Regret Bounds for Episodic Risk-Sensitive Linear Quadratic Regulator

Wenhao Xu   Xuefeng Gao   Xuedong He

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong

Introduction

What is LQR? Why we need to study risk-sensitive LQR?

Introduction

What is LQR? Why we need to study risk-sensitive LQR?

- LQR: optimize a **quadratic cost function** subject to linear dynamics.
- Abbasi-Yadkori and Szepesvári (2011); Mania et al. (2019); Cohen et al. (2019); Simchowitz and Foster (2020) have studied regret bounds for the **risk-neutral** LQR in the **infinite-horizon average reward setting**.
- Basei et al. (2022) is among the first to establish regret bounds for the **risk-neutral continuous time finite-horizon** LQR in the episodic setting.

Introduction

- Risks exist in many applications, e.g. finance, robotics and healthcare.
- The **linear exponential-of-quadratic regulator** (LEQR) problem is one of the most fundamental problems in risk-sensitive optimal control, and there is extensive literature on this topic (Jacobson, 1973; Whittle, 1990; Zhang et al., 2021).
- We design two algorithms for the LEQR problem in the episodic finite-horizon setting: one requiring a specific identifiability assumption and the other relaxing this assumption. We also provide the regret guarantee for the algorithms.

The LEQR Problem

- Linear dynamics: $x_{t+1} = Ax_t + Bu_t + w_t$, where the state vector $x_t \in \mathbb{R}^n$, the control vector $u_t \in \mathbb{R}^m$, the matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and the process noise $w_t \in \mathbb{R}^n$ form a sequence of i.i.d. Gaussian random vectors.
- Policy $\pi = \{u_0, u_1, \cdots, u_{T-1}\}$.
- The exponential risk-sensitive cost: $J^\pi(x_0) = \frac{1}{\gamma} \log \mathbb{E} \exp\left(\frac{\gamma}{2}\left(\sum_{t=0}^{T-1} c_t(x_t, u_t) + c_T(x_T)\right)\right)$, where $c_t(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t$, $c_T(x_T) = x_T^\top Q_T x_T$, $Q \succeq 0$, $Q_T \succeq 0$, $R \succ 0$, and γ is the risk-sensitivity parameter.
- The goal in the finite-horizon LEQR problem is to choose a control policy π to minimize $J^\pi(x_0)$.

The LEQR Problem

- When the system parameters are all **known**, under the assumption that $I - \gamma P_{t+1} \succ 0$ for all $t = 0, 1, \cdots, T-1$, the optimal feedback control is $u_t^\star = K_t x_t$, where (K_t) can be solved from the modified Riccati equation:

$$
\begin{aligned}
P_T &= Q_T, \\
\widetilde{P}_{t+1} &= P_{t+1} + \gamma P_{t+1} \left(I_n - \gamma P_{t+1}\right)^{-1} P_{t+1}, \\
K_t &= -(B^\top \widetilde{P}_{t+1} B + R)^{-1} B^\top \widetilde{P}_{t+1} A, \\
P_t &= Q + K_t^\top R K_t + (A + BK_t)^\top \widetilde{P}_{t+1} (A + BK_t), \\
t &= 0, 1, \cdots, T-1.
\end{aligned}
$$

The Least-Squares Greedy Algorithm

- We consider **episodic learning** with N episodes and use **total regret** to measure the performance of our algorithm.
- Divide N episodes into L epochs, where the l-th epoch has m_l episodes, thus $\sum_{l=1}^{L} m_l = N$.
- Apply the **least-squares estimation** for the unknown system matrices $\theta = [A\ B]^\top$ by using the data from the $(l-1)$-th epoch.
- Execute the policy according to the **greedy strategy**.

The Least-Squares Greedy Algorithm

Algorithm 1 The Least-Squares Greedy Algorithm

Input: Parameters $L, T, m_1, \theta^1, Q, Q_T, R$
for $l = 1, \cdots, L$ **do**
  $m_l = 2^{l-1} m_1$
  Compute (K_t^l) for all t by using θ^l.
  for $k = 1, \cdots, m_l$ **do**
    for $t = 0, \cdots, T-1$ **do**
      Play $u_t^{l,k} \leftarrow K_t^l x_t^{l,k}$.
    end for
  end for
  Obtain θ^{l+1} from the l_2-regularized least-squares estimation.
end for

Theoretical Result

- **Assumption 1 (Persistence of Excitation).** For the sequence of the controller (K_t), we assume that
$$\left\{ v \in \mathbb{R}^{n+m} \middle| \left[I_n \; K_t^\top \right] v = 0, \forall t = 0, \cdots, T-1 \right\} = \{0\}.$$
- Suppose Assumption 1 holds, the regret upper bound of the Least-Squares Greedy Algorithm is logarithmic in the number of episodes N.

The Least-Squares-Based Algorithm with Exploration Noise

- What would happen if the identifiability condition is **not satisfied**? In particular, is $\sqrt{N}$ regret achievable?

The Least-Squares-Based Algorithm with Exploration Noise

- What would happen if the identifiability condition is **not satisfied**? In particular, is $\sqrt{N}$ regret achievable?
- The answer is yes. The Least-Squares-Based Algorithm with Exploration Noise can achieve this bound without the identifiability condition.
- Apply the **least-squares estimation** for the unknown system matrices $\theta = [A\ B]^\top$ by using the data from the previous $(k-1)$ episodes.
- Execute the control with decaying exploration noise (g_t^k).

The Least-Squares-Based Algorithm with Exploration Noise

Algorithm 2 The Least-Squares-Based Algorithm with Exploration Noise

Input: Parameters $T, N, \theta^1, Q, Q_T, R, \lambda$
for $k = 1, \cdots, N$ **do**
  Compute (K_t^k) for all t by using θ^k.
  for $t = 0, \cdots, T-1$ **do**
    Play $u_t^k \leftarrow K_t^k x_t^k + g_t^k, g_t^k \sim \mathcal{N}(0, \frac{1}{\sqrt{k}} I_m)$.
  end for
  Obtain θ^{k+1} from the l_2-regularized least-squares estimation.
end for

Theoretical Results

- The Least-Squares-Based Algorithm with Exploration Noise can achieve $\sqrt{N}$-regret without the identifiability assumption.

Future Direction

- Study regret bounds for LEQR in the non-episodic setting.
- Study regret bounds for online LQR with other risk measures.
- Study the regret lower bounds for online LEQR.
- Study regret bounds for LEQR with partially observable states.

Thank You!

Acknowledgement:
HK GRF No. 14201424, 14200123, 14212522, 14218022

Paper QR code:

Abbasi-Yadkori, Y. and Szepesvári, C. (2011). Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26. JMLR Workshop and Conference Proceedings.

Basei, M., Guo, X., Hu, A., and Zhang, Y. (2022). Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *The Journal of Machine Learning Research*, 23(1):8015–8048.

Cohen, A., Koren, T., and Mansour, Y. (2019). Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1300–1309. PMLR.

Jacobson, D. (1973). Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic control*, 18(2):124–131.

Mania, H., Tu, S., and Recht, B. (2019). Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32.

Simchowitz, M. and Foster, D. (2020). Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR.

Whittle, P. (1990). Risk-sensitive optimal control. *Wiley*.

Zhang, K., Zhang, X., Hu, B., and Basar, T. (2021). Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity. *Advances in Neural Information Processing Systems*, 34:2949–2964.