The RBMLE method for Reinforcement Learning

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

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

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

***Abstract*—The Reward Biased Maximum Likelihood Estimate (RBMLE) method was proposed about four decades ago for the adaptive control of unknown Markov Decision Processes, and later studied for more general Controlled Markovian Systems and Linear Quadratic Gaussian systems. It showed that if one could bias the Maximum Likelihood Estimate in favor of parameters with larger rewards then one could obtain long-term average optimality. It provided a reason for preferring parameters with larger rewards based on the fact that generally one can only identify the behavior of a system under closed-loop, and therefore any limiting parameter estimate has to necessarily have lower reward than the true parameter. It thereby provided a reason for what his now called "optimism in the face of uncertainty". It similarly preceded the definition of "regret", and it is only in the last three years that it has been analyzed for its regret performance, both analytically, and in comparative simulation testing. This paper provides an account of the RBMLE method for reinforcement learning.**

***Index Terms*—Reinforcement Learning, LQG Systems, MDPs, Multi-armed bandits.**

I. Introduction

The problem of controlling an unknown dynamical system so as to maximize a reward criterion has been of much interest in adaptive control and learning [1, 2, 3, 4, 5, 6, 7]. Consider a stochastic dynamical system with state and controls at time t denoted by $x_t \in X$ and $u_t \in U$ respectively, and with the state evolution described by

$$x_{t+1} = f_\theta(x_t, u_t, w_{t+1})),\ t = 1, 2, \ldots, \tag{1}$$

where w_t is independent and identically distributed (iid) "noise" at time t, or by its conditional distributions

$$\mathbb{P}(x_{t+1} \in B | \{x_s\}_{s=1}^t, \{u_s\}_{s=1}^t) = P_\theta(x_t, u_t, B), \tag{2}$$

for all Borel sets B; see [8]. The system dynamics is parameterized by θ. The true value of the parameter, denoted by $\theta^\star$ is not known to the system operator, who knows only that it belongs to a compact set Θ. The operator would nevertheless like to choose controls $\{u_t\}$ based on running observations of the state $\{x_t\}$ so as to maximize the expected long-term average reward

$$\liminf_{T\to\infty} \frac{1}{T}\ \mathbb{E}\sum_{t=1}^{T} r(x_t, u_t). \tag{3}$$

If $\theta^\star$ were known, then, under some conditions, one could derive an optimal stationary control policy $\phi^\star_{\theta^\star} : X \mapsto U$ [8]. Since however $\theta^\star$ is not known, the controls $\{u_t\}$ have to serve the following two purposes:

1) Performance optimization: Maximize rewards.
2) Exploring the system behaviour: Collect "enough" information about θ so that one could generate optimal controls at future times.

The dual objectives listed above are typically conflicting, and hence one needs to perform a trade-off between these two.

II. The Certainty Equivalence Approach

The classical approach to dealing with an unknown system has been to make an estimate $\hat{\theta}_t$ of the unknown parameter $\theta^\star$ at each time t, based on the past observations of the state, and then to apply a control input u_t that would be optimal if indeed $\hat{\theta}_t$ were the true parameter. This approach is called "certainty equivalence" (CE), and is based on an approach of separating the estimation and control tasks.

For simplicity of exposition, let us start with the context of a Markov Decision Process (MDP), where the sets X, U, Θ are all finite. The dynamics are described by controlled transition probabilities, $P_\theta(i, u, j) = \mathbb{P}(x_{t+1} = j | x_t = i, u_t = u, \theta$ is the true parameter). Given a trace $\{x_0, u_0, x_1, u_1, \ldots, x_t\}$, let

$$\ell(\theta; t) := \sum_{s=1}^{t} \log P_\theta(x_{s-1}, u_{s-1}, x_s), \tag{4}$$

denote the log-likelihood of the parameter θ based on the information available at time t. The maximum likelihood estimate (MLE) of the unknown system parameter, denoted $\hat{\theta}_t$ satisfies

$$\hat{\theta}_t \in \arg\max \ell(\theta; t). \tag{5}$$

The CE approach applies the control $u_t = \phi^\star_{\hat{\theta}_t}(x_t)$ at time t.

It was shown in [9] that the MLE converges to the true parameter $\theta^\star$, and consequently CE yields the optimal average

This material is based upon work partially supported by NSF Tripods CCF-1934904, NSF CMMI-2038625, U.S. Army Research Office W911NF-18-10331, U.S. Army Research Office W911NF-21-20064, U.S. ONR N00014-21-1-2385, and U.S. Department of Homeland Security 70RSAT20CB0000017. The work of Rahul Singh was partially supported by the SERB Grant SRG/2021/002308.

reward, if the following distinguishability condition is satisfied: For any two parameters $\theta_1, \theta_2 \in \Theta$, with $\theta_1 \neq \theta_2$,

$$P_{\theta_1}(x, u, \cdot) \neq P_{\theta_2}(x, u, \cdot). \tag{6}$$

The condition (6) is however very restrictive; it is not even satisfied in Multi-Armed Bandits (MAB) [10].

In general, [11] showed that the MLE converges to a (random) value $\hat{\theta}(\infty)$ which only satisfies the property,

$$P_{\hat{\theta}(\infty)}(x, \phi^\star_{\hat{\theta}(\infty)}(x), \cdot) = P_{\theta^\star}(x, \phi^\star_{\hat{\theta}(\infty)}(x), \cdot), \tag{7}$$

called "closed-loop identification". It says that *under the limiting controller* $\phi^\star_{\hat{\theta}(\infty)}$, the behaviour of the true system with parameter $\theta^\star$ is the same as the behavior of the system with parameter $\hat{\theta}(\infty)$. It does not imply that $\phi^\star_{\hat{\theta}(\infty)}$ is optimal for $\theta^\star$. The inability to determine $\theta^\star$ under an adaptive control law is called the "closed-loop identifiability problem".

III. The Reward-Biased MLE Method for Long-term Average Optimality

The "Reward-Biased MLE" (RBMLE) method [12] was designed to overcome the closed-loop identifiability problem and converge to an optimal controller. The first observation motivating its design was that the optimal reward for the system described by $\hat{\theta}(\infty)$ is less than that for the true parameter $\theta^\star$. To see this, let $J(\phi, \theta)$ denote the average reward (3) when the stationary policy ϕ is applied to the system θ, and let $J^\star(\theta) := \sup_\phi J(\phi, \theta) = J(\phi^\star_\theta, \theta)$ denote the corresponding optimal average reward. It follows from (7) that

$$J(\phi^\star_{\hat{\theta}(\infty)}, \theta^\star) = J(\phi^\star_{\hat{\theta}(\infty)}, \hat{\theta}(\infty)) = J^\star(\hat{\theta}(\infty)). \tag{8}$$

Since $\phi^\star_{\hat{\theta}(\infty)}$ need not be optimal for $\theta^\star$, we also have

$$J^\star(\theta^\star) \geq J(\phi^\star_{\hat{\theta}(\infty)}, \theta^\star). \tag{9}$$

Upon combining these two inequalities, we infer

$$J^\star(\theta^\star) \geq J^\star(\hat{\theta}(\infty)). \tag{10}$$

Thus MLE converges to a parameter for which the optimal reward is less than the true optimal average reward.

[12] therefore proposed adding a counter-bias, a term proportional to $J^\star(\theta)$, to the log-likelihood, favoring parameters with larger rewards. The resulting RBMLE is

$$\theta_{RB}(t) \in \arg\max_{\theta \in \Theta} \alpha(t) g(J^\star(\theta)) + \ell(\theta; t) \tag{11}$$

where g is any strictly monotonic increasing positive function and $\alpha(t)$ is a weighting factor. RBMLE implements $\phi^\star_{\theta_{RB}(t)}(x_t)$ at time t. [12] showed that if $\alpha(t)$ satisfies $\lim_{t\to\infty} \alpha(t) = \infty$ and $\lim_{t\to\infty} \frac{\alpha(t)}{t} = 0$ then $\theta_{RB}(t)$ converges in a Cesaro sense to $\theta^\star$, and that the average reward obtained is consequently equal to the optimal value $J^\star(\theta^\star)$.

RBMLE was extended to a variety of settings where the long-term average optimality was established:

- Finite state, finite action MDPs where the transition probabilities were not restricted to a finite set in [13].
- More general state spaces with finite parameter set [14].
- Linear, Quadratic, Gaussian (LQG) systems with a finite parameter set in [14].
- Bernoulli bandit problems [15], where it results in an index policy with an explicit simple index.
- Countable state space, with compact action and parameter sets in [16].
- Controlled diffusions in [17] and [18].

IV. Regret and the Upper Confidence Bound Method

In [19], a more stringent criterion of "regret" was introduced to assess the performance of a policy ϕ:

$$R_\phi(T) := TJ^\star(\theta^\star) - E_{\phi,\theta^\star} \sum_{t=1}^{T} r(x_t, u_t). \tag{12}$$

A long-term average optimal policy satisfies $R_{\phi_{\theta^\star}}(T) := o(T)$. However, there could be several policies having a regret of order $o(T)$, and it is of interest to determine the minimal achievable regret, as well as an optimal policy.

The pioneering work of [19] resolved both these issues, which we illustrate below in a simplified context of Bernoulli bandits. Consider N Bernoulli bandits, with probabilities of success $\theta_1 > \theta_2 \geq \theta_3 >\geq \theta_4 \ldots \geq \theta_N$. When Arm i is played it yields a reward of 1 with probability θ_i, and 0 with probability $1-\theta_i$. Clearly the best arm is Arm 1. In the absence of knowing the best arm, [19] showed that for any policy ϕ,

$$\liminf_{T\to+\infty} \frac{1}{\log T} R_\phi(T) \geq \sum_{i=2}^{N} \frac{(\theta_{\max} - \theta_i)}{KL(\theta_{\max}, \theta_i)}, \tag{13}$$

where $KL(p, q)$ is the Kullback-Leibler divergence.

Moreover, [19] also designed a policy that attained the lower-bound (13). It consists of constructing a confidence interval $[0, \text{UCB}_i(t)]$ for each θ_i at each time t, based on the history of past plays. Then it plays that arm i for which $\text{UCB}_i(t)$ is largest. The value of the probability associated with the confidence intervals above is $1 - O(1/t)$. (Several details are omitted and can be found in [19]). This policy is now called an "Upper Confidence Bound" (UCB) policy.

The UCB approach has been used to design near regret-optimal policies for several reinforcement learning problems, including multi-armed bandits [20], contextual bandits where there is a context vector determining a bandit's success [21], MDPs [22, 23], Bayesian Optimization [24], LQG systems [25] and constrained MDPs [26, 27].

V. Optimism in the Face of Uncertainty

The UCB is described as "Optimism in the Face of Uncertainty" (OFU), since it uses the largest of the plausible values to chooses an arm to play [20, 21, 22, 23, 25].

The RBMLE approach provides a rationale for such optimism. Specifically, any reasonable estimator should be able to at least identify the closed-loop transition behavior, yielding (7). This necessarily results in the chain of equalities (8) vis-a-vis the average rewards. Moreover, by its very definition, the Certainty Equivalence approach satisfies (9). As a consequence

one has a one-sided bias (10). This motivates the "optimistic" approach to overcome this bias.

Indeed, the RBMLE approach could also be called optimistic in the face of uncertainty, except that it does so differently from UCB by directly giving increased weight to parameters with larger rewards in the estimation criterion.

VI. Regret Performance of RBMLE

RBMLE was developed before the advent of regret in [19], when the focus in control was on the long-term average criterion. After the initial wave of activity, there was no further work examining its performance with respect to the subsequently defined criterion of regret. Since 2019 there has been a resurgence of activity to theoretically establish its regret performance, as well as its performance in simulation testing against other extensively tested algorithms such as UCB.

Another contender that has also been well studied is the Thompson Sampling approach. This was actually the very first work on bandits [28], though that terminology was not used. It predates even the work of Robbins [29], who was apparently unaware of it. In the context of bandits, it adopted a Bayesian approach with a prior probability distribution of success probability of arms, that is updated to a posterior distribution based on an arm's history of failures and successes. It advocated choosing an arm to play with a probability equal to the posterior probability that it is the best arm. This prescription has performed well in a variety of other contexts [30, 31, 32, 33].

A. Multi-Armed Bandits

The first work in which the regret performance of RBMLE was examined was [34], in the context of multi-armed bandits. Suppose that there are N "arms" [35] and the agent has to "pull" an arm at each time t. Upon pulling arm i, it receives a random reward r_t with distribution $\mathcal{D}_i$, and mean θ_i.

Because many real-world problems such as clinical trials, website optimization, recommendation systems [21], resource allocation in networks [36, 37], etc., can be posed as MABs, this problem has been extensively studied [10, 35]. Since the work [19], efforts have been directed towards deriving versions of UCB such as UCBT [20] and KLUCB [38], with provable finite time bounds of the order of $\mathcal{O}(\log(T))$. A method based on Thompson Sampling [30] has also been shown to attain $\mathcal{O}(\log(T))$ regret. A heuristic called "Information-Directed Sampling" (IDS) [39] shows excellent empirical performance.

The regret of RBMLE algorithm when it is applied to MAB was studied in [34]. It considered the case where the reward distribution $D_i(\theta)$ is a single parameter exponential family of distributions. RBMLE can then be simplified to an index policy, where each arm is characterized at each time by a single number called its index, with the arm having the highest index being chosen for playing, as shown in Table I[1]. It was shown that the expected regret of RBMLE is upper bounded by $C \log T$, where C is a problem-dependant constant.

[1] $N_i(t)$:= number of plays of arm i, $p_i(t)$:= empirical mean for arm i at time t, $\tilde{p}_i(t) := \min\left\{p_i(t) + \frac{\alpha(t)}{N_i(t)}, 1\right\}$, and H is the binary Bernoulli entropy.

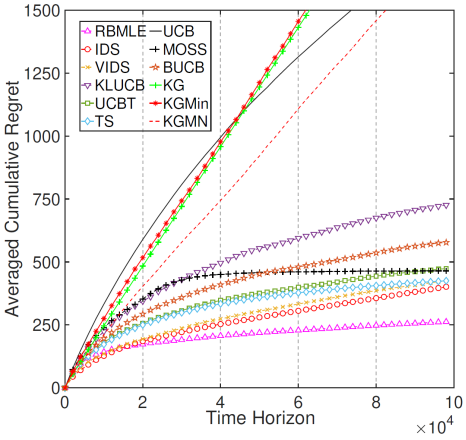

Fig. 1. Bernoulli Bandits with $T = 10^5$ and Number of arms = 10.

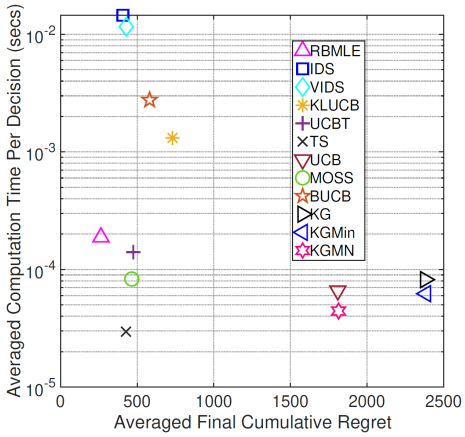

Fig. 2. Average computation time per decision vs. averaged cumulative regret.

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

TABLE I
RBMLE Index for common distributions

Also, the empirical performance of RBMLE was studied against a number of leading contenders in [34]. As shown in Figure 1, it is very competitive against existing state-of-the-art algorithms. Another important consideration is the scalability of an algorithm with respect to the number of arms, with respect to computational complexity. Due to the simple form of its index, RBMLE has low computational complexity as shown in Figure 2, which gives it an advantage over IDS [39].

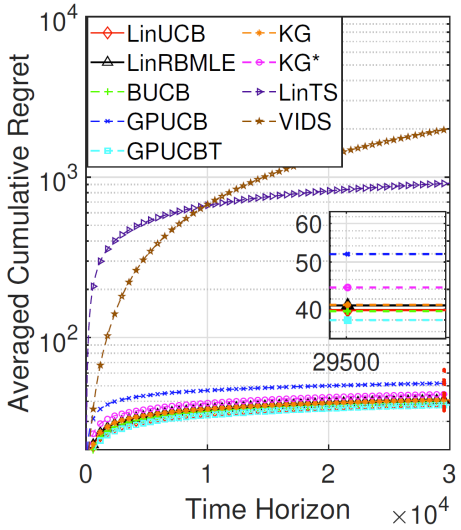

Fig. 3. Linear Contextual Bandits with time varying context vectors, $K = 10$ and $T = 3 \times 10^4$ (Cumulative regret averaged over 50 trials).

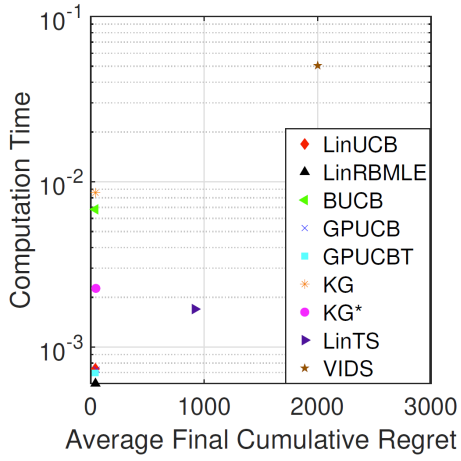

Fig. 4. Average computation time per decision vs. averaged cumulative regret.

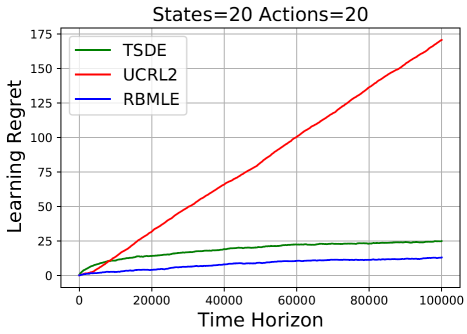

Fig. 5. Comparision of RBMLE with UCRL2 and Thompson Sampling for a randomly generated MDP with 20 states and 20 actions.

B. Contextual Bandits

In many practical scenarios, the rewards of various arms are dependent upon each other. Thus, gaining information about a single arm also helps the agent to infer something about a subset of the remaining arms. At each time t an adversary generates a "context vector" $x_t = \{x_{t,i} \in \mathbb{R}^d : i \in [1, K]\}$. There exists a unknown parameter $\theta^\star$ at time t such that the conditional mean reward given the past is ${\theta^\star}^T x_t + \eta_t$, where η_t is the noise in the reward observation at time t. The choice of an optimal arm at t is given by $a_t^\star \in \arg\max {\theta^\star}^T x_{t,i}$. Since $\theta^\star$ is unknown, the goal is to decide which arm to play at time t so as to minimize the expected psuedo-regret $R(T) = \sum_{t=1}^T {\theta^\star}^T x_t^\star - {\theta^\star}^T x_t$, where $x_t^\star = x_t^{a_t^\star}$.

[40] used the RBMLE approach to solve the contextual bandit problem. The resulting algorithm, dubbed LinRBMLE, was shown to be an index-based policy that achieves a regret of $\mathcal{O}(\sqrt{T}\log T)$. Notably, this regret compares well with Thompson Sampling (LinTS) [**?**], GPUCB with linear kernel [41] and SupLinUCB [42].

In Figure 3 and 4, the performance of LinRBMLE as compared with other popular algorithms in [40] is shown. Similar to the case of stochastic MAB setup, LinRBMLE involves a scalable and efficient computational procedure that also yields a competitive empirical regret.

C. Discrete Markov Decision Processes

Consider an MDP with finite state space X and finite action space U. In the general setting setting described in Section I, the parameter θ is equal to $p(\cdot,\cdot,\cdot)$. It lies in $\Theta = \text{Simplex}^{|U|}(|X|)$, where $\text{Simplex}(n)$ is the probability simplex in n dimensions. This is called an undiscounted reinforcement learning of a discrete MDP [7, 10].

UCB based algorithms such as UCRL [22] and UCRL2 [43] have been shown to achieve $O(\log T)$ regret, as have Thompson Sampling methods such as [44]. The finite-time regret analysis of RBMLE remained has recently been resolved in [45], which showed that RBMLE enjoys an instance-dependent $O(\log T)$ regret.

Figure 5 presents the results of a comparative simulation study of RBMLE with UCRL2 [23] and Thompson Sampling (TSDE) [46] conducted in [45]. The average regret of RBMLE can be seen to be significantly lower than UCLR2 and TSDE.

Extended value iteration [23] is used in to compute the RBMLE estimate (11) in [45], though with no guarantee of convergence to the global maximum. Developing efficient computational methods for (11) in MDP setting is an important open problem.

D. Linear Quadratic Gaussian (and sub-Gaussian) Systems

Consider a linear system

$$x_{t+1} = A^\star x_t + B^\star u_t + w_{t+1}, \tag{14}$$

where the noise w_t is sub-Gaussian [10]. The instantaneous reward incurred at time t is

$$r(x, u) := -(x^T Q x + u^T R u), \tag{15}$$

where $Q \geq 0$ and $R > 0$ are known matrices. When the system parameter $\theta^\star = (A^\star, B^\star)$ is unknown, this is popularly

known as the adaptive LQG (the "G" is strictly only applicable when the noise is Gaussian) control problem or reinforcement learning of an LQG system. The RBMLE approach was analyzed in [47, 48, 49, 50] where the long-term optimality of RBMLE was established in the Gaussian case.

[25] proposed an algorithm called OFU that is based on the UCB approach. At each time t it chooses a parameter estimate with maximum average reward within a "confidence set",

$$\mathcal{C}_t(\delta) := \{\theta = (A, B) : V_t(\theta) \leq \gamma_t(\delta)\}, \tag{16}$$

where $V_t(\theta) := \sum_{s=0}^{t-1} (x_{s+1} - Ax_s - Bu_s)^2$ is the squared fitting error of $\theta = (A, B)$. It was shown that it achieves $\mathcal{O}(\sqrt{T})$ regret.

In recent work, [51] has proposed an algorithm, Augmented RBMLE-UCB, which combines the fundamental idea behind RBMLE as well as OFU. The Augmented RBMLE algorithm in [51] chooses a parameter estimate

$$\theta_t \in \arg \max_{\theta \in \Theta \cap C_t(\delta)} \{-V_t(\theta) + \alpha(t) J^\star(\theta)\}. \tag{17}$$

[51] show that this modified RBMLE has a similar $\mathcal{O}(\sqrt{T})$ regret bound.

VII. Concluding Remarks

We have provided an overview of the RBMLE algorithm that was developed more than four decades ago in [12] to overcome the fundamental challenge of closed-loop identifiability in adaptive control. Its optimality with respect to the long-term average reward criterion was established for a variety of systems. However it was not analyzed for its performance with respect to the the more stringent criterion of "regret" that was subsequently proposed in [19]. Recently there has been a resurgence of work examining its regret performance both theoretically as well in simulation studies against the leading state-of-the-art algorithms, including UCB and its variants, Thompson sampling-based strategies, and heuristics. RBMLE generally has state-of-the-science theoretically established regret, and appears to be very competitive with respect to regret performance in simulations. The reasoning behind the design of RBMLE provides a justification for the use of optimism in reinforcement learning. It provides a systematic approach to the design of reinforcement learning strategies.

References

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

[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," 1986.

[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] O. Hernández-Lerma and J. B. Lasserre, *Further topics on discrete-time Markov control processes*. Springer Science & Business Media, 2012, vol. 42.

[9] P. Mandl, "Estimation and control in markov chains," *Advances in Applied Probability*, pp. 40–60, 1974.

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

[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. 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.

[13] P. R. Kumar and W. Lin, "Optimal adaptive controllers for unknown markov chains," *IEEE Transactions on Automatic Control*, vol. 27, no. 4, pp. 765–774, 1982.

[14] P. R. Kumar, "Simultaneous identification and adaptive control of unknown systems over finite parameter sets," *IEEE Transactions on Automatic Control*, vol. 28, no. 1, pp. 68–76, 1983.

[15] A. Becker and P. Kumar, "Optimal strategies for the n-armed bandit problem," *Univ. Maryland. Baltimore County, Math. Res. Rep*, pp. 81–1, 1981.

[16] V. Borkar, "The Kumar-Becker-Lin scheme revisited," *Journal of optimization theory and applications*, vol. 66, no. 2, pp. 289–309, 1990.

[17] V. S. Borkar, "Self-tuning control of diffusions without the identifiability condition," *J. Optim. Theory Appl.*, vol. 68, no. 1, p. 117–138, jan 1991.

[18] T. Duncan, B. Pasik-Duncan, and L. Stettner, "Almost self-optimizing strategies for the adaptive control of diffusion processes," *Journal of optimization theory and applications*, vol. 81, no. 3, pp. 479–507, 1994.

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

[20] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[21] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 661–670.

[22] P. Auer and R. Ortner, "Logarithmic online regret bounds for undiscounted reinforcement learning," in *Advances in Neural Information Processing Systems*, 2007, pp. 49–56.

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

[24] 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.

[25] 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*. JMLR Workshop and Conference Proceedings, 2011, pp. 1–26.

[26] L. Zheng and L. Ratliff, "Constrained upper confidence reinforcement learning," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 620–629.

[27] R. Singh, A. Gupta, and N. B. Shroff, "Learning in Markov decision processes under constraints," *arXiv preprint arXiv:2002.12435*, 2020.

[28] 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.

[29] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[30] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conference on learning theory*. JMLR Workshop and Conference Proceedings, 2012, pp. 39–1.

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

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

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

[34] 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*. PMLR, 2020, pp. 6248–6258.

[35] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[36] R. Singh and P. R. Kumar, "Throughput optimal decentralized scheduling of multihop networks with end-to-end deadline constraints: Unreliable links," *IEEE Transactions on Automatic Control*, vol. 64, no. 1, pp. 127–142, 2018.

[37] R. Singh and P. Kumar, "Adaptive csma for decentralized scheduling of multi-hop networks with end-to-end deadline constraints," *IEEE/ACM Transactions on Networking*, vol. 29, no. 3, pp. 1224–1237, 2021.

[38] S. Filippi, O. Cappé, and A. Garivier, "Optimism in reinforcement learning and kullback-leibler divergence," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2010, pp. 115–122.

[39] 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.

[40] Y.-H. Hung, P.-C. Hsieh, X. Liu, and P. R. Kumar, "Reward-biased maximum likelihood estimation for linear stochastic bandits," *arXiv preprint arXiv:2010.04091*, 2020.

[41] 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.

[42] 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*. JMLR Workshop and Conference Proceedings, 2011, pp. 208–214.

[43] P. Auer, T. Jaksch, and R. Ortner, "Near-optimal regret bounds for reinforcement learning," in *Advances in neural information processing systems*, 2009, pp. 89–96.

[44] I. Osband, B. Van Roy, and D. Russo, "(more) efficient reinforcement learning via posterior sampling," *Advances in Neural Information Processing Systems*, 2013.

[45] A. Mete, R. Singh, X. Liu, and P. R. Kumar, "Reward biased maximum likelihood estimation for reinforcement learning," in *Learning for Dynamics and Control*. PMLR, 2021, pp. 815–827.

[46] Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain, "Learning unknown Markov Decision Processes: A Thompson Sampling approach," *arXiv preprint arXiv:1709.04570*, 2017.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] A. Mete, R. Singh, and P. R. Kumar, "Augmented RBMLE-UCB Approach for Adaptive Control of Linear Quadratic Systems," *arXiv preprint arXiv:2201.10542*, 2022.