Augmented RBMLE-UCB Approach for Adaptive Control of Linear Quadratic Systems

Akshay Mete
Texas A & M University
College Station, Texas, USA
akshaymete@tamu.edu

Rahul Singh
Indian Institute of Science
Bengaluru, Karnataka, India
rahulsingh@iisc.ac.in

P. R. Kumar
Texas A & M University
College Station, Texas, USA
prk@tamu.edu

Abstract

We consider the problem of controlling an unknown stochastic linear system with quadratic costs – called the adaptive LQ control problem. We re-examine an approach called "Reward-Biased Maximum Likelihood Estimate" (RBMLE) [1] that was proposed more than forty years ago, and which predates the "Upper Confidence Bound" (UCB) method as well as the definition of "regret" for bandit problems [2]. It simply added a term favoring parameters with larger rewards to the criterion for parameter estimation. We show how the RBMLE and UCB methods can be reconciled, and thereby propose an Augmented RBMLE-UCB algorithm that combines the penalty of the RBMLE method with the constraints of the UCB method [3], uniting the two approaches to optimism in the face of uncertainty. We establish that theoretically, this method retains $\tilde{\mathcal{O}}(\sqrt{T})$ regret, the best known so far. We further compare the empirical performance of the proposed Augmented RBMLE-UCB and the standard RBMLE (without the augmentation) with UCB, Thompson Sampling, Input Perturbation, Randomized Certainty Equivalence and StabL on many real-world examples including flight control of Boeing 747 and Unmanned Aerial Vehicle. We perform extensive simulation studies showing that the Augmented RBMLE consistently outperforms UCB, Thompson Sampling and StabL by a huge margin, while it is marginally better than Input Perturbation and moderately better than Randomized Certainty Equivalence.

1 Introduction

Consider a linear stochastic system $x_{t+1} = A^\star x_t + B^\star u_t + w_{t+1}$, where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^m$ and $w_t \in \mathbb{R}^n$ are the state, control applied, and state noise, respectively, at time t. We study the adaptive control/reinforcement learning problem [4, 5, 6, 7] where the controller's goal is to minimize the expected finite horizon quadratic cost $\mathbb{E}\sum_{t=1}^{T}(x_t^\intercal Q x_t + u_t^\intercal R u_t)$ by choosing the control input based on observing the past states and controls, without the knowledge of "true parameter" $\theta^\star := (A^\star, B^\star)^\intercal$. The deviation from what would have been optimally possible had the true parameter been known is measured by the "regret," [2], $R(T) := \sum_{t=1}^{T}(x_t^\intercal Q x_t + u_t^\intercal R u_t) - T\, J^\star(\theta^\star)$, where $J^\star(\theta^\star) := \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\sum_{t=1}^{T}(x_t^\intercal Q x_t + u_t^\intercal R u_t)$ is the optimal average cost achievable when $(A^\star, B^\star)$ are known.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

For this adaptive linear-quadratic (LQ) control problem, there are broadly four classical approaches: the Reward-Biased Maximum Likelihood Estimate (RBMLE) approach [1], the Diminishing Excitation (DE) approach [8], the Upper Confidence Bound (UCB) approach [2], and the Thompson Sampling (TS) [9] approach based on sampling from the posterior distribution. The RBMLE and UCB approaches are "certainty equivalent" (CE) in the sense that they make an estimate $\hat{\theta}_t$ of the unknown true parameter, and then take an action $u_t = K(\hat{\theta}_t)x_t$ that would be optimal if the estimate $\hat{\theta}_t$ were indeed the true parameter. They only differ in what parameter estimate they choose. The DE approach applies $u_t = K(\hat{\theta}_t)x_t + v_t$, where v_t is an added "excitation," an independent noise, of diminishing variance and $\hat{\theta}_t$ is the maximum likelihood estimate. The Randomized Certainty Equivalence (RCE) [10] adds excitation to the parameter estimate.

1.1 The Contributions

1. We unite the two approaches to "optimism under uncertainty," RBMLE [1] and UCB [2], by showing that the RBMLE method is a penalty version of the constrained optimization problem of UCB for the case of linear quadratic systems [11, 12, 13, 14, 3]. Based on this we propose an Augmented RBMLE (ARBMLE) method that combines the penalty and constrained versions so that on the one hand it retains the analytical tractability of UCB and on the other hand provides the performance of RBMLE.
2. We determine how to choose the biasing factor $\alpha(t)$ for ARBMLE, and establish a finite time regret bound $\tilde{\mathcal{O}}(\sqrt{T})$, the same as the OFULQ algorithm of [3], the best order available to date.
3. We perform extensive comparative simulation studies of the performance of:
   (a) ARBMLE and the standard RBMLE.
   (b) OFULQ [3], which is the UCB-approach adapted to the LQ problem.
   (c) TS [15] which is the Thompson sampling approach adapted to the LQ problem
   (d) Input Perturbation (IP) [16] which is a recent reincarnation of DE for the adaptive LQ problem that additionally assumes apriori knowledge of a stabilizing controller.
   (e) Stabl [17], a modified OFULQ that adds "excitation" to the input for initialization.
   (f) RCE [10], which adds excitation to the Least Squares Estimate (LSE).

   The examples used for our simulation study have been used in many recent papers [18, 19, 17], namely (a) the longitudinal flight control of Boeing 747 with linearized dynamics [17],(b) Unmanned Aerial Vehicle (UAV) [20, 17] (c) unstable Laplacian dynamics [18], and (d) large transient dynamics [18]. Our simulation results show that ARBMLE outperforms OFULQ, TS and StabL by a large margin, which is primarily due to lack of stabilization experienced by OFULQ and TS [17]. RCE also exhibits a higher regret than ARBMLE. While the empirical performance of IP is marginally worse than ARBMLE. The results show that the ARBMLE has the same performance as the original RBMLE [21], and they both outperform all the above other algorithms. Notably the choices made by ARBMLE and OFULQ within the confidence interval are very far apart, with ARBMLE outperforming OFULQ by a large margin.

1.2 Previous Works

Prior work on RBMLE has concentrated on establishing long-term average optimality [22, 23, 1, 11, 24, 25, 12, 13, 14, 26, 27, 28]. Recently, its regret performance has been addressed for Multi-Armed Bandits (MABs) [29], linear contextual bandits [30], and MDPs [31]. Forced exploration techniques, somewhat similar in spirit to the ϵ_t-greedy learning algorithm of [7, 32], are studied in [33, 34, 35] vis-a-vis ensuring that they do not suffer from the insufficient exploration. An adaptive LQG control based on the UCB approach, called OFULQ, is proposed in [3] which establishes a regret of $\tilde{\mathcal{O}}(\sqrt{T})$. A similar algorithm to address the adaptive LQG control problem is also designed in [36]. A computationally efficient algorithm called ROBUST with a regret of $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$ is proposed in [18]. An alternative approach for designing learning algorithms is Thompson sampling [9], and [37, 38] have established an expected regret of $\tilde{\mathcal{O}}(\sqrt{T})$ in a Bayesian context. More recently, new DE-based algorithms, IP [16] and RCE [10], have been proposed that make an additional assumption, not

made in ARBMLE, OFULQ or TS, that one has access to a stabilizing controller for the unknown system, and establish $\tilde{O}(\sqrt{T})$ regret.

1.3 The RBMLE Approach

The RBMLE approach, proposed four decades ago [1], was the first approach not resorting to forced choices as in [39]. We begin by giving an informal description of it in the context of the adaptive LQ problem. Since the true parameter $\theta^\star := (A^\star, B^\star)^T$ is not known, one can make a Least-Squares Estimate (LSE) $\hat{\theta}_t = (\hat{A}_t, \hat{B}_t)^\intercal$ of them:

$$\hat{\theta}_t \in \text{ArgMin}_{\theta=(A,B)^\intercal} \sum_{s=0}^{t-1} ||x_{s+1} - Ax_s - Bu_s||^2. \tag{1}$$

Under the certainty-equivalence approach, the control input applied is $u_t = K(\hat{\theta}_t)x_t$ where $K(\theta)$ is the optimal linear feedback gain for the LQ problem when the system is described by $\theta = (A, B)^\intercal$. Suppose now that these estimates were to converge as $t \to \infty$ to $\hat{\theta}_\infty = (\hat{A}_\infty, \hat{B}_\infty)^\intercal$. Then, asymptotically, the input applied is $u_t \approx K(\hat{\theta}_\infty)x_t$. The closed-loop system therefore settles down to behaving according to

$$x_{t+1} \approx (A^\star + B^\star K(\hat{\theta}_\infty))x_t + w_{t+1}. \tag{2}$$

As it does so, one loses the ability to identify the matrices $(A^\star, B^\star)$, and asymptotically one can only identify the closed-loop gain $A^\star + B^\star K(\hat{\theta}_\infty)$. The problem is that as the parameter estimates begin to converge to $\hat{\theta}_\infty$, the control gain converges to $K(\hat{\theta}_\infty)$ and further exploration ceases, and one ends up only identifying the behavior of the system under the limiting gain $K(\hat{\theta}_\infty)$ being applied to the system. Since the limiting policy $u_t = K(\hat{\theta}_\infty)x_t$ need not be optimal for the long-term average cost for the true system $(A^\star, B^\star)$, the CE rule leads to a sub-optimal performance. Indeed, this problem goes by various names in different fields – the dual control problem [40, 41], the closed-identifiability problem [42, 43, 44], or exploration vs. exploitation dilemma [45].

This problem was resolved in [1] without resorting to forced exploration as in [39]. The key observation made there was that under CE the limiting parameter estimate $\hat{\theta}(\infty)$ has a one-sided bias. Specifically, $J^\star(\hat{\theta}_\infty) \geq J^\star(\theta^\star)$, i.e., the limiting parameter estimate has an optimal cost that is larger than the optimal cost of $\theta^\star$. To see this, denote by $J(K, \theta)$ the long-term cost of using the control gain K when the parameter is θ. Then the fact that the models $\hat{\theta}_\infty$ and $\theta^\star$ have identical behavior under the gain $K(\hat{\theta}_\infty)$ implies that $J(K(\hat{\theta}_\infty), \hat{\theta}_\infty) = J(K(\hat{\theta}_\infty), \hat{\theta}^\star)$. Now note that $J(K(\hat{\theta}_\infty), \hat{\theta}_\infty) = J^\star(\hat{\theta}_\infty)$ since the gain $K(\hat{\theta}(\infty))$ is optimal for $\hat{\theta}_\infty$. However the gain $K(\hat{\theta}_\infty)$ is not necessarily optimal for $\theta^\star$, and so $J(K(\hat{\theta}_\infty), \hat{\theta}^\star) \geq J^\star(\theta^\star)$. Therefore,

$$J^\star(\hat{\theta}_\infty) = J(K(\hat{\theta}_\infty), \hat{\theta}_\infty) = J(K(\hat{\theta}_\infty), \hat{\theta}^\star) \geq J^\star(\theta^\star). \tag{3}$$

Following from this observation, it was reasoned in [1] that if one could slightly bias the MLE to favor models with lower optimal costs $J^\star(\theta)$ so as to obtain $J^\star(\hat{\theta}_\infty) \leq J^\star(\theta^\star)$, then one would have equality throughout (3), to obtain $J(K(\hat{\theta}_\infty), \hat{\theta}^\star) = J^\star(\theta^\star)$, yielding the desired result that the gain $K(\hat{\theta}_\infty)$ is optimal for $\theta^\star$. So motivated, [1] proposed RBMLE[1] which in the context of the LSE suggests choosing the parameter estimate $\hat{\theta}_{RBMLE}(t) = (\hat{A}_{RBMLE}, \hat{B}_{RBMLE})^\intercal$ as:

$$\hat{\theta}_{RBMLE}(t) \in \text{ArgMin}_{\theta=(A,B)^\intercal} \Big[\alpha(t)J^\star(\theta) + \sum_{s=0}^{t-1} ||x_{s+1} - Ax_s - Bu_s||^2\Big]. \tag{4}$$

Generally $\alpha(t) = o(t)$ with the precise growth rate of $\alpha(t) \nearrow +\infty$ dependent on the context.

[1] It has been called the Cost-Biased MLE, as in [12], or Reward-Biased MLE (RBMLE) as in [29], depending on whether one minimizes a cost or maximizes a reward.

1.4 The UCB Approach

The "Upper Confidence Bound" (UCB) approach was first proposed in the context of Multi-Armed Bandit (MAB) Problems in [2]. In the context of Bernoulli bandits, it essentially consists of constructing, for each arm i at each time t, a confidence interval $(\theta(i,1;t), \theta(i,2;t))$ of its payoff probability with confidence $(1-\delta(t))$, for $\delta(t) = o(\frac{1}{t})$, and playing the arm i with the highest value of $\theta(i,1;t)$. This approach has been generalized to a variety of contexts including linear contextual bandits [46], Gaussian Processes [47], MDPs [48, 49], and LQ systems [3].

Particularized to LQ systems, the OFULQ algorithm of [3] that is based on the UCB approach suggests choosing $\hat{\theta}_{OFLUQ}(t) = (\hat{A}_{OFULQ}, \hat{B}_{OFULQ})^\intercal$ as the minimizer of:

$$\min_{\theta \in \mathcal{C}_t(\delta)} J^\star(\theta), \text{ where} \tag{5}$$

$$\mathcal{C}_t(\delta) := \left\{ \theta = (A,B)^\intercal : \sum_{s=0}^{t-1} ||x_{s+1} - Ax_s - Bu_s||^2 \leq \gamma_T(\delta) \right\} \tag{6}$$

is a $(1-\delta)$-high confidence set of parameters for an appropriate choice of $\gamma_T(\delta)$ [3].

2 The Augmented RBMLE-UCB Method

The UCB approach has been called "Optimism in the Face of Uncertainty" (OFU), since it chooses an optimistic arm after calculating the confidence intervals. The RBMLE version is also the same, though it does it in a different way by directly giving preference to parameters that can yield better rewards. It arrives at optimism in a very systematic way by noticing that closed-loop identification leads to the chain of inequalities (3), which could then be made into equalities by ensuring that $J^\star(\hat{\theta}_\infty) \leq J^\star(\theta^\star)$.

We now show how one may reconcile the two approaches to optimism in the face of uncertainty in the context of the LQ problem, and then combine them to obtain a method that has experimentally superior empirical performance while also allowing a proof that it achieves the currently best known order of regret.

It can be seen that the RBMLE approach (4) can be considered an unconstrained penalty version of the constrained optimization problem (5)-(6), with a penalty factor $\frac{1}{\alpha(t)}$ for constraint violation. This provides a justification for employing "optimism" in the UCB approach.

This raises the question of whether we can take advantage of these synergies to fashion a superior algorithm, that is superior from the viewpoint of being able to provide a theoretical guarantee of the best known order of regret to date, as well as superior from the point of view of providing the best experimental performance of the algorithms to date. One can draw inspiration here from the Augmented Lagrangian Method of [50, 51]. Given a constrained problem as in (5)-(6), it adds a penalty for constraint violation, and also adjoins the constraint through a Lagrange multiplier, thus involving the constraint twice. In our case, we add the penalty and also retain the constraint. This leads to the Augmented RBMLE-UCB method:

$$\hat{\theta}_{\text{Aug}}(t) \in \text{ArgMin}_{\theta \in \mathcal{C}_t(\delta)} \Big[\alpha(t) J^\star(\theta) + \sum_{s=0}^{t-1} ||x_{s+1} - Ax_s - Bu_s||^2 \Big]. \tag{7}$$

The advantage of retaining the constrained optimization is in enabling theoretical analysis, in that we can make use of the bounds on parameter estimates from concentration inequalities, as we shall show in the sequel. We thereby prove in Section 4 that it also has $\tilde{\mathcal{O}}(\sqrt{T})$ order of regret, which is the best known so far [3]. Moreover, simulations reported in Section 5 show that its performance is also much improved. In all cases, the simulation performance of ARBMLE is the same as RBMLE, raising the open problem of proving that the original RBMLE for the LQ problem also has the best known order of regret.

3 Problem Formulation

As introduced in section 1, we consider the following linear system,

$$x_{t+1} = A^\star x_t + B^\star u_t + w_{t+1} \tag{8}$$

where, $A^\star \in \mathbb{R}^{n\times n}$ and $B^\star \in \mathbb{R}^{n\times m}$ are unknown system parameters. Define $z_t^\top := (x_t^\top, u_t^\top)$. Then, the linear stochastic system can equivalently be written as

$$x_{t+1} = {\theta^\star}^\top z_t + w_{t+1}. \tag{9}$$

The system incurs a cost at time t given by, $x_t^\intercal Q x_t + u_t^\intercal R u_t$. We assume that $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{m\times m}$ are known positive semi-definite and positive definite matrices respectively. We make the following assumptions on $\theta^\star = (A^\star, B^\star)^\intercal$.

Assumption 1. *There exists a known positive constant c such that $\theta^\star \in \mathcal{S} = \mathcal{S}_0 \cap \mathcal{S}_1$ where, $\mathcal{S}_0 = \{\theta = (A,B)^\intercal \,|\, (A,B)$ is stabilizable and $\left(A, Q^{1/2}\right)$ is detectable$\}$[2] [54], and $\mathcal{S}_1 = \{\theta = (A,B)^\intercal \,|trace(\theta^\top\theta) \le c\}$.*

Assumption 2. *Let $\mathcal{F}_t := \sigma((x_0,u_0),(x_1,u_1),\cdots,(x_t,u_t))$ denote the history of states and inputs until time t. The state noise $\{w_t\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}$, with $\mathbb{E}[w_{t+1}w_{t+1}^\intercal|\mathcal{F}_t] = I_n$, and is element-wise sub-Gaussian, i.e., for any $\gamma \in \mathbb{R}$, there exists $L > 0$ with $\mathbb{E}\left[exp(\gamma w_{t,j})|\mathcal{F}_t\right] \le exp\left(\frac{\gamma^2 L^2}{2}\right)$ $\forall\, j$ and $\forall\, t$.*

For a matrix M, we use $\|M\|$ to denote its operator norm induced from the ℓ_2-norm.

3.1 Controller Design for a Known LQG System

For a discrete time linear system, $x_{t+1} = \theta^\intercal z_t + w_{t+1}$, where $\theta = (A,B)^\intercal \in \mathcal{S}$, there is an unique positive semidefinite matrix $P(\theta)$ that satisfies the Riccati equation (see [6])

$$P(\theta) = Q + A^\intercal P(\theta)A - A^\intercal P(\theta)B(B^\intercal P(\theta)B + R)^{-1}B^\intercal P(\theta)A.$$

The optimal control law which minimizes the long-term average quadratic cost is $u_t = K(\theta)x_t$, where the "gain matrix" is $K(\theta) := -(B^\intercal P(\theta)B + R)^{-1}B^\intercal P(\theta)A$. The optimal average cost, $J^\star(\theta)$ is equal to $trace(P(\theta))$. As a consequence of Assumption 1, the parameter set $\mathcal{S}$ is bounded, hence one can show that $P(\theta)$ is bounded as well:

$$D := \sup_{\theta\in\mathcal{S}} ||P(\theta)|| < \infty. \tag{10}$$

The following assumption is commonly made in online LQR learning problem in which the knowledge of a stabilizing controller is not made [3, 38].

Assumption 3.

$$\rho := \sup_{\theta\in\mathcal{S}} ||A + BK(\theta)|| < 1, \text{ and} \tag{11}$$

$$c_0 := \sup_{\theta\in\mathcal{S}} ||K(\theta)|| < +\infty. \tag{12}$$

Note that there are recent works which relax (11) slightly to $\sup_{\theta\in\mathcal{S}} \rho(A + BK(\theta)) < 1$, however they assume that the learner has access to a stabilizing controller [18, 55, 56].

3.2 Construction of Confidence Interval

The ℓ_2-regularized squared fitting error with parameter $\lambda > 0$ is given by:

$$V_t(\theta) = \lambda||\theta||_2^2 + \sum_{s=0}^{t-1} ||x_{s+1} - \theta^\intercal z_s||_2^2. \tag{13}$$

Let $\hat{\theta}_t$ be the ℓ_2 regularized least-squares estimate of $\theta^\star$, i.e., $\hat{\theta}_t \in \arg\min_{\theta\in\mathcal{S}} V_t(\theta)$. Next, given the history $\mathcal{F}_t$, we construct a "high-probability confidence ball," $\mathcal{C}_t(\delta)$ around $\hat{\theta}_t$, i.e., a set of plausible system parameters that contains the true parameter $\theta^\star$ with a high probability. Let

$$\mathcal{C}_t(\delta) := \left\{\theta : trace\left((\theta - \hat{\theta}_t)^\intercal Z_t(\theta - \hat{\theta}_t)\right) \le \beta_t(\delta)\right\}, \tag{14}$$

[2] We note that even though [3] requires each θ from $\mathcal{S}_0$ to satisfy the stronger reachability-observability assumption, the results therein hold true if this is replaced by the weaker stabilizability-detectability assumption [52, p.61]. In fact all that is required for the proofs to go through, is that for each $\theta \in \mathcal{S}_0$, the corresponding Riccati equation have a unique solution. It is well-known [53] that stabilizability and detectability, as above, is sufficient for this.

where $Z_t := \lambda I_{n+m} + \sum_{s=0}^{t} z_s z_s^\intercal$, and

$$\beta_t(\delta) := \left(nL\sqrt{2\log\left(\frac{\sqrt{det(Z_t)det(\lambda I)}}{\delta}\right)} + \sqrt{\lambda}c \right)^2 . \tag{15}$$

Define

$$\mathcal{E}_1(t) = \{\theta^\star \in \mathcal{C}_s(\delta/4),\ \forall\ s = 1, 2, \cdots, t\} \text{ and } \mathcal{E}_2(t) = \{||x_s|| \leq d_t,\ \forall s \leq t\}, \tag{16}$$

where d_t is defined in the Appendix. Let $\mathcal{E}_1 := \mathcal{E}_1(T)$ and $\mathcal{E}_2 := \mathcal{E}_2(T)$. Then, as in [3], $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - \frac{\delta}{2}$. Moreover, on the event $\mathcal{E}_2(t)$ defined in (16), the following holds,

$$\max_{1\leq s\leq t} \|x_s\| \leq X_t = Y_t^{n+m+1},\ \forall t = 1, 2, \ldots, \text{where,} \tag{17}$$

$$Y_t := \max\Bigg(e, \lambda(n+d)(e-1)\times$$
$$4\left(c_1 \log\left(\frac{1}{\delta}\right) + c_2 \log\left(\frac{t}{\delta}\right)\right)\Bigg) \times \left(\log^2\left(4c_1 \log\left(\frac{1}{\delta}\right) + 4c_2 \log\left(\frac{t}{\delta}\right)\right)\right),$$

with c_1 and c_2 being problem dependent constants.

4 The Augmented RBMLE-UCB Algorithm

We employ a version of the Augmented RBMLE-UCB (ARBMLE) algorithm that proceeds in an episodic manner. Let t_k denote the starting time of the k-th episode. Then, during episode k, it implements the control policy, $u_t = K(\theta_{t_k})x_t,\ \forall\ t \in \{t_k, t_k+1, \ldots, t_{k+1}-1\}$. where, θ_{t_k} is obtained by solving the following optimization problem,

$$\theta_{t_k} \in \arg\min_{\theta\in\mathcal{S}\cap\mathcal{C}_{t_k}(\delta)} \{V_{t_k}(\theta) + \alpha(t_k)J^\star(\theta)\}, \tag{18}$$

where the bias-term, $\alpha(t) = \alpha_0\sqrt{T}, \forall t$, for $\alpha_0 > 0$.

Algorithm 1 Augmented RBMLE-UCB (ARBMLE)

Initialize: $t = 0$, $Z_0 = \lambda I_{n+m}$
for $k = 0, 1, \cdots$ **do**
  if $det(Z_t) > 2det(Z_{t_{k-1}})$ **then**
    solve the following optimization to obtain θ_{t_k},

$$\theta_t \in \arg\min_{\theta\in\mathcal{S}\cap\mathcal{C}_{t_k}(\delta)} \{V_{t_k}(\theta) + \alpha(t_k)J^\star(\theta)\},$$

  else
    $\theta_t = \theta_{t-1}$
  end if
  $u_t = K(\theta_t)x_t$
  $Z_{t+1} = Z_t + z_t z_t^\intercal$
  $t \to t+1$
end for

In Theorem 4.1, we show that regret for ARBMLE is upper bounded by $\tilde{\mathcal{O}}\left(\sqrt{T\log\frac{1}{\delta}}\right)$ which is same order as OFULQ [3].

Theorem 4.1. *For any $\delta \in (0,1)$ and $T > 0$, with a probability at least $(1-\delta)$, the regret of the RBMLE Algorithm is upper-bounded by $R(T) \leq \tilde{\mathcal{O}}\left(\sqrt{T\log\frac{1}{\delta}}\right)$.*

Proof. Appendix A.1. □

Ex.	RBMLE	ARBMLE	OFULQ	TS	IP	RCE	STABL
(a)	3233	3233	1.2×10^6	4.2×10^{10}	3251	3408	1.8×10^6
(b)	5930	5930	5.4×10^{12}	2.8×10^{13}	5955	6396	1.9×10^{10}
(c)	16144	16135	2.1×10^{12}	1.1×10^{20}	16164	180639	$1,2 \times 10^9$
(d)	540297	528805	4.9×10^6	8.2×10^{11}	540248	2.2×10^{14}	1.4×10^7

Table 1: Average Regret Performance at $T = 500$.

5 Empirical Performance

We evaluate the empirical performance of ARBMLE as well as standard (unaugmented) RBMLE. We compare these algorithms with OFULQ [3], Thompson Sampling (TS) [15], Input Perturbations (IE) [16], Randomized Certainty Equivalence (RCE) [10], and Stabl [17]. The results shown here are for the following examples of linear systems that have appeared in the recent literature on adaptive control of linear systems:

1. Unstable Laplacian dynamics [18, 17, 19].
2. Large transient dynamics [18].
3. Unmanned Aerial Vehicle (UAV) [20, 17].
4. Longitudinal Flight Control of Boeing 747 [17].

The details of these examples are provided in the Appendix.

Each simulation experiment is performed for a time horizon of 500 steps, and repeated 50 times. The reported results are the averaged values over the 100 runs. In Figure 1, we compare the $\log_{10}$ of the averaged regret of ARBMLE, OFULQ, TS and Stabl. In Figure 2, we plot the averaged regret of RBMLE, ARBMLE, IP and RCE. We summarize the results using the regret values at $T = 500$, averaged over the 50 runs,in Table 1. More results for more examples are provided in the Appendix. Details of the implementations can be found in the Appendix.

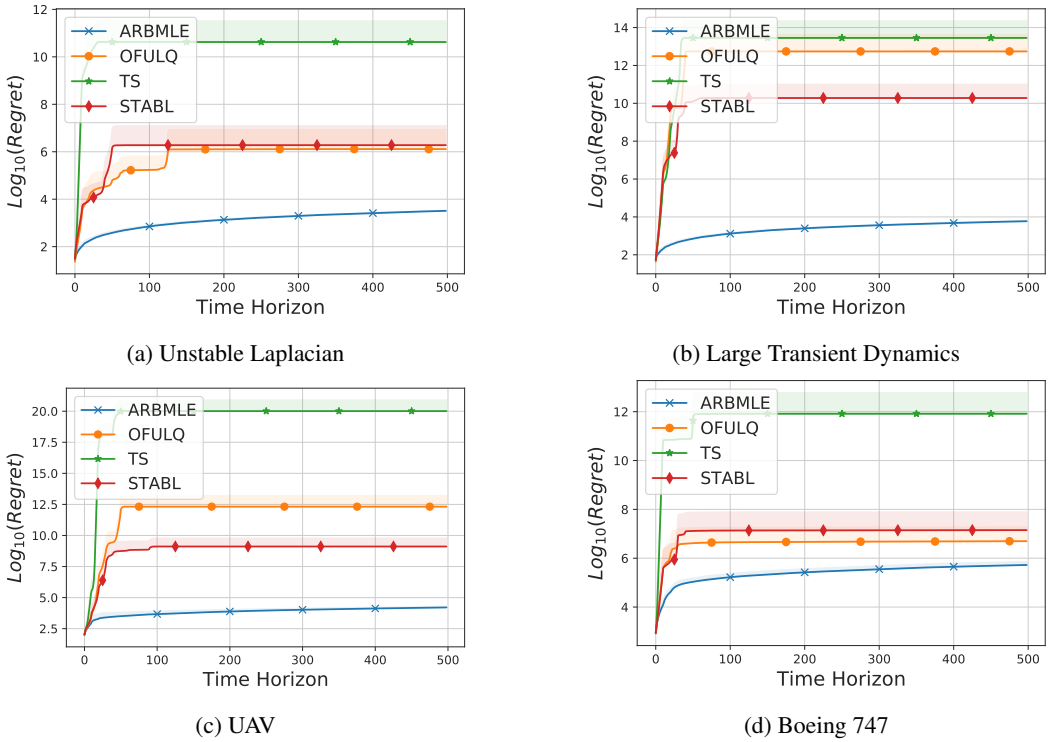
We highlight the key observations from above experiments:

- ARBMLE and RBMLE were always found to have the same empirical performance in most cases. More specifically, both these algorithms choose the same estimate θ_t, and this suggests that $\frac{1}{\alpha(t)}$ needs to be greater than the Lagrange multiplier for the ball constraint $\theta \in \mathcal{C}_{t_k}(\delta)$. This also motivates future study of the regret of the (unaugmented) RBMLE.
- As can be seen in Figure 1, ARBMLE/RBMLE outperform OFULQ and TS by a huge margin. We conjecture that this is due to temporary instability when OFULQ, TS are employed [17]. ARBMLE/RBMLE also has a signficantly lower regret as compared with StabL.
- Figure 2 shows that ARBMLE/RBMLE also outperforms RCE moderately, and IP marginally.

Remark 1. *Our simulations for ARBMLE, OFULQ and TS are based on the confidence interval $\beta_t(\delta)$ as defined in* (15)*. Instead, recent works [17, 18] use $\beta_t(\delta) := trace\left((\theta^\star - \hat{\theta}_t)^\intercal Z_t(\theta^\star - \hat{\theta}_t)\right)$. However, one may note that $\theta^\star$ in $\beta_t(\delta)$ is not known to the learning agent, and so such a definition of $\beta_t(\delta)$ is not a viable implementation. The effect of the choice of $\beta_t(\delta)$ on the regret performance is shown in the appendix.*

6 Concluding Remarks

We reconcile the RBMLE and UCB approaches by showing that RBMLE is an unconstrained penalty version of the constrained optimization problem of UCB. Showing that UCB is a constrained version of RBMLE also explains why the optimism embodied in UCB-based schemes is justified. In particular, it is justified by the goal of nullifying the one-sided bias that results from the closed-loop identification of dynamics.



(a) Unstable Laplacian

(b) Large Transient Dynamics

(c) UAV

(d) Boeing 747

Figure 1: Logarithm of the Averaged Regret over 50 runs of RBMLE, ARBMLE, OFULQ and TS for various example systems.

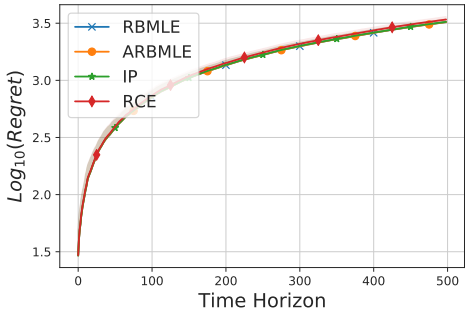
Building on this, we propose an Augmented RBMLE-UCB method that not only matches the best known order of regret, $\tilde{\mathcal{O}}(\sqrt{T})$, to date – that of the OFULQ algorithm – but also outperforms OFULQ, TS and Stabl by a significant margin in simulation experiments. In fact, for the UAV experiment, the regret of OFULQ is $\approx 10^8$, TS $\approx 10^{16}$ and StabL $\approx 10^5$ times that of RBMLE. It outperforms RCE also, but the gains are moderate, while with respect to IP the gains are smaller. The simulations were carried out on real-world examples such as Boeing 747, UAV, Unstable Laplacian and Large Transient Dynamics which were taken from recent works [18, 55, 19, 17, 20] on adaptive LQG control.

This work further extends recent studies of RBMLE for MDPs [31], stochastic Multi-Armed Bandits [29] and Contextual bandits [30] establishing state-of-art regret results as well as empirically good performance.

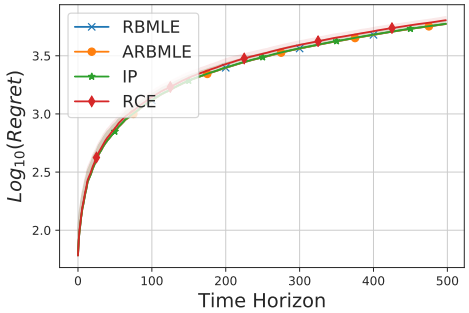
There remain several open questions.

Currently we study only an augmented form of RBMLE, i.e. ARBMLE, in which the agent searches for a parameter value that optimizes a certain reward-biased maximum likelihood objective function within a "high-confidence ball." However, (unaugmented) RBMLE optimizes this over the larger set $\mathcal{S}$ (Assumption 1) that is known to contain the true parameter. The reason for this augmentation is simply that currently we are unable to prove regret bounds for RBMLE without it. By including it, we can capitalize on the nice technical results in [3] in the analysis of OFULQ. Simulations show that the constraint is loose, and performance is the same with/without the constraint. In fact, as shown by the simulations, the choices of θ_{t_k} made by the standard RBMLE, and the Augmented RBMLE are the same. It remains to be seen whether similar regret guarantees can be obtained by the (unaugmented) RBMLE, which remains an open problem.

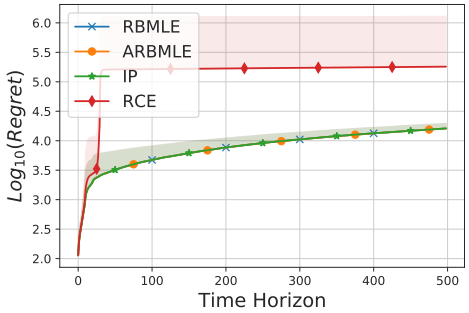
One may note that ARBMLE and OFULQ differ in their choices of decisions. Denoting their estimates of the unknown parameter at a given time by θ_{ARBMLE} and θ_{OFULQ} respectively, θ_{OFULQ} always lies at the boundary of the UCB-ball constraint. This is easily seen to be true for bandits and MDPs, and empirically observed to be so for LQ systems. In contrast, θ_{ARBMLE} is most often in the interior of the UCB-ball. Moreover, while OFULQ treats all models within the UCB-ball equally and only



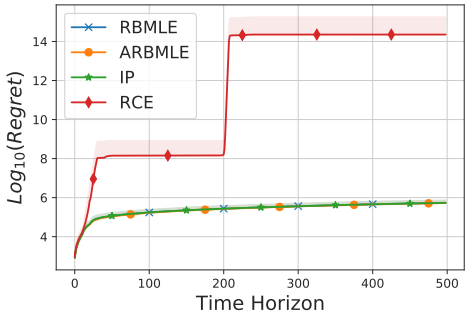
(a) Unstable Laplacian



(b) Large Transient Dynamics



(c) UAV



(d) Boeing 747

Figure 2: Logarithm of the Averaged Regret over 50 runs of RBMLE, ARBMLE, IP and RCE for various example systems. Table 1 provides the quantative results for a finer comparative evaluation of the regrets.

assesses them by their cost, ARBMLE prefers models that are closer to the Least Squares Estimate (θ_{LSE}). For example, if $J^\star(\theta_1) = J^\star(\theta_2)$, then ARBMLE prefers the θ_i that is closer to θ_{LSE}. We conjecture that this is the reason why ARBMLE has a significantly better performance than OFULQ – but we are unable to prove it.

It has been shown in previous works on adaptive LQG control [11, 12] that under appropriate formulations RBMLE is guaranteed to stabilize an unknown LQG system over an infinite horizon, i.e., $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(\|x(t)\|^2 + \|u(t)\|^2) < \infty$ a.s.. Morover, the sample path performance cost $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(x^\intercal(t)Qx(t) + u^\intercal(t)Ru(t))$ is (a.s.) equal to the optimal performance that could be attained if the system parameters $(A^\star, B^\star)$ were known. In fact, stability is a prerequisite before one can establish the latter type of result, as the decades of work on stochastic adaptive control in the nineteen seventies to the nineties has shown. While recent work has tended to study the finer performance measure of regret, it is usually over a finite horizon, and more attention to the stability of the learning process over an infinite horizon appears well deserved. Similarly, robustness which also was rigorously formulated and addressed in earlier decades needs to be re-examined [57].

Acknoweldgments

This material is based upon work partially supported by the US Army Contracting Command under W911NF-22-1-0151, US Office of Naval Research under N00014-21-1-2385; 4/21-22 DARES: Army Research Office W911NF-21-20064 US National Science Foundation under CMMI-2038625, The views expressed herein and conclusions contained in this document are those of the authors and should not be interpreted as representing the views or official policies, either expressed or implied, of the U.S. Army Contracting Command, ONR, ARO, NSF, or the United States Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. The work of Rahul Singh was partially supported by the SERB Grant SRG/2021/002308, and PC 39010B.

References

[1] P. R. Kumar and A. Becker. A new family of optimal adaptive controllers for Markov chains. IEEE Transactions on Automatic Control, 27(1):137–146, 1982.

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

[3] 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, pages 1–26. JMLR Workshop and Conference Proceedings, 2011.

[4] Karl J Åström and Björn Wittenmark. Adaptive control. Courier Corporation, 2013.

[5] P. R. Kumar. A survey of some results in stochastic adaptive control. SIAM Journal on Control and Optimization, 23(3):329–380, 1985.

[6] P. R. Kumar and Pravin Varaiya. Stochastic systems: Estimation, identification, and adaptive control. Prentice-Hall, Englewood Cliffs, NJ, 1986.

[7] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.

[8] Han-Fu Chen and Lei Guo. Identification and stochastic adaptive control, volume 5. Springer Science & Business Media, 1991.

[9] 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.

[10] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. On adaptive linear-quadratic regulators. Automatica, 117:108982, 2020.

[11] P. R. Kumar. Optimal adaptive control of linear-quadratic-gaussian systems. SIAM Journal on Control and Optimization, 21(2):163–178, 1983.

[12] Marco C. Campi and P. R. Kumar. Adaptive linear quadratic Gaussian control: the cost-biased approach revisited. SIAM Journal on Control and Optimization, 36(6):1890–1907, 1998.

[13] Maria Prandini and Marco C. Campi. Adaptive lqg control of input-output systems—a cost-biased approach. SIAM Journal on Control and Optimization, 39(5):1499–1519, 2000.

[14] Sergio Bittanti, Marco C. Campi, et al. Adaptive control of linear time invariant systems: the "Bet on the Best" principle. Communications in Information & Systems, 6(4):299–320, 2006.

[15] Marc Abeille and Alessandro Lazaric. Thompson sampling for linear-quadratic control problems. In Artificial Intelligence and Statistics, pages 1246–1254. PMLR, 2017.

[16] Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Input perturbations for adaptive control and learning. Automatica, 117:108950, 2020.

[17] Sahin Lale, Kamyar Azizzadenesheli, Babak Hassibi, and Animashree Anandkumar. Reinforcement learning with fast stabilization in linear dynamical systems. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, volume 151 of Proceedings of Machine Learning Research, pages 5354–5390. PMLR, 28–30 Mar 2022.

[18] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. Advances in Neural Information Processing Systems, 31, 2018.

[19] Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Model-free linear quadratic control via reduction to expert prediction. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 3108–3117. PMLR, 2019.

[20] Feiran Zhao, Keyou You, and Tamer Başar. Infinite-horizon risk-constrained linear quadratic regulator with average cost. In Proceedings of the 60th IEEE Conference on Decision and Control (CDC), pages 390–395. IEEE, 2021.

[21] Marco C. Campi and P. R. Kumar. Adaptive linear quadratic Gaussian control: the cost-biased approach revisited. SIAM Journal on Control and Optimization, 36(6):1890–1907, 1998.

[22] P. R. Kumar. Adaptive control with a compact parameter set. SIAM Journal on Control and Optimization, 20(1):9–13, 1982.

[23] P. R. Kumar and Woei Lin. Optimal adaptive controllers for unknown Markov chains. IEEE Transactions on Automatic Control, 27(4):765–774, 1982.

[24] P. R. Kumar. Simultaneous identification and adaptive control of unknown systems over finite parameter sets. IEEE Transactions on Automatic Control, 28(1):68–76, 1983.

[25] V. S. Borkar. The Kumar-Becker-Lin scheme revisited. Journal of optimization theory and applications, 66(2):289–309, 1990.

[26] Łukasz Stettner. On nearly self-optimizing strategies for a discrete-time uniformly ergodic adaptive model. Applied Mathematics and Optimization, 27(2):161–177, 1993.

[27] V. S. Borkar. Self-tuning control of diffusions without the identifiability condition. Journal of optimization theory and applications, 68(1):117–138, 1991.

[28] TE Duncan, B Pasik-Duncan, and L Stettner. Almost self-optimizing strategies for the adaptive control of diffusion processes. Journal of optimization theory and applications, 81(3):479–507, 1994.

[29] Xi Liu, Ping-Chun Hsieh, Yu Heng Hung, Anirban Bhattacharya, and P. R. Kumar. Exploration through reward biasing: Reward-biased maximum likelihood estimation for stochastic multi-armed bandits. In International Conference on Machine Learning, pages 6248–6258. PMLR, 2020.

[30] Yu-Heng Hung, Ping-Chun Hsieh, Xi Liu, and P. R. Kumar. Reward-biased maximum likelihood estimation for linear stochastic bandits. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 7874–7882, 2021.

[31] Akshay Mete, Rahul Singh, Xi Liu, and P. R. Kumar. Reward Biased Maximum Likelihood Estimation for Reinforcement Learning. In Learning for Dynamics and Control, pages 815–827. PMLR, 2021.

[32] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. Machine learning, 47(2):235–256, 2002.

[33] Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. The Annals of Statistics, 10(1):154–166, 1982.

[34] Tze Leung Lai and Ching-Zong Wei. Asymptotically efficient self-tuning regulators. SIAM Journal on Control and Optimization, 25(2):466–481, 1987.

[35] Han-Fu Chen and Lei Guo. Optimal adaptive control and consistent parameter estimates for armax model with quadratic cost. SIAM Journal on Control and Optimization, 25(4):845–867, 1987.

[36] Morteza Ibrahimi, Adel Javanmard, and Benjamin Van Roy. Efficient reinforcement learning for high dimensional linear quadratic systems. In NIPS, pages 2645–2653, 2012.

[37] Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In UAI, pages 1–11. Citeseer, 2015.

[38] Yi Ouyang, Mukul Gagrani, and Rahul Jain. Control of unknown linear systems with thompson sampling. In Proceedings of the 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pages 1198–1205. IEEE, 2017.

[39] Herbert Robbins. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 58(5):527–535, 1952.

[40] Aleksandr Aronovich Feldbaum. Dual control theory. i. Avtomatika i Telemekhanika, 21(9):1240–1249, 1960.

[41] Björn Wittenmark. Adaptive dual control methods: An overview. Adaptive Systems in Control and Signal Processing 1995, pages 67–72, 1995.

[42] Vivek Borkar and P. Varaiya. Adaptive control of Markov chains, i: Finite parameter set. IEEE Transactions on Automatic Control, 24(6):953–957, 1979.

[43] Donald A Berry and Bert Fristedt. Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). London: Chapman and Hall, 5(71-87):7–7, 1985.

[44] John Gittins, Kevin Glazebrook, and Richard Weber. Multi-armed bandit allocation indices. John Wiley & Sons, 2011.

[45] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.

[46] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

[47] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. arXiv preprint arXiv:0912.3995, 2009.

[48] Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In Advances in neural information processing systems, pages 49–56, 2007.

[49] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. Journal of Machine Learning Research, 11(4), 2010.

[50] Magnus R Hestenes. Multiplier and gradient methods. Journal of optimization theory and applications, 4(5):303–320, 1969.

[51] Michael JD Powell. A method for nonlinear constraints in minimization problems. Optimization, pages 283–298, 1969.

[52] Yasin Abbasi-Yadkori. Online learning for linearly parametrized control problems. PhD thesis, University of Alberta, Edmonton, Alberta, 2013.

[53] Frank L Lewis, Draguna Vrabie, and Vassilis L Syrmos. Optimal control. John Wiley & Sons, 2012.

[54] Thomas Kailath. Linear systems, volume 156. Prentice-Hall Englewood Cliffs, NJ, 1980.

[55] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. Foundations of Computational Mathematics, 20(4):633–679, 2020.

[56] Max Simchowitz and Dylan Foster. Naive exploration is optimal for online LQR. In International Conference on Machine Learning, pages 8937–8948. PMLR, 2020.

[57] L Praly, S-F Lin, and P. R. Kumar. A robust adaptive minimum variance controller. SIAM journal on control and optimization, 27(2):235–266, 1989.

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

[59] Tadashi Ishihara, Hai-Jiao Guo, and Hiroshi Takeda. A design of discrete-time integral controllers with computation delays via loop transfer recovery. Automatica, 28(3):599–603, 1992.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Material: Augmented RBMLE-UCB Approach for Adaptive Control of Linear Quadratic Systems

Appendix

A Regret Analysis

We now prove the $\tilde{\mathcal{O}}\left(\sqrt{T\log\frac{1}{\delta}}\right)$ upper-bound on regret of Augmented RBMLE-UCB that was claimed in Section 4.

Lemma A.1. *The regret $R(T)$ of the Augmented RBMLE-UCB learning algorithm can be decomposed as $R(T) = R_1 + R_2 + R_3 + R_4$, where*

$$
\begin{aligned}
R_1 &:= \sum_{t=0}^{T}\left\{x_t^\intercal P(\theta_t)x_t - \mathbb{E}\left[x_{t+1}^\intercal P(\theta_{t+1})x_{t+1}|\mathcal{F}_t\right]\right\},\\
R_2 &:= \sum_{t=0}^{T}\mathbb{E}\big[x_t^\intercal\left(P(\theta_{t+1}) - P(\theta_t)\right)x_t|\mathcal{F}_t\big],\\
R_3 &:= \sum_{t=0}^{T}\left\{\left(A^\star x_t + B^\star u_t\right)^\intercal P(\theta_t)\left(A^\star x_t + B^\star u_t\right) - \left(A_t x_t + B_t u_t\right)^\intercal P(\theta_t)\left(A_t x_t + B_t u_t\right)\right\},\\
R_4 &:= \sum_{t=0}^{T}\left(J(\theta_t) - J(\theta^\star)\right).
\end{aligned}
\tag{19}
$$

Proof. Consider an algorithm that implements $u_t = K(\theta_t)x_t$ at time t. Note that $x_{t+1} = A^\star x_t + B^\star u_t + w_{t+1}$. Define $\tilde{x}_{t+1}^u := A_t x_t + B_t u + w_{t+1}$. Then, the Bellman optimality equation for the Linear Quadratic control problem can be written as follows,

$$
\begin{aligned}
J^\star(\theta_t) + x_t^\intercal P(\theta_t)x_t &= \min_u\left(x_t^\intercal Q x_t + u^\intercal R u + \mathbb{E}\left[(\tilde{x}_{t+1}^u)^\intercal P(\theta_t)\tilde{x}_{t+1}^u|\mathcal{F}_t\right]\right)\\
&= \left(x_t^\intercal Q x_t + u_t^\intercal R u_t + \mathbb{E}\left[(\tilde{x}_{t+1}^{u_t})^\intercal P(\theta_t)\tilde{x}_{t+1}^{u_t}|\mathcal{F}_t\right]\right).
\end{aligned}
$$

Upon substituting the value of $\tilde{x}_{t+1}^{u_t}$ in the above, we get

$$
\begin{aligned}
&J^\star(\theta_t) + x_t^\intercal P(\theta_t)x_t\\
&= (x_t^\intercal Q x_t + u_t^\intercal R u_t) + \mathbb{E}\Big[(A_t x_t + B_t u_t + w_{t+1})^\intercal P(\theta_t)(A_t x_t + B_t u_t + w_{t+1})|\mathcal{F}_t\Big].
\end{aligned}
\tag{20}
$$

Note that $w_{t+1} = x_{t+1} - (A^\star x_t + B^\star u_t)$ and w_t is a martingale difference sequence (Assumption 2). Thus, the l.h.s. of (20) can be written as follows,

$$
\begin{aligned}
J^\star(\theta_t) + x_t^\intercal P(\theta_t)x_t - (x_t^\intercal Q x_t + u_t^\intercal R u_t) =& J^\star(\theta_t) - J^\star(\theta^\star) + x_t^\intercal P(\theta_t)x_t\\
&- (x_t^\intercal Q x_t + u_t^\intercal R u_t - J^\star(\theta^\star))\\
&= \mathbb{E}\left[x_{t+1}^\intercal P(\theta_{t+1})x_{t+1}|\mathcal{F}_t\right]\\
&- \mathbb{E}\big[x_t^\intercal\left(P(\theta_{t+1}) - P(\theta_t)\right)x_t|\mathcal{F}_t\big]\\
&- \left(A^\star x_t + B^\star u_t\right)^\intercal P(\theta_t)\left(A^\star x_t + B^\star u_t\right)\\
&+ \left(A_t x_t + B_t u_t\right)^\intercal P(\theta_t)\left(A_t x_t + B_t u_t\right).
\end{aligned}
$$

Therefore by taking a sum from $t = 0$ to $t = T$ on both sides, we get $R(T) = R_1 + R_2 + R_3 + R_4$. □

Lemma A.2. *On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have $R_1 \leq 2DW^2\sqrt{2T\log\frac{8}{\delta}} + n\sqrt{B_\delta}$, where $B_\delta := b\log\left(\frac{4n\sqrt{b}}{v\delta}\right)$, $b := v + T(cDX_T)^2(1 + c_0^2)$, $W = nL\sqrt{2n\log\left(\frac{8nT}{\delta}\right)}$ and D is as in* (10).

Proof. The proof is the same as that of Lemma 7 in [3], and hence omitted. □

Lemma A.3. *On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have $R_2 \leq 2DX_T^2 \log_2 T$, where X_T is defined in* (17).

Proof. The term in the summation $\sum_{t=0}^{T} \mathbb{E}\left[x_t^\intercal \left(P(\theta_{t+1}) - P(\theta_t)\right) x_t | \mathcal{F}_t\right]$ (19) corresponding to time t is non-zero only when a change in the policy occurs at t. There are $(n+m)\log_2\left(1 + TX_T^2 \frac{1+c_0^2}{\lambda}\right)$ episodes till time T (Lemma 8, [3]). Therefore, there are $(n+m)\log_2\left(1 + TX_T^2 \frac{1+c_0^2}{\lambda}\right)$ non-zero terms and each of them is bounded by $2DX_T^2$. □

Lemma A.4. *On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have $R_3 \leq \frac{8cX_T^2 D(1+c_0^2)}{\sqrt{\lambda}} \sqrt{T\beta_T\left(\frac{\delta}{8}\right) \log \frac{det(Z_T)}{det(\lambda I)}}$, where $Z_T := \max_{0 \leq t \leq T} \|z_t\|$, and $\beta_T\left(\frac{\delta}{8}\right)$ is defined in* (15).

Proof. The proof is the same as that of Lemma 13 in [3], and hence omitted. □

Lemma A.5. *On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, $R_4 \leq \frac{1}{\alpha_0}\left(\beta_T\left(\frac{\delta}{4}\right) + \lambda c^2\right)\sqrt{T}$.*

Proof. As defined in Lemma A.1, we have,

$$R_4 = \sum_{t=0}^{T} \left(J^\star(\theta_t) - J^\star(\theta^\star)\right).$$

During the k-th episode, the algorithm chooses $u_t = K(\theta_{t_k})x_t$, $\forall\, t = t_k, t_k+1, \ldots, t_{k+1}$, where, θ_{t_k} is as in (18), and obtained by solving the corresponding optimization problem at the beginning of the episode at time t_k. Therefore R_4 can be written as :

$$R_4 = \sum_{k=0}^{K} \Delta_k, \text{ where } \Delta_k := (t_{k+1} - t_k)\left(J^\star(\theta_{t_k}) - J^\star(\theta^\star)\right).$$

Δ_k is bounded as follows:

$$\Delta_k = (t_{k+1} - t_k)\left(J(\theta_{t_k}) - J(\theta^\star)\right) \leq \frac{(t_{k+1} - t_k)}{\alpha(t_k)}\left(V_{t_k}(\theta^\star) - V_{t_k}(\theta_{t_k})\right), \tag{21}$$

where the inequality holds since θ_{t_k} is a minimizer of $V_{t_k}(\theta) + \alpha(t_k)J^\star(\theta)$ (18). Moreover,

$$\begin{aligned} V_{t_k}(\theta^\star) - V_{t_k}(\theta_{t_k}) &= V_{t_k}(\theta^\star) + V_{t_k}(\hat{\theta}_{t_k}) - V_{t_k}(\hat{\theta}_{t_k}) - V_{t_k}(\theta_{t_k}) \\ &\leq V_{t_k}(\theta^\star) - V_{t_k}(\hat{\theta}_{t_k}), \end{aligned}$$

where the inequality follows since $\hat{\theta}_{t_k}$ is a minimizer of $V_{t_k}(\cdot)$.

Since $\theta^\star \in C_{t_k}(\delta)$, it follows from the definition of the confidence ball that $V_{t_k}(\theta^\star) - V_{t_k}(\hat{\theta}_{t_k}) \leq \beta_{t_k}\left(\frac{\delta}{4}\right)$. Since $\beta_t(\delta/4) \leq \beta_T(\delta/4)$, we have $V_{t_k}(\theta^\star) - V_{t_k}(\hat{\theta}_{t_k}) \leq \beta_T\left(\frac{\delta}{4}\right)$. Therefore,

$$V_{t_k}(\theta^\star) - V_{t_k}(\theta_{t_k}) \leq \beta_T\left(\frac{\delta}{4}\right). \tag{22}$$

Setting $\alpha(t) = a_0\sqrt{T}$, we get

$$\sum_{k=1}^{K} \Delta_k \leq \beta_T\left(\frac{\delta}{4}\right) \sum_{k=1}^{K} \frac{t_{k+1} - t_k}{\alpha_0\sqrt{T}} = \frac{1}{\alpha_0}\beta_T\left(\frac{\delta}{4}\right)\sqrt{T}.$$

□

A.1 Proof of Theorem 4.1

Proof. To analyze regret on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we substitute individual bounds on $R_1,\ R_2,\ R_3$ and R_4 in order to obtain

$$R(T) \leq 2DW^2\sqrt{2T\log\frac{8}{\delta}} + n\sqrt{B_\delta} + 2DX_T^2\log_2 T + \frac{8X_T^2 SD(1+C^2)}{\sqrt{\lambda}}\sqrt{T\beta_T\left(\frac{\delta}{8}\right)\log\frac{det(Z_T)}{det(\lambda I)}} + \frac{1}{\alpha_0}\beta_T\left(\frac{\delta}{4}\right)\sqrt{T}.$$

□

B Definition of d_t

The quantity d_t in the definition of $\mathcal{E}_2(t)$ in (16) is defined as follows,

$$\begin{aligned}
d_t :&= \frac{1}{1-\rho}\left(\frac{\eta}{\rho}\right)^{n+m}\left[2L\sqrt{n\log\frac{4nt(t+1)}{\delta}} + G\mathcal{Z}_T^{\frac{n+m}{n+m+1}}\beta_t\left(\frac{\delta}{4}\right)^{(2(n+d+1))^{-1}}\right],\\
\eta :&= \max\left\{1, \sup_{\theta\in\mathcal{S}}||A^\star + B^\star K(\theta)||\right\},\\
\mathcal{Z}_T :&= \max_{0\leq t\leq T}\|z_t\|,\\
G :&= 2\left(\frac{2c(n+m)^{n+m+0.5}}{\sqrt{U}}\right)^{(n+m+1)^{-1}},\\
U :&= \frac{U_0}{H},\\
U_0 :&= \frac{1}{16^{n+m-2}\max\{1, c^{2(n+m-2)}\}},\\
&H \text{ is constant such that } H > \max\left\{16, \frac{4c^2M^2}{(n+m)U_0}\right\},\\
M :&= \sup_{Y\geq 0}\frac{nL\sqrt{(n+m)\log\left(\frac{1+\frac{TY}{\lambda}}{\delta}\right)} + \lambda^{1/2}c}{Y}.
\end{aligned}$$

C Simulation Experiments

In this section, we provide the details on the simulation experiments, along with some additional results. The code and instructions for replicating the presented results are provided in the supplementary material.

1. We begin by describing the linear systems used for our experiments in Section 5.
   (a) **Unmanned Aerial Vehicle (UAV):** This system represents a linearized dynamics of an unmanned aerial vehicle (UAV) in a two-dimensional plane, which has been recently studied in the context of reinforcement learning in [17, 20]. The first and third states represent the positions, while the second and fourth states represent the velocities in each dimension. The inputs are accelerations in each dimension.

$$A^\star = \begin{bmatrix}1 & 0.5 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0.5\\ 0 & 0 & 0 & 1\end{bmatrix}, B^\star = \begin{bmatrix}0.125 & 0\\ 0.5 & 0\\ 0 & 0.125\\ 0 & 0.5\end{bmatrix}, Q = diag(1, 0.1, 2, 0.2), R = I_2.$$

   (b) **Unstable Laplacian Dynamics** This represents a Laplacian system where the adjacent nodes are weakly connected. The lack of stablility (i.e., $\lambda_{max}(A^\star) \geq 1$) makes it a

challenging example for system identification and hence it has been studied recently in [18, 19, 55, 58, 17]. The system matrices are as follows:

$$A^\star = \begin{bmatrix} 1.01 & 0.01 & 0 \\ 0.01 & 1.01 & 0.01 \\ 0 & 0.01 & 1.01 \end{bmatrix}, B^\star = I_3, Q = I_3, R = I_3.$$

(c) **Large transient dynamics:** We also consider the following unstable system which additionally exhibits large transients.

$$A^\star = \begin{bmatrix} 1 & 0 & 0 \\ 1.1 & 1 & 0 \\ 0 & 1.1 & 1 \end{bmatrix}, B^\star = I_3, Q = I_3, R = I_3.$$

(d) **Longitudinal Flight Control of Boeing 747:** This represents the linearized dynamics of Boeing 747 at 40,000 ft altitute and speed of 774 ft/sec, which was first introduced in [59]. The empirical performance of OFULQ, TS and StabL for this system was recently studied in [17]. The four states represent velocity of aircraft along the body axis, velocity perpendicular to the body axis, angle of the body axis with horizontal and the angular velocity. The inputs are elevator angle and thrust of the aircraft. The system matrices are as follows:

$$A^\star = \begin{bmatrix} 0.99 & 0.03 & -0.02 & -0.32 \\ 0.01 & 0.47 & 4.7 & 0 \\ 0.02 & -0.06 & 0.4 & 0 \\ 0.01 & -0.04 & 0.72 & 0.99 \end{bmatrix}, B^\star = \begin{bmatrix} 0.01 & 0.99 \\ -3.44 & 1.66 \\ -0.83 & 0.44 \\ -0.47 & 0.25 \end{bmatrix}, Q = I_4, R = I_4.$$

2. In our experiments, we compared the empirical performance of Augmented RBMLE-UCB and RBMLE with following algorithms: (1) OFULQ [3], (2) Thompson Sampling [15], (3) StabL [17], (4) Randomized Certainty Equivalence (RCE) [10], and (5) Input Perturbations [16]. The pseudo-code for all of the implemented algorithms is given in Algorithm 2, where the choice of θ_{t_k} and u_t made by each algorithm are described in Table 2. The optimization problems for ARBMLE, RBMLE, OFULQ and StabL described in Table 2 are non-convex problems. We used projected gradient descent to solve the optimization problems. Expression for gradient of the RBMLE objective with respect to θ can be obtained explicitly as in [52].

Algorithm 2 Reinforcement Learning for LQ systems.

Initialize: $t = 0$, $Z_0 = \lambda I_{n+m}$
for $k = 0, 1, \cdots$ **do**
  if $det(Z_t) > 2det(Z_{t_{k-1}})$ **then**
    Calculate θ_t as defined by the RL algorithm (See Table 2).
  else
    $\theta_t = \theta_{t-1}$
  end if
  $u_t = f(K(\theta_t), x_t)$ (See Table 2).
  $Z_{t+1} = Z_t + z_s z_s^T$
end for

3. Initially, the controls are chosen as follows in order to obtain a initial estimates of the system:

$$u_t = K_{\text{init}} x_t + \eta_t \text{ for } 0 \leq t \leq T_{\text{init}} \text{ and } \eta_t \text{ is } \mathcal{N}(0, 1),$$

i.i.d., K_{init} is a stabilizing controller and $x_0 = 0$. The noise is pre-generated, ensuring that initialization is uniform across algorithms. The definition of confidence interval in ARBMLE, TS and StabL depends on the choice of confidence parameter δ and a constant c such that $trace({\theta^\star}^T \theta^\star) < c^2$. StabL algorithm uses an excitation $\mathcal{N}(0, \sigma_w^2)$ for $T < T_w$. The values of various hyper-parameters used in our experiments are described in Table 3.

Algorithm	θ_{t_k}	$u_t,\ \forall\, t\ \in \{t_k, t_k+1, \ldots, t_{k+1}-1\}$
ARBMLE	$\arg\min_{\theta\in\mathcal{S}\cap\mathcal{C}_{t_k}(\delta)} \left\{V_{t_k}(\theta) + \alpha(t_k)J^\star(\theta)\right\},$	$K(\theta_{t_k})x_t$
RBMLE	$\arg\min_{\theta\in\mathcal{S}} \left\{V_{t_k}(\theta) + \alpha(t)J^\star(\theta)\right\},$	$K(\theta_{t_k})x_t$
OFULQ	$\arg\min_{\theta\in\mathcal{S}\cap\mathcal{C}_{t_k}(\delta)} J^\star(\theta)$	$K(\theta_{t_k})x_t$
TS	$\hat{\theta}_{t_k} + \beta_{t_k}(\delta)Z_{t_k}^{-1/2}\mathcal{N}(0,1)$	$K(\theta_{t_k})x_t$
IP	$\hat{\theta}_{t_k}$	$K(\theta_{t_k})x_t + \eta_t^{IP}$
RCE	$\hat{\theta}_{t_k} + \eta_t^{RCE}$	$K(\theta_{t_k})x_t$
StabL	$\arg\min_{\theta\in\mathcal{S}\cap\mathcal{C}_{t_k}(\delta)} J^\star(\theta)$	$\begin{cases} K(\theta_{t_k})x_t + \mathcal{N}(0,\sigma_w^2) \text{ if } T < T_w \\ K(\theta_{t_k})x_t \text{ otherwise} \end{cases}$

Table 2: Choices of θ_{t_k} and u_t for various algorithms.

Parameter	T_{init}	T	δ	λ	α_0	σ_w	T_w
Value	50	500	10^{-4}	10^{-4}	10^{-2}	2	35

Table 3: Values of various parameters

4. We provide simulation results for following additional examples. Figure 3 includes the comparison between ARBMLE, OFULQ, TS and StabL. Figure 4 provides a comparison between ARBMLE, RBMLE, RCE and IP. The average regret values for these systems at $T = 500$ are shown in Table 4.

5. (a) Stabilizable but Not Controllable System: We consider a system studied in [17] which is stabilizable but not controllable. Lack of controllability is challenging for system identification. ARBMLE/RBMLE outperforms OFULQ, TS, StabL and RCE by a significant margin.

$$A^\star = \begin{bmatrix} -2 & 0 & 1.1 \\ 1.5 & 0.9 & 1.3 \\ 0 & 0 & 0.5 \end{bmatrix}, B^\star = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, Q = R = I_3.$$

   (b) Chained Integrator Dynamics: We consider a simple chained integrator system with 2-dimensional states and 2-dimensional input.

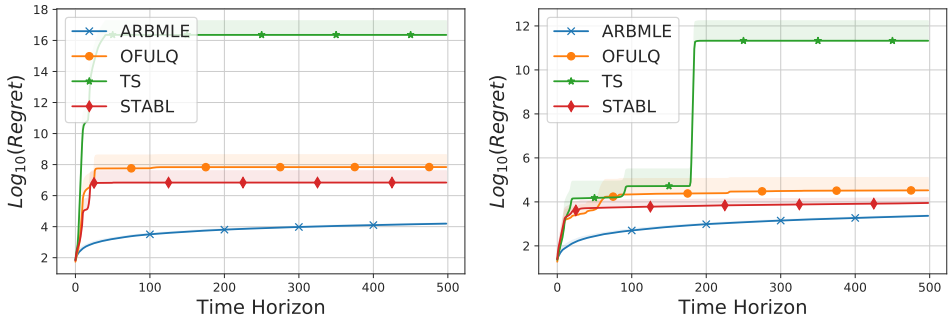
$$A^\star = \begin{bmatrix} 1 & 0.1 \\ 0 & 1 \end{bmatrix}, B^\star = I_2, Q = I_2, R = I_2.$$

Ex.	RBMLE	ARBMLE	OFULQ	TS	IP	RCE	STABL
(a)	15665	15663	6.9×10^7	2.2×10^{16}	15628	39593	6.9×10^6
(b)	2322	2322	33449	2.1×10^{11}	2337	2402	8927

Table 4: Average Regret Performance at $T = 500$.

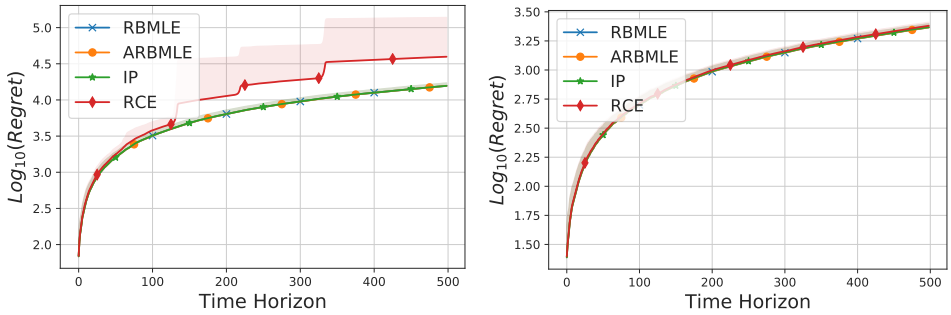
6. Additonal Remarks:

- We use a stabilizing controller K_{init} for initialization of our simulation experiments similar to [18]. Note that our theoretical regret analysis does not assume knowledge of a stabilizing controller, unlike some recent works on adaptive control of LQG systems including [18, 10, 56].
- As demonstrated in the simulation results, OFULQ and Thompson Sampling have a very large initial regret indicating poor initial estimates of system parameters (also highlighted in [17]). ARBMLE/RBMLE, IP and RCE show much better initial regret performance compared to OFULQ and TS.
- Implementation of ARBMLE, OFULQ, StabL and TS involve definition $\beta_t(\delta)$ which denotes boundary of confidence interval. Our simulations for ARBMLE, OFULQ, StabL and TS are based on $\beta_t(\delta)$ as defined in (15). Instead, recent works [18, 17] use $\beta_t(\delta) := trace\left((\theta^\star - \hat{\theta}_t)^\intercal Z_t(\theta^\star - \hat{\theta}_t)\right)$. However, one may note that $\theta^\star$ in $\beta_t(\delta)$ is not known to the learning agent, and so such a definition of $\beta_t(\delta)$ is not a viable for implementation. The effect of the choice of $\beta_t(\delta)$ on the regret performance is shown in the Figure 5.



(a) Stabilizable, not controllable (b) Chained Integrator Dynamics

Figure 3: Logarithm of the Averaged Regret over 50 runs of ARBMLE, OFULQ, TS. and StabL.

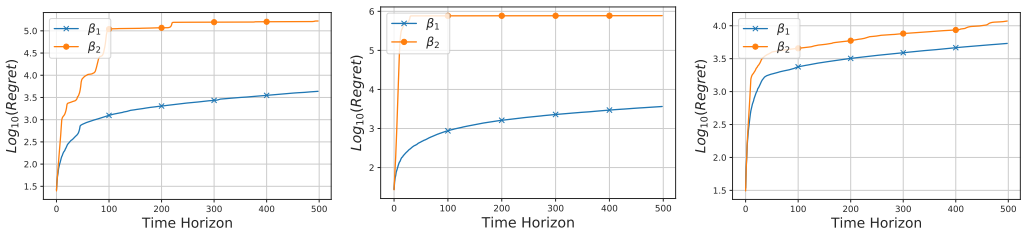


(a) Stabelizable, not controllable (b) Chained Integrator Dynamics

Figure 4: Logarithm of the Averaged Regret over 50 runs of ARBMLE, RBMLE, RCE, TS

- The estimates of OFULQ lies on boundary, while the estimates of ARBMLE/RBMLE, IP and RCE are closer to the least squared estimate. Note that RBMLE can be seen as Lagrangian version of OFULQ, indicating that $\alpha(t)$ may be much smaller than the implicit Lagrange multiplier for OFULQ.

7. The code for our simulation experiments is provided in the supplementary material. The seed values for random number generation are set appropriately for replication of the results. The instruction for the code are provided in supplementary material.



(a) OFULQ (b) TS (c) Stabl

Figure 5: Effect of choice of confidence interval definition on performance. β_1: confidence interval as defined in 15. β_2: confidence interval as defined in [18]