An Iterative Riccati Algorithm for Online Linear Quadratic Control

Mohammad Akbari[1]
Bahman Gharesifard[2]
Tamas Linder[3]

Abstract

We study an online setting of the linear quadratic Gaussian optimal control problem on a sequence of cost functions, where similar to classical online optimization, the future decisions are made by only knowing the cost in hindsight. We introduce a modified online Riccati update that under some boundedness assumptions, leads to logarithmic regret bounds, improving the best known square-root bound. In particular, for the scalar case we achieve the logarithmic regret without any boundedness assumption. As opposed to earlier work, proposed method does not rely on solving semi-definite programs at each stage.

1 Introduction

The problem of prediction and decision making has many applications in engineering, economy and social sciences, for instance, portfolio selection [1, 19], transportation and traffic control [21], power engineering [29], manufacturing and supply chain management; and it has received substantial attention in recent years, see [3], [23], and [8]. The subject under study in this work sits within this general theme of a class of decision making problems, and particularly is related to online optimization. The literature on online optimization is extremely rich and its connections to many other areas of learning has been explored in recent years [8, 13, 24, 14, 15, 11, 6].

Unlike the classical setting of online optimization, where the decisions of the learner are solely chosen according to a cost function, in many realistic scenarios learner's decisions are inputs to a *control system*. Examples include power supply management in the presence of time-varying energy costs due to demand fluctuations and tracking of an adversarial target. In such scenarios, decisions are usually assumed to be a function of current state which is referred to as a *policy*. As usual, the regret is defined as the difference between the accumulated costs incurred by 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. Of course, the online optimization problem discussed above would reduce to the classical optimal control problem if the cost functions were available to the decision maker. Our work is closely related to the recent work of [9] where an online version of linear quadratic Gaussian control is studied. In particular, 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

[1] Department of Mathematics and Statistics at Queen's University, 13mav1@queensu.ca.

[2] Department of Mathematics and Statistics at Queen's University, bahman.gharesifard@queensu.ca.

[3] Department of Mathematics and Statistics at Queen's University, tamas.linder@queensu.ca.

the time horizon. Another closely related work is [2], where the cost functions are assume to be general convex and globally Lipschitz functions. In contrast to [9], the noise is adversarial. The generated control actions, which lead to a sublinear regret bound, are linear feedbacks which rely on a finite history of the past disturbances. As will become apparent later in this paper, the study of online optimization problems with dynamic/control constraints is a rich complex issue, because the impact of the current decision can propagate through all future times via the underlying dynamics.

Before we state our contributions, it is worth pointing 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* [10]. 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 [28, 20, 17].

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 [26]. 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 [27, 18]. Finally, our setting is also related to online optimization in dynamic environments [12], where the decisions are constrained in dynamics chosen by the environment. However, the objective of [12] 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.

Contributions. Similar to [9] we consider the online linear quadratic Gaussian optimal control problem, where the cost function only becomes available in hindsight. In contrast to that work, where an online algorithm using semi-definite programming update is employed to generate the control inputs, we take a control-theoretic approach and employ a modified version of the classical Riccati update, using averaged past data, to generate control policies. Our main result is a $\mathcal{O}(\log T)$ regret bound for the online linear quadratic Gaussian optimal control problem, improving the best known $\mathcal{O}(\sqrt{T})$ bound [9] for time horizon T, under some boundedness assumption. 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 particular, for the scalar case, we prove a stronger result that initializing the control policy to be stable is enough to guarantee boundedness of the solutions of the proposed online Riccati update.

Notation. Throughout the paper, we use the following mathematical notion. Let $\mathbb{R}$ denote the set of real numbers. 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 $\mathrm{Tr}(A)$. If A is an $n \times n$ real matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$, then the spectral radius of $\rho(A)$ of A is $\rho(A) = \max\{|\lambda_1|, \ldots, |\lambda_n|\}$. We use $A \succeq B$ to indicate that $A - B$ is positive semi-definite.

2 Problem Formulation

We start by describing the general problem of online optimization in control systems. We focus on a special class of control systems where the system dynamics are linear and the cost functions are quadratic. 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 [25]: 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

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

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

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

We consider controllers of the form $u_t = \pi_t(x_t)$, where the function $\pi_t : \mathbb{R}^n \to \mathbb{R}^m$ is called a policy. This assumption does not place any restriction, as the optimal policy will provably be of this form [25]. It is well-known that under the assumption that the control system is stabilizable, and cost matrices Q_t and R_t are positive definite, the optimal policy is a stable linear feedback of the state, which will be described in Section 3.

2.2 Problem Setting

We now define the problem we study in this work, following [9]. 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 stable linear feedback, we use the set of stable linear feedback functions as the set of admissible policies. This setting is formally presented next.

Let $x_t \in \mathbb{R}^n$ and $u_t \in \mathbb{R}^m$ be the control state and controller action at time $t \geq 1$. At each time t, 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

$$J_t(K_t) = \mathbb{E}\Big[x_t^\top Q_t x_t + u_t^\top R_t u_t\Big]. \tag{2}$$

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

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

where $\mathcal{K}$ is the set of stable 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.

3 Iterative Methods for Solving the 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 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:

$$P_t = A^\top P_{t+1} A - A^\top P_{t+1} B (B^\top P_{t+1} B + R_t)^{-1} B^\top P_{t+1} A + Q_t \tag{3}$$

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) the pair (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) the pair (A, C), where $Q = C^\top C$, is detectable [i.e., if $u_t \to 0$ and $Cx_t \to 0$ then, $x_t \to 0$],

it is well-known that the optimal policy is unique, time invariant, and is a linear function of the state [5], i.e., $u_t = -K^\star x_t$. Here $K^\star$ is given by

$$K^\star = (B^\top P^\star B + R)^{-1} B^\top P^\star A, \tag{4}$$

where $P^\star$ satisfies the discrete algebraic Riccati equation (DARE):

$$P^\star = A^\top P^\star A - A^\top P^\star B (B^\top P^\star B + R)^{-1} B^\top P^\star A. \tag{5}$$

Moreover, P_t given by (3) converges to $P^\star$ as $t \to \infty$ [25]. By using the policy $K^\star$, we have that $x_{t+1} = (A - BK^\star)x_t + w_t$. The optimal policy $K^\star$ is guaranteed to be stable i.e. $\rho(A - BK^\star) < 1$. 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 $(A - BK^\star)x + w_t$, so that we have $\mathbb{E}[x] = \mathbb{E}[(A - BK^\star)x + w_t]$, which implies $\mathbb{E}[x] = 0$, and the covariance matrix $X = \mathbb{E}[xx^\top]$ satisfies $X = (A - BK^\star)X(A - BK^\star)^\top + W$, see e.g., [9].

Several methods for solving DARE exist in the literature, including iterative methods [7], algebraic methods [22], and semi-definite programming [4]. Our work is based on iterative methods, and in particular, two techniques that we review here. The first is given in [7], 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 [16], uses the following idea: Let P_t be the solution of the equation

$$P_t = (A - BK_t)^\top P_t (A - BK_t) + K_t^\top R K_t + Q, \tag{6}$$

where

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

starting from a stable 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.,

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

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 stable policies. Similar to [9], 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.

4 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 [9]. The notion of strong stability is defined as follows.

Definition 4.1. *A policy K is called stable if $\rho(A-BK)<1$. A policy K is (κ,γ)-strongly stable (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 stable policy K is stable, since the matrices $A-BK$ and L are similar and hence $\rho(A-BK)=\rho(L)\leq(1-\gamma)$. [9, Lemma B.1.], included as Lemma A.1 in the Appendix, shows that every stable policy is (κ,γ)-strongly stable for some $\kappa>0$ and $0<\gamma\leq 1$. 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 $\widehat{X}$, which satisfies

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

Lemma A.2 in the Appendix provides the details. Note that Lemma A.2 only applies to the controllers with a fixed policy. In order to obtain a similar result for the variation of the state covariance matrices using a sequence of different (κ,γ)-strongly stable policies $\{K_t\}_{t\geq 1}$, we need to define a notion of sequential strong stability, which is presented next.

Definition 4.2. *A sequence of policies $\{K_t\}_{t\geq 1}$ is sequentially (κ,γ)-strongly stable, 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*

$$A-BK_t=H_tL_tH_t^{-1}$$

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 A.3 of the Appendix. 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

$$P_t=(A-BK_t)^\top P_t(A-BK_t)+\bar{Q}_t+K_t^\top\bar{R}_tK_t, \tag{7}$$

where

$$K_{t+1}=(B^\top P_tB+\bar{R}_t)^{-1}B^\top P_tA \tag{8}$$

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 stable policy. The reason for this update will become clear as part of our algorithm in Section 5. 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 stable, with appropriate choices of κ and γ.

Proposition 4.3. *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 (7), and assume that the policy K_t given by (8) is stable for $t\geq 1$. 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 stable.*

We provide a proof of this result in the Appendix. 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 A.6), one can obtain that the sequence $\{K_t\}_{t\geq 1}$ is sequentially strongly stable.

Proposition 4.4. *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 (7), and assume that the policy K_t given by (8) is stable 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 stable.*

We postpone the proof to the Appendix. Note 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. We believe that this property should hold only by assuming stability of K_1 for the general case, but have not been able to prove this. Nevertheless, we prove the result for the scalar case in Proposition A.4 in the Appendix.

5 The Online Riccati Algorithm

We outline our main algorithm in this section. We consider the set of admissible policies $\mathcal{K}$ to be the set of stable policies. We propose an algorithm to generate stable policies for the linear quadratic control problem. Before that, we state our assumptions.

Assumption 5.1. *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 $\mathrm{Tr}(Q_t) \leq \sigma$, $\mathrm{Tr}(R_t) \leq \sigma$, for some $\sigma > \mu > 0$ for all $t \geq 1$.*
- *For the noise covariance matrix W we have that $\omega = \mathrm{Tr}(W) < \infty$.*

We first provide an informal description of the algorithm; a formal description is given in Algorithm 1. We start from a stable 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_ts and R_ts through (6). 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 in our algorithm.

Remark 5.2 (Parameters used in Algorithm 1)**.** Our algorithm naturally uses the parameters μ, and σ, stated in Assumption 5.1, and correspond to an estimate on the space where the time-varying matrices Q_t and R_t can be selected from. For the reset step, we also need (an estimate on) the strong stability parameters κ and γ which are defined in Algorithm 1. Proposition 4.3 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.4; 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.4 to vector cases which is an avenue of our current research will remove this restriction all together.

Algorithm 1 Online Riccati Update

```
Input: The system matrices $A$ and $B$, initial state $x_1$, time horizon $T$, parameters $\nu, \mu, \kappa = \sqrt{\nu/\mu}, \gamma = 1/(2\kappa^2), \sigma$
Output: A sequence of stable policies $\{K_t\}_{t=1}^T$
Initialize $K_1$ to be stable
for each $t = 1, 2, \cdots, T$ :
    receive $x_t$
    use controller $u_t = -K_t x_t$ and receive $Q_t$ and $R_t$
    update $\bar{R}_t = \frac{t-1}{t}\bar{R}_{t-1} + \frac{1}{t}R_t$, $\bar{Q}_t = \frac{t-1}{t}\bar{Q}_{t-1} + \frac{1}{t}Q_t$
    update $P_t$ as the solution of
        $$P_t = (A - BK_t)^\top P_t (A - BK_t) + \bar{Q}_t + K_t^\top \bar{R}_t K_t$$
    Reset:
    if $t = t^\star := \lceil \frac{4\kappa^3\|B\|}{\gamma\mu}(2\sigma\kappa + \frac{2\kappa^3\|B\|\sigma(1+\kappa^2)}{\gamma}) + 1 \rceil$ :
        Initialize $\ell = 0$, $\widehat{P}_0 = P_{t^\star}$, and $\widehat{K}_0 = K_{t^\star}$
        while $\|\widehat{P}_\ell - \widehat{P}_{\ell-1}\| > (\frac{2\sigma}{\|B\|} + \frac{4\kappa^2\sigma(1+\kappa^2)}{\gamma})/t^\star$ :
            $\ell \leftarrow \ell + 1$
            $\widehat{K}_\ell = (B^\top \widehat{P}_{\ell-1} B + \bar{R}_{t^\star})^{-1} B^\top \widehat{P}_{\ell-1} A$
            $\widehat{P}_\ell$ satisfies $\widehat{P}_\ell = (A - B\widehat{K}_\ell)^\top \widehat{P}_\ell (A - B\widehat{K}_\ell) + \bar{Q}_{t^\star} + \widehat{K}_\ell^\top \bar{R}_{t^\star} \widehat{K}_\ell$
        return $P_{t^\star} = \widehat{P}_\ell$
    return $K_{t+1} = (B^\top P_t B + \bar{R}_t)^{-1} B^\top P_t A$
```

6 Main Results

We are now in a position to state our main contribution.

Theorem 6.1. *Suppose that the tuple $(A, B, \{Q_t\}_{t=1}^T, \{R_t\}_{t=1}^T, W)$ satisfies Assumption 5.1. Suppose that the matrices P_t generated by Algorithm 1 are uniformly bounded, i.e., $P_t \le \nu I$. Let $\kappa = \sqrt{\frac{\nu}{\mu}}$ and $\gamma = 1/2\kappa^2$. Then for $T \ge \frac{4\kappa^3\|B\|}{\gamma\mu}(2\sigma\kappa + \frac{2\kappa^3\|B\|\sigma(1+\kappa^2)}{\gamma}) + 1$, we have that*

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

where $t^\star = \frac{4\kappa^3\|B\|}{\gamma\mu}(2\sigma\kappa + \frac{2\kappa^3\|B\|\sigma(1+\kappa^2)}{\gamma}) + 1$,

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

and $\hat{m}$ and l are constants defined in Lemmas A.6 and A.9, respectively. Consequently,

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

Sketch of the proof of Theorem 6.1. The proof is provided in the Appendix, and is quite involved. We provide a brief sketch here. Our first technical result Lemma A.5 shows that Algorithm 1, as long as it is initialized at an stable policy, iteratively produces stable polices. This step is analogous to the classical result of [16] 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 4.3. A technical part of this proof demonstrates the reason that we needed 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 A.8, A.9, and A.10 which eventually yield the result. □

Note that the assumption of (κ, γ)-strongly stability in Theorem 6.1 will be satisfied as long as the solutions to the online Riccati equation are uniformly bounded, see Proposition 4.3. In particular, we do not need this assumption for the scalar case, see Proposition A.4.

References

[1] A. Agarwal, E. Hazan, S. Kale, and R. E. Schapire. Algorithms for portfolio management based on the Newton method. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 9–16, 2006.

[2] N. Agarwal, B. Bullins, E. Hazan, S. Kakade, and K. Singh. Online control with adversarial disturbances. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 111–119, 2019.

[3] O. Anava, E. Hazan, S. Mannor, and O. Shamir. Online learning for time series prediction. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 172–184, 2013.

[4] V. Balakrishnan and L. Vandenberghe. Semidefinite programming duality and linear time-invariant systems. *IEEE Transactions on Automatic Control*, volume 48, pages 30–41, 2003.

[5] D. P. Bertsekas. Stable optimal control and semicontractive dynamic programming. *SIAM Journal on Control and Optimization*, volume 56, pages 231–252, 2018.

[6] A. Blum and Y. Mansour. From external to internal regret. *Journal of Machine Learning Research*, volume 8, pages 1307–1324, 2007.

[7] P. E. Caines and D. Q. Mayne. On the discrete time matrix Riccati equation of optimal control. *International Journal of Control*, volume 12, pages 785–794, 1970.

[8] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[9] A. Cohen, A. Hasidim, T. Koren, N. Lazic, Y. Mansour, and K. Talwar. Online linear quadratic control. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1029–1038, 2018.

[10] C. E. Garcia, D. M. Prett, and M. Morari. Model predictive control: Theory and practice; a survey. *Automatica*, volume 25, pages 335–348, 1989.

[11] E. Gofer, N. Cesa-Bianchi, C. Gentile, and Y. Mansour. Regret minimization for branching experts. In *Conference on Learning Theory*, pages 618–638, 2013.

[12] E. C. Hall and R. M. Willett. Online convex optimization in dynamic environments. *IEEE Journal of Selected Topics in Signal Processing*, volume 9, pages 647–662, 2015.

[13] E. Hazan. Introduction to online convex optimization. *Foundation and Trends in Optimization*, volume 2, pages 157–325, 2016.

[14] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, volume 69, pages 169–192, 2007.

[15] E. Hazan and S. Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, volume 15, pages 2489–2512, 2014.

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

[17] R. Jenatton, J. Huang, and C. Archambeau. Adaptive algorithms for online convex optimization with long-term constraints. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 402–411, 2016.

[18] A. Karimi and C. Kammer. A data-driven approach to robust control of multivariable systems by convex optimization. *Automatica*, volume 85, pages 227 – 233, 2017.

[19] H. Luo, C. Wei, and K. Zheng. Efficient online portfolio with logarithmic regret. In *Advances in Neural Information Processing Systems 31*, pages 8235–8245. Curran Associates, Inc., 2018.

[20] M. J. Neely and H. Yu. Online convex optimization with time-varying constraints. *arXiv preprint arXiv:1702.04783*, 2017.

[21] M. Patel and N. Ranganathan. IDUTC: an intelligent decision-making system for urban traffic-control applications. *IEEE Transactions on Vehicular Technology*, volume 50, pages 816–829, 2001.

[22] L. Rodman and P. Lancaster. *Algebraic Riccati Equations*. Oxford Mathematical Monographs. 1995.

[23] S. Ross, G. Gordon, and D. Bagnell. 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*, pages 627–635, 2011.

[24] S. Shalev-Shwartz. *Online Learning and Online Convex Optimization*, volume 12 of *Foundations and Trends in Machine Learning*. Now Publishers Inc, 2012.

[25] T. Soderstrom. *Discrete-Time Stochastic Systems: Estimation and Control*. Springer-Verlag, 2nd edition, 2002.

[26] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[27] Y. Yang, Z. Guo, H. Xiong, D. Ding, Y. Yin, and D. C. Wunsch. Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[28] H. Yu, M. Neely, and X. Wei. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems 30*, pages 1428–1438. 2017.

[29] J. Zhai, Y. Li, and H. Chen. An online optimization for dynamic power management. In *2016 IEEE International Conference on Industrial Technology (ICIT)*, pages 1533–1538, 2016.

A Appendix

This section includes the proofs of our main results. We start with recalling two results from [9].

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

We refer the reader to [9, Lemma B.1] for a proof of this lemma.

Lemma A.2. *Let the pair (A, B) be stabilizable, and assume the controller uses a fixed (κ, γ)-strongly stable 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 $\widehat{X}$, and in particular, for any $t \geq 1$,*

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

We refer the reader to [9, Lemma 3.2] for a proof.

Lemma A.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 stable with $\kappa > 0$ and $0 < \gamma \leq 1$. For each K_t, let $\widehat{X}_t$ be the corresponding steady-state covariance matrix, i.e., $\widehat{X}_t$ satisfies $\widehat{X}_t = (A - BK_t)\widehat{X}_t(A - BK_t)^\top + W$ and assume that $\|\widehat{X}_{t+1} - \widehat{X}_t\| \leq \eta_t$ 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 $X_1 \succeq 0$. Then for $t \geq 1$,*

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

The proof is similar to [9, Lemma 3.5], but we include it for completeness.

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

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

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

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

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

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

Taking the norms yeilds

$$\begin{aligned} \|\Delta_{t+1}\| \leq& \|L_t\|^2 \|H_{t+1}^{-1} H_t\|^2 \|\Delta_t\| + \|H_{t+1}^{-1}\|^2 \|\widehat{X}_t - \widehat{X}_{t+1}\| \\ \leq& (1-\gamma)^2(1+\gamma)^2 \|\Delta_t\| + \frac{\eta_t}{\alpha^2} \\ \leq& (1-\gamma^2)^2 \|\Delta_t\| + \frac{\eta_t}{\alpha^2}, \end{aligned}$$

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

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

which concludes the proof. □

Proposition 4.3. By the assumption of stability and since $Q_t \succeq \mu I$, we have that

$$\begin{aligned}P_t = & (A - BK)^\top P_t (A - BK) + \bar{Q}_t + K^\top \bar{R}_t K \\ \succeq & (A - BK)^\top P_t (A - BK) + \mu I, \end{aligned} \tag{9}$$

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

$$\mu I \preceq P_t \preceq \nu I. \tag{10}$$

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 (9) 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 $\|L_t\| \le \sqrt{1 - \bar{\kappa}^{-2}} \le 1 - \bar{\kappa}^{-2}/2$. Also, using (10) we have that

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

which finishes the proof. □

Proposition 4.4. Proceeding as in the proof of Proposition 4.3, one can show that the matrix $L_t = P_t^{1/2}(A - BK_t)P_t^{-1/2}$ satisfies $\|L_t\| \le 1 - 1/2\bar{\kappa}^2$ with $\|P_t^{1/2}\| \le \sqrt{\nu}$ and $\|P_t^{-1/2}\| \le 1/\sqrt{\mu}$. To establish the sequential strong stability stated by Definition 4.2 it thus suffices to show that $\|P_{t+1}^{-1/2}P_t^{1/2}\| \le 1 + 1/2\bar{\kappa}^2$ for $t \ge 1$. To this end, observe that $\|P_{t+1} - P_t\| \le \eta$, and that

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

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

Proposition A.4. *Let $n = m = 1$ and let $\{P_t\}_{t=1}^T$ be a sequence of positive numbers generated by Equation* (7) *and* (8) *recursively, and assume that policy K_t is stable for all $t \ge 1$. Assume that $A \ge 3$ or $A \le 1$ and $|\bar{Q}_t - \bar{Q}_{t-1}| \le \frac{2\max_{t\ge1}\bar{Q}_t}{t}$ and $|\bar{R}_t - \bar{R}_{t-1}| \le \frac{2\max_{t\ge1}\bar{R}_t}{t}$. Then there exists $\nu > 0$ such that $P_t \le \nu$ for all $t \ge 1$.*

Proof. Note that

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

Since K_t is stable using the stability of K_1, c.f. Lemma A.5, we have that

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

By the stability of K_t and positiveness of P_{t-1} we have that $P_{t-1} > \frac{(|A|-1)\bar{R}_{t-1}}{B^2}$. Let y_t be a real-valued function on $[\frac{(|A|-1)\bar{R}_{t-1}}{B^2}, \infty)$ defined by

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

The function y_t has a local minimum on the interval $[\frac{(|A|-1)\bar{R}_{t-1}}{B^2}, \infty)$, and a horizontal asymptote as P_{t-1} goes to infinity. By doing calculation we have that

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

and

$$\lim_{P_{t-1}\to\infty} y_t(P_{t-1}) = \bar{Q}_t + \frac{A^2}{B^2}\bar{R}_t$$

Using these, we conclude that for $P_{t-1} \geq \frac{\bar{Q}_t \bar{R}_{t-1}}{2\bar{R}_t} + \frac{A^2-1}{2B^2}\bar{R}_{t-1}$, we have that

$$P_t = y_t(P_{t-1}) \leq \bar{Q}_t + \frac{A^2}{B^2}\bar{R}_t \leq \max\{\bar{Q}_t + \frac{A^2}{B^2}\bar{R}_t\}.$$

In order to show that $P_{t-1} \geq \frac{\bar{Q}_t \bar{R}_{t-1}}{2\bar{R}_t} + \frac{A^2-1}{2B^2}\bar{R}_{t-1}$, we use the following calculation.

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

where we have used $K_{t-1}^2 \geq \frac{(|A|-1)^2}{B^2}$ using the stability assumption of K_{t-1}. Now for $A > 3$ or $A < 1$ we have that

$$\frac{(|A|-1)^2}{B^2}\bar{R}_{t-1} > \frac{A^2-1}{2B^2}\bar{R}_{t-1},$$

and since $\bar{Q}_t - \bar{Q}_{t-1} \leq 2/t \max_{t\geq 1} \bar{Q}_t$ and $\bar{R}_t - \bar{R}_{t-1} \leq 2/t \max_{t\geq 1} \bar{R}_t$, it can be shown that for $t \geq \hat{t} = \frac{\max_{t\geq 1} \bar{Q}_t}{2\min_{t\geq 1} \bar{Q}_t}(1 + \frac{\max_{t\geq 1} \bar{R}_t}{\min_{t\geq 1} \bar{R}_t})$, we have that

$$\bar{Q}_{t-1} \geq \frac{\bar{Q}_t \bar{R}_{t-1}}{2\bar{R}_t},$$

which proves $P_{t-1} \geq \frac{\bar{Q}_t \bar{R}_{t-1}}{2\bar{R}_t} + \frac{A^2-1}{2B^2}\bar{R}_{t-1}$. Using $\nu = \max\{\max_{t\leq\hat{t}} P_t, \max_{t\geq 1}\{\bar{Q}_t + \frac{A^2}{B^2}\bar{R}_t\}\}$ we have that $P_t \leq \nu$, as claimed. □

A.1 Proof of Theorem 6.1:

First we give a straightforward reformulation of the regret function. For matrices A and B of appropriate size, let $A \bullet B = \mathrm{Tr}(A^\top B)$. Then

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

$$= \sum_{t=1}^{T} (Q_t + K_t^\top R_t K_t) \bullet (X_t - \widehat{X}_t) \tag{11}$$

$$+ \sum_{t=1}^{T} (Q_t + K_t^\top R_t K_t) \bullet \widehat{X}_t - \sum_{t=1}^{T} (Q_t + K^{\star\top} R_t K^\star) \bullet \widehat{X}^\star \tag{12}$$

$$+ \sum_{t=1}^{T} (Q_t + K^{\star\top} R_t K^\star) \bullet \widehat{X}^\star - \sum_{t=1}^{T} (Q_t + K^{\dagger\top} R_t K^\dagger) \bullet \widehat{X}^\dagger \tag{13}$$

$$+ \sum_{t=1}^{T} (Q_t + K^{\dagger\top} R_t K^\dagger) \bullet (\widehat{X}^\dagger - X_t^\dagger), \tag{14}$$

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

$$\widehat{X}^\dagger = (A - BK^\dagger)\widehat{X}^\dagger(A - BK^\dagger)^\top + W. \tag{15}$$

$K^\star$ 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 can show that (13) 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 in the second equality, the cyclic property of the trace in the third equality, and (15) in the forth equality. By [16], $P^\star \preceq P^\dagger$ and we have the result.

We start with our first technical result, which shows that Algorithm 1 produces stable polices. This step is similar to the classical result of [16] 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 4.3.

Lemma A.5. *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 stable policy K_1. Then policy K_t remains stable for all $t \geq 1$.*

Proof. We proceed by an induction argument. First, since the system is stabilizable, there exists a stable policy and hence we can choose K_1 to be stable, i.e. such that $\rho(A - BK_1) < 1$. Assume now that K_t is stable, for some $t \geq 1$. Then, using (7), P_t is uniquely determined by

$$P_t = \sum_{i=0}^{\infty} (A_t^\top)^i (\bar{Q}_t + K_t^\top \bar{R}_t K_t) A_t^i. \tag{16}$$

By a straightforward computation, we have that

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

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 (7), we have that

$$P_t = A_{t+1}^\top P_t A_{t+1} + V, \tag{17}$$

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,

$$P_t = \sum_{i=0}^{\infty} (A_{t+1}^\top)^i (V) A_{t+1}^i, \tag{18}$$

It is easy to observe that V is positive definite. Now, using (16), since K_t is stable, the matrix P_t is finite. Using (18), and the fact that the left side of (18) is finite, we have that $\rho(A_{t+1}) < 1$, i.e., K_{t+1} is stable, otherwise the sum on the right side of (18) will diverges. □

In order to get a $\log(T)$ regret bound, we need to have bounds of order $\mathcal{O}(1/t)$ on $\|P_t - P_{t-1}\|$, $\|\widehat{X}_t - \widehat{X}_{t-1}\|$ and $\|K_t - K_{t-1}\|$. Also, recall that such bounds are essential for obtaining sequential strong stability using Proposition 4.4. The next lemma and its corollary serves this purpose.

Lemma A.6. *Suppose that $\mu I \preceq Q_t, R_t$ and $\mathrm{Tr}(Q_t), \mathrm{Tr}(R_t) \leq \sigma$. 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 stable. Then we have $\|P_{t+1} - P_t\| \leq m/t$ for some $m > 0$, for $t \geq 1$.*

Proof. Note that using (17), we have

$$\begin{aligned} P_{t+1} - P_t =& A_{t+1}^\top (P_{t+1} - P_t) A_{t+1} + K_{t+1}^\top (\bar{R}_{t+1} - \bar{R}_t) K_{t+1} + (\bar{Q}_{t+1} - \bar{Q}_t) \\ & - (K_{t+1} - K_t)^\top (B^\top P_t B + \bar{R}_t)(K_{t+1} - K_t). \end{aligned} \tag{19}$$

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

$$K_{t+1} - K_t = (B^\top P_t B + \bar{R}_t)^{-1} \big[B^\top (P_t - P_{t-1}) A_t + (\bar{R}_{t-1} - \bar{R}_t) K_t \big]. \tag{20}$$

Using this along with (19), we have that

$$\begin{aligned} P_{t+1} - P_t =& A_{t+1}^\top (P_{t+1} - P_t) A_{t+1} + K_{t+1}^\top (\bar{R}_{t+1} - \bar{R}_t) K_{t+1} + (\bar{Q}_{t+1} - \bar{Q}_t) \\ & - \big[B^\top (P_t - P_{t-1}) A_t + (\bar{R}_{t-1} - \bar{R}_t) K_t \big]^\top (B^\top P_t B + \bar{R}_t)^{-1} \qquad (21) \\ & \times \big[B^\top (P_t - P_{t-1}) A_t + (\bar{R}_{t-1} - \bar{R}_t) K_t \big]. \qquad (22) \end{aligned}$$

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

$$\begin{aligned} P_{t+1} - P_t =& \sum_{i=0}^{\infty} (A_{t+1}^\top)^i M_t A_{t+1}^i \\ & \leq \|M_t\| \sum_{i=0}^{\infty} (A_{t+1}^\top)^i A_{t+1}^i, \end{aligned} \tag{23}$$

where

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

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

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

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

$$\begin{aligned} \|M_t\| =& \|K_{t+1}^\top (\bar{R}_{t+1} - \bar{R}_t) K_{t+1} + (\bar{Q}_{t+1} - \bar{Q}_t)\| \\ & + \|(B^\top P_t B + \bar{R}_t)^{-1}\| (\|B\| \|A_t\| \|P_t - P_{t-1}\| + \|(\bar{R}_{t-1} - \bar{R}_t) K_t\|)^2. \end{aligned} \tag{24}$$

Using (23) and (24), we also have

$$z_{t+1} \leq c_t (h_t z_t + d_t)^2 + e_{t+1}, \tag{25}$$

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

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

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

$$c_t \leq \kappa^2/\gamma\mu,\ d_t \leq \frac{2\sigma\kappa}{t}, \text{ and } e_t \leq \frac{2\kappa^2\sigma(1+\kappa^2)}{\gamma t},\ h_t \leq \|B\|\kappa, \tag{26}$$

for $t \geq 1$. We next claim that there exists a time $t^\star$ and a constant $m > 0$ such that $z_t \leq m/t$ for all $t > t^\star$. 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 $t \geq t^\star = \frac{4\kappa^3\|B\|}{\gamma\mu}(2\sigma\kappa + \frac{2\kappa^3\|B\|\sigma(1+\kappa^2)}{\gamma}) + 1$, using an elementary calculation, one can observe that

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

The claim then follows by noting that

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

where we have used (26).

It remains to show that the condition we placed to obtain the last inequality, i.e., that $z_{t^\star+1} \leq m/(t^\star + 1)$, is satisfied. To proceed with this, first note that $t^\star$ 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 (19). 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 $\|\widehat{P}_\ell - \widehat{P}_{\ell-1}\| \leq m/t^\star$, 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 $t^\star$ the algorithm implements $P_{t^\star} = \widehat{P}_\ell$. In the next time step $t^\star + 1$, the algorithm updates $P_{t^\star+1}$ as usual, using (7). We know by the previous part of the proof that $\|P_{t^\star+1} - P_{t^\star}\| \leq m/(t^\star+1)$, which shows that $z_{t^\star+1} \leq m/(t^\star+1)$ is satisfied. To conclude the proof, note that we can show that $z_t \leq \hat{m}/t$, for all $t \geq 1$, simply by selecting $\hat{m} = \max\{m, tz_t | t \leq t^\star\}$. □

Corollary A.7. *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

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

Note that Lemma A.6 can be used to bound $K_t - K_{t-1}$. Using (20), we have that

$$\begin{aligned}\|K_{t+1} - K_t\| \le& \|(B^\top P_t B + \bar{R}_t)^{-1}\| \left[\|B\|\|P_t - P_{t-1}\|\|A_t\| + \|\bar{R}_{t-1} - \bar{R}_t\|\|K_t\|\right] \\ \le& \frac{\kappa}{\mu}(\|B\|\hat{m} + 2\sigma)/t, \end{aligned} \tag{27}$$

where we have used $\|(B^\top P_t B + \bar{R}_t)^{-1}\| \le \mu^{-1}$, $\|A_t\| \le \kappa$, $\|K_t\| \le \kappa$, and $\hat{m}$ is given in the proof of Lemma A.6. Using this $\|\widehat{X}_t - \widehat{X}_{t+1}\|$ is bounded by $M/t + M'/t^2$, where

$$M = \frac{2\kappa^6\omega}{\mu\gamma^2}\|B\|\big(\|B\|\hat{m} + 2\sigma\big), \tag{28}$$

and

$$M' = \frac{\kappa^6\omega}{\mu\gamma^2}\|B\|^2(\|B\|\hat{m} + 2\sigma)^2, \tag{29}$$

where we have used

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

□

The following lemmas will be used to derive bounds on the redundancy terms (11), (12), and (14).

Lemma A.8. *Suppose that the tuple $(A, B, \{Q_t\}_{t=1}^T, \{R_t\}_{t=1}^T, W)$ satisfies Assumption 5.1. Suppose that the matrices P_t generated by Algorithm 1 are uniformly bounded, i.e., $P_t \le \nu I$. 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*

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

Proof. For $t \ge t^\star$, we have that $\|P_{t+1} - P_t\| \le m/t \le \mu/\kappa^2$. Then, using Proposition 4.4, the matrices K_t are sequentially (κ, γ)-strongly stable for $t \ge t^\star$ ($\kappa = \sqrt{\frac{\nu}{\mu}}$ and $\gamma = 1/(2\kappa^2)$). Using this by Lemma A.3, we conclude that for $t \ge t^\star$

$$\|X_{t+1} - \widehat{X}_{t+1}\| \le \kappa^2 e^{-2\gamma^2(t+1-t^\star)}\|X_{t^\star} - \widehat{X}_{t^\star}\| + \kappa^2 \sum_{s=0}^{t-t^\star} e^{-2\gamma^2 s}\eta_{t-s}; \tag{30}$$

hence we can separate (11) into two parts as follows:

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

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

$$\begin{aligned} \sum_{t=1}^{t^\star}(Q_t + K_t^\top R_t K_t)\bullet(X_t-\widehat{X}_t) &= \sum_{t=1}^{t^\star}\mathrm{Tr}\big[(Q_t+K_t^\top R_t K_t)(X_t-\widehat{X}_t)\big] \\ &\le \sum_{t=1}^{t^\star}\mathrm{Tr}(Q_t+K_t^\top R_tK_t)\|(X_t-\widehat{X}_t)\| \\ &\le t^\star\sigma(1+\kappa^2)\max_{0<t\le t^\star}\|(X_t-\widehat{X}_t)\|, \end{aligned} \tag{31}$$

where we have used

$$\mathrm{Tr}(Q_t+K_t^\top R_tK_t)\le\sigma(1+\kappa^2)$$

Using (30), we have that

$$\begin{aligned} \sum_{t=t^\star}^{T}(Q_t + K_t^\top R_t K_t)\bullet(X_t-\widehat{X}_t) &\le \sum_{t=t^\star}^{T}\mathrm{Tr}(Q_t+K_t^\top R_tK_t)\|(X_t-\widehat{X}_t) \\ &\le \sum_{t=t^\star}^{T}\sigma(1+\kappa^2)\|X_t-\widehat{X}_t\| \\ &\le (\sigma(1+\kappa^2))\kappa^2\sum_{t=t^\star}^{T}\Big(e^{-2\gamma^2 t}\|X_{t^\star}-\widehat{X}_{t^\star}\|+\sum_{s=0}^{t-t^\star}e^{-2\gamma^2 s}\eta_{t-s}\Big) \\ &\le 2\kappa^4\sigma(\|X_{t^\star}-\widehat{X}_{t^\star}\|\frac{e^{-2\gamma^2 t^\star}}{1-e^{-2\gamma^2}}+\sum_{t=t^\star}^{T}\sum_{s=0}^{t-t^\star}e^{-2\gamma^2 s}\eta_{t-s}). \end{aligned}$$

Note that by using Corollary A.7, we have $\eta_t = M/t + M'/t$, where M and M' are given by (28) and (29). Consequently,

$$\begin{aligned} \sum_{t=t^\star}^{T}(Q_t + K_t^\top R_t K_t)\bullet(X_t-\widehat{X}_t) \le& 2\kappa^4\sigma\|X_{t^\star}-\widehat{X}_{t^\star}\|\frac{e^{-2\gamma^2 t^\star}}{1-e^{-2\gamma^2}} \\ &+2\kappa^4\sigma\sum_{t=t^\star}^{T}\sum_{s=0}^{t-t^\star}e^{-2\gamma^2 s}\Big(\frac{M}{t-s}+\frac{M'}{(t-s)^2}\Big). \end{aligned} \tag{32}$$

Next, by changing the order of summation we obtain

$$\begin{aligned} \sum_{t=t^\star}^{T}\sum_{s=0}^{t-t^\star}e^{-2\gamma^2 s}\Big(\frac{M}{t-s}+\frac{M'}{(t-s)^2}\Big) &= \sum_{s=0}^{T-t^\star}e^{-2\gamma^2 s}\sum_{t=s+t^\star}^{T}\Big(\frac{M}{t-s}+\frac{M'}{(t-s)^2}\Big) \\ &\le \sum_{s=0}^{T-t^\star}e^{-2\gamma^2 s}\Big(M\log\Big(\frac{T-s}{t^\star}\Big)+\frac{M'\pi^2}{6}\Big) \\ &\le \frac{M'\pi^2}{6(1-e^{-2\gamma^2})}+\sum_{s=0}^{T-t^\star}Me^{-2\gamma^2 s}\log\Big(\frac{T}{t^\star}\Big) \\ &\le \frac{M'\pi^2}{6(1-e^{-2\gamma^2})}+\frac{M}{1-e^{-2\gamma^2}}\log\Big(\frac{T}{t^\star}\Big), \end{aligned}$$

where we have used a logarithmic upper bound for $\sum_{t=t^\star}^{T-s}1/t$ 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 (32) we obtain

$$\begin{aligned} \sum_{t=t^\star}^{T}(Q_t + K_t^\top R_t K_t)\bullet(X_t-\widehat{X}_t) \le& 2\kappa^4\sigma\Big(\|X_{t^\star}-\widehat{X}_{t^\star}\|\frac{e^{-2\gamma^2 t^\star}}{1-e^{-2\gamma^2}}+\frac{M'\pi^2}{6(1-e^{-2\gamma^2})} \\ &+\frac{M}{1-e^{-2\gamma^2}}\log\Big(\frac{T}{t^\star}\Big)\Big) \end{aligned} \tag{33}$$

The result follow by adding (31) and (33).

□

Lemma A.9. *Suppose that the tuple* $(A, B, \{Q_t\}_{t=1}^T, \{R_t\}_{t=1}^T, W)$ *satisfies Assumption 5.1. Suppose that the matrices* P_t *generated by Algorithm 1 are uniformly bounded, i.e.,* $P_t \le \nu I$*. Let* $\kappa = \sqrt{\frac{\nu}{\mu}}$ *and* $\gamma = 1/2\kappa^2$*. Then the covariance matrices* $\widehat{X}_t$ *and* $\widehat{X}^\star$ *satisfy*

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

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

$$\begin{aligned}(Q_t + K_t^\top R_t K_t) =& (t\bar{Q}_t - (t-1)\bar{Q}_{t-1}) + K_t^\top (t\bar{R}_t - (t-1)\bar{R}_{t-1})K_t \\ =& t(\bar{Q}_t + K_t^\top \bar{R}_t K_t) - (t-1)(\bar{Q}_{t-1} + K_t^\top \bar{R}_{t-1} K_t) \\ =& t(P_t - A_t^\top P_t A_t) - (t-1)(P_{t-1} - A_t^\top P_{t-1} A_t) \\ &+ (t-1)(K_t - K_{t-1})^\top (B^\top P_{t-1} B + \bar{R}_{t-1})^{-1}(K_t - K_{t-1}),\end{aligned} \tag{34}$$

where we have used (7) and (17) in the third equality. Note that

$$\begin{aligned}A_t^\top P_t A_t \bullet \widehat{X}_t =& \mathrm{Tr}(A_t^\top P_t A_t \widehat{X}_t) \\ =& \mathrm{Tr}(P_t A_t \widehat{X}_t A_t^\top) \\ =& P_t \bullet A_t \widehat{X}_t A_t^\top \\ =& P_t \bullet (\widehat{X}_t - W) \\ =& P_t \bullet \widehat{X}_t - P_t \bullet W.\end{aligned} \tag{35}$$

Therefore, by multiplying (34) and $\widehat{X}_t$ we obtain

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

where we have used (35) to cancel out some terms. Summing over t and using the telescopic series for $tP_t \bullet W - (t-1)P_{t-1} \bullet W$, we obtain

$$\begin{aligned}\sum_{t=1}^{T}(Q_t + K_t^\top R_t K_t) \bullet \widehat{X}_t \le& \, TP_T \bullet W \\ &+ \sum_{t=1}^{T}(t-1)(K_t - K_{t-1})^\top (B^\top P_{t-1} B + \bar{R}_{t-1})^{-1}(K_t - K_{t-1}) \bullet \widehat{X}_t.\end{aligned} \tag{36}$$

On the other hand,

$$\begin{aligned}\sum_{t=1}^{T}(Q_t + K^{\star\top} R_t K^\star) \bullet \widehat{X}^\star =& T(\bar{Q}_T + K^{\star\top} \bar{R}_T K^\star) \bullet \widehat{X}^\star \\ =& T(P^\star - A^{\star\top} P^\star A^\star) \bullet \widehat{X}^\star \\ =& T(P^\star \bullet \widehat{X}^\star - P^\star \bullet A^\star \widehat{X}^\star A^{\star\top}) \\ =& T(P^\star \bullet \widehat{X}^\star - P^\star \bullet \widehat{X}^\star + P^\star \bullet W) \\ =& TP^\star \bullet W.\end{aligned} \tag{37}$$

Therefore, by subtracting (37) from (36) we have

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

Note that $P^\star$ 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 [16], i.e. there exists $C > 0$ such that for all $t \geq 2$,

$$\|P_t - P^\star\| \leq C\|P_{t-1} - P^\star\|^2 \tag{38}$$

and by a similar analysis, we also have

$$\|P_{t+1} - P_t\| \leq C\|P_t - P_{t-1}\|^2. \tag{39}$$

Here we use a similar technique to bound $\|P_T - P^\star\|$. We can update the sequence P_t after time T by starting at P_T using (7), with $\bar{Q}_t = \bar{Q}_T$ and $\bar{R}_t = \bar{R}_T$ fixed for $t \geq T$. We hence have that

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

where we have used (38). For $T \geq t^\star$, $C\|P_T - P_{T+1}\| < 1$ 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

$$T(P_T - P^\star) \bullet W \leq T\omega\|P_T - P^\star\| \leq T\omega l\|P_T - P_{T+1}\| \leq \frac{T\omega l\hat{m}}{T} = \omega l\hat{m}, \tag{40}$$

where we have used $\omega = \mathrm{Tr}(W)$ and $\|P_T - P_{T+1}\| \leq \hat{m}/T$ by Lemma A.6. We now proceed by noting that

$$\begin{aligned}
&\sum_{t=2}^{T}(t-1)(K_t - K_{t-1})^\top (B^\top P_{t-1} B + \bar{R}_{t-1})^{-1}(K_t - K_{t-1}) \bullet \widehat{X}_t \\
&\qquad\leq \sum_{t=2}^{T}(t-1)\mathrm{Tr}(\widehat{X}_t)\frac{1}{(t-1)^2\mu^3}(\|B\|\kappa\hat{m} + 2\sigma\kappa)^2 \\
&\qquad\leq \frac{\kappa^2}{\gamma}\omega(\|B\|\kappa\hat{m} + 2\sigma\kappa)^2 \sum_{t=2}^{T}\frac{1}{(t-1)\mu^3} \\
&\qquad\leq \frac{\kappa^4\omega}{\gamma\mu^3}(\|B\|\hat{m} + 2\sigma)^2(1 + \log(T)),
\end{aligned} \tag{41}$$

where we have used the bound in (27) on $\|K_t - K_{t-1}\|$, and the bound for $\mathrm{Tr}(\widehat{X}_t)$. Adding (41) and (40) completes the proof. □

Lemma A.10. *Suppose that the tuple $(A, B, \{Q_t\}_{t=1}^T, \{R_t\}_{t=1}^T, W)$ satisfies Assumption 5.1. 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 we have*

$$\sum_{t=1}^{T}(Q_t + K^{\star\top} R_t K^\star) \bullet (\widehat{X}^\star - X_t^\star) \leq \frac{\sigma(1+\kappa^2)\kappa^2}{1-e^{-2\gamma}} \|\widehat{X}^\star - X_1^\star\|. \tag{42}$$

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

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

as claimed. □

To conclude, by summing the right hand side of (31), (33), (41), (40), and (42), we obtain the regret bound as follows,

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

which finishes the proof of Theorem 6.1