Methods in Extending RL to LQR Adaptive Control: A Survey

Vikram Ramanathan [1]

Abstract

Learning-theoretic work on extending fundamental RL algorithms to LQR Adaptive Control has recently (over the last decade) received more attention. Inspired by the significant progress made independently in RL and adaptive control, this fusion of approaches shows significant promise with potential applications in robotics, plant operations, and game theory (to name a few). This survey is intended to provide the reader with an overview of the recent algorithms and significant theoretical results obtained in extending RL to LQR Adaptive control. This study is done with regards to the benchmark discrete-time, infinite-horizon LQR with unknown dynamics. Regret analyses are presented, and detailed discussions that compare and contrast the methods are provided.

1. Introduction

Reinforcement Learning (RL) methods have largely been studied in the context of finite discrete MDPs and Multi-Armed Bandits. In order to extend these methods to the adaptive control problem setting, these methods must be analyzed and modified to account for continuous state-action spaces, infinite horizons, on-policy learning, and feasible (stabilizable/controllable) sampling. Unfortunately, this task is not trivial. In the past decade, as the works cited in this survey point out, research in this area has born fruit, specifically in the context of the benchmark discrete-time, infinite-horizon, time-invariant Linear Quadratic Regulator (LQR) with unknown dynamics setting.

Note that the adaptive control problem setting is one of on-policy learning i.e. the agent must execute, evaluate and incrementally improve upon its control policy (unlike off-policy learning, where the optimal policy is determined independently of the agent's actions). Thus, the fundamental results proving the convergence of off-policy learning techniques such as Q-learning and policy iteration (Bradtke et al., 1994), and Temporal Difference (TD) learning (Tu & Recht, 2018) do not apply to this problem setting. This survey is primarily concerned with regret analyses, specifically those results that establish finite-time (non-asymptotic) convergence.

With all the recent work in this area, we believe that the papers covered in this survey will provide a strong foundation for future learning-theoretic developments at the interface of RL and adaptive control. We hope this survey is able to:

- Provide significant insights, assumptions, and regret bounds seen in literature on extending well-known methods such as Optimism in the Face of Uncertainty (OFU) and Thompson Sampling (TS) to Adaptive LQR Control
- Compare and contrast the numerous approaches proposed in literature

2. Problem Setting

We are primarily concerned with discrete-time, infinite-horizon, time-invariant LQR,

$$x_{t+1} = A_* x_t + B_* u_t + w_{t+1} \tag{1}$$

where $A_* \in \mathbb{R}^{n \times n}$, $B_* \in \mathbb{R}^{n \times m}$ are unknown matrices. $t = 0, 1, ...$, w_{t+1} is noise. Define the corresponding quadratic cost function as follows,

$$c_t(x_t, u_t) = x_t' Q x_t + u_t' R u_t \tag{2}$$

where $Q \in S_{++}^n$, $R \in S_{++}^m$ and are known. The average expected cost is then given by,

$$J(u_0, u_1, ...) = lim_{T \to \infty} sup \frac{1}{T} \sum_{t=0}^{T} \mathbb{E}[c_t] \tag{3}$$

Therefore, as stated previously, the problem is to design a controller that minimizes the regret, which is defined as,

$$\mathcal{R}(T) = \sum_{t=0}^{T} (c_t - J_*) \tag{4}$$

[1]Harvard University. Correspondence to: Vikram Ramanathan <vramanathan@g.harvard.edu>.

STAT 234 Final Paper, 2021

where J_* is the lowest (optimal) average cost.

Let $\theta'_* = (A_*, B_*)$. Then, the goal of the learning problem is to learn θ'_* and the associated optimal linear control gain, U_*, which can be found using the well-known closed-form solution (Sage, 1968),

$$u_t^* = -(B'_* P_{t+1} B_* + R)^{-1}(B'_* P_{t+1} A_*)x_t = U_* x_t = \pi_t(x_t) \tag{5}$$

where π_t is simply notation for the policy at time t. The matrix P_t is given by the Discrete Algebraic Riccati Equation (DARE),

$$P_t = A'_*(P_{t+1} - P_{t+1}B_*(B'_* P_{t+1} B_* + R)^{-1} B'_* P_{t+1})A_* + Q \tag{6}$$

In general, given that x_t and u_t are random, even when applying the optimal policy, regret, $\mathcal{R}(T)$ is $O(\sqrt{T})$ (given by Prop. 1 below) (Faradonbeh et al., 2017). Thus, this effectively establishes a lower regret bound for any other algorithm.

Proposition 1. Applying optimal control action $u_t^* = U(\theta_*)x_t$, the distribution of $lim_{t\to\infty} \frac{\mathcal{R}(T)}{T^{1/2}}$ is Gaussian centered at zero.

3. Background

3.1. Early Principles

One of the fundamental problems in adaptive control (in general) is the ability to identify open loop system dynamics, θ'_*, from closed-loop measurements, $\theta_* U(\theta_*)$. Attempting to overcome this problem, very early (70s and 80s) research in this area proposed the certainty equivalence principle and forced exploration schemes.

3.1.1. Certainty Equivalence

The Certainty Equivalence (CE) principle states that the closed-loop optimal control can be obtained by using estimated model dynamics as the true model dynamics. For example, a simple least squares estimate in each time step, t, (given by the following optimization problem) can be taken as the true system dynamics and subsequently, utilized to derive a corresponding closed-loop control policy.

$$\min_{A,B} \quad \sum_{i=1}^{t} ||x_t - Ax_{t-1} - Bu_{t-1}||^2 \tag{7}$$

The reader may note that there is a very apparent problem with this approach that most likely arises from the lack of exploration. Granted that the observed state in each time step, x_t, is a direct consequence of the action chosen in the previous time step, u_{t-1}, it is definitely possible for incorrect estimates to cause persistent errant/suboptimal behavior. In fact, it was shown that in the case that θ_* belongs to a known finite set, the least squares estimate can converge with positive probability to a false estimate, causing *strictly* suboptimal long-term average cost (Kumar, 1983). As a result, the authors proposed a *cost-biased* maximum likelihood estimator as an alternative to the simple maximum likelihood parameter estimator (least squares). This approach included an additional term that favors estimates with smaller average costs. While this solution was proven to provide optimal performance for linear systems with quadratic costs, it was under the assumption that the parameter set is finite. Following this work, Campi and Kumar (Campi & Kumar, 1998) extend this result for a cost-biased parameter estimator to the case of infinite yet compact parameter sets, proving asymptotic convergence and optimality. Optimism in the Face of Uncertainty (OFU) principle builds on this result.

3.1.2. Forced Exploration Schemes

Forced exploration schemes differ from the "self-tuning" regulator (i.e. CE-based learning algorithm described in 3.1) in that they actively explore the state-space to find an approximation of the system parameters. Fietcher (Fiechter, 1997b) proposed an episode-based exploration phase in which open-loop control is used to obtain a set of observations, i.e. $\{(x_0, u_0), (x_1, u_1), ..., (x_M, u_M)\}$ where M is the number of exploration episodes. These observations are then used as data points to compute a least squares regression estimator for the system dynamics, θ'_*. Evidently, this is an off-policy learning algorithm. So, the accompanying asymptotic analysis is done in a PAC (Probably Approximately Correct) learning framework. To convert this algorithm to an on-policy algorithm that learns, executes and evaluates at the same time, Feitcher points to the results of (Fiechter, 1997a) which essentially "flips a coin" to choose between exploration or exploitation. Another forced-exploration technique is the inclusion of some perturbation or dither signal to the control action (Caines & Lafortune, 1984). Lai and Ying (Lai & Ying, 1991) introduced "occassional excitation" by adding a dither signal occasionally.

Note that the results presented are proven in the context of asymptotic convergence and lack strong regret bounds.

3.2. Sample Complexity and Regret

An important question for any reinforcement learning algorithm is: How do we measure its performance? Sample complexity and regret are two commonly used frameworks used to study this question. While we encourage the reader

to read the cited papers below for a more detailed understanding of these frameworks, we present a simple overview of the concepts needed for this survey.

Sample complexity is loosely defined as a measure of how much data must be collected to achieve an approximately optimal policy. Formally, Kakade (Kakade, 2003) defines the sample complexity of an algorithm (in the context of RL), $\mathcal{A}$ to be the number of timesteps, t, such that the non-stationary policy at time t, μ_t, is not ϵ-optimal from the current state, s_t, at time t. One important concept in bounding the sample complexity is Probably Approximately Correct (PAC).

In the context of episodic, offline LQR, an algorithm, $\mathcal{A}$ is said to be (ϵ, δ)-PAC if after T episodes it satisfies,

$$Pr[N_\epsilon > (n, m, \frac{H}{\epsilon}, \frac{1}{\delta}] \leq \delta \tag{8}$$

H is the horizon (may be infinite). N_ϵ is the number of episodes for which the policy, π, is not ϵ-optimal ($J_\pi > J_* + \epsilon$). It has been shown that LQR is episodic PAC-learnable using least-squares system identification and robust controller synthesis (Dean et al., 2019). By definition, proving that an algorithm satisfies the PAC-learnable criterion (Eq. 8 is sufficient to bound the sample complexity for that algorithm. However, the PAC framework is only applicable in the context of off-policy learning algorithms and hence, is not relevant to our problem setting.

On the other hand, the regret of an algorithm (defined in the case of LQR in Eq. 4) is loosely defined as the cumulative reward loss incurred by the process of learning. Generally, regret is analyzed in the context of on-policy learning and attempts to quantify the exploration-exploitation trade-off.

In this paper, we also distinguish between Bayesian and Frequentist Regret. Bayesian Regret, as its name implies, adopts the Bayesian view, i.e. we start with a prior over the unknown parameters. More formally,

$$\text{Bayes Regret}(\mathcal{A}, T; \theta) = \mathbb{E}_{\theta \sim p_\theta}\Big[\sum_{t=1}^{T} \mathbb{E}[c_t - J_* | \theta]\Big] \tag{9}$$

$\mathcal{A}$ is some algorithm, T is the time step of concern and p_θ is a prior distribution of the system parameters.

Similar to Bayesian regret, frequentist regret assumes a true but unknown set of parameters but does not start with a prior of these unknown parameters. It is given by Eq. 4.

4. OFU-LQ

Optimism in the Face of Uncertainty (OFU) is a principle that prescribes a "smarter" method than that of CE or forced exploration. Specifically, an algorithm using this principle selects a control law in each update step such that it has the lowest long-term average cost among those that are still unfalsified (i.e. feasible) by the past observations. Originally, OFU was first introduced by Lai and Robbins (Lai & Robbins, 1985) in the context of the exploration/exploitation problem in multi-armed bandits. When applying OFU to adaptive LQR (OFU-LQ), we first establish high-probability confidence set for the model parameters. Then, an optimal controller is designed for each member in the confidence set (in the case of LQR, this is found easily using Eq. 5). Finally, the controller whose average cost is the smallest is chosen (Abbasi-Yadkori & Szepesvári, 2011). In fact, OFU-LQ approaches share the following 3 main ingredients:

- Optimistic parameter estimates
- Lazy/Infrequent updates
- Concentration inequalities for least-squares estimator

These ingredients are evident in the OFU-LQ blueprint shown in Algorithm 1. In each time step, an OFU-LQ algorithm chooses to update the estimated system parameters based on the *update criterion*. If true, some (generally, least-squares) estimate, $\hat{\theta}_t$ is computed. Using $\hat{\theta}_t$, a confidence set is constructed using some pre-defined confidence set constructor, ϕ. Then, $\tilde{\theta}_t$, the element in the confidence set with least long-term average cost is determined. With the optimistic estimated system dynamics, $\tilde{\theta}_t$, we can use Eq. 5, to obtain the corresponding control, u_t. Finally, u_t is executed, the new state, x_{t+1} is observed and the tuple (x_t, u_t, x_{t+1}) is added to the set, $\mathcal{H}_t$. We will look at particular examples of OFU-LQ algorithms later in this section.

Algorithm 1 OFU-LQ Blueprint

Initialize: confidence set constructor ϕ
for *timestep t=0,1,...* **do**
  if *update criterion* **then**
    Compute estimated $\hat{\theta}_t$ (using history, $\mathcal{H}_t$)
    Construct confidence set $\mathcal{S}_t = \phi(\hat{\theta}_t)$
    Compute $\tilde{\theta}_t \in argmin_{\theta \in \mathcal{S}_t} J(\theta)$
  else
    $\tilde{\theta}_t = \tilde{\theta}_{t-1}$
  end
  Calculate $u_t = U(\tilde{\theta}_t) x_t$
  Execute u_t, observe new state x_{t+1}
  Update history $\mathcal{H}_t = \mathcal{H}_t \cup (x_t, u_t, x_{t+1})$
end

Now, we will list some of the important assumptions made in the reviewed papers.

Assumption 1. Unknown true dynamics, θ_* and subsequently, sampled $\tilde{\theta}_t$ (by construction) are controllable and observable

Assumption 2a/b/c. The random variables, w_t, are component-wise (a) sub-Gaussian, (b) Gaussian, (c) sub-Weibull

Assumption 3. [Identifiable] Let $q = n + m$. For a k-sparse matrix $\theta_0 = [A_0, B_0] \in \mathbb{R}^{n \times q}$ and $U \in \mathbb{R}^{n \times m}$, define $\tilde{U} = [I, -U']' \in \mathbb{R}^{q \times n}$ and let $H = \tilde{U}\Lambda\tilde{U}'$ where Λ is given by the solution of, $\Lambda - \theta_0\tilde{U}\Lambda\tilde{U}'\theta_0' = I$. U is (ρ, C_{min}, α) identifiable (with respect to θ_0) i.e. it satisfies the following conditions for all $S \subseteq [q]$, $|S| \leq k$,
(1) $||A_0 - B_0 U|| \leq \rho \leq 1$ [Asymptotically stable]
(2) $\lambda_{min}(H_{SS}) \geq C_{min}$ [Mutual Incoherence]
(3) $||H_{S^cS}H_{SS}^{-}1||_\infty \leq 1 - \alpha$ [Irrepresentable]

Assumption 4. [Stabilizable] $\theta_0 = [A_0, B_0]$ is stabilizable i.e. $\exists U \in \mathbb{R}^{m \times n}$ s.t.,

$$|\lambda_{max}(A_0 + B_0 U)| < 1$$

In Assumption 3, we provide a point of clarification regarding the notation used. For $M \in \mathbb{R}^{m \times n}$, $S \subseteq [m]$, M_{SJ} is the submatrix of M formed by the rows in S and columns in J.

In this survey, we study 3 seminal OFU-LQ papers. Before we delve into the details of each, we present Table 1 which summarizes the important points in each paper. This tabular format provides an effective mechanism to compare and contrast the various methods as well. Abbasi-Yadkori and Szepesvari (2011) (Abbasi-Yadkori & Szepesvári, 2011) is a foundational OFU-LQ paper that was the first to prove $\tilde{O}(\sqrt{T})$ regret (here, $\tilde{O}$ excludes logarithmic factors). They define,

$$\theta_*' = (A_*, B_*) \tag{10}$$

$$z_t = \begin{pmatrix} x_t \\ u_t \end{pmatrix} \tag{11}$$

Therefore, the state transition can be written as,

$$x_{t+1} = \theta_*' z_t + w_{t+1} \tag{12}$$

To calculate the l2-Regularized Least Squares (RLS) estimator, they define the square error, $e(\cdot)$, $\forall \theta \in \mathcal{S}_0$ ($\mathcal{S}_0$ is the set of controllable and observable system parameters)

$$e(\theta) = \lambda Tr(\theta'\theta) + \sum_{s=0}^{t-1} Tr((x_{s+1} - \theta' z_s)(x_{s+1} - \theta' z_s)') \tag{13}$$

Finally, they set $\hat{\theta}_t = argmin_\theta e(\theta) = (Z'Z + \lambda I)^{-1}Z'X$ where $Z = \{z_0', ..., z_{t-1}'\}$ and $X = \{x_1', ..., x_t'\}$ where λ is the regularization coefficient.

The "lazy" update used in the algorithm follows from the definition of the regularized design matrix (which underlies the covariates) given in Eq. 14. Specifically, *update criterion* is given as $det(V_t) > 2det(V_0)$. Note that V_0 is initialized as λI and if the update condition is satisfied, V_0 is set to V_t (please refer to (Abbasi-Yadkori & Szepesvári, 2011) for more details). Intuitively, the update criterion is a measure of how much the update is worth in the given timestep.

$$V_t = \lambda I + \sum_{i=0}^{t-1} z_i z_i' \tag{14}$$

One limitation of Abbasi-Yadkori and Szepesvari's regret bound is that it scales super-exponentially with the dimension of the system dynamics (i.e. m and n). Ibrahami et. al. (2012) (Ibrahimi et al., 2012) attempt to address this drawback by considering OFU-LQ in applications where the system dynamics are sparse and large in dimension (specifically, larger than the time horizon of interest, T). They introduce concepts from sparse signal recovery literature to support their arguments. Their OFU-LQ algorithm and associated regret analysis boast a regret bound that is linear in dimension, n, and still $\tilde{O}(\sqrt{T})$. Additionally, unlike (Abbasi-Yadkori & Szepesvári, 2011), they use $l1$ RLS parameter estimation for $\tilde{\theta}_t$. Ibrahimi et al. utilize an episodic update, i.e., an update is performed once every τ iterations.

Both papers (Abbasi-Yadkori & Szepesvári, 2011) (Ibrahimi et al., 2012) utilize Assumption 1 in their regret analyses. Ibrahimi et al. introduce Assumption 3 i.e. the control is identifiable to aid with their analysis. While Abbasi-Yadkori and Szepesvari use Assumption 2a to characterize the noise, w_t, Ibrahimi et al. use Assumption 2b. This difference is not particularly significant in the derivation of the regret bounds. However, the difference between Assumptions 2a/b and 2c are important since the use of 2c provides a more general result. Fardoneh et. al. (2017) (Faradonbeh et al., 2017) points out that Assumption 1 is a strong assumption which fails when the true dynamics is controllable but still stabilizable (common in sparse, large dimensional systems). Then, the paper proves $\tilde{O}(\sqrt{T})$ by introducing the weaker Assumption 4 (stabilizability). They utilize a simple Least-Squares Estimator (LSE) and an episodic update with τ changing according to a predefined formulation (refer to original paper for details).

We now provide some insight into why OFU works for the LQR adaptive control setting. Let $\tilde{U}' = [I_n, U]$, then $A_0 + B_0 U = \theta_0 \tilde{U}$. Note that by applying linear control gain, U, observations of state vectors will lead to identifi-

cation of the closed-loop matrix, $\theta_* \tilde{U}(\theta_*)$ not θ_*. So why is OFU effective? When the estimator, $\tilde{\theta}$, is *optimistically* chosen and history gets larger with time, the estimation of the closed-loop matrix becomes more precise over time i.e. eventually, $\tilde{\theta}\tilde{U}(\tilde{\theta}) \approx \theta_* \tilde{U}(\tilde{\theta})$. By Lemma 1 (Faradonbeh et al., 2017), $\tilde{U}$ is a near-optimal linear feedback for the system dynamics given by θ_*.

Lemma 1. If $J_*(\theta_1) \leq J_*(\theta_0)$ and $\theta_1 \tilde{U}(\theta_1) = \theta_0 \tilde{U}(\theta_1)$, then $U(\theta_1)$ is an optimal linear feedback for the system evolving according to θ_0

5. TS-LQ

Recent developments in Thompson Sampling (TS) approaches for LQR adaptive control stem from the foundational RL result that posterior sampling generally outperforms OFU-RL algorithms (Osband & Van Roy, 2017). This was shown to hold for finite horizon MDPs with discrete state-action spaces. However, the authors conjecture that these results extend to the ∞-dimensional case. The shortcomings of OFU-RL are apparent:

1. There may be a lack of statistical efficiency, which emerges from possible sub-optimal construction of the confidence sets
2. The computation of the confidence sets along with the complex optimization problem to find $\tilde{\theta}$ may be intractable

TS-LQ can be thought of as a "stochastically optimistic algorithm" that attempts to address these shortcomings in the LQR adaptive control setting. The blueprint of the TS-LQ algorithm is shown in Algorithm 2. In the TS-LQ scheme, similar to OFU-LQ, we can have infrequent updates given by the *update criterion*. When $\tilde{\theta}$ is to be updated, the estimated system parameters, $\hat{\theta}$ are calculated first using least-squares (similar to OFU-LQ). Then, $\tilde{\theta}$ is sampled from the rejection sampling set. The rejection sampling set, $\mathcal{R}_{\mathcal{S}}(f(\hat{\theta}))$ is an operator that samples the given distribution, f, until a feasible i.e. stabilizable element is obtained. Finally, the control, u_t can simply be calculated using Eq. 5 and the set, $\mathcal{H}_t$ updated. As seen in the blueprint algorithm above, TS-LQ is different from OFU-LQ in that it does not perform the computationally intensive task of solving for the confidence set and solving the subsequent optimization for $\tilde{\theta}_t$. Instead, TS-LQ relies on a generic random process informed by the posterior.

The major results in literature for TS-LQ are summarized in Table 2.

In their first paper, Abeille, Lazaric (2017) (Abeille & Lazaric, 2017) comment that the large regret bound, $\tilde{O}(T^{2/3})$, is likely the result of a tradeoff between the parameter sample frequency and the cumulative regret incurred every time the control policy changes. Intuitively, regret can grow unbounded if parameter sampling is not done frequent enough since TS relies on random sampling to generate optimistic models (unlike OFU, where each update is rare and guaranteed to be optimistic). Hence, TS favors short episodes. The algorithm in Abeille, Lazaric (2017) is similar to lazy PSRL (Abbasi-Yadkori & Szepesvari, 2014) but uses a generic random process and throws away the Bayesian structure and Gaussian prior assumption. The "lazy" update is similar to that of Abbasi-Yadkori and Szepesvari (given by $det(V_t) > 2det(V_0)$). With the RLS-estimate, $\hat{\theta}_t$ and the design matrix, V_t, this TS-LQ algorithm samples a perturbed parameter, $\tilde{\theta}$.

Algorithm 2 TS-LQ Blueprint

Initialize: $\hat{\theta_0}$, rejection sampling set $\mathcal{R}_{\mathcal{S}}$
for *timestep t=0,1,...* **do**
  if *update criterion* **then**
    Compute estimated $\hat{\theta}_t$ (using history, $\mathcal{H}_t$)
    Sample $\tilde{\theta} \sim \mathcal{R}_{\mathcal{S}}(f(\hat{\theta}))$
  else
    $\tilde{\theta}_t = \tilde{\theta}_{t-1}$
  end
  Calculate $u_t = U(\tilde{\theta}_t)x_t$
  Execute u_t, observe new state x_{t+1}
  Update history $\mathcal{H}_t = \mathcal{H}_t \cup (x_t, u_t, x_{t+1})$
end

$$\tilde{\theta}_t = \mathcal{R}_{\mathcal{S}}(\hat{\theta}_t + \beta_t(\delta_1) W_t \eta_t) \tag{15}$$

where $W_t = V_t^{-1/2}$, every coordinate of $\eta_t \in \mathbb{R}^{(n+m)\times(n+m)}$ is a random sample drawn i.i.d. from $\mathcal{N}(0,1)$. $\beta_t(\delta_1)$ and δ_1 are given by Thm 2. in (Abbasi-Yadkori et al., 2011) (provided as reference for the interested reader; however, we will not be delving into this further). Now, we will be covering some important (in our opinion) theoretical arguments used in the regret bound proof of this paper (specifically, regret decomposition). First, the paper introduces 3 *concentration events*, i.e. high probabiltity events, $\hat{E}, \tilde{E}, \bar{E}$.

Definition 1. [$\hat{E}_t$ and $\tilde{E}_t$] Let $\delta \in (0,1)$ and $\delta_1 = \delta/(8T)$ and $t \in [0,T]$. We define the event, $\hat{E}_t = \{\forall s \leq t, ||\hat{\theta}_s - \theta_*||_{V_s} \leq \beta_s(\delta_1)\}$ and the event $\tilde{E}_t = \{\forall s \leq t, ||\tilde{\theta}_s - \hat{\theta}_s||_{V_s} \leq \gamma_s(\delta_1)\}$

Definition 2. [$\bar{E}_t$] Let $\delta \in (0,1)$, X_1, X_2 be 2 problem dependent positive constants and $t \in [0,T]$. We define the event, $\bar{E}_t = \{\forall s \leq t, ||x_s|| \leq X_1 log\frac{X_2}{\delta}\}$

Note that the definition of $\hat{E}_t$ is an RLS estimate concen-

Paper	Update	Major Assump.	Estimation Scheme	Regret Bound
Abbasi-Yadkori, Szepesvari (2011)	Lazy	A1, A2a	$l2$ RLS	$\tilde{O}(\sqrt{T})$
Ibrahimi et. al. (2012)	Episodic	A1, A2b, A3	$l1$ RLS	$\tilde{O}(n\sqrt{T})$
Faradoneh et. al. (2017)	Episodic	A2c, A4	LSE	$\tilde{O}(\sqrt{T})$

Table 1. OFU-LQ Results

Paper	Update	Major Assump.	Regret	Regret Bound
Abeille, Lazaric (2017)	Lazy	A2a, A4	Frequentist	$\tilde{O}(T^{2/3})$
Ouyang et. al. (2017)	Lazy	A4	Bayesian	$\tilde{O}(\sqrt{T})$
Abeille, Lazaric (2018)	Always	A4	Frequentist	$\tilde{O}(\sqrt{T})$

Table 2. TS-LQ Results

tration event; $\tilde{E}_t$ ensures that $\tilde{\theta}_s$ concentrates around $\hat{\theta}_s$; and $\bar{E}_t$ bounds the states. Also, $\hat{E} = \hat{E}_T \subset \cdots \subset \hat{E}_1$; $\tilde{E} = \tilde{E}_T \subset \cdots \subset \tilde{E}_1$; $\bar{E} = \bar{E}_T \subset \cdots \subset \bar{E}_1$.

Assumption 5. $\{w_t\}_t$ is a $\mathcal{F}_t$-martingale difference sequence, where $\mathcal{F}_t$ is the filtration which represents the information knowledge up to time t.

The paper shows that with Assumption 5. and conditioned on the filtration, $\mathcal{F}_t$ and the event, $E_t = \hat{E}_t \cap \tilde{E}_t \cap \bar{E}_t$, the regret can be decomposed as follows,

$$\mathcal{R}(T) = \mathcal{R}^{TS} + (\mathcal{R}_1^{RLS} + \mathcal{R}_2^{RLS} + \mathcal{R}_3^{RLS})1\{E_t\} \quad (16)$$

where,

$$\mathcal{R}^{TS} = \sum_{t=0}^{T} \{J(\tilde{\theta}_t) - J(\theta_*)\}1\{E_t\} \quad (17)$$

$$\mathcal{R}_1^{RLS} = \sum_{t=0}^{T} \{\mathbb{E}[x'_{t+1}P(\tilde{\theta}_{t+1}x_{t+1}|\mathcal{F}_t] - x'_tP(\tilde{\theta}_t)x_t\} \quad (18)$$

$$\mathcal{R}_2^{RLS} = \sum_{t=0}^{T} \mathbb{E}[x'_{t+1}(P(\tilde{\theta}_t) - P(\tilde{\theta}_{t+1}))x_{t+1}|\mathcal{F}_t] \quad (19)$$

$$\mathcal{R}_3^{RLS} = \sum_{t=0}^{T} \{z'_t\tilde{\theta}_tP(\tilde{\theta}_t)\tilde{\theta}'_tz_t - z'_t\theta_*P(\tilde{\theta}_t)\theta'_*z_t\} \quad (20)$$

Given that the RLS estimator used in this TS-LQ algorithm is the same as that of OFU, the regret terms, $\mathcal{R}_1^{RLS}$ and $\mathcal{R}_3^{RLS}$ can be bounded using the results of Abbasi-Yadkori, Szepesvári(Abbasi-Yadkori & Szepesvári, 2011). $\mathcal{R}_2^{RLS}$ is affected by changes in $\tilde{\theta}$ between time steps and hence is called the *consistency regret*. $\mathcal{R}^{TS}$ simply evaluates the difference in the optimal average expected cost between the true system parameters, θ_*, and the sampled, $\tilde{\theta}_t$. Hence, $\mathcal{R}^{TS}$ is referred to as the *optimality regret.*. The paper goes on to develop bounds for these regret terms, stating that to do so one must traverse the trade-off between lazy updates to bound consistency regret and frequent updates to bound optimality regret.

Ouyang et. al. (2017) (Ouyang et al., 2017) use similar update criteria as Abeille and Lazaric (2017) (i.e. choosing to update model parameters lazily). Unlike Abeille and Lazaric (2017) (which provides a frequentist regret bound), they prove $\tilde{O}(\sqrt{T})$ Bayesian regret (i.e. under the assumption that there is a prior over the model parameters). Note that in the context of TS-LQ, since the sampled, $\tilde{\theta}_t$ is drawn from a posterior distribution that was constructed from the prior for θ_*, $J(\tilde{\theta}_t$ is the same as $J)\theta_*$ in expectation, i.e. $\mathbb{E}[\mathcal{R}_t^{TS}] = 0$ (Abeille & Lazaric, 2017). Hence, moving from Bayesian regret to Frequentist regret requires $\mathcal{R}_t^{TS}$ to be bounded, which may not be trivial to derive. Frequentist regret bounds are more desired since they don't rely on the selection of a suitable prior.

In 2018, Abeille and Lavaric wrote a second paper (Abeille & Lazaric, 2018) that improved the regret bound to near optimal, $\tilde{O}(\sqrt{T})$, assuming only a lower-bound on the probability of optimistic sampling. This new result stems from a novel bound on the consistency regret, $\mathcal{R}_2^{RLS}$, through the use of frequent updates ($\tilde{\theta}$ updated every time step). They note that by doing so, in the worst case, such an implementation may incur linear regret in $\mathcal{R}_2^{RLS}$ since the difference between the value function for any two different sampled parameters (in successive time steps) is bounded by a constant. However, Lemma 1 in the paper proves that the expected difference in regret decomposition bounds $\mathcal{R}_2^{RLS}$ by $O(\sqrt{T})$. They also update the model and policy every timestep (unlike the other two papers) i.e. *update criterion=True* always.

6. Remarks

This paper specifically highlights OFU and TS based RL approaches in the LQR adaptive control setting. It is clear from our discussion that both approaches have benefits and drawbacks. Regret bounds and the important theoretical arguments used to obtain them were presented as well. Notably, we see that the recent regret bounds for both OFU-LQ (Faradonbeh et al., 2017) and TS-LQ (Abeille & Lazaric, 2018) are near optimal, $\tilde{O}(\sqrt{T})$. However, as noted in the introduction of this paper, regret itself is not a sufficient criterion for optimal performance. There are several other important criteria. For example, from a practical viewpoint, while TS-LQ is the easily implemented (compared to OFU-LQ), it is important to note that the TS-LQ rejection sampling technique may not always be efficient. There are newer methods that continue to emerge, promising more efficient optimistic sampling. These include ideas from robust controller synthesis (Leurent et al., 2020) and coarse-ID control (Dean et al., 2018). However, these methods are out of the scope of this work.

7. Acknowledgements

I would like to thank Susan, Peng, and Kelly for their invaluable advice and support during this project.

References

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

Abbasi-Yadkori, Y. and Szepesvari, C. Bayesian optimal control of smoothly parameterized systems: The lazy posterior sampling algorithm. *arXiv preprint arXiv:1406.3926*, 2014.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pp. 2312–2320, 2011.

Abeille, M. and Lazaric, A. Thompson sampling for linear-quadratic control problems. In *Artificial Intelligence and Statistics*, pp. 1246–1254. PMLR, 2017.

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

Bradtke, S. J., Ydstie, B. E., and Barto, A. G. Adaptive linear quadratic control using policy iteration. In *Proceedings of 1994 American Control Conference-ACC'94*, volume 3, pp. 3475–3479. IEEE, 1994.

Caines, P. and Lafortune, S. Adaptive control with recursive identification for stochastic linear systems. *IEEE Transactions on Automatic Control*, 29(4):312–321, 1984.

Campi, M. C. and Kumar, P. Adaptive linear quadratic gaussian control: the cost-biased approach revisited. *SIAM Journal on Control and Optimization*, 36(6):1890–1907, 1998.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. Regret bounds for robust adaptive control of the linear quadratic regulator. *arXiv preprint arXiv:1805.09388*, 2018.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, pp. 1–47, 2019.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. Finite time analysis of optimal adaptive policies for linear-quadratic systems. *arXiv preprint arXiv:1711.07230*, 2017.

Fiechter, C.-N. Expected mistake bound model for on-line reinforcement learning. In *ICML*, volume 97, pp. 116–124. Citeseer, 1997a.

Fiechter, C.-N. Pac adaptive control of linear systems. In *Proceedings of the tenth annual conference on Computational learning theory*, pp. 72–80, 1997b.

Ibrahimi, M., Javanmard, A., and Roy, B. V. Efficient reinforcement learning for high dimensional linear quadratic systems. In *Advances in Neural Information Processing Systems*, pp. 2636–2644, 2012.

Kakade, S. M. *On the sample complexity of reinforcement learning*. PhD thesis, UCL (University College London), 2003.

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

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

Lai, T. L. and Ying, Z. Parallel recursive algorithms in asymptotically efficient adaptive control of linear stochastic systems. *SIAM journal on control and optimization*, 29(5):1091–1127, 1991.

Leurent, E., Efimov, D., and Maillard, O.-A. Robust-adaptive control of linear systems: beyond quadratic costs. In *NeurIPS 2020-34th Conference on Neural Information Processing Systems*, 2020.

Osband, I. and Van Roy, B. Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*, pp. 2701–2710. PMLR, 2017.

Ouyang, Y., Gagrani, M., and Jain, R. Learning-based control of unknown linear systems with thompson sampling. *arXiv preprint arXiv:1709.04047*, 2017.

Sage, A. P. Optimum systems control. Technical report, SOUTHERN METHODIST UNIV DALLAS TX INFORMATION AND CONTROL SCIENCES CENTER, 1968.

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