{P}_0 = P^*$*, we conclude that* $\lim_{t\to\infty}|\hat{P}_t - P^*| = |P^*_{(\tilde{A},\tilde{B})} - P^*| \leq \delta_1$*, where* $P^*_{(\tilde{A},\tilde{B})}$ *is the optimal solution to* (6) *corresponding to* $(\tilde{A}, \tilde{B}, Q, R)$*. Thus, we can conclude, for any* $\tilde{A}$ *and* $\tilde{B}$ *satisfying* $|\tilde{A} - A| \leq \bar{a}_p$ *and* $|\tilde{B} - B| \leq \bar{b}_p$*, then* $P^*_{(\tilde{A},\tilde{B})} \in \mathcal{B}_{\delta_1}(P^*)$*.*

When the maximum estimation error $\overline{\Delta\theta}(\hat{\theta}_0, \bar{D}) \leq \min\{\bar{a}_p, \bar{b}_p\}$*, then we have*

$$\begin{aligned}\gamma_1(\|a\|_\infty) + \gamma_2(\|b\|_\infty) &= \frac{\bar{p}_a\|a\|_\infty + \bar{p}_b\|b\|_\infty}{1-\sigma}\\ &\leq \frac{(\bar{p}_a + \bar{p}_b)\|\Delta\theta\|_\infty}{1-\sigma}.\end{aligned} \tag{54}$$

Together with (24)*, when Assumption 4 holds and the initial policy* $\hat{K}_0$ *is selected as the solution to the* (6) *using* $(\hat{A}_0, \hat{B}_0, Q, R)$*, the conditions required by Theorem 5 are satisfied. Substituting* (54) *into* (53)*, we conclude* (35a)*. Now we turn to prove* (35b)*. If the data sequence* $\{d_t\}$ *is bounded, then we can directly use Theorem 3 to prove* (35b)*.*

For matrices, $|\cdot|_2$ denotes their induced-2 norm. Based on Theorem 6, $\hat{K}_t$ is stabilizing, for all $t \in \mathbb{Z}_+$. Then we can define:

$$\bar{K}_{\mathrm{cl}} := \sup_{\substack{|\hat{A}-A| \leq \bar{a}_p, \\ |\hat{B}-B| \leq \bar{b}_p, \\ P \in \mathcal{B}_{\delta_1}(P^*)}} |\hat{A} + \hat{B}(\hat{B}^\top P \hat{B} + R)^{-1} \hat{B}^\top P \hat{A}|_2 \tag{55}$$

and we have $\bar{K}_{\mathrm{cl}} \in [0, 1)$. The additional excitation term e_t satisfies $\|e_t\| \leq \bar{e},\ \forall\, t \in \mathbb{Z}_+$ (27). Additionally, we have

$$\begin{aligned} |\hat{K}_t| &= |(R + B_t^\top \hat{P}_t B_t)^{-1} \hat{B}_t^\top \hat{P}_t \hat{A}_t| \\ &\leq \underbrace{|R^{-1}|(|B| + \|\Delta\theta\|_\infty)(|P^*| + \delta_1)(|A| + \|\Delta\theta\|_\infty)}_{=:\bar{K}} \end{aligned} \tag{56}$$

Then we can introduce the following lemma, which shows the boundedness of x_t:

***Lemma 1 (Boundedness of state** x_t**)** Given the system (1) with noise satisfying 2 and with the control input $u_t = \hat{K}_t x_t + e_t$, where $\hat{K}_t$ is the stabilizing gain from ORLS+PI and e_t satisfies (27), the state of system (1) remains bounded:*

$$|x_t| \leq \max\left(\frac{|B|\bar{e} + \|w\|_\infty}{1 - \bar{K}_{\mathrm{cl}}}, |x_0|\right) =: \bar{x},\ \forall t \in \mathbb{Z}_+, \tag{57}$$

where $\bar{K}_{\mathrm{cl}}$ is defined in (55) and $\bar{e}$ is defined in (27).

Proof 5 (Proof of Lemma 1) *For the case $|x_t| \geq \frac{|B|\bar{e}+\|w\|_\infty}{1-\bar{K}_{\mathrm{cl}}}$,*

$$\begin{aligned} |x_{t+1}| &\leq |A + B\hat{K}_t|_2 |x_t| + |B|\bar{e} + \|w\|_\infty \\ &\leq \bar{K}_{\mathrm{cl}} \frac{|B|\bar{e} + \|w\|_\infty}{1 - \bar{K}_{\mathrm{cl}}} + |B|\bar{e} + \|w\|_\infty \\ &= \frac{|B|\bar{e} + \|w\|_\infty}{1 - \bar{K}_{\mathrm{cl}}}. \end{aligned}$$

Together with the upper bound on the initialization, we conclude the proof.

Further, we can also derive the bound of the data d_t:

***Lemma 2 (Boundedness of data** d_t**)** Given the system (1) with noise satisfying (2) and with the control input $u_t = \hat{K}_t x_t + e_t$ where $\hat{K}_t$ is the stabilizing gain from ORLS+PI and e_t satisfies (27), the data d_t, which is employed for RLS, is bounded:*

$$\begin{aligned} |d_t| = \left|\begin{bmatrix} x_t \\ u_t \end{bmatrix}\right| &\leq \left|\begin{bmatrix} I \\ \hat{K}_t \end{bmatrix}\right| |x_t| + \left|\begin{bmatrix} 0 \\ e_t \end{bmatrix}\right| \\ &\leq (1 + \bar{K})\bar{x} + \bar{e} =: \bar{D} \end{aligned} \tag{58}$$

where $\bar{K}$ is defined in (56) and $\bar{x}$ is defined in Lemma 1.

Using the upper bound of the data sequence $\{d_t\}$ and together with Theorem 3, we can conclude (35b).

References

[1] Zhong-Sheng Hou and Zhuo Wang. From model-based control to data-driven control: Survey, classification and perspective. *Information Sciences*, 235:3–35, 2013.

[2] A. Khaki-Sedigh. *An Introduction to Data-Driven Control Systems*. Wiley, 2023.

[3] Damoon Soudbakhsh, Anuradha M. Annaswamy, Yan Wang, Steven L. Brunton, Joseph Gaudio, Heather Hussain, Draguna Vrabie, Jan Drgona, and Dimitar Filev. Data-driven control: Theory and applications. In *2023 American Control Conference (ACC)*, 2023.

[4] Florian Dörfler. Data-driven control: Part two of two: Hot take: Why not go with models? *IEEE Control Systems Magazine*, 43(6):27–31, 2023.

[5] Timm Faulwasser, Ruchuan Ou, Guanru Pan, Philipp Schmitz, and Karl Worthmann. Behavioral theory for stochastic systems? a data-driven journey from willems to wiener and back again. *Annual Reviews in Control*, 55:92–117, 2023.

[6] Julian Berberich and Frank Allgöwer. An overview of systems-theoretic guarantees in data-driven model predictive control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2024.

[7] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20, 10 2017.

[8] Mina Ferizbegovic, Jack Umenberger, Håkan Hjalmarsson, and Thomas B. Schön. Learning robust LQ-controllers using application oriented exploration. *IEEE Control Systems Letters*, 4(1):19–24, 2020.

[9] Nicolas Chatzikiriakos, Robin Strässer, Frank Allgöwer, and Andrea Iannelli. End-to-end guarantees for indirect data-driven control of bilinear systems with finite stochastic data. arXiv preprint arXiv:2409.18010, 2024.

[10] Anastasios Tsiamis, Ingvar M Ziemann, Manfred Morari, Nikolai Matni, and George J. Pappas. Learning to control linear systems can be hard. In *Proceedings of Thirty Fifth Conference on Learning Theory*, 2022.

[11] L. Ljung. *System Identification: Theory for the User*. Prentice Hall information and system sciences series. Prentice Hall PTR, 1999.

[12] K.J. Åström and B. Wittenmark. *Adaptive Control*. Dover Books on Electrical Engineering. Dover Publications, 2008.

[13] Anuradha M. Annaswamy. Adaptive control and intersections with reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 6(Volume 6, 2023):65–93, 2023.

[14] Frank L Lewis, Draguna Vrabie, and Vassilis L Syrmos. *Optimal control*. John Wiley & Sons, 2012.

[15] D. Bertsekas. *Abstract Dynamic Programming: 3rd Edition*. Athena Scientific., 2022.

[16] Youngsuk Park, Ryan A. Rossi, Zheng Wen, Gang Wu, and Handong Zhao. Structured policy iteration for linear quadratic regulator. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[17] Donghwan Lee. Convergence of dynamic programming on the semidefinite cone for discrete-time infinite-horizon LQR. *IEEE Transactions on Automatic Control*, 67(10):5661–5668, 2022.

[18] Bo Pang, Tao Bian, and Zhong-Ping Jiang. Robust policy iteration for continuous-time linear quadratic regulation. *IEEE Transactions on Automatic Control*, 67(1):504–511, 2022.

[19] Dimitri Bertsekas. Newton's method for reinforcement learning and model predictive control. *Results in Control and Optimization*, 7:100121, 2022.

[20] Bowen Song, Chenxuan Wu, and Andrea Iannelli. Convergence and robustness of value and policy iteration for the linear quadratic regulator. In *2025 European Control Conference (ECC)*, 2025, available as Preprint arXiv:2411.04548.

[21] Tae Yoon Chun, Jae Young Lee, Jin Bae Park, and Yoon Ho Choi. Stability and monotone convergence of generalised policy iteration for discrete-time linear quadratic regulations. *International Journal of Control*, 89(3):437–450, 2016.

[22] Farnaz Adib Yaghmaie, Fredrik Gustafsson, and Lennart Ljung. Linear quadratic control using model-free reinforcement learning. *IEEE Transactions on Automatic Control*, 68(2):737–752, Feb 2023.

[23] Lorenzo Sforni, Guido Carnevale, Ivano Notarnicola, and Giuseppe Notarstefano. On-policy data-driven linear quadratic regulator via combined policy iteration and recursive least squares. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, 2023.

[24] Bowen Song and Andrea Iannelli. The role of identification in data-driven policy iteration: A system theoretic study. *International Journal of Robust and Nonlinear Control*, 2024.

[25] Asaf Cassel, Alon Cohen, and Tomer Koren. Logarithmic regret for learning linear quadratic regulators efficiently. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1328–1337. PMLR, 13–18 Jul 2020.

[26] Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 1–26, Budapest, Hungary, 09–11 Jun 2011. PMLR.

[27] Max Simchowitz and Dylan Foster. Naive exploration is optimal for online LQR. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8937–8948. PMLR, 13–18 Jul 2020.

[28] Marco Borghesi, Alessandro Bosso, and Giuseppe Notarstefano. On-policy data-driven linear quadratic regulator via model reference adaptive reinforcement learning. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 32–37, 2023.

[29] Florian Dörfler, Zhiyu He, Giuseppe Belgioioso, Saverio Bolognani, John Lygeros, and Michael Muehlebach. Toward a systems theory of algorithms. *IEEE Control Systems Letters*, 8:1198–1210, 2024.

[30] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1467–1476. PMLR, 10–15 Jul 2018.

[31] Shokichi Takakura and Kazuhiro Sato. Structured output feedback control for linear quadratic regulator using policy gradient method. *IEEE Transactions on Automatic Control*, 69(1):363–370, 2024.

[32] Guido Carnevale, Nicola Mimmo, and Giuseppe Notarstefano. Data-driven LQR with finite-time experiments via extremum-seeking policy iteration. arXiv preprint arXiv:2412.02758, 2024.

[33] Feiran Zhao, Florian Dörfler, Alessandro Chiuso, and Keyou You. Data-enabled policy optimization for direct adaptive learning of the LQR. arXiv preprint arXiv:2401.14871, 2024.

[34] Adam L. Bruce, Ankit Goel, and Dennis S. Bernstein. Convergence and consistency of recursive least squares with variable-rate forgetting. *Automatica*, 119:109052, 2020.

[35] Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Proceedings of the Thirty-Second Conference on Learning Theory*, 2019.

[36] Yifan Xie, Julian Berberich, and Frank Allgöwer. Data-driven min-max mpc for linear systems. In *2024 American Control Conference (ACC)*, 2024.

[37] Janani Venkatasubramanian, Johannes Köhler, Mark Cannon, and Frank Allgöwer. Towards targeted exploration for non-stochastic disturbances. *20th IFAC Symposium on System Identification SYSID*, 2024.

[38] G. Hewer. An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Transactions on Automatic Control*, 16(4):382–384, 1971.

[39] K. B. Petersen and M. S. Pedersen. The matrix cookbook. Technical University of Denmark, October 2008. Version 20081110.

[40] Eduardo D. Sontag. *Input to State Stability: Basic Concepts and Results*, pages 163–220. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[41] Zhong-Ping Jiang and Yuan Wang. Input-to-state stability for discrete-time nonlinear systems. *Automatica*, 37(6):857–869, 2001.

[42] Andrea Iannelli and Roy S. Smith. A multiobjective LQR synthesis approach to dual control for uncertain plants. *IEEE Control Systems Letters*, 4(4):952–957, 2020.