Exploration Through Reward Biasing: Reward-Biased Maximum Likelihood Estimation for Stochastic Multi-Armed Bandits

Xi Liu[*1] Ping-Chun Hsieh[*2] Yu-Heng Hung[2] Anirban Bhattacharya[3] P. R. Kumar[1]

Abstract

Inspired by the Reward-Biased Maximum Likelihood Estimate method of adaptive control, we propose RBMLE – a novel family of learning algorithms for stochastic multi-armed bandits (SMABs). For a broad range of SMABs including both the *parametric* Exponential Family as well as the *non-parametric* sub-Gaussian/Exponential family, we show that RBMLE yields an index policy. To choose the bias-growth rate $\alpha(t)$ in RBMLE, we reveal the nontrivial interplay between $\alpha(t)$ and the regret bound that generally applies in both the Exponential Family as well as the sub-Gaussian/Exponential family bandits. To quantify the finite-time performance, we prove that RBMLE attains *order-optimality* by adaptively estimating the unknown constants in the expression of $\alpha(t)$ for Gaussian and sub-Gaussian bandits. Extensive experiments demonstrate that the proposed RBMLE achieves empirical regret performance competitive with the state-of-the-art methods, while being more computationally efficient and scalable in comparison to the best-performing ones among them.

1. Introduction

Controlling an unknown system to maximize long-term average reward is a well studied adaptive control problem (Kumar, 1985). For unknown Markov Decision Processes (MDPs), Mandl (1974) proposed the certainty equivalent (CE) method that at each stage makes a maximum likelihood estimate (MLE) of the unknown system parameters and then applies the action optimal for that estimate. Specifically, consider an MDP with state space $\mathcal{S}$, action space $\mathcal{U}$, and controlled transition probabilities $p(i,j,u;\boldsymbol{\eta})$ denoting the probability of transition to a next state $s(t+1)=j$ when the current state $s(t)=i$ and action $u(t)=u$ is applied at time t, indexed by a parameter $\boldsymbol{\eta}$ in a set Ξ. The true parameter is $\boldsymbol{\eta}^0 \in \Xi$, but is unknown. A reward $r(i,j,u)$ is obtained when the system transitions from i to j under u. Consider the goal of maximizing the long-term average reward $\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1} r(s(t),s(t+1),u(t))$. Let $J(\phi,\boldsymbol{\eta})$ denote the long-term average reward accrued by the stationary control law $\phi:\mathcal{S}\to\mathcal{U}$ which chooses the action $u(t)=\phi(s(t))$. Let $J_{opt}(\boldsymbol{\eta}) := \max_{\phi} J(\phi,\boldsymbol{\eta}) = J(\phi_{\boldsymbol{\eta}},\boldsymbol{\eta})$ be the optimal long-term reward, and $\phi_{\boldsymbol{\eta}}$ an optimal control law, if the true parameter is $\boldsymbol{\eta}$. Denote by $\widehat{\boldsymbol{\eta}}_t \in \operatorname{argmax}_{\boldsymbol{\eta}\in\Xi}\prod_{\tau=0}^{t-1} p(s(\tau),s(\tau+1),u(\tau);\boldsymbol{\eta})$ a MLE of the true parameter $\boldsymbol{\eta}^0$. Under the CE approach, the action taken at time t is $u(t)=\phi_{\widehat{\boldsymbol{\eta}}_t}(s(t))$. This approach was shown to be sub-optimal by Borkar & Varaiya (1979) since it suffers from the "closed-loop identifiability problem": the parameter estimates $\widehat{\boldsymbol{\eta}}_t$ converge to a $\boldsymbol{\eta}^*$ for which it can only be guaranteed that:

$$p(i,j,\phi_{\boldsymbol{\eta}^*}(i);\boldsymbol{\eta}^*) = p(i,j,\phi_{\boldsymbol{\eta}^*}(i);\boldsymbol{\eta}^0) \text{ for all } i,j, \tag{1}$$

and, in general, $\phi_{\boldsymbol{\eta}^*}$ need not be optimal for $\boldsymbol{\eta}^0$.

To solve this fundamental problem, Kumar & Becker (1982) noticed that due to (1), $J(\phi_{\boldsymbol{\eta}^*},\boldsymbol{\eta}^*) = J(\phi_{\boldsymbol{\eta}^*},\boldsymbol{\eta}^0)$. Since $J(\phi_{\boldsymbol{\eta}^*},\boldsymbol{\eta}^0) \le J_{opt}(\boldsymbol{\eta}^0)$, but $J(\phi_{\boldsymbol{\eta}^*},\boldsymbol{\eta}^*) = J_{opt}(\boldsymbol{\eta}^*)$ due to $\phi^*_{\boldsymbol{\eta}}$ being optimal for $\boldsymbol{\eta}^*$, it follows that $J_{opt}(\boldsymbol{\eta}^*) \le J_{opt}(\boldsymbol{\eta}^0)$, i.e., the parameter estimates are biased in favor of parameters with smaller optimal reward. To undo this bias, with f denoting any strictly monotone increasing fucntion, they suggested employing the RBMLE estimate:

$$\widehat{\boldsymbol{\eta}}_t^{\text{RBMLE}} = \operatorname*{argmax}_{\boldsymbol{\eta}\in\Xi} f(J_{opt}(\boldsymbol{\eta}))^{\alpha(t)} \prod_{\tau=0}^{t-1} p(s(\tau),s(\tau+1),u(\tau),\boldsymbol{\eta}), \tag{2}$$

in the CE scheme, with $u(t)=\phi_{\widehat{\boldsymbol{\eta}}_t^{\text{RBMLE}}}(s(t))$. In (2), $\alpha(t):[1,\infty)\to\mathbb{R}_+$ is allowed to be any function that satisfies $\lim_{t\to\infty}\alpha(t)=\infty$ and $\lim_{t\to\infty}\alpha(t)/t=0$. This method

[*]Equal contribution [1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas, USA [2]Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan [3]Department of Statistics, Texas A&M University, College Station, Texas, USA. Correspondence to: Ping-Chun Hsieh <pinghsieh@nctu.edu.tw>, Xi Liu <xiliu.tamu@gmail.com>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

was shown to yield the optimal long-term average reward in a variety of settings (Kumar & Lin, 1982; Kumar, 1983b;a; Borkar, 1990; 1991; Stettner, 1993; Duncan et al., 1994; Campi & Kumar, 1998; Prandini & Campi, 2000). For the case of Bernoulli bandits, it was shown in Becker & Kumar (1981) that the RBMLE approach provides an index policy where each arm has a simple index, and the policy is to just play the arm with the largest index.

The structure of (2) has a few critical properties that contribute the success of RBMLE. First, the bias term $J_{opt}(\boldsymbol{\eta})^{\alpha(t)}$ multiplying the likelihood encourages active exploration by favoring $\boldsymbol{\eta}$s with potentially higher maximal long-term rewards. Second, the effect of the bias term gradually diminishes as $\alpha(t)$ grows like $o(t)$, which makes the exploitation dominate the estimation at later stages.

Several critical questions need to be answered to tailor the RBMLE to the stochastic multi-armed bandits (SMABs).

1. The average reward optimality proved in prior RBMLE studies is a gross measure implying only that regret (defined below) is $o(T)$, while in SMABs a much stronger $O(\log(T))$ *finite-time* order-optimal regret is desired. What is the regret bound of the RBMLE algorithms in SMABs?

2. The above options for $\alpha(t)$ are very broad, and not all of them lead to order-optimality of regret. How should one choose the function $\alpha(t)$ to attain order-optimality?

3. What are the advantages of RBMLE algorithms compared to the existing methods? Recall that the Upper Confidence Bounds (UCB) approach pioneered by (Lai & Robbins, 1985), and streamlined by (Auer et al., 2002), is conductive to establish the regret bound, but suffers from much higher empirical regret than its counterparts, while the Information Directed Sampling (IDS) (Russo & Van Roy, 2014; 2018a) approach appears to achieve the smallest empirical regret in various bandits, but suffers from high computational complexity with resulting poor scalability due to the calculation or estimation of high dimensional integrals.

The major contributions of this paper are:

1. We show that RBMLE yields an "index policy" for Exponential Family SMABs, and explicitly determine the indices. (An index policy is one where each arm has an index that depends only on its own past performance history, and one simply plays the policy with the highest index). We also propose novel RBMLE learning algorithms for SMABs from sub-Gaussian/Exponential non-parametric families.

2. We reveal the general interplay between the choice of $\alpha(t)$ and the regret bound. When a lower bound on the "minimum gap," the difference between the means of the best and second best arms, and an upper bound on the maximal mean reward are known, simple closed-form indices as well as $O(\log(T))$ order-optimal regret are achieved for reward distributions for both parametric Exponential families as well as sub-Gaussian/Exponential non-parametric families.

3. When the two bounds are unknown, the proposed RBMLE algorithms still attain order-optimality in the Gaussian and sub-Gaussian cases by adaptively estimating them in the index expressions on the fly.

4. We evaluate the empirical performance of RBMLE algorithms in extensive experiments. They demonstrate competitive performance in regret as well as scalability against the current best policies.

2. Problem Setup

Consider an N-armed bandit, where each arm i is characterized by its reward distribution $\mathcal{D}_i$ with mean $\theta_i \in \Theta$, where Θ denotes the set of possible values for the mean rewards. Without loss of generality, let $\theta_1 > \theta_2 \geq \cdots \geq \theta_N \geq 0$. For each arm i, let $\Delta_i := \theta_1 - \theta_i$ be the gap between its mean reward and that of the optimal arm, and $\Delta := \Delta_2$ the "minimum gap." Let $\boldsymbol{\theta}$ denote the vector $(\theta_1, \cdots, \theta_N)$. At each time $t = 1, \cdots, T$, the decision maker chooses an arm $\pi_t \in [N] := \{1, \cdots, N\}$ and obtains the corresponding reward X_t, which is independently drawn from the distribution $\mathcal{D}_{\pi_t}$. Let $N_i(t)$ and $S_i(t)$ be the total number of trials of arm i and the total reward collected from pulls of arm i up to time t, respectively. Define $p_i(t) := S_i(t)/N_i(t)$ as the empirical mean reward up to t. Denote by $\mathcal{H}_t = (\pi_1, X_1, \pi_2, X_2, \cdots, \pi_t, X_t)$ the history of all the choices of the decision maker and the reward observations up to time t. Let $L(\mathcal{H}_t; \{\mathcal{D}_i\})$ denote the likelihood of the history $\mathcal{H}_t$ under the reward distributions $\{\mathcal{D}_i\}$. The objective is to minimize the *regret* defined as $\mathcal{R}(T) := T\theta_1 - \mathbb{E}[\sum_{t=1}^{T} X_t]$, where the expectation is taken with respect to the randomness of the rewards and the employed policy.

3. The RBMLE Policy for Exponential Families and its Indexability

Let the probability density function of the reward obtained from arm i be a one-parameter Exponential Family distribution:

$$p(x; \eta_i) = A(x) \exp\big(\eta_i x - F(\eta_i)\big), \qquad (3)$$

where $\eta_i \in \mathcal{N}$ is the canonical parameter, $\mathcal{N}$ is the parameter space, $A(\cdot)$ is a real-valued function, and $F(\cdot)$ is a real-valued twice-differentiable function. Then (see (Jordan, 2010)), $\theta_i = \dot{F}(\eta_i)$ ("dot" denoting derivative) is the mean of the reward distribution, and its variance is $\ddot{F}(\eta_i)$. Also, (i) $F(\cdot)$ is strictly *convex*, and hence (ii) the function $\dot{F}(\eta)$ is strictly *increasing*, hence invertible. A critical property that will be used to derive the RBMLE is that $\eta_i = \dot{F}^{-1}(\theta_i)$ is strictly monotone increasing in the mean reward θ_i.

We consider the case where the reward distribution of each arm i has the density function $p(\cdot;\eta_i)$, with $F(\cdot)$ and $A(\cdot)$ being identical across all the arms $1 \leq i \leq N$. Let $\eta := (\eta_1, \eta_2, \dots, \eta_N)$ denote the parameter vector that collectively describes the set of all arms. Based on (3), if X_τ is the reward obtained at time τ by pulling arm π_τ, then the likelihood of $\mathcal{H}_t$ at time t under the parameter vector $\boldsymbol{\eta}$ is

$$L(\mathcal{H}_t;\boldsymbol{\eta}) = \prod_{\tau=1}^{t} A(X_\tau)\exp\big(\eta_{\pi_\tau}X_\tau - F(\eta_{\pi_\tau})\big). \tag{4}$$

Now let us consider the reward bias term $f(J_{opt}(\boldsymbol{\eta}))^{\alpha(t)}$ in (2). The optimal reward obtainable from $\boldsymbol{\eta}$ is the maximum of the mean rewards of the arms, i.e., $J_{opt}(\boldsymbol{\eta}) = \max_{1\leq i\leq N}\theta_i$, where $\theta_i = \dot{F}(\eta_i)$ is the mean reward from arm i. By choosing the strictly monotone increasing function $f(\cdot) = \exp(\dot{F}^{-1}(\cdot))$, the reward-bias term reduces to

$$f(J_{opt}(\boldsymbol{\eta}))^{\alpha(t)} := \max_{i\in[N]}\big\{\exp(\eta_i\alpha(t))\big\}. \tag{5}$$

Multiplying the reward-bias term and the likelihood term, the RBMLE estimator is

$$\widehat{\boldsymbol{\eta}}_t^{\text{RBMLE}} := \operatorname*{argmax}_{\boldsymbol{\eta}:\eta_j\in\mathcal{N},\forall j}\big\{L(\mathcal{H}_t;\boldsymbol{\eta})\max_{i\in[N]}\exp(\eta_i\alpha(t))\big\}. \tag{6}$$

The corresponding arm chosen by RBMLE at time t is

$$\pi_t^{\text{RBMLE}} = \operatorname*{argmax}_{i\in[N]}\big\{\widehat{\eta}_{t,i}^{\text{RBMLE}}\big\}. \tag{7}$$

By combining (7) with (6), we have the following index strategy equivalent to (7):

$$\pi_t^{\text{RBMLE}} = \operatorname*{argmax}_{i\in[N]}\Big\{\max_{\boldsymbol{\eta}:\eta_j\in\mathcal{N},\forall j}\big\{L(\mathcal{H}_t;\boldsymbol{\eta})\exp(\eta_i\alpha(t))\big\}\Big\}. \tag{8}$$

The proof of the above result is provided in Appendix A.

3.1. The RBMLE Indices for Exponential Families

An index policy is one where each arm i has an index $I_i(t)$ that is only a function of the history of that arm i up to time t, and the policy is to simply play the arm with the largest index $I_i(t)$ at time t (Whittle, 1980). For the RBMLE policy as written in (8) it is not quite obvious that it is an index policy since the term $L(\mathcal{H}_t;\boldsymbol{\eta})\exp(\eta_i\alpha(t))$ for arm i depends on the combined history $\mathcal{H}_t$ of all arms, including arms other than i. It was recognized in (Becker & Kumar, 1981) that RBMLE yields an index policy for the case of Bernoulli bandits. The following proposition shows that RBMLE is indeed an index policy for the Exponential Family described above, i.e., RBMLE is "indexable," and explicitly identifies what the index of an arm is.[1]

Proposition 1 *The arm selected at time t by the RBMLE algorithm for Exponential Family rewards is*

$$\pi_t^{RBMLE} = \operatorname*{argmax}_{i\in[N]} I(p_i(t), N_i(t), \alpha(t)), \text{ where} \tag{9}$$

$$I(\nu,n,\alpha(t)) = \big(n\nu+\alpha(t)\big)\dot{F}^{-1}\big(\big[\nu+\frac{\alpha(t)}{n}\big]_\Theta\big) \tag{10}$$

$$-n\nu\dot{F}^{-1}(\nu) - nF\big(\dot{F}^{-1}\big(\big[\nu+\frac{\alpha(t)}{n}\big]_\Theta\big)\big) + nF\big(\dot{F}^{-1}(\nu)\big), \tag{11}$$

with $[\cdot]_\Theta$ denoting the clipped value within the set Θ.

Remark 1 The clipping ensures that the input of $\dot{F}^{-1}$ is within $[0,1]$, since, for example, under Bernoulli reward distributions, $p_i(t)+\alpha(t)/N_i(t)$ could be larger than 1.

The indices for three commonly-studied distributions are provided below.

Corollary 1 *For Bernoulli distributions, with $F(\eta) = \log(1+e^\eta)$, $\dot{F}(\eta) = \frac{e^\eta}{1+e^\eta}$, $\dot{F}^{-1}(\theta) = \log(\frac{\theta}{1-\theta})$, $F(\dot{F}^{-1}(\theta)) = \log(\frac{1}{1-\theta})$, and with $\tilde{p}_i(t) := \min\{p_i(t)+\alpha(t)/N_i(t), 1\}$, the RBMLE index is*

$$I(p_i(t), N_i(t), \alpha(t)) = \tag{12}$$

$$N_i(t)\big\{\tilde{p}_i(t)\log\tilde{p}_i(t) + (1-\tilde{p}_i(t))\log(1-\tilde{p}_i(t)) \tag{13}$$

$$-p_i(t)\log(p_i(t)) - (1-p_i(t))\log(1-p_i(t))\big\}. \tag{14}$$

Corollary 2 *For Gaussian reward distributions with the same variance σ^2 across arms, $F(\eta_i) = \sigma^2\eta_i^2/2$, $\dot{F}(\eta_i) = \sigma^2\eta_i$, $\dot{F}^{-1}(\theta_i) = \theta_i/\sigma^2$, and $F(\dot{F}^{-1}(\theta_i)) = \theta_i^2/2\sigma^2$, for each arm i, the RBMLE index is*

$$I(p_i(t), N_i(t), \alpha(t)) = p_i(t) + \frac{\alpha(t)}{2N_i(t)}. \tag{15}$$

Corollary 3 *For Exponential distributions, the index is*

$$I(p_i(t), N_i(t), \alpha(t)) = N_i(t)\log\big(\frac{N_i(t)p_i(t)}{N_i(t)p_i(t)+\alpha(t)}\big). \tag{16}$$

Remark 2 The RBMLE indices derived for parametric distributions can also be applied to *non-parametric* distributions. As shown in Propositions 4 and 5, they still achieve $O(\log(T))$ regret.

Table 1 compares the RBMLE indices with other policies. That these new indices have performance competitive with state-of-the-art (Section 5), is of potential interest.

[1] The proofs of all Lemmas, Corollaries and Propositions are provided in the Appendices.

Table 1: Comparison of indices produced by RBMLE with other approaches. Below, $H(p)$ is the binary entropy, $\overline{V}_t(i)$ is the upper bound on the variance, and the other quantities are defined in Sections 2 and 3.

Algorithm	Index
BMLE	(Bernoulli) $N_i(t)\big(H(p_i(t)) - H(\tilde{p}_i(t))\big)$
	(Gaussian) $p_i(t) + \alpha(t)/(2N_i(t))$
	(Exponential) $N_i(t) \log\big(\frac{N_i(t)p_i(t)}{N_i(t)p_i(t)+\alpha(t)}\big)$
UCB	$p_i(t) + \sqrt{2\log t/N_i(t)}$
UCB-Tuned	$p_i(t) + \sqrt{\min\{\frac{1}{4}, \overline{V}_t(i)\} \log(t)/N_i(t)\}}$
MOSS	$p_i(t) + \sqrt{\max(\log(\frac{T}{N_i(t)\cdot N}), 0)/N_i(t)}$

3.2. Properties of the RBMLE Indices

We introduce several useful properties of the index $I(\nu, n, \alpha(t))$ in (10)-(11) to better understand the behavior of the derived RBMLE indices and prepare for regret analysis in subsequent sections. To begin with, we discuss the dependence of $I(\nu, n, \alpha(t))$ on ν and n.

Lemma 1 *(i) For a fixed $\nu \in \Theta$ and $\alpha(t) > 0$, $I(\nu, n, \alpha(t))$ is strictly decreasing with n, for all $n > 0$.*

(ii) For a fixed $n > 0$ and $\alpha(t) > 0$, $I(\nu, n, \alpha(t))$ is strictly increasing with ν, for all $\nu \in \Theta$.

Since the RBMLE index is $I(p_i(t), N_i(t), \alpha(t))$, Lemma 1.(ii) suggests that the index of an arm increases with its empirical mean reward $p_i(t)$.

To prepare for the following lemmas, we first define a function $\xi(k; \nu) : \mathbb{R}_{++} \to \mathbb{R}$ as

$$\xi(k;\nu) = k\Big[\big(\nu + \frac{1}{k}\big)\dot{F}^{-1}(\nu + \frac{1}{k}) - \nu\dot{F}^{-1}(\nu)\Big] \tag{17}$$

$$- k\Big[F\Big(\dot{F}^{-1}(\nu + \frac{1}{k})\Big) - F(\dot{F}^{-1}(\nu))\Big]. \tag{18}$$

It is easy to verify that $I(\nu, k\alpha(t), \alpha(t)) = \alpha(t)\xi(k; \nu)$. By Lemma 1.(i), we know $\xi(k; \nu)$ is strictly decreasing with k. Moreover, define a function $K^*(\theta', \theta'')$ as

$$K^*(\theta', \theta'') = \inf\{k : \dot{F}^{-1}(\theta') > \xi(k; \theta'')\}. \tag{19}$$

Lemma 2 *Given any pair of real numbers $\mu_1, \mu_2 \in \Theta$ with $\mu_1 > \mu_2$, for any real numbers n_1, n_2 that satisfy $n_1 > 0$ and $n_2 > K^*(\mu_1, \mu_2)\alpha(t)$ (with $K^*(\mu_1, \mu_2)$ being finite), we have $I(\mu_1, n_1, \alpha(t)) > I(\mu_2, n_2, \alpha(t))$.*

Lemma 3 *Given any real numbers $\mu_0, \mu_1, \mu_2 \in \Theta$ with $\mu_0 > \mu_1$ and $\mu_0 > \mu_2$, for any real numbers n_1, n_2 that satisfy $n_1 \le K^*(\mu_0, \mu_1)\alpha(t)$ and $n_2 > K^*(\mu_0, \mu_2)\alpha(t)$, we have $I(\mu_1, n_1, \alpha(t)) > I(\mu_2, n_2, \alpha(t))$.*

Remark 3 *Lemmas 2 and 3 show how RBMLE naturally engages in the exploration vs. exploitation tradeoff. Lemma 2 shows that RBMLE indeed tends to avoid an arm with a smaller empirical mean reward after sufficient exploration, as quantified in terms of $\alpha(t)$ by $n_2 > K^*(\mu_1, \mu_2)\alpha(t)$. On the other hand, Lemma 3 suggests that RBMLE is designed to continue exploration even if the empirical mean reward is initially fairly low (which is reflected by the fact that there is no restriction on the ordering between μ_1 and μ_2 in Lemma 3), when there has been insufficient exploration, as quantified by $n_1 \le K^*(\mu_0, \mu_1)\alpha(t)$. These properties emerge naturally out of the reward biasing.*

4. Regret Analysis of the RBMLE Algorithm

We now analyze the finite-time performance of the proposed RBMLE algorithm. The KL divergence KL$(\eta' \,||\, \eta'')$ between two distributions can be expressed as

$$\mathrm{KL}(\eta' \,||\, \eta'') = F(\eta'') - [F(\eta') + \dot{F}(\eta')(\eta'' - \eta')]. \tag{20}$$

Define $D(\theta', \theta'') : \Theta \times \Theta \to \mathbb{R}_+$ by

$$D(\theta', \theta'') := \mathrm{KL}(\dot{F}^{-1}(\theta') \,||\, \dot{F}^{-1}(\theta'')). \tag{21}$$

4.1. Interplay of Bias-Growth Rate $\alpha(t)$ and Regret

We determine the regret bounds for several classes of distributions, both parametric as well as non-parametric.

4.1.1. Lower-Bounded Exponential Family

We consider the regret performance of RBMLE for Exponential Families with a known lower bound $\underline{\theta}$ on the mean. E.g., $\underline{\theta} = 0$ for Bernoulli distributions. Such a collection includes commonly-studied Exponential Families that are defined on the positive half real line, such as Exponential, Binomial, Poisson, and Gamma (with a fixed shape parameter).

Proposition 2 *For any Exponential Family with a lower bound $\underline{\theta}$ on the mean, for any $\varepsilon \in (0, 1)$, the regret of RBMLE using (10)-(11) with $\alpha(t) = C_\alpha \log t$ and $C_\alpha \ge 4/(D(\theta_1 - \frac{\varepsilon\Delta}{2}, \theta_1)K^*(\theta_1 - \frac{\varepsilon\Delta}{2}, \underline{\theta}))$ satisfies*

$$\mathcal{R}(T) \le \sum_{a=2}^{N} \Delta_a \big[\max\big\{\frac{4}{D(\theta_a + \frac{\varepsilon\Delta_a}{2}, \theta_a)}, \tag{22}$$

$$C_\alpha K^*(\theta_1 - \frac{\varepsilon\Delta_a}{2}, \theta_a + \frac{\varepsilon\Delta_a}{2})\big\} \log T + 1 + \frac{\pi^2}{3}\big]. \tag{23}$$

4.1.2. Gaussian Distributions

Proposition 3 *For Gaussian reward distributions with variance bounded by σ^2 for all arms, the regret of RBMLE using (15) with $\alpha(t) = C_\alpha \log t$ and $C_\alpha \ge \frac{256\sigma^2}{\Delta}$ satisfies*

$$\mathcal{R}(T) \le \sum_{a=2}^{N} \Delta_a \big[\frac{2}{\Delta_a} C_\alpha \log T + \frac{2\pi^2}{3}\big]. \tag{24}$$

4.1.3. Beyond Parametric Distributions

RBMLE indices derived for Exponential Families can be readily applied to other non-parametric distributions. Moreover, the regret proofs in Propositions 2-3 can be readily extended if the non-parametric rewards also satisfy proper concentration inequalities. We consider two classes of reward distributions, namely sub-Gaussian and sub-Exponential (Wainwright, 2019):

Definition 1 *A random variable X with mean $\mu = \mathbb{E}[X]$ is σ-sub-Gaussian if there exists $\sigma > 0$ such that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}},\ \forall \lambda \in \mathbb{R}. \tag{25}$$

Definition 2 *A random variable X with mean $\mu = \mathbb{E}[X]$ is (ρ, κ)-sub-Exponential if there exist $\rho, \kappa \geq 0$ such that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\rho^2\lambda^2}{2}},\ \forall |\lambda| < \frac{1}{\kappa}. \tag{26}$$

Proposition 4 For any σ-sub-Gaussian reward distributions, RBMLE using (15) with $\alpha(t) = C_\alpha \log t$ and $C_\alpha \geq \frac{256\sigma^2}{\Delta}$ yields $\mathcal{R}(T) \leq \sum_{a=2}^{N} \Delta_a \big[\frac{2}{\Delta_a} C_\alpha \log T + \frac{2\pi^2}{3}\big]$.

Proof The proof of Proposition 3 still holds for Proposition 4 without any change as Hoeffding's inequality directly works for sub-Gaussian distributions.

Proposition 5 For any (ρ, κ)-sub-Exponential reward distribution defined on the positive half line with a lower bound $\underline{\theta}$ on the mean, RBMLE using (10)-(11) with $\alpha(t) = C_\alpha \log t$ and $C_\alpha \geq 16(\kappa\varepsilon\Delta + 2\rho^2)/((\varepsilon\Delta)^2 K^*(\theta_1 - \frac{\varepsilon\Delta}{2}, \underline{\theta}))$ achieves a regret bound

$$\mathcal{R}(T) \leq \sum_{a=2}^{N} \Delta_a \big[1 + \frac{\pi^2}{3} + \max\big\{\frac{16(\kappa\varepsilon\Delta + 2\rho^2)}{(\varepsilon\Delta_a)^2}, \tag{27}$$

$$C_\alpha K^*(\theta_1 - \frac{\varepsilon\Delta_a}{2}, \theta_a + \frac{\varepsilon\Delta_a}{2})\big\} \log T\big]. \tag{28}$$

Remark 4 We highlight that the Propositions 2-5 aim to provide the relationship between the upper bound on regret and the gap Δ. We find that to establish the $O(\log(T))$ regret bound, the pre-constant C_α has to be large enough. One of our major technical contributions is to quantify the non-trivial relationship between C_α and Δ as well as the dependency between the bound and Δ. A similar dependency also exists in other algorithms such as UCB, KL-UCB, and IDS, etc, due to the sharp characterization of the pre-constant by (Lai & Robbins, 1985). Moreover, we consider adaptive estimation of the gap as illustrated in Algorithms 1-3. In practice, we show that such adaptive scheme is sufficient to achieve excellent performance (see Section 5).

4.2. RBMLE with Adaptive Estimation of C_α

In this section, we provide the pseudo code of the experiments in Section 5. As discussed above, RBMLE achieves logarithmic regret by estimating C_α in $\alpha(t) = C_\alpha \log t$, where the estimation of C_α involves the minimum gap Δ and the largest mean θ_1. We consider the following adaptive scheme that gradually learns Δ and θ_1.

Algorithm 1 Adaptive Scheme with Estimation of C_α in Gaussian Bandits

```
1: Input: $N$, $\sigma$, and $\beta(t)$
2: for $t = 1, 2, \cdots$ do
3:   for $i = 1$ to $N$ do
4:     $U_i(t) = p_i(t) + \sqrt{2\sigma^2(N+2)\log t / N_i(t)}$ // upper confidence bound of the empirical mean
5:     $L_i(t) = p_i(t) - \sqrt{2\sigma^2(N+2)\log t / N_i(t)}$ // lower confidence bound of the empirical mean
6:   end for
7:   $\hat{\Delta}_t = \max_i \Big\{ \max\big(0, L_i(t) - \max_{j \neq i} U_j(t)\big) \Big\}$
8:   Calculate $\hat{C}_\alpha(t) = \frac{256\sigma^2}{\hat{\Delta}_t}$
9:   $\alpha(t) = \min\{\hat{C}_\alpha(t), \beta(t)\} \log t$
10: end for
```

- **Estimate Δ and θ_1**: Note that Δ can be expressed as $\max_{1 \leq i \leq N}\{\theta_i - \max_{j \neq i} \theta_j\}$. For each arm i, construct $U_i(t)$ and $L_i(t)$ as the upper and lower confidence bounds of $p_i(t)$ based on proper concentration inequalities. Then, construct an estimator of Δ as $\hat{\Delta}_t := \max_{1 \leq i \leq N} \big\{ \max\big(0, L_i(t) - \max_{j \neq i} U_j(t)\big)\big\}$. Meanwhile, we use $U_{\max}(t) := \max_{1 \leq i \leq N} U_i(t)$ as an estimate of θ_1. Based on the confidence bounds, we know $\hat{\Delta}_t \leq \Delta$ and $U_{\max}(t) \geq \theta_1$, with high probability.
- **Construct the bias using estimators**: We construct $\alpha(t) = \min\{\widehat{C}_\alpha(t), \beta(t)\} \log t$, where $\widehat{C}_\alpha(t)$ estimates $C_\alpha(t)$ by replacing Δ with $\hat{\Delta}_t$ and θ_1 with $U_{\max}(t)$, with $\beta(t)$ a non-negative strictly increasing function satisfying $\lim_{t\to\infty} \beta(t) = \infty$ (e.g. $\beta(t) = \sqrt{\log t}$ in the experiments in Section 5). With high probability, $\widehat{C}_\alpha(t)$ gradually approaches the target value $C_\alpha(t)$ from above as time evolves. On the other hand, $\beta(t)$ guarantees smooth exploration initially and will ultimately exceed $\widehat{C}_\alpha(t)$.

4.2.1. (Sub) Gaussian Distributions

To further illustrate the overall procedure, we first use the C_α in Propositions 3 and 4 as an example, with the pseudo code provided in Algorithm 1. The main idea is to learn the appropriate C_α considered in the regret analysis by gradually increasing C_α until it is sufficiently large. This is accomplished by setting $\alpha(t) = \min\{\hat{C}_\alpha(t), \beta(t)\} \log t$ (Line 9 in Algorithm 1) where $\hat{C}_\alpha(t)$ serves as an over-estimate of the minimum required C_α based on the estimators $\hat{\Delta}_t$ for Δ (Lines 3-8 in Algorithm 1). $\hat{\Delta}_t$ is a conservative estimate of Δ in the sense that $\hat{\Delta}_t \leq \Delta$, conditioned on the high probability events $\theta_i \in [L_i(t), U_i(t)]$, for all i. Here the confidence bounds $L_i(t)$ and $U_i(t)$ are constructed with the help of Hoeffding's inequality. For small t, it is expected that $\hat{\Delta}_t$ is very close to zero and hence $\hat{C}_\alpha(t)$ is large. Therefore, initially $\beta(t)$ serves to gradually increase C_α and guarantees enough exploration after $\beta(t)$ exceeds the minimum required C_α. Given sufficient exploration enabled by $\beta(t)$, the estimate $\hat{\Delta}_t$ gets accurate (i.e. $\hat{\Delta}_t \approx \Delta$), and subsequently $\hat{C}_\alpha(t)$ is clamped at some value slightly larger than the minimum required C_α.

Next, we quantify the regret performance of RBMLE in Algorithm 1 in Proposition 6 as follows.

Proposition 6 *For any σ-sub-Gaussian reward distributions, the regret of RBMLE given by Algorithm 1 satisfies*

$$\mathcal{R}(T) \leq \sum_{a=2}^{N} \Delta_a \Big[\max\Big\{ \frac{1024\sigma^2(N+2)}{\Delta^2} \log T, T_0 \Big\} + \frac{N\pi^2}{3} \Big], \tag{29}$$

where $T_0 := \min\{t \in \mathbb{N} : \beta(t) \geq \frac{256\sigma^2(N+2)}{\Delta}\} < \infty$.

4.2.2. Bernoulli Distributions

As above, Algorithm 2 shows the pseudo code for estimating the C_α of $\alpha(t)$ in Bernoulli bandits. Similar to the Gaussian case, C_α is estimated based on $\hat{\Delta}_t$ and $U_{\max}(t)$ with the help of Hoeffding's inequality. In addition, as the calculation of $\hat{C}_\alpha(t)$ involves the subroutine of searching for the value $K^*(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2}, 0)$, we can accelerate the adaptive scheme by first checking if it is possible to have $\hat{C}_\alpha(t) \geq \beta(t)$. Equivalently, this can be done by quickly verifying whether $\xi(\frac{N+2}{2(\varepsilon\hat{\Delta}_t)^2\beta(t)}, 0) < \dot{F}^{-1}(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2})$ (Line 9 in Algorithm 2).

Algorithm 2 Adaptive Scheme with Estimation of C_α in Bernoulli Bandits

```
1: Input: $N$, $\varepsilon \in (0, \frac{1}{2})$, and $\beta(t)$
2: for $t = 1, 2, \cdots$ do
3:   for $i = 1$ to $N$ do
4:     $U_i(t) = \min\big(p_i(t) + \sqrt{(N+2)\log t / N_i(t)}, 1\big)$
       // upper confidence bound of the empirical mean
5:     $L_i(t) = \max\big(p_i(t) - \sqrt{(N+2)\log t / N_i(t)}, 0\big)$
       // lower confidence bound of the empirical mean
6:   end for
7:   $U_{\max}(t) = \max_{i=1,\cdots,N} U_i(t)$
8:   $\hat{\Delta}_t = \max_i \Big\{ \max\big(0, L_i(t) - \max_{j \neq i} U_j(t)\big) \Big\}$
9:   if $\xi\big(\frac{N+2}{2(\varepsilon\hat{\Delta}_t)^2\beta(t)}, 0\big) < \dot{F}^{-1}(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2})$ then
10:    $\alpha(t) = \beta(t)\log t$ // In this case, we know $\hat{C}_\alpha(t) > \beta(t)$
11:  else
12:    Find $\hat{C}_\alpha(t) = \frac{N+2}{2(\varepsilon\hat{\Delta}_t)^2 K^*(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2}, 0)}$ by solving the minimization problem of (19) for $K^*(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2}, 0)$.
13:    $\alpha(t) = \min\{\hat{C}_\alpha(t), \beta(t)\}\log t$
14:  end if
15: end for
```

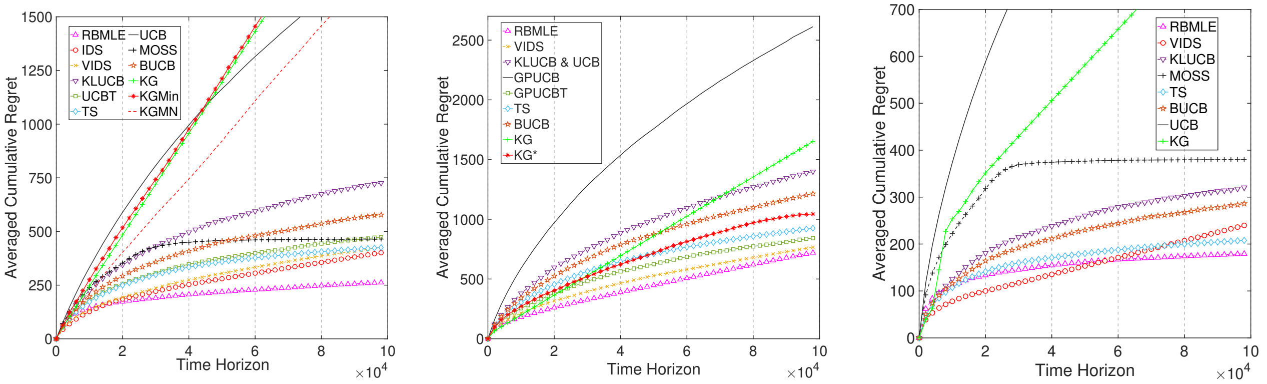
4.2.3. Exponential Distributions

Algorithm 3 demonstrates the pseudo code for selecting $\alpha(t)$ in exponential bandits. Compared to the Bernoulli case, the main difference in the exponential case lies in the construction of the confidence bounds (Lines 4-5 in Algorithm 3), which leverage the sub-exponential tail bounds instead of Hoeffding's inequality.

5. Simulation Experiments

To evaluate the performance of the proposed RBMLE algorithms, we conduct a comprehensive empirical comparison with other state-of-the-art methods vis-a-vis three aspects: effectiveness (cumulative regret), efficiency (computation time per decision vs. cumulative regret), and scalability (in number of arms). We paid particular attention to the fairness of comparisons and reproducibility of results. To ensure sample-path sameness for all methods, we considered each method over a pre-prepared dataset containing the context of each arm and the outcomes of pulling each arm over all rounds. Hence, the outcome of pulling an arm is obtained by querying the pre-prepared data instead of calling the random generator and changing its state. A few benchmarks such as Thompson Sampling (TS) and Variance-based Information Directed Sampling (VIDS) that rely on outcomes of random sampling in each round of decision-making are separately evaluated with the same prepared data and with the same seed. To ensure reproducibility of experimental results, we set up the seeds for the random number generators at the beginning of each experiment.

The benchmark methods compared include UCB (Auer et al., 2002), UCB-Tuned (UCBT) (Auer et al., 2002), KLUCB (Cappé et al., 2013), MOSS (Audibert & Bubeck, 2009), Bayes-UCB (BUCB) (Kaufmann et al., 2012a), GPUCB (Srinivas et al., 2012), GPUCB-Tuned (GPUCBT) (Srinivas et al., 2012), TS (Agrawal & Goyal, 2012), Knowledge Gradient (KG) (Ryzhov et al., 2012), KG* (Ryzhov et al., 2010), KGMin (Kamiński, 2015) KGMN (Kamiński, 2015), IDS (Russo & Van Roy, 2018b), and VIDS, (Russo & Van Roy, 2018b). A detailed review of these methods is presented in Section 6. The values of their hyper-parameters are as follows. In searching for a solution of KLUCB and $\hat{C}_\alpha(t)$ in RBMLE, the maximum number of iterations is set to be 100. Following the suggestion in the original papers, we take $c = 0$ in KLUCB and BUCB. We take $\delta = 10^{-5}$ in GPUCB. We tune the parameter c in GPUCBT



(a) (b) (c)

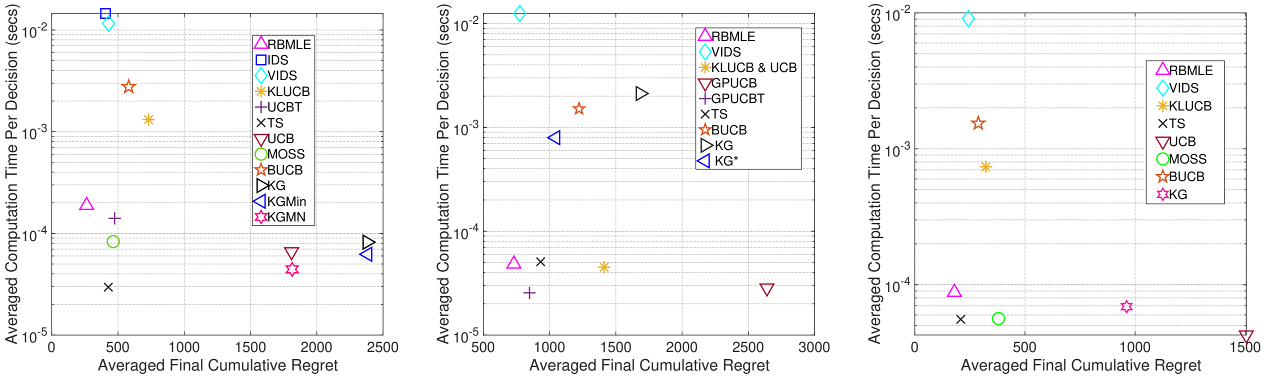
Figure 1: Averaged cumulative regret: (a) Bernoulli bandits with $(\theta_i)_{i=1}^{10} = (0.66, 0.67, 0.68, 0.69, 0.7, 0.61, 0.62, 0.63, 0.64, 0.65)$ & $\Delta = 0.01$; (b) Gaussian bandits with $(\theta_i)_{i=1}^{10} = (0.41, 0.52, 0.66, 0.43, 0.58, 0.65, 0.48, 0.67, 0.59, 0.63)$ & $\Delta = 0.01$; (c) Exponential bandits with $(\theta_i)_{i=1}^{10} = (0.31, 0.1, 0.2, 0.32, 0.33, 0.29, 0.2, 0.3, 0.15, 0.08)$ & $\Delta = 0.01$.

Algorithm 3 Adaptive Scheme with Estimation $\alpha(t)$ in Exponential Bandits

```
Input: $N, \varepsilon \in (0, \frac{1}{2})$, and $\beta(t)$
for $t = 1, 2, \cdots$ do
    for $i = 1$ to $N$ do
        $U_i(t) = p_i(t) + \frac{\kappa(N+2)\log t+\sqrt{\kappa^2(N+2)^2(\log t)^2+2\rho^2(N+2)\log t}}{N_i(t)}$
        // upper confidence bound
        $L_i(t) = \max\left(p_i(t) - \frac{\kappa(N+2)\log t+\sqrt{\kappa^2(N+2)^2(\log t)^2+2\rho^2(N+2)\log t}}{N_i(t)}, 0\right)$
        // lower confidence bound
    end for
    $U_{\max}(t) = \max_{i=1,\cdots,N} U_i(t)$
    $\hat{\Delta}_t = \max_i \left\{ \max\left(0, L_i(t) - \max_{j\neq i} U_j(t)\right) \right\}$
    if $\xi\left(\frac{16(\kappa\varepsilon\hat{\Delta}_t+2\rho^2)}{(\varepsilon\hat{\Delta}_t)^2\beta(t)}, 0\right) < \dot{F}^{-1}(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2})$ then
        $\alpha(t) = \beta(t)\log t$ // In this case, we know $\hat{C}_\alpha(t) > \beta(t)$
    else
        Find $\hat{C}_\alpha(t) = \frac{16(\kappa\varepsilon\hat{\Delta}_t+2\rho^2)}{(\varepsilon\hat{\Delta}_t)^2 K^*(U_{\max}(t)-\frac{\varepsilon\hat{\Delta}_t}{2},0)}$ by solving the minimization problem of (19) for $K^*(U_{\max}(t) - \frac{\varepsilon\hat{\Delta}_t}{2}, 0)$.
        $\alpha(t) = \min\{\hat{C}_\alpha(t), \beta(t)\}\log t$
    end if
end for
```

for each experiment and choose $c = 0.9$ that achieves the best performance. In the comparison with IDS and VIDS, we uniformly sampled 100 points over the interval $[0, 1]$ for Bernoulli and Exponential Bandits and sampled 1000 points for Gaussian bandits (the q in Algorithm 4 in (Russo & Van Roy, 2018b)) and take $M = 10^4$ in sampling (Algorithm 3 in (Russo & Van Roy, 2018b)). The conjugate priors for Bayesian-family methods are Beta distribution $\mathcal{B}(1, 1)$ for Bernoulli bandits, $\mathcal{N}(0, 1)$ for Gaussian bandits with $\sigma = 1$, and Gamma distribution $\Gamma(1, 1)$ for Exponential bandits with $\rho = 10$ and $\kappa = 10$. The average is taken over 100 trials. The time horizon in the experiments for effectiveness and efficiency is 10^5, and for scalability is 10^4.

Effectiveness. Figures 1, 3–4 and Tables 2-10 (some are in Appendix M.1) illustrate the effectiveness of RBMLE with respect to the cumulative regret as well as quantiles. Note that in the (b) sub-figures of these figures, since KLUCB shares the same closed-form index as UCB in Gaussian bandits, their curves coincide. We observe that for all three types of bandits, Bernoulli, Gaussian and Exponential, RBMLE achieves competitive performance, often slightly better than the best performing benchmark method. IDS or VIDS are often the closest competitors to RBMLE. However, the computational complexity of RBMLE is much lower compared to IDS and VIDS, which need to compute high dimensional integrals or estimate them through sampling. One other advantage of RBMLE over some benchmark methods is that it is “time horizon agnostic”, i.e., the computation of RBMLE index does not need the knowledge of time horizon T. In contrast, BUCB, MOSS, GPUCBT, and KG-family algorithms (KG, KG*) need to know T. It is worth mentioning that in Bernoulli bandits, KG, KGMin, and KGMN perform poorly as they explore insufficiently. This is not surprising as several papers have pointed out the limitations of KG-family methods when observations are discrete (Kamiński, 2015).



(a)

(b)

(c)

Figure 2: Averaged computation time per decision vs. averaged final cumulative regret: (a) Figure 1(a); (b) Figure 1(b); (c) Figure 1(c).

Efficiency Figures 2 and 5–6 (some are in Appendix M.2) present the efficiency of RBMLE in terms of averaged computation time per decision vs. averaged final cumulative regret. The former averaged the total time spent in all trials over the number of trials and number of rounds. The latter averaged the total cumulative regret in all trials over the number of trials. RBMLE is seen to provide competitive performance compared to other benchmark methods. It achieves slightly better regret, and does so with orders of magnitude less computation time, than IDS, VIDS and KLUCB. This is largely because IDS and VIDS need to estimate several integrals in each round, and KLUCB often relies on Newton's method or bisection search to find the index for an arm, except in Gaussian bandtis, where KLUCB has a simple closed-form solution. It is also observed that the computation time of RBMLE is larger than some benchmark methods such as UCB, GPUCB, and KG, which enjoy a simple closed-form index. However, their regret performance is far worse than RBMLE's. In the comparison of efficiency, the closest competitors to RBMLE are TS, MOSS, UCBT, and GPUCBT. Compared to them, RBMLE still enjoys salient advantages in different aspects. Compared to TS, RBMLE follows the frequentist formulation and thus its performance does not deteriorate when a bad prior is chosen. Compared to MOSS, RBMLE does not rely on the knowledge of T to compute its index. Compared to UCBT as well as GPUCBT, RBMLE has a order-optimal regret bound, as proved in the earlier sections.

Scalability Tables 11-13 show the scalability of RBMLE as the number of arms is increased. This is illustrated through comparing different methods' averaged computation time per decision averaged computation time per decision under varying numbers of arms. We observe that RBMLE scales well for various reward distributions as the number of arms increases, often demonstrating performance comparable to the most scalable ones among the benchmark methods. The averaged computation time per decision stays at a few 10^{-4} seconds even when the number of arms reaches 70. In contrast, it can be as high as thousands of 10^{-4} seconds for IDS and VIDS, and it is often tens of times higher for KLUCB.

6. Related Work

The RBMLE approach proposed in (Kumar & Becker, 1982) has been examined in a variety of settings, including MDPs and Linear-Quadratic-Gaussian systems, in (Kumar & Lin, 1982; Kumar, 1983b;a; Borkar, 1990; 1991; Stettner, 1993; Duncan et al., 1994; Campi & Kumar, 1998; Prandini & Campi, 2000). The simple index for the case of Bernoulli bandits was derived in (Becker & Kumar, 1981).

However, prior studies have been limited to focusing on long-term average reward optimality, which corresponds to a loose $o(T)$ bound on regret. This paper aims to tailor RBMLE to more general SMAB problems, and to prove its finite-time regret performance.

Learning algorithms for SMAB problems have been extensively studied. Most prior studies can be categorized into frequentist approaches or Bayesian approaches. In the frequentist settings, the family of UCB algorithms, including UCB (Auer et al., 2002), UCBT (Auer et al., 2002), and MOSS (Audibert & Bubeck, 2009; Degenne & Perchet, 2016), are among the most popular ones given their simplicity in implementation and the established regret bounds. An upper confidence bound can be directly derived from concentration inequalities or constructed with the help of other information measures, such as the Kullback–Leibler divergence used by KLUCB (Filippi et al., 2010; Garivier & Cappé, 2011; Cappé et al., 2013). The concept of upper confidence bound has later been extended to various types of models, such as contextual linear bandits (Chu et al., 2011; Abbasi-Yadkori et al., 2011; Rusmevichientong & Tsitsiklis, 2010), Gaussian process bandit optimization (GPUCB and GPUCBT) (Srinivas et al., 2012), and model-based reinforcement learning (Jaksch et al., 2010). The above list is by no means exhaustive but is mainly meant to illustrate the wide applicability of the UCB approach in different settings. While being a simple and generic index-type algorithm, UCB-based methods sometimes suffer from much higher regret than their counterparts (Russo & Van Roy, 2014; Chapelle & Li, 2011). Different from the UCB solutions, the proposed RBMLE algorithm addresses the exploration and exploitation tradeoff by directly operating with the likelihood function to navigate the exploration, and therefore it makes better use of the information of the parametric distributions.

On the other hand, the Bayesian approach studies the setting where the unknown reward parameters are drawn from an underlying prior distribution. As one of the most popular Bayesian bandit algorithms, TS (Scott, 2010; Chapelle & Li, 2011; Agrawal & Goyal, 2012; Korda et al., 2013; Kaufmann et al., 2012b) follows the principle of probability matching by continuously updating the posterior distribution based on a prior. In addition to strong theoretical guarantees (Agrawal & Goyal, 2012; Kaufmann et al., 2012b), TS has been reported to achieve superior empirical performance to its counterparts (Chapelle & Li, 2011; Scott, 2010). While being a powerful bandit algorithm, TS can be sensitive to the choice of the prior (Korda et al., 2013; Liu & Li, 2016). Another popular Bayesian algorithm is BUCB (Kaufmann et al., 2012a), which combines the Bayesian interpretation of bandit problems and the simple closed-form expression of UCB-type algorithms. In contrast, RBMLE does not rely on a prior and hence completely obviates the potential issues arising from an inappropriate prior choice. Another line of investigation takes advantage of the Bayesian update in information-related measures. KG (Ryzhov et al., 2012) and its variant KG* (Ryzhov et al., 2010), KGMin, and KGMN (Kamiński, 2015; Ryzhov et al., 2012; 2010) proceed by making a greedy one-step look-ahead measurement for exploration, as suggested by their names. While KG has been shown to empirically perform well for Gaussian distributions (Ryzhov et al., 2010; Wang et al., 2016), its performance is not readily quantifiable, and it does not always converge to optimality (Russo & Van Roy, 2014). Another competitive solution is IDS and its variant,VIDS, proposed by Russo & Van Roy (2018b). Different from the KG algorithm, IDS blends in the concept of information gain by looking at the ratio between the square of expected immediate regret and the expected reduction in the entropy of the target. Moreover, it has been reported in (Russo & Van Roy, 2014; 2018a) that IDS achieves state-of-the-art results in various bandit models. However, IDS and VIDS suffer from high computational complexity and poor scalability due to the excessive sampling required for estimating high dimensional integrals. Compared to these regret-competitive solutions, the proposed RBMLE algorithms can achieve comparable performance both theoretically and empirically, but at the same time it retains computational efficiency.

7. Concluding Remarks

The RBMLE method, developed in the general study of adaptive control four decades ago, provides a scheme for optimal control of general Markovian systems. It exploits the observation that the MLE has a one-sided bias favoring parameters with smaller optimal rewards, and so delicately steers the scheme to an optimal estimate by biasing the MLE in the reverse way. Just like the later Upper Confidence Bound policy, it also can be regarded as exemplifying the philosophy of being "optimistic in the face of uncertainty," but does it in a very different way. The resulting indices (Table 1) are very different. Over the four decades since its introduction, the RBMLE method has not been further analyzed or empirically evaluated for its regret performance.

This paper takes an initial step in this direction. It shows how indices can be derived naturally for the Exponential Family of reward distributions, and how these indices can even be applied to other non-parametric distributions. It studies the interplay between the choice of the growth rate of the reward-bias and the resulting regret. It exposes the important role played by the knowledge of the minimum gap in the choice of the reward-bias growth rate. When this minimum gap is not known, it shows how it can be adaptively estimated. It empirically shows that RBMLE attains excellent regret performance compared with other state-of-art methods, while requiring low computation time per decision compared to other methods with comparable regret performance, and scales well as the number of arms is increased.

Being a general purpose approach for optimal decision making under certainty, RBMLE holds potential for a number of such problems, including contextual bandits, adversarial bandits, Bayesian optimization and more beyond.

Acknowledgments

We sincerely thank all anonymous reviewers and meta reviewers for their insightful suggestions during rebuttal! This work is partially supported by the Ministry of Science and Technology of Taiwan under Contract No. MOST 108-2636-E-009-014. This material is also based upon work partially supported by the U.S. Army Research Office under Contract No. W911NF-18-10331, U.S. ONR under Contract No. N00014-18-1-2048, and NSF Science & Technology Center Grant CCF-0939370.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Agrawal, S. and Goyal, N. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory (COLT)*, pp. 39–1, 2012.

Audibert, J.-Y. and Bubeck, S. Minimax policies for adversarial and stochastic bandits. In *COLT*, pp. 217–226, 2009.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3): 235–256, 2002.

Becker, A. and Kumar, P. R. Optimal strategies for the N-armed bandit problem. Technical report, Mathematics Research Report No. 81-1, Department of Mathematics, University of Maryland Baltimore County, Jan 1981.

Borkar, V. and Varaiya, P. Adaptive control of Markov chains, I Finite parameter set. *IEEE Transactions on Automatic Control*, 24(6):953–957, 1979.

Borkar, V. S. The Kumar-Becker-Lin scheme revisited. *Journal of Optimization Theory and Applications*, 66(2):289–309, 1990.

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

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

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., Stoltz, G., et al. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

Chapelle, O. and Li, L. An empirical evaluation of Thompson sampling. In *Advances in neural information processing systems*, pp. 2249–2257, 2011.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 208–214, 2011.

Degenne, R. and Perchet, V. Anytime optimal algorithms in stochastic multi-armed bandits. In *International Conference on Machine Learning*, pp. 1587–1595, 2016.

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

Filippi, S., Cappé, O., and Garivier, A. Optimism in reinforcement learning and Kullback-Leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 115–122, 2010.

Garivier, A. and Cappé, O. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory (COLT)*, pp. 359–376, 2011.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Jordan, M. Chapter 8: The Exponential family: Basics, 2010. URL https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/chapter8.pdf.

Kamiński, B. Refined knowledge-gradient policy for learning probabilities. *Operations Research Letters*, 43(2):143–147, 2015.

Kaufmann, E., Cappé, O., and Garivier, A. On Bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics (AISTATS)*, pp. 592–600, 2012a.

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: an asymptotically optimal finite-time analysis. In *Proceedings of the 23rd international conference on Algorithmic Learning Theory*, pp. 199–213. Springer-Verlag, 2012b.

Korda, N., Kaufmann, E., and Munos, R. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pp. 1448–1456, 2013.

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

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

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

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

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

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

Liu, C.-Y. and Li, L. On the prior sensitivity of Thompson sampling. In *International Conference on Algorithmic Learning Theory (ALT)*, pp. 321–336, 2016.

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

Prandini, M. and Campi, M. C. Adaptive LQG control of input-output systems—A cost-biased approach. *SIAM Journal on Control and Optimization*, 39(5):1499–1519, 2000.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

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

Russo, D. and Van Roy, B. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018a.

Russo, D. and Van Roy, B. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018b.

Ryzhov, I. O., Frazier, P. I., and Powell, W. B. On the robustness of a one-period look-ahead policy in multi-armed bandit problems. *Procedia Computer Science*, 1(1):1635–1644, 2010.

Ryzhov, I. O., Powell, W. B., and Frazier, P. I. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195, 2012.

Scott, S. L. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

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

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, Y., Wang, C., and Powell, W. The knowledge gradient for sequential decision making with stochastic binary feedbacks. In *International Conference on Machine Learning (ICML)*, pp. 1138–1147, 2016.

Whittle, P. Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2): 143–149, 1980.