The Reward Biased Method: An Optimism based Approach for Reinforcement Learning

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

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

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

***Abstract*—The exploration-exploitation trade-off, also known as the "dual control problem" or the "closed-loop identifiability problem" is a fundamental challenge in reinforcement learning. One of the initial approaches proposed for this problem consisted of adding a bias term that favored models with larger rewards to the likelihood function. This "Reward-Biased" approach was shown to be asymptotically optimal in a variety of contexts including Multi-Armed Bandits (MABs), Markov Decision Processes (MDPs), Linear Quadratic Gaussian (LQG) systems, nonlinear systems, and controlled diffusions. Recent results on regret guarantees and empirical experiments highlight the performance advantage of the Reward-Biased Method. This paper provides an account of recent developments on the finite time analysis of RBMLE along with insights on the reason for its competitive advantage, and identifies some open problems.**

***Index Terms*—Reinforcement Learning, Adaptive Control, Optimism, MDPs, Contextual Bandits, Stochastic Bandits, Linear Quadratic control.**

I. Introduction

Reinforcement learning (RL)/adaptive control focuses on controlling an unknown stochastic system in order to maximize a reward criterion [1]–[9]. Consider a stochastic dynamical system with state $s_t \in \mathcal{S}$ and controls $a_t \in \mathcal{A}$ at time t. The stochastic system is parameterized by $M^\star$ which governs its state transitions:

$$s_{t+1} = f_{M^\star}(s_t, a_t, w_{t+1}),\ t = 1, 2, \ldots,$$

where w_t is noise. For example, a linear system can be parameterized by $M^\star = [A^\star, B^\star]$ where $s_{t+1} = A^\star s_t + B^\star a_t + w_{t+1}$. The stochastic system returns a reward r_t at time t. The learner's goal is to maximize a suitably defined scalar measure J of its reward stream, such as the expected reward over a finite time interval, or the average or discounted reward.

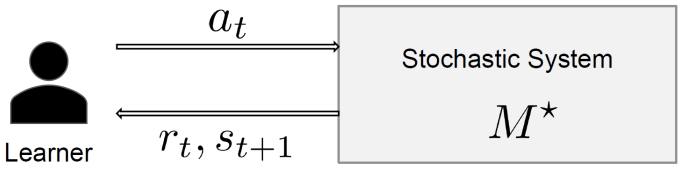

Fig. 1. General Framework for Reinforcement Learning

If $M^\star$ is known, the learner could choose the policy $\pi^\star$ in Π, the set of all non-anticipative policies, which maximizes its objective function. In this paper, we restrict our discussion to the long-term average reward $J(\pi, M^\star)$, although it can be extended to other objectives such as discounted reward, or the "episodic" case involving repeated finite intervals of time.

When $M^\star$ is not known, the learner is faced with a fundamental challenge in reinforcement learning, variously called the dual control problem [10], the closed-loop identifiability problem [11], or the exploration-exploitation trade-off [12]. The learner is faced between the following tactically conflicting choices:

1) Exploration: Collect more information to learn $M^\star$ accurately in order to learn the optimal controller $\pi^\star$.
2) Exploitation: Choose the best controller based on information already gathered.

A wide variety of solutions to this exploration-exploitation trade-off have been proposed in the RL/adaptive control literature. One of the earliest approaches, proposed in [13], consisted of adding a bias term that favored model parameters with larger expected reward to the log-likelihood function, and then using a certainty equivalent control law. It was shown that this optimistic approach yields a long-term average cost reward that is optimal for the unknown system. This result was established under various conditions in a variety of contexts including Multi-Armed Bandits (MABs) [14], Markov Decision Processes (MDPs) [15], Linear Quadratic Gaussian (LQG) systems [16], nonlinear systems, and controlled diffusions. This method is called the "Reward-Biased Method" (RBM). A detailed historical account on the development of the RBM class of algorithms can be found in [14]–[17].

Another notable approach, also based on optimism, the "upper confidence bound (UCB)" approach, was proposed in the seminal work [18]. Rather than using a point estimate it considered the most optimistic parameter in a confidence interval or set. It also introduced the notion of "regret", a finer measure than long-term average reward. Since then, the UCB approach has been adapted for various RL settings including MDPs [8], [9], LQG [19], MABs [12], linear contextual bandits [20], linear MDPs [21], and constrained MDPs [22].

With the advent of reinforcement learning, the focus has intensified on finer objectives such as finite-time performance,

and on issues of computational complexity. The finite time "regret" is defined as

$$R(T) = TJ(\pi^\star, M^\star) - \sum_{t=1}^{T} r_t.$$

After a gap of many years, the Reward-Biased Method has been re-examined vis-a-vis the issues of more contemporaneous interest, its regret guarantees and empirical performance, in the context of various reinforcement learning scenarios:

- Markov Decision Processes [15]
- Linear Quadratic System [16]
- Linear Contextual Bandits [23]
- Stochastic multi-armed bandits [14].

It has been shown to achieve order-optimal regret in these cases. In this paper, we provide an account of these recent developments on RBM. We highlight some potential future research directions of interest and the associated challenges.

II. A General Reinforcement Learning Framework

We consider a general model-based reinforcement learning framework. The stochastic system $M^\star$ has the following state evolution:

$$s_{t+1} = f_{M^\star}(s_t, a_t, w_{t+1}),\ t = 1, 2, \ldots,$$

where s_t is the observed state, a_t is the control action at time t, and w_t is the system noise. The stochastic system returns a reward r_t at time t.

The learner is assumed to have the knowledge of:

1) a compact set $\mathcal{M}$ such that $M^\star \in \mathcal{M}$,
2) a decision space Π such that $\pi^\star \in \Pi$.

At time t, the learner chooses a model estimate M_t from the set $\mathcal{M}$ and policy π_t from the policy space Π.

Such a formulation captures a wide range of reinforcement learning setups including Tabular MDPs, structured bandits and linear control systems. Similar general framework known as Decision making with structured observations (DMSO) has been proposed in [24].

III. The Reward-Biased Method

We consider the average-reward reinforcement learning setting with periodic estimate updates. The Reward-Biased estimate is computed at the beginning of each episode. The length of episode k is denoted by E_k. There are several variants of the RBM. Generally, the RBM estimate M_t of $M^\star$ at time t takes the form:

$$M_t \in \arg\max_{M \in \mathcal{M}} \{\alpha(t) J^\star(M) - D(M, \mathcal{F}_t)\}. \tag{1}$$

where, $\mathcal{F}_t = \{\{s_u\}_{u=1}^t, \{a_u\}_{u=1}^{t-1}\}$ is the collection of states and control inputs observed till t. $\alpha(t)$ is a positive bias-term that grows with t, and $D(\cdot,\cdot) \geq 0$ is a fitting criterion that measures how closely the model M fits the observed data $\mathcal{F}_t$. The certainty-equivalence control policy implemented by the learner is

$$\pi_t \in \arg\max_{\pi \in \Pi} J(\pi, M_t).$$

Algorithm 1 RBM for Reinforcement Learning

Input: $\mathcal{M}, \Pi, \mathcal{S}, \mathcal{A}$
Initialize: $t = 1$.
for k=1,2,...

$$M_t \in \arg\max_{M \in \mathcal{M}} \{\alpha(t) J^\star(M) - D(M, \mathcal{F}_t)\}$$
$$\pi_t \in \arg\max_{\pi \in \Pi} J(\pi, M_t)$$

while $e_k < E_k$ **do**
Implement the control input $a_t = \pi_t(s_t)$
Observe the reward r_t and state s_{t+1}
Set $M_{t+1} = M_t$ and $\pi_{t+1} = \pi_t$
Update $t \to t + 1$
end while do

end for

A. Design of $\alpha(t)$ and $D(\cdot,\cdot)$

The empirical performance and the regret bounds are dependant on the choice of the bias term $\alpha(t)$ and the fitting criterion $D(\cdot,\cdot)$.

- [25] showed that if $\alpha(t)$ is chosen such that

$$\lim_{t\to\infty} \alpha(t) \to \infty \ \text{ and } \lim_{t\to\infty} \frac{\alpha(t)}{t} \to 0,$$

  then the RBMLE algorithm achieves the long-term average optimality of the reward for tabular MDPs. A finer regret analysis can suggest how to choose the bias term $\alpha(t)$ to reduce regret.
- The fitting criterion $D(\cdot,\cdot)$ captures how well the model M fits the observed data. The initial Reward-Biased Maximum Likelihood Estimate (RBMLE) in [25] chose the log-likelihood ratio for D. A fine regret analysis can suggest a preferable choice of D.

The optimal choices of $\alpha(t)$ and $D(\cdot,\cdot)$ for various RL scenarios are discussed in next section.

IV. Finite Time Analysis of RBMLE

In this section, we summarize the recent results on finite time performance of RBMLE in various reinforcement learning contexts.

A. Tabular MDPs

Tabular MDPs with finite states and finite actions have been studied in reinforcement learning, e.g., [8], [9]. Here the goal of the learner is to minimize the regret.

$$M^\star = \{p^\star(s, s', a) :\ \forall\ s, s' \in \mathcal{S} \text{ and } a \in \mathcal{A}\}.$$

In the case where $M^\star$ is known, the optimal average reward and optimal policy can be obtained from dynamic programming [4]. In the RL setup, when $M^\star$ is unknown, the RBMLE algorithm for a tabular MDP chooses M_t as follows:

$$M_t \in \arg\max_{M \in \mathcal{M}} \{\alpha_0 \log t\ J^\star(M) - D(M, \mathcal{F}_t)\}. \tag{2}$$

TABLE I
RBMLE FOR TABULAR MDP: NOTATION

$\mathcal{S}$	$\{1, 2, \cdots, S\}$
$\mathcal{A}$	$\{1, 2, \cdots, A\}$
$\mathcal{M}$	$\Delta^{S\times S\times A}$
Π	Set of stationary policies
r_t	$r(s_t, a_t) \in [0, 1]$
a_t	$\pi_t(s_t)$
$D(M, \mathcal{F}_t)$	log-likelihood of M at time t

Since an optimal policy for an MDP is a stationary policy, the RBMLE optimization can be reduced to an index-based algorithm, where each policy is associated with an index and the algorithm simply chooses the policy with highest value of index.

Theorem 1 (Theorem 9, [15]): The regret of the RBMLE algorithm for Tabular MDPs is upper bounded as:

$$\mathbb{E}[R(T)] \leq C \log T$$

where, C is an instance-dependent constant.

Remark 1: Choice of D: The RBMLE is analyzed with $D(\cdot, \cdot)$ chosen based on the log-likelihood of M; this leads to a sub-optimal pre-constant in the regret bound provided in Theorem 1. The regret analysis show that the pre-constant can be improved by choosing $D(\cdot, \cdot)$ based on the L_1-distance between M and the maximum likelihood estimate $\hat{M}_t$.

Empirical Performance: Figures 2 and 3 compare the empirical regret performance of RBMLE with its UCB and Thompson Sampling [26]–[28] counterparts, namely UCRL2 [9] and TSDE [29], respectively. The regret of RBMLE is seen to be lower than that of UCRL2 and TSDE.

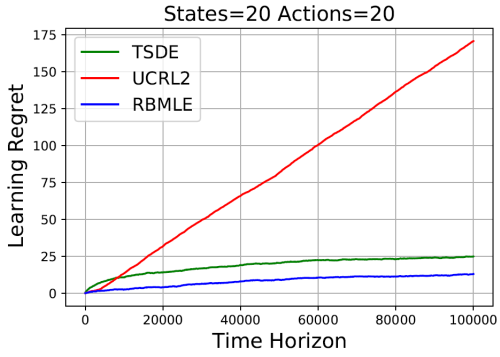

Fig. 2. Regret Performance of RBMLE, UCLR2 and TSDE for randomly generated MDPs with 20 states and 20 actions (Figure 1, [15]).

B. Linear Quadratic Control [16]

The adaptive control of a linear system with a quadratic cost is one of the most extensively studied problems in adaptive control [30], [31]. (A special case where the weight on control is zero, called the self-tuning regulator, is predominant in the control literature [32], [33]). Consider the following linear system:

$$s_{t+1} = A^\star s_t + B^\star a_t + w_{t+1}, \tag{3}$$

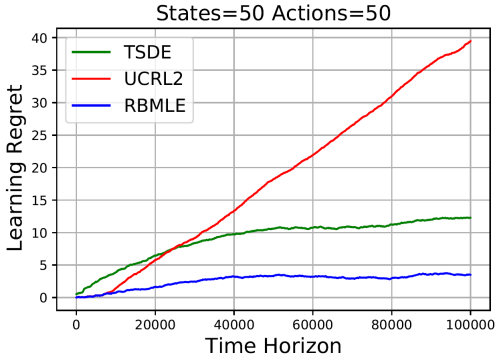

Fig. 3. Regret Performance of RBMLE, UCLR2 and TSDE for a randomly generated MDP with 50 states and 50 actions(Figure 1, [15]).

where the noise w_t is i.i.d. and component-wise sub-Gaussian [34]. The reward at time t is $r(s, a) := -(s^T Q s + a^T R a)$, where $Q \geq 0$ and $R > 0$ are known matrices. When the system parameter $M^\star$ is known, the optimal average reward and optimal controller can be found using Riccati equations [4]. When the system parameter $M^\star = [A^\star, B^\star]$

TABLE II
RBMLE FOR LINEAR QUADRATIC SYSTEM: NOTATION

$\mathcal{S}$	$\mathbb{R}^m$
$\mathcal{A}$	$\mathbb{R}^n$
w_t	Sub-gaussian, A martingale difference sequence wrt $\{F_t\}$
$\mathcal{M}$	$[A, B] : A \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{m\times n}$; [A, B] is stabilizable
Π	$K \in \mathbb{R}^{n\times m}$
r_t	$-(s_t^T Q s_t + a_t^T R a_t)$
a_t	$K_t s_t$
$D(M, \mathcal{F}_t)$	regularized least-squared error at time t

is unknown, the long-term average optimality of RBMLE was established in [30], [31], [35], [36]. The RBMLE estimate is

$$M_t \in \arg\max_{M\in\mathcal{M}} \{\alpha(t) J^\star(M) - V_t(M)\}.$$

where $V_t(M) := \sum_{s=0}^{t-1} (x_{s+1} - Ax_s - Bu_s)^2$ is the squared fitting error of $M = [A, B]$. Reference [19] has proposed an algorithm called OFU (Optimism in the Face of Uncertainty) that is based on the UCB approach. At each time t, it chooses a parameter estimate with maximum average reward within a "confidence set",

$$\mathcal{C}_t(\delta) := \{M = [A, B] : V_t(M) \leq \gamma_t(\delta)\}. \tag{4}$$

In a recent work, [16] has proposed an algorithm, called Augmented RBMLE-UCB (ARBMLE), that brings the fundamental ideas behind RBMLE and OFU together. The ARBMLE algorithm [16] is a constrained version of RBMLE,

$$M_t \in \arg\max_{M\in\mathcal{M}\cap C_t(\delta)} \left\{\alpha_0 \sqrt{t}\ J^\star(M) - V_t(M)\right\}. \tag{5}$$

Theorem 2 (Theorem 4.1, [16]): For any $\delta \in (0,1)$ and $T > 0$, with a probability at least $(1-\delta)$ the regret of the ARBMLE Algorithm is upper-bounded by

$$R(T) = \tilde{\mathcal{O}}\left(\sqrt{T \log \frac{1}{\delta}}\right).$$

Empirical Performance: The empirical regrets of RBMLE as well as ARBMLE are compared with several proposed RL algorithms including OFULQ [37], Thompson Sampling (TS) [38], Input Perturbations (IE) [39], Randomized Certainty Equivalence (RCE) [40], and Stabl [41]. Figures 4 and Fig 5 show the regret performance of these algorithms for a linear model of an Unmanned Aerial Vehicle (UAV) [16]. It is noteworthy that RBMLE and ARBMLE exhibit the "almost" same empirical performance, suggesting that the confidence interval only adds a loose constraint for ARBMLE. This suggests that the constraint that $M \in \cap C_t(\delta)$ in (5) can be deleted. However, we have been unable to prove the regret bounds without this constraint. This remains an open challenge.

Both RBMLE and ARBMLE outperform OFULQ, StabL and TS. The empirical performance of RBMLE/ARBMLE is marginally better than IP and RCE.

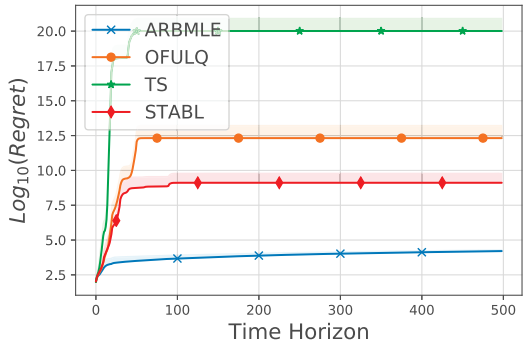

Fig. 4. Regret of ARBMLE, TS, OFULQ, StabL (Figure 1-c, [16]).

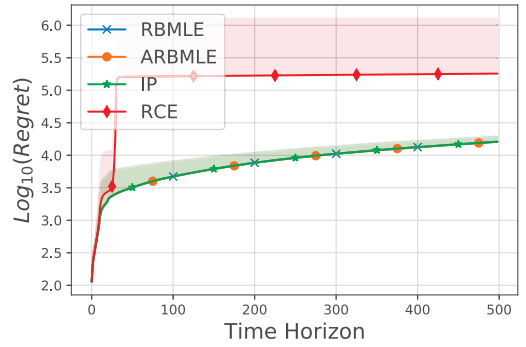

Fig. 5. Regret of RBMLE, ARBMLE, IP, RCE (Figure 2-c, [16]).

Remark 2: Theorem 2 provides regret guarantees only for the ARBMLE algorithm. Providing a regret guarantee for the RBMLE algorithm for LQ control still remains an open problem. As noted earlier, simulation results show the exact same empirical performance for the RBMLE and ARBMLE algorithms. Based on these simulations, one expects a similar regret bound for RBMLE.

C. Linear Contextual Bandits [23]

Linear contextual bandits have found applications in advertisement recommendations, and clinical trials [27]. At each time t a "context vector" $s_t = \{s_{t,i} \in \mathbb{R}^d : i \in [1, A]\}$ is observed. There exists an unknown parameter $M^\star$ such that the conditional mean reward given the past is

TABLE III
RBMLE FOR LINER CONTEXTUAL BANDITS: NOTATION

$\mathcal{S}$	$\varnothing$
$\mathcal{A}$	$\{1, 2, \cdots, A\}$
r_t	${M^\star}^T s_t + w_t$
w_t	Sub-Gaussian noise at time t
a_t	$a_t \in \{1, 2, \cdots, A\}$
$D(M, \mathcal{F}_t)$	log-likelihood of M at time t
λ	regularization parameter

$$\mathbb{E}[r_t|\mathcal{F}_t] = {M^\star}^T s_t.$$

Since $M^\star$ is not known, the learner aims to minimize the "pseudo regret" which is defined as:

$$R_{pseudo}(T) = \sum_{t=1}^{T} {M^\star}^T s_t^\star - {M^\star}^T s_t.$$

The RBMLE algorithm for linear contextual bandits, called as LinRBMLE is an index-based policy. The index of arm a at time t is given by:

$$I_{t,a} = \max_M \{\alpha(t) s_{t,a} - \lambda ||M||^2 + D(M, \mathcal{F}_t)\} \qquad (6)$$

The LinRBMLE algorithm then simply chooses the arm with highest index.

Theorem 3 (Theorem 1, [23]): The regret of LinRBMLE algorithm for linear contextual bandits is

$$R_{psuedo}(T) = \mathcal{O}(d\sqrt{T} \log T).$$

Empirical Performance: In Figures IV-C and IV-C, the performance of LinRBMLE is compared with other popular algorithms including LinUCB [20], LinTS [27], BUCB [42], GPUCB [43], GPUCB Tuned [44], KG, KG* [45]. The computation time for each arm pull as well as the regret are shown. It can be seen that LinRBMLE performs better than all algorithms except GPUCBT. LinRBMLE involves a scalable and efficient computational procedure that also yields a competitive empirical regret.

D. Stochastic Multi-armed Bandits [14]

Stochastic Multi-armed Bandits (MABs) represent perhaps the most simplified and most extensively studied RL setting. A large variety of learning algorithms have been proposed for stochastic MABs. Suppose there are K arms with unknown reward distribution with mean μ_k for arm k.

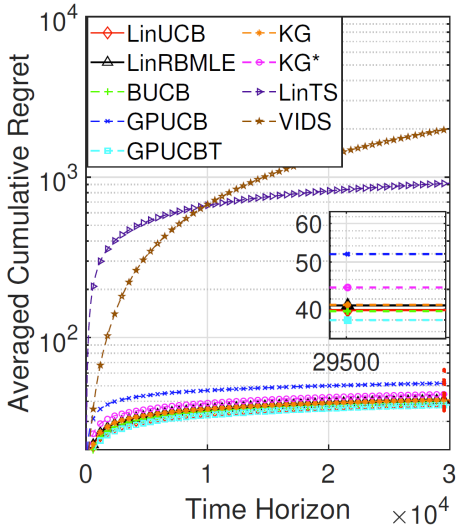

Fig. 6. Linear Contextual Bandits with time varying context vectors, $K = 10$ and $T = 3 \times 10^4$ (Figure 1, [23]).

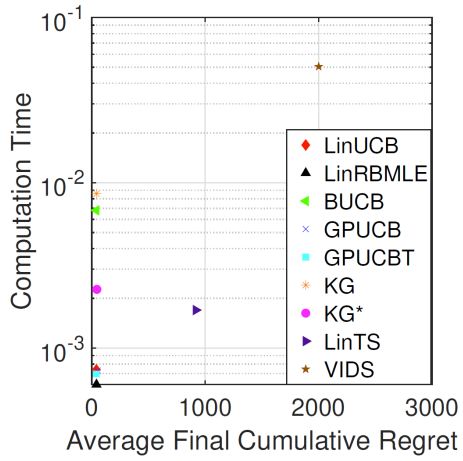

Fig. 7. Average computation time per decision vs. averaged cumulative regret. (Figure 2 , [23]).

The RBMLE algorithm for stochastic bandits can be simplified to an index based policy. The algorithm chooses the arm with highest index value. The RBMLE indices for common reward distributions are provided in Table V.

where, $N_i(t)$ is the number of plays of arm i till t, $H(p)$ is the binary entropy, and $\hat{p}_i(t)$ is MLE estimate of μ_i.

Theorem 4: (Proposition 4, [14]) The regret of RBMLE for a finite family of multi-armed bandita with sub-Gaussian reward distributions is given by:

$$\mathbb{E}[R(T)] \leq C_1 \log T + C_2$$

where C_1, C_2 are problem-dependant constants.

TABLE IV
RBMLE FOR MULTI-ARMED BANDITS: NOTATION

$\mathcal{S}$	$\varnothing$
$\mathcal{A}$	$\{1, 2, \cdots, A\}$
r_t	Sub-Gaussian with mean μ_{a_t}
a_t	$a_t \in \mathcal{A}$

TABLE V
RBMLE INDEX FOR COMMON DISTRIBUTIONS FOR MABS (TABLE 1, [14])

Distribution	RBMLE Index
Bernoulli	$N_i(t)\left(H(p_i(t)) - H(\hat{p}_i(t))\right)$
Exponential	$N_i(t) \log \frac{N_i(t)p_i(t)}{N_i(t)p_i(t)+\alpha(t)}$
Gaussian	$p_i(t) + \frac{\alpha(t)}{2N_i(t)}$

Empirical Performance: In Figure 8, the empirical performance of RBMLE and other leading bandit algorithm is compared. RBMLE outperforms these state-of-the-art bandit algorithms. In Fig 9, the average computational time per pull is plotted against the average regret. Due to the simple form of its index (shown in Table V), RBMLE has low computational complexity. This gives RBMLE an edge over algorithms like IDS and VIDS [46].

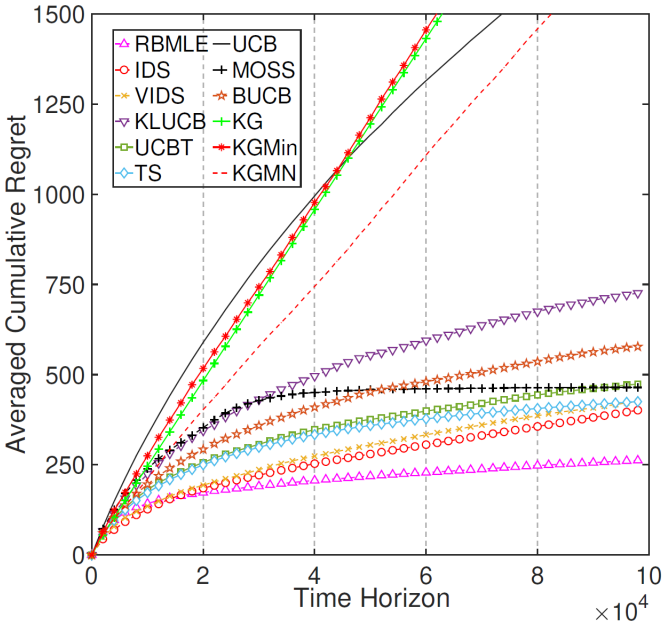

Fig. 8. Regret Performance of various algorithms for stochastic MABs (Figure 1, [14]).

V. DISCUSSION

The RBMLE algorithms overall appear to show a promising empirical performance when compared to state-of-the-art algorithms in various RL scenarios including MDPs [15], LQ control [16], and Stochastic Bandits [14] as well as Linear bandits [23]. This motivates further study of RBMLE. We now outline some insights gleaned from the study and performance of RBMLE, as well as several outstanding problems.

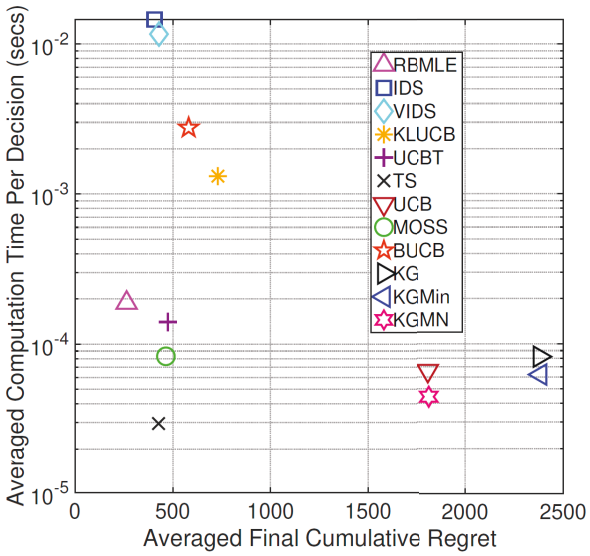

Fig. 9. Average Computational Time vs Average Cumulative regret (Figure 2, [14]).

A. RBMLE vs UCB: Some insights about UCB

RBMLE and UCB algorithms are both based on "optimism under uncertainty". UCB can be regarded as Primal problem, while RBMLE can be regarded as a Lagrangian with a very specific choice of Lagrange multiplier. The UCB algorithm chooses M_t^{UCB} as the solution of the following optimization problem:

$$\max_{M\in\mathcal{M}} J^\star(M)$$
$$\text{such that: } V_t(M) \leq C_t(\delta). \tag{7}$$

On the other hand, the RBMLE algorithm can be written as:

$$\max_{M\in\mathcal{M}} \{J^\star(M) - \frac{1}{\alpha(t)} V_t(M)\}. \tag{8}$$

If one takes (7) as the Primal optimization problem, then (8) is simply the Lagrangian of (7),

$$\max_{M\in\mathcal{M}} \{J^\star(M) - \lambda V_t(M)\},$$

where RBMLE specifically chooses $\lambda = \frac{1}{\alpha(t)}$ as its Lagrange multiplier. However, UCB chooses a different Lagrange multiplier corresponding to whatever is the optimal solution for the Dual of (7).

To compare the Lagrange multipliers chosen by UCB and RBMLE, one can compare their degrees of optimism. Fig. 10 plots their estimation errors. RBMLE's estimation error is smaller than that of UCB. While the UCB solution lies on the boundary of the confidence ellipsoid, the RBMLE solution typically lies strictly in the interior of the confidence ellipsoid, much closer to the true model. Thus, while UCB chooses the *most* optimistic model within the confidence ellipsoid, the degree of optimism in RBMLE is *controlled* by the bias-term $\alpha(t)$, and it chooses a lesser degree of optimism by choosing a larger Lagrange multiplier. Since RBMLE provides a superior performance this suggests that perhaps the optimism of UCB needs to be reduced to obtain better performance.

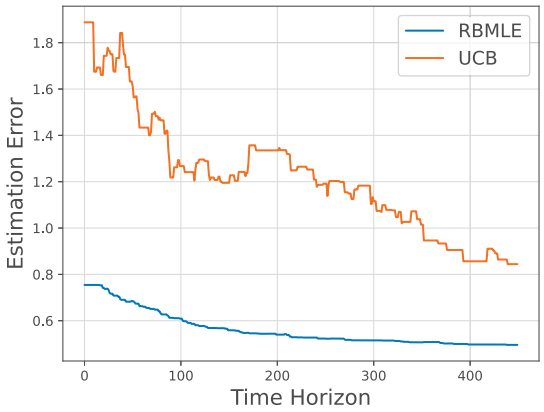

Fig. 10. $||M_t - M^\star||_2$ for RBMLE and UCB (OFULQ) for a linear system.

B. Computational Complexity

The RBMLE algorithm can be reduced to a simple index based policy in the special cases of stochastic multi-armed bandits and the contextual bandits. However due to the non-convex nature of the average reward function in the cases of MDPs and LQG, an efficient computational procedure for RBMLE remains elusive.

C. Other Reinforcement Learning Settings

The Reward Biased approach is a general model-based RL approach. Similar to the Upper Confidence Bound approach that has been widely adopted in a variety of RL settings, RBMLE can also be adapted to a wide-range of reinforcement learning setups including Constrained MDPs [22], linear MDPs [21], Lipschitz Bandits [47], etc.

The RBMLE has been exclusively studied in the long-term average reward criteria until now. Analysis of RBMLE for other popular reward scenarios such as discounted reward and episodic rewards would be an interesting extension.

D. Instant Independent Regret Bounds

All the regret results presented in [14]–[16], [23] are instant-dependant regret bounds. Since [9], there has been significant interest in worst-case regret guarantees especially in MDPs setups. The worst-case regret analysis of RBMLE is an open problem.

VI. Acknowledgement

This material is based upon work partially supported by the US Army Contracting Command under W911NF-22-1-0151 and W911NF2120064, US National Science Foundation under CMMI-2038625, and US Office of Naval Research under N00014-21-1-2385. 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. NSF,

ONR, ARO, or the United States Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Rahul Singh's work was partially supported by the Science and Engineering Research Board through the grant SRG/2021/002308.

REFERENCES

[1] Y. Z. Tsypkin and Z. J. Nikolic, *Adaptation and learning in automatic systems*, vol. 73. Academic Press New York, 1971.

[2] K. J. Åström and B. Wittenmark, *Adaptive control*. Courier Corporation, 2013.

[3] P. R. Kumar, "A survey of some results in stochastic adaptive control," *SIAM Journal on Control and Optimization*, vol. 23, no. 3, pp. 329–380, 1985.

[4] P. R. Kumar and P. Varaiya, *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015.

[5] G. C. Goodwin and K. S. Sin, *Adaptive filtering prediction and control*. Courier Corporation, 2014.

[6] B. Wittenmark, "Adaptive dual control methods: An overview," *Adaptive Systems in Control and Signal Processing 1995*, pp. 67–72, 1995.

[7] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[8] P. Auer and R. Ortner, "Logarithmic online regret bounds for undiscounted reinforcement learning," in *Advances in neural information processing systems*, pp. 49–56, 2007.

[9] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning.," *Journal of Machine Learning Research*, vol. 11, no. 4, 2010.

[10] A. A. Feldbaum, "Dual control theory. i," *Avtomatika i Telemekhanika*, vol. 21, no. 9, pp. 1240–1249, 1960.

[11] V. Borkar and P. Varaiya, "Adaptive control of Markov chains, i: Finite parameter set," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 953–957, 1979.

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2, pp. 235–256, 2002.

[13] A. Becker, P. R. Kumar, and C.-Z. Wei, "Adaptive control with the stochastic approximation algorithm: Geometry and convergence," *IEEE Transactions on Automatic Control*, vol. 30, no. 4, pp. 330–338, 1985.

[14] X. Liu, P.-C. Hsieh, Y. H. Hung, A. 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*, pp. 6248–6258, PMLR, 2020.

[15] A. Mete, R. Singh, X. Liu, and P. R. Kumar, "Reward Biased Maximum Likelihood Estimation for Reinforcement Learning," in *Learning for Dynamics and Control*, pp. 815–827, PMLR, 2021.

[16] A. Mete, R. Singh, and P. R. Kumar, "Augmented rbmle-ucb approach for adaptive control of linear quadratic systems," in *Advances in Neural Information Processing Systems*, 2022.

[17] A. Mete, R. Singh, and P. Kumar, "The rbmle method for reinforcement learning," in *2022 56th Annual Conference on Information Sciences and Systems (CISS)*, pp. 107–112, IEEE, 2022.

[18] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

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

[20] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214, JMLR Workshop and Conference Proceedings, 2011.

[21] L. Yang and M. Wang, "Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound," in *International Conference on Machine Learning*, pp. 10746–10756, PMLR, 2020.

[22] Y. Efroni, S. Mannor, and M. Pirotta, "Exploration-exploitation in constrained mdps," *arXiv preprint arXiv:2003.02189*, 2020.

[23] Y.-H. Hung, P.-C. Hsieh, X. Liu, and P. R. Kumar, "Reward-biased maximum likelihood estimation for linear stochastic bandits," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7874–7882, 2021.

[24] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin, "The statistical complexity of interactive decision making," *arXiv preprint arXiv:2112.13487*, 2021.

[25] P. R. Kumar and A. Becker, "A new family of optimal adaptive controllers for Markov chains," *IEEE Transactions on Automatic Control*, vol. 27, no. 1, pp. 137–146, 1982.

[26] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[27] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *Proceedings of the 30th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.), Proceedings of Machine Learning Research, (Atlanta, Georgia, USA), pp. 127–135, PMLR, 17–19 Jun 2013.

[28] A. Gopalan and S. Mannor, "Thompson sampling for learning parameterized Markov decision processes," in *Conference on Learning Theory*, pp. 861–898, PMLR, 2015.

[29] Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain, "Learning unknown Markov decision processes: A thompson sampling approach," *arXiv preprint arXiv:1709.04570*, 2017.

[30] P. R. Kumar, "Optimal adaptive control of linear-quadratic-gaussian systems," *SIAM Journal on Control and Optimization*, vol. 21, no. 2, pp. 163–178, 1983.

[31] M. Prandini and M. C. Campi, "Adaptive lqg control of input-output systems—a cost-biased approach," *SIAM Journal on Control and Optimization*, vol. 39, no. 5, pp. 1499–1519, 2000.

[32] K. J. Åström, "Computer control of a paper machine—an application of linear stochastic control theory," *IBM Journal of research and development*, vol. 11, no. 4, pp. 389–405, 1967.

[33] R. Singh, A. Mete, A. Kar, and P. Kumar, "Finite time regret bounds for minimum variance control of autoregressive systems with exogenous inputs," *arXiv preprint arXiv:2305.16974*, 2023.

[34] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[35] M. C. Campi and P. R. Kumar, "Adaptive linear quadratic Gaussian control: the cost-biased approach revisited," *SIAM Journal on Control and Optimization*, vol. 36, no. 6, pp. 1890–1907, 1998.

[36] S. Bittanti, M. C. Campi, *et al.*, "Adaptive control of linear time invariant systems: the "Bet on the Best" principle," *Communications in Information & Systems*, vol. 6, no. 4, pp. 299–320, 2006.

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

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

[39] M. K. Shirani Faradonbeh, A. Tewari, and G. Michailidis, "Input perturbations for adaptive control and learning," *Automatica*, vol. 117, p. 108950, 2020.

[40] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "On adaptive linear-quadratic regulators," *Automatica*, vol. 117, p. 108982, 2020.

[41] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Reinforcement learning with fast stabilization in linear dynamical systems," in *International Conference on Artificial Intelligence and Statistics*, pp. 5354–5390, PMLR, 2022.

[42] E. Kaufmann, O. Cappé, and A. Garivier, "On bayesian upper confidence bounds for bandit problems," in *Artificial intelligence and statistics*, pp. 592–600, PMLR, 2012.

[43] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," *arXiv preprint arXiv:0912.3995*, 2009.

[44] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," *Oper. Res.*, vol. 66, p. 230–252, jan 2018.

[45] I. O. Ryzhov, W. B. Powell, and P. I. Frazier, "The knowledge gradient algorithm for a general class of online learning problems," *Operations Research*, vol. 60, no. 1, pp. 180–195, 2012.

[46] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1583–1591, 2014.

[47] S. Magureanu, R. Combes, and A. Proutiere, "Lipschitz bandits: Regret lower bound and optimal algorithms," in *Conference on Learning Theory*, pp. 975–999, PMLR, 2014.