



Logarithmic regret in online linear quadratic control using Riccati updates

Mohammad Akbari¹ · Bahman Gharesifard² · Tamas Linder¹

Received: 6 January 2021 / Accepted: 6 March 2022

© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract

An online policy learning problem of linear control systems is studied. In this problem, the control system is known and linear, and a sequence of quadratic cost functions is revealed to the controller in hindsight, and the controller updates its policy to achieve a sublinear regret, similar to online optimization. A modified online Riccati algorithm is introduced that under some boundedness assumption leads to logarithmic regret bound. In particular, the logarithmic regret for the scalar case is achieved without boundedness assumption. Our algorithm, while achieving a better regret bound, also has reduced complexity compared to earlier algorithms which rely on solving semi-definite programs at each stage.

Keywords Online linear quadratic control · Online optimal control · Riccati equation · Newton–Hewer Riccati update · Logarithmic regret

1 Introduction

Decision making based on predictions is a cornerstone of engineering, economy, and social sciences. Examples, including portfolio selection [2, 3], transportation and traf-

Research supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada. An incomplete version of this paper was presented at the 2nd Conference on Learning for Dynamics and Control (L4DC) and has appeared as [1].

✉ Mohammad Akbari
13mav1@queensu.ca

Bahman Gharesifard
gharesifard@ucla.ca

Tamas Linder
tamas.linder@queensu.ca

¹ Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada

² Department of Electrical and Computer Engineering, University of California, Los Angeles, CA 90095-1594, USA

fic control [4], power engineering [5], manufacturing and supply chain management, show promising achievements and have received considerable attention in recent years, see, e.g., [6–8]. The problem setting that we address in this paper belongs to a class of decision making problems known as *online optimization*. The literature on this subject is extremely rich and its connections to many other areas of learning have been explored in recent years [8–14].

Unlike the general setting of online optimization, where the decisions of the learner are solely chosen according to a time-varying cost function, in many realistic scenarios the learner's decisions involve a *control system*, where the decisions not only incur a cost at the present time, but also change the system state and incur a cost at the future. Examples include power supply management in the presence of time-varying energy costs due to demand fluctuations and tracking of an adversarial target. In this problem setting, decisions are assumed to be a function of the current state which is referred to as a *policy*. In order to assess the performance of policies over time, a regret is defined as the difference between the accumulated costs incurred by the control actions made in hindsight using previous states and the cost incurred by the best fixed admissible policy when all the cost functions are known in advance. Similar to online optimization, the objective is to design algorithms to generate policies which make the regret function grow sublinearly. Clearly if the cost functions were available to the decision maker, the problem discussed above would reduce to the classical optimal control problem.

The problem setting here is similar to the one studied in [15] where an online version of linear quadratic Gaussian control is introduced. In particular, in [15] an online gradient descent algorithm with a fixed learning rate is proposed, where in each iteration, a projection onto a bounded set of positive-definite matrices is taken, which itself relies on solving a semi-definite program. Under the assumptions that the underlying system is controllable, the cost functions are bounded, and the covariance of the disturbance is positive-definite, it is proved that the regret is sublinear, and grows as $\mathcal{O}(\sqrt{T})$, where T is the time horizon. Other closely related works are [16] and [17], where the cost functions are assumed to be general convex and globally Lipschitz functions. In contrast to [15], the noise assumed in [16] is adversarial, and [17] achieves a regret bound of $\mathcal{O}((\log(T))^7)$. In these works, the generated control actions, which lead to a sublinear regret bound, are linear feedbacks which rely on a finite history of the past disturbances. Similarly, in another recent work [18] a fixed (known) system with adversarial disturbances and fixed (known) quadratic cost functions is assumed, and a regret bound of $\mathcal{O}((\log(T))^3)$ is achieved.

Here, we point out a wider set of literature related to our work. First, we note that one can think about the underlying control system as a dynamical constraint on the optimization problem. Considering control systems as constraints is also classical in the context of *model predictive control* [19]. Although we tackle dynamic constraints in this work, we should emphasize that online optimization problems with static constraints, known only in hindsight, also play a key role in various settings and have generated interest in recent years [20–22].

Our work is also related to the framework of Markov decision processes (MDPs), where the system transition to the next state is defined through a probability distribution. Moreover, a reward is given to the decision maker for each action at each state.

This framework is classical in *reinforcement learning*, where the objective is to learn the optimal policy which yields the maximum reward [23]. It is also worth pointing out that there is another key role that regret minimization has played recently, bringing learning and control theory together, in the context of robust control, adaptive control, and system identification. Here, the regret enters through the lack of perfect knowledge of the model, and research efforts focus on generating algorithms for updating models in a data-driven fashion [24, 25]. Finally, our setting is also related to online optimization in dynamic environments [26], where the decisions are constrained in dynamics chosen by the environment. However, the objective of [26] is to study the impact of model mismatch on the overall regret, whereas in this paper the decisions are input to a control system, which impacts the way the decisions affect the future outcomes through its dynamics.

1.1 Contributions

We consider the problem of online linear quadratic Gaussian optimal control, where the control system is linear and known and the cost function is quadratic and time-varying and only becomes available in hindsight. In contrast to [15], where an online algorithm using semi-definite programming update is designed to generate the control policies, we employ a control-theoretic approach and introduce an online version of a classical iterative Riccati update, known as the Newton–Hewer [27] update. Using this update, which is less known than the classical Riccati difference equation [28], is key in developing our algorithm. This algorithm, which employs only a few matrix addition and multiplication operations in each time step, has reduced complexity compared to the one using semi-definite programming in each time step and is easier to implement. Our main result is a $\mathcal{O}(\log T)$ regret bound for the online linear quadratic Gaussian optimal control problem, improving the $\mathcal{O}(\sqrt{T})$ bound of [15] and the $\mathcal{O}((\log(T))^7)$ bound of [17] for time horizon T , under some boundedness assumption. Indeed, the technical part of our result relies on characterizing the interplay between a notion of stability for the sequence of control policies and boundedness of the solutions of the proposed Riccati update. In addition to our prior work [1], our observation shows that the latter boundedness property, which follows for the Riccati difference equation from monotonicity with respect to the underlying parameters, cannot be obtained using monotonicity; in fact, the Newton–Hewer updates can fail to be monotone in this setting [29]. This being said, for the scalar case, we are able to prove the boundedness property, yielding the stronger result that initializing the control policy to be stabilizing is enough to guarantee boundedness of the solutions of the proposed online Riccati update. We will demonstrate why the argument for the scalar case cannot be extended to the non-scalar case. We provide numerical results to illustrate this issue while adding different numerical results to show that the uniform boundedness assumption is not a strong assumption for our algorithm which achieves a sublinear regret in practical problems. In addition, we compare the regret of our algorithm with the regret of the Follow-the-Leader Algorithm through numerical examples.

1.2 Notation

We let $\mathbb{R}$ denote the set of real numbers and $\mathbb{R}^{n \times m}$ denote the set of $n \times m$ real matrices. We use lowercase letters for vectors and uppercase letters for matrices. We denote by $\|\cdot\|$ the Euclidean norm on vectors and its corresponding operator norm on real matrices. We denote by $A^\top$ the transpose of matrix A . Thus $\|A\| = \sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^\top A)}$, where $\sigma_{\max}(A)$ is the largest singular value of A and $\lambda_{\max}(A^\top A)$ is the largest eigenvalue of $A^\top A$. Trace of matrix A is denoted by $\text{Tr}(A)$. If A is an $n \times n$ real matrix with eigenvalues $\lambda_1, \dots, \lambda_n$, then the spectral radius of $\rho(A)$ of A is $\rho(A) = \max\{|\lambda_1|, \dots, |\lambda_n|\}$. We use $A \succeq B$ to indicate that $A - B$ is positive semi-definite.

2 Problem formulation

We start by describing the problem of online optimization for the class of linear control systems with quadratic cost. Let us recall this setting.

2.1 Discrete-time linear quadratic Gaussian control

The discrete-time linear quadratic Gaussian (LQG) control problem is defined as follows, see for instance [30]: Let $x_t \in \mathbb{R}^n$ and $u_t \in \mathbb{R}^m$ be the control state and the control action at time t , respectively, with initial state x_1 . The system dynamics are given by

$$x_{t+1} = Ax_t + Bu_t + w_t, \quad t \geq 1 \quad (2.1)$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, and $\{w_t\}_{t \geq 1}$ are i.i.d. Gaussian noise vectors with zero mean and covariance $W \in \mathbb{R}^{n \times n}$ ($w_t \sim \mathcal{N}(0, W)$). It is assumed that the initial value is Gaussian $x_1 \sim \mathcal{N}(m, X_1)$ and is independent of the noise sequence $\{w_t\}_{t \geq 1}$. The cost incurred in each time step t is a quadratic function of the state and control action given by $x_t^\top Q_t x_t + u_t^\top R_t u_t$, where $Q_t \in \mathbb{R}^{n \times n}$ and $R_t \in \mathbb{R}^{m \times m}$ are positive-definite matrices. The total cost after T time steps is given by

$$J_T(x_1, u_1, \dots, u_T) = \mathbb{E} \left[x_T^\top Q_T x_T + \sum_{t=1}^{T-1} (x_t^\top Q_t x_t + u_t^\top R_t u_t) \right].$$

We consider controllers of the form $u_t = \pi_t(x_t)$, where the function $\pi_t : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is called a policy. This assumption does not restrict generality, as the optimal policy will provably be of this form [30]. It is well known that under the assumption that the control system is stabilizable, and the cost matrices Q_t and R_t are positive-definite, the optimal policy is a stabilizing linear feedback of the state, which will be described next.

2.2 Discrete algebraic Riccati equation

In the classical LQG problem, where all the cost functions are known, the optimal policy can be obtained by dynamic programming, and is a linear function of the state. In particular, $u_t = -K_t x_t$, where K_t is given by the equation

$$K_t = (B^\top P_{t+1} B + R_t)^{-1} B^\top P_{t+1} A,$$

and P_{t+1} is a sequence of positive-definite matrices obtained iteratively, backwards in time, from the dynamic Riccati equation:

$$P_t = A^\top P_{t+1} A - A^\top P_{t+1} B (B^\top P_{t+1} B + R_t)^{-1} B^\top P_{t+1} A + Q_t \quad (2.2)$$

with the terminal condition $P_T = Q_T$.

For the infinite-horizon problem with the assumption that $Q_t = Q$ and $R_t = R$ are fixed, and under the assumptions that

- (i) R is positive-definite
- (ii) (A, B) is stabilizable, i.e., there exists a linear policy $\pi(x) = -Kx$ such that the closed-loop system $x_{t+1} = (A - BK)x_t$ is asymptotically stable: $\rho(A - BK) < 1$,
- (iii) (A, C) is detectable where $Q = C^\top C$, [i.e., if $u_t \rightarrow 0$ and $Cx_t \rightarrow 0$ then, $x_t \rightarrow 0$],

it is well known that the optimal policy is unique, time invariant, and is a linear function of the state [31], i.e., $u_t = -K^* x_t$. Here K^* is given by

$$K^* = (B^\top P^* B + R)^{-1} B^\top P^* A, \quad (2.3)$$

where P^* satisfies the discrete algebraic Riccati equation (DARE):

$$P^* = A^\top P^* A - A^\top P^* B (B^\top P^* B + R)^{-1} B^\top P^* A + Q. \quad (2.4)$$

Moreover, P_t given by (2.2) converges to P^* as $t \rightarrow \infty$ [30]. By using the policy K^* , we have that $x_{t+1} = (A - BK^*)x_t + w_t$. The optimal policy K^* is guaranteed to be stabilizing i.e. $\rho(A - BK^*) < 1$. Here, x_t converges to a stationary distribution, i.e., x_t converges weakly to a random variable x which has the same distribution as $(A - BK^*)x + w_t$, so that we have $\mathbb{E}[x] = \mathbb{E}[(A - BK^*)x + w_t]$, which implies $\mathbb{E}[x] = 0$, and the covariance matrix $X = \mathbb{E}[xx^\top]$ satisfies $X = (A - BK^*)X(A - BK^*)^\top + W$, see e.g., [15].

2.3 Problem setting

We now define the problem we study in this work, following [15]. In *online linear quadratic control*, the sequence of cost matrices $\{Q_t\}_{t \geq 1}$ and $\{R_t\}_{t \geq 1}$ are not known in advance and Q_t and R_t are only revealed after choosing the control action u_t . Since it is not possible to find the optimal policy before observing the whole sequence of cost

matrices $\{Q_t\}_{t \geq 1}$ and $\{R_t\}_{t \geq 1}$, the decision maker faces a *regret*. Here, we assume that the control system (A, B) is stabilizable, and the cost matrices Q_t and R_t are positive-definite and uniformly bounded over $t \geq 1$. As the optimal policy for the system with these assumptions is given by a stabilizing linear feedback, we use the set of stabilizing linear feedback functions as the set of admissible policies.

Let $x_t \in \mathbb{R}^n$ and $u_t \in \mathbb{R}^m$ be the control state and controller action at time $t \geq 1$. The controller uses a linear feedback policy $u_t = -K_t x_t$ and commits to this action after observing x_t . Then the controller receives the positive-definite matrices $Q_t \in \mathbb{R}^{n \times n}$ and $R_t \in \mathbb{R}^{m \times m}$, and suffers the cost

$$J_t(K_t) = \mathbb{E}[x_t^\top Q_t x_t + u_t^\top R_t u_t]. \quad (2.5)$$

The objective is to design an algorithm to generate a sequence of policies $\{K_t\}_{t \geq 1}$ such that the regret function, which is defined as

$$R(T) = \sum_{t=1}^T J_t(K_t) - \min_{K \in \mathcal{K}} \sum_{t=1}^T J_t(K), \quad (2.6)$$

where $\mathcal{K}$ is the set of stabilizing policies, grows sublinearly in T . In other words, the average regret over time converges to zero. Before stating our main results, we provide a brief review of the iterative Riccati updates that we employ to design our main algorithm.

2.4 Iterative methods for solving the discrete algebraic Riccati equation

Several methods for solving DARE exist in the literature, including iterative methods [28], algebraic methods [32], and semi-definite programming [33]. Our work is based on iterative methods, and in particular, two techniques that we review here. The first is given in [28], where one runs the recursion

$$P_{t+1} = A^\top P_t A - A^\top P_t B (B^\top P_t B + R)^{-1} B^\top P_t A + Q.$$

It is shown that under the assumption that (A, B) is stabilizable and (A, C) is detectable, where $Q = C^\top C$, the sequence $\{P_t\}$ converges to the unique solution of DARE.

A second approach, studied by [27], uses the following idea: Let P_t be the solution of the equation

$$P_t = (A - BK_t)^\top P_t (A - BK_t) + K_t^\top R K_t + Q, \quad (2.7)$$

where

$$K_t = (B^\top P_{t-1} B + R)^{-1} B^\top P_{t-1} A,$$

starting from a stabilizing policy K_1 . Then under the assumption that (A, B) is stabilizable and (A, C) is detectable, where $Q = C^\top C$, the sequence $\{P_t\}$ converges to the solution of DARE and the rate of convergence is quadratic, i.e.,

$$\|P_t - P^*\| \leq C \|P_{t-1} - P^*\|^2$$

where $C > 0$ is a constant. In what follows, we modify this algorithm and use it for the online linear quadratic Gaussian problem. We present our algorithm after reviewing some salient properties of stabilizing policies. Similar to [15], we use the notion of *strong stability*, which allows us to analyze the rate of convergence of the state covariance matrices under our proposed algorithm.

3 Strong stability

A key property that we require before introducing our algorithm is the notion of strong stability and sequential strong stability which are similar to the ones in [15]. The notion of strong stability is defined as follows.

Definition 1 A policy K is called stabilizing if $\rho(A - BK) < 1$. A policy K is (κ, γ) -strongly stabilizing (for $\kappa > 0$ and $0 < \gamma \leq 1$) if $\|K\| \leq \kappa$, and there exist matrices L and H such that $A - BK = HLH^{-1}$, with $\|L\| \leq 1 - \gamma$ and $\|H\| \|H^{-1}\| \leq \kappa$.

Note that every (κ, γ) -strongly stabilizing policy K is stabilizing, since the matrices $A - BK$ and L are similar and hence $\rho(A - BK) = \rho(L) \leq (1 - \gamma)$. Lemma 1 shows that every stabilizing policy is (κ, γ) -strongly stabilizing for some $\kappa > 0$ and $0 < \gamma \leq 1$.

Lemma 1 [15, Lemma B.1.] Suppose that for a linear system defined by A, B , a policy K is stabilizing. Then there are parameters $\kappa > 0$, $0 < \gamma \leq 1$ for which it is (κ, γ) -strongly stabilizing.

We refer the reader to [15, Lemma B.1] for a proof of this lemma.

Under the assumption of (κ, γ) -strong stability of policy K , the state covariance matrices $X_t = \mathbb{E}[x_t x_t^\top]$ converge exponentially to a steady-state covariance matrix $\hat{X}$, which satisfies

$$\hat{X} = (A - BK)\hat{X}(A - BK)^\top + W.$$

Lemma 2 provides the details.

Lemma 2 [15, Lemma 3.2] Let the pair (A, B) be stabilizable, and assume the controller uses a fixed (κ, γ) -strongly stabilizing policy K , i.e., for $t \geq 1$, we have $u_t = -Kx_t$. Let X_t be the covariance matrix of x_t . Then the sequence $\{X_t\}_{t \geq 1}$ converges to the steady-state covariance matrix $\hat{X}$, and in particular, for any $t \geq 1$,

$$\|X_{t+1} - \hat{X}\| \leq \kappa^2 e^{-2\gamma t} \|X_1 - \hat{X}\|.$$

We refer the reader to [15, Lemma 3.2] for a proof.

In order to obtain a similar result for the change of the state covariance matrices using a sequence of different (κ, γ) -strongly stabilizing policies $\{K_t\}_{t \geq 1}$, we need to define a notion of sequential strong stability, which is presented next.

Definition 2 A sequence of policies $\{K_t\}_{t \geq 1}$ is sequentially (κ, γ) -strongly stabilizing, for $\kappa > 0$ and $0 < \gamma \leq 1$, if there exist sequences of matrices $\{H_t\}_{t \geq 1}$ and $\{L_t\}_{t \geq 1}$ such that

$$A - BK_t = H_t L_t H_t^{-1}$$

for all $t \geq 1$, with the following properties:

- $\|L_t\| \leq 1 - \gamma$ and $\|K_t\| \leq \kappa$;
- $\|H_t\| \leq \beta$ and $\|H_t^{-1}\| \leq 1/\alpha$ with $\kappa = \beta/\alpha$ and $\alpha > 0$ and $\beta > 0$;
- $\|H_{t+1}^{-1} H_t\| \leq 1 + \gamma$.

The importance of this notion of stability is demonstrated in Lemma 3.

Lemma 3 Let the pair (A, B) be stabilizable, and suppose that the controller uses $u_t = -K_t x_t$ for $t \geq 1$ and where $\{K_t\}_{t \geq 1}$ is sequentially (κ, γ) -strongly stabilizing with $\kappa > 0$ and $0 < \gamma \leq 1$. For each K_t , let $\hat{X}_t$ be the corresponding steady-state covariance matrix, i.e., $\hat{X}_t$ satisfies $\hat{X}_t = (A - BK_t)\hat{X}_t(A - BK_t)^\top + W$ and assume that $\|\hat{X}_{t+1} - \hat{X}_t\| \leq \eta_t$ with $\eta_t > 0$, for all $t \geq 1$. Let X_t be the corresponding state covariance matrix at time t , starting from some initial $X_1 \geq 0$. Then for $t \geq 1$,

$$\|X_{t+1} - \hat{X}_{t+1}\| \leq \kappa^2 e^{-2\gamma^2 t} \|X_1 - \hat{X}_1\| + \kappa^2 \sum_{s=0}^{t-1} e^{-2\gamma^2 s} \eta_{t-s}.$$

The proof is similar to [15, Lemma 3.5], but we include it for completeness.

Proof By definition, for all $t \geq 1$, we have that

$$\begin{aligned} X_{t+1} &= (A - BK_t)X_t(A - BK_t)^\top + W, \\ \hat{X}_t &= (A - BK_t)\hat{X}_t(A - BK_t)^\top + W. \end{aligned}$$

Subtracting the equations, substituting $A - BK_t = H_t L_t H_t^{-1}$ and rearranging yields

$$H_t^{-1}(X_{t+1} - \hat{X}_t)(H_t^{-1})^\top = L_t H_t^{-1}(X_t - \hat{X}_t)(H_t^{-1})^\top L_t^\top.$$

Let $\Delta_t = H_t^{-1}(X_t - \hat{X}_t)(H_t^{-1})^\top$ for all $t \geq 1$. Then the above can be written as

$$\begin{aligned} \Delta_{t+1} &= (H_{t+1}^{-1} H_t L_t) \Delta_t (H_{t+1}^{-1} H_t L_t)^\top \\ &\quad + (H_{t+1}^{-1})(\hat{X}_t - \hat{X}_{t+1})(H_{t+1}^{-1})^\top. \end{aligned}$$

Taking the norms yields

$$\|\Delta_{t+1}\| \leq \|L_t\|^2 \|H_{t+1}^{-1} H_t\|^2 \|\Delta_t\| + \|H_{t+1}^{-1}\|^2 \|\hat{X}_t - \hat{X}_{t+1}\|$$

$$\begin{aligned} &\leq (1 - \gamma)^2(1 + \gamma)^2 \|\Delta_t\| + \frac{\eta_t}{\alpha^2} \\ &\leq (1 - \gamma^2)^2 \|\Delta_t\| + \frac{\eta_t}{\alpha^2}, \end{aligned}$$

and by unfolding the recursion, we obtain

$$\begin{aligned} \|\Delta_{t+1}\| &\leq (1 - \gamma^2)^{2t} \|\Delta_1\| + \frac{1}{\alpha^2} \sum_{s=0}^{t-1} (1 - \gamma^2)^{2s} \eta_{t-s} \\ &\leq e^{-2\gamma^2 t} \|\Delta_1\| + \frac{1}{\alpha^2} \sum_{s=0}^{t-1} e^{-2\gamma^2 s} \eta_{t-s}. \end{aligned}$$

Using $X_t - \hat{X}_t = H_t \Delta_t H_t^\top$ now, we have that

$$\begin{aligned} \|X_{t+1} - \hat{X}_{t+1}\| &\leq e^{-2\gamma^2 t} \|\Delta_1\| \|H_{t+1}\|^2 + \frac{\|H_{t+1}\|^2}{\alpha^2} \sum_{s=0}^{t-1} e^{-2\gamma^2 s} \eta_{t-s} \\ &\leq \kappa^2 e^{-2\gamma^2 t} \|X_1 - \hat{X}_1\| + \kappa^2 \sum_{s=0}^{t-1} e^{-2\gamma^2 s} \eta_{t-s}, \end{aligned}$$

which concludes the proof. $\square$

We now proceed with some key results that we later use to ensure strong stability for the sequence of policies generated. Suppose that a sequence of positive-definite matrices P_t is generated recursively as

$$P_t = (A - BK_t)^\top P_t (A - BK_t) + \bar{Q}_t + K_t^\top \bar{R}_t K_t, \quad (3.1)$$

where

$$K_{t+1} = (B^\top P_t B + \bar{R}_t)^{-1} B^\top P_t A \quad (3.2)$$

and where $\bar{R}_t \in \mathbb{R}^{m \times m}$ and $\bar{Q}_t \in \mathbb{R}^{n \times n}$ are given positive-definite matrices for all $t \geq 1$, and K_1 is an initial stabilizing policy. The reason for this update will become clear as part of our algorithm in Sect. 4. The key point we wish to make here is that under the assumption of uniform boundedness of the matrix sequence $\{P_t\}_{t \geq 1}$, and the stability of matrix K_t , for all $t \geq 1$, the sequence $\{K_t\}_{t \geq 1}$ is uniformly (κ, γ) -strongly stabilizing, with appropriate choices of κ and γ .

Proposition 1 *Assume that for $t \geq 1$, $Q_t, R_t \geq \mu I$ and $P_t \leq \nu I$, where $\mu, \nu > 0$ and $\{P_t\}_{t \geq 1}$ is the sequence of matrices obtained as the solution of (3.1), and assume that the policy K_t given by (3.2) is stabilizing for all $t \geq 1$. Define $\bar{\kappa} = \sqrt{\frac{\nu}{\mu}}$. Then the sequence $\{K_t\}_{t \geq 1}$ is uniformly $(\bar{\kappa}, 1/2\bar{\kappa}^2)$ -strongly stabilizing.*

Proof By the assumption of stability and since $Q_t \succeq \mu I$, we have that

$$\begin{aligned} P_t &= (A - BK)^\top P_t (A - BK) + \bar{Q}_t + K^\top \bar{R}_t K \\ &\succeq (A - BK)^\top P_t (A - BK) + \mu I, \end{aligned} \quad (3.3)$$

where we have used the positive-definiteness of $K^\top \bar{R}_t K$. In particular, this means that $P_t \succeq \mu I$ for all t . On the other hand, assuming $P_t \preceq \nu I$, we have

$$\mu I \preceq P_t \preceq \nu I. \quad (3.4)$$

Given that P_t is positive-definite and nonsingular, we can define $L_t = P_t^{1/2}(A - BK)P_t^{-1/2}$. Multiplying (3.3) by $P_t^{-1/2}$ from both sides, we obtain $I \succeq L_t^\top L_t + \mu P_t^{-1} \succeq L_t^\top L_t + \bar{\kappa}^{-2}I$. Thus $L_t^\top L_t \preceq (1 - \bar{\kappa}^{-2})I$, so $\|L_t\| \leq \sqrt{1 - \bar{\kappa}^{-2}} \leq 1 - \bar{\kappa}^{-2}/2$. Also, using (3.4) we have that

$$\|P_t^{1/2}\| \|P_t^{-1/2}\| \leq \bar{\kappa},$$

which finishes the proof. $\square$

We now present a second useful result, where we show that under the additional property that the rate of changes of sequence P_t is small (which we will be able to establish for our proposed algorithm, see Lemma 5), one can obtain that the sequence $\{K_t\}_{t \geq 1}$ is sequentially strongly stabilizing.

Proposition 2 Assume that for $t \geq 1$, $Q_t, R_t \succeq \mu I$ and $P_t \preceq \nu I$, where $\mu, \nu > 0$ and $\{P_t\}_{t \geq 1}$ is the sequence of matrices obtained as the solution of (3.1), and assume that the policy K_t given by (3.2) is stabilizing for $t \geq 1$. Let $\bar{\kappa} = \sqrt{\frac{\nu}{\mu}}$, and suppose that $\|P_{t+1} - P_t\| \leq \eta$ for $t \geq 1$ for some $\eta \leq \mu/\bar{\kappa}^2$. Then the sequence $\{K_t\}_{t \geq 1}$ is sequentially $(\bar{\kappa}, 1/2\bar{\kappa}^2)$ -strongly stabilizing.

Proof Proceeding as in the proof of Proposition 1, one can show that the matrix $L_t = P_t^{1/2}(A - BK_t)P_t^{-1/2}$ satisfies $\|L_t\| \leq 1 - 1/2\bar{\kappa}^2$ with $\|P_t^{1/2}\| \leq \sqrt{\nu}$ and $\|P_t^{-1/2}\| \leq 1/\sqrt{\mu}$. To establish the sequential strong stability stated by Definition 2 it thus suffices to show that $\|P_{t+1}^{-1/2}P_t^{1/2}\| \leq 1 + 1/2\bar{\kappa}^2$ for $t \geq 1$. To this end, observe that $\|P_{t+1} - P_t\| \leq \eta$, and that

$$\begin{aligned} \|P_{t+1}^{-1/2}P_t^{1/2}\|^2 &= \|P_{t+1}^{-1/2}P_tP_{t+1}^{-1/2}\| \\ &\leq \|P_{t+1}^{-1/2}P_{t+1}P_{t+1}^{-1/2}\| + \|P_{t+1}^{-1/2}(P_{t+1} - P_t)P_{t+1}^{-1/2}\| \\ &\leq 1 + \|P_{t+1}^{-1/2}\|^2\|P_{t+1} - P_t\| \\ &\leq 1 + \frac{\eta}{\mu}, \end{aligned}$$

where the second inequality follow by the sub-multiplicative of matrix operator norm. Hence, since $\eta \leq \mu/\bar{\kappa}^2$, then $\|P_{t+1}^{-1/2}P_t^{1/2}\| \leq \sqrt{1 + 1/\bar{\kappa}^2} \leq 1 + 1/2\bar{\kappa}^2$ as required. $\square$

The above results rely on uniform boundedness of the sequence $\{P_t\}_{t \geq 1}$, which we assume throughout the paper. However, we can show that stability of K_1 is enough to guarantee this property in the scalar case, see Proposition A.1 in the Appendix. Based on our extensive simulation studies, one of which is shown in Example A.3, we believe that this property should hold only by assuming stability of K_1 for the general case. One of the main reason for the difficulty of establishing this result is the lack of monotonicity of the evolutions of the Newton–Hewer dynamics with respect to the underlying system parameter, a sharp contrast with the Riccati difference updates [28], which we have recently reported in [29]. In this sense, the proof of Proposition A.1 for the scalar case establishes the boundedness property of the sequence $\{P_t\}_{t \geq 1}$ without relying on monotonicity. We have outlined further details in Remark A.2.

4 The online Riccati algorithm

We outline our main algorithm in this section. Our assumptions are as follows:

Assumption 3 Throughout we assume that

- The pair (A, B) is stabilizable.
- The cost matrices Q_t and R_t are positive-definite and $\mu I \preceq Q_t$, $\mu I \preceq R_t$, and $\text{Tr}(Q_t) \leq \sigma$, $\text{Tr}(R_t) \leq \sigma$, for some $\sigma > \mu > 0$ for all $t \geq 1$.
- For the noise covariance matrix W we have that $\omega = \text{Tr}(W) < \infty$.

Algorithm 1 Online Riccati Update

Input: The system matrices A and B , initial state x_1 , time horizon T , parameters $\nu, \mu, \kappa = \sqrt{\nu/\mu}$, $\gamma = 1/(2\kappa^2)$, σ

Output: A sequence of stabilizing policies $\{K_t\}_{t=1}^T$

- 1: **Initialize** K_1 to be stabilizing
- 2: **for** each $t = 1, 2, \dots, T$:
- 3: receive x_t
- 4: use controller $u_t = -K_t x_t$ and receive Q_t and R_t
- 5: update $\bar{R}_t = \frac{t-1}{t} \bar{R}_{t-1} + \frac{1}{t} R_t$, $\bar{Q}_t = \frac{t-1}{t} \bar{Q}_{t-1} + \frac{1}{t} Q_t$
- 6: update $\hat{P}_t$ as the solution of

$$P_t = (A - BK_t)^\top P_t (A - BK_t) + \bar{Q}_t + K_t^\top \bar{R}_t K_t$$

7: **Reset:**

8: **if** $t = t^* := \lceil \frac{4\kappa^3 \|B\|}{\gamma\mu} \left(2\sigma\kappa + \frac{2\kappa^3 \|B\| \sigma (1+\kappa^2)}{\gamma} \right) + 1 \rceil$:

9: Initialize $\ell = 0$, $\hat{P}_0 = P_{t^*}$, and $\hat{K}_0 = K_{t^*}$

10: **while** $\|\hat{P}_\ell - \hat{P}_{\ell-1}\| > \left(\frac{2\sigma}{\|B\|} + \frac{4\kappa^2 \sigma (1+\kappa^2)}{\gamma} \right) / t^*$:

11: $\ell \leftarrow \ell + 1$

12: $\hat{K}_\ell = (B^\top \hat{P}_{\ell-1} B + \bar{R}_{t^*})^{-1} B^\top \hat{P}_{\ell-1} A$

13: $\hat{P}_\ell$ satisfies $\hat{P}_\ell = (A - B\hat{K}_\ell)^\top \hat{P}_\ell (A - B\hat{K}_\ell) + \bar{Q}_{t^*} + \hat{K}_\ell^\top \bar{R}_{t^*} \hat{K}_\ell$

14: **return** $P_{t^*} = \hat{P}_\ell$

15: **return** $K_{t+1} = (B^\top P_t B + \bar{R}_t)^{-1} B^\top P_t A$

A formal description is given in Algorithm 1. We provide an informal description. We start from a stabilizing policy K_1 ; the existence of K_1 is provided by the assumption of stabilizability of the control system. At each time step $t \geq 1$, the controller uses the policy $u_t = -K_t x_t$ after observing x_t , then the cost matrices Q_t and R_t are revealed, and the controller updates P_t and K_t using the average of the history of Q_t s and R_t s through (2.7). There is a technical step in our algorithm, which we call the “reset” step and describe in detail later in the proof; this step allows us to show that using these updates the change of the norm of the policies is $\mathcal{O}(1/t)$, and this gives a regret bound $\mathcal{O}(\log(T))$. Before we state the algorithm, we need to elaborate on the parameters used.

Remark 1 (Parameters used in Algorithm 1) Our algorithm naturally uses parameters μ and σ , stated in Assumption 3. For the reset step, we also need (an estimate on) the strong stability parameters κ and γ , which are defined in Algorithm 1. Proposition 1 plays a key role in that regard, as it states that as long as we can estimate a uniform bound on the sequence P_t , we can obtain these parameters. In the scalar case, we know this uniform bound by Proposition A.1; in other cases, given that the parameters are not needed in the early steps of the algorithm, one can envision that we can run our algorithm with a large estimate on this bound and adjust it if necessary. Extending Proposition A.1 to vector cases, which is an avenue of our current research, will remove this restriction all together.

5 Main results

We are now in a position to state our main contribution, providing a logarithmic bound for the regret (2.6). We have opted not to explicitly display the bound as part of the statement; this can be found in (5.34).

Theorem 1 *Suppose that the tuple $(A, B, \{Q_t\}_{t=1}^T, \{R_t\}_{t=1}^T, W)$ satisfies Assumption 3. Suppose that the matrices P_t generated by Algorithm 1 are uniformly bounded. Then we have that*

$$\mathcal{R}(T) = \mathcal{O}(\log(T)).$$

The rest of this section is devoted to proving Theorem 1. The proof is quite involved, and for this reason we find it useful to provide a brief description to help the reader navigate through the proof. Our first technical result Lemma 4 shows that Algorithm 1, as long as it is initialized at an stabilizing policy, iteratively produces stabilizing policies. This step is analogous to the classical result of [27] for the case where the cost objective matrices Q_t and R_t are fixed. Recall that, by Proposition 1, stability of policies K_t is required to establish strong stability. A technical part of this proof demonstrates the reason why we need the reset step of the algorithm to ensure that the sequence of policies $\{P_{t+1} - P_t\}$ decay as m/t , for some $m > 0$. Using this and by rewriting the regret using trace products, we establish a set of bounds in Lemmas 6, 7, and 8 which eventually yield the result.

Proof of Theorem 1:

We first provide a straightforward reformulation of the regret function. For matrices A and B of appropriate size, let $A \bullet B = \text{Tr}(A^\top B)$. Then

$$\begin{aligned} R(T) &= \sum_{t=1}^T \mathbb{E} \left[x_t^\top Q_t x_t + u_t^\top R_t u_t \right] - \sum_{t=1}^T \mathbb{E} \left[x_t^{\dagger\top} Q_t x_t^\dagger + x_t^{\dagger\top} K^{\dagger\top} R_t K^\dagger x_t^\dagger \right] \\ &= \sum_{t=1}^T (Q_t + K_t^\top R_t K_t) \bullet X_t - \sum_{t=1}^T (Q_t + K^{\dagger\top} R_t K^\dagger) \bullet X_t^\dagger \end{aligned} \quad (5.1)$$

As a result, we have that

$$R(T) = \sum_{t=1}^T (Q_t + K_t^\top R_t K_t) \bullet (X_t - \widehat{X}_t) \quad (5.2)$$

$$+ \sum_{t=1}^T (Q_t + K_t^\top R_t K_t) \bullet \widehat{X}_t - \sum_{t=1}^T (Q_t + K^{\star\top} R_t K^\star) \bullet \widehat{X}^\star \quad (5.3)$$

$$+ \sum_{t=1}^T (Q_t + K^{\star\top} R_t K^\star) \bullet \widehat{X}^\star - \sum_{t=1}^T (Q_t + K^{\dagger\top} R_t K^\dagger) \bullet \widehat{X}^\dagger \quad (5.4)$$

$$+ \sum_{t=1}^T (Q_t + K^{\dagger\top} R_t K^\dagger) \bullet (\widehat{X}^\dagger - X_t^\dagger), \quad (5.5)$$

where $K^\dagger$ is the fixed optimal policy for the system $(A, B, \{Q_t\}_{t=1}^T, \{R_t\}_{t=1}^T, W)$, $X_t = \mathbb{E}[x_t x_t^\top]$ is the covariance matrix of x_t when the system follows policies K_t generated by Algorithm 1, $\widehat{X}_t$ is the steady-state covariance matrix using the policy K_t , i.e. $\widehat{X}_t$ satisfies

$$\widehat{X}_t = (A - BK_t) \widehat{X}_t (A - BK_t)^\top + W,$$

and

$$X_t^\dagger = \mathbb{E}[x_t^\dagger x_t^{\dagger\top}]$$

is the covariance matrix of the state $x_t^\dagger$ at time t when the system uses policy $K^\dagger$ at each time t ; similarly, $\widehat{X}^\dagger$ is the steady-state covariance matrix using the policy $K^\dagger$, i.e., $\widehat{X}^\dagger$ satisfies

$$\widehat{X}^\dagger = (A - BK^\dagger) \widehat{X}^\dagger (A - BK^\dagger)^\top + W. \quad (5.6)$$

$K^\star$ is the solution to DARE and $\widehat{X}^\star$ is the steady-state covariance matrix using policy $K^\star$. From now on, we use the notation $A_t = A - BK_t$ to simplify the presentation.

Note that by the following computation we show that (5.4) is negative. Since $\widehat{X}^\star$ and $\widehat{X}^\dagger$ are fixed, we have that

$$\begin{aligned}
 & \sum_{t=1}^T (Q_t + K^{\star\top} R_t K^\star) \bullet \widehat{X}^\star - \sum_{t=1}^T (Q_t + K^{\dagger\top} R_t K^\dagger) \bullet \widehat{X}^\dagger \\
 &= T(\bar{Q}_T + K^{\star\top} \bar{R}_T K^\star) \bullet \widehat{X}^\star - T(\bar{Q}_T + K^{\dagger\top} \bar{R}_T K^\dagger) \bullet \widehat{X}^\dagger \\
 &= T(P^\star - A^{\star\top} P^\star A^\star) \bullet \widehat{X}^\star - T(P^\dagger - A^{\dagger\top} P^\dagger A^\dagger) \bullet \widehat{X}^\dagger \\
 &= T(P^\star \bullet \widehat{X}^\star - P^\star \bullet A^\star \widehat{X}^\star A^{\star\top}) - T(P^\dagger \bullet \widehat{X}^\dagger - P^\dagger \bullet A^\dagger \widehat{X}^\dagger A^{\dagger\top}) \\
 &= T(P^\star \bullet \widehat{X}^\star - P^\star \bullet (\widehat{X}^\star - W)) - T(P^\dagger \bullet \widehat{X}^\dagger - P^\dagger \bullet (\widehat{X}^\dagger - W)) \\
 &= T(P^\star - P^\dagger) \bullet W \leq 0,
 \end{aligned}$$

where $P^\star$ and $P^\dagger$ satisfies $P = (A - BK)^\top P(A - BK) + \bar{Q}_T + K^\top \bar{R}_T K$ for $K = K^\star$ and $K = K^\dagger$, respectively, and we have used this fact in the second equality, the cyclic property of the trace in the third equality, and (5.6) in the fourth equality. By [27, Theorem 1], $P^\star \preceq P^\dagger$ and we have the result.

We start with our first technical result, which shows that Algorithm 1 produces stabilizing policies. This step is similar to the classical result of [27] for the case where the cost objective matrices Q_t and R_t are fixed. Recall that stability of policies K_t is required to establish strong stability, see Proposition 1.

Lemma 4 *Suppose that the pair (A, B) is stabilizable and let the sequence $\{K_t\}_{t \geq 1}$ be generated by Algorithm 1, starting from a stabilizing policy K_1 . Then policy K_t remains stabilizing for all $t \geq 1$.*

Proof We proceed by an induction argument. First, since the system is stabilizable, there exists a stabilizing policy and hence we can choose K_1 to be stabilizing, i.e. such that $\rho(A - BK_1) < 1$. Assume now that K_t is stabilizing, for some $t \geq 1$. Then, using (3.1), P_t is uniquely determined by

$$P_t = \sum_{i=0}^{\infty} (A_t^\top)^i (\bar{Q}_t + K_t^\top \bar{R}_t K_t) A_t^i. \quad (5.7)$$

By a straightforward computation, we have that

$$\begin{aligned}
 & A_t^\top P_t A_t + K_t^\top \bar{R}_t K_t \\
 &= (A - BK_t)^\top P_t (A - BK_t) + K_t^\top \bar{R}_t K_t \\
 &= A^\top P_t A - K_t^\top B^\top P_t A - A^\top P_t B K_t + K_t^\top (B^\top P_t B + \bar{R}_t) K_t \\
 &= A^\top P_t A - K_t^\top (B^\top P_t B + \bar{R}_t) K_{t+1} - K_{t+1}^\top (B^\top P_t B + \bar{R}_t) K_t \\
 &\quad + K_t^\top (B^\top P_t B + \bar{R}_t) K_t \\
 &= A^\top P_t A + (K_{t+1} - K_t)^\top (B^\top P_t B + \bar{R}_t) (K_{t+1} - K_t) \\
 &\quad - K_{t+1}^\top (B^\top P_t B + \bar{R}_t) K_{t+1}
 \end{aligned}$$

$$\begin{aligned}
 &= A^\top P_t A + (K_{t+1} - K_t)^\top (B^\top P_t B + \bar{R}_t)(K_{t+1} - K_t) \\
 &\quad - K_{t+1}^\top B^\top P_t A - A^\top P_t B K_{t+1} + K_{t+1}^\top (B^\top P_t B + \bar{R}_t) K_{t+1} \\
 &= A_{t+1}^\top P_t A_{t+1} + K_{t+1}^\top \bar{R}_t K_{t+1} + (K_{t+1} - K_t)^\top (B^\top P_t B + \bar{R}_t)(K_{t+1} - K_t),
 \end{aligned}$$

where we have used $(B^\top P_t B + \bar{R}_t) K_{t+1} = B^\top P_t A$ in the third and fifth equalities. Therefore, using this and (3.1), we have that

$$P_t = A_{t+1}^\top P_t A_{t+1} + V, \quad (5.8)$$

where

$$V = K_{t+1}^\top \bar{R}_t K_{t+1} + (K_{t+1} - K_t)^\top (B^\top P_t B + \bar{R}_t)(K_{t+1} - K_t) + \bar{Q}_t.$$

As a result,

$$P_t = \sum_{i=0}^{\infty} (A_{t+1}^\top)^i (V) A_{t+1}^i, \quad (5.9)$$

It is easy to observe that V is positive-definite. Now, using (5.7), since K_t is stabilizing, the matrix P_t is finite. Using (5.9), and the fact that the left side of (5.9) is finite, we have that $\rho(A_{t+1}) < 1$, i.e., K_{t+1} is stabilizing, otherwise the sum on the right side of (5.9) will diverge. $\square$

In order to get a $\log(T)$ regret bound, we need to have bounds of order $\mathcal{O}(1/t)$ on $\|P_t - P_{t-1}\|$, $\|\hat{X}_t - \hat{X}_{t-1}\|$ and $\|K_t - K_{t-1}\|$. Also, recall that such bounds are essential for obtaining sequential strong stability using Proposition 2. The next lemma and its corollary serves this purpose.

Lemma 5 *Suppose that $\mu I \preceq Q_t$, R_t and $\text{Tr}(Q_t), \text{Tr}(R_t) \leq \sigma$. Let $\{P_t\}_{t \geq 1}$ and $\{K_t\}_{t \geq 1}$ be the sequences of matrices generated by Algorithm 1, and assume that the sequence $\{K_t\}_{t \geq 1}$ is (κ, γ) -strongly stabilizing. Then we have $\|P_{t+1} - P_t\| \leq m/t$ for some $m > 0$, for $t \geq 1$.*

Proof Note that using (5.8), we have

$$\begin{aligned}
 P_{t+1} - P_t &= A_{t+1}^\top (P_{t+1} - P_t) A_{t+1} + K_{t+1}^\top (\bar{R}_{t+1} - \bar{R}_t) K_{t+1} + (\bar{Q}_{t+1} - \bar{Q}_t) \\
 &\quad - (K_{t+1} - K_t)^\top (B^\top P_t B + \bar{R}_t)(K_{t+1} - K_t).
 \end{aligned} \quad (5.10)$$

By the definition of K_t , we have the following identity:

$$K_{t+1} - K_t = (B^\top P_t B + \bar{R}_t)^{-1} [B^\top (P_t - P_{t-1}) A_t + (\bar{R}_{t-1} - \bar{R}_t) K_t]. \quad (5.11)$$

Using this along with (5.10), we have that

$$P_{t+1} - P_t = A_{t+1}^\top (P_{t+1} - P_t) A_{t+1} + K_{t+1}^\top (\bar{R}_{t+1} - \bar{R}_t) K_{t+1} + (\bar{Q}_{t+1} - \bar{Q}_t)$$

$$- [B^\top (P_t - P_{t-1})A_t + (\bar{R}_{t-1} - \bar{R}_t)K_t]^\top (B^\top P_t B + \bar{R}_t)^{-1} \quad (5.12)$$

$$\times [B^\top (P_t - P_{t-1})A_t + (\bar{R}_{t-1} - \bar{R}_t)K_t]. \quad (5.13)$$

By the stability of K_{t+1} , we have that

$$\begin{aligned} P_{t+1} - P_t &= \sum_{i=0}^{\infty} (A_{t+1}^\top)^i M_t A_{t+1}^i \\ &\leq \|M_t\| \sum_{i=0}^{\infty} (A_{t+1}^\top)^i A_{t+1}^i, \end{aligned} \quad (5.14)$$

where

$$\begin{aligned} M_t &= K_{t+1}^\top (\bar{R}_{t+1} - \bar{R}_t) K_{t+1} + (\bar{Q}_{t+1} - \bar{Q}_t) \\ &\quad - [B^\top (P_t - P_{t-1})A_t + (\bar{R}_{t-1} - \bar{R}_t)K_t]^\top (B^\top P_t B + \bar{R}_t)^{-1} \\ &\quad \times [B^\top (P_t - P_{t-1})A_t + (\bar{R}_{t-1} - \bar{R}_t)K_t]. \end{aligned}$$

Given the strong stability of K_{t+1} , we can write $A_{t+1} = H_{t+1} L_{t+1} H_{t+1}^{-1}$. Hence, we have that

$$\begin{aligned} \left\| \sum_{i=0}^{\infty} (A_{t+1}^\top)^i A_{t+1}^i \right\| &\leq \sum_{i=0}^{\infty} \|(A_{t+1}^\top)^i A_{t+1}^i\| \\ &\leq \sum_{i=0}^{\infty} \|H_{t+1}\|^2 \|H_{t+1}^{-1}\|^2 \|L_{t+1}\|^{2i} \\ &\leq \sum_{i=0}^{\infty} \kappa^2 (1 - \gamma)^{2i} = \frac{\kappa^2}{1 - (1 - \gamma)^2} \leq \frac{\kappa^2}{\gamma}, \end{aligned}$$

where we used $\|H_{t+1}\| \|H_{t+1}^{-1}\| \leq \kappa$ and $\|L_{t+1}\| \leq 1 - \gamma$. We now proceed to bound M_t . We can write

$$\begin{aligned} \|M_t\| &= \|K_{t+1}^\top (\bar{R}_{t+1} - \bar{R}_t) K_{t+1} + (\bar{Q}_{t+1} - \bar{Q}_t)\| \\ &\quad + \|(B^\top P_t B + \bar{R}_t)^{-1} \|(\|B\| \|A_t\| \|P_t - P_{t-1}\| + \|(\bar{R}_{t-1} - \bar{R}_t)K_t\|)\|^2. \end{aligned} \quad (5.15)$$

Using (5.14) and (5.15), we also have

$$z_{t+1} \leq c_t (h_t z_t + d_t)^2 + e_{t+1}, \quad (5.16)$$

where $z_t = \|P_t - P_{t-1}\|$, and

$$c_t = \frac{\kappa^2}{\gamma} \|(B^\top P_t B + \bar{R}_t)^{-1}\|$$

$$\begin{aligned}d_t &= \|(\bar{R}_{t-1} - \bar{R}_t)K_t\| \\e_{t+1} &= \frac{\kappa^2}{\gamma} \|K_{t+1}^\top (\bar{R}_{t+1} - \bar{R}_t)K_{t+1} + (\bar{Q}_{t+1} - \bar{Q}_t)\| \\h_t &= \|B\| \|A_t\|.\end{aligned}$$

Using the fact that

$$\|\bar{Q}_{t+1} - \bar{Q}_t\| = \frac{1}{t+1} \|(Q_t - \bar{Q}_t)\| \leq \frac{2}{t+1} \max_{t \geq 0} \|Q_t\| \leq \frac{2\sigma}{t+1},$$

along with

$$\|\bar{R}_{t+1} - \bar{R}_t\| = \frac{1}{t+1} \|(R_t - \bar{R}_t)\| \leq \frac{2}{t+1} \max_{t \geq 0} \|R_t\| \leq \frac{2\sigma}{t+1},$$

and

$$\|(B^\top P_t B + \bar{R}_t)^{-1}\| \leq (\lambda_{\min}(R_t))^{-1} \leq \mu^{-1},$$

and $\|A_t\| \leq \kappa$, we conclude

$$c_t \leq \kappa^2/\gamma\mu, \quad d_t \leq \frac{2\sigma\kappa}{t}, \quad \text{and} \quad e_t \leq \frac{2\kappa^2\sigma(1+\kappa^2)}{\gamma t}, \quad h_t \leq \|B\|\kappa, \quad (5.17)$$

for $t \geq 1$. We next claim that there exists a time t^* and a constant $m > 0$ such that $z_t \leq m/t$ for all $t > t^*$. We use an inductive argument to prove this statement. The base case will be proved later. Assume now that $z_t \leq m/t$; we show that $z_{t+1} \leq m/(t+1)$. First, note that if

$$m \leq \frac{2\sigma}{\|B\|} + \frac{4\kappa^2\sigma(1+\kappa^2)}{\gamma},$$

for $t \geq t^* = \frac{4\kappa^3\|B\|}{\gamma\mu}(2\sigma\kappa + \frac{2\kappa^3\|B\|\sigma(1+\kappa^2)}{\gamma}) + 1$, using an elementary calculation, one can observe that

$$\frac{\kappa^2}{\gamma\mu} \left(\kappa\|B\| \frac{m}{t} + \frac{2\sigma\kappa}{t} \right)^2 + \frac{2\kappa^2\sigma(1+\kappa^2)}{\gamma(t+1)} \leq \frac{m}{t+1}.$$

The claim then follows by noting that

$$z_{t+1} \leq c_t(h_t z_t + d_t)^2 + e_t \leq \frac{\kappa^2}{\gamma\mu} \left(\kappa\|B\| \frac{m}{t} + \frac{2\sigma\kappa}{t} \right)^2 + \frac{2\kappa^2\sigma(1+\kappa^2)}{\gamma(t+1)},$$

where we have used (5.17).

It remains to show that the condition we placed to obtain the last inequality, i.e., that $z_{t^*+1} \leq m/(t^* + 1)$, is satisfied. To proceed with this, first note that t^* is exactly the reset time in Algorithm 1. Also, the evolution of $\widehat{P}_\ell$ in the reset part of the algorithm is still according to (5.10). Since the matrices Q_t and R_t are fixed in the reset part, $\{\widehat{P}_\ell\}$ is a Cauchy sequence. Hence, by choosing ℓ large enough, we have that $\|\widehat{P}_\ell - \widehat{P}_{\ell-1}\| \leq m/t^*$, terminating the reset stage of the algorithm; with slight abuse of notation, we let $\widehat{P}_\ell$ be the outcome of the reset part of the algorithm. Note that at time t^* the algorithm implements $P_{t^*} = \widehat{P}_\ell$. In the next time step $t^* + 1$, the algorithm updates P_{t^*+1} as usual, using (3.1). We know by the previous part of the proof that $\|P_{t^*+1} - P_{t^*}\| \leq m/(t^* + 1)$, which shows that $z_{t^*+1} \leq m/(t^* + 1)$ is satisfied. To conclude the proof, note that we can show that $z_t \leq \widehat{m}/t$, for all $t \geq 1$, simply by selecting $\widehat{m} = \max\{m, tz_t | t \leq t^*\}$. $\square$

Corollary 1 *Let $\widehat{X}_t$ be the steady-state covariance matrix using policy K_t generated by Algorithm 1. Then we have $\|\widehat{X}_t - \widehat{X}_{t-1}\| \leq M/t + M'/t^2$ for some $M > 0$ and $M' > 0$ and for $t \geq 1$.*

Proof By the definition of $\widehat{X}_t$, we have that

$$\begin{aligned} \widehat{X}_t - \widehat{X}_{t-1} &= A_t \widehat{X}_t A_t^\top - A_{t-1} \widehat{X}_{t-1} A_{t-1}^\top \\ &= A_t (\widehat{X}_t - \widehat{X}_{t-1}) A_t^\top + (A_t - A_{t-1}) \widehat{X}_{t-1} (A_t - A_{t-1})^\top \\ &\quad + A_{t-1} \widehat{X}_{t-1} (A_t - A_{t-1})^\top + (A_t - A_{t-1}) \widehat{X}_{t-1} A_{t-1}^\top \\ &= A_t (\widehat{X}_t - \widehat{X}_{t-1}) A_t^\top + B(K_t - K_{t-1}) \widehat{X}_{t-1} (K_t - K_{t-1})^\top B^\top \\ &\quad + A_{t-1} \widehat{X}_{t-1} (K_{t-1} - K_t)^\top B^\top + B(K_{t-1} - K_t) \widehat{X}_{t-1} A_{t-1}^\top. \end{aligned}$$

Note that Lemma 5 can be used to bound $K_t - K_{t-1}$. Using (5.11), we have that

$$\begin{aligned} \|K_{t+1} - K_t\| &\leq \|(B^\top P_t B + \bar{R}_t)^{-1}\| [\|B\| \|P_t - P_{t-1}\| \|A_t\| + \|\bar{R}_{t-1} - \bar{R}_t\| \|K_t\|] \\ &\leq \frac{\kappa}{\mu} (\|B\| \widehat{m} + 2\sigma)/t, \end{aligned} \quad (5.18)$$

where we have used $\|(B^\top P_t B + \bar{R}_t)^{-1}\| \leq \mu^{-1}$, $\|A_t\| \leq \kappa$, $\|K_t\| \leq \kappa$, and $\widehat{m}$ is given in the proof of Lemma 5. Using this $\|\widehat{X}_t - \widehat{X}_{t+1}\|$ is bounded by $M/t + M'/t^2$, where

$$M = \frac{2\kappa^6 \omega}{\mu \gamma^2} \|B\| (\|B\| \widehat{m} + 2\sigma), \quad (5.19)$$

and

$$M' = \frac{\kappa^6 \omega}{\mu \gamma^2} \|B\|^2 (\|B\| \widehat{m} + 2\sigma)^2, \quad (5.20)$$

where we have used

$$\|\widehat{X}_{t-1}\| \leq \|W\| \sum_{i=0}^{\infty} \|(A_{t-1}^\top)^i (A_{t-1}^i)\| \leq \frac{\omega \kappa^2}{\gamma}$$

□

The following lemmas will be used to derive bounds on the redundancy terms (5.2), (5.3), and (5.5).

Lemma 6 Suppose that the tuple $(A, B, \{Q_t\}_{t=1}^T, \{R_t\}_{t=1}^T, W)$ satisfies Assumption 3. Suppose that the matrices P_t generated by Algorithm 1 are uniformly bounded, i.e., $P_t \leq \nu I$. Let $\kappa = \sqrt{\frac{\nu}{\mu}}$ and $\gamma = 1/2\kappa^2$. Then for the covariance matrices X_t and $\widehat{X}_t$, we have

$$\begin{aligned} \sum_{t=1}^T (Q_t + K_t^\top R_t K_t) \bullet (X_t - \widehat{X}_t) &\leq t^\star \sigma (1 + \kappa^2) \max_{0 < t \leq t^\star} \|X_t - \widehat{X}_t\| \\ &\quad + 2\kappa^4 \sigma \left(\|X_{t^\star} - \widehat{X}_{t^\star}\| \frac{e^{-2\gamma^2 t^\star}}{1 - e^{-2\gamma^2}} + \frac{M' \pi^2}{6(1 - e^{-2\gamma^2})} \right. \\ &\quad \left. + \frac{M}{1 - e^{-2\gamma^2}} \log \left(\frac{T}{t^\star} \right) \right). \end{aligned}$$

Proof For $t \geq t^\star$, we have that $\|P_{t+1} - P_t\| \leq m/t \leq \mu/\kappa^2$. Then, using Proposition 2, the matrices K_t are sequentially (κ, γ) -strongly stabilizing for $t \geq t^\star$ ($\kappa = \sqrt{\frac{\nu}{\mu}}$ and $\gamma = 1/(2\kappa^2)$). Using this by Lemma 3, we conclude that for $t \geq t^\star$

$$\|X_{t+1} - \widehat{X}_{t+1}\| \leq \kappa^2 e^{-2\gamma^2(t+1-t^\star)} \|X_{t^\star} - \widehat{X}_{t^\star}\| + \kappa^2 \sum_{s=0}^{t-t^\star} e^{-2\gamma^2 s} \eta_{t-s}; \quad (5.21)$$

hence we can separate (5.2) into two parts as follows:

$$\begin{aligned} \sum_{t=1}^T (Q_t + K_t^\top R_t K_t) \bullet (X_t - \widehat{X}_t) &= \sum_{t=1}^{t^\star} (Q_t + K_t^\top R_t K_t) \bullet (X_t - \widehat{X}_t) \\ &\quad + \sum_{t=t^\star}^T (Q_t + K_t^\top R_t K_t) \bullet (X_t - \widehat{X}_t). \end{aligned}$$

By stability of policies K_t , the matrices X_t and $\widehat{X}_t$ are bounded and we have that

$$\sum_{t=1}^{t^\star} (Q_t + K_t^\top R_t K_t) \bullet (X_t - \widehat{X}_t) = \sum_{t=1}^{t^\star} \text{Tr}[(Q_t + K_t^\top R_t K_t)(X_t - \widehat{X}_t)]$$

$$\begin{aligned}
 &\leq \sum_{t=1}^{t^*} \text{Tr}(Q_t + K_t^\top R_t K_t) \|(X_t - \hat{X}_t)\| \\
 &\leq t^* \sigma (1 + \kappa^2) \max_{0 < t \leq t^*} \|(X_t - \hat{X}_t)\|, \quad (5.22)
 \end{aligned}$$

where we have used

$$\text{Tr}(Q_t + K_t^\top R_t K_t) \leq \sigma(1 + \kappa^2)$$

Using (5.21), we have that

$$\begin{aligned}
 &\sum_{t=t^*}^T (Q_t + K_t^\top R_t K_t) \bullet (X_t - \hat{X}_t) \\
 &\leq \sum_{t=t^*}^T \text{Tr}(Q_t + K_t^\top R_t K_t) \|(X_t - \hat{X}_t)\| \\
 &\leq \sum_{t=t^*}^T \sigma(1 + \kappa^2) \|X_t - \hat{X}_t\| \\
 &\leq (\sigma(1 + \kappa^2)) \kappa^2 \sum_{t=t^*}^T \left(e^{-2\gamma^2 t} \|X_{t^*} - \hat{X}_{t^*}\| + \sum_{s=0}^{t-t^*} e^{-2\gamma^2 s} \eta_{t-s} \right) \\
 &\leq 2\kappa^4 \sigma (\|X_{t^*} - \hat{X}_{t^*}\| \frac{e^{-2\gamma^2 t^*}}{1 - e^{-2\gamma^2}} + \sum_{t=t^*}^T \sum_{s=0}^{t-t^*} e^{-2\gamma^2 s} \eta_{t-s}).
 \end{aligned}$$

Note that by using Corollary 1, we have $\eta_t = M/t + M'/t$, where M and M' are given by (5.19) and (5.20). Consequently,

$$\begin{aligned}
 \sum_{t=t^*}^T (Q_t + K_t^\top R_t K_t) \bullet (X_t - \hat{X}_t) &\leq 2\kappa^4 \sigma \|X_{t^*} - \hat{X}_{t^*}\| \frac{e^{-2\gamma^2 t^*}}{1 - e^{-2\gamma^2}} \\
 &\quad + 2\kappa^4 \sigma \sum_{t=t^*}^T \sum_{s=0}^{t-t^*} e^{-2\gamma^2 s} \left(\frac{M}{t-s} + \frac{M'}{(t-s)^2} \right). \quad (5.23)
 \end{aligned}$$

Next, by changing the order of summation we obtain

$$\begin{aligned}
 \sum_{t=t^*}^T \sum_{s=0}^{t-t^*} e^{-2\gamma^2 s} \left(\frac{M}{t-s} + \frac{M'}{(t-s)^2} \right) &= \sum_{s=0}^{T-t^*} e^{-2\gamma^2 s} \sum_{t=s+t^*}^T \left(\frac{M}{t-s} + \frac{M'}{(t-s)^2} \right) \\
 &\leq \sum_{s=0}^{T-t^*} e^{-2\gamma^2 s} \left(M \log \left(\frac{T-s}{t^*} \right) + \frac{M' \pi^2}{6} \right)
 \end{aligned}$$

$$\begin{aligned} &\leq \frac{M'\pi^2}{6(1-e^{-2\gamma^2})} + \sum_{s=0}^{T-t^*} M e^{-2\gamma^2 s} \log\left(\frac{T}{t^*}\right) \\ &\leq \frac{M'\pi^2}{6(1-e^{-2\gamma^2})} + \frac{M}{1-e^{-2\gamma^2}} \log\left(\frac{T}{t^*}\right), \end{aligned}$$

where we have used a logarithmic upper bound for $\sum_{t=t^*}^{T-s} 1/t$ and the identity $\sum_{t=1}^{\infty} 1/t^2 = \pi^2/6$ in the second inequality. The third and fourth inequalities follow by manipulating geometric series. Therefore, by substituting this inequality in Equation (5.23) we obtain

$$\begin{aligned} \sum_{t=t^*}^T (Q_t + K_t^\top R_t K_t) \bullet (X_t - \hat{X}_t) &\leq 2\kappa^4 \sigma \left(\|X_{t^*} - \hat{X}_{t^*}\| \frac{e^{-2\gamma^2 t^*}}{1-e^{-2\gamma^2}} + \frac{M'\pi^2}{6(1-e^{-2\gamma^2})} \right. \\ &\quad \left. + \frac{M}{1-e^{-2\gamma^2}} \log\left(\frac{T}{t^*}\right) \right) \end{aligned} \quad (5.24)$$

The result follow by adding (5.22) and (5.24). $\square$

Lemma 7 Suppose that the tuple $(A, B, \{Q_t\}_{t=1}^T, \{R_t\}_{t=1}^T, W)$ satisfies Assumption 3. Suppose that the matrices P_t generated by Algorithm 1 are uniformly bounded, i.e., $P_t \leq \nu I$. Let $\kappa = \sqrt{\frac{\nu}{\mu}}$ and $\gamma = 1/2\kappa^2$. Then the covariance matrices $\hat{X}_t$ and $\hat{X}^*$ satisfy

$$\begin{aligned} \sum_{t=1}^T (Q_t + K_t^\top R_t K_t) \bullet \hat{X}_t - \sum_{t=1}^T (Q_t + K^{\star\top} R_t K^*) \bullet \hat{X}^* &\leq \omega l \hat{m} \\ &\quad + \frac{\kappa^4 \omega}{\gamma \mu^3} (\|B\| \hat{m} + 2\sigma)^2 (1 + \log(T)). \end{aligned}$$

Proof Using the fact that $Q_t = t\bar{Q}_t - (t-1)\bar{Q}_{t-1}$ and $R_t = t\bar{R}_t - (t-1)\bar{R}_{t-1}$, we have

$$\begin{aligned} (Q_t + K_t^\top R_t K_t) &= (t\bar{Q}_t - (t-1)\bar{Q}_{t-1}) + K_t^\top (t\bar{R}_t - (t-1)\bar{R}_{t-1}) K_t \\ &= t(\bar{Q}_t + K_t^\top \bar{R}_t K_t) - (t-1)(\bar{Q}_{t-1} + K_t^\top \bar{R}_{t-1} K_t) \\ &= t(P_t - A_t^\top P_t A_t) - (t-1)(P_{t-1} - A_t^\top P_{t-1} A_t) \\ &\quad + (t-1)(K_t - K_{t-1})^\top (B^\top P_{t-1} B + \bar{R}_{t-1})^{-1} (K_t - K_{t-1}), \end{aligned} \quad (5.25)$$

where we have used (3.1) and (5.8) in the third equality. Note that

$$\begin{aligned} A_t^\top P_t A_t \bullet \hat{X}_t &= \text{Tr}(A_t^\top P_t A_t \hat{X}_t) \\ &= \text{Tr}(P_t A_t \hat{X}_t A_t^\top) \\ &= P_t \bullet A_t \hat{X}_t A_t^\top \end{aligned}$$

$$\begin{aligned}
 &= P_t \bullet (\widehat{X}_t - W) \\
 &= P_t \bullet \widehat{X}_t - P_t \bullet W.
 \end{aligned} \tag{5.26}$$

Therefore, by multiplying (5.25) and $\widehat{X}_t$ we obtain

$$\begin{aligned}
 (Q_t + K_t^\top R_t K_t) \bullet \widehat{X}_t &= t P_t \bullet W - (t-1) P_{t-1} \bullet W \\
 &\quad + (t-1)(K_t - K_{t-1})^\top (B^\top P_{t-1} B + \bar{R}_{t-1})^{-1} (K_t - K_{t-1}) \bullet \widehat{X}_t,
 \end{aligned}$$

where we have used (5.26) to cancel out some terms. Summing over t and using the telescopic series for $t P_t \bullet W - (t-1) P_{t-1} \bullet W$, we obtain

$$\begin{aligned}
 \sum_{t=1}^T (Q_t + K_t^\top R_t K_t) \bullet \widehat{X}_t &\leq T P_T \bullet W \\
 &\quad + \sum_{t=1}^T (t-1)(K_t - K_{t-1})^\top (B^\top P_{t-1} B + \bar{R}_{t-1})^{-1} (K_t - K_{t-1}) \bullet \widehat{X}_t.
 \end{aligned} \tag{5.27}$$

On the other hand,

$$\begin{aligned}
 \sum_{t=1}^T (Q_t + K^{\star\top} R_t K^*) \bullet \widehat{X}^* &= T(\bar{Q}_T + K^{\star\top} \bar{R}_T K^*) \bullet \widehat{X}^* \\
 &= T(P^* - A^{\star\top} P^* A^*) \bullet \widehat{X}^* \\
 &= T(P^* \bullet \widehat{X}^* - P^* \bullet A^* \widehat{X}^* A^{\star\top}) \\
 &= T(P^* \bullet \widehat{X}^* - P^* \bullet \widehat{X}^* + P^* \bullet W) \\
 &= T P^* \bullet W.
 \end{aligned} \tag{5.28}$$

Therefore, by subtracting (5.28) from (5.27) we have

$$\begin{aligned}
 \sum_{t=1}^T (Q_t + K_t^\top R_t K_t) \bullet \widehat{X}_t - \sum_{t=1}^T (Q_t + K^{\star\top} R_t K^*) \bullet \widehat{X}^* &= T(P_T - P^*) \bullet W + \\
 &\quad + \sum_{t=1}^T (t-1)(K_t - K_{t-1})^\top (B^\top P_{t-1} B + \bar{R}_{t-1})^{-1} (K_t - K_{t-1}) \bullet \widehat{X}_t.
 \end{aligned}$$

Note that P^* is the solution of DARE when the cost matrices $Q_t = \bar{Q}_T$ and $R_t = \bar{R}_T$ are chosen to be fixed; it is the limit of the sequence P_t when Q_t and R_t are chosen to be $\bar{Q}_T$ and $\bar{R}_T$, respectively. The rate of convergence is quadratic [27], i.e. there exists $C > 0$ such that for all $t \geq 2$,

$$\|P_t - P^*\| \leq C \|P_{t-1} - P^*\|^2 \tag{5.29}$$

and by a similar analysis, we also have

$$\|P_{t+1} - P_t\| \leq C\|P_t - P_{t-1}\|^2. \quad (5.30)$$

Here we use a similar technique to bound $\|P_T - P^*\|$. We can update the sequence P_t after time T by starting at P_T using (3.1), with $\bar{Q}_t = \bar{Q}_T$ and $\bar{R}_t = \bar{R}_T$ fixed for $t \geq T$. We hence have that

$$\begin{aligned} \|P_T - P^*\| &= \lim_{t \rightarrow \infty} \|P_T - P_t\| \\ &\leq \lim_{t \rightarrow \infty} \sum_{i=0}^{t-1} \|P_{T+i} - P_{T+i+1}\| \\ &\leq \lim_{t \rightarrow \infty} \sum_{i=0}^{t-1} C^{2^i-1} \|P_T - P_{T+1}\|^{2^i} \\ &= \|P_T - P_{T+1}\| \lim_{t \rightarrow \infty} \sum_{i=0}^{t-1} C^{2^i-1} \|P_T - P_{T+1}\|^{2^i-1}, \end{aligned}$$

where we have used (5.29). For $T \geq t^*$, $C\|P_T - P_{T+1}\| < 1$ and thus the sum $\sum_{i=0}^{\infty} C^{2^i-1} \|P_T - P_{T+1}\|^{2^i-1}$ is bounded by some finite value $l > 0$. Hence we have

$$T(P_T - P^*) \bullet W \leq T\omega\|P_T - P^*\| \leq T\omega l\|P_T - P_{T+1}\| \leq \frac{T\omega l \hat{m}}{T} = \omega l \hat{m}, \quad (5.31)$$

where we have used $\omega = \text{Tr}(W)$ and $\|P_T - P_{T+1}\| \leq \hat{m}/T$ by Lemma 5. We now proceed by noting that

$$\begin{aligned} &\sum_{t=2}^T (t-1)(K_t - K_{t-1})^\top (B^\top P_{t-1} B + \bar{R}_{t-1})^{-1} (K_t - K_{t-1}) \bullet \hat{X}_t \\ &\leq \sum_{t=2}^T (t-1) \text{Tr}(\hat{X}_t) \frac{1}{(t-1)^2 \mu^3} (\|B\| \kappa \hat{m} + 2\sigma \kappa)^2 \\ &\leq \frac{\kappa^2}{\gamma} \omega (\|B\| \kappa \hat{m} + 2\sigma \kappa)^2 \sum_{t=2}^T \frac{1}{(t-1) \mu^3} \\ &\leq \frac{\kappa^4 \omega}{\gamma \mu^3} (\|B\| \hat{m} + 2\sigma)^2 (1 + \log(T)), \end{aligned} \quad (5.32)$$

where we have used the bound in (5.18) on $\|K_t - K_{t-1}\|$, and the bound for $\text{Tr}(\hat{X}_t)$. Adding (5.32) and (5.31) completes the proof. $\square$

Lemma 8 Suppose that the tuple $(A, B, \{Q_t\}_{t=1}^T, \{R_t\}_{t=1}^T, W)$ satisfies Assumption 3. Suppose that the matrices P_t generated by Algorithm 1 are uniformly bounded, i.e.,

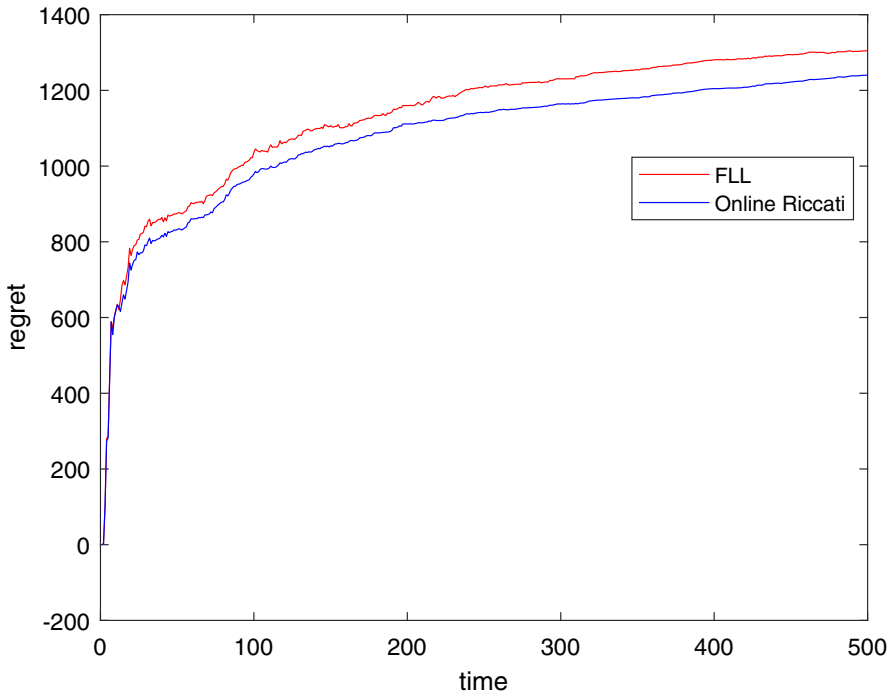


Fig. 1 The regret over time using the policies generated by Algorithm 1 and FLL algorithm

$P_t \leq vI$. Let $\kappa = \sqrt{\frac{v}{\mu}}$ and $\gamma = 1/2\kappa^2$. Then we have

$$\sum_{t=1}^T (Q_t + K^{\star\top} R_t K^{\star}) \bullet (\hat{X}^{\star} - X_t^{\star}) \leq \frac{\sigma(1 + \kappa^2)\kappa^2}{1 - e^{-2\gamma}} \|\hat{X}^{\star} - X_1^{\star}\|. \quad (5.33)$$

Proof $\|K^{\star}\| \leq \kappa$ implies that $\text{Tr}(Q_t + K^{\star\top} R_t K^{\star}) \leq \sigma(1 + \kappa^2)$. Moreover, by Lemma 2, we have that

$$\begin{aligned} \sum_{t=1}^T (Q_t + K^{\star\top} R_t K^{\star}) \bullet (\hat{X}^{\star} - X_t^{\star}) &\leq \sigma(1 + \kappa^2) \sum_{t=1}^T \|\hat{X}^{\star} - X_t^{\star}\| \\ &\leq \sigma(1 + \kappa^2) \sum_{t=1}^T \kappa^2 e^{-2\gamma(t-1)} \|\hat{X}^{\star} - X_1^{\star}\| \\ &\leq \sigma(1 + \kappa^2)\kappa^2 \frac{1}{1 - e^{-2\gamma}} \|\hat{X}^{\star} - X_1^{\star}\|, \end{aligned}$$

as claimed. $\square$

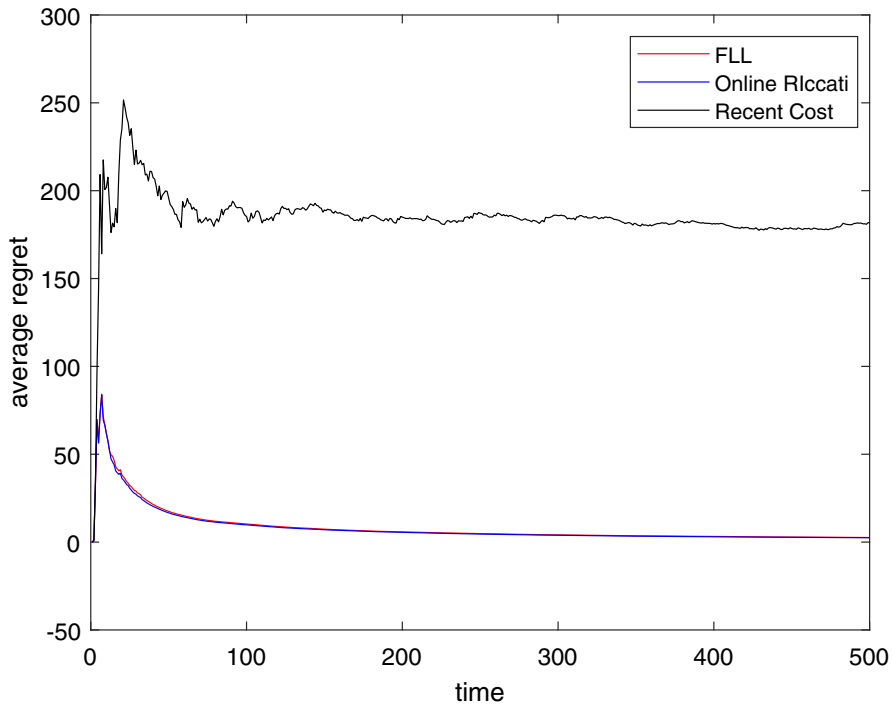


Fig. 2 The average regret over time using the policies generated by Algorithm 1, FLL algorithm and recent cost policy for the first experiment

To conclude, by summing the right hand side of (5.22), (5.24), (5.32), (5.31), and (5.33), we obtain the regret bound as follows,

$$\begin{aligned}
 R(T) \leq & \left(2\kappa^4\sigma \frac{M}{1-e^{-2\gamma^2}} + \frac{\kappa^4\omega}{\gamma\mu^3} (\|B\|\hat{m} + 2\sigma)^2 \right) \log(T) \\
 & - 2\kappa^4\sigma \frac{M}{1-e^{-2\gamma^2}} \log(t^*) + t^*\sigma(1+\kappa^2) \max_{0 < t \leq t^*} \|X_t - \hat{X}_t\| \\
 & + 2\kappa^4\sigma \left(\|X_{t^*} - \hat{X}_{t^*}\| \frac{e^{-2\gamma^2 t^*}}{1-e^{-2\gamma^2}} + \frac{M'\pi^2}{6(1-e^{-2\gamma^2})} \right) + \omega l \hat{m} \\
 & + \frac{\kappa^4\omega}{\gamma\mu^3} (\|B\|\hat{m} + 2\sigma)^2 + \frac{\sigma(1+\kappa^2)\kappa^2}{1-e^{-2\gamma}} \|\hat{X}^* - X_1^*\|, \quad (5.34)
 \end{aligned}$$

which finishes the proof of Theorem 1

Note that the assumption of (κ, γ) -strongly stability in Theorem 1 will be satisfied as long as the solutions to the online Riccati equation are uniformly bounded. In particular, we do not need this assumption for the scalar case, see Proposition A.1.

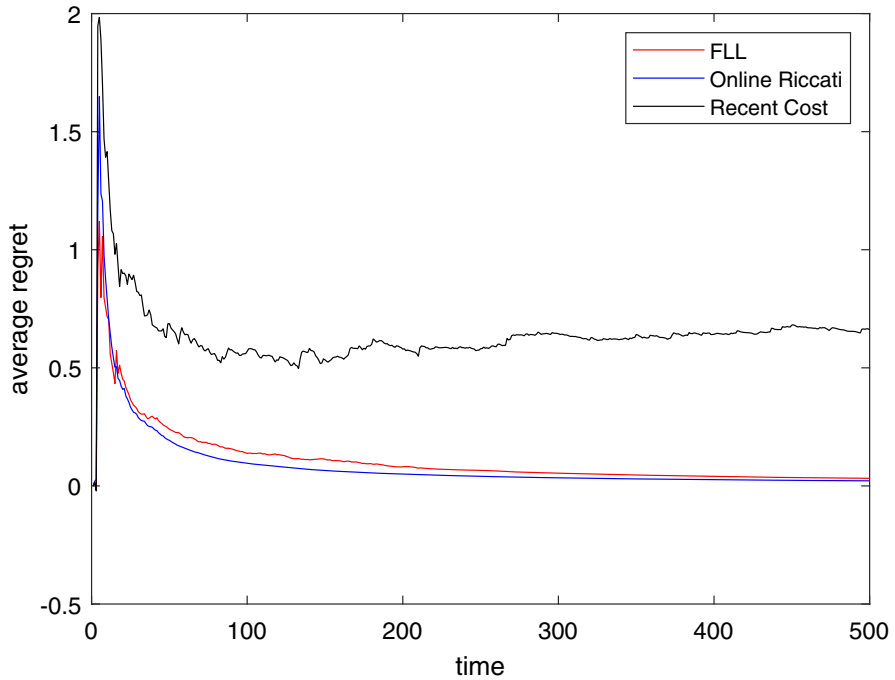


Fig. 3 The average regret over time using the policies generated by Algorithm 1, FLL algorithm and recent cost policy for the second experiment

6 Simulation results

We provide simulation results for the proposed algorithm to illustrate its performance. The control system dynamics are given by $x_{t+1} = Ax_t + Bu_t + w_t$, where the pair (A, B) is stabilizable, and $A \in \mathbb{R}^{10 \times 10}$ and $B \in \mathbb{R}^{10 \times 7}$, and w_t is a Gaussian noise. The matrices A and B are chosen randomly with entry-wise i.i.d uniform distribution on $[-3, 3]$ and $[-2, 2]$ respectively. We have considered three scenarios for the cost functions. For the first experiment, the matrices Q_t and R_t are generated randomly with the Wishart distribution with unit variance and 20 degrees of freedom. For the second and third experiment, we followed the experiment setting of [15], where $Q_t = Q$ is fixed as the identity matrix, while R_t is diagonal where some diagonal entries are 1, while others are r_t . For the second experiment, we assume that r_t is randomly changing over time with i.i.d uniform distribution on $[0.1, 1]$, and for the third experiment, we assume that r_t is changing over time according to a random walk restricted to $[0.1, 1]$ taking steps of size 0.1, -0.1 , 0, with probability 0.1, 0.1, 0.8, respectively. We ran the algorithm with the stabilizing matrix K_0 and $X_0 = 0$ to generate a sequence of matrices K_t , and we computed the regret and the average regret over time. We compared our results with the ones stated in [15].

Figure 1 shows the regret over time for the online Riccati algorithm and the follow the lazy leader (FLL) algorithm given in [15] for the first experiment. The results show

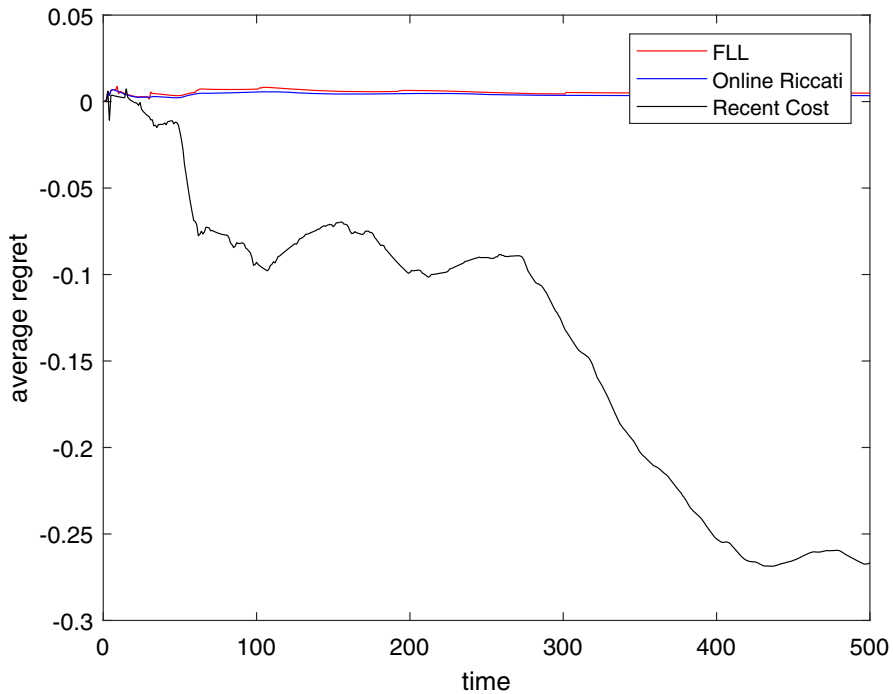


Fig. 4 The average regret over time using the policies generated by Algorithm 1, FLL algorithm and recent cost policy for the third experiment

that both algorithms behave similarly. Although [15] found a regret bound of $\mathcal{O}(\sqrt{T})$ while we have achieved a regret bound of $\mathcal{O}(\log T)$, this simulation result is expected, since FLL uses the average cost matrices Q_t and R_t over time and finds the optimal K_t and uses it for the next time step, and the online Riccati algorithm uses a Riccati update of the average cost matrices Q_t and R_t over time.

Figures 2, 3 and 4 show the average regret of the online Riccati algorithm, FLL algorithm, and the recent cost policy, where the optimal policy of recent cost matrices is used for the next time step. The graphs show that the online Riccati algorithm works well for different scenarios, and as expected the recent cost policy is not a good strategy and only works for the random walk scenario where the change in the cost function is slow, as also indicated in [15], and we have plotted these for comparison.

A Appendix

Proposition A.1 *Let $n = m = 1$ and let $\{P_t\}_{t=1}^T$ be a sequence of positive numbers generated by Equation (3.1) and (3.2) recursively, and assume that policy K_t is stabilizing for all $t \geq 1$. Then there exists $\nu > 0$ such that $P_t \leq \nu$ for all $t \geq 1$.*

Proof Note that

$$P_t = (A - BK_t)^2 P_t + \bar{Q}_t + K_t^2 \bar{R}_t,$$

Since K_t is stabilizing using the stability of K_1 , c.f. Lemma 4, we have that

$$P_t = \frac{\bar{Q}_t + K_t^2 \bar{R}_t}{1 - (A - BK_t)^2}.$$

Now if you consider P_t as a function of K_t , by taking derivative of P_t with respect to K_t and setting it to zero, we have that

$$K_t = \frac{-\bar{R}_t + A^2 \bar{R}_t - B^2 \bar{Q}_t + \sqrt{(\bar{R}_t - A^2 \bar{R}_t + B^2 \bar{Q}_t)^2 + 4A^2 B^2 \bar{R}_t \bar{Q}_t}}{2AB \bar{R}_t}$$

minimizes the P_t and the minimum admissible P_t which we denote by $\check{P}_t$ is given by

$$\check{P}_t = \frac{A^2 \bar{R}_t - \bar{R}_t + \bar{Q}_t B^2 + \sqrt{(\bar{R}_t - A^2 \bar{R}_t - \bar{Q}_t B^2)^2 + 4B^2 \bar{Q}_t \bar{R}_t}}{2B^2}.$$

Now if we write P_{t+1} as a function of P_t we have that

$$\begin{aligned} P_{t+1} &= \frac{\bar{Q}_{t+1} + K_{t+1}^2 \bar{R}_{t+1}}{1 - (A - BK_{t+1})^2} \\ &= \frac{\bar{Q}_{t+1} + ((B^2 P_t + \bar{R}_t)^{-1} B P_t A)^2 \bar{R}_{t+1}}{1 - (A \bar{R}_t (B^2 P_t + \bar{R}_t)^{-1})^2} \\ &= \frac{\bar{Q}_{t+1} (B^2 P_t + \bar{R}_t)^2 + B^2 P_t^2 A^2 \bar{R}_{t+1}}{(B^2 P_t + \bar{R}_t)^2 - A^2 \bar{R}_t^2}. \end{aligned}$$

By taking derivative of P_{t+1} with respect to P_t , we conclude that for the admissible P_t , i.e., $P_t \geq \check{P}_t$, the function P_{t+1} is decreasing for $P_t \leq \check{P}_t$ and increasing for $P_t \geq \check{P}_t$ [29], where $\check{P}_t$ is given by

$$\check{P}_t = \frac{(A^2 - 1) \bar{R}_{t+1} \bar{R}_t + B^2 \bar{Q}_{t+1} \bar{R}_t + \sqrt{((A^2 - 1) \bar{R}_{t+1} \bar{R}_t + B^2 \bar{Q}_{t+1} \bar{R}_t)^2 + 4B^2 \bar{Q}_{t+1} \bar{R}_{t+1} \bar{R}_t^2}}{2B^2 \bar{R}_{t+1}}.$$

Since P_{t+1} is decreasing for $P_t \leq \check{P}_t$ and increasing for $P_t \geq \check{P}_t$, its maximum is achieved on the boundary. So we will check the value of P_{t+1} for the point P_t at infinity and at its admissible minimum $\check{P}_t$. Now letting P_t goes to infinity, we have

$$P_{t+1} = \lim_{P_t \rightarrow \infty} \frac{\bar{Q}_{t+1} (B^2 P_t + \bar{R}_t)^2 + B^2 P_t^2 A^2 \bar{R}_{t+1}}{(B^2 P_t + \bar{R}_t)^2 - A^2 \bar{R}_t^2} = \frac{A^2}{B^2} \bar{R}_{t+1} + \bar{Q}_{t+1},$$

and for $P_t = \tilde{P}_t$, we have

$$P_{t+1} = \frac{\bar{Q}_{t+1}(B^2 \tilde{P}_t + \bar{R}_t)^2 + B^2 \tilde{P}_t^2 A^2 \bar{R}_{t+1}}{(B^2 \tilde{P}_t + \bar{R}_t)^2 - A^2 \bar{R}_t^2}$$

One can observe that P_{t+1} as a function of R_t has a similar behaviour. So for P_{t+1} to achieve its maximum, $(\tilde{P}_t, R_t)$ should be minimum and (Q_{t+1}, R_{t+1}) should be maximum. So if we let $Q_{\max} = \max\{\bar{Q}_1, \bar{Q}_2, \dots, \bar{Q}_T\}$, $Q_{\min} = \min\{\bar{Q}_1, \bar{Q}_2, \dots, \bar{Q}_T\}$, $R_{\max} = \max\{\bar{R}_1, \bar{R}_2, \dots, \bar{R}_T\}$, $R_{\min} = \min\{\bar{R}_1, \bar{R}_2, \dots, \bar{R}_T\}$, and

$$\tilde{P}_{\min} = \frac{A^2 R_{\min} - R_{\min} + Q_{\min} B^2 + \sqrt{(R_{\min} - A^2 R_{\min} - Q_{\min} B^2)^2 + 4 B^2 Q_{\min} R_{\min}}}{2 B^2},$$

we obtain that for all $t > 0$

$$P_t \leq \max \left\{ \frac{A^2}{B^2} R_{\max} + Q_{\max}, \frac{Q_{\max} (B^2 \tilde{P}_{\min} + R_{\min})^2 + B^2 \tilde{P}_{\min}^2 A^2 R_{\max}}{(B^2 \tilde{P}_{\min} + R_{\min})^2 - A^2 R_{\min}^2} \right\}$$

□

We illustrate in the next remark as to why the argument that we have used above cannot be readily extended to non-scalar cases.

Remark A.2 The procedure that we have used above to prove boundedness of P_t relied on studying the evolutions of P_{t+1} as a function of P_t . When these quantities are not scalars, one naturally aims to consider the norm of P_{t+1} as a function of the norm of P_t . However, an example can be constructed where P_{t+1} as a function of P_t becomes unbounded as P_t approaches the boundary of the set positive-definite matrices that make K_{t+1} unstabilizing. This does not happen in the scalar case since this boundary is smaller than $\tilde{P}_t$, the minimum achievable P_t . Figure 5 depicts the norm of P_{t+1} for different trials of selecting P_t . For each trial, the P_t is chosen as $P_t = P^* + \Omega$, where P^* is the minimum achievable P_t for a stabilizing matrix K_t , and Ω is a positive definite matrix. It can be seen that the norm of P_{t+1} for some trials gets very large. For example, for P_t

$$P_t = \begin{pmatrix} 18714 & -312 & 291 \\ -312 & 82149 & -144 \\ 291 & -144 & 14220 \end{pmatrix},$$

the matrix $A - BK_{t+1}$ has the eigenvalues

$$\lambda(A - BK_{t+1}) = \begin{pmatrix} -0.999996 \\ 0.002971 \\ -0.000047 \end{pmatrix},$$

and the first eigenvalue that is near 1, which makes the norm of P_{t+1} around the order of 7.7×10^8 . However, in several simulations of online Riccati algorithm, we observed

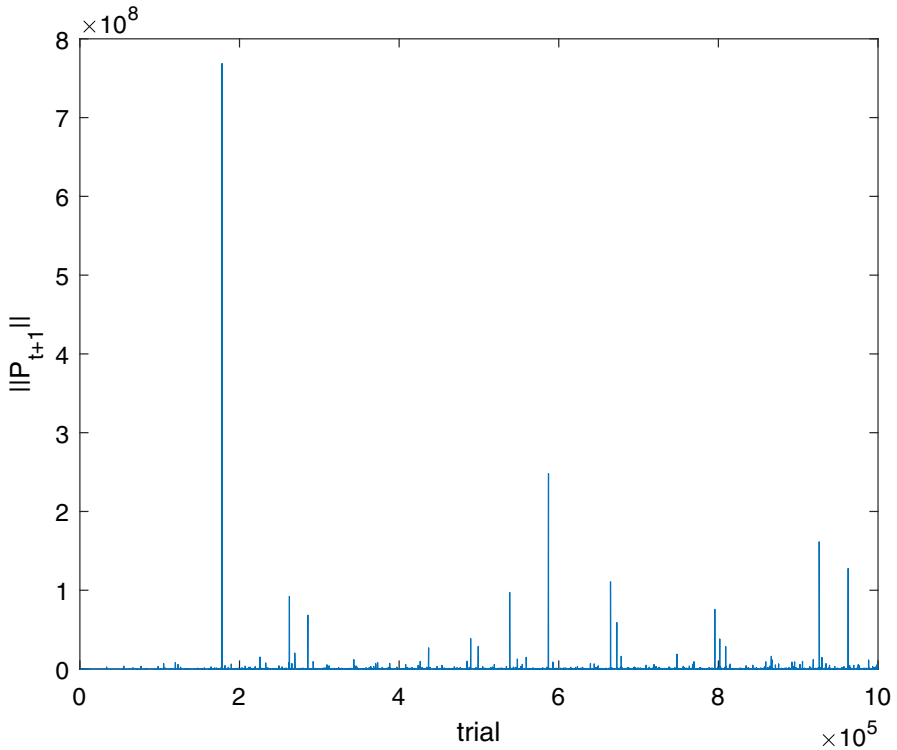


Fig. 5 This graph shows the norm of P_{t+1} for different values of $P_t \geq P^*$. P_t can be near the boundary that makes K_{t+1} unstabilizing, and hence P_{t+1} gets very large

that changes in P_t as a result of changes in bounded $\bar{Q}_t$ and $\bar{R}$ do not make K_{t+1} to get close to the unstabilizing policy boundary, and hence P_{t+1} cannot get unbounded. We will show this behaviour in the following experiment.

Example A.3 In order to observe the behaviour of matrices P_t over time, a linear discrete-time control system with $n = 7$ states and $m = 5$ control actions is considered, where the matrices (A, B) are fixed.

We used several trials, where for each trial a sequence of positive definite random matrices Q_t and R_t with Wishart distribution is generated and we used the online Riccati algorithm with different initialization K_1 to generate the sequence P_t . Figure 6 shows the graph of the norm of P_t over time for each trial. Clearly, P_t stays bounded. Similar property is observed in all our simulation studies. Understanding why this boundedness occurs and if this is generally true is an important open problem, and appears to be difficult in light of the previous remark.

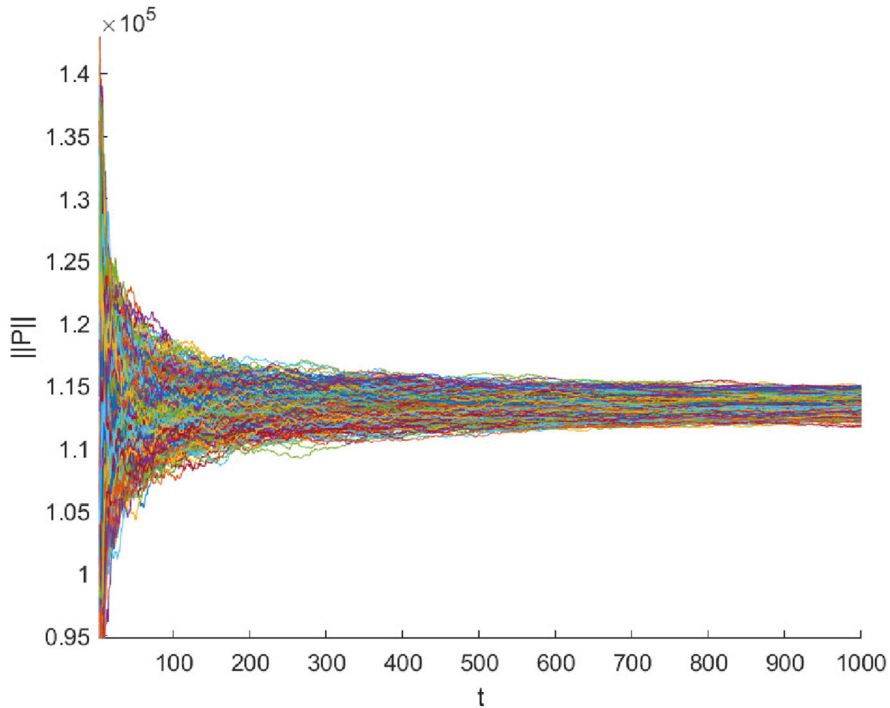


Fig. 6 The norm of P_t over time for 1000 trials is shown. For each trial, a sequence of matrices Q_t and R_t with Wishart distribution is generated and the sequence P_t is generated using the online Riccati algorithm

References

1. Akbari M, Ghahesifard B, Linder T (2020) “Riccati updates for online linear quadratic control,” In *proceedings of the 2nd conference on learning for dynamics and control*, vol. 120, pp. 476–485, PMLR
2. Agarwal A, Hazan E, Kale S, Schapire RE (2006) “Algorithms for portfolio management based on the Newton method,” In *proceedings of the 23rd international conference on machine learning, ICML '06*, pp. 9–16
3. Luo H, Wei C, Zheng K (2018) “Efficient online portfolio with logarithmic regret,” In *Advances in neural information processing systems 31*, pp. 8235–8245, Curran Associates, Inc
4. Patel M, Ranganathan N (2001) IDUTC: an intelligent decision-making system for urban traffic-control applications. *IEEE Trans Vehicular Technol* 50(3):816–829
5. Zhai J, Li Y, Chen H (2016) “An online optimization for dynamic power management,” In *2016 IEEE international conference on industrial technology (ICIT)*, pp. 1533–1538
6. Anava O, Hazan E, Mannor S, Shamir O (2013) “Online learning for time series prediction,” in *proceedings of the 26th annual conference on learning theory*, vol. 30, pp. 172–184
7. Ross S, Gordon G, Bagnell D (2011) “A reduction of imitation learning and structured prediction to no-regret online learning,” In *proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635
8. Cesa-Bianchi N, Lugosi G (2006) *Prediction, learning, and games*. Cambridge University Press, Cambridge
9. Hazan E (2016) Introduction to online convex optimization. *Foundation Trends Optim* 2(3–4):157–325
10. Shalev-Shwartz S (2012) *Online Learning and Online Convex Optimization*, vol. 12 of *Foundations and Trends in Machine Learning*. Now Publishers Inc

11. Hazan E, Agarwal A, Kale S (2007) Logarithmic regret algorithms for online convex optimization. *Mach Learn* 69(2–3):169–192
12. Hazan E, Kale S (2014) Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *J Mach Learn Res* 15(1):2489–2512
13. Gofer E, Cesa-Bianchi N, Gentile C, Mansour Y (2013) “Regret minimization for branching experts,” In *conference on learning theory*, pp. 618–638
14. Blum A, Mansour Y (2007) From external to internal regret. *J Mach Learn Res* 8:1307–1324
15. Cohen A, Hasidim A, Koren T, Lazic N, Mansour Y, Talwar K (2018) “Online linear quadratic control,” In *proceedings of the 35th international conference on machine learning*, vol. 80, pp. 1029–1038
16. Agarwal N, Bullins B, Hazan E, Kakade S, Singh K (2019) “Online control with adversarial disturbances,” In *proceedings of the 36th international conference on machine learning*, vol. 97, pp. 111–119
17. Agarwal N, Hazan E, Singh K (2019) “Logarithmic regret for online control,” <http://arxiv.org/abs/1909.05062>
18. Foster D, Simchowitz M (2020) “Logarithmic regret for adversarial online control,” In *proceedings of the 37th international conference on machine learning*, vol. 119, pp. 3211–3221, PMLR
19. Garcia CE, Prett DM, Morari M (1989) Model predictive control: theory and practice; a survey. *Automatica* 25(3):335–348
20. Yu H, Neely M, Wei X (2017) Online convex optimization with stochastic constraints. *Adv Neural Inf Process Syst* 30:1428–1438
21. Neely, MJ, Yu H (2017) “Online convex optimization with time-varying constraints,” <http://arxiv.org/abs/1702.04783>
22. Jenatton R, Huang J, Archambeau C (2016) “Adaptive algorithms for online convex optimization with long-term constraints,” In *proceedings of the 33rd international conference on machine learning*, vol. 48, pp. 402–411
23. Sutton RS, Barto AG (2018) Reinforcement learning: an introduction. The MIT Press, Cambridge
24. Yang Y, Guo Z, Xiong H, Ding D, Yin Y, Wunsch DC (2019) Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning. *IEEE Trans Neural Netw Learn Syst* 2:87
25. Karimi A, Kammer C (2017) A data-driven approach to robust control of multivariable systems by convex optimization. *Automatica* 85:227–233
26. Hall EC, Willett RM (2015) Online convex optimization in dynamic environments. *IEEE J Select Top Signal Process* 9(4):647–662
27. Hewer G (1971) An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Trans Autom Control* 16(4):382–384
28. Caines PE, Mayne DQ (1970) On the discrete time matrix Riccati equation of optimal control. *Int J Control* 12(5):785–794
29. Akbari M, Ghahsifard B, Linder T (2020) “On the lack of monotonicity of Newton-Hewer updates for Riccati equations,” <http://arxiv.org/abs/2010.15983>
30. Soderstrom T (2002) Discrete-time stochastic systems: estimation and control, 2nd edn. Springer-Verlag, New York
31. Bertsekas DP (2018) Stable optimal control and semicontractive dynamic programming. *SIAM J Control Optim* 56(1):231–252
32. Rodman L, Lancaster P (1995) *Algebraic Riccati Equations*. Oxford Mathematical Monographs
33. Balakrishnan V, Vandenberghe L (2003) Semidefinite programming duality and linear time-invariant systems. *IEEE Trans Autom Control* 48(1):30–41