Efficient Optimistic Exploration in Linear-Quadratic Regulators via Lagrangian Relaxation

Marc Abeille [1] **Alessandro Lazaric** [2]

Abstract

We study the exploration-exploitation dilemma in the linear quadratic regulator (LQR) setting. Inspired by the extended value iteration algorithm used in optimistic algorithms for finite MDPs, we propose to relax the optimistic optimization of OFU-LQ and cast it into a constrained *extended* LQR problem, where an additional control variable implicitly selects the system dynamics within a confidence interval. We then move to the corresponding Lagrangian formulation for which we prove strong duality. As a result, we show that an ϵ-optimistic controller can be computed efficiently by solving at most $O\big(\log(1/\epsilon)\big)$ Riccati equations. Finally, we prove that relaxing the original OFU problem does not impact the learning performance, thus recovering the $\widetilde{O}(\sqrt{T})$ regret of OFU-LQ. To the best of our knowledge, this is the first computationally efficient confidence-based algorithm for LQR with worst-case optimal regret guarantees.

1. Introduction

Exploration-exploitation in Markov decision processes (MDPs) with continuous state-action spaces is a challenging problem: estimating the parameters of a generic MDP may require many samples, and computing the corresponding optimal policy may be computationally prohibitive. The linear quadratic regulator (LQR) model formalizes continuous state-action problems, where the dynamics is linear and the cost is quadratic in state and action variables. Thanks to its specific structure, it is possible to efficiently estimate the parameters of the LQR by least-squares regression and the optimal policy can be computed by solving a Riccati equation. As a result, several exploration strategies have been adapted to the LQR to obtain effective learning algorithms.

[1]Criteo AI Lab [2]Facebook AI Research. Correspondence to: Marc Abeille <m.abeille@criteo.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Confidence-based exploration. Bittanti et al. (2006) introduced an adaptive control system based on the "bet on best" principle and proved asymptotic performance guarantees showing that their method would eventually converge to the optimal control. Abbasi-Yadkori & Szepesvári (2011) later proved a finite-time $\widetilde{O}(\sqrt{T})$ regret bound for OFU-LQ, later generalized to less restrictive stabilization and noise assumptions by Faradonbeh et al. (2017). Unfortunately, neither exploration strategy comes with a computationally efficient algorithm to solve the optimistic LQR, and thus they cannot be directly implemented. On the TS side, Ouyang et al. (2017) proved a $\widetilde{O}(\sqrt{T})$ regret for the Bayesian regret, while Abeille & Lazaric (2018) showed that a similar bound holds in the frequentist case but restricted to 1-dimensional problems. While TS-based approaches require solving a single (random) LQR, the theoretical analysis of Abeille & Lazaric (2018) suggests that a new LQR instance should be solved at each time step, thus leading to a computational complexity growing linearly with the total number of steps. On the other hand, OFU-based methods allow for "lazy" updates, which require solving an optimistic LQR only a *logarithmic* number of times w.r.t. the total number of steps. A similar lazy-update scheme is used by Dean et al. (2018), who leveraged robust control theory to devise the first learning algorithm with polynomial complexity and sublinear regret. Nonetheless, the resulting adaptive algorithm suffers from a $\widetilde{O}(T^{2/3})$ regret, which is significantly worse than the $\widetilde{O}(\sqrt{T})$ achieved by OFU-LQ.

To the best of our knowledge, the only efficient algorithm for confidence-based exploration with $\widetilde{O}(\sqrt{T})$ regret has been recently proposed by Cohen et al. (2019). Their method, called OSLO, leverages an SDP formulation of the LQ problem, where an optimistic version of the constraints is used. As such, it translates the original non-convex OFU-LQ optimization problem into a *convex* SDP. While solving an SDP is known to have *polynomial* complexity, no explicit analysis is provided and it is said that the runtime may scale polynomially with LQ-specific parameters and the time horizon T (Cor. 5), suggesting that OSLO may become impractical for moderately large T. Furthermore, OSLO requires an initial system identification phase of length $\widetilde{O}(\sqrt{T})$ to properly initialize the method. This strong requirement effectively reduces OSLO to an explore-then-commit strategy, whose

regret is dominated by the length of the initial phase.

Perturbed certainty equivalence exploration. A recent stream of research (Faradonbeh et al., 2018b; Mania et al., 2019; Simchowitz & Foster, 2020) studies variants of the perturbed certainty equivalence (CE) controller (i.e., the optimal controller for the estimated LQR) and showed that this simple exploration strategy is sufficient to reach worst-case optimal regret $\widetilde{O}(\sqrt{T})$. Since the CE controller is not recomputed at each step (i.e., lazy updates) and the perturbation is obtained by sampling from a Gaussian distribution, the resulting methods are computationally efficient. Nonetheless, these methods rely on an isotropic perturbation (i.e., all control dimensions are equally perturbed) and they require the variance to be *large enough* so as to *eventually* reduce the uncertainty on the system estimate along the dimensions that are not naturally "excited" by the CE controller and the environment noise. Being agnostic to the uncertainty of the model estimate and its impact on the average cost, may lead this type of approaches to have longer (unnecessary) exploration and larger regret. On the other hand, confidence-based methods relies on exploration controllers that are explicitly designed to excite more the dimensions with higher uncertainty and impact on the performance. As a result, they are able to perform more effective exploration. We further discuss this difference in Sect. 6.

In this paper, we introduce a novel instance of OFU, for which we derive a computationally efficient algorithm to solve the optimistic LQR with explicit computational complexity and $\widetilde{O}(\sqrt{T})$ regret guarantees. Our approach is inspired by the extended value iteration (EVI) used to solve a similar optimistic optimization problem in finite state-action MDPs (e.g. Jaksch et al., 2010). Relying on an initial estimate of the system obtained after a *finite* number of system identification steps, we first relax the confidence ellipsoid constraints and we cast the OFU optimization problem into a constrained LQR with extended control. We show that the relaxation of the confidence ellipsoid constraint does not impact the regret and we recover a $\widetilde{O}(\sqrt{T})$ bound. We then turn the constrained LQR into a regularized optimization problem via Lagrangian relaxation. We prove strong duality and show that we can compute an ϵ-optimistic and ϵ-feasible solution for the constrained LQR by solving only $O(\log(1/\epsilon))$ algebraic Riccati equations. As a result, we obtain the *first efficient worst-case optimal confidence-based algorithm for LQR*. In deriving these results, we introduce a novel derivation of explicit conditions on the accuracy of the system identification phase leveraging tools from Lyapunov stability theory that may be of independent interest.

2. Preliminaries

We consider the discrete-time linear quadratic regulator (LQR) problem. At any time t, given state $x_t \in \mathbb{R}^n$ and control $u_t \in \mathbb{R}^d$, the next state and cost are obtained as

$$\begin{aligned} x_{t+1} &= A_* x_t + B_* u_t + \epsilon_{t+1}; \\ c(x_t, u_t) &= x_t^\mathsf{T} Q x_t + u_t^\mathsf{T} R u_t, \end{aligned} \tag{1}$$

where A_*, B_*, Q, R are matrices of appropriate dimension and $\{\epsilon_{t+1}\}_t$ is the process noise. Let $\mathcal{F}_t = \sigma(x_0, u_0, \dots, x_t, u_t)$ be the filtration up to time t, we rely on the following assumption on the noise.[1]

Assumption 1. *The noise $\{\epsilon_t\}_t$ is a martingale difference sequence w.r.t. the filtration $\mathcal{F}_t$ and it is componentwise conditionally sub-Gaussian, i.e., there exists $\sigma > 0$ such that $\mathbb{E}(\exp(\gamma \epsilon_{t+1,i})|\mathcal{F}_t) \leq \exp(\gamma^2\sigma^2/2)$ for all $\gamma \in \mathbb{R}$. Furthermore, we assume that the covariance of ϵ_t is the identity matrix.*

The dynamics parameters are summarized in $\theta_*^\mathsf{T} = (A_*, B_*)$ and the cost function can be written as $c(x_t, u_t) = z_t^\mathsf{T} C z_t$ with $z_t = (x_t, u_t)^\mathsf{T}$ and the cost matrix

$$C = \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}. \tag{2}$$

The solution to an LQ is a stationary deterministic policy $\pi : \mathbb{R}^n \to \mathbb{R}^d$ mapping states to controls minimizing the infinite-horizon average expected cost

$$J_\pi(\theta_*) = \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T} c(x_t, u_t)\right], \tag{3}$$

with $x_0 = 0$ and $u_t = \pi(x_t)$. We assume that the LQR problem is "well-posed".

Assumption 2. *The cost matrices Q and R are symmetric p.d. and known, and (A_*, B_*) is stabilizable, i.e., there exists a controller K, such that $\rho(A_* + B_* K) < 1$.*[2]

In this case, Thm.16.6.4 in (Lancaster & Rodman, 1995) guarantees the existence and uniqueness of an optimal policy $\pi_* = \arg\min_\pi J_\pi(\theta_*)$, which is linear in the state, i.e., $\pi_*(x) = K(\theta_*)x$, where,

$$\begin{aligned} K(\theta_*) &= -\big(R + B_*^\mathsf{T} P(\theta_*) B_*\big)^{-1} B_*^\mathsf{T} P(\theta_*) A_*, \\ P(\theta_*) &= Q + A_*^\mathsf{T} P(\theta_*) A_* + A_*^\mathsf{T} P(\theta_*) B_* K(\theta_*). \end{aligned} \tag{4}$$

For convenience, we will denote $P_* = P(\theta_*)$. The optimal average cost is $J_* = J_{\pi_*}(\theta_*) = \mathrm{Tr}(P_*)$. Further, let $L(\theta_*)^\mathsf{T} = \begin{pmatrix} I & K(\theta_*)^\mathsf{T} \end{pmatrix}$, then the closed-loop matrix $A^\mathsf{c}(\theta_*) = A_* + B_* K(\theta_*) = \theta_*^\mathsf{T} L(\theta_*)$ is asymptotically stable.

[1] As shown by Faradonbeh et al. (2017), this can be relaxed to Weibull distributions with known covariance.

[2] $\rho(A)$ is the spectral radius of the matrix A, i.e., the largest absolute value of the eigenvalues of A.

While Asm. 2 guarantees the existence of an optimal linear controller, its optimal cost J_* may still grow unbounded when θ_* is nearly unstable. A popular solution is to introduce a "strong" stability assumption (i.e., $\rho(A^c(\theta_*)) \le \overline{\rho} < 1$). Nonetheless, this imposes stability uniformly over all state dimensions, whereas, depending on the cost matrices Q and R, some dimensions may be less sensitive than others in terms of their impact on the cost. Here we prefer imposing an assumption directly on the optimal cost.[3]

Assumption 3. *There exists $D > 0$ such that $J_* = Tr(P_*) \le D$ and D is known.*

Finally, we introduce $\kappa = D/\lambda_{\min}(C)$, a quantity that will characterize the complexity of many aspects of the learning problem in the following. Intuitively, κ measures the cost of controlling w.r.t. the minimum cost incurred if the uncontrolled system was perfectly stable.

The learning problem. We assume that Q and R are known, while θ_* needs to be estimated from data. We consider the online learning problem where at each step t the learner observes the current state x_t, it executes a control u_t and it incurs the associated cost $c(x_t, u_t)$; the system then transitions to the next state x_{t+1} according to Eq. 1. The learning performance is measured by the cumulative regret over T steps defined as $\mathcal{R}_T(\theta_*) = \sum_{t=0}^{T} \big(c_t - J_*(\theta_*)\big)$. Exploiting the linearity of the dynamics, the unknown parameter θ_* can be directly estimated from data by regularized least-squares (RLS). For any sequence of controls $(u_0, \dots, u_t)$ and the induced states $(x_0, x_1, \dots, x_{t+1})$, let $z_t = (x_t, u_t)^\mathsf{T}$, the RLS estimator with a regularization bias θ_0 and regularization parameter $\lambda \in \mathbb{R}_+^*$ defined as[4]

$$\begin{aligned}\widehat{\theta}_t &= \operatorname*{arg\,min}_{\theta \in \mathbb{R}^{(n+d)\times n}} \sum_{s=0}^{t-1} \|x_{s+1} - \theta^\mathsf{T} z_s\|^2 + \lambda \|\theta - \theta_0\|_F^2 \\ &= V_t^{-1}\Big(\lambda\theta_0 + \sum_{s=0}^{t-1} z_s x_{s+1}^\mathsf{T}\Big),\end{aligned} \tag{5}$$

where $V_t = \lambda I + \sum_{s=0}^{t-1} z_s z_s^\mathsf{T}$ is the design matrix. The RLS estimator concentrates as follows (see App. C.3).

Proposition 1 (Thm. 1 in Abbasi-Yadkori & Szepesvári 2011)**.** *For any $\delta \in (0,1)$ and any $\mathcal{F}_t$-adapted sequence $(z_0, \dots, z_t)$, the RLS estimator $\hat{\theta}_t$ is such that*

$$\|\theta_* - \widehat{\theta}_t\|_{V_t} \le \beta_t(\delta) \qquad \textit{where} \tag{6}$$

$$\beta_t(\delta) = \sigma\sqrt{2n \log\Big(\frac{\det(V_t)^{1/2} n}{\det(\lambda I)^{1/2}\delta}\Big)} + \lambda^{1/2}\|\theta_0 - \theta_*\|_F,$$

w.p. $1-\delta$ (w.r.t. the noise $\{\epsilon_{t+1}\}_t$ and any randomization in the choice of the control).

[3] An alternative assumption may bound the operator norm of P_*, (see e.g., Simchowitz & Foster (2020)).

[4] For $\theta_0 = 0$, this reduces to the standard estimator. The need for a "centered" regularization term is explained in the next section.

Finally, we recall a standard result of RLS.

Proposition 2 (Lem. 10 in Abbasi-Yadkori & Szepesvári 2011)**.** *Let $\lambda \ge 1$, for any arbitrary $\mathcal{F}_t$-adapted sequence $(z_0, z_1, \dots, z_t)$, let V_{t+1} be the corresponding design matrix, then*

$$\sum_{s=0}^{t} \min\big(\|z_s\|^2_{V_s^{-1}}, 1\big) \le 2\log\frac{\det(V_{t+1})}{\det(\lambda I)}.$$

Moreover when $\|z_t\| \le Z$ for all $t \ge 0$, then

$$\sum_{s=0}^{t} \|z_s\|^2_{V_s^{-1}} \le \Big(1 + \frac{Z^2}{\lambda}\Big)(n+d)\log\Big(1 + \frac{(t+1)Z^2}{\lambda(n+d)}\Big). \tag{7}$$

3. A Sequentially Stable Variant of OFU-LQ

In this section we introduce a variant of the original OFU-LQ of Abbasi-Yadkori & Szepesvári (2011) that we refer to as OFU-LQ++. Similar to (Faradonbeh et al., 2017), we use an initial system identification phase to initialize the system and we provide *explicit* conditions on the accuracy required to guarantee sequential stability thereafter. This result is obtained leveraging tools from Lyapunov stability theory which may be of independent interest.

Faradonbeh et al. (2018a) showed that it is possible to construct a set $\Theta_0 = \{\theta : \|\theta - \theta_0\| \le \epsilon_0\}$ containing the true parameters θ^* with high probability, through a system identification phase where a randomized sequence of linear controllers is used to accurately estimate the dynamics. In particular, they proved that a set Θ_0 with accuracy ϵ_0 can be obtained by running the system identification phase for as long as $T_0 = \Omega(\epsilon_0^{-2})$ steps.[5]

After the initial phase, OFU-LQ++ uses the estimate θ_0 to regularize the RLS as in (5) and it proceeds through episodes. At the beginning of episode k it computes a parameter

$$\theta_k = \operatorname*{arg\,min}_{\theta \in \mathcal{C}_k} J(\theta), \tag{8}$$

where t_k is the step at which episode k begins and the constrained set $\mathcal{C}_k$ is defined as

$$\mathcal{C}_k = \mathcal{C}(\beta_{t_k}, V_{t_k}) := \{\theta : \|\theta - \widehat{\theta}_{t_k}\|_{V_{t_k}} \le \beta_{t_k}\}, \tag{9}$$

where $\beta_t = \beta_t(\delta/4)$ is defined in (6). Then the corresponding optimal control $K(\theta_k)$ is computed (4) and the associated policy is executed until $\det(V_t) \ge 2\det(V_{t_k})$.

Lemma 1. *Let $\Theta_0 = \{\theta : \|\theta - \theta_0\| \le \epsilon_0\}$ be the output of the initial system identification phase of Faradonbeh*

[5] An alternative scheme for system identification requires access to a stable controller K_0 and to perturb the corresponding controls to returned a set Θ_0 of desired accuracy ϵ_0 (see e.g., Simchowitz & Foster 2020).

et al. (2018a). For all $t \geq 0$, consider the confidence ellipsoid $\mathcal{C}_t := \{\theta : \|\theta - \widehat{\theta}_t\|_{V_t} \leq \beta_t\}$, where $\widehat{\theta}_t$ and V_t are defined in (5) with regularization bias θ_0 (the center of Θ_0), regularization parameter

$$\lambda = \frac{2n\sigma^2}{\epsilon_0^2}\Big(\log(4n/\delta) + (n+d)\log\Big(1 + \kappa X^2 T\Big)\Big), \tag{10}$$

and β_t is defined in (6) where $\|\theta - \theta_*\|$ is replaced by its upper-bound ϵ_0. Let $\{K(\theta_t)\}_{t\geq 1}$ be the sequence of optimistic controllers generated by OFU-LQ++ *and let $\{x_t\}_{t\geq 0}$ be the induced state process (Eq. 1). If $\epsilon_0 \leq O(1/\kappa^2)$, then with probability at least $1 - \delta/2$, for all $t \leq T$,*

$$\begin{cases} \theta_* \in \mathcal{C}_t \\ \|x_t\| \leq X := 20\sigma\sqrt{\kappa\|P_*\|_2 \log(4T/\delta)/\lambda_{\min}(C)}. \end{cases} \tag{11}$$

OFU-LQ++ has some crucial differences w.r.t. the original algorithm. OFU-LQ receives as input a Θ_0 such that for any $\theta \in \Theta_0$ the condition $\|\theta^\mathsf{T} L(\theta)\| < 1$ holds. While this condition ensures that all the LQ systems in Θ_0 are indeed stable, it does not immediately imply that the optimal controllers $K(\theta)$ stabilize the *true* system θ_*. Nonetheless, Abbasi-Yadkori & Szepesvári (2011) proved that the sequence of controllers generated by OFU-LQ naturally defines a state process $\{x_t\}_t$ which remains bounded at any step t with high probability. Unfortunately, their analysis suffers from several drawbacks: **1)** the state bound scales exponentially with the dimensionality, **2)** as θ_* is required to belong to Θ_0, it should satisfy itself the condition $\|\theta_*^\mathsf{T} L(\theta_*)\| < 1$, which significantly restricts the type of LQR systems that can be solved by OFU-LQ, **3)** the existence of Θ_0 is stated *by assumption* and no concrete algorithm to construct it is provided.

Furthermore, OFU-LQ requires solving (8) under the constraint that θ belongs to the intersection $\mathcal{C}_k \cap \Theta_0$, while OFU-LQ++ only uses the confidence set $\mathcal{C}_k$ to guarantee that the controllers $K(\widetilde{\theta}_k)$ generated through the episodes induces a sequentially stable state process. Although the resulting optimization problem is still non-convex and difficult to solve directly, removing the constraint of Θ_0 enables the relaxation that we introduce in the next section. Finally, we notice that our novel analysis of the sequential stability of OFU-LQ++ leads to a tighter bound on the state, more explicit conditions on ϵ_0, and lighter assumptions than Faradonbeh et al. (2018a).

As a result, we can refine the analysis of OFU-LQ and obtain a much sharper regret bound for OFU-LQ++.

Lemma 2. *For any LQR (A_*, B_*, Q, R) satisfying Asm. 1, 2, and 3, after T steps* OFU-LQ++*, if properly initialized and tuned as in Lem. 1, suffers a regret*

$$\mathcal{R}(T) = \widetilde{O}\Big(\big(\kappa\|P_*\|_2^2 + \sqrt{\kappa}\|P_*\|_2^{3/2}(n+d)\sqrt{n}\big)\sqrt{T}\Big). \tag{12}$$

4. An Extended Formulation of OFU-LQ++

The optimization in (8) is non-convex and it cannot be solved directly. In this section we introduce a relaxed constrained formulation of (8) and show that its solution is an optimistic controller with similar regret as OFU-LQ++ at the cost of requiring a slightly more accurate initial exploration phase (i.e., smaller ϵ_0).

4.1. The Extended Optimistic LQR with Relaxed Constraints

Our approach is directly inspired by the extended value iteration (EVI) used to solve a similar optimistic optimization problem in finite state-action MDPs (e.g. Jaksch et al., 2010). In EVI, the choice of dynamics θ from $\mathcal{C}$ is added as an additional control variable, thus obtaining an extended policy $\widetilde{\pi}$. Exploiting the specific structure of finite MDPs, it is shown that optimizing over policies $\widetilde{\pi}$ through value iteration is equivalent to solving a (finite) MDP with the same state space and an extended (compact) control space and the resulting optimal policy, which prescribes both actions and a choice of the model θ, is indeed optimistic w.r.t. the original MDP. Leveraging a similar idea, we "extend" the LQR with estimated parameter $\widehat{\theta}_t$ by introducing an additional control variable w corresponding to a specific choice of $\theta \in \mathcal{C}_k$. In the following we remove the dependency of $\widehat{\theta}$, β, V, and $\mathcal{C}$ on the learning step t_k and episode k, while we use a generic time s to formulate the extended LQR.

Let $\theta \in \mathcal{C}$ such that $\theta = \widehat{\theta} + \delta_\theta = (A, B) = (\widehat{A} + \delta_A, \widehat{B} + \delta_B)$, then the dynamics of the corresponding LQR is

$$\begin{aligned} x_{s+1} &= Ax_s + Bu_s + \epsilon_{s+1} \\ &= \widehat{A}x_s + \widehat{B}u_s + \delta_A x_s + \delta_B u_s + \epsilon_{s+1}, \\ &= \widehat{A}x_s + \widehat{B}u_s + \delta_\theta z_s + \epsilon_{s+1}, \end{aligned} \tag{13}$$

where we isolate the "perturbations" δ_A and δ_B applied to the current estimates. We replace the perturbation associated to θ with a novel control variable w_s, the *perturbation control variable*, which effectively plays the role of "choosing" the parameters of the perturbed LQR, thus obtaining

$$\begin{aligned} x_{s+1} &= \widehat{A}x_s + \widehat{B}u_s + w_s + \epsilon_{s+1}, \\ &= \widehat{A}x_s + \widetilde{B}\widetilde{u}_s + \epsilon_{s+1}, \end{aligned} \tag{14}$$

where we conveniently introduced $\widetilde{B} = [\widehat{B}, I]$ and $\widetilde{u}_s = [u_s, w_s]$.[6] This *extended* system has the same state variables as the original LQ, while the number of control variables moves from d to $n + d$. Since perturbations δ_θ are such that $\theta = \widehat{\theta} + \delta_\theta \in \mathcal{C}$, we introduce a constraint on the perturbation control such that $\|w_s\| = \|\delta_\theta^\mathsf{T} z_s\| \leq \beta\|z_s\|_{V^{-1}}$

[6] In the following we use tilde-notation such as $\widetilde{\pi}$ and $\widetilde{wtB}$ to denote quantities in the extended LQR.

(see Prop. 1). We refer to the resulting system as the *extended LQR with hard constraints*. Unfortunately, this constrained system is no longer a "standard" LQR structure, as the constraint should be verified *at each step*. To overcome this difficulty, we relax the previous constraint and define

$$g_{\widetilde{\pi}}(\widehat{\theta},\beta,V)=\lim_{S\to\infty}\frac{1}{S}\mathbb{E}\Big(\sum_{s=0}^{S}\|w_s\|^2-\beta^2\|z_s\|^2_{V^{-1}}\Big), \quad (15)$$

where $\widetilde{\pi}=(\pi^u,\pi^w)$ is an *extended* policy defining both standard and perturbation controls, so that $u_s=\pi^u(x_s)$ and $w_s=\pi^w(x_s)$, the expectation is w.r.t. the noise ϵ_{s+1}, and the dynamics of x_s follows (14). As a result, we translate the original constraint $\theta\in\mathcal{C}$, which imposed a per-step condition on w_s to $g_{\widetilde{\pi}}(\widehat{\theta},\beta,V)\le 0$, which considers the asymptotic average behavior of w_s. We are now ready to define the *extended LQR with relaxed constraints* as

$$\begin{aligned}\min_{\widetilde{\pi}}\mathcal{J}_{\widetilde{\pi}}(\widehat{\theta},\beta,V):=&\limsup_{S\to\infty}\frac{1}{S}\mathbb{E}\Big[\sum_{s=0}^{S}c(x_s,\pi^u(x_s))\Big]\\ \text{subject to}\quad & x_{s+1}=\widehat{A}x_s+\widetilde{B}\widetilde{u}_s+\epsilon_{s+1} \quad (16)\\ & g_{\widetilde{\pi}}(\widehat{\theta},\beta,V)\le 0,\end{aligned}$$

where $\mathcal{J}_{\widetilde{\pi}}$ is the average cost of (14) when controlled with $\widetilde{\pi}$ and c is the cost of the original LQ. We also denote by $\mathcal{J}_*(\widehat{\theta},\beta,V)$ the minimum of (16). Once the constrained LQR is solved, the component $\widetilde{\pi}^u$ relative to the variable u is used to control the real system for the whole episode until the termination condition is met. When $\widetilde{\pi}$ is linear, we denote by $\widetilde{K}$ the associated gain (i.e., $\widetilde{\pi}(x)=\widetilde{K}x$), and we use K_u (resp. K_w) to refer to the block of $\widetilde{K}$ corresponding to the control u (resp. the perturbation control w).

4.2. Optimism and Regret

We show that the optimization in (16) preserves the learning guarantees of the original OFU-LQ algorithm at the cost of a slightly stronger requirement on ϵ_0. This is not obvious as (16) is relaxing the constraints imposed by the confidence set used in (8) and solving the extended LQR might lead to a perturbation control $\widetilde{\pi}^w$ that does not actually correspond to any feasible model θ in $\mathcal{C}$. Intuitively, we need the constraint $g_{\widetilde{\pi}}$ to be loose enough so as to guarantee optimism and tight enough to preserve good regret performance. We start by showing that optimizing (16) gives an optimistic solution.

Lemma 3. *Under Asm. 2, and 3, whenever $\theta_*\in\mathcal{C}$, the optimal solution to* (16) *is optimistic, i.e.,*

$$\mathcal{J}_*(\widehat{\theta},\beta,V)\le J_*, \quad (17)$$

The lemma above shows that the optimal controller in the extended LQR has an average cost (in the extended LQR) that is smaller than the true optimal average cost, thus certifying its optimistic nature of (16). This is expected, since (16) is a relaxed version of the original OFU-LQ++ problem, which returns optimistic solutions by definition. Then we show that applying the optimistic extended controllers induce a sequentially stable state process.

Lemma 4. *Given the same system identification phase and RLS estimator of* OFU-LQ++ *(see Lem. 1), let $\{\widetilde{K}_t\}_{t\ge1}$ be the sequence of extended optimistic controllers generated by solving* (16) *and $\{x_t\}_{t\ge0}$ be the state process (Eq. 1) induced when by the sequence of controllers $\{K_{u,t}\}_{t\ge0}$. If $\epsilon_0\le O(1/\kappa^{3/2})$, then with probability at least $1-\delta/2$, for all $t\le T$,*

$$\begin{cases}\theta_*\in\mathcal{C}_t\\ \|x_t\|\le X:=20\sigma\sqrt{\kappa\|P_*\|_2\log(4T/\delta)/\lambda_{\min}(C)}\end{cases} \quad (18)$$

This lemma is the counterpart of Lem. 1 for the extended LQR and it illustrates that, due to the relaxed constraint, the condition on ϵ_0 is tighter by a factor $1/\sqrt{\kappa}$, while the bound on the state remains the same and this, in turn, leads to the same regret as OFU-LQ++ (Lem. 2) but for problem dependent constants.

Theorem 1. *Let (A_*,B_*,Q,R) be any LQR satisfying Asm. 1, 2, and 3. If the conditions in Lem. 4 are satisfied and the extended LQR with relaxed constrained* (16) *is solved exactly at each episode, then w.p. at least $1-\delta$,*

$$\mathcal{R}(T)=\widetilde{O}\big((n+d)\sqrt{n}\kappa^{3/2}\|P_*\|_2^2\sqrt{T}\big). \quad (19)$$

5. Efficient Solution to the Constrained Extended LQR via Lagrangian Relaxation

We introduce the Lagrangian formulation of (16). Let $\mu\in\mathbb{R}$ be the Lagrangian parameter, we define

$$\mathcal{L}_{\widetilde{\pi}}(\widehat{\theta},\beta,V;\mu):=\mathcal{J}_{\widetilde{\pi}}(\widehat{\theta},\beta,V)+\mu g_{\widetilde{\pi}}(\widehat{\theta},\beta,V). \quad (20)$$

Since both average cost $\mathcal{J}_{\widetilde{\pi}}$ and constraint $g_{\widetilde{\pi}}$ measure asymptotic average quantities, we can conveniently define the matrices ($C_\dagger$ being the bordering of matrix C in (2))

$$C_\dagger=\begin{pmatrix}Q&0&0\\0&R&0\\0&0&0\end{pmatrix};\quad C_g=\begin{pmatrix}-\beta^2V^{-1}&0\\0&I\end{pmatrix},$$

and write the Lagrangian as

$$\mathcal{L}_{\widetilde{\pi}}(\widehat{\theta},\beta,V;\mu)=\lim_{S\to\infty}\frac{1}{S}\mathbb{E}\Big[\sum_{s=0}^{S-1}\begin{pmatrix}x_s^\intercal & \tilde{u}_s^\intercal\end{pmatrix}C_\mu\begin{pmatrix}x_s\\ \tilde{u}_s\end{pmatrix}\Big],$$

with $C_\mu=C_\dagger+\mu C_g$. This formulation shows that $\mathcal{L}_{\widetilde{\pi}}(\mu)$ can be seen as the average cost of an extended LQR problem with state x_s, control $\widetilde{u}_s$, linear dynamics $(\widehat{A},\widetilde{B})$ and quadratic cost with matrix C_μ. As a result, we introduce

the *Lagrangian extended LQR* problem associated to the extended LQR with relaxed constraints of (16) as

$$\begin{aligned} \mathcal{L}_*(\widehat{\theta}, \beta, V) = \sup_{\mu \in \mathcal{M}} \min_{\widetilde{\pi}} \mathcal{L}_{\widetilde{\pi}}(\widehat{\theta}, \beta, V; \mu) \\ \text{subject to} \quad x_{s+1} = \widehat{A} x_s + \widetilde{B} \widetilde{u}_s + \epsilon_{s+1} \end{aligned}, \tag{21}$$

where $\mathcal{M} = [0, \widetilde{\mu})$ is the domain of the Lagrangian parameter (more details on $\widetilde{\mu}$ are reported in App. G.2). We prove the following fundamental result.

Theorem 2. *For any extended LQR parametrized by $\widehat{\theta}$, V, β, and psd cost matrices Q, R, there exists a domain $\mathcal{M} = [0, \widetilde{\mu})$ with $\widetilde{\mu} \in \mathbb{R}_+$, such that strong duality between the relaxed optimistic optimization in (16) and its Lagrangian formulation (21) holds:*

$$\mathcal{J}_*(\widehat{\theta}, \beta, V) = \mathcal{L}_*(\widehat{\theta}, \beta, V).$$

Supported by the strong duality above, we provide a more detailed characterization of $\mathcal{L}_{\widetilde{\pi}}(\widehat{\theta}, \beta, V; \mu)$, which motivates the design of an efficient algorithm to optimize over $\widetilde{\pi}$, μ. For ease of notation, in the following we consider $\widehat{\theta}$, β, and V as fixed and we drop them from the definition of $\mathcal{L}_{\widetilde{\pi}}(\mu)$, which we study as a function of $\widetilde{\pi}$ and μ.

5.1. The Lagrangian Dual Function

We introduce the Lagrangian dual function, for any $\mu \in \mathbb{R}_+$,

$$\begin{aligned} \mathcal{D}(\mu) = \min_{\widetilde{\pi}} \mathcal{L}_{\widetilde{\pi}}(\mu) \\ \text{s.t.} \;\; x_{s+1} = \widehat{A} x_s + \widetilde{B} \widetilde{u}_s + \epsilon_{s+1} \end{aligned}, \tag{22}$$

and we denote by $\widetilde{\pi}_\mu$ the corresponding extended optimal policy. For *small enough* μ, the cost matrix C_μ is p.d., which allows solving (22) using standard Riccati theory. The main technical challenge arises for larger values of μ when the solution of the dual Lagrangian function may not be computable by Riccati equations or it may not even be defined. Fortunately, the following lemma shows that within the domain $\mathcal{M}$ where Thm. 2 holds, there always exists a Riccati solution for (22).

Lemma 5. *For any extended LQR parametrized by $\widehat{\theta}$, V, β, and psd cost matrices Q, R, consider the domain $\mathcal{M} = [0, \widetilde{\mu})$ where Thm. 2 holds, then for any $\mu \in \mathcal{M}$*

1. *The extended LQ in (22) is controllable and it admits a unique solution*

$$\widetilde{\pi}_\mu = \arg\min_{\widetilde{\pi}} \mathcal{L}_{\widetilde{\pi}}(\mu), \tag{23}$$

obtained by solving the generalized discrete algebraic Riccati equation (DARE) (Molinari, 1975)[7] associated with the Lagrange LQR $(\widehat{A}, \widetilde{B}, C_\mu)$. Let $C_\mu = (R_\mu \;\; N_\mu; N_\mu \;\; Q_\mu)$ be the canonical formulation for the cost matrix, then

$$\begin{aligned} D_\mu &= R_\mu + \tilde{B}^\mathsf{T} P_\mu \tilde{B} \\ P_\mu &= Q_\mu + A^\mathsf{T} P_\mu A \\ &\quad - [A^\mathsf{T} P_\mu \tilde{B} + N_\mu^\mathsf{T}] D_\mu^{-1} [\tilde{B}^\mathsf{T} P_\mu A + N_\mu], \end{aligned} \tag{24}$$

and the optimal control is $\widetilde{K}_\mu = -D_\mu^{-1}[\tilde{B}^\mathsf{T} P_\mu A + N_\mu]$, while the dual function is $\mathcal{D}(\mu) = Tr(P_\mu)$.

2. *$\mathcal{D}(\mu)$ is concave and continuously differentiable.*

3. *The derivative $\mathcal{D}'(\mu) = g_{\widetilde{\pi}_\mu}$, i.e., it is equal to the constraint evaluated at the optimal extended policy for μ. As a result, the Lagrangian dual can be written as*

$$\mathcal{L}_{\widetilde{\pi}_\mu}(\mu) = \mathcal{D}(\mu) = \mathcal{J}_{\widetilde{\pi}_\mu} + \mu \mathcal{D}'(\mu). \tag{25}$$

The previous lemma implies that in order to solve (21) we may need to evaluate the dual function $\mathcal{D}(\mu)$ only where the the optimal control can be computed by solving a DARE.

Since $\mathcal{D}(\mu)$ is concave and smooth, we can envision using a simple dichotomy approach to optimize $\mathcal{D}(\mu)$ over $\mathcal{M}$ and solve (21). Nonetheless, we notice that Thm. 2 only provides strong duality in a sup/min sense, which means that the optimum may not be attainable within $\mathcal{M}$. Furthermore, even when there exists a maximum, computing an ϵ-optimal solution in terms of the Lagrangian formulation, i.e., finding a pair $\mu, \widetilde{\pi}$ such that $|\mathcal{L}_{\widetilde{\pi}}(\mu) - \mathcal{L}_*| \leq \epsilon$, may not directly translate in a policy with desirable performance in terms of its average cost $\mathcal{J}_{\widetilde{\pi}}$ and feasibility w.r.t. the constraint $g_{\widetilde{\pi}}$.

We illustrate this issue in the example in Fig. 1. We display a qualitative plot of the Lagrangian dual $\mathcal{D}(\mu)$ and its derivative $\mathcal{D}'(\mu)$ when (21) admits a maximum at μ^* and the dichotomy search returned values μ_l and μ_r that are ϵ-close and $\mu^* \in [\mu_l, \mu_r]$. We consider the case where the algorithm returns μ_l as the candidate solution. By concavity and the fact that $\mathcal{D}'(0) > 0$, the function $\mathcal{D}(\mu)$ is Lipschitz in the interval $[0, \mu^*)$ with constant bounded by $\mathcal{D}'(0)$. Thus the accuracy $\mu^* - \mu_l \leq \epsilon$ translates into an equivalent ϵ-optimality in $\mathcal{D}$ (i.e., $\mathcal{D}(\mu^*) - \mathcal{D}(\mu_l) = \mathcal{L}^* - \mathcal{L}_{\widetilde{\pi}_{\mu_l}}(\mu_l) \leq \mathcal{D}'(0)\epsilon$). Nonetheless, this does not imply a similar guarantee for $D'(\mu)$. If the second derivative of $\mathcal{D}(\mu)$ (i.e., the curvature of the function) is large close to μ^*, the original error ϵ can be greatly amplified when evaluating $\mathcal{D}'(\mu_l)$. For instance, if $\mathcal{D}''(\mu) \gg 1/\epsilon$, then $\mathcal{D}'(\mu_l) = \Omega(1)$. Given the last point of Lem. 5, this means that despite returning an ϵ-optimal solution in the sense of (21), $\widetilde{\pi}_{\mu_l}$ may significantly violate the constraint (as $\mathcal{D}'(\mu_l) = g_{\widetilde{\pi}_{\mu_l}} = \Omega(1)$). While Eq. (25) implies that $\widetilde{\pi}_{\mu_l}$ is still optimistic (i.e., $\mathcal{J}_{\widetilde{\pi}_{\mu_l}} \leq J^\star$, as in Lem. 3), the regret accumulated by $\widetilde{\pi}_{\mu_l}$ cannot be controlled anymore, since the pertubration control w_s may be arbitrarily outside the confidence interval. Interestingly, the

[7] The need for generalized DARE is due to the fact that for some $\mu \in \mathcal{M}$, the associated cost C_μ may not be p.s.d.

curvature becomes larger and larger as the optimum shifts to the extremum of $\mathcal{M}$ and, in the limit, (21) only admits a supremum. In this case, no matter how close μ_l is to μ^*, the associated policy $\widetilde{\pi}_{\mu_l}$ may perform arbitrarily bad.

More formally, we have the following lemma (the explicit value of α is reported in Lem. 14).

Lemma 6. *For any LQR parametrized by $\widehat{\theta}$, V, β, and psd cost matrices Q, R, consider the domain $\mathcal{M} = [0, \widetilde{\mu})$ where Thm. 2 holds. Let $\mathcal{M}_+$ be a subset of $\mathcal{M}$ such that $\mathcal{M}_+ = \{\mu \in \mathcal{M} \text{ s.t. } \mathcal{D}'(\mu) \geq 0\}$. Then, $\mathcal{D}$ has Lipschitz gradient, i.e., there exists a constant α depending on $\widehat{\theta}$, V, β, and the cost matrices Q, R, such that for all $(\mu_1, \mu_2) \in \mathcal{M}_+^2$,*

$$|\mathcal{D}'(\mu_1) - \mathcal{D}'(\mu_2)| \leq |\mu_1 - \mu_2| \frac{\alpha}{\lambda_{\min}(D_{\mu_1})},$$

where D_μ is defined in (24).

This result shows that even when $|\mu_1 - \mu_2| \leq \epsilon$, the difference in gradients may be arbitrarily large when $\lambda_{\min}(D_\mu) \ll \epsilon$ (i.e., large curvature). In the next section, we build on this lemma to craft an adaptive stopping condition for the dichotomy search and to detect that case of large curvature.

5.2. An Efficient Dichotomy Search

The algorithm we propose, DS-OFU, seeks to find a value of μ of zero gradient $\mathcal{D}'(\mu)$ by dichotomy search. While $\mathcal{D}(\mu)$ is a 1-dim function and Lem. 5 guarantees that it is concave in $\mathcal{M}$, there are three major challenges to address: **1)** Thm. 2 does not provide any explicit value for $\widetilde{\mu}$; **2)** The algorithm needs to evaluate $\mathcal{D}'(\mu)$; **3)** For any ϵ, DS-OFU must return a policy $\widetilde{\pi}_\epsilon$ that is ϵ-optimistic and ϵ-feasible for the extended LQR with relaxed constraints (16).

DS-OFU starts by checking the sign of the gradient $\mathcal{D}'(0)$. If $\mathcal{D}'(0) \leq 0$, the algorithm ends and outputs the optimal policy $\widetilde{\pi}_0$ since by concavity 0 is the arg-max of $\mathcal{D}$ and $\widetilde{\pi}_0$ is the exact solution to (21). If $\mathcal{D}'(0) > 0$, the dichotomy starts with accuracy ϵ and a *valid*[8] search interval $[0, \mu_{\max}]$, where $\mu_{\max}$ is defined as follows.

Lemma 7. *Let $\mu_{\max} := \beta^{-2} \lambda_{\max}(C) \lambda_{\max}(V)$, then $\mathcal{D}'(\mu_{\max}) < 0$.*

The previous lemma does not imply that $[0, \mu_{\max}] \supseteq \mathcal{M}$, but it provides an explicit value of μ with negative gradient, thus defining a bounded and valid search interval for the dichotomy process. At each iteration, DS-OFU updates either μ_l or μ_r so that the interval $[\mu_l, \mu_r]$ is always valid.

The second challenge is addressed in the following proposition, which illustrates how the derivative $\mathcal{D}'(\mu)$ (equivalently the constraint $g_{\widetilde{\pi}_\mu}$) can be efficiently computed.

[8] We say that $[\mu_l, \mu_r]$ is valid if $\mathcal{D}'(\mu_l) \geq 0$ and $\mathcal{D}'(\mu_r) \leq 0$.

Proposition 3. *For any $\mu \in \mathcal{M}$, let $\widetilde{\pi}_\mu$ (Eq. 23) have an associated controller $\widetilde{K}_\mu$ that induces a closed-loop dynamics $A^{\mathsf{c}}(\widetilde{K}_\mu) = \widehat{A} + \widehat{B}\widetilde{K}_\mu$ then $\mathcal{D}'(\mu) = g_{\widetilde{\pi}_\mu} = Tr(G_\mu)$, where G_μ is the unique solutions of the Lyapunov equation*

$$G_\mu = \left(A^{\mathsf{c}}(\widetilde{K}_\mu)\right)^{\mathsf{T}} G_\mu A^{\mathsf{c}}(\widetilde{K}_\mu) + \begin{pmatrix} I \\ \widetilde{K}_\mu \end{pmatrix}^{\mathsf{T}} C_g \begin{pmatrix} I \\ \widetilde{K}_\mu \end{pmatrix},$$

This directly from the fact that $g_{\widetilde{\pi}}$ is an asymptotic average quadratic quantity (as much as the average cost J), and it is thus the solution of a Lyapunov equation of dimension n.

The remaining key challenge is to design an adaptive stopping condition that is able to keep refining the interval $[\mu_l, \mu_r]$ until either an accurate enough solution is returned, or, the curvature is too large (or even infinite). In the latter case, the algorithm switches to a failure mode, for which we design an ad-hoc solution.

Since the objective is to achieve an ϵ-feasible solution (i.e., $g_{\widetilde{\pi}_{\overline{\mu}}} \leq \epsilon$), we leverage Lem. 6 and we interrupt the dichotomy process whenever $(\mu_r - \mu_l)\alpha / \lambda_{\min}(D_{\mu_l}) \leq \epsilon$. Nonetheless, when the optimum of (21) is not attainable in $\mathcal{M}$, the previous stopping condition may never be verified and the algorithm would never stop. As a result, we interrupt the dichotomy process when $\lambda_{\min}(D_{\mu_l}) \leq \lambda_0 \epsilon^2$ for a given constant λ_0. In this case, the dichotomy fails to return a viable solution and we need to revert to a *backup* strategy, which consists in either modifying the controller found at μ_l or applying a suitable perturbation to the original cost matrix $C_\dagger$. In the latter case, we design a perturbation such that **1)** the optimization problem (21) associated to the system with the perturbed cost $C'_\dagger$ admits a maximum and can be efficiently solved by the same dichotomy process illustrated before and **2)** the corresponding solution $\widetilde{\pi}'$ is ϵ-optimistic and ϵ-feasible in the *original* system. The explicit backup strategy is reported in App. I.

Theorem 3. *For any LQR parametrized by $\widehat{\theta}$, V, β, and psd cost matrices Q, R, and any accuracy $\epsilon \in (0, 1/2)$, there exists values of α and λ_0 and a backup strategy such that*

1. DS-OFU *outputs an ϵ-optimistic and ϵ-feasible policy $\widetilde{\pi}_\epsilon$ given by the linear controller $\widetilde{K}_\epsilon$ such that*

$$\mathcal{J}_{\widetilde{\pi}_\epsilon} \leq \mathcal{J}_* + \epsilon \quad \text{and} \quad g_{\widetilde{\pi}_\epsilon} \leq \epsilon.$$

2. DS-OFU *terminates within at most $N = O\big(\log(\mu_{\max}/\epsilon)\big)$ iterations, each solving one Riccati and one Lyapunov equation for the extended Lagrangian LQR, both with complexity $O(n^3)$.*

This result shows that DS-OFU returns a solution to (16) at any level of accuracy ϵ in a total runtime $O(n^3 \log(1/\epsilon))$. We refer to the algorithm resulting by plugging DS-OFU into the OFU-LQ++ learning scheme as LAGLQ (Lagrangian-LQ). By running DS-OFU with $\epsilon = 1/\sqrt{t}$ provides the regret guarantee of Thm. 1.

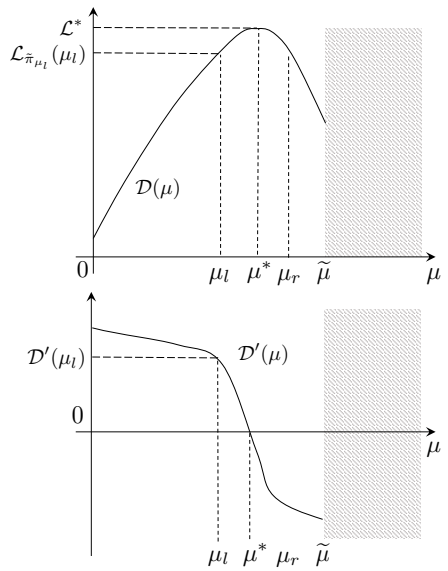

Figure 1. Qualitative plots of $\mathcal{D}(\mu)$ and its derivative $\mathcal{D}'(\mu)$.

Input: $\widehat{\theta}$, β, V, ϵ, α, λ_0
if $\mathcal{D}(0) \leq 0$ **then**
  Set $\overline{\mu} = 0$ and $\widetilde{\pi}_\epsilon = \widetilde{\pi}_{\overline{\mu}}$
else
  Set $\mu_l = 0$, $\mu_r = \mu_{\max}$ (Lem. 7)
  while $\alpha \frac{\mu_r - \mu_l}{\lambda_{\min}(D_{\mu_l})} \geq \epsilon$ **or** $\lambda_{\min}(D_{\mu_l}) \geq \lambda_0 \epsilon^2$ **do**
    Set $\overline{\mu} = (\mu_l + \mu_r)/2$
    if $\mathcal{D}'(\overline{\mu}) > 0$ **then**
      $\mu_l = \overline{\mu}$
    else
      $\mu_r = \overline{\mu}$
    end if
  end while
end if
if $\alpha \frac{\mu_r - \mu_l}{\lambda_{\min}(D_{\mu_l})} < \epsilon$ **then**
  Set $\overline{\mu} = \mu_l$ and $\widetilde{\pi}_\epsilon = \widetilde{\pi}_{\overline{\mu}}$
else
  Set $\widetilde{\pi}_\epsilon$ to the control return by the *backup procedure*
end if
return Control policy $\widetilde{\pi}_\epsilon$

Figure 2. The DS-OFU algorithm to solve (21).

6. Discussion

We investigate the difference between confidence-based and isotropic exploration in term of complexity, bounds and empirical performance. While not conclusive, we believe this discussion sheds light on how confidence-based methods may be better at adapting to the structure of the problem. As a representative for isotropic exploration, we refer to CECCE (Simchowitz & Foster, 2020), which offers the tightest regret guarantee among the CE strategies. For confidence-based exploration, we discuss the guarantee of both OFU-LQ++ and LAGLQ, but limit the computational and experiment comparisons with LAGLQ only.

Computational complexity. Both LAGLQ and CECCE proceeds through episodes of increasing length. LAGLQ relies on the standard determinant-based rule ($\det(V_t) \geq 2\det(V_{t_k})$) to decide when to stop, while in CECCE the length of each episode is twice longer than the previous one. In both cases, the length of the episodes is increasing exponentially over time, thus leading to $O(\log T)$ updates. Given the complexity analysis in Thm. 3, we notice that DS-OFU and computing the CE controller have the same order of complexity, where CECCE solves *one* Riccati equation, while DS-OFU solves as many as $\log(1/\epsilon)$ Riccati and Lyapunov equations. On systems of moderate side and given the small number of recomputations, the difference between the two approaches is relatively narrow.

Regret. We limit the comparison to the main order term $\widetilde{O}(\sqrt{T})$ and the dependencies on dimensions n and d, and problem-dependent constants such as κ and $\|P^*\|_2$,

$$\mathcal{R}_{\text{CECCE}} = \widetilde{O}\big(\|P^*\|_2^{11/2} d\sqrt{nT}\big),$$
$$\mathcal{R}_{\text{OFU-LQ++}} = \widetilde{O}\big(\kappa^{1/2}\|P_*\|_2^{3/2}(n+d)\sqrt{n}\sqrt{T}\big),$$
$$\mathcal{R}_{\text{LAGLQ}} = \widetilde{O}\big(\kappa^{3/2}\|P_*\|_2^{2}(n+d)\sqrt{n}\sqrt{T}\big).$$

The first difference is that CECCE has worst-case optimal dependency $d\sqrt{n}$ on the dimension of the problem, while optimistic algorithms OFU-LQ++ and LAGLQ are slightly worse, scaling with $(n+d)\sqrt{n}$. While this shows that OFU-LQ++ and LAGLQ are worst-case optimal when $n \approx d$, it is an open question whether $\epsilon-$greedy is by nature superior to confidence-based method when $d \ll n$ or whether it is due to a loose analysis. In fact, those dependencies are mostly inherited from the confidence intervals in (1) which is treated differently in (Simchowitz & Foster, 2020), thanks to a *refined* bound and a *different* regret decomposition. This suggests that a finer analysis for OFU-LQ++ and LAGLQ may close this gap.

The main difference lies in the dependency on complexity-related quantities κ and $\|P^*\|_2$. While there is no strict ordering between them,[9] they both measure the cost of controlling the system. In this respect, CECCE suffers from a significantly larger dependency than optimistic algorithms: OFU-LQ++ offers the best performance while LAGLQ is slightly worse than OFU-LQ++, due to the use of a relaxed constraint to obtain tractability. We believe this difference in performance may be intrinsic in the fact that methods

[9]The definition of $\kappa = D/\lambda_{\min}Q$, with $D \geq \text{Tr}(P^*)$ may suggest $\kappa > \|P^*\|_2$, but the smallest eigenvalue of Q may be large enough so that $\kappa \leq \|P^*\|_2$.

based on isotropic perturbations of the CE are less effective in adapting to the actual structure of the problem. As the isotropic perturbation is tuned to guarantee a sufficient estimation in all directions of the state-control space, it leads to over-exploration w.r.t. some directions as soon as there is an asymmetry in the cost sensitivity in the estimation error. On the other hand, OFU-LQ++ and LAGLQ further leverage this asymmetry from the confidence set, and by optimism, do not waste exploration to learn accurately directions which have little to no impact on the performance.

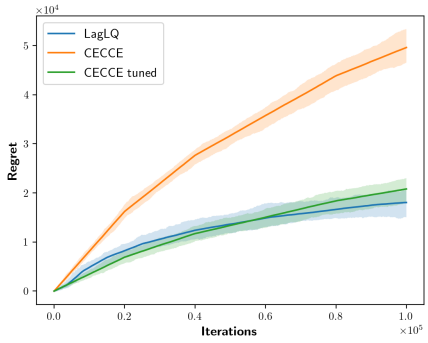

Figure 3. Regret curves for CECCE and LAGLQ.

Empirical comparison. We conclude with a simple numerical simulation (details in App. J). We compare CECCE with the variance parameter (σ_{in}^2) set as suggested in the original paper and a tuned version where we shrink it by a factor $\sqrt{\|P_*\|_2}$, and LAGLQ where the confidence interval is set according to (1). Both algorithms receive the same set Θ_0 obtained from an initial system identification phase. In Fig. 3 we see that LAGLQ performs better than both the original and tuned versions of CECCE. More interestingly, while CECCE is "constrained" to have a $O(\sqrt{T})$ regret by the definition of the perturbation itself, which scales as $1/\sqrt{t}$, it seems LAGLQ's regret is $o(\sqrt{T})$, suggesting that despite the worst-case lower bound $\Omega(\sqrt{T})$, LAGLQ may be adapt to the structure of the problem and achieve better regret.

References

Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *COLT*, pp. 1–26, 2011.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

Abeille, M. and Lazaric, A. Improved regret bounds for thompson sampling in linear quadratic control problems. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 1–9, 2018.

Bittanti, S., Campi, M., et al. Adaptive control of linear time invariant systems: the "bet on the best" principle. *Communications in Information & Systems*, 6(4):299–320, 2006.

Cohen, A., Koren, T., and Mansour, Y. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. *CoRR*, abs/1902.06223, 2019. URL http://arxiv.org/abs/1902.06223.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. Regret bounds for robust adaptive control of the linear quadratic regulator. *CoRR*, abs/1805.09388, 2018. URL http://arxiv.org/abs/1805.09388.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Finite time analysis of optimal adaptive policies for linear-quadratic systems. *CoRR*, abs/1711.07230, 2017. URL http://arxiv.org/abs/1711.07230.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Finite time adaptive stabilization of LQ systems. *CoRR*, abs/1807.09120, 2018a. URL http://arxiv.org/abs/1807.09120.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Input perturbations for adaptive regulation and learning. *CoRR*, abs/1811.04258, 2018b. URL http://arxiv.org/abs/1811.04258.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11:1563–1600, August 2010.

Lancaster, P. and Rodman, L. *Algebraic riccati equations*. Oxford University Press, 1995.

Mania, H., Tu, S., and Recht, B. Certainty Equivalent Control of LQR is Efficient. *arXiv e-prints*, art. arXiv:1902.07826, Feb 2019.

Molinari, B. P. The stabilizing solution of the discrete algebraic riccati equation. *Automatic Control, IEEE Transactions on*, 20 (3):396–399, Jun 1975.

Ouyang, Y., Gagrani, M., and Jain, R. Learning-based control of unknown linear systems with thompson sampling. *CoRR*, abs/1709.04047, 2017. URL http://arxiv.org/abs/1709.04047.

Rockafellar, R. T. *Convex analysis*. Number 28. Princeton university press, 1970.

Rugh, W. J. *Linear system theory*, volume 2. prentice hall Upper Saddle River, NJ, 1996.

Simchowitz, M. and Foster, D. J. Naive Exploration is Optimal for Online LQR. *arXiv e-prints*, art. arXiv:2001.09576, Jan 2020.

Van Dooren, P. A generalized eigenvalue approach for solving riccati equations. *SIAM Journal on Scientific and Statistical Computing*, 2(2):121–135, 1981.