# Adaptive LQR Control through Bandit-Inspired Gain Exploration

Federico Baldan
University of Washington
fbaldan@uw.edu

## 1. Approach and Study

This project investigate adaptive control of linear quadratic regulator (LQR) problems by recasting the online selection of feedback gains as a regret-minimization problem inspired by multi-armed bandit frameworks. Through this perspective, we explore how different exploration-exploitation strategies impact performance metrics such as cumulative regret, fuel usage, and safety constraints within a simplified satellite attitude control scenario. First of all, we have to formulate online attitude control as an adaptive LQR bandit problem. At each timestep $t$, we can define a controller which selects a feedback gain $K_t$, applies the control $u_t = K_t x_t$, and incurs the cost $c_t = x_t^\top Q x_t + u_t^\top R u_t$. The benchmark used is the optimal infinite-horizon LQR cost for real system dynamics $(A_\star, B_\star)$:

$$Regret(T) = \mathbb{E}\left[\sum_{t=0}^{T-1}(c_t - c^\star)\right], \quad c^\star = x^\top P x \quad (1)$$

where $P$ is given by the discrete Riccati equation [3]. While our formulation allows us to treat the feedback gains selection as a MAB problem, it is important to note that the considered system is a linear dynamical system where each action affects future states and rewards. As highlighted by Simchowitz & Foster (2020), the problem is more properly an online learning or adaptive control in linear dynamical systems, where regret minimization is measured against the optimal LQR controller for the true system parameters [5]. The true dynamics are sampled around a nominal model:

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t, \quad w_t \sim \mathcal{N}(0, W) \quad (2)$$

We set $d_x = 4$ (three angular rates and one attitude error), and $d_u = 3$ (thruster controls). To account for uncertainty and errors, we perturb $A_0, B_0$ with $\delta \sim \mathcal{U}([-0.3, 0.3])$:

$$A_\star = (1+\delta)A_0, \quad B_\star = (1+\delta)B_0 \quad (3)$$

which tries to captures typical on-orbit scenarios. To better capture environmental variability, at each step we randomly select the noise variance $\sigma_t^2$ from a uniform distribution between $10^{-3}$ and $10^{-1}$. This means the process noise $w_t$ can fluctuate over time, reflecting the changing disturbance levels a satellite might encounter in orbit.

To facilitate algorithmic comparisons inspired by bandit methods, we discretize the space of feedback gains as follows:

$$\mathcal{K}_{\text{disc}} = \{K^{(i)}\}_{i=1}^{100}, \quad (4)$$

$$K^{(i)} = -(R + B_i^\top P_i B_i)^{-1} B_i^\top P_i A_i \quad (5)$$

These gains are precomputed on a grid of $\delta$ values and form a discrete set of stabilizing controllers.

$$\mathcal{K}_{\text{cont}} = \left\{ K \in \mathbb{R}^{d_u \times d_x} \mid \rho(A_\star + B_\star K) < 1 \right\} \quad (6)$$

In this case, candidate controllers are projected back into the stabilizing set using a convex QP [4]. Note that even if our focus is primarily theoretical, we present the approach and provide an illustrative example of the problem. No real dataset is used, as suitable data are not available and the aerospace scenario is intentionally simplified for clarity. For evaluation we use:

- **Cumulative regret**
- **Total fuel** $\sum_t \|u_t\|$: check control effort and propellant usage overtime due to the control.
- **Safety violations** $\frac{1}{T}\sum_t \mathbb{I}\{\|x_t\| > \theta_{\max}\}$: indicates robustness with respect to the constraints.

## 2. Experiments and Study

Reffering to the linear–quadratic dynamics above, we compare adaptive algorithms for selecting feedback gains online. While this setup is inspired by bandit methods, the system's stateful dynamics mean that rewards are not independent, so the problem is better viewed as online adaptive control [5] as discussed previously. We present three core algorithms, each with a distinct exploration–exploitation approach to attitude regulation.

**CE–$\varepsilon$ Greedy:** Following the self-bounding ODE analysis of Simchowitz & Foster [5], we implement an epoch-based certainty-equivalent strategy. At each epoch of length $\tau_k = 2^k$, we update the system estimate $(\hat{A}_t, \hat{B}_t)$ via ”recursive” least squares (RLS) and solve the discrete Riccati equation given by:

$$P = A^\top PA - A^\top PB(R + B^\top PB)^{-1}B^\top PA + Q, \tag{7}$$

with $(A, B) = (\hat{A}_t, \hat{B}_t)$, yielding the controller

$$K_t = -(R + B^\top PB)^{-1}B^\top PA. \tag{8}$$

The control input is then applied with decaying Gaussian exploration:

$$u_t = K_t x_t + \eta_t, \qquad \eta_t \sim \mathcal{N}\left(0,\ \frac{\varepsilon_0}{\sqrt{t}} I\right), \tag{9}$$

where the $t^{-1/2}$ decay means exploration is stronger at the beginning to help learning, but gradually decreases so the controller can prioritize exploitation as $t$ increases. The Riccati equation (7) is solved at each epoch using the latest parameter estimates.

**OFU–LQR (Optimism in Face of Uncertainty):** Following the OFU principle [2, 5], we maintain an ellipsoidal confidence set

$$\mathcal{C}_t = \left\{\theta : (\theta - \hat{\theta}_t)^\top V_t\,(\theta - \hat{\theta}_t) \le \beta_t^2\right\}, \tag{10}$$

where $\theta = \mathrm{vec}(A, B)$ and $V_t$ is the RLS design matrix. This set is updated at each step to represent the uncertainty about the true system parameters, ”shrinking” over time as more data is collected and keeping only those parameter with plausible values. In fact, at each step, we select the most optimistic model within this set:

$$(\tilde{A}, \tilde{B}) = \arg \min_{(A,B)\in\mathcal{C}_t} J_{\mathrm{LQR}}(A, B), \tag{11}$$

where $J_{\mathrm{LQR}}(A, B)$ is the infinite-horizon cost given by the unique solution $P$ of the discrete Riccati equation (7). The corresponding optimal gain $K_t$ is then applied as $u_t = K_t x_t$. This approach allows the controller to focus its exploration on the most uncertain aspects of the system, theoretically leading to regret that grows at most on the order of $\tilde{O}(\sqrt{T})$ [5].

**TS–LQR** Here, we use a linear-Gaussian Thompson sampling approach. For each row $\theta_i$ of $(A, B)$, we keep a Gaussian posterior:

$$\theta_i \sim \mathcal{N}(\hat{\theta}_{i,t},\ \Sigma_t), \qquad \Sigma_t = V_t^{-1}. \tag{12}$$

At every step, we sample $\tilde{\theta}_t$ from this posterior, reshape it into $(\tilde{A}, \tilde{B})$, solve the Riccati equation (7), and apply the resulting gain $u_t = K_t x_t$. The updates

$$V_{t+1} = V_t + \phi_t\phi_t^\top, \qquad \hat{\theta}_{t+1} = \hat{\theta}_t + V_{t+1}^{-1}\phi_t\bigl(x_{t+1} - \hat{\theta}_t^\top\phi_t\bigr) \tag{13}$$

will reduce overtime the uncertainty in our model as more data is collected, so we expect the amount of exploration to decrease as the parameters become more precise.

## 3. Results

We evaluated each algorithm using a Monte Carlo approach, running $M = 10$ independent random seeds to capture both average performance and variability. For every run, we generated a new set of system parameters $(A_\star, B_\star)$, noise realizations, and initial conditions, aiming to reflect consistent algorithmic behavior rather than a single lucky (or unlucky) scenario. To summarize performance, we plot the mean across seeds, with shaded regions indicating 95% confidence intervals. In the context of satellite attitude control, with extremely high reliability and safety, the 95% level is chosen to ensure that the performance bands might reflect not only typical outcomes but also potential catastrophic or unexpected events.

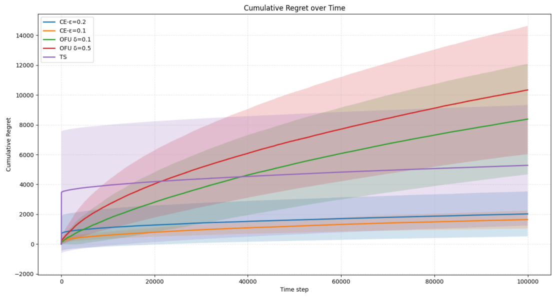

Figure 1. Cumulative regret over time (10 seeds, $T = 10^5$).

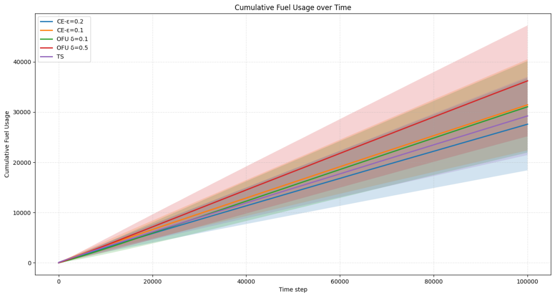

Figure 2. Fuel usage (10 seeds, $T = 10^5$).

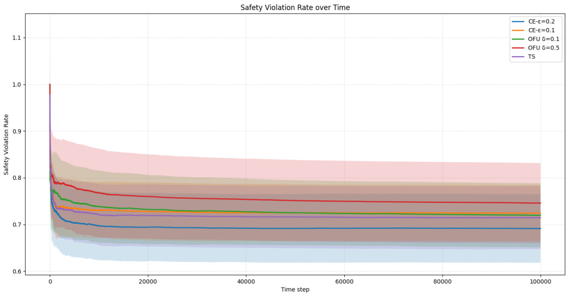

Figure 3. Safety violation rate (10 seeds, $T = 10^5$).

All curves grow roughly as $\sqrt{T}$, which matches the classical $\tilde{O}(\sqrt{T})$ regret bound for adaptive LQR. Among the methods, CE–$\varepsilon = 0.2$ converges fastest, because its decaying Gaussian noise quickly reduces exploration overhead as $t$ increases. OFU, by always selecting the most optimistic model in its confidence set, tends to apply more aggressive gains, which can lead to higher regret—especially for larger values of $\delta$ (e.g., $\delta = 0.5$). Thompson sampling sits between those scenarios: posterior sampling focuses exploration more effectively than CE–$\varepsilon$, but without the conservatism of low-$\delta$ OFU. As expected, achieving lower regret often requires more aggressive control actions, which can result in higher fuel consumption.

Moreover, figure 3 highlights how these exploration strategies impact the safety violation rate. OFU–LQR achieves the lowest rate of safety violations, motivated by the fact that the confidence-ellipsoid approach allows more conservative gains and thus reduces the risk of exceeding attitude constraints. In contrast, CE–$\varepsilon$ and TS show higher safety violation rates, since their more aggressive or less precisely identified gains increase the chance of constraint violations.

## 4. Applicability and limitations

In conclusion, we approached the satellite attitude control problem as an LQR with unknown linear dynamics, considering gain selection as a multi-armed bandit problem over a set of linear feedback controllers. This abstraction allowed us to study how different sequential learning algorithms adapt and perform under uncertainty. However, our experimental setup relies on significant simplifications of real aerospace dynamics. In practice, satellite attitude evolves according to nonlinear equations, and the safety constraints make the kind of exploratory gain updates used by bandit algorithms too risky and not reliable because prone to transient violation of attitude or power limits. Moreover, the bandit abstraction assumes stationarity and independence that may not fully hold in closed-loop control, as highlighted in [1, 5].

That said, I think these methods remain interesting for other domains where real-time adaptation is valuable and safety constraints are less rigid. For instance, in drone control or robotics, controllers following the same logic applied, can be deployed for on-line evaluation of feedback gains, allowing the controller to potentially adapt in real time to system dynamics changes.

## References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011. 3

[2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Regret analysis for stochastic linear bandits. *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pages 1–26, 2011. 2

[3] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2012. 1

[4] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Sewoong Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *arXiv preprint arXiv:1802.08334*, 2019. 1

[5] Max Simchowitz and Dylan J. Foster. Improved regret bounds for linear quadratic regulator. *arXiv preprint arXiv:1812.01251*, 2020. 1, 2, 3