Regret Analysis of Online LQR Control via Trajectory Prediction and Tracking

Yitian Chen[*] YITIAN.CHEN@ANU.EDU.AU
Timothy L. Molloy[*] TIMOTHY.MOLLOY@ANU.EDU.AU
Tyler Summers[†] TYLER.SUMMERS@UTDALLAS.EDU
Iman Shames[*] IMAN.SHAMES@ANU.EDU.AU
[*]*CIICADA Lab, The Australian National University* [†]*The University of Texas at Dallas*

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

In this paper, we propose and analyse a new method for online linear quadratic regulator (LQR) control with a priori unknown time-varying cost matrices. The cost matrices are revealed sequentially with the potential for future values to be previewed over a short window. Our novel method involves using the available cost matrices to predict the optimal trajectory, and a tracking controller to drive the system towards it. We adopted the notion of dynamic regret to measure the performance of this proposed online LQR control method, with our main result being that the (dynamic) regret of our method is upper bounded by a constant. Moreover, the regret upper bound decays exponentially with the preview window length, and is extendable to systems with disturbances. We show in simulations that our proposed method offers improved performance compared to other previously proposed online LQR methods.

Keywords: Online LQR, Dynamic Regret, Trajectory tracking.

1. Introduction

Optimal control problems arise in many fields such as econometrics (Björk et al., 2021; Radneantu, 2009), robotics (Hampsey et al., 2023; Renganathan et al., 2020), physics (Liu et al., 2021) and machine learning (Westenbroek et al., 2020). The Linear Quadratic Regulator (LQR) problem is the archetypal optimal control problem with vector-valued states and controls, and is reviewed in the following. Consider a controllable linear time-invariant system

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

where t is a nonegative integer, m and n are positive integers, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $x_t, w_t \in \mathbb{R}^n$, and $x_0 = \bar{x}_0$ for some $\bar{x}_0 \in \mathbb{R}^n$, and $u_t \in \mathbb{R}^m$. For a given finite time horizon $T \geq 2$ and initial condition $\bar{x}_0$, the control decisions $\{u_t\}_{t=0}^{T-2}$ are computed to minimise the quadratic cost function

$$J_T(\{x_t\}_{t=0}^{T-1}, \{u_t\}_{t=0}^{T-2}) := \sum_{t=0}^{T-2} x_t^\intercal Q_t x_t + u_t^\intercal R_t u_t + x_{T-1}^\intercal Q_{T-1} x_{T-1}, \tag{2}$$

where $Q_t \in \mathbb{S}_+^n$ and $R_t \in \mathbb{S}_{++}^m$ are time-varying cost matrices and $\mathbb{S}_+^n$ and $\mathbb{S}_{++}^n$ denote the sets of positive semi-definite symmetric and positive definite symmetric matrices, respectively. The states x_t and controls u_t minimising (2) must satisfy (1). When the cost matrices $\{Q_t\}_{t=0}^{T-1}$ and $\{R_t\}_{t=0}^{T-2}$

© 2023 Y. Chen[*], T. L. Molloy[*], T. Summers[†] & I. Shames[*].

are known *a priori*, the controls minimising (2) subject to (1) can be found in closed form, cf. (Anderson and Moore, 2007, Chapter 2). However, in many real-world applications, such as power systems (Kouro et al., 2009), chemistry (Chen et al., 2012) and mechatronics (Vukov et al., 2015), full information about the cost matrices over the whole time horizon is unavailable to the decision maker (in advance).

In our work, for a given time horizon T and preview window length $0 \leq W \leq T-2$, we suppose that at any time t where $0 \leq t < T-2-W$, only the initial condition of the system (1) and the (partial) sequences of cost matrices $\{Q_i\}_{i=0}^{t+W}$ and $\{R_i\}_{i=0}^{t+W}$ are known. Let the cost-function information available to the decision maker at time t be

$$\mathcal{H}_t := \{\{Q_i\}_{i=0}^{t+W}, \{R_i\}_{i=0}^{t+W}, \bar{x}_0\}, \tag{3}$$

where $\mathcal{H}_t$ contains the full temporal information about the cost matrices for $t \geq T-2-W$. The main focus of our work is to propose a novel control policy that generates u_t using the information available at time t, and investigate its performance. We specifically consider a feedback control policy $\pi(\cdot,\cdot)$ of the form

$$u_t = \pi(x_t, \mathcal{H}_t), \tag{4}$$

and adopt the notion of regret to measure its performance. Several different notions of regret have been well studied and explored in the online optimization problem, including static regret (Zinkevich, 2003; Shalev-Shwartz, 2012), dynamic regret (Jadbabaie et al., 2015). In our work, performance is measured by dynamic regret. For any control sequence $\{u_t\}_{t=0}^{T-2}$ and associated state sequence $\{x_t\}_{t=0}^{T-1}$, the dynamic regret is defined as

$$\text{Regret}_T(\{u_t\}_{t=0}^{T-2}) := J_T(\{x_t\}_{t=0}^{T-1}, \{u_t\}_{t=0}^{T-2}) - J_T(\{x_t^*\}_{t=0}^{T-1}, \{u_t^*\}_{t=0}^{T-2}), \tag{5}$$

where

$$\{u_t^*\}_{t=0}^{T-2} := \operatorname*{argmin}_{\{\upsilon_i\}_{i=0}^{T-2}} J_T(\{\xi_i\}_{i=0}^{T-2}, \{\upsilon_i\}_{i=0}^{T-2}), \tag{6}$$

and $\{x_t^*\}_{t=0}^{T-1}$ satisfy the system dynamics (1) for input sequence $\{u_t^*\}_{t=0}^{T-2}$.

1.1. Related Works

Similar investigations of regret in online LQR problems have recently been conducted in Cohen et al. (2018), Zhang et al. (2021), and Akbari et al. (2022), with additional studies focusing on properties of the Riccati operator in such problems conducted in (Sun and Cantoni, 2023a,b). Cohen et al. (2018) and Akbari et al. (2022) considered a different notion of regret involving comparison with controls $\tilde{u}_t = -K\tilde{x}_t$ (instead of u_t^*) generated by a fixed gain K from the set of $(\bar{\kappa}, \bar{\gamma})$-strongly stable gains denoted by $\mathcal{K}$. More precisely, $\mathcal{K}$ is the set of all gains where for any $K \in \mathcal{K}$, there exist matrices L and H such that $A + BK = HLH^{-1}$, with $\|L\| \leq 1 - \bar{\gamma}$ and $\|H\|, \|H^{-1}\| \leq \bar{\kappa}$ for prescribed scalars $\bar{\kappa}$ and $\bar{\gamma}$[1]. For a sequence of controls $\{u_t\}_{t=0}^{T-1}$, the notion of regret for time horizon T and controls $\{u_t\}_{t=0}^{T-1}$ from these works is

$$\text{StablisingRegret}_T(\{u_t\}_{t=0}^{T-2}) := J_T(\{x_t\}_{t=0}^{T-1}, \{u_t\}_{t=0}^{T-2}) - J_T(\{\tilde{x}_t\}_{t=0}^{T-1}, \{\tilde{K}\tilde{x}_t\}_{t=0}^{T-2}), \tag{7}$$

where $\tilde{K} \in \operatorname{argmin}_{K\in\mathcal{K}} J_T(\{\tilde{x}_t\}_{t=0}^{T-1}, \{K\tilde{x}_t\}_{t=0}^{T-2})$ and $\{\tilde{x}_t\}_{t=0}^{T-1}$ satisfies (1).

[1] We use $\|\cdot\|$ to denote either the 2-norm of a vector or the spectral norm of a matrix, depending on its argument.

Cohen et al. (2018) proposed an online LQR algorithm that yields controls with a theoretical regret upper bound of $\text{StablisingRegret}_T(\{u_t\}_{t=0}^{T-1}) \leq O(\sqrt{T})$. However, the algorithm involves a computationally expensive projection step at each time t, and the projection set can become empty for some controllable systems when the covariance of the system disturbances w_t is positive definite[2]. Thus, this method is not applicable to all controllable linear time-invariant systems. Moreover, the theoretical stabilising regret upper bound is proportional to the inverse of the cube of the lower bound of covariance of system disturbances, i.e., $\text{StablisingRegret}_T(\{u_t\}_{t=0}^{T-1}) = O(\frac{1}{\sigma^3})$, where the covariance of disturbances from (1) is lower bounded by $\sigma^2 I$. If $\sigma = 0$, the theoretical regret upper bound is undefined. Akbari et al. (2022) proposed an Online Riccati Update algorithm that obtains $\text{StablisingRegret}_T(\{u_t\}_{t=0}^{T-1}) = O(\sigma^2 \log(T))$. The result avoids the undefined regret upper bound of Cohen et al. (2018) when the covariance matrix is not lower bounded by a positive σ. However, like Cohen et al. (2018), the performance of the algorithm proposed in Akbari et al. (2022) is only guaranteed to achieve sublinear *stabilising regret* (7) against the best *fixed* control gain K from the set $\mathcal{K}$. This notion of regret is not suitable for dynamic non-stationary environments. For example, a self-driving car may operate in different environments such as high-wind areas, or high and low-friction road surfaces. For the best performance in such environments, we need to use time-varying control gains. A suitable notion of regret captures the discrepancy between the performance of the aforementioned gains and the best time-varying policies chosen in hindsight.

Zhang et al. (2021) investigated the dynamic regret (5) offered by an online LQR approach inspired by model predictive control. Future cost matrices and predicted disturbances are assumed to be available over a short future preview window of length $W \geq 0$, and the following assumption is made.

Assumption 1 *There exist symmetric positive definite matrices $Q_{min}, Q_{max}, R_{min}, R_{max}$ such that for time $0 \leq t \leq T-2$,*

$$\begin{aligned} 0 \prec Q_{min} \preceq Q_t \preceq Q_{max}, \\ 0 \prec R_{min} \preceq R_t \preceq R_{max}, \end{aligned} \tag{8}$$

where $F \prec G$ denotes $G - F$ being positive definite for symmetric matrices F and G.

Under Assumption 1, Zhang et al. (2021) proposed an online LQR algorithm for selecting controls u_t at time t by solving

$$\min_{\{u_k\}_{k=t}^{t+W}} \sum_{k=t}^{t+W} x_k^\mathsf{T} Q_k x_k + u_k^\mathsf{T} R_k u_k + x_{t+W+1}^\mathsf{T} P_{max} x_{t+W+1}$$

subject to (1) where P_{max} is the solution of the algebraic Riccati equation for the infinite-horizon LQR problem with cost matrices Q_{max} and R_{max}. The dynamic regret (5) of control sequences generated by this method is shown to be upper bounded by a quantity that shrinks exponentially as the preview window length increases. However, the estimate of the tail cost at each time step (i.e., $x_{t+W+1}^\mathsf{T} P_{max} x_{t+W+1}$) can be too pessimistic due to its reliance on P_{max} and the matrices Q_{max} and R_{max} from the bounds given in Assumption 1.

[2] For example, the set is empty if $A = \begin{pmatrix} 1 & 2 \\ 6 & 9 \end{pmatrix}$, $B = \begin{pmatrix} 9 \\ 6 \end{pmatrix}$, and the disturbances are distributed according to a multivariate Gaussian with mean zero and covariance matrix I_2.

1.2. Contributions

The key contributions of this paper are:

- The proposal of a method for solving the online LQR problem that is independent of the given upper or lower bounds on the cost matrices;
- Development of a regret bound for the disturbance-free case and proof that our proposed control policy yields sublinear regret;
- Provision of sufficient conditions under which our regret bound is less than that of the state-of-the-art methodology; and
- Analysis of our regret bound in the presence of disturbances.

Outline. The rest of the paper is organised as follows. In Section 2, we state the online LQR problem that we consider. In Section 3, we introduce our proposed online LQR algorithm and bound its dynamic regret. In Section 4, we provide numerical results for the simulation of our proposed algorithm. Concluding remarks are presented in the last section.

2. Problem Formulation

In this paper, we consider the following problem.

Problem 1 (Online LQR) *Consider the controllable system* (1)*. Let the cost matrices in* (5) *satisfy Assumption 1 for any given $T \geq 2$ and $W < T - 2$. At time $0 \leq t \leq T - W - 2$, the available information to the decision maker is given by $\mathcal{H}_t$ as defined in* (3) *and the current state x_t. It is desired to design a control policy $\pi(\cdot, \cdot)$ of the form* (4) *that yields a regret, as defined by* (5)*, that is independent of the bounds given in Assumption 1. Moreover, we seek to establish appropriate regret bounds for the following cases:*

a) The case where $w_t = 0$ for $0 \leq t \leq T - 1$;

b) The case where the disturbances w_t for $0 \leq t \leq T - 1$ are independent and identically distributed (i.i.d.) random variables such that $\mathbf{E}(w_t) = 0$ and $\mathbf{E}(w_t w_t^\mathsf{T}) = W_d$ with $\mathbf{E}(\cdot)$ being the expectation operator and $W_d \in \mathbb{S}^n_+$.

Specifically, for part a) of Problem 1 we show that the regret (as defined in (5)) associated with our proposed control policy is sublinear with respect to the time horizon T for the case where $w_t = 0$ for $0 \leq t \leq T - 1$, i.e.,

$$\text{Regret}_T(\{u_t\}_{t=0}^{T-2}) \leq o(T). \tag{9}$$

For part b), we define the notion of "expected regret" as

$$\text{ExpectedRegret}_T(\{u_t\}_{t=0}^{T-2}) := \mathbf{E}(J_T(\{x_t\}_{t=0}^{T-1}, \{u_t\}_{t=0}^{T-2}) - J_T(\{x_t^*\}_{t=0}^{T-1}, \{u_t^*\}_{t=0}^{T-2})), \tag{10}$$

and show that our proposed control policy yields control that satisfy

$$\text{ExpectedRegret}_T(\{u_t\}_{t=0}^{T-2}) \leq C_{ER} T \gamma^{2W}$$

for positive scalars C_{ER} and γ[3]. In what follows we address this problem.

[3] The exact definition of γ will be presented in Theorem 1.

3. Approach and Regret Analysis

Our proposed online LQR approach involves first using the available information $\mathcal{H}_t$ at each time t to predict the optimal state x^*_{t+1} solving the full information LQR problem described in (6). We then select controls to track this prediction. At time $0 \leq t \leq T-1$, we only know the information in $\mathcal{H}_t$. Let $x_{t+1|t+W}$ denote the estimate of the optimal state at time $t+1$ based on $\mathcal{H}_t$ and the current state x_t. We aim to track to the state $x_{t+1|t+W}$ at time $t+1$.

Prediction. At each time t, we plan an optimal trajectory starting from the initial state $\bar{x}_0$ using the known cost matrices up to time $t+W$ and setting all the future matrices to be equal to their known values for time $t+W$. Specifically, at time t where $0 \leq t < T-W$, define $J_{t+W}(\cdot,\cdot)$ as

$$\begin{aligned} J_{t+W}(\{\xi_i\}_{i=0}^{T-1}, \{\upsilon_i\}_{i=0}^{T-2}) := &\sum_{k=0}^{t+W} [\xi_k^\mathsf{T} Q_k \xi_k + \upsilon_k^\mathsf{T} R_k \upsilon_k] \\ &+ \sum_{k=t+1+W}^{T-2} [\xi_k^\mathsf{T} Q_{t+W} \xi_k + \upsilon_k^\mathsf{T} R_{t+W} \upsilon_k] + \xi_{T-1}^\mathsf{T} Q_{t+W} \xi_{T-1}, \end{aligned} \tag{11}$$

and

$$J_{t+W}(\{\xi_i\}_{i=0}^{T-1}, \{\upsilon_i\}_{i=0}^{T-2}) := J_T(\{\xi_i\}_{i=0}^{T-1}, \{\upsilon_i\}_{i=0}^{T-2}) \tag{12}$$

for $T-W \leq t \leq T-1$.

Then, we find the predicted optimal control sequence for all $0 \leq j \leq T-2$ by solving

$$\begin{aligned} \left(\{x_{j|t+W}\}_{j=0}^{T-1}, \{u_{j|t+W}\}_{j=0}^{T-2}\right) = &\underset{(\{\xi_i\}_{i=0}^{T-1}, \{\upsilon_i\}_{i=0}^{T-2})}{\operatorname{argmin}} \quad J_{t+W}(\{\xi_i\}_{i=0}^{T-1}, \{\upsilon_i\}_{i=0}^{T-2}) \\ &\text{subject to} \quad \xi_{i+1} = A\xi_i + B\upsilon_i, \quad \xi_0 = \bar{x}_0. \end{aligned} \tag{13}$$

Prediction Tracking. We propose the following feedback control policy

$$\pi(x_t, \mathcal{H}_t) = K(x_t - x_{t|t+W}) + u_{t|t+W}, \tag{14}$$

where $K \in \mathbb{R}^{m \times n}$ is a control matrix such that $\rho(A+BK) < 1$, and $\rho(\cdot)$ denotes the matrix spectral radius. Intuitively, such control matrix K leads to contraction of the distance between x_{t+1} and $x_{t+1|t+W}$, respectively given by (1) and (13).

3.1. Regret Analysis for the Disturbance-free Case

In the following theorem, we present the result for the case of Problem 1a) that the control sequence generated by (14) incurs a sublinear upper bound regret with respect to time horizon T. Here, with a slight abuse of notation, for a sequence of matrices $\{\Sigma_i\}_{i=0}^N$, we define $\max_{0 \leq t \leq N} \Sigma_t := \{\Sigma_\tau \mid 0 \leq \tau \leq N, \Sigma_\tau \succeq \Sigma_k \text{ for all } 0 \leq k \leq N\}$ and $\min_{0 \leq t \leq N} \Sigma_t := \{\Sigma_\tau \mid 0 \leq \tau \leq N, \Sigma_\tau \preceq \Sigma_k \text{ for all } 0 \leq k \leq N\}$. This enables us to define *cost matrix sequence extrema* as $\bar{R}_{max} := \max_{0 \leq t \leq T-2} R_t$, $\bar{Q}_{max} := \max_{0 \leq t \leq T-1} Q_t$, $\bar{R}_{min} := \min_{0 \leq t \leq T-2} R_t$, and $\bar{Q}_{min} := \min_{0 \leq t \leq T-1} Q_t$. For any matrix Γ, we further define $\lambda_{min}(\Gamma)$ as the minimum eigenvalue of Γ and $\lambda_{max}(\Gamma)$ as the maximum eigenvalue of Γ.

Theorem 1 (Main Result) *Consider the linear system defined by* (1). *For a given time horizon $T \geq 2$ and preview window length $0 \leq W \leq T-2$. Suppose that at time $0 \leq t \leq T-2$ the control input u_t is generated by policy $\pi(\cdot,\cdot)$ as given by (14). Under Assumption 1, the regret defined by* (5) *satisfies*

$$\begin{aligned}\mathrm{Regret}_T(\{u_t\}_{t=0}^{T-2}) \leq \frac{10D\gamma^{2W}\|\bar{x}_0\|^2}{3}\Bigg[&(\alpha_1+\alpha_2)(\frac{C^2C_K\gamma}{(\gamma-1)})^2\Bigg(\gamma^2 S_T(\eta^2\gamma^2) - 2\gamma S_T(\eta^2\gamma) \\ &+ S_T(\eta^2)) + \frac{10C_f^2}{3}((\frac{\eta\gamma}{q(q-\eta\gamma)} - \frac{\eta}{q(q-\eta)})^2 S_T(q^2) \\ &+ \frac{(\eta\gamma)^2 S_T(\eta^2\gamma^2)}{q^2(q-\eta\gamma)^2} + \frac{\eta^2 S_T(\eta^2)}{q^2(q-\eta)^2})\Bigg) + (C_K C^2)^2 S_T(\eta^2)\Bigg],\end{aligned} \tag{15}$$

where $\bar{P}_{max}$ satisfies

$$\bar{P}_{max} = \bar{Q}_{max} + A^\intercal \bar{P}_{max} A - A^\intercal \bar{P}_{max} B(\bar{R}_{max} + B^\intercal \bar{P}_{max} B)^{-1} B^\intercal \bar{P}_{max} A,$$

$D = \left\|\bar{R}_{max} + B^\intercal \bar{P}_{max} B\right\|$, $C_K = \left\|(\bar{R}_{min} + B^\intercal \bar{Q}_{min} B)^{-1}\right\|^2 \left\|\bar{R}_{max} B^\intercal\right\| \frac{\lambda_{max}^2(\bar{P}_{max})}{\lambda_{min}(\bar{Q}_{min})}$, $C = \frac{\lambda_{max}(\bar{P}_{max})}{\lambda_{min}(\bar{Q}_{min})}$, $\eta = \sqrt{1 - \frac{\lambda_{min}(\bar{Q}_{min})}{\lambda_{max}(\bar{P}_{max})}}$, $\alpha = \max_{\substack{0\leq i\leq t-1\\0\leq t\leq T-2}} \{\lambda_{max}(A^\intercal P_{i+1}^* A), \lambda_{max}(A^\intercal P_{i+1|t} A)\}$, $\beta = \min_{0\leq t\leq T-2} \lambda_{min}(Q_t)$, $\gamma = \frac{\alpha}{\alpha+\beta}$, $S_T(z) = \sum_{t=0}^{T-1} z^t$, $\alpha_1 = \max_t \left\|K_{t|t+W} - K\right\|^2$, $\alpha_2 = \max_t 2\left\|K_t^* - K\right\|^2$, $C_f = \max_{n\geq 0} \frac{\|(A+BK)^n\|}{(q+\varepsilon)^n}$, $q = \rho(A+BK)+\varepsilon$, *and* $0 \leq \varepsilon < 1-\rho(A+BK)$.

Proof See (Chen et al., 2022, Appendix A). ■

Remark 2 *For any $z \in [0,1)$ there exists a $\Lambda \in \mathbb{R}$, such that $\lim_{T\to\infty} S_T(z) = \Lambda$. Consequently, $\overline{\lim}_{T\to\infty} \frac{\mathrm{Regret}_T(\{u_t\}_{t=0}^{T-2})}{T} = 0$, which implies that the control sequence described by (14) yields sublinear regret.*

Remark 3 *Let $F(\bar{x}_0, A, B, T, \bar{R}_{max}, \bar{R}_{min}, \bar{Q}_{max}, \bar{Q}_{min}, K)$ denote the right hand side (RHS) of* (15). *By stating almost identical lemmas to (Chen et al., 2022, Lemmas 8 and 9) using the bounds given in Assumption 1 instead of the cost matrices sequence extrema values, one can arrive at a regret bound in terms of these bounds analogous to* (15)*:*

$$\mathrm{Regret}_T(\{u_t\}_{t=0}^{T-2}) \leq F(\bar{x}_0, A, B, T, R_{max}, R_{min}, Q_{max}, Q_{min}, K).$$

In the following proposition, we state a condition in terms of the bounds given in Assumption 1 and the cost matrices sequence extrema where it is guaranteed that the bound given in the above theorem is smaller than that of (Zhang et al., 2021, Theorem 1, Equation (15)). Obviously, there might be other conditions, the exploration of which is left to future work.

Proposition 4 *Adopt the hypothesis of Theorem 1. If*

$$\lambda_{max}^{10}(Q_{max}) \geq \frac{5\left[(1+\frac{\alpha_1+\alpha_2}{(1-\gamma)^2})(\frac{1}{1-\eta^2}) + \frac{10C_f^2}{q^2(q-\eta\gamma)^2(q-\eta)^2(1-\eta^2)(1-\eta^2\gamma^2)(1-q^2)}\right]}{6(C_K^2\lambda_{min}^2(\bar{R}_{min})\lambda_{min}^4(\bar{Q}_{min}))^{-1}\|A\|^2\|B\|^2\|B\bar{R}_{min}^{-1}B^\intercal\|^2}, \tag{16}$$

where Q_{max} is given in Assumption 1, then the RHS of inequality in (Zhang et al., 2021, Theorem 1, Equation (15)) is greater than the RHS of inequality (5) *in Theorem 1.*

Proof See (Chen et al., 2022, Appendix B). ■

The RHS of (16) is independent of the matrices $Q_{min}, Q_{max}, R_{min}, R_{max}$ given in Assumption 1. On the other hand, the upper bound of regret for control decisions generated by (Zhang et al., 2021, Algorithm 1) does depend on these values and even if the actual sequence of the cost matrices remains bounded away from these bounds, the method still explicitly uses the bounds and this is a potential source of conservatism.

3.2. Regret Analysis in the Presence of Disturbances

The result presented in the following theorem address Problem 1 case b). Note that at time t, $\{w_k\}_{k=0}^{t}$ is the available sequence of disturbances to the decision maker. In this case, we still consider a policy $\pi(\cdot,\cdot)$ as given by (14) with the only difference that $x_{t|t+W}$ is obtained by solving the following optimisation problem:

$$\begin{aligned}\left(\{x_{j|t+W}\}_{j=0}^{T-1}, \{u_{j|t+W}\}_{j=0}^{T-2}\right) = &\underset{(\{\xi_i\}_{i=0}^{T-1}, \{\upsilon_i\}_{i=0}^{T-2})}{\operatorname{argmin}} && J_{t+W}(\{\xi_i\}_{i=0}^{T-1}, \{\upsilon_i\}_{i=0}^{T-2})\\ &\text{subject to} && \xi_{i+1} = A\xi_i + B\upsilon_i + w_i \quad \text{for } 0 \leq i \leq t,\\ & && \xi_{i+1} = A\xi_i + B\upsilon_i \quad \text{for } i > t, \quad \xi_0 = \bar{x}_0.\end{aligned} \tag{17}$$

Theorem 5 *Consider the system defined by* (1)*. For a given time horizon $T \geq 2$ and preview window length $0 \leq W \leq T-2$. Suppose that at time $0 \leq t \leq T-2$ the control input u_t is generated by policy $\pi(\cdot,\cdot)$ as given by (14). Under Assumption 1, the expected regret defined by* (10) *satisfies*

$$\text{ExpectedRegret}_T(\{u_t\}_{t=0}^{T-2}) \leq C_{ER} T \gamma^{2W} \tag{18}$$

where C_{ER} is a positive scalar and γ is given in Theorem 1.

Proof See (Chen et al., 2022, Appendix C). ■

In the next section, we investigate the performance of the proposed algorithm for different scenarios.

4. Numerical Simulations

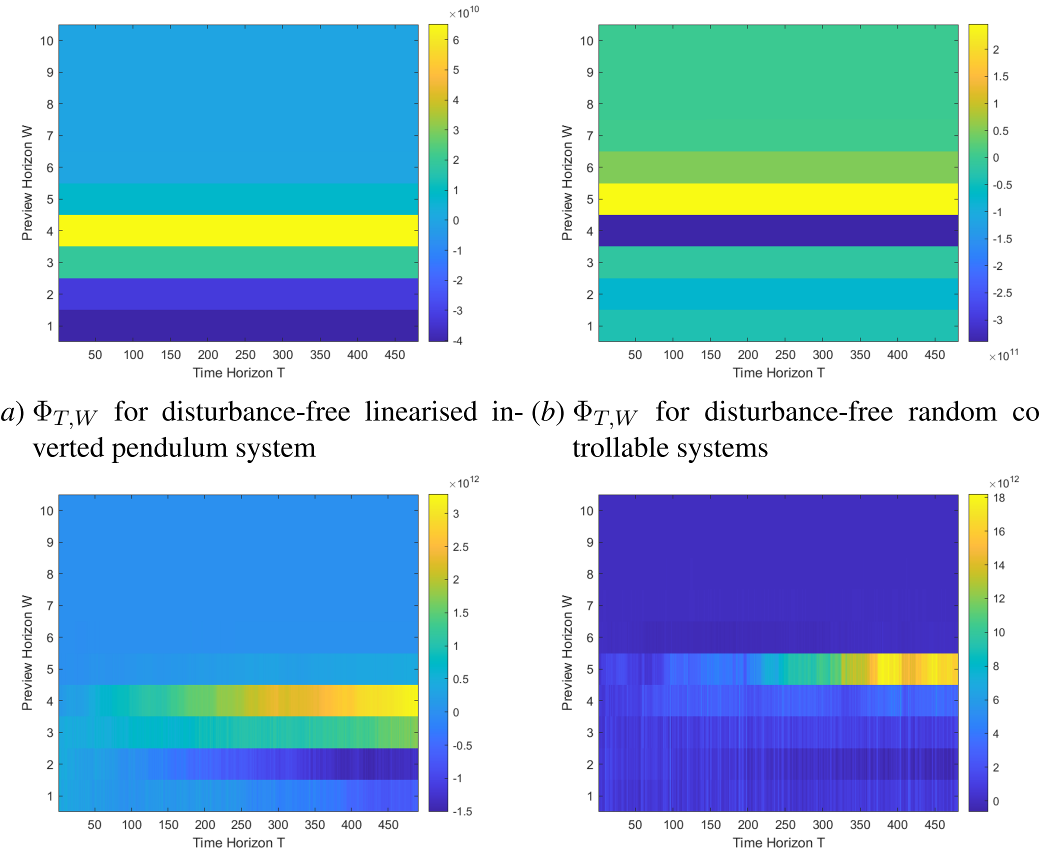
In this section, we numerically demonstrate the performance of the proposed algorithm.[4] To this end, define $\Phi_{T,W} := \text{Regret}_T(\{u_t'\}_{t=0}^{T-2}) - \text{Regret}_T(\{u_t\}_{t=0}^{T-2})$, where $\{u_t'\}_{t=0}^{T-2}$ is generated from (Zhang et al., 2021, Algorithm 1) and $\{u_t\}_{t=0}^{T-2}$ is generated by the policy described in (14), under preview window length of W.

4.1. Linearised Inverted Pendulum

Consider the following linearised inverted pendulum system (Franklin et al., 2020, Chapter 2.13):

$$x_{t+1} = \begin{pmatrix} 0 & 1 & 0 & 0\\ 0 & -0.1818 & 2.6727 & 0\\ 0 & 0 & 0 & 1\\ 0 & -18.1818 & 31.1818 & 0 \end{pmatrix} x_t + \begin{pmatrix} 0\\ 1.8182\\ 0\\ 4.5455 \end{pmatrix} u_t. \tag{19}$$

[4] Code can be found at https://gitlab.anu.edu.au/u7361886/l4dcsimulation.git.



(a) $\Phi_{T,W}$ for disturbance-free linearised inverted pendulum system

(b) $\Phi_{T,W}$ for disturbance-free random controllable systems

(c) $\Phi_{T,W}$ for linearised inverted pendulum system with disturbances

(d) $\Phi_{T,W}$ for random controllable system with disturbances

Figure 1: Performance measure $\Phi_{T,W}$ for simulated systems.

In the following experiments, the preview horizon W ranges from 0 to 19 and the time horizon T ranges from 19 to 500. The cost matrices are chosen uniformly satisfying by Assumption 1 with $Q_{min} = 8 \times 10^3 I_{4\times4}$, $Q_{max} = 3.2 \times 10^4 I_{4\times4}$, $R_{min} = 2 \times 10^3$, and $R_{max} = 9.8 \times 10^4$. The fixed controller from (14) is chosen by placing the poles at the location of $(1, 6, 4, 3) \times 10^{-3}$. We repeat the experiment in 200 trials. Figure 1(a)subfigure demonstrates $\Phi_{T,W}$, under preview window length from 0 to 19 and time horizon from 19 to 500. As the preview window length is greater than 2, our method outperforms (Zhang et al., 2021, Algorithm 1).

4.2. Random Linear Systems

In this experiment, the linear system is randomly chosen where all elements of A and B are drawn uniformly within the range of $(0, 10)$ and ensure the pairs of (A, B) are controllable. The setting of preview window length, time horizon, cost matrices and the pole location for the control matrix K from (14) are identical to what we have chosen in Section 4.1. The plot in Figure 1(b)subfigure demonstrates the subtraction between the regret of control decision generated by (Zhang et al., 2021, Algorithm 1) and the regret of control decision generated by our proposed method, by averaging the regret over 200 trials. As the preview window length exceeds 4, our method outperforms (Zhang et al., 2021, Algorithm 1).

The plots from Figure 1(*a*)subfigure and 1(*b*)subfigure demonstrate that, as the preview window length exceeds the rank of the system, which is the least number of steps required to steer the state of the system to a designated state, the proposed method outperforms the method from Zhang et al. (2021).

4.3. Linear Systems with disturbances

The following experiments repeat the ones considered in Section 4.1 and 4.2, using the system defined in (19) and in the presence of disturbance $w_t \sim \mathcal{N}(0, 25I_{4\times4})$. The setting of the preview window length, time horizon, cost matrices and the pole location for the control matrix K is identical to what we have chosen from the experiment in Section 4.1. The method of finding $x_{t|t+W}$ and $u_{t|t+W}$ can be referred to Remark 3.2. The plots in Figures 1(*c*)subfigure and 1(*d*)subfigure depicts the average value of $\Phi_{T,W}$ after 200 random trials.

5. Conclusions and Future Work

This paper proposes a new online LQR control policy that achieves sublinear dynamic regret for the disturbance-free case where the cost matrices are sequentially revealed as time progresses. The proposed method and consequently its regret has been demonstrated to be, contrary to the state-of-the-art, independent of the *ex-ante* upper and lower bound of the cost matrices. To exhibit the effect of such independence, a sufficient condition is provided under which the regret upper bound of the proposed method is guaranteed to be smaller than that of (Zhang et al., 2021, Theorem 1). This paper leads to many interesting research directions which are briefly discussed below. It would be interesting to devise a methodology for selecting a time-varying feedback gain matrix in (14) instead of a fixed K in order to further minimise the regret. Moreover, one can extend the algorithm to the case of time-varying A_t and B_t for the system matrices and via differential dynamic programming for nonlinear dynamics with control constraints, and establish new dynamic regret bounds.

Acknowledgments

This work was supported in part by the Australian Research Council's Discovery Projects Funding Scheme under Grant DP210102454 and via grant AUSMURIB000001 associated with ONR MURI grant N00014-19-1-2571.

References

Mohammad Akbari, Bahman Gharesifard, and Tamas Linder. Logarithmic regret in online linear quadratic control using Riccati updates. *Mathematics of Control, Signals, and Systems*, April 2022. ISSN 0932-4194, 1435-568X. doi: 10.1007/s00498-022-00323-4.

Brian D. O. Anderson and John B. Moore. *Optimal Control: Linear Quadratic Methods*. Courier Corporation, February 2007. ISBN 978-0-486-45766-6.

Tomas Björk, Mariana Khapko, and Agatha Murgoci. *Time-Inconsistent Control Theory with Finance Applications*. Springer Finance. Springer International Publishing, Cham, 2021. ISBN 978-3-030-81842-5 978-3-030-81843-2. doi: 10.1007/978-3-030-81843-2.

Xianzhong Chen, Mohsen Heidarinejad, Jinfeng Liu, and Panagiotis D. Christofides. Distributed economic MPC: Application to a nonlinear chemical process network. *Journal of Process Control*, 22(4):689–699, April 2012. ISSN 0959-1524. doi: 10.1016/j.jprocont.2012.01.016.

Yitian Chen, Timothy L Molloy, Tyler Summers, and Iman Shames. Regret Analysis of Online LQR Control via Trajectory Prediction and Tracking: Extended Version. November 2022.

Alon Cohen, Avinatan Hassidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online Linear Quadratic Control. *arXiv:1806.07104 [cs, stat]*, June 2018. arXiv: 1806.07104.

Gene F Franklin, Abbas Emami-Naeini, and J. David Powell. *Feedback control of dynamic systems Gene F. Franklin, Stanford University, J. David Powell, Stanford University, Abbas Emami-Naeini, SC Solutions, Inc.* Pearson, New York, NY, eighth edition, global edition edition, 2020. ISBN 1-292-27452-2. Publication Title: Feedback control of dynamic systems.

Matthew Hampsey, Pieter van Goor, Tarek Hamel, and Robert Mahony. Exploiting different symmetries for trajectory tracking control with application to quadrotors. *IFAC-PapersOnLine*, 56 (1):132–137, 2023. 12th IFAC Symposium on Nonlinear Control Systems NOLCOS 2022.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online Optimization : Competing with Dynamic Comparators. *arXiv:1501.06225 [cs, math, stat]*, January 2015. arXiv: 1501.06225.

Samir Kouro, Patricio Cortes, RenÉ Vargas, Ulrich Ammann, and JosÉ Rodriguez. Model Predictive Control—A Simple and Powerful Method to Control Power Converters. *IEEE Transactions on Industrial Electronics*, 56(6):1826–1838, June 2009. ISSN 1557-9948. doi: 10.1109/TIE.2008.2008349. Conference Name: IEEE Transactions on Industrial Electronics.

Yang Liu, Jian Feng Yang, Ren De Qi, and Ning Ning Meng. Nonlinear control of active power filter based on LQR control. *Journal of Physics: Conference Series*, 1748(5):052061, January 2021. ISSN 1742-6596. doi: 10.1088/1742-6596/1748/5/052061. Publisher: IOP Publishing.

Nicoleta Radneantu. Making the Invisible Visible: the Intangible Assets Recognition, the Valuation and Reporting in Romania. *Annals of the University of Petrosani, Economics*, 9:6–6, January 2009.

Venkatraman Renganathan, Iman Shames, and Tyler H. Summers. Towards Integrated Perception and Motion Planning with Distributionally Robust Risk Constraints. *IFAC-PapersOnLine*, 53(2): 15530–15536, January 2020. ISSN 2405-8963. doi: 10.1016/j.ifacol.2020.12.2396.

Shai Shalev-Shwartz. *Online learning and online convex optimization*. Number 4:2 in Foundations and trends in machine learning. Now, Boston, 2012. ISBN 978-1-60198-546-0.

Jintao Sun and Michael Cantoni. On receding-horizon approximation in time-varying optimal control, May 2023a. arXiv:2305.06010 [cs, eess, math].

Jintao Sun and Michael Cantoni. On Riccati contraction in time-varying linear-quadratic control, May 2023b. arXiv:2305.06003 [cs, eess, math].

M. Vukov, S. Gros, G. Horn, G. Frison, K. Geebelen, J. B. Jørgensen, J. Swevers, and M. Diehl. Real-time nonlinear MPC and MHE for a large-scale mechatronic application. *Control Engineering Practice*, 45:64–78, December 2015. ISSN 0967-0661. doi: 10.1016/j.conengprac.2015.08.012.

Tyler Westenbroek, David Fridovich-Keil, Eric Mazumdar, Shreyas Arora, Valmik Prabhu, S. Sastry, and Claire Tomlin. Feedback Linearization for Uncertain Systems via Reinforcement Learning. pages 1364–1371, May 2020. doi: 10.1109/ICRA40945.2020.9197158.

Runyu Zhang, Yingying Li, and Na Li. On the Regret Analysis of Online LQR Control with Predictions. In *2021 American Control Conference (ACC)*, pages 697–703, May 2021. doi: 10.23919/ACC50511.2021.9483108.

Martin Zinkevich. Online Convex Programming and Generalized Infinitesimal Gradient Ascent. 2, April 2003.