The Confusing Instance Principle for Online Linear Quadratic Control

Anonymous authors
Paper under double-blind review

Keywords: Model-based, linear quadratic regulator, exploration, minimum empirical divergence.

Summary

We revisit the problem of controlling linear systems with quadratic cost under unknown dynamics within model-based reinforcement learning. Traditional methods like Optimism in the Face of Uncertainty and Thompson Sampling, rooted in multi-armed bandits (MABs), face practical limitations. In contrast, we propose an alternative based on the *Confusing Instance* (CI) principle, which underpins regret lower bounds in MABs and discrete Markov Decision Processes (MDPs) and is central to the *Minimum Empirical Divergence* (MED) family of algorithms, known for their asymptotic optimality in various settings. By leveraging the structure of LQR policies along with sensitivity and stability analysis, we develop `MED-LQ`. This novel control strategy extends CI and MED principles beyond small-scale settings.

Our work addresses a crucial research gap by exploring whether the CI principle can improve exploration strategies in continuous MDPs. While the exploration-exploitation dilemma is well understood in discrete settings, the curse of dimensionality makes this challenge significantly harder in continuous spaces. `MED-LQ` overcomes these challenges by efficiently searching for confusing instances through rank-one and entry-wise perturbations while avoiding intractable confidence bounds. Benchmarks on a comprehensive control suite demonstrate that `MED-LQ` achieves competitive performance across various scenarios, establishing foundations for a fresh perspective on exploration in continuous MDPs and opening new avenues for structured exploration in complex control problems.

Contribution(s)

1. We formulate the Confusing Instance (CI) principle as an optimization problem in the LQR setting, extending this concept beyond MABs and discrete MDPs for the first time.
Context: The CI principle has previously been applied only in discrete settings, primarily in multi-armed bandits and tabular MDPs (Honda & Takemura, 2010; 2015; Pesquerel & Maillard, 2022; Balagopalan & Jun, 2024).
2. We develop `MED-LQ`, a novel control strategy that implements the Minimum Empirical Divergence (MED) framework for online LQR, and show his numerical competitiveness.
Context: Prior work established MED algorithms in discrete MDPs settings, with `IMED-RL` for ergodic case (Pesquerel & Maillard, 2022) and `IMED-KD` for the communicating case (Saber et al., 2024).
3. We develop a novel computational approach for building confusing instances in continuous systems through sensitivity analysis of rank-one perturbations.
Context: Prior work limited confusing instances to discrete settings and linear bandits. Our sensitivity analysis for continuous control systems represents the first extension of this principle to linear dynamical systems.
4. We introduce `linquax`, a library for efficient research in online LQR problems, built with JAX to leverage automatic differentiation and provide GPU/TPU compatibility.
Context: Prior to our work, no *modern* open-source library existed specifically for online LQR, creating a significant barrier to reproducible research in this domain.

The Confusing Instance Principle for Online Linear Quadratic Control

Anonymous authors
Paper under double-blind review

Abstract

We revisit the problem of controlling linear systems with quadratic cost under unknown dynamics with model-based reinforcement learning. Traditional methods like Optimism in the Face of Uncertainty and Thompson Sampling, rooted in multi-armed bandits (MABs), face practical limitations. In contrast, we propose an alternative based on the *Confusing Instance* (CI) principle, which underpins regret lower bounds in MABs and discrete Markov Decision Processes (MDPs) and is central to the *Minimum Empirical Divergence* (MED) family of algorithms, known for their asymptotic optimality in various settings. By leveraging the structure of LQR policies along with sensitivity and stability analysis, we develop `MED-LQ`. This novel control strategy extends the principles of CI and MED beyond small-scale settings. Our benchmarks on a comprehensive control suite demonstrate that `MED-LQ` achieves competitive performance in various scenarios while highlighting its potential for broader applications in large-scale MDPs.

1 Introduction

In Reinforcement Learning (RL), the exploration-exploitation dilemma is well understood in small-scale settings like multi-armed bandits (MABs) and discrete Markov Decision Processes (MDPs), for which strong theoretical guarantees exist. The curse of dimensionality impacts this dilemma in continuous or high-dimensional spaces, where analyzing this trade-off becomes significantly harder, and traditional exploration strategies struggle to scale. This is evident in deep RL, which, despite its empirical success, e.g. Osband et al. (2016); Bellemare et al. (2016); Burda et al. (2018); Sekar et al. (2020); Ladosz et al. (2022), often lacks theoretical foundations. In this work, we study the exploration-exploitation dilemma in the online *Linear Quadratic Regulator* (LQR) problem where dynamics are *unknown*, in the same setting of Abbasi-Yadkori & Szepesvári (2011). Widely used in control applications such as robotics and autonomous systems, LQR enables explicit analysis in continuous, structured MDPs (Cohen et al., 2018; Tu & Recht, 2018; 2019; Maran et al., 2025).

Research gap. Traditional exploration strategies, such as *Optimism in the Face of Uncertainty* (OFU), have been widely applied to LQR and beyond, providing upper regret bounds that evaluate the worst-case performance of a learner, typically scaling as $\widetilde{O}(\sqrt{T})$, but suffer from inherent limitations (Lattimore & Szepesvari, 2017). On the other hand, lower regret bounds establish fundamental performance limits for any learner on a given problem instance. A key tool in deriving these bounds is the *Confusing Instance* (CI) principle, which constructs hard-to-distinguish problem instances that directly appear in regret lower bound analysis. The *Minimum Empirical Divergence* (MED) family of algorithms is explicitly designed to match these regret lower bounds, leveraging the CI principle to guide exploration efficiently. MED-based methods achieve asymptotic and instance-dependent optimality, often outperforming numerically OFU-based approaches in various settings. Although characterizing regret lower bounds beyond discrete MDPs remains an open research problem and not in the scope of this work, we provide empirical evidence to address the following question, *Can the Confusing Instance principle improve exploration strategies in continuous MDPs?*

To the best of our knowledge, this paper is the first to explore the potential of the CI principle in continuous MDPs, through the online LQR setting as an entry point which presents both simplifications and challenges. This work paves the way for novel exploration strategies in large spaces.

From MABs to large MDPs. RL exploration strategies generally follow a similar evolution. Initially, an idea emerges in discrete MABs. This idea is then extended to linear bandits in parallel with discrete MDPs. The concepts are then applied to continuous MDPs, typically in the LQR setting. Finally, heuristics are developed to tackle high-dimensional problems in deep RL. The evolution of the OFU principle begins with the Upper Confidence Bounds (`UCB`) algorithm in MABs (Auer et al., 2002), followed by `OFUL` in the linear case (Abbasi-Yadkori et al., 2011). It then extends to `UCRL` in discrete MDPs (Auer & Ortner, 2006; Auer et al., 2008; Bourel et al., 2020), `OFULQ` in LQR (Abbasi-Yadkori & Szepesvári, 2011; Abeille & Lazaric, 2020; Lale et al., 2022; Mete et al., 2022), and finally, in deep RL (Bellemare et al., 2016; Curi et al., 2020). Thompson Sampling (`TS`) emerged as a more efficient alternative to the OFU principle, relying implicitly on confidence bounds, allowing for analysis similar to OFU. It started with MABs (Thompson, 1933; Kaufmann et al., 2012), then extended to linear MABs with `LinTS` (Agrawal & Goyal, 2013; Abeille & Lazaric, 2017a), discrete MDPs with `PSRL` (Osband et al., 2013; Osband & Van Roy, 2017), in LQR (Abeille & Lazaric, 2017b; 2018; Kargin et al., 2022), and finally to deep RL through Bayesian or ensemble neural networks (Osband et al., 2016; Azizzadenesheli et al., 2018). The MED principle[1] has seen more recent developments, with its foundation rooted in the regret lower bounds introduced by Lai & Robbins (1985) and Burnetas & Katehakis (1996; 1997). First proposed by Honda & Takemura (2010; 2011), the MED principle has been applied to various MABs settings (Honda & Takemura, 2015; Saber et al., 2021; Pesquerel et al., 2021; Bian & Jun, 2022; Saber & Maillard, 2024), and to linear MABs (Bian & Tan, 2024; Balagopalan & Jun, 2024). In discrete MDPs, `IMED-RL` (Pesquerel & Maillard, 2022) emerges as a state-of-the-art algorithm under ergodic assumptions. In communicating MDPs, novel promising strategies explore the MED principle but face the NP-hard challenge of finding CIs (Saber et al., 2024; Boone & Maillard, 2025). In our paper, we propose to continue the evolution of MED by extending it beyond MABs and discrete MDPs.

Outline and contributions. Our paper makes several key contributions to RL for unknown LQ systems. After formalizing the problem setup in Section 2, we present a novel formulation of CIs as an optimization problem in Section 3, developing an efficient solution method specifically engineered for LQR. Section 4 introduces our main algorithmic contribution, `MED-LQ`, which leverages these CIs to enable principled exploration while maintaining computational tractability through careful sensitivity and stability analysis. In Section 5, we present comprehensive empirical evaluations across both classical control benchmarks and industrial applications, demonstrating that `MED-LQ` matches state-of-the-art performance while overcoming the practical limitations of OFU approaches. Our work bridges an important gap between theoretical optimality and practical implementation in continuous control settings, with broader implications for exploration in large-scale MDPs.

2 Setup and Background material

The optimal control problem. Consider a linear time-invariant system written in state-space form, where the state $x_t \in \mathbb{R}^d$ evolves according to the discrete-time dynamics (Bertsekas, 2012)

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

upon receiving control $u_t \in \mathbb{R}^k$, where the system matrices $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{d\times k}$ govern the dynamics of the system, and $w_t \sim \mathcal{N}(0, \Omega)$ represents an i.i.d. centered Gaussian noise with known covariance Ω. We further assume that $\Omega = \sigma_w^2 I_d$. The quadratic cost associated to this control is $c(x_t, u_t) = x_t^\intercal Q x_t + u_t^\intercal R u_t$, where $Q \in \mathbb{R}^{d\times d}$ and $R \in \mathbb{R}^{k\times k}$ are positive definite matrices. For

[1] Baudry et al. (2023b) shows that `MED` and `TS` can be analyzed following a common methodology.

the rest of the paper, we summarize the system's unknown parameters in $\Theta = (A, B)^\intercal$. The infinite horizon average cost function for a policy π specifying the control u in each state x is

$$J_\pi(\Theta) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} c(x_t, u_t)\right]. \tag{2}$$

Further, a policy π is classically parameterized by a *gain* matrix $K \in \mathbb{R}^{k\times d}$ as $\pi(x_t) = -Kx_t$, making it a linear function of the state, with associated cost (2) denoted $J_K(\Theta)$. Optimal planning can be achieved by solving the Discrete Algebraic Ricatti Equation (DARE), $P = A^\intercal PA + Q - A^\intercal PB\left(B^\intercal PB + R\right)^{-1} B^\intercal PA$. We denote the solution of the DARE, $P^\star(\Theta)$, and the optimal gain that minimizes Eq. (2) is given as $K^\star(\Theta) = -\left(B^\intercal P^\star(\Theta)B + R\right)^{-1} B^\intercal P^\star(\Theta)A$, which achieves the minimal cost $J^\star(\Theta) = J_{K^\star(\Theta)}(\Theta)$. When Θ is clear from context, we simply write $P^\star$, $K^\star$, $J^\star$.

The learning problem. We follow the model-based RL setting of Abbasi-Yadkori & Szepesvári (2011), where parameter $\Theta^\star$ is unknown and Q and R are assumed known. We assume that the system is part of the *stabilizable* set $\mathcal{S}_0$, meaning there exists a gain matrix K such that $A - BK$ is stable, that is with all eigenvalues confined to the interval $(-1, 1)$. It is convenient to introduce the constraint set $\mathcal{S} \subseteq \mathcal{S}_0 = \{\Theta \in \mathbb{R}^{(k+d)\times d} : J^\star(\Theta) \leq D, \mathrm{Tr}(\Theta\Theta^\intercal) \leq S^2\}$. At each time t the learner chooses a policy π_t, observes the current state x_t, executes a control $u_t = \pi_t(x_t)$ and incurs the associated cost $c_t = x_t^\intercal Qx_t + u_t^\intercal Ru_t$; the system then transitions to the next state x_{t+1}. The learning performance is measured by the cumulative regret over T steps defined as $\mathcal{R}(T) = \sum_{t=0}^{T}(c_t - J^\star(\Theta^\star))$. The unknown parameter $\Theta^\star$ can be directly estimated from sequences $\{x_t, u_t, x_{t+1}\}$ using regularized least-squares (RLS). Let $z_t = (x_t, u_t)^\intercal$, for any regularization paramameter $\lambda \in \mathbb{R}^+$, the design matrix and the RLS estimate are defined as

$$V_t = \lambda I + \sum_{s=0}^{t-1} z_s z_s^\intercal, \tag{3}$$

$$\widehat{\Theta}_t = V_t \sum_{s=0}^{t-1} z_s x_{s+1}^\intercal. \tag{4}$$

Using Theorem 1 from Abbasi-Yadkori & Szepesvári (2011), for any $\theta \in (0, 1)$, for all $0 \leq t \leq T$, the underlying parameter Θ lives in the ellipsoid $\mathcal{E}_t(\delta)$ with probability at least $1 - \delta$ where

$$\mathcal{E}_t(\delta) = \left\{\Theta^\star \in \mathcal{S} : \|\Theta^\star - \widehat{\Theta}_t\|_{V_t} \leq \beta_t(\delta)\right\}, \text{ with } \beta_t(\delta) = n\sigma_w\sqrt{2\log\left(\frac{\det(V_t)^{1/2}}{\det(\lambda I)^{1/2}\delta}\right)} + \lambda^{1/2}S.$$

Policy evaluation. From the form of the policies, it is convenient to introduce $A_K = A - BK$, known as the closed-loop system of K. Indeed using this notation, transitions under policy K rewrite $x_{t+1} = A_K x_t + w_t$, and the discrete-time Bellman equation writes $P_K(\Theta) = Q_K + A_K^\intercal P_K(\Theta)A_K$, where $Q_K = Q + K^\intercal RK$ and $P_K(\Theta)$ is the solution to a discrete-time Lyapunov equation. We denote the *spectral radius* of a matrix M as $\rho(M)$. If K stabilizes the system, then $\rho(A_K) < 1$, the cost of K is finite, and $x_t \to 0$ at a geometric rate. Under the objective (2), for a gain K, a better gain K' ensures $J_{K'}(\Theta) \leq J_K(\Theta)$, with $J_K(\Theta) = \sigma_w^2\,\mathrm{Tr}(P_K(\Theta))$, the average cost of K in Θ.

Optimal MABs strategies. The Minimum Empirical Divergence (`MED`) algorithm, introduced by Honda & Takemura (2010), achieves asymptotic optimality for MABs. `MED` derives directly from the fundamental regret lower bound established by Burnetas & Katehakis (1996), which states that for any suboptimal arm $a \in \mathcal{A}$ (where $\mu_a < \mu_\star$, with $\mu_\star$ being the optimal mean), the expected number of pulls $N_a(T)$ must satisfy: $\liminf_{T\to\infty} \mathbb{E}[N_a(T)]/\log T \geq 1/\mathcal{K}_a(b_a, \mu_\star)$. Here, $b_a \in \mathcal{D}_a$ represents the reward distribution of arm a, and $\mathcal{K}_a(b_a, \mu_\star)$ captures the minimum information cost needed to confuse the algorithm between arm a and a better arm. This is formalized as $\mathcal{K}_a(b_a, \mu_\star) = \inf\left\{\mathrm{KL}(b_a\|b) : b \in \mathcal{D}_a,\ \mathbb{E}_{X\sim b}[X] > \mu_\star\right\}$, where KL denotes the Kullback-Leibler divergence. At each time step t, `MED` elegantly transforms this information-theoretic principle into an exploration strategy by sampling arm a with probability proportional to $\exp(-N_a(t)\mathcal{K}_a(\hat{b}_a, \hat{\mu}_\star))$, where notation with ˆ denotes empirical estimates. The cornerstone of the `MED` framework is identifying the *confusing instance*, the alternative model that minimizes the KL divergence while appearing more rewarding than the currently best arm. In the following section, we extend this powerful concept to the substantially more complex setting of LQR.

3 Efficient Confusing Instance Search for LQR

In this section, we now discuss the main insight of our contribution, borrowing the notion of confusing instances originating from MAB theory to the LQR framework.

Intuition. The central element revealing the structure of a sequential decision problem appears when deriving lower bounds on the regret performance of any consistent learner, namely a learner able to achieve optimality on a class of decision problems $\mathbb{M}$ rather than a single instance $\mathbf{M} \in \mathbb{M}$. The high-level idea is easy to get, and consists of considering, for a given $\mathbf{M} \in \mathbb{M}$ a policy π that isn't optimal in $\mathbf{M}$, hence does not achieve minimal cost $J_\star(\mathbf{M})$, where here $\star$ is optimal in $\mathbf{M}$. We then want to build another MDP $\widetilde{\mathbb{M}}$ in which π achieves better gain, that is $J_\pi(\widetilde{\mathbf{M}}) \leq J_\star(\mathbf{M})$. Given the multitude of possible MDPs satisfying these conditions, we naturally seek those informationally closest to our initial estimate $\mathbf{M}$. More precisely, the rationale is that if $\mathbf{M}$ and $\widetilde{\mathbf{M}}$ are hard to distinguish from playing optimally in $\mathbf{M}$, say, from a hypothesis-testing perspective, then any learner that must be optimal in both environments should deviate from playing $\star$.

Formally, let $\Pi^\star(\mathbf{M}) = \{\pi \in \Pi : J_\pi(\mathbf{M}) \leq J_{\pi'}(\mathbf{M}) \forall \pi' \in \Pi\}$ denote optimal policies for $\mathbf{M}$, and alternative models as $\mathrm{Alt}(\mathbf{M}) = \{\widetilde{\mathbf{M}} \in \mathbb{M} : \Pi^\star(\mathbf{M}) \cap \Pi^\star(\widetilde{\mathbf{M}}) = \emptyset\}$. Introducing $d(\mathbf{M}, \widetilde{\mathbf{M}})$ to be e.g. the expected log-likelihood ratio of a trajectory generated from $\Pi^\star(\mathbf{M})$ in both models, we then look for $\widetilde{\mathbf{M}} \in \mathrm{Alt}(\mathbf{M})$ minimizing $d(\mathbf{M}, \widetilde{\mathbf{M}})$. Such an instance is called confusing or model $\mathbf{M}$.

Specializing this approach to LQ systems introduces both simplifications and challenges. Interestingly, given $\mathbf{M}(\Theta)$, $\Pi^\star(\mathbf{M})$ reduces to $\pi_{K^\star}$, hence we can consider the expected log-likelihood ratio along the trajectory from $K^\star$ in both systems. Note that $K^\star$ must stabilize both systems.

Proposition 1 (Asymptotic per-step expected log-likelihood ratio for LQR)**.** *Given a gain K that is stabilizing for the two systems Θ and $\widetilde{\Theta}$, and assuming both systems share the same covariance matrix Ω, the asymptotic per-step expected log-likelihood under the two systems is*

$$\mathbf{d}_K(\Theta \| \widetilde{\Theta}) \overset{def}{=} \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_\Theta \left[\log \frac{\mathbf{p}(\tau_T)}{\tilde{\mathbf{p}}(\tau_T)}\right] = \frac{1}{2} \operatorname{Tr}\left((A_K - \widetilde{A}_K)^\intercal \Omega^{-1} (A_K - \widetilde{A}_K) \Sigma_K(\Theta)\right). \tag{5}$$

where τ_T denotes a trajectory of length T from π_K and the stationary distribution $\Sigma_K(\Theta)$ induced by K satisfies a discrete-time Lyapunov equation $\Sigma_K(\Theta) = \mathbb{E}_\Theta [x_t^\intercal x_t | K] = \Omega + A_K \Sigma_K(\Theta) A_K^\intercal$.

The proof of this proposition is given in Appendix A.1. We now have the necessary elements to tackle the challenge of identifying the most confusing instances in LQR.

3.1 The Challenge of Approaching the Most Confusing Instance

Finding the most confusing instance and its associated sub-optimality cost is generally NP-hard. This section introduces key simplifications that yield a computationally efficient approximation.

At a high level, rather than optimizing $\mathbf{d}_K(\Theta \| \widetilde{\Theta})$ over all possible confusing $\tilde{\Theta}$, we will proceed in Section 4 by sampling a finite set of perturbations $\Theta'_1, \ldots, \Theta'_n$ around a base configuration Θ and then optimize within the convex hull of these anchor points. To justify our approach, we analyze an optimization concerning a single perturbation parameter Θ' of the system. Thanks to the explicit form of optimal policies in LQR, the optimization problem can be formulated as

$$\underline{\mathbf{K}}(\Theta \| \Theta') = \inf_{\widetilde{\Theta}} \{\mathbf{d}_K(\Theta, \widetilde{\Theta}) \quad \text{subject to} \quad J_{K'}(\widetilde{\Theta}) < J_K(\widetilde{\Theta})\}, \tag{6}$$

where we consider two *close* stabilizable instances Θ and Θ', with their respective optimal gains $K = K^\star(\Theta)$ and $K' = K^\star(\Theta')$. This objective function is strictly convex in $\widetilde{\Theta}$ for fixed Θ. However, the constraint is non-convex, as the set of stable matrices is generally non-convex. To search for solutions that are both stable with controlled cost, we first observe that as each stabilizing LQ system is guaranteed to have a unique optimal gain that minimizes the associated cost, the cost of

K' cannot exceed that of K in Θ', ensuring that $J_{K'}(\Theta') \leq J_K(\Theta')$. (In particular, Θ' is a feasible solution of the optimization problem defined in Equation (6) and, we get the crude upper bound $\underline{\mathbf{K}}(\Theta\|\Theta') \leq \mathbf{d}_K(\Theta, \Theta')$.) From the preceding initial remark, we thus observed that $J_{K'}(\Theta') - J_K(\Theta') < 0$, while $J_{K'}(\Theta) - J_K(\Theta) > 0$. This justifies performing a line search interpolating between Θ and Θ', effectively reducing the optimization to a one-dimensional search problem, and yielding a reduced upper bound on the sub-optimality cost. More formally, we introduce the analytic curve connecting these instances, parametrized by $\alpha \in [0, 1]$ and expressed as $\Theta(\alpha) = (A + \alpha\Delta_A, B + \alpha\Delta_B)$ with $\Delta_A = A' - A$ and $\Delta_B = B' - B$. We then form the following key result.

Proposition 2 (Sub-optimality cost refinement)**.** *Using the linear interpolation parametrization, a valid upper-bound on* $\underline{\mathbf{K}}(\Theta\|\Theta')$ *can be obtained by finding the root of*

$$\mathcal{L}(\alpha) = J_{K'}(\Theta(\alpha)) - J_K(\Theta(\alpha)) = 0. \tag{7}$$

Proof. Assuming Θ and Θ' yield different dynamics, $\mathcal{L}(\alpha)$ is a continuous function in $[0, 1]$. Now, by definition $\mathcal{L}(0) \cdot \mathcal{L}(1) = (J_{K'}(\Theta) - J_K(\Theta)) \cdot (J_{K'}(\Theta') - J_K(\Theta')) < 0$, because $J_{K'}(\Theta) - J_K(\Theta) > 0$ and $J_{K'}(\Theta') - J_K(\Theta') < 0$. This implies that a root exists according to Bolzano's theorem. We can show that $L(\alpha)$ is not convex, but has no local optima, which allows global convergence, as demonstrated in Section 3 of Fazel et al. (2018). The objective function increases monotonically as $\widetilde{\Theta}$ diverges from Θ since its derivative equals t times the trace of positive definite matrices' product, ensuring positivity for all $t > 0$. Thus, finding the unique root that satisfies the cost constraint is the solution of Eq. (6), on the linear curve. $\square$

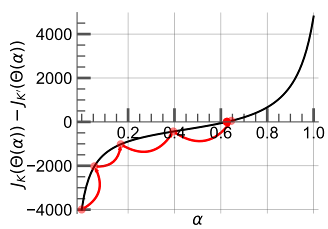


Figure 1: Optimization landscape of the objective $\mathcal{L}(\alpha)$ on two Inverted Pendulum system (Barto et al., 1983) parametrized by the mass and the length of the pendulum. Θ is parametrized by (.1, .4) and Θ' by (.3, 1.). Red arrows indicate the Newton steps taken during the optimization process.

This objective is non-convex due to the potential non-stability of the interpolated closed-loop system (Lin & Antsaklis, 2009), at the boundary between stable and unstable policies, the objective function quickly becomes infinity. However, $\mathcal{L}(\alpha)$ is *almost* smooth (see Lemma 6 in Fazel et al. (2018)) when the closed-loop systems are not close to the boundary of stability, that allows in practice, to deploy the Newton method, that can solve the objective in few steps, as shown in the Figure 1.

3.2 Fast Approximate Solution for Small Perturbed Systems

To find the root of (7), we introduce a Taylor approximation[2] of the objective function. For sufficiently small perturbations Θ' around Θ, the interpolation between closed-loop systems $A(\alpha) - B(\alpha)K$ and $A(\alpha) - B(\alpha)K'$ remains stable, thanks to the existence of a stability radius (Hinrichsen & Pritchard, 1986). This stability property, supported by perturbation theory, allows us to employ a first-order Taylor expansion to derive a closed-form approximation of Eq. (7).

Proposition 3 (Sub-optimality cost refinement under small perturbations)**.** *Assume the closed-loop system undergoes perturbations* Δ_A *and* Δ_B *that are sufficiently small (e.g.,* $\|\Delta_A\|, \|\Delta_B\| \leq \epsilon$ *for a small* $\epsilon > 0$*) so that higher-order terms can be neglected, we denote* $\Delta_K = \Delta_A + \Delta_B K$ *the perturbation on the closed loop system, and the objective* $\mathcal{L}(\alpha)$ *can be approximated with*

$$\mathcal{L}(\alpha) \approx (p_K - p_{K'}) - \alpha(\overline{p}_K - \overline{p}_{K'}) + \alpha^2(\overline{\overline{p}}_K - \overline{\overline{p}}_{K'}), \tag{8}$$

with $p_K = \mathrm{Tr}(P_K(\Theta))$, $\overline{p}_K = \mathrm{Tr}(\overline{P}_K(\Theta))$, $\overline{\overline{p}}_K = \mathrm{Tr}(\overline{\overline{P}}_K(\Theta))$, *and* $\overline{P}_K(\Theta) = A_K^\intercal \overline{P}_K(\Theta) A_K + A_K^\intercal P_K(\Theta)\Delta_K + \Delta_K^\intercal P_K(\Theta) A_K$, $\overline{\overline{P}}_K(\Theta) = A_K^\intercal \overline{\overline{P}}_K(\Theta) A_K + \Delta_K^\intercal P_K(\Theta)\Delta_K$. *This is a second-degree polynomial whose coefficients correspond to the trace of the solution of discrete-time Lyapunov equations. The solution can be obtained by identifying the positive root.*

The proof of this proposition is in A.2. We now have all the elements to design an efficient algorithm.

[2] Taylor approximation for finding confusing instance has already been explored by Baudry et al. (2023a) in MABs.

4 Towards Minimum Empirical Divergence Strategies for Online LQR

In this section, we introduce `MED-LQ`, our novel algorithm that extends the asymptotically optimal `MED` strategy of Honda & Takemura (2011) from MABs to LQ systems. Our approach incorporates several adaptations specifically crafted to address the unique challenges of continuous dynamics.

4.1 The `MED-LQ` Algorithm

Algorithm 1: `MED-LQ`: Minimum Empirical Divergence for Linear Quadratic Systems

Input: Q, R, $\widehat{\Theta}_0$, $V_0 = \lambda I$, $\delta > 0$, T, n, σ_η, σ_v, ϵ.

1 **for** $t = 0, \ldots, T$ **do**
2   **if** $\det(V_t) > 2\det(V_0)$ **then**
3   Compute $\widehat{\Theta}_t$ via RLS (4) and set $\widehat{K}_t = K(\widehat{\Theta}_t)$;
4   Generate n perturbations $\left\{W_i = \eta_i\, e_j e_k^\top \mid j, k \sim \mathcal{U}(\{1, \ldots, n\}),\ \eta_i \sim \mathcal{U}(-\sigma_\eta, \sigma_\eta)\right\}$;
5   Form the candidate sets $\{\overline{\Theta}_i = \widehat{\Theta}_t + W_i\}$ and $\{K_i = K(\Theta_i)\}$;
6   Define the mask $m_i = m(\overline{\Theta}_i, \widehat{\Theta}_t; \epsilon) \in \{0, 1\}$ (9);
7   For each candidate with $m_i = 1$, compute the $h_i = \mathbf{H}(\widehat{\Theta}_t \parallel \overline{\Theta}_i; V_t)$ (10);
8   Set $\widetilde{\Theta}_t = \widehat{\Theta}_t + \sum_{i=1}^n \omega_i\, W_i$ with $\omega_i = \exp(h_i) m_i \Big/ \sum_{j=1}^n \exp(h_j) m_j$ and $V_0 = V_t$;
9   **else**
10   Set $\widetilde{\Theta}_t = \widetilde{\Theta}_{t-1}$;
11   **end**
12   Compute the optimal empirical gain $\widetilde{K}_t = K(\widetilde{\Theta}_t)$;
13   Apply $u_t = \widetilde{K}_t x_t$ **if** ($\widetilde{K}_t$ stabilize $\widetilde{\Theta}_t$) **else** $\widetilde{K}_t x_t + \nu_t$, with $\nu_t \sim \mathcal{N}(0, \sigma_\nu^2)$;
14   Obtain x_{t+1} and record (z_t, x_{t+1}) and update $V_{t+1} = V_t + z_t z_t^\top$;
15 **end**

`MED-LQ` is an online learning algorithm that carefully balances exploration and exploitation in linear dynamical systems. Inspired by the standard learning framework of Abbasi-Yadkori & Szepesvári (2011), the algorithm proceeds in rounds over a finite horizon. At each time step t, it first checks whether the accumulated information, quantified by the determinant of the design matrix $\det(V_t)$ has doubled (line 2). When it does, a new optimal empirical parameter $\widehat{\Theta}_t$ is computed using RLS, and the corresponding control gain is derived $\widehat{K}_t = K(\widehat{\Theta}_t)$. To enhance exploration, `MED-LQ` generates a collection of n candidate parameters $\forall i \in \{0, \cdots, n\}$, $\overline{\Theta}_i = \widehat{\Theta}_t + W_i$ by applying random rank-one perturbations W_i to the RLS estimate (line 4,5). Rank-one perturbations simplify the stability analysis, making it tractable (Laffey et al., 2002), and are inspired by the local-policy search from (Pesquerel et al., 2021). Each candidate is then filtered through a set of constraints (line 6), to ensure that the most confusing instance search (6) is well-defined. The search for the most confusing instance is well-defined when the following constraints defined by $m(\overline{\Theta}, \widehat{\Theta}; \epsilon)$ hold

$$\mathbb{I}\left\{\underbrace{\rho(\widehat{A}_{\widehat{K}}) < 1 \wedge \rho(\overline{A}_{\overline{K}}) < 1}_{\text{Closed-loop stability.}} \wedge \underbrace{\widehat{A}_{\widehat{K}}\widehat{A}_{\overline{K}} \succeq 0 \wedge \overline{A}_{\widehat{K}}\overline{A}_{\overline{K}} \succeq 0}_{\text{Linear interpolation stability.}} \wedge \underbrace{J_{\widehat{K}}(\widehat{\Theta}) - J_{\overline{K}}(\widehat{\Theta}) > \epsilon}_{\text{Alternative set membership.}}\right\}, \tag{9}$$

where $X \succeq 0$ denotes positive semi-definiteness, and ϵ is a small threshold value. The first two conditions ensure closed-loop stability. The next two follow from Theorem 1 of Laffey et al. (2002), and check that the linear curve between the two closed-loop systems is stable. The last condition checks if $\overline{\Theta}$ belongs to the alternative set of $\widehat{\Theta}$. The linear interpolation stability condition, enabled by our rank-one perturbations, represents a conservative approach. While ensuring stability across the entire interpolation interval exceeds technical requirements, removing this constraint would necessitate computing confusing costs for more instances and implementing careful post-filtering mechanisms. We recommend this filtering criterion for computational efficiency, especially in systems with small to moderate dimensions. For those candidates that pass the stability check, the algorithm evaluates their *Minimum Empirical Divergence* (line 8), which captures the cost of making a perturbed system optimal. This quantity is inspired by MED and LinMED strategies.

Definition 4 (Minimum Empirical Divergence coefficients for LQR)**.** *During the learning process, where V_t represent the design matrix at time t, $\widehat{\Theta}_t$ the empirical optimal RLS estimate and Θ an alternative parameter, the minimum empirical divergence is given by*

$$\mathbf{H}_t(\Theta) = -\frac{\mathbf{K}(\widehat{\Theta}_t \| \Theta)}{\|\Theta\|^2_{V_t^{-1}}}. \tag{10}$$

MED-LQ generates exponential weights (line 8) to create a weighted combination of perturbations, biasing parameter estimates toward candidates with lower divergence values. Finally, the corresponding control gain is applied to the system. We introduce additional isotropic exploration noise ν_t, similarly to Tu & Recht (2019); Lale et al. (2022); Kargin et al. (2022), when the empirical gain fails to stabilize the empirical estimate, which intuitively happens mainly in the early rounds. This noise provides excitation, ensuring the identifiability of the system dynamics by exploring the state-space in all directions. Finally, new state data is collected to update the design matrix, thus refining the parameter estimates over time. The full algorithm is summarized in Algorithm 1.

4.2 Intuition and design elements

Let us now provide insights and sketch the main ideas supporting the soundness of this strategy.

MED-LQ extends the asymptotically optimal IMED-RL algorithm (Pesquerel & Maillard, 2022) for ergodic discrete MDPs to the LQR setting while incorporating continuous aspects developed in LinMED (Balagopalan & Jun, 2024) for linear sub-Gaussian MABs. Both methods leverage regret lower bounds to achieve superior efficiency compared to OFU-based approaches, with IMED being the deterministic counterpart of MED.

Ergodicity and information gain. In IMED-RL, ergodicity ensures that every policy eventually visits all states, enabling efficient information gathering across the state space. For linear dynamical systems, the situation is comparable: observing a single state can provide global insights about system dynamics, similar to the information transfer in linear bandits. However, since quantifying the per-step information gain is challenging, we execute each chosen policy for multiple steps until a significant change in information volume occurs (line 2).

Policy improvement. A cornerstone of IMED-RL is exploiting the policy improvement property from Puterman (2014), which guarantees that in discrete ergodic MDPs, any sub-optimal policy can be improved through a *local* (single-state) modification, a convenient property not universally applicable. This approach efficiently identifies confusing instances by searching only over local policy modifications, with central analysis demonstrating a high probability of policy improvement. For linear-quadratic systems, we identify single entry-wise perturbations of the system matrix as the natural equivalent to single-state modifications. This approach yields substantial computational benefits, as candidate perturbations become straightforward to generate. However, rather than directly applying single-entry perturbations, which alone may be insufficient to guarantee policy improvement, we form convex combinations of candidates weighted by MED coefficients. This strategic convex combination substantially expands the search space volume, significantly increasing the probability of discovering effective policy improvements.

Policy gradient. While IMED-RL estimates an empirical MDP, applies value iteration, and selects actions minimizing the IMED index, MED-LQ follows a parallel approach. We estimate system parameters via RLS, solve the DARE to capture the value function, and define our minimum empirical divergence analogously to IMED-RL. Inspired by LinMED, the term $1/\|\Theta\|^2_{V_t^{-1}}$ effectively functions as a visitation count analog. Conceptually, where IMED-RL implements policy iteration, MED-LQ adopts an approximate policy gradient approach. The fundamental intuition is that a policy's selection likelihood should at least match its posterior probability of optimality, with perturbations directing exploration toward promising regions of the policy space.

Continuum policy and ϵ-optimality. The policy improvement lemma from Puterman (2014) applies to discrete, ergodic MDPs, where finitely many policies ensure that a finite number of improvement steps reach optimality. This property doesn't extend to continuous MDPs with infinite policy sets. By introducing the parameter ϵ in our filtering condition, we effectively consider ϵ-near-optimal policies rather than strictly optimal ones, implicitly covering the policy space with finitely many level sets. This approach ensures that finitely many ϵ-policy-improvement steps yield a near-optimal policy. In practice, ϵ requires careful calibration: not too small (to ensure non-empty filtered sets) and not too large (to avoid requiring excessive policy-improvement steps). We recommend ϵ that scales between $O(1/T)$ and $O(1/\log^2(T))$.

Excitation. A well-known challenge in LQR is the initial information scarcity that impedes the invertibility of matrices defining stable policies. This challenge dissipates after sufficient observations span the entire state space, after adequate system excitation. In line (13), we introduce noise ν_t to enforce excitation whenever the control fails to stabilize the confusing instance. This mechanism primarily induces additional exploration during early rounds, while in the asymptotic regime, all selected policies naturally stabilize the system, eliminating the need for artificial excitation. In our Section 5, we study the effect of excitation under the name of "auto-stabilization".

Following our insights, a rigorous regret analysis of `MED-LQ` presents unique theoretical challenges distinct from `MED`, `IMED-RL`, or `LinMED`. Two critical questions emerge: (1) establishing that `MED-LQ` guarantees high-probability policy improvements with sufficient margin at each iteration, and (2) determining the precise magnitude of entry-wise perturbations needed to ensure policy improvements exist within local neighborhoods. These challenges require adapting policy-improvement arguments to continuous settings, a non-trivial extension demanding specialized analysis beyond this paper's scope. While `MED-LQ` deliberately addresses these challenges through techniques such as combining multiple single-entry perturbations, we reserve a comprehensive theoretical analysis for future work.

5 Numerical Experiments

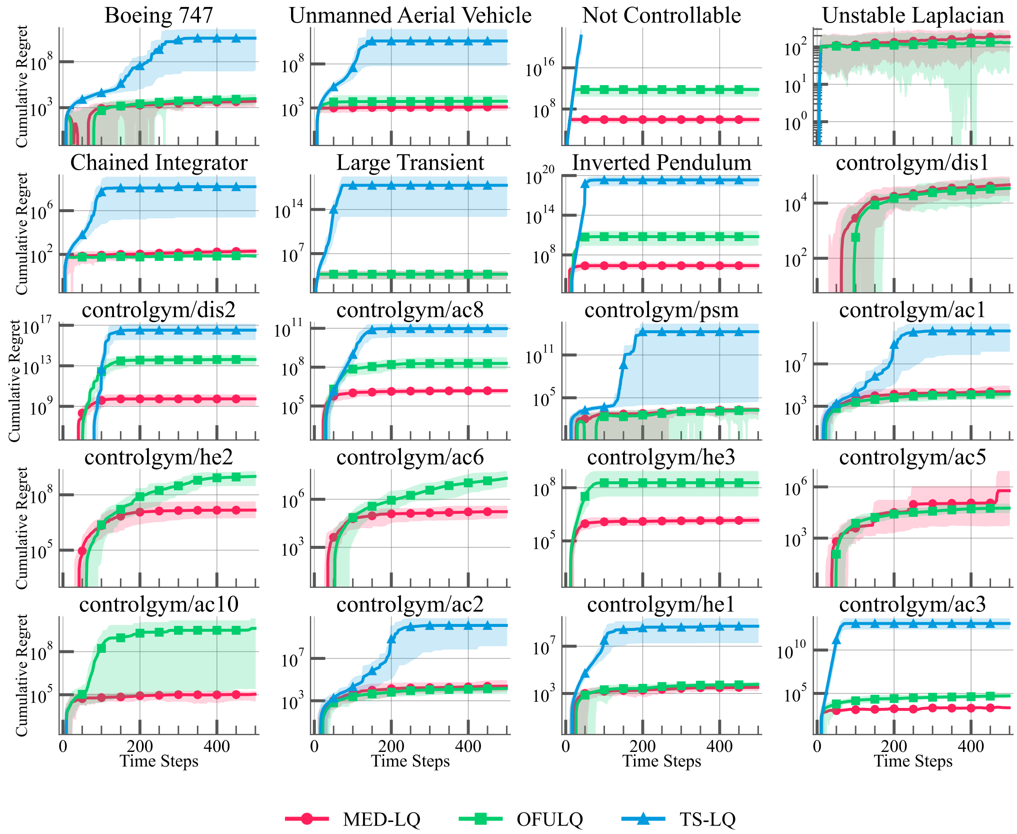
To study the numerical potential of `MED-LQ`, we evaluate it on a control suite that includes classic environments from the online LQR literature, such as the Boeing 747 and Unmanned Aerial Vehicle, as well as additional industrial control problems from `controlgym` (Zhang et al., 2023), inspired by real-world applications. All environments are subject to a normal noise $\mathcal{N}(0,1)$ and have a moderate size (from 2 to 10 dims). We assess the performance of `MED-LQ` in two distinct scenarios.

Scenario 1: Stable Initialization. In this setting, we initialize the algorithm with a stable controller and seed the dataset with a trajectory of 50 time steps. We compare `MED-LQ` against `OFULQ` (Abbasi-Yadkori & Szepesvári, 2011) and `TS-LQ` (Abeille & Lazaric, 2017b). The stable initialization allows us to assess the exploration efficiency and convergence properties when the system starts in a well-controlled regime. The results are shown in Figure 2a.

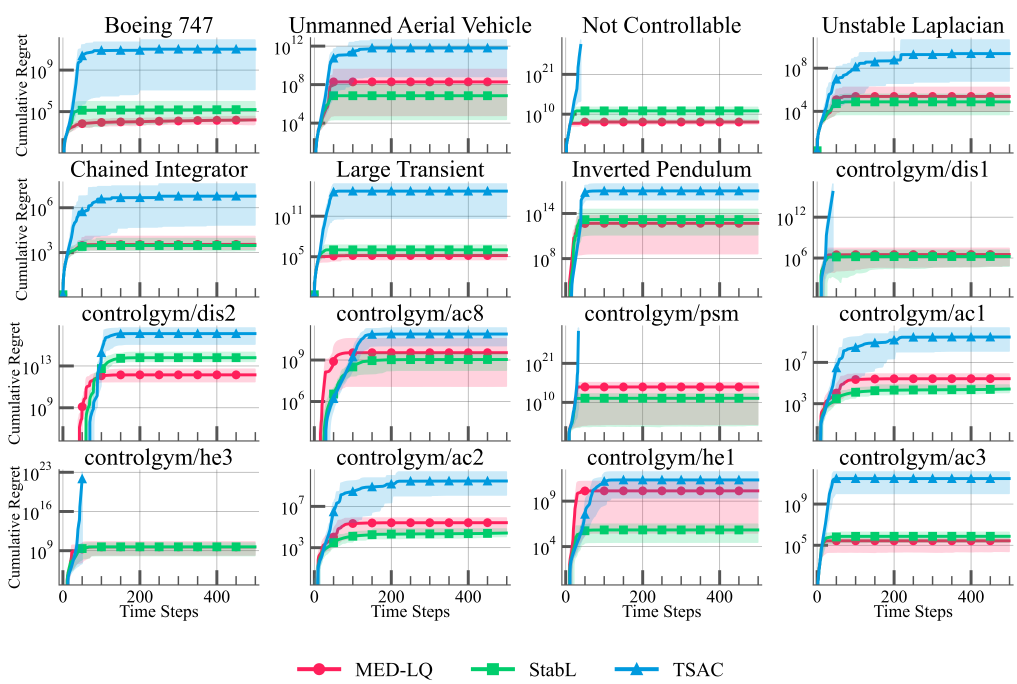
Scenario 2: Auto-Stabilization Here, `MED-LQ` is deployed with an initial parameter estimate $\widehat{\Theta}_0 = \mathbf{0}$. To facilitate auto-stabilization, the policy is executed with isotropic noise $w \sim \mathcal{N}(0,1)$ for the first 35 time steps, as in Lale et al. (2022). We compare `MED-LQ` against `StabL` (Lale et al., 2022) and `TSAC` (Kargin et al., 2022), the auto-stabilizing counterparts of `OFULQ` and `TS-LQ`, respectively. The results are shown in Figure 2b.

Implementation details. We implement all baselines within the JAX framework (Bradbury et al., 2018) using a new library, `linquax`[3], which delivers highly performant online LQR algorithms with GPU/TPU support and automatic differentiation. In our implementation, `OFULQ` and `StabL`

[3] A WIP version of the library is available in `https://anonymous.4open.science/r/linquax-4FCF/`.



(a) Comparison of `MED-LQ`, `OFULQ`, and `TS-LQ` initialized with a stable controller.



(b) Comparison of `MED-LQ`, `StabL`, and `TSAC` in the auto-stabilization scenario

Figure 2: Performance comparison of `MED-LQ` under two distinct initialization scenarios.

are optimized via projected gradient descent, while `TS-LQ` and `TSAC` employ a rejection sampling operator. In addition to the doubling trick, we enforce a minimum patience period of 10 steps to prevent excessive early updates that can lead to increased regret. All algorithms share common hyperparameters, chosen after previous work, with $\lambda = 1 \times 10^{-4}$ and $\delta = 1 \times 10^{-4}$. For `MED-LQ`, we define without hyperparameter search the number of candidates $n = 128$ and $\sigma_\eta = 1$. Experiments were conducted in less than 1 hour, on a CPU-only cluster equipped with four 64-core AMD Zen3 processors. For classic environments, we used 64 random seeds, and for `controlgym` environments, 48 seeds. Performance metrics are reported as the interquartile mean along with the 25th percentile and 75th percentile for each experiment.

Discussion of results. We compare `MED-LQ` against `OFULQ`, `TS-LQ`, `StabL`, and `TSAC`. Our experimental evaluation reveals that `MED-LQ` demonstrates strong performance across environments. With stable initialization, `MED-LQ` shows rapid convergence to low cumulative regret, validating that CI-guided exploration effectively balances exploration and exploitation. All algorithms benefit from stable initialization, allowing them to focus on policy refinement rather than basic stabilization. In zero-knowledge settings requiring auto-stabilization, `MED-LQ` quickly discovers stabilizing policies. It consistently outperforms Thompson Sampling methods, which occasionally fail to find stabilizing controllers even after 10,000 rejection sampling attempts. Compared to state-of-the-art methods `OFULQ` and `StabL`, `MED-LQ` demonstrates superior efficiency in most environments, matching `StabL`'s performance in others, with the sole exception being the `controlgym/he1` environment under auto-stabilization. These results establish `MED-LQ` as a competitive and reliable alternative to OFU-based and Thompson Sampling approaches for online LQR tasks.

Sample size study. We now examine how the sample size used in `MED-LQ` affects both regret and execution time in the Inverted Pendulum environment. Experiments were run on a NVIDIA A100 GPU. Figure 3 presents the results. The plot on the left shows that runtime remains relatively constant across different sample sizes (0.3-0.5 seconds), highlighting the parallelization capabilities of our GPU implementation. The right plot shows that increasing the sample size leads to slightly lower regret until approximately 64 samples, after which the performance plateaus. This suggests that in Inverted Pendulum 64 samples are sufficient to adequately span the space of candidate policies.

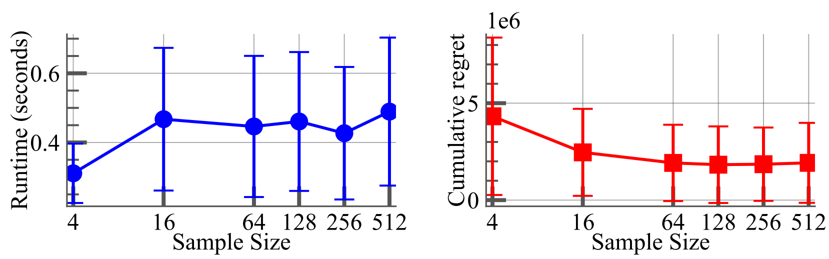


Figure 3: Study of the sample size on the Inverted Pendulum environment.

6 Conclusion

In this work, we introduced the Confusing Instance (CI) principle as a novel approach to exploration in online Linear Quadratic Control (LQR). By extending the Minimum Empirical Divergence (MED) framework beyond discrete settings, we developed `MED-LQ`, the first method to apply the confusing instance principle beyond tabular MDPs. Our approach employs strategically designed rank-one and entry-wise perturbations that enable efficient identification of confusing instances while maintaining computational feasibility. Notably, `MED-LQ` avoids confidence bounds (intractable in large spaces) and instead relies on the policy iteration framework. Our methodology is generalizable to other settings: compute empirical optimal policy, generate candidates, approximate confusing instances, compute the minimum empirical divergence, and update policy toward areas minimizing this divergence. Benchmarks demonstrate that `MED-LQ` matches state-of-the-art performance, overcoming limitations of existing methods such as OFU and TS.

We believe that the CI principle deserves greater attention as it introduces a fresh perspective on exploration in continuous MDPs. Our work establishes the foundations for this promising approach, opening new avenues for exploration strategies in complex problems.

Future Work. Future research should refine `MED-LQ`'s theoretical foundations by establishing formal regret bounds and analyzing the minimal perturbation magnitudes needed for guaranteed policy improvements. A particularly promising direction is to extend the CI principle to high-dimensional problems in deep RL, where efficient exploration remains challenging. The principles established here provide a foundation for novel exploration strategies in both continuous control and complex decision-making tasks.

Broader Impact Statement

Our work on efficient exploration in LQR systems has potential applications in robotics, autonomous vehicles, and industrial control systems. While our algorithm enables more efficient learning in these domains, it could also accelerate the deployment of autonomous systems with inherent safety considerations. We advocate for robust safety validation before deploying such learning-based controllers in critical applications.

A Proofs of the main propositions

In this section, we detail the proof of Proposition 1 that provides the form of the asymptotic per-step expected log-likelihood ratio when following a given policy with control K. We then detail the proof of Proposition 2 which provides an approximation of the cost function to be optimized in the regime of small perturbations, which yields a closed-form approximate solution.

A.1 Asymptotic per-step expected log-likelihood ratio for LQR

Proof of Proposition 1. The one-step likelihood of observing x_{t+1} given x_t under Θ (ignoring constants) is $\mathbf{p}(x_{t+1}|x_t) \propto \exp\left(-\frac{1}{2}(x_{t+1} - A_K x_t)^\intercal \Omega^{-1}(x_{t+1} - A_K x_t)\right)$. We denote by $\tilde{\mathbf{p}}$ the transition probability under $\widetilde{\Theta}$. Thus the one-step likelihood ratio is

$$\begin{aligned}
\ell_t &= \log \frac{\mathbf{p}(x_{t+1}|x_t)}{\tilde{\mathbf{p}}(x_{t+1}|x_t)} \\
&= \frac{1}{2}\left((x_{t+1} - \widetilde{A}_K x_t)^\intercal \Omega^{-1}(x_{t+1} - \widetilde{A}_K x_t) - (x_{t+1} - A_K x_t)^\intercal \Omega^{-1}(x_{t+1} - A_K x_t)\right) \\
&= \frac{1}{2}\left(\left((A_K - \widetilde{A}_K)x_t + w_t\right)^\intercal \Omega^{-1}\left((A_K - \widetilde{A}_K)x_t + w_t\right) - w_t^\intercal \Omega^{-1} w_t\right) \\
&= \frac{1}{2}\left(x_t^\intercal (A_K - \widetilde{A}_K)^\intercal \Omega^{-1}(A_K - \widetilde{A}_K)x_t + 2 w_t^\intercal \Omega^{-1}(A_K - \widetilde{A}_K)x_t\right),
\end{aligned} \tag{11}$$

taking the expectation, the second term vanishes, and we have

$$\begin{aligned}
\mathbb{E}_\Theta[\ell_t] &= \frac{1}{2}\mathbb{E}_\Theta\left[x_t^\intercal (A_K - \widetilde{A}_K)^\intercal \Omega^{-1}(A_K - \widetilde{A}_K)x_t\right] \\
&= \frac{1}{2}\operatorname{Tr}\left((A_K - \widetilde{A}_K)^\intercal \Omega^{-1}(A_K - \widetilde{A}_K)\Sigma_K(\Theta)\right),
\end{aligned} \tag{12}$$

where the stationary distribution $\Sigma_K(\Theta) = \mathbb{E}_\Theta\left[x_t x_t^\intercal | K\right] = \Omega + A_K \Sigma_K(\Theta) A_K^\intercal$, satisfies a discrete-time Lyapunov equation. For a trajectory τ of T steps, the total expected log-likelihood ratio is

$$\mathbb{E}_\Theta\left[\log \frac{\mathbf{p}(\tau)}{\tilde{\mathbf{p}}(\tau)}\right] = \sum_{t=1}^{T} \mathbb{E}_\Theta[\ell_t] = \frac{T}{2}\operatorname{Tr}\left((A_K - \widetilde{A}_K)^\intercal \Omega^{-1}(A_K - \widetilde{A}_K)\Sigma_K(\Theta)\right). \tag{13}$$

Taking the limit as $T \to \infty$, we see that the total expected log-likelihood ratio diverges linearly, while the per-step average converges to

$$\mathbf{d}_K(\Theta \| \widetilde{\Theta}) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_\Theta \left[\log \frac{\mathbf{p}(\tau)}{\tilde{\mathbf{p}}(\tau)}\right] = \frac{1}{2}\,\mathrm{Tr}\left((A_K - \widetilde{A}_K)^\intercal \Omega^{-1}(A_K - \widetilde{A}_K)\Sigma_K(\Theta)\right). \quad (14)$$

□

A.2 Sub-optimality cost refinement under small perturbations

Proof of Proposition 2. We begin by expressing the cost for the perturbed system $J_K(\Theta(\alpha))$, as

$$\sigma_w^2\ \mathrm{Tr}\big(P_K(\Theta(\alpha))\big) = \sigma_w^2\, \mathbf{i}^\top\ \mathrm{vec}\big(P_K(\Theta(\alpha))\big) = \sigma_w^2\, \mathbf{i}^\top \left(I_{d^2} - A_K^\top(\alpha) \otimes A_K^\top(\alpha)\right)^{-1} \mathbf{q}_K, \quad (15)$$

with $\mathbf{i} = \mathrm{vec}(I_d)$ and $\mathbf{q}_K = \mathrm{vec}(Q_K)$. The closed-loop dynamics for the interpolated system is

$$A_K(\alpha) = A - BK + \alpha\big(\Delta_A + \Delta_B K\big) = A_K + \alpha\,\Delta_K. \quad (16)$$

Its Kronecker square naturally expands as a quadratic function of α,

$$A_K^\top(\alpha) \otimes A_K^\top(\alpha) = X_K + \alpha\,\overline{X}_K + \alpha^2\,\overline{\overline{X}}_K, \quad (17)$$

where $X_K = A_K^\top \otimes A_K^\top$, $\overline{X}_K = (A_K \otimes \Delta_K + \Delta_K \otimes A_K)^\intercal$ and $\overline{\overline{X}}_K = (\Delta_K \otimes \Delta_K)^\intercal$. Thus, the inverse appearing in the cost can be written in terms of perturbation, as

$$\left(I_{d^2} - A_K^\top(\alpha) \otimes A_K^\top(\alpha)\right)^{-1} = \left(I_{d^2} - X_K - \widetilde{X}_K(\alpha)\right)^{-1}, \quad (18)$$

with $\widetilde{X}_K(\alpha) = \alpha\,\overline{X}_K + \alpha^2\,\overline{\overline{X}}_K$. Assuming that the perturbations are small, we apply a first-order expansion of the infinite series, as described in Section 2.2.4 of Stewart & Sun (1990), to obtain

$$\left(I_{d^2} - X_K - \widetilde{X}_K(\alpha)\right)^{-1} \approx Y_K - Y_K\,\widetilde{X}_K(\alpha)\,Y_K, \quad (19)$$

where $Y_K = (I_{d^2} - X_K)^{-1}$. For clarity, we introduce the scalar coefficients $p_K = \mathbf{i}^\top Y_K\,\mathbf{q}_K$, $\overline{p}_K = \mathbf{i}^\top Y_K\,\overline{X}_K\,Y_K\,\mathbf{q}_K$, and $\overline{\overline{p}}_K = \mathbf{i}^\top Y_K\,\overline{\overline{X}}_K\,Y_K\,\mathbf{q}_K$. Hence, the cost function is simplified to

$$\mathbf{i}^\top \left(I_{d^2} - A_K^\top(\alpha) \otimes A_K^\top(\alpha)\right)^{-1} \mathbf{q}_K \approx p_K - \alpha\,\overline{p}_K + \alpha^2\,\overline{\overline{p}}_K. \quad (20)$$

Repeating the derivation for another gain K' and equating the two expressions for $\mathcal{L}(\alpha)$ leads to

$$\begin{gathered}(p_K - p_{K'}) - \alpha\,(\overline{p}_K - \overline{p}_{K'}) + \alpha^2\,(\overline{\overline{p}}_K - \overline{\overline{p}}_{K'}) = 0,\\ \boxed{\alpha = \frac{(\overline{p}_K - \overline{p}_{K'}) \pm \sqrt{(\overline{p}_K - \overline{p}_{K'})^2 - 4(\overline{\overline{p}}_K - \overline{\overline{p}}_{K'})(p_K - p_{K'})}}{2(\overline{\overline{p}}_K - \overline{\overline{p}}_{K'})}.}\end{gathered} \quad (21)$$

Choosing the positive solution completes the derivation. Finally, using the identities $\mathrm{vec}(AXB) = (B^\intercal \otimes A)\,\mathrm{vec}(X)$, and $\mathrm{vec}(I_d)^\intercal\,\mathrm{vec}(X) = \mathrm{Tr}(X)$, and the Neumann series expansion, Kronecker products and vectorizations simplify and complete the proof. □

References

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

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

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pp. 176–184. PMLR, 2017a.

Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. In *Artificial intelligence and statistics*, pp. 1246–1254. PMLR, 2017b.

Marc Abeille and Alessandro Lazaric. Improved regret bounds for thompson sampling in linear quadratic control problems. In *International Conference on Machine Learning*, pp. 1–9. PMLR, 2018.

Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. In *International Conference on Machine Learning*, pp. 23–31. PMLR, 2020.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pp. 127–135. PMLR, 2013.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19, 2006.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9. IEEE, 2018.

Kapilan Balagopalan and Kwang-Sung Jun. Minimum empirical divergence for sub-gaussian linear bandits. *arXiv preprint arXiv:2411.00229*, 2024.

Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846, 1983.

Dorian Baudry, Fabien Pesquerel, Rémy Degenne, and Odalric-Ambrym Maillard. Fast asymptotically optimal algorithms for non-parametric stochastic bandits. *Advances in Neural Information Processing Systems*, 36:11469–11514, 2023a.

Dorian Baudry, Kazuya Suzuki, and Junya Honda. A general recipe for the analysis of randomized multi-armed bandit algorithms. *arXiv preprint arXiv:2303.06058*, 2023b.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012.

Jie Bian and Kwang-Sung Jun. Maillard sampling: Boltzmann exploration done optimally. In *International Conference on Artificial Intelligence and Statistics*, pp. 54–72. PMLR, 2022.

Jie Bian and Vincent YF Tan. Indexed minimum empirical divergence-based algorithms for linear bandits. *arXiv preprint arXiv:2405.15200*, 2024.

Victor Boone and Odalric-Ambrym Maillard. The regret lower bound for communicating markov decision processes. *arXiv preprint arXiv:2501.13013*, 2025.

Hippolyte Bourel, Odalric Maillard, and Mohammad Sadegh Talebi. Tightening exploration in upper confidence reinforcement learning. In *International Conference on Machine Learning*, pp. 1056–1066. PMLR, 2020.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Alon Cohen, Avinatan Hasidim, Tomer Koren, Nevena Lazic, Yishay Mansour, and Kunal Talwar. Online linear quadratic control. In *International Conference on Machine Learning*, pp. 1029–1038. PMLR, 2018.

Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. *Advances in Neural Information Processing Systems*, 33:14156–14170, 2020.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pp. 1467–1476. PMLR, 2018.

Diederich Hinrichsen and Anthony J Pritchard. Stability radius for structured perturbations and the algebraic riccati equation. *Systems & Control Letters*, 8(2):105–113, 1986.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pp. 67–79. Citeseer, 2010.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85:361–391, 2011.

Junya Honda and Akimichi Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *J. Mach. Learn. Res.*, 16:3721–3756, 2015.

Taylan Kargin, Sahin Lale, Kamyar Azizzadenesheli, Animashree Anandkumar, and Babak Hassibi. Thompson sampling achieves $\tilde{O}(\sqrt{T})$ regret in linear quadratic control. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 3235–3284. PMLR, 02–05 Jul 2022.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pp. 199–213. Springer, 2012.

Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 85:1–22, 2022.

Thomas Laffey, Robert Shorten, and Fiacre O Cairbre. On the stability of convex sums of rank-1 perturbed matrices. In *Proceedings of the 2002 American Control Conference (IEEE Cat. No. CH37301)*, volume 2, pp. 1246–1247. IEEE, 2002.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Animashree Anandkumar. Reinforcement learning with fast stabilization in linear dynamical systems. In *International Conference on Artificial Intelligence and Statistics*, pp. 5354–5390. PMLR, 2022.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pp. 728–737. PMLR, 2017.

Hai Lin and Panos J Antsaklis. Stability and stabilizability of switched linear systems: a survey of recent results. *IEEE Transactions on Automatic control*, 54(2):308–322, 2009.

Davide Maran, Alberto Maria Metelli, Matteo Papini, and Marcello Restelli. Local linearity: the key for no-regret reinforcement learning in continuous mdps. *Advances in Neural Information Processing Systems*, 37:75986–76029, 2025.

Akshay Mete, Rahul Singh, and PR Kumar. Augmented rbmle-ucb approach for adaptive control of linear quadratic systems. *Advances in Neural Information Processing Systems*, 35:9302–9314, 2022.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pp. 2701–2710. PMLR, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

Fabien Pesquerel and Odalric-Ambrym Maillard. Imed-rl: Regret optimal learning of ergodic markov decision processes. *Advances in Neural Information Processing Systems*, 35:26363–26374, 2022.

Fabien Pesquerel, Hassan Saber, and Odalric-Ambrym Maillard. Stochastic bandits with groups of similar arms. *Advances in Neural Information Processing Systems*, 34:19461–19472, 2021.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Hassan Saber and Odalric-Ambrym Maillard. Bandits with multimodal structure. In *Reinforcement Learning Conference*, volume 1, pp. 39, 2024.

Hassan Saber, Pierre Ménard, and Odalric-Ambrym Maillard. Indexed minimum empirical divergence for unimodal bandits. *Advances in Neural Information Processing Systems*, 34:7346–7356, 2021.

Hassan Saber, Fabien Pesquerel, Odalric-Ambrym Maillard, and Mohammad Sadegh Talebi. Logarithmic regret in communicating mdps: Leveraging known dynamics with bandits. In *Asian Conference on Machine Learning*, pp. 1167–1182. PMLR, 2024.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International conference on machine learning*, pp. 8583–8592. PMLR, 2020.

G.W. Stewart and J. Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Elsevier Science, 1990. ISBN 9780126702309.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, pp. 5005–5014. PMLR, 2018.

Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on learning theory*, pp. 3036–3083. PMLR, 2019.

Xiangyuan Zhang, Weichao Mao, Saviz Mowlavi, Mouhacine Benosman, and Tamer Başar. Controlgym: Large-scale control environments for benchmarking reinforcement learning algorithms. *arXiv preprint arXiv:2311.18736*, 2023.