Work Report: Implementation of Thompson Sampling for Linear Quadratic Regulators (LQR)

Saurav Kumar
21D070063

August 3, 2025

1 Introduction

This report describes the implementation of Thompson Sampling for Linear Quadratic Regulators (LQR), following the algorithm presented in the paper by Marc Abeille and Alessandro Lazaric. The Thompson Sampling approach is designed to handle the exploration-exploitation trade-off in LQR control problems, where system dynamics are unknown and must be learned interactively. The goal is to compute an optimal control policy that minimizes the cumulative cost over time while efficiently estimating the system's parameters.

2 Problem Setup

We consider an LQR problem with linear system dynamics of the form:

$$x_{t+1} = A^* x_t + B^* u_t + \epsilon_{t+1}, \quad \epsilon_{t+1} \sim \mathcal{N}(0, I), \tag{1}$$

$$c(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t \tag{2}$$

where:

- $x_t \in \mathbb{R}^n$ is the system state at time step t,
- $u_t \in \mathbb{R}^d$ is the control input at time step t,
- $A^* \in \mathbb{R}^{n \times n}$ is the unknown state-transition matrix,
- $B^* \in \mathbb{R}^{n \times d}$ is the unknown control matrix,
- $Q \in \mathbb{R}^{n \times n}$ and $R \in \mathbb{R}^{d \times d}$ are the state and control cost matrices,
- $\epsilon_{t+1} \in \mathbb{R}^n$ is Gaussian noise with zero mean and unit covariance.

The objective is to design a policy $\pi(x_t)$ that minimizes the cumulative cost:

$$J^\pi(A^*, B^*) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} c(x_t, u_t)\right]$$

with unknown parameters A^* and B^*.

3 Algorithm Overview

Input: $\hat{\theta}_0$, $V_0 = \lambda I$, δ, T, τ, $t_0 = 0$

Set $\delta' = \delta/(8T)$
for $t = \{0, \ldots, T\}$ **do**
  if $\det(V_t) > 2\det(V_0)$ **or** $t \geq t_0 + \tau$ **then**
    while $\widetilde{\theta}_t \notin \mathcal{S}$ **do**
      Sample $\eta_t \sim \mathcal{D}^{\mathrm{TS}}$
      Compute $\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta_t$
    end while
    Let $V_0 = V_t$, $t_0 = t$,
  else
    $\widetilde{\theta}_t = \widetilde{\theta}_{t-1}$
  end if
  Execute control $u_t = K(\widetilde{\theta}_t)x_t$
  Move to state x_{t+1}, receive cost $c(x_t, u_t)$
  Compute V_{t+1} and $\widehat{\theta}_{t+1}$
end for

Figure 1: Thompson sampling algorithm.

4 Equations and Estimation Procedure

4.1 Regularized Least Squares

At time step t, the system parameters are estimated using the data collected so far. Define the input-output data matrices as:

$$Z_t = \begin{bmatrix} x_0^\top & u_0^\top \\ \vdots & \vdots \\ x_{t-1}^\top & u_{t-1}^\top \end{bmatrix}, \quad X_{t+1} = \begin{bmatrix} x_1^\top \\ \vdots \\ x_t^\top \end{bmatrix}$$

The least-squares estimate of the parameters $\theta = \begin{bmatrix} A^\top & B^\top \end{bmatrix}$ is computed as:

$$\hat{\theta}_t = (Z_t^\top Z_t + \lambda I)^{-1} Z_t^\top X_{t+1}$$

4.2 Confidence Bound Calculation

The confidence bound for the estimated parameters is calculated as:

$$\beta_t(\delta) = \left(nL_s \sqrt{2 \log \left(\frac{\sqrt{\det V_t}}{\delta \sqrt{\det \lambda I}} \right)} + \lambda^{1/2} S \right) \tag{3}$$

where:

- L_s: The bound on the process noise, representing the sub-Gaussian noise in the system.
- λ: Regularization parameter used in the least squares estimation.
- S: A bound on the system parameters $\mathbf{A}$ and $\mathbf{B}$, ensuring they lie within a certain norm.
- δ: Confidence level parameter, ensuring the true system parameters are within the confidence set with high probability.
- $V_t = Z_t^\top Z_t + \lambda I$: The regularized design matrix.

The confidence set is defined by the following inequality:

$$\text{trace}\left((\hat{\Theta}_t - \Theta^*)^\top V_t (\hat{\Theta}_t - \Theta^*) \right) \leq \beta_t(\delta)^2 \tag{4}$$

where $\hat{\Theta}_t$ is the estimated system parameter and Θ^* is the true parameter.

4.3 Thompson Sampling Perturbation

Thompson Sampling perturbs the RLS estimate by sampling from the ellipsoid E_t defined by the covariance matrix:

$$\tilde{\theta}_t = \hat{\theta}_t + \beta_t(\delta) W_t \eta_t, \quad \eta_t \sim \mathcal{N}(0, I)$$

where W_t is the Cholesky decomposition of V_t^{-1}. The sampled parameter $\tilde{\theta}_t$ is used to compute the control policy for the current episode.

- $W_t \in \mathbb{R}^{(n+d)\times(n+d)}$ is the system state at time step t,
- $\eta_t \in \mathbb{R}^{(n+d)\times n}$ is the control input at time step t, .

4.4 Control Policy via Riccati Equation

Given the perturbed parameters $(\tilde{A}_t, \tilde{B}_t)$, the control policy is computed by solving the DARE:

$$P = Q + \tilde{A}_t^\top P \tilde{A}_t - \tilde{A}_t^\top P \tilde{B}_t (R + \tilde{B}_t^\top P \tilde{B}_t)^{-1} \tilde{B}_t^\top P \tilde{A}_t \tag{5}$$

$$K_t = -(R + \tilde{B}_t^\top P \tilde{B}_t)^{-1} \tilde{B}_t^\top P \tilde{A}_t \tag{6}$$

The control action applied is $u_t = K_t x_t$.

5 Performance Metrics

Cost Function:

$$c_t = x_t^\top Q x_t + u_t^\top R u_t$$

Regret Calculation:

$$R(T) = \sum_{t=0}^{T-1} (c_t - c_t^{\text{opt}})$$

6 Simulation Results

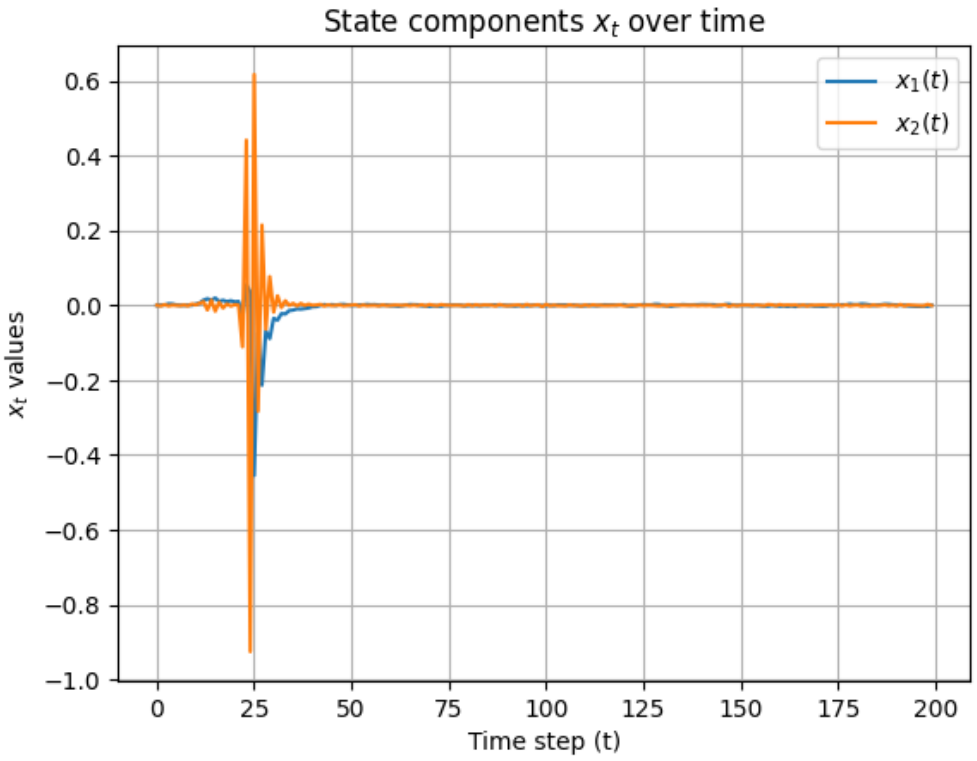
For the simulation, we used the following system parameters and constants:

- State Dimension: $n = 2$
- Control Dimension: $d = 1$
- True System Matrices:
$$\mathbf{A} = \begin{bmatrix} 1.0 & 0.40 \\ 0.005 & -0.99 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0.2 \\ 0.5 \end{bmatrix}$$
- Time Steps: $T = 200$
- Cost Matrices: $Q = I_n$, $R = I_d$ (identity matrices)
- Regularization Parameter: $\lambda = 10^{-4}$
- Bound on Process Noise: $L_s = 0.001$
- Bound on System Parameter: $S = 1.0$
- Confidence Level Parameter: $\delta = 0.1$
- Episode length: $\tau = 1.0$

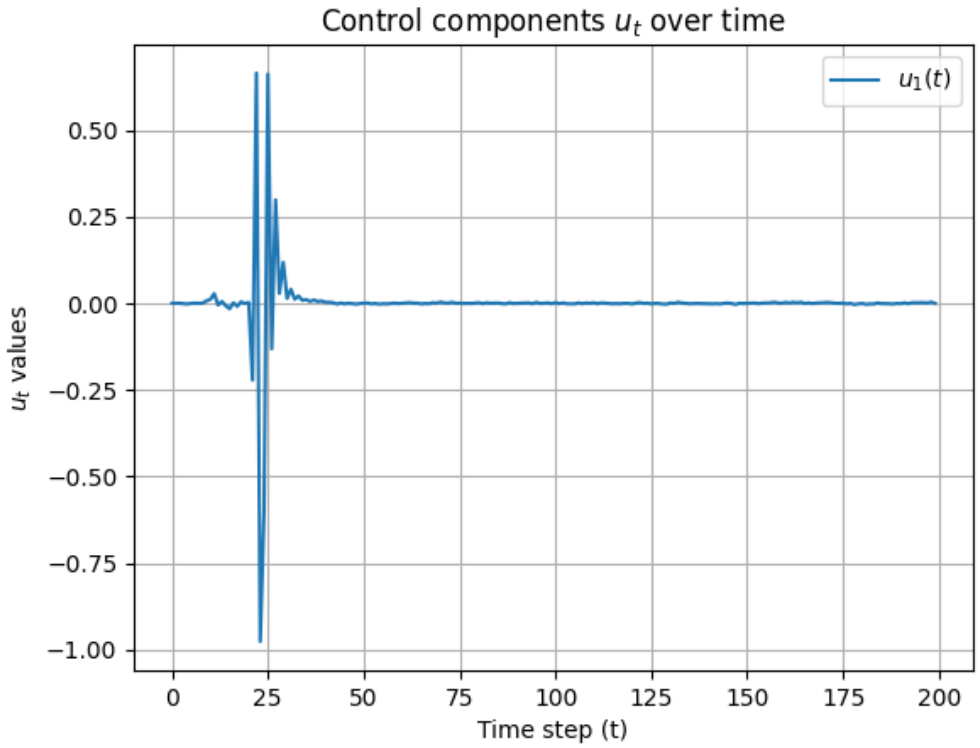
The Thompson Sampling LQR controller was run for 200 time steps with these parameters. Below are the results for the estimated system matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ at the end of the simulation:

$$\tilde{\mathbf{A}} = \begin{bmatrix} 1.0057 & 0.3844 \\ 0.0054 & -0.9749 \end{bmatrix}, \quad \tilde{\mathbf{B}} = \begin{bmatrix} 0.2058 \\ 0.4901 \end{bmatrix}$$

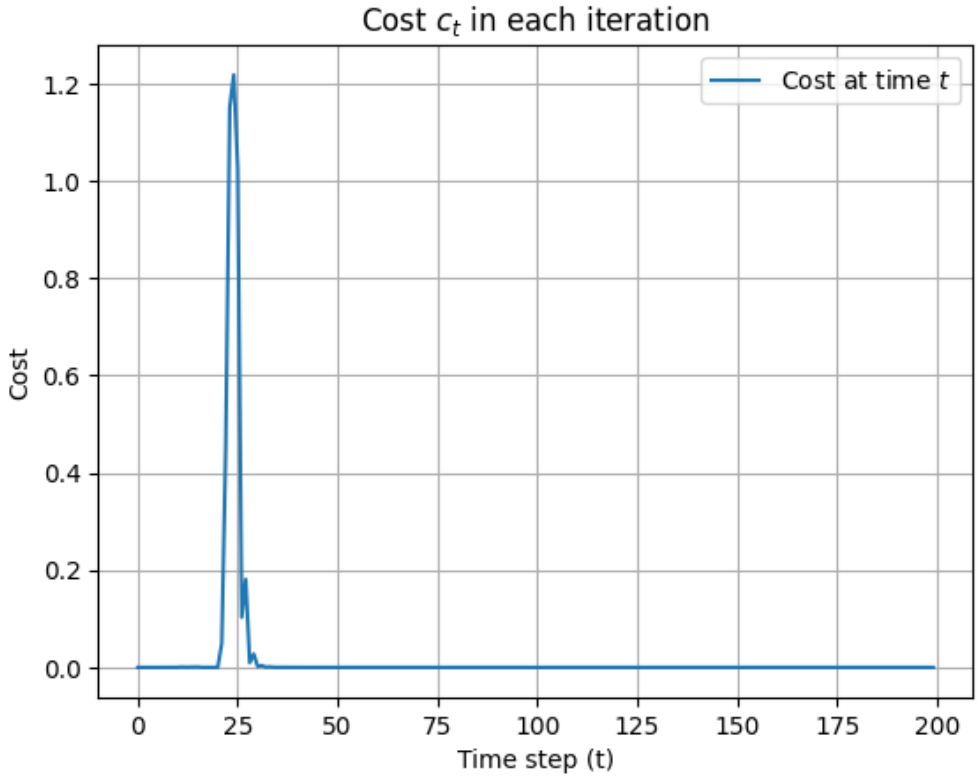
6.1 State Evolution Plot



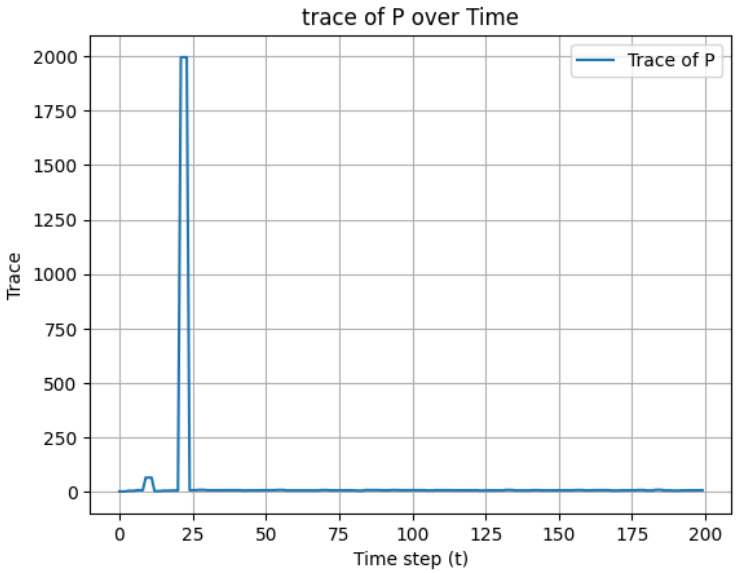
6.2 Control Input Plot



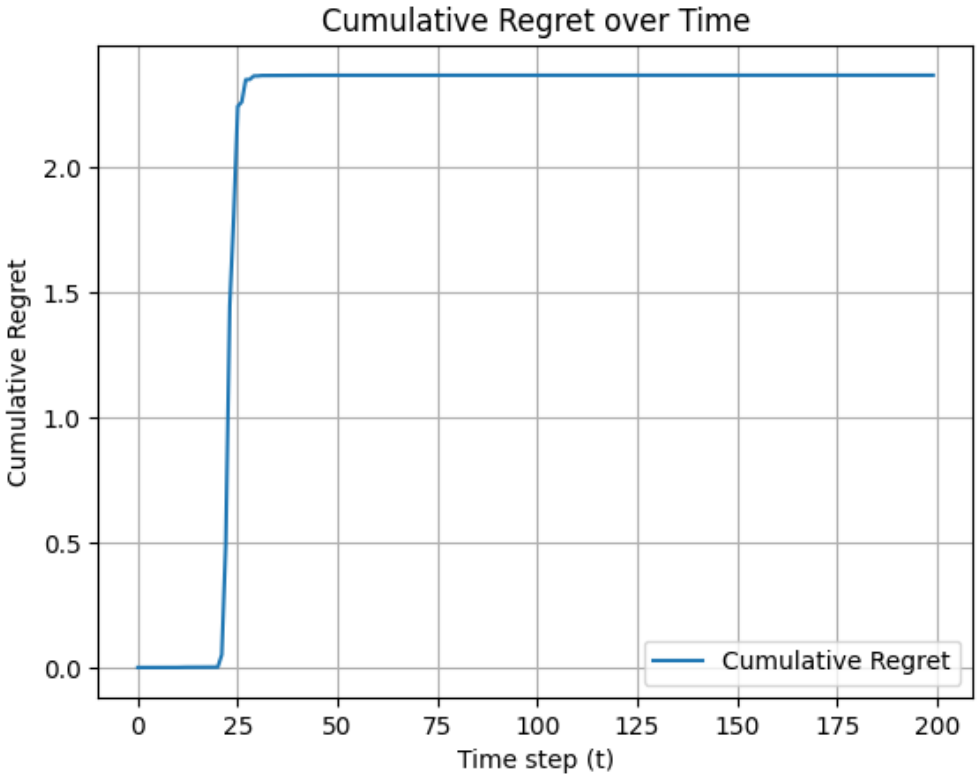
6.3 Instantaneous Cost Plot



6.4 Trace of Riccati Matrix Plot



6.5 Cumulative Regret Plot



7 Python Code

Google drive link:
https://drive.google.com/file/d/1e9GjhCuv_caMM_GgvoFJx67pZi1AeJho/view?usp=drive_link

8 Conclusion

The Thompson Sampling algorithm for LQR control was successfully implemented, and the results demonstrate its ability to efficiently explore and exploit the system dynamics. The cumulative regret followed the expected theoretical bounds, showing that the policy converged to near-optimal behavior.

References

- Marc Abeille and Alessandro Lazaric, Thompson Sampling for Linear-Quadratic Control Problems
- Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction. MIT Press.