Appendix

A. Proof of the Index Strategy in (8)

Recall that $\widehat{\boldsymbol{\eta}}_t^{\text{RBMLE}} = (\widehat{\eta}_{t,1}^{\text{RBMLE}}, \cdots, \widehat{\eta}_{t,N}^{\text{RBMLE}})$ is the reward-biased MLE for $\boldsymbol{\eta}$ and from (6) that $\widehat{\boldsymbol{\eta}}_t^{\text{RBMLE}}$ is a maximizer of the following problem:

$$\max_{\boldsymbol{\eta}:\eta_j \in \mathcal{N}, \forall j} \big\{ L(\mathcal{H}_t; \boldsymbol{\eta}) \max_{i \in [N]} \exp(\eta_i \alpha(t)) \big\}. \tag{30}$$

Define an index set and a parameter set as

$$\mathcal{I}_t := \operatorname*{argmax}_{i \in [N]} \{ \widehat{\eta}_{t,i}^{\text{RBMLE}} \} \tag{31}$$

$$H_{t,i} := \operatorname*{argmax}_{\boldsymbol{\eta}:\eta_j \in \mathcal{N}, \forall j} \big\{ L(\mathcal{H}_t; \boldsymbol{\eta}) \exp(\eta_i \alpha(t)) \big\} \tag{32}$$

Note that at each time t, RBMLE would select an arm from the index set $\mathcal{I}_t$, as shown in (7). For each arm i, consider an estimator $\bar{\boldsymbol{\eta}}_t^{(i)} = (\bar{\eta}_{t,1}^{(i)}, \cdots, \bar{\eta}_{t,N}^{(i)}) \in H_{t,i}$. Accordingly, we further define an index set

$$\mathcal{I}_t' := \operatorname*{argmax}_{i \in [N]} \{ L(\mathcal{H}_t; \bar{\boldsymbol{\eta}}_t^{(i)}) \exp(\bar{\eta}_{t,i}^{(i)} \alpha(t)) \}. \tag{33}$$

Next, we show that the two index sets are identical, i.e. $\mathcal{I}_t = \mathcal{I}_t'$. Since $L(\mathcal{H}_t; \widehat{\boldsymbol{\eta}}_t^{\text{RBMLE}})$ does not depend on i, we know

$$\operatorname*{argmax}_{i \in [N]} \{ \widehat{\eta}_{t,i}^{\text{RBMLE}} \} = \operatorname*{argmax}_{i \in [N]} \big\{ L(\mathcal{H}_t; \widehat{\boldsymbol{\eta}}_t^{\text{RBMLE}}) \exp(\widehat{\eta}_{t,i}^{\text{RBMLE}} \alpha(t)) \big\}. \tag{34}$$

Moreover, we have

$$\begin{aligned}
\max_{i \in [N]} \Big\{ L(\mathcal{H}_t; \widehat{\boldsymbol{\eta}}_t^{\text{RBMLE}}) \exp(\widehat{\eta}_{t,i}^{\text{RBMLE}} \alpha(t)) \Big\} &= L(\mathcal{H}_t; \widehat{\boldsymbol{\eta}}_t^{\text{RBMLE}}) \cdot \max_{i \in [N]} \exp(\widehat{\eta}_{t,i}^{\text{RBMLE}} \alpha(t)) && (35) \\
&= \max_{\boldsymbol{\eta}:\eta_j \in \mathcal{N}, \forall j} \Big\{ L(\mathcal{H}_t; \boldsymbol{\eta}) \cdot \max_{i \in [N]} \exp(\eta_i \alpha(t)) \Big\} && (36) \\
&= \max_{\boldsymbol{\eta}:\eta_j \in \mathcal{N}, \forall j} \Big\{ \max_{i \in [N]} \big\{ L(\mathcal{H}_t; \boldsymbol{\eta}) \cdot \exp(\eta_i \alpha(t)) \big\} \Big\} && (37) \\
&= \max_{i \in [N]} \Big\{ \max_{\boldsymbol{\eta}:\eta_j \in \mathcal{N}, \forall j} \big\{ L(\mathcal{H}_t; \boldsymbol{\eta}) \cdot \exp(\eta_i \alpha(t)) \big\} \Big\} && (38) \\
&= \max_{i \in [N]} \big\{ L(\mathcal{H}_t; \bar{\boldsymbol{\eta}}_t^{(i)}) \exp(\bar{\eta}_{t,i}^{(i)} \alpha(t)) \big\}, && (39)
\end{aligned}$$

where (35) follows from the fact that $L(\mathcal{H}_t; \widehat{\boldsymbol{\eta}}_t^{\text{RBMLE}})$ does not depend on i, (36) holds by the definition of $\widehat{\boldsymbol{\eta}}_t^{\text{RBMLE}}$, (37)-(38) follow from that interchanging the order of the two $\max$ operations does not change the optimal value and the optimizers, and (39) follows from the definitions of $H_{t,i}$ and $\bar{\boldsymbol{\eta}}_t^{(i)}$. By (32), (33), and (35)-(39), we conclude that $\mathcal{I}_t = \mathcal{I}_t'$, and hence (8) indeed holds.

B. Proof of Proposition 1

Recall from (8) that

$$\pi_t^{\text{RBMLE}} = \operatorname*{argmax}_{i \in [N]} \Big\{ \max_{\boldsymbol{\eta}:\eta_j \in \mathcal{N}, \forall j} \big\{ L(\mathcal{H}_t; \boldsymbol{\eta}) \exp(\eta_i \alpha(t)) \big\} \Big\} \tag{40}$$

By plugging $L(\mathcal{H}_t; \boldsymbol{\eta})$ into (40) using the density function of the Exponential Families and taking the logarithm of (40),

$$\pi_t^{\text{BMLE}} = \operatorname*{argmax}_{i \in \{1, \cdots, N\}} \Big\{ \max_{\boldsymbol{\eta}:\eta_j \in \mathcal{N}, \forall j} \Big\{ \underbrace{\sum_{\tau=1}^{t} \big(\eta_{\pi_\tau} X_\tau - F(\eta_{\pi_\tau}) \big) + \eta_i \alpha(t)}_{=: \ell_i(\mathcal{H}_t; \boldsymbol{\eta})} \Big\} \Big\}. \tag{41}$$

Note that the inner maximization problem for $\ell_i(\mathcal{H}_t;\boldsymbol{\eta})$ over $\boldsymbol{\eta}$ is convex since $F(\cdot)$ is a convex function. Recall that $N_i(t)$ and $S_i(t)$ denote the total number of trials of arm i and the total reward collected from pulling arm i up to time t, as defined in Section 2. By taking the partial derivatives of $\ell_i(\mathcal{H}_t;\boldsymbol{\eta})$ with respect to each η_i, we know that $\ell_i(\mathcal{H}_t;\boldsymbol{\eta})$ is maximized when $\dot{F}(\eta_i)=\big[\frac{S_i(t)+\alpha(t)}{N_i(t)}\big]_\Theta$ and $\dot{F}(\eta_j)=\frac{S_j(t)}{N_j(t)}$, for $j\neq i$, where $[\cdot]_\Theta$ denotes the clipped value within the set Θ. For each $i=1,\cdots,N$, we then define

$$\eta_i^* := \dot{F}^{-1}\big(\frac{S_i(t)}{N_i(t)}\big), \tag{42}$$

$$\eta_i^{**} := \dot{F}^{-1}\Big(\Big[\frac{S_i(t)+\alpha(t)}{N_i(t)}\Big]_\Theta\Big). \tag{43}$$

By substituting $\{\eta_i^*\}$ and $\{\eta_i^{**}\}$ into (41), we have

$$\pi_t^{\text{BMLE}} = \underset{i\in\{1,\cdots,N\}}{\text{argmax}}\ \Big\{\ell_i\big(\mathcal{H}_t;\eta_i^{**},\{\eta_j^{**}\}_{j\neq i}\big)\Big\} \tag{44}$$

$$= \underset{i\in\{1,\cdots,N\}}{\text{argmax}}\ \Big\{\ell_i\big(\mathcal{H}_t;\eta_i^{**},\{\eta_j^{*}\}_{j\neq i}\big)-\ell_i\big(\mathcal{H}_t;\{\eta_j^{*}\}_{j=1,\cdots,N}\big)\Big\} \tag{45}$$

$$= \underset{i\in\{1,\cdots,N\}}{\text{argmax}}\ \Big\{\Big[\big((S_i(t)+\alpha(t))\big)\eta_i^{**}-N_i(t)F(\eta_i^{**})\Big]-\Big[S_i(t)\eta_i^{*}-N_i(t)F(\eta_i^{*})\Big]\Big\}. \tag{46}$$

By substituting $N_i(t)p_i(t)$ for $S_i(t)$ in (46), we then arrive at the index as

$$I(p_i(t),N_i(t),\alpha(t)) = \Big[\big((N_i(t)p_i(t)+\alpha(t))\big)\eta_i^{**}-N_i(t)F(\eta_i^{**})\Big]-\Big[N_i(t)p_i(t)\eta_i^{*}-N_i(t)F(\eta_i^{*})\Big]. \tag{47}$$

□

C. Proof of Corollary 1

Recall from (47) that for the Exponential Family rewards, the BMLE index is

$$I(p_i(t),N_i(t),\alpha(t)) = \Big[\big((N_i(t)p_i(t)+\alpha(t))\big)\eta_i^{**}-N_i(t)F(\eta_i^{**})\Big]-\Big[N_i(t)p_i(t)\eta_i^{*}-N_i(t)F(\eta_i^{*})\Big]. \tag{48}$$

For the Bernoulli case, we know $F(\eta)=\log(1+e^\eta)$, $\dot{F}(\eta)=\frac{e^\eta}{1+e^\eta}$, $\dot{F}^{-1}(\theta)=\log(\frac{\theta}{1-\theta})$, and $F(\dot{F}^{-1}(\theta))=\log(\frac{1}{1-\theta})$. Since $\Theta=[0,1]$ for Bernoulli rewards, we need to analyze the following two cases when substituting the above $\dot{F}^{-1}(\theta)$ and $F(\dot{F}^{-1}(\theta))$ into (48):

- **Case 1:** $\alpha(t)<N_i(t)(1-p_i(t))$ (or equivalently $\tilde{p}_i(t)<1$)

  We have

$$I(p_i(t),N_i(t),\alpha(t)) \tag{49}$$

$$=\big(N_i(t)p_i(t)+\alpha(t)\big)\log\Big(\frac{N_i(t)p_i(t)+\alpha(t)}{N_i(t)-(N_i(t)p_i(t)+\alpha(t))}\Big)-N_i(t)\log\Big(\frac{N_i(t)}{N_i(t)-(N_i(t)p_i(t)+\alpha(t))}\Big) \tag{50}$$

$$-N_i(t)p_i(t)\log\Big(\frac{N_i(t)p_i(t)}{N_i(t)-N_i(t)p_i(t)}\Big)+N_i(t)\log\Big(\frac{N_i(t)}{N_i(t)-N_i(t)p_i(t)}\Big) \tag{51}$$

$$=N_i(t)\Big\{\big(p_i(t)+\frac{\alpha(t)}{N_i(t)}\big)\log\Big(p_i(t)+\frac{\alpha(t)}{N_i(t)}\Big)+\big(1-(p_i(t)+\frac{\alpha(t)}{N_i(t)})\big)\log\Big(1-(p_i(t)+\frac{\alpha(t)}{N_i(t)})\Big) \tag{52}$$

$$-p_i(t)\log(p_i(t))-(1-p_i(t))\log(1-p_i(t))\Big\}, \tag{53}$$

  where (52)-(53) are obtained by reorganizing the terms in (50)-(51).

- **Case 2:** $\alpha(t)\geq N_i(t)(1-p_i(t))$ (or equivalently $\tilde{p}_i(t)=1$)

  In this case, the index would be the same as the case where $p_i(t)+\alpha(t)/N_i(t)=1$. Therefore, we simply have

$$I(p_i(t),N_i(t),\alpha(t)) = N_i(t)\Big\{-p_i(t)\log(p_i(t))-(1-p_i(t))\log(1-p_i(t))\Big\}. \tag{54}$$

□

D. Proof of Corollary 2

Recall from (47) that for the Exponential Family rewards, the BMLE index is

$$I(p_i(t), N_i(t), \alpha(t)) = \Big[\big((N_i(t)p_i(t) + \alpha(t))\big)\eta_i^{**} - N_i(t)F(\eta_i^{**})\Big] - \Big[N_i(t)p_i(t)\eta_i^{*} - N_i(t)F(\eta_i^{*})\Big], \tag{55}$$

where $\eta_i^* = \dot{F}^{-1}\big(\frac{S_i(t)}{N_i(t)}\big)$ and $\eta_i^{**} = \dot{F}^{-1}\big(\frac{S_i(t)+\alpha(t)}{N_i(t)}\big)$. For Gaussian rewards with the same variance σ^2 for all arms, we have $F(\eta_i) = \sigma^2\eta_i^2/2$, $\dot{F}(\eta_i) = \sigma^2\eta_i$, $\dot{F}^{-1}(\theta_i) = \theta_i/\sigma^2$, and $F(\dot{F}^{-1}(\theta_i)) = \theta_i^2/2\sigma^2$, for each arm i. Therefore, the BMLE index becomes

$$I(p_i(t), N_i(t), \alpha(t)) \tag{56}$$

$$= \frac{S_i(t)+\alpha(t)}{\sigma^2 N_i(t)}(S_i(t)+\alpha(t)) - N_i(t)\frac{\sigma^2}{2}\Big(\frac{S_i(t)+\alpha(t)}{\sigma^2 N_i(t)}\Big)^2 - S_i(t)\frac{S_i(t)}{\sigma^2 N_i(t)} + N_i(t)\frac{\sigma^2}{2}\Big(\frac{S_i(t)}{\sigma^2 N_i(t)}\Big)^2 \tag{57}$$

$$= \frac{2S_i(t)\alpha(t) + \alpha(t)^2}{2\sigma^2 N_i(t)}. \tag{58}$$

Equivalently, for the Gaussian rewards, the selected arm at each time t is

$$\pi_t^{\text{BMLE}} = \underset{i\in\{1,\cdots,N\}}{\text{argmax}} \Big\{p_i(t) + \frac{\alpha(t)}{2N_i(t)}\Big\}. \tag{59}$$

□

E. Proof of Corollary 3

Recall from (47) that for the Exponential Family distributions, the BMLE index is

$$I(p_i(t), N_i(t), \alpha(t)) = \Big[\big((N_i(t)p_i(t) + \alpha(t))\big)\eta_i^{**} - N_i(t)F(\eta_i^{**})\Big] - \Big[N_i(t)p_i(t)\eta_i^{*} - N_i(t)F(\eta_i^{*})\Big], \tag{60}$$

where $\eta_i^* = \dot{F}^{-1}\big(\frac{S_i(t)}{N_i(t)}\big)$ and $\eta_i^{**} = \dot{F}^{-1}\big(\frac{S_i(t)+\alpha(t)}{N_i(t)}\big)$. For the exponential distribution, we have $F(\eta_i) = \log(\frac{-1}{\eta_i})$, $\dot{F}(\eta_i) = \frac{-1}{\eta_i}$, $\dot{F}^{-1}(\theta_i) = \frac{-1}{\theta_i}$, and $F(\dot{F}^{-1}(\theta_i)) = \log\theta_i$, for each arm i. Therefore, the BMLE index becomes

$$I(p_i(t), N_i(t), \alpha(t)) \tag{61}$$

$$= (N_i(t)p_i(t) + \alpha(t)) \cdot \Big(-\frac{N_i(t)}{N_i(t)p_i(t)+\alpha(t)}\Big) - N_i(t)\log\Big(\frac{N_i(t)p_i(t)+\alpha(t)}{N_i(t)}\Big) \tag{62}$$

$$- \Big(N_i(t)p_i(t)\big(-\frac{1}{p_i(t)}\big)\Big) + N_i(t)\log p_i(t) \tag{63}$$

$$= N_i(t)\log\big(\frac{N_i(t)p_i(t)}{N_i(t)p_i(t)+\alpha(t)}\big). \tag{64}$$

□

F. Proof of Lemma 1

(i) Recall that

$$I(\nu, n, \alpha(t)) = \big(n\nu + \alpha(t)\big)\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big) - n\nu\dot{F}^{-1}(\nu) - nF\Big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)\Big) + nF\big(\dot{F}^{-1}(\nu)\big).$$

By taking the partial derivative of $I(\nu, n, \alpha(t))$ with respect to n, we have

$$\frac{\partial I}{\partial n} = \nu\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big) + \big(n\nu + \alpha(t)\big)\frac{\partial \dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)}{\partial n} - \nu\dot{F}^{-1}(\nu) \tag{65}$$

$$- F\Big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)\Big) - n\dot{F}\Big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)\Big)\frac{\partial \dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)}{\partial n} + F\big(\dot{F}^{-1}(\nu)\big) \tag{66}$$

$$= \nu \cdot \Big[\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big) - \dot{F}^{-1}(\nu)\Big] - \Big[F\Big(\dot{F}^{-1}\big(\nu + \frac{\alpha(t)}{n}\big)\Big) - F\big(\dot{F}^{-1}(\nu)\big)\Big]. \tag{67}$$

Since $\dot{F}(\cdot)$ is strictly increasing for the Exponential Families, we know $\dot{F}^{-1}(\cdot)$ is also strictly increasing and $\dot{F}^{-1}(\nu + \alpha(t)/n) > \dot{F}^{-1}(\nu)$. Moreover, by the strict convexity of $F(\cdot)$, we have

$$F\Big(\dot{F}^{-1}\big(\nu+\frac{\alpha(t)}{n}\big)\Big) - F\big(\dot{F}^{-1}(\nu)\big) > \Big(\dot{F}^{-1}\big(\nu+\frac{\alpha(t)}{n}\big) - \dot{F}^{-1}(\nu)\Big)\cdot \underbrace{\dot{F}\big(\dot{F}^{-1}(\nu)\big)}_{=\nu}. \tag{68}$$

Therefore, by (65)-(68), we conclude that $\frac{\partial I}{\partial n} < 0$ and hence $I(\nu, n, \alpha(t))$ is strictly decreasing with n.

(ii) Recall that

$$I(\nu,n,\alpha(t)) = \big(n\nu+\alpha(t)\big)\dot{F}^{-1}\big(\nu+\frac{\alpha(t)}{n}\big) - n\nu\dot{F}^{-1}(\nu) - nF\Big(\dot{F}^{-1}\big(\nu+\frac{\alpha(t)}{n}\big)\Big) + nF\big(\dot{F}^{-1}(\nu)\big).$$

By taking the partial derivative of $I(\nu, n, \alpha(t))$ with respect to ν, we have

$$\frac{\partial I}{\partial \nu} = n\dot{F}^{-1}\big(\nu+\frac{\alpha(t)}{n}\big) + (n\nu+\alpha(t))\frac{\partial \dot{F}^{-1}\big(\nu+\frac{\alpha(t)}{n}\big)}{\partial \nu} - \Big(n\dot{F}^{-1}(\nu) + n\nu\frac{\partial \dot{F}^{-1}(\nu)}{\partial \nu}\Big) \tag{69}$$

$$- n\underbrace{\dot{F}\Big(\dot{F}^{-1}\big(\nu+\frac{\alpha(t)}{n}\big)\Big)}_{\le \nu+\alpha(t)/n}\frac{\partial \dot{F}^{-1}\big(\nu+\frac{\alpha(t)}{n}\big)}{\partial \nu} + n\underbrace{\dot{F}\Big(\dot{F}^{-1}(\nu)\Big)}_{=\nu}\frac{\partial \dot{F}^{-1}(\nu)}{\partial \nu} \tag{70}$$

$$\ge n\cdot\Big[\dot{F}^{-1}\big(\nu+\frac{\alpha(t)}{n}\big) - \dot{F}^{-1}(\nu)\Big] \tag{71}$$

$$> 0, \tag{72}$$

where the last inequality follows from the fact that $\dot{F}^{-1}(\cdot)$ is strictly increasing for the Exponential Families. Therefore, we can conclude that $I(\nu, n, \alpha(t))$ is strictly increasing with ν, for all $\alpha(t) > 0$ and for all $n > 0$.

G. Proof of Lemma 2

Recall that we define

$$\xi(k;\nu) = k\Big[\big(\nu+\frac{1}{k}\big)\dot{F}^{-1}(\nu+\frac{1}{k}) - \nu\dot{F}^{-1}(\nu)\Big] - k\Big[F\Big(\dot{F}^{-1}\big(\nu+\frac{1}{k}\big)\Big) - F(\dot{F}^{-1}(\nu))\Big], \tag{73}$$

$$K^*(\theta',\theta'') = \inf\{k : \dot{F}^{-1}(\theta') > \xi(k;\theta'')\}. \tag{74}$$

Moreover, we have $I(\mu_1, k\alpha(t), \alpha(t)) = \alpha(t)\xi(k;\mu_1)$. By Lemma 1.(i), we know that $I(\mu_1, k\alpha(t), \alpha(t))$ decreases with k, for all $k > 0$. Let $z = \frac{1}{k}$. Under any fixed $\mu_1 \in \Theta$ and $\alpha(t) > 0$, we also know that

$$\lim_{k\to\infty}\xi(k;\mu_1) = \lim_{z\downarrow 0}\frac{\big[(\mu_1+z)\dot{F}^{-1}(\mu_1+z) - \mu_1\dot{F}^{-1}(\mu_1)\big] - \big[F\big(\dot{F}^{-1}(\mu_1+z)\big) - F(\dot{F}^{-1}(\mu_1))\big]}{z} \tag{75}$$

$$= \lim_{z\downarrow 0}\dot{F}^{-1}(\mu_1+z) + (\mu_1+z)\frac{\partial \dot{F}^{-1}(\mu_1+z)}{\partial z} - \dot{F}\big(\dot{F}^{-1}(\mu_1+z)\big)\frac{\partial \dot{F}^{-1}(\mu_1+z)}{\partial z} \tag{76}$$

$$= \dot{F}^{-1}(\mu_1), \tag{77}$$

where (75) is obtained by replacing $1/k$ with z, and (76) follows from L'Hôpital's rule. Therefore, we have

$$\lim_{k\to\infty} I(\mu_1, k\alpha(t), \alpha(t)) = \alpha(t)\cdot\dot{F}^{-1}(\mu_1). \tag{78}$$

By Lemma 1.(i) and (78), we know

$$I(\mu_1, k\alpha(t), \alpha(t)) \ge \alpha(t)\dot{F}^{-1}(\mu_1), \quad \text{for all } k > 0. \tag{79}$$

For any $n_2 > K^*(\mu_1, \mu_2)\alpha(t)$, we have

$$I(\mu_1, n_1, \alpha(t)) \ge \alpha(t)\dot{F}^{-1}(\mu_1) \tag{80}$$

$$\ge I(\mu_2, K^*(\mu_1,\mu_2)\alpha(t), \alpha(t)) \tag{81}$$

$$> I(\mu_2, n_2, \alpha(t)), \tag{82}$$

where (80) follows from (79), (81) holds from the definition of $K^*(\cdot,\cdot)$, and (82) holds due to Lemma 1.(i). Finally, we show that $K^*(\mu_1,\mu_2)$ is finite given that $\mu_1 > \mu_2$. We consider the limit of $\xi(k;\mu_2)$ when k approaches zero and again let $z = \frac{1}{k}$:

$$\lim_{k\downarrow 0}\xi(k;\mu_2) = \lim_{z\to\infty}\frac{\big[(\mu_2+z)\dot{F}^{-1}(\mu_2+z) - \nu\dot{F}^{-1}(\mu_2)\big] - \big[F\big(\dot{F}^{-1}\big(\mu_2+z\big)\big) - F(\dot{F}^{-1}(\mu_2))\big]}{z} \tag{83}$$

$$= \lim_{z\to\infty}\dot{F}^{-1}(\mu_2+z) + (\mu_2+z)\underbrace{\frac{\partial\dot{F}^{-1}(\mu_2+z)}{\partial z}}_{\geq 0} - \underbrace{\dot{F}\big(\dot{F}^{-1}(\mu_2+z)\big)}_{\leq \mu_2+z}\underbrace{\frac{\partial\dot{F}^{-1}(\mu_2+z)}{\partial z}}_{\geq 0} \tag{84}$$

$$\geq \lim_{z\to\infty}\dot{F}^{-1}(\mu_2+z) \tag{85}$$

$$\geq \dot{F}^{-1}(\mu_1), \tag{86}$$

where (84) follows from L'Hôpital's rule and (86) holds due to the fact that $\dot{F}^{-1}$ is increasing. By (83)-(86) and since $\xi(k;\mu_2)$ is continuous and strictly decreasing with k, we know there must exist a finite $k' \geq 0$ such that $\dot{F}^{-1}(\mu_1) = \xi(k';\mu_2)$. This implies that $K^*(\mu_1,\mu_2)$ is finite given that $\mu_1 > \mu_2$. □

H. Proof of Lemma 3

Similar to the proof of Lemma 2, we leverage the function $K^*(\cdot,\cdot)$ as defined in (74). By (74), we know that for any $k > K^*(\mu_0,\mu_2)$, we have $\xi(k;\mu_2) < \dot{F}^{-1}(\mu_0)$. Therefore, if $n_2 > K^*(\mu_0,\mu_2)\alpha(t)$,

$$I(\mu_2, n_2, \alpha(t)) < I(\mu_2, K^*(\mu_0,\mu_2), \alpha(t)) \tag{87}$$

$$= \alpha(t)\xi(K^*(\mu_0,\mu_2);\mu_2) \tag{88}$$

$$= \alpha(t)\dot{F}^{-1}(\mu_0). \tag{89}$$

Similarly, for any $k \leq K^*(\mu_0,\mu_1)$, we have $\xi(k;\mu_1) \geq \dot{F}^{-1}(\mu_0)$. Then, if $n_1 \leq K^*(\mu_0,\mu_1)\alpha(t)$, we know

$$I(\mu_1, n_1, \alpha(t)) \geq I(\mu_1, K^*(\mu_0,\mu_1), \alpha(t)) \tag{90}$$

$$= \alpha(t)\xi(K^*(\mu_0,\mu_1);\mu_1) \tag{91}$$

$$= \alpha(t)\dot{F}^{-1}(\mu_0). \tag{92}$$

Hence, by (87)-(92), we conclude that $I(\mu_1,n_1,\alpha(t)) > I(\mu_2,n_2,\alpha(t))$, for all $n_1 \leq K^*(\mu_0,\mu_1)\alpha(t)$ and $n_2 > K^*(\mu_0,\mu_2)\alpha(t)$. □

I. Proof of Proposition 2

Proof Sketch: Our target is to quantify the expected number of trials of each sub-optimal arm a up to time T. The regret bound proof starts with a similar demonstration as for UCB1 (Auer et al., 2002) by studying the probability of the event $\{I(p_1(t),N_1(t),\alpha(t)) \leq I(p_a(t),N_a(t),\alpha(t))\}$, using the Chernoff bound for Exponential Families. However, it is significantly different from the original proof as the dependency between the level of exploration and the bias term $\alpha(t)$ is technically more complex, compared to the straightforward confidence interval used by the conventional UCB-type policies. Specifically, the main challenge lies in characterizing the behavior of the RBMLE index for both regimes where $N_1(t)$ is small compared to $\alpha(t)$, as well as when it is large compared to $\alpha(t)$. Such a challenge is handled by considering three cases separately: (i) Consider $N_1(t) > \frac{4}{D(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_1)}\log t$ and apply Lemma 2; (ii) Consider $N_1(t) \leq \frac{4}{D(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_1)}\log t$ and $N_1(t) \leq K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\underline{\theta})\alpha(t)$ and apply Lemma 3; (iii) Use Lemma 3 to show that $\{N_1(t) \leq \frac{4}{D(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_1)}\log t\}$ and $\{N_1(t) > K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\underline{\theta})\alpha(t)\}$ cannot occur simultaneously.

To begin with, for each arm i, we define $p_{i,n}$ to be the empirical average reward collected in the first n pulls of arm i. For any Exponential Family reward distribution, the empirical mean of each arm i satisfies the following concentration inequalities (Korda et al., 2013): For any $\delta > 0$,

$$\mathbb{P}(p_{i,n} - \theta_i \geq \delta) \leq \exp(-nD(\theta_i+\delta,\theta_i)), \tag{93}$$

$$\mathbb{P}(\theta_i - p_{i,n} \geq \delta) \leq \exp(-nD(\theta_i-\delta,\theta_i)). \tag{94}$$

Next, for each arm i, we define the following confidence intervals for each pair of $n, t \in \mathbb{N}$:

$$\delta_i^+(n,t) := \inf\Big\{\delta : \exp(-nD(\theta_i+\delta,\theta_i)) \le \frac{1}{t^4}\Big\}, \tag{95}$$

$$\delta_i^-(n,t) := \inf\Big\{\delta : \exp(-nD(\theta_i-\delta,\theta_i)) \le \frac{1}{t^4}\Big\}. \tag{96}$$

Accordingly, for each arm i and for each pair of $n, t \in \mathbb{N}$, we define the following events:

$$G_i^+(n,t) = \Big\{p_{i,n} - \theta_i \le \delta_i^+(n,t)\Big\}, \tag{97}$$

$$G_i^-(n,t) = \Big\{\theta_i - p_{i,n} \le \delta_i^-(n,t)\Big\}. \tag{98}$$

By the concentration inequality considered in Section 2, we have

$$\mathbb{P}(G_i^+(n,t)^{\mathsf{c}}) \le e^{-nD(\theta_i+\delta_i^+(n,t),\theta_i)} \le \frac{1}{t^4}, \tag{99}$$

$$\mathbb{P}(G_i^-(n,t)^{\mathsf{c}}) \le e^{-nD(\theta_i-\delta_i^-(n,t),\theta_i)} \le \frac{1}{t^4}. \tag{100}$$

Consider the bias term $\alpha(t) = C_\alpha \log t$ with $C_\alpha \ge 4/(D(\theta_1 - \frac{\varepsilon}{2}\Delta, \theta_1) \cdot K^*(\theta_1 - \frac{\varepsilon}{2}\Delta, \underline{\theta}))$ and $\varepsilon \in (0,1)$. Recall that we assume arm 1 is the unique optimal arm. Our target is to quantify the total number of trials of each sub-optimal arm. Define

$$Q_a(T) := \max\Big\{\frac{4}{D(\theta_a + \frac{\varepsilon}{2}\Delta_a, \theta_a)}, C_\alpha K^*(\theta_1 - \frac{\varepsilon}{2}\Delta_a, \theta_a + \frac{\varepsilon}{2}\Delta_a)\Big\}\log T + 1. \tag{101}$$

We start by characterizing $\mathbb{E}[N_a(T)]$ for each $a = 2, \cdots, N$:

$$\mathbb{E}[N_a(T)] \tag{102}$$

$$\le Q_a(T) + \mathbb{E}\Bigg[\sum_{t=Q_a(T)+1}^{T} \mathbb{I}\big(I(p_a(t), N_a(t), \alpha(t)) \ge I(p_1(t), N_1(t), \alpha(t), N_a(t) \ge Q_a(T))\big)\Bigg] \tag{103}$$

$$= Q_a(T) + \sum_{t=Q_a(T)+1}^{T} \mathbb{P}\Big(I\big(p_a(t), N_a(t), \alpha(t)\big) \ge I\big(p_1(t), N_1(t), \alpha(t)\big), N_a(t) \ge Q_a(T)\Big) \tag{104}$$

$$\le Q_a(T) + \sum_{t=Q_a(T)+1}^{T} \mathbb{P}\Big(\max_{Q_a(T)\le n_a\le t} I\big(p_{a,n_a}, n_a, \alpha(t)\big) \ge \min_{1\le n_1\le t} I\big(p_{1,n_1}, n_1, \alpha(t)\big)\Big) \tag{105}$$

$$\le Q_a(T) + \sum_{t=Q_a(T)+1}^{T}\sum_{n_1=1}^{t}\sum_{n_a=Q_a(T)}^{t} \mathbb{P}\Big(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \ge I\big(p_{1,n_1}, n_1, \alpha(t)\big)\Big) \tag{106}$$

$$\le Q_a(T) + \sum_{t=Q_a(T)+1}^{T}\sum_{n_1=1}^{t}\sum_{n_a=Q_a(T)}^{t} \Big(\underbrace{\mathbb{P}\left(G_1^-(n_1,t)^{\mathsf{c}}\right)}_{\le \frac{1}{t^4}} + \underbrace{\mathbb{P}\left(G_a^+(n_a,t)^{\mathsf{c}}\right)}_{\le \frac{1}{t^4}}\Big) \tag{107}$$

$$+ \sum_{t=Q_a(T)+1}^{T}\sum_{n_1=1}^{t}\sum_{n_a=Q_a(T)}^{t} \mathbb{P}\Big(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \ge I\big(p_{1,n_1}, n_1, \alpha(t)\big), G_1^-(n_1,t), G_a^+(n_a,t)\Big) \tag{108}$$

$$\le Q_a(T) + \frac{\pi^2}{3} + \sum_{t=Q_a(T)+1}^{T}\sum_{n_1=1}^{t}\sum_{n_a=Q_a(T)}^{t} \mathbb{P}\Big(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \ge I\big(p_{1,n_1}, n_1, \alpha(t)\big), G_1^-(n_1,t), G_a^+(n_a,t)\Big), \tag{109}$$

where the last equation follows from the fact that $\sum_{t=Q_a(T)+1}^{T}(\frac{1}{t^2}) \le \pi^2/6$ and (103) can be obtained by taking the expectation on both sides of the first inequality of (6) in (Auer et al., 2002) and using the fact that arm i is chosen implies that i's index is larger than the optimal arm's. Next, to provide an upper bound for (109), we need to consider the following three cases separately. As suggested by (109), we can focus on the case where $n_a \ge Q_a(T)$.

- **Case 1:** $n_1 > \frac{4}{D(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_1)}\log t$

  Since $n_1 > \frac{4}{D(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_1)}\log t$, we have $p_{1,n_1} \geq \theta_1 - \frac{\varepsilon}{2}\Delta$ on the event $G_1^-(n_1,t)$. Similarly, as $n_a \geq Q_a(T) > \frac{4}{D(\theta_a+\frac{\varepsilon}{2}\Delta_a,\theta_a)}\log t$, we have $p_{a,n_a} \leq \theta_a + \frac{\varepsilon}{2}\Delta_a$ on the event $G_a^+(n_a,t)$. Therefore, we know

  $$p_{1,n_1} - p_{a,n_a} > (1-\varepsilon)\Delta. \tag{110}$$

  Then, we have

  $$\begin{aligned} I(p_{1,n_1},n_1,\alpha(t)) &> I(\theta_1 - \frac{\varepsilon}{2}\Delta, n_1, \alpha(t)) && (111)\\ &\geq I(\theta_a - \frac{\varepsilon}{2}\Delta, K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_a+\frac{\varepsilon}{2}\Delta)\alpha(t),\alpha(t)) && (112)\\ &\geq I(\theta_a - \frac{\varepsilon}{2}\Delta_a, K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_a+\frac{\varepsilon}{2}\Delta)\alpha(t),\alpha(t)) && (113)\\ &\geq I(p_{a,n_a}, K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_a+\frac{\varepsilon}{2}\Delta)\alpha(t),\alpha(t)) && (114)\\ &\geq I(p_{a,n_a}, Q_a(T),\alpha(t)) && (115)\\ &\geq I(p_{a,n_a}, n_a,\alpha(t)), && (116) \end{aligned}$$

  where (111) and (113)-(114) hold by Lemma 1.(i) (i.e., $K^*(\theta,\theta')$ is strictly decreasing with respect to θ and strictly increasing with respect to θ'), (112) holds by Lemma 2, and (115)-(116) follow from Lemma 1.(i). Hence, in Case 1, we always have $I\big(p_{1,n_1},n_1,\alpha(t)\big) > I\big(p_{a,n_a},n_a,\alpha(t)\big)$.

- **Case 2:** $n_1 \leq \frac{4}{D(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_1)}\log t$ and $n_1 \leq K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\underline{\theta})\alpha(t)$

  Similar to Case 1, since $n_a \geq Q_a(T) > \frac{4}{D(\theta_a,\theta_a+\frac{\varepsilon}{2}\Delta_a)}\log t$, we have $p_{a,n_a} \leq \theta_a + \frac{\varepsilon}{2}\Delta_a$ on the event $G_a^+(n_a,t)$. Moreover, as $n_1 \leq K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\underline{\theta})\alpha(t)$ and $n_a \geq Q_a(T) > K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_a+\frac{\varepsilon}{2}\Delta)\alpha(t)$, by Lemma 3 we know

  $$I(\underline{\theta},n_1,\alpha(t)) > I(\theta_a+\frac{\varepsilon}{2}\Delta,n_a,\alpha(t)). \tag{117}$$

  Therefore, we obtain that

  $$\begin{aligned} I\big(p_{1,n_1},n_1,\alpha(t)\big) &> I\big(\underline{\theta},n_1,\alpha(t)\big) && (118)\\ &> I(\theta_a+\frac{\varepsilon}{2}\Delta,n_a,\alpha(t)) && (119)\\ &> I\big(p_{a,n_a},n_a,\alpha(t)\big), && (120) \end{aligned}$$

  where (118) and (120) follow from Lemma 1.(ii), and (119) is a direct result of (117). Hence, in Case 2, we still have $I\big(p_{1,n_1},n_1,\alpha(t)\big) > I\big(p_{a,n_a},n_a,\alpha(t)\big)$.

- **Case 3:** $n_1 \leq \frac{4}{D(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_1)}\log t$ and $n_1 > K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\underline{\theta})\alpha(t)$

  Recall that $\alpha(t) = C_\alpha \log t$ with $C_\alpha \geq 4/(D(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_1)\cdot K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\underline{\theta}))$. Therefore, the two events $\{n_1 \leq \frac{4}{D(\theta_1-\frac{\varepsilon}{2}\Delta,\theta_1)}\log t\}$ and $\{n_1 > K^*(\theta_1-\frac{\varepsilon}{2}\Delta,\underline{\theta})\alpha(t)\}$ cannot happen at the same time.

To sum up, in all the above three cases, we have

$$\mathbb{P}\left(I\big(p_{a,n_a},n_a,\alpha(t)\big) \geq I\big(p_{1,n_1},n_1,\alpha(t)\big), G_1^-(n_1,t), G_a^+(n_a,t)\right) = 0. \tag{121}$$

By (109) and (121), we conclude that $E[N_a(T)] \leq Q_a(T) + \frac{\pi^2}{3}$, for every $a \neq 1$.

Finally, the total regret can be upper bounded as

$$\begin{aligned} \mathcal{R}(T) &\leq \sum_{a=2}^{N} \Delta_a \cdot E[N_a(T)] && (122)\\ &= \sum_{a=2}^{N} \Delta_a \left[\max\left\{\frac{4}{D(\theta_a+\frac{\varepsilon}{2}\Delta_a,\theta_a)}, C_\alpha K^*(\theta_1-\frac{\varepsilon}{2}\Delta_a,\theta_a+\frac{\varepsilon}{2}\Delta_a)\right\}\log T + 1 + \frac{\pi^2}{3}\right]. && (123) \end{aligned}$$

□

J. Proof of Proposition 3

Proof Sketch: We extend the proof procedure of Proposition 2 for Gaussian rewards, with the help of Hoeffding's inequality. We then prove an additional lemma, which shows that conditioned on the "good" events, the RBMLE index of the optimal arm (i.e., arm 1) is always larger than that of a sub-optimal arm a if $N_a(t) \geq \frac{2}{\Delta_a}\alpha(t)$ and $\alpha(t) \geq \frac{256\sigma^2}{\Delta_a}$, regardless of $N_1(t)$.

We extend the proof of Proposition 2 to the case of Gaussian rewards. To begin with, we define the confidence intervals and the "good" events. Recall that for each arm i, we define $p_{i,n}$ to be the empirical average reward collected in the first n pulls of arm i. For each arm i, for each pair of $n, t \in \mathbb{N}$, we define

$$\delta_i(n,t) := \inf\Big\{\delta : \max\big\{\exp(-nD(\theta_i+\delta,\theta_i)), \exp(-nD(\theta_i-\delta,\theta_i))\big\} \leq \frac{1}{t^4}\Big\}. \tag{124}$$

Accordingly, for each arm i and for each pair of $n, t \in \mathbb{N}$, we define the following events:

$$G_i(n,t) = \Big\{|p_{i,n} - \theta_i| \leq \delta_i(n,t)\Big\}, \tag{125}$$

For the Gaussian rewards, we can leverage Hoeffding's inequality for sub-Gaussian distributions as follows:

Lemma J.1 *Under σ-sub-Gaussian rewards for all arms, for any $n \in \mathbb{N}$, we have*

$$\mathbb{P}(|p_{i,n} - \theta_i| \geq \delta) \leq 2\exp(-\frac{n}{2\sigma^2}\delta^2). \tag{126}$$

Proof of Lemma J.1: This is a direct result of Proposition 2.5 in (Wainwright, 2019). □

Based on Lemma J.1, we focus on the case $D(\theta', \theta'') = \frac{1}{2\sigma^2}(|\theta' - \theta''|)^2$ and $\delta_i(n,t) = \sqrt{(8\sigma^2 \log t)/n}$. For ease of notation, we use γ_* to denote the constant $8\sigma^2$.

Before providing the regret analysis, we first introduce the following useful lemma.

Lemma J.2 *Suppose $\gamma > 0$ and $\mu_1, \mu_2 \in \mathbb{R}$ with $\mu_1 > \mu_2$. Given $\alpha(t) = c\log t$ with $c \geq \frac{32\gamma}{\mu_1-\mu_2}$, for any $n_2 \geq \frac{2}{\mu_1-\mu_2}\alpha(t)$ and any $n_1 > 0$, we have $I(\mu_1 - \sqrt{(\gamma \log t)/n_1}, n_1, \alpha(t)) > I(\mu_2 + \sqrt{(\gamma \log t)/n_2}, n_2, \alpha(t))$.*

Proof of Lemma J.2: We start by considering $n_2 \geq M\alpha(t)$, for some $M > 0$. Then, note that

$$I\Big(\mu_1 - \sqrt{\frac{\gamma \log t}{n_1}}, n_1, \alpha(t)\Big) = \mu_1 - \sqrt{\frac{\gamma \log t}{n_1}} + \frac{\alpha(t)}{2n_1}, \tag{127}$$

$$I\Big(\mu_2 + \sqrt{\frac{\gamma \log t}{n_2}}, n_2, \alpha(t)\Big) = \mu_2 + \sqrt{\frac{\gamma \log t}{n_2}} + \frac{\alpha(t)}{2n_2}. \tag{128}$$

For ease of notation, we use x_1 and x_2 to denote $\sqrt{(\gamma \log t)/n_1}$ and $\sqrt{(\gamma \log t)/n_2}$, respectively. Then, we know

$$I\Big(\mu_1 - \sqrt{\frac{\gamma \log t}{n_1}}, n_1, \alpha(t)\Big) - I\Big(\mu_2 + \sqrt{\frac{\gamma \log t}{n_1}}, n_2, \alpha(t)\Big) \geq (\mu_1 - \mu_2) - (x_1 + x_2) + \frac{c}{2\gamma}(x_1^2 - x_2^2) \tag{129}$$

$$\geq (\mu_1 - \mu_2) - x_1 - \sqrt{\frac{\gamma}{cM}} + \frac{c}{2\gamma}x_1^2 - \frac{1}{2M}, \tag{130}$$

where (130) follows from $n_2 \geq M\alpha(t)$. Define $w(x_1) := (\mu_1 - \mu_2) - x_1 - \sqrt{\frac{\gamma}{cM}} + \frac{c}{2\gamma}x_1^2 - \frac{1}{2M}$. The quadratic polynomial $w(x_1)$ remains positive for all $x_1 \in \mathbb{R}$ if the discriminant of $w(x_1)$, denoted by $\text{Disc}(w(x_1))$, is negative. Indeed, we have

$$\text{Disc}(w(x_1)) = 1 - 4 \cdot \frac{c}{2\gamma} \cdot (-\sqrt{\frac{\gamma}{cM}} - \frac{1}{2M} + (\mu_1 - \mu_2)) \leq -39, \tag{131}$$

where the last inequality follows from $c \geq \frac{32\gamma}{\mu_1-\mu_2}$ and $M = \frac{2}{\mu_1-\mu_2}$. □

Now, we are ready to prove Proposition 3: Consider the bias term $\alpha(t) = C_\alpha \log t$ with $C_\alpha \geq \frac{32\gamma^*}{\Delta}$, where $\gamma^* = 8\sigma^2$. Recall that we assume arm 1 is the unique optimal arm. Our target is to quantify the total number of trials of each

sub-optimal arm. Next, we characterize the expected total number of trials of each sub-optimal arm, i.e., $\mathbb{E}[N_a(T)]$. We define $Q_a^*(T) = \frac{2}{\Delta_a} C_\alpha \log T$. By using a similar argument to (102)-(109), we have

$$\mathbb{E}[N_a(T)] \leq Q_a^*(T) + \sum_{t=Q_a^*(T)+1}^{T} \mathbb{P}\left(I\big(p_a(t), N_a(t), \alpha(t)\big) \geq I\big(p_1(t), N_1(t), \alpha(t)\big), N_a(t) \geq Q_a^*(T)\right) \tag{132}$$

$$\leq Q_a^*(T) + \sum_{t=Q_a^*(T)+1}^{T} \sum_{n_1=1}^{t} \sum_{n_a=Q_a^*(T)}^{t} \mathbb{P}\left(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \geq I\big(p_{1,n_1}, n_1, \alpha(t)\big)\right) \tag{133}$$

$$\leq Q_a^*(T) + \sum_{t=Q_a^*(T)+1}^{T} \sum_{n_1=1}^{t} \sum_{n_a=Q_a^*(T)}^{t} \Big(\underbrace{\mathbb{P}\left(G_1(n_1,t)^{\mathsf{c}}\right)}_{\leq \frac{2}{t^4}} + \underbrace{\mathbb{P}\left(G_a(n_a,t)^{\mathsf{c}}\right)}_{\leq \frac{2}{t^4}}\Big) \tag{134}$$

$$+ \sum_{t=Q_a^*(T)+1}^{T} \sum_{n_1=1}^{t} \sum_{n_a=Q_a^*(T)}^{t} \mathbb{P}\left(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \geq I\big(p_{1,n_1}, n_1, \alpha(t)\big), G_1(n_1,t), G_a(n_a,t)\right) \tag{135}$$

$$\leq Q_a^*(T) + \frac{2\pi^2}{3} + \sum_{t=Q_a^*(T)+1}^{T} \sum_{n_1=1}^{t} \sum_{n_a=Q_a^*(T)}^{t} \mathbb{P}\left(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \geq I\big(p_{1,n_1}, n_1, \alpha(t)\big), G_1(n_1,t), G_a(n_a,t)\right). \tag{136}$$

Conditioned on the events $G_i(n_1,t)$ and $G_a(n_a,t)$, we obtain that

$$I\big(p_{1,n_1}, n_1, \alpha(t)\big) \geq I(\theta_1 - \sqrt{(\gamma_* \log t)/n_1}, n_1, \alpha(t)) \tag{137}$$

$$> I(\theta_a + \sqrt{(\gamma_* \log t)/n_a}, n_a, \alpha(t)) \tag{138}$$

$$\geq I\big(p_{a,n_a}, n_a, \alpha(t)\big), \tag{139}$$

where (137) and (139) follow from Lemma 1.(i), and (138) follows from Lemma J.2. Hence, for $n_1 > 0$ and $n_a \geq Q_a^*(T)$,

$$\mathbb{P}\left(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \geq I\big(p_{1,n_1}, n_1, \alpha(t)\big), G_1(n_1,t), G_a(n_a,t)\right) = 0. \tag{140}$$

By (136) and (140), we know $E[N_a(T)] \leq Q_a^*(T) + \frac{2\pi^2}{3}$, for every $a \neq 1$. Hence, the total regret can be upper bounded as

$$\mathcal{R}(T) \leq \sum_{a=2}^{N} \Delta_a \big[\frac{2}{\Delta_a} C_\alpha \log T + \frac{2\pi^2}{3}\big]. \tag{141}$$

□

K. Proof of Proposition 5

The proof of Proposition 2 can be easily extended to Proposition 5 by replacing the Chernoff bound with the sub-Exponential tail bound. For sub-exponential reward distributions, we consider the sub-exponential tail bound as follows:

Lemma K.1 *Under (ρ, κ)-sub-exponential rewards for all arms, for any $n \in \mathbb{N}$, we have*

$$\mathbb{P}(p_{i,n} - \theta_i \geq \delta) \leq \exp\Big(-\frac{n^2\delta^2}{2(n\kappa\delta + \rho^2)}\Big). \tag{142}$$

Similar to the proof of Proposition 2, we consider the bias term $\alpha(t) = C_\alpha \log t$, but with $C_\alpha \geq 16(\kappa\varepsilon\Delta + 2\rho^2)/((\varepsilon\Delta)^2 K^*(\theta_1 - \frac{\varepsilon\Delta}{2}, \underline{\theta}))$. Note that here we simply replace $D(\theta_1 - \frac{\varepsilon\Delta}{2}, \theta_1)$ with $\frac{(\varepsilon\Delta)^2}{4(\kappa\varepsilon\Delta + 2\rho^2)}$ by comparing (142) with (93). Similarly, we define

$$\tilde{Q}_a(T) := \max\Big\{\frac{16(\kappa\varepsilon\Delta + 2\rho^2)}{(\varepsilon\Delta_a)^2}, C_\alpha K^*(\theta_1 - \frac{\varepsilon}{2}\Delta_a, \theta_a + \frac{\varepsilon}{2}\Delta_a)\Big\} \log T + 1. \tag{143}$$

Note that the proof of Proposition 2 relies only on Lemmas 1-3, and these lemmas are tied to the distributions for deriving the RBMLE index, not to the underlying true reward distributions. Therefore, it is easy to verify that the same proof procedure still holds here by replacing $Q_a(T)$ with $\tilde{Q}_a(T)$. □

L. Proof of Proposition 6

Proof Sketch: An $O(\log T)$ regret bound can be obtained by considering the extensions as follows:

- By extending Lemma J.2, we show that for any two arms i and j, there exist constants $M_1 > 0$ and $M_2 > 0$ such that $I_i > I_j$ for any $n_i \leq M_1 \log T$ and $n_j \geq M_2 \log T$.
- We then extend (132)-(133) by using the fact that arm a is chosen implies that its index is larger than all the other arm's.
- Extend (134)-(136) by considering the good events across all arms (instead of just arm a and the optimal arm).
- Finally, we extend (137)-(139) by using Lemma J.2 and the fact that under the good events, the estimated $\hat{\Delta}_t$ is between $\Delta/2$ and Δ.

In Algorithm 1, since gradually learning the minimal gap Δ involves all the arms (see Line 7 of Algorithm 1), we need to extend Lemma J.2 to remove the assumption $\mu_1 > \mu_2$. The extension is conducted in Lemma L.1 below.

Lemma L.1 *Let γ be a positive constant. For any two arms i and j with $\mu_i, \mu_j \in \mathbb{R}$ and $\delta := \mu_j - \mu_i$, given $\alpha(t) = c \log t$ with $c \geq \frac{32\gamma(N+2)}{\Delta}$, then for any $n_i \leq \frac{1}{8} M \frac{\Delta}{\max(\delta,\Delta)} \log(t)$ and $n_j \geq M \log t$, we have $I(\mu_i - \sqrt{(\gamma \log t)/n_i}, n_i, \alpha(t)) > I(\mu_j + \sqrt{(\gamma \log t)/n_j}, n_j, \alpha(t))$, where $M = \frac{32\gamma(N+2)}{\Delta^2}$.*

Proof of Lemma L.1: For ease of notation, we use I_i and I_j to denote $I(\mu_i - \sqrt{(\gamma \log t)/n_i}, n_i, \alpha(t))$ and $I(\mu_j + \sqrt{(\gamma \log t)/n_j}, n_j, \alpha(t))$, respectively, within this proof. Then, we have

$$I_i = \mu_i - \sqrt{\frac{\gamma \log t}{n_i}} + \frac{\alpha(t)}{2n_i}, \tag{144}$$

$$I_j = \mu_j + \sqrt{\frac{\gamma \log t}{n_j}} + \frac{\alpha(t)}{2n_j}. \tag{145}$$

Using x to denote $\sqrt{(\gamma \log t)/n_i}$, we have

$$I_i - I_j \geq (\mu_i - \mu_j) - x - \sqrt{\frac{\gamma}{M}} + \frac{c}{2\gamma} x^2 - \frac{c}{2\gamma} \Big(\sqrt{\frac{\gamma}{M}} \Big)^2 \tag{146}$$

$$= -\delta - \sqrt{\frac{\gamma}{M}} - \frac{c}{2M} + \frac{c}{2\gamma} x^2 - x, \tag{147}$$

where (146) follows from $n_j \geq M \log(t)$. Since $n_i \leq \frac{1}{8} M \frac{\Delta}{\max(\delta,\Delta)} \log(t)$, we have

$$x = \sqrt{\frac{\gamma \log t}{n_i}} \geq \sqrt{\frac{8\gamma \log t}{\frac{32\gamma(N+2)}{\Delta^2} \frac{\Delta}{\max(\delta,\Delta)}}} = \sqrt{\frac{\Delta \cdot \max(\delta,\Delta)}{4(N+2)}}. \tag{148}$$

By (148) and the fact that the $cx^2/(2\gamma) - x$ has its minimum at $x = \gamma/c \leq \Delta/(32(N+2))$, we can construct a lower bound for $cx^2/(2\gamma) - x$ in (147):

$$\frac{c}{2\gamma} x^2 - x \geq \frac{c}{2\gamma} \frac{8\gamma}{M} \frac{\max(\delta,\Delta)}{\Delta} - \sqrt{\frac{8\gamma}{M} \frac{\max(\delta,\Delta)}{\Delta}}. \tag{149}$$

Then we can obtain a lower bound of $I_i - I_j$:

$$I_i - I_j \geq -\delta - \sqrt{\frac{\gamma}{M}} - \frac{c}{2M} + \frac{c}{2M} \cdot 8 \frac{\max(\delta,\Delta)}{\Delta} - \sqrt{\frac{8\gamma}{M} \cdot \frac{\max(\delta,\Delta)}{\Delta}} \tag{150}$$

$$= -\delta - \Big(\sqrt{\frac{\gamma}{M}} + \sqrt{\frac{8\gamma}{M} \cdot \frac{\max(\delta,\Delta)}{\Delta}} \Big) - \Big(\frac{c}{2M} - \frac{c}{2M} \cdot 8 \frac{\max(\delta,\Delta)}{\Delta} \Big) \tag{151}$$

$$\geq -\delta - (\sqrt{8}+1) \sqrt{\frac{\gamma}{M} \cdot \frac{\max(\delta,\Delta)}{\Delta}} + \frac{c}{2M} \cdot 7 \cdot \frac{\max(\delta,\Delta)}{\Delta} \tag{152}$$

$$\geq -\delta - (\sqrt{8}+1) \sqrt{\frac{\Delta \max(\delta,\Delta)}{32(N+2)}} + \frac{7}{2} \max(\delta,\Delta), \tag{153}$$

$$> 0, \tag{154}$$

where (152)-(153) follow from that $c \geq \frac{32\gamma(N+2)}{\Delta}$ and $M = \frac{32\gamma(N+2)}{\Delta^2}$. □

Proof of Proposition 6: First, we set $\gamma = 8\sigma^2$. Recall that $T_0 := \min\{t \in \mathbb{N} : \beta(t) \geq \frac{32\gamma(N+2)}{\Delta}\} < \infty$. Within this proof, we take $\delta_i(n,t) = \sqrt{(2\sigma^2(N+2)\log t)/n}$ and define the good events as $G_i(n,t) := \{|p_{i,n} - \theta_i| \leq \delta_i(n,t)\}$. By Lemma J.1, we know $\mathbb{P}(G_i(n_i,t)^c) \leq 2/t^{N+2}$, for all t and i. We also define $Q_a^*(T) := \max\{4 \cdot \frac{32\gamma(N+2)}{\Delta^2} \log T, T_0\}$. Denote by $\mathbb{E}[N_a(T)]$ the expected total number of trials of arm a. Then for each a, we can construct an upper bound of $\mathbb{E}[N_a(T)]$ by:

$$\mathbb{E}[N_a(T)] \leq Q_a^*(T) + \sum_{t=Q_a^*(T)+1}^{T} \mathbb{P}\left(I\big(p_a(t), N_a(t), \alpha(t)\big) \geq I\big(p_i(t), N_i(t), \alpha(t)\big), \forall i \neq a, N_a(t) \geq Q_a^*(T)\right) \tag{155}$$

$$\leq Q_a^*(T) + \sum_{t=Q_a^*(T)+1}^{T} \sum_{\substack{n_i=1 \\ i\neq a}}^{t} \sum_{n_a=Q_a^*(T)}^{t} \mathbb{P}\left(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \geq I\big(p_{i,n_i}, n_i, \alpha(t), \forall i \neq a\big)\right) \tag{156}$$

$$\leq Q_a^*(T) + \sum_{t=Q_a^*(T)+1}^{T} \sum_{\substack{n_i=1 \\ i\neq a}}^{t} \sum_{n_a=Q_a^*(T)}^{t} \Big(\underbrace{\sum_{i}^{N} \mathbb{P}\left(G_i(n_i,t)^c\right)}_{\leq 2N/t^{N+2}}\Big) \tag{157}$$

$$+ \sum_{t=Q_a^*(T)+1}^{T} \sum_{\substack{n_i=1 \\ i\neq a}}^{t} \sum_{n_a=Q_a^*(T)}^{t} \mathbb{P}\left(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \geq I\big(p_{i,n_i}, n, \alpha(t)\big), \forall i \neq a, G_1(n_1,t), G_2(n_2,t), \cdots, G_N(n_N,t)\right) \tag{158}$$

$$\leq Q_a^*(T) + \frac{N\pi^2}{3} \tag{159}$$

$$+ \sum_{t=Q_a^*(T)+1}^{T} \sum_{n_1=1}^{t} \sum_{n_a=Q_a^*(T)}^{t} \mathbb{P}\left(I\big(p_{a,n_a}, n_a, \alpha(t)\big) \geq I\big(p_{i,n_i}, n_i, \alpha(t)\big), \forall i \neq a, G_1(n_1,t), G_2(n_2,t), \cdots, G_N(n_N,t)\right). \tag{160}$$

- **Case 1:** $n_i \geq \frac{1}{8}\frac{32\gamma(N+2)}{\Delta^2}\frac{\Delta}{\max(\theta_a - \theta_i, \Delta)}\log t$ for all $i \neq a$. Since n_i is large enough, it is easy to check that the following inequality holds under the good events for all $i \neq a$:

$$|p_{i,n_i} - \theta_i| \leq \sqrt{\frac{2\sigma^2 \log t}{\frac{1}{8}\frac{32\gamma(N+2)}{\Delta^2}\frac{\Delta}{\max(\theta_a - \theta_i)}\log t}} \leq \frac{\max(\theta_a - \theta_i, \Delta)}{4\sqrt{N+2}} \leq \frac{\max(\theta_a - \theta_i, \Delta)}{8}. \tag{161}$$

  Similarly, we have $|p_{a,n_a} - \theta_a| \leq \frac{\Delta}{8\sqrt{2}\sqrt{N+2}}$. Moreover, by checking $L_i(t)$ and $U_i(t)$ under the good events, we also know $\widehat{\Delta}_t \geq \frac{\Delta}{2}$. This also implies that $\widehat{C}_\alpha(t) \leq \frac{32\gamma(N+2)}{\Delta/2}$ and hence $\alpha(t) \leq \frac{32\gamma(N+2)}{\Delta/2}\log t$. Therefore by Lemma J.2, we know that $I\big(p_{1,n_1}, n_1, \alpha(t)\big) \geq I\big(p_{a,n_a}, n_a, \alpha(t)\big)$.

- **Case 2:** There exists some $i \neq a$ such that $n_i < \frac{1}{8}\frac{32\gamma(N+2)}{\Delta^2}\frac{\Delta}{\max(\theta_a - \theta_i, \Delta)}\log t$. By Lemma L.1, this implies that $I\big(p_{i,n_i}, n_i, \alpha(t)\big) > I\big(p_{a,n_a}, n_a, \alpha(t)\big)$.

Therefore, we have for every $a \neq 1$,

$$E[N_a(T)] \leq Q_a^*(T) + \frac{N\pi^2}{3} \tag{162}$$

$$= \max\left\{\frac{128\gamma(N+2)}{\Delta^2}\log T, T_0\right\} + \frac{N\pi^2}{3}. \tag{163}$$

Hence, the total regret can be upper bounded as

$$\mathcal{R}(T) \leq \sum_{a=2}^{N} \Delta_a \Big[\max\left\{\frac{128\gamma(N+2)}{\Delta^2}\log T, T_0\right\} + \frac{N\pi^2}{3}\Big]. \tag{164}$$

□

M. Additional Empirical Results

In this subsection, we present additional empirical results for more examples to demonstrate the effectiveness, efficiency and scalability of the proposed RBMLE algorithm.

M.1. Effectiveness

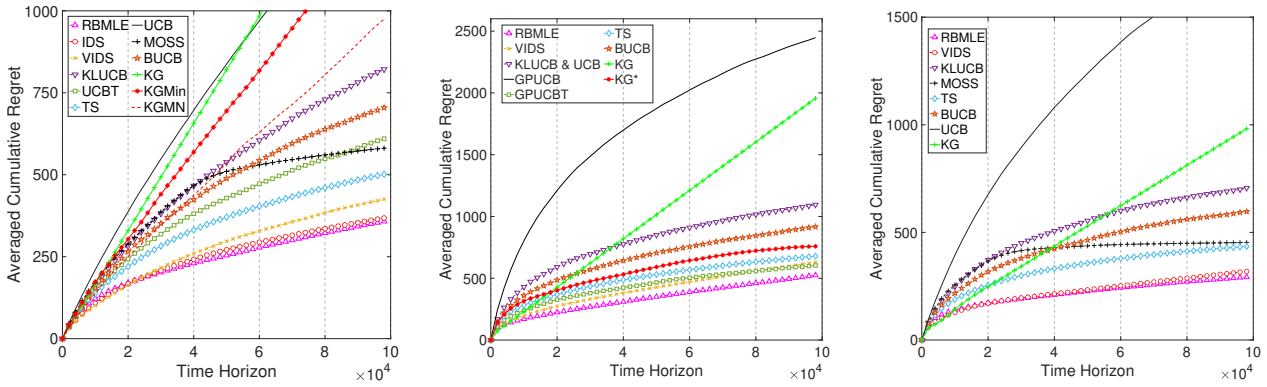
Figures 3-4 illustrate the effectiveness of RBMLE with respect to the cumulative regret, under a different set of parameters, for the three types of bandits. Tables 2-10 provide detailed statistics, including the mean as well as the standard deviation and quantiles of the final regrets, with the row-wise smallest values highlighted in boldface. From the Tables, we observe that RBMLE tends to have the smallest value of regret at medium to high quantiles, and comparable to the smallest values at other lower quantiles among those that have comparable mean values (e.g., IDS, VIDS, KLUCB). Along with the presented statistics of standard deviation, they suggest that RBMLE's performance enjoys comparable robustness as those baselines that achieve similar mean regret.

M.2. Efficiency

Figures 5-6 present the efficiency of RBMLE in terms of averaged computation time per decision (ACTPD) vs. averaged final cumulative regret. The computation times are measured on a Linux server with (i) an Intel Xeon E7 v4 server operating at a maximal clock rate of 3.60 GHz, and (ii) a total of 528 GB memory. While there are 64 cores in the server, we force the program to run on just one core for a fair comparison.

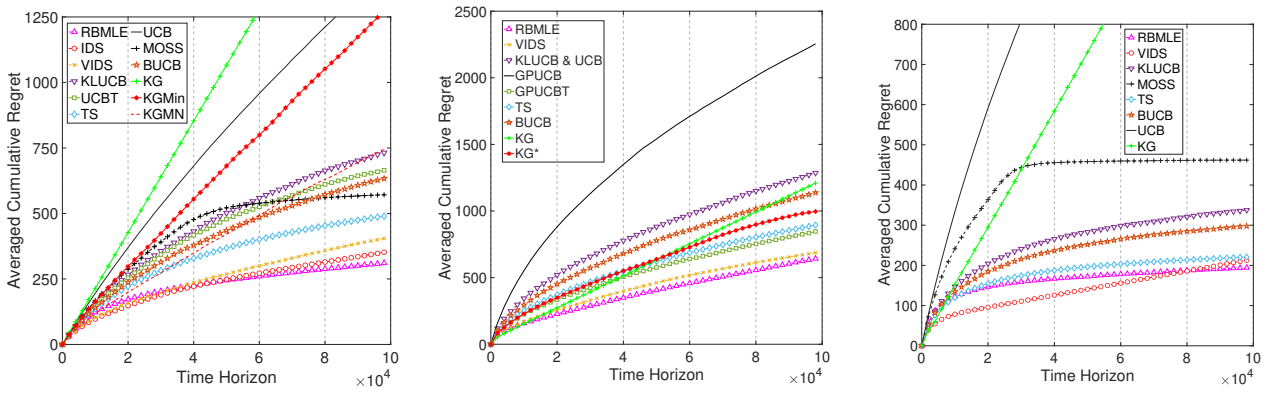
M.3. Scalability

Tables 11-13 show the computation time per decision of different policies under varying numbers of arms.



(a) (b) (c)

Figure 3: Averaged cumulative regret: (a) Bernoulli bandits with $(\theta_i)_{i=1}^{10} = (0.655, 0.6, 0.665, 0.67, 0.675, 0.68, 0.685, 0.69, 0.695, 0.7)$ & $\Delta = 0.005$; (b) Gaussian bandits with $(\theta_i)_{i=1}^{10} = (0.5, 0.75, 0.4, 0.6, 0.55, 0.76, 0.68, 0.41, 0.52, 0.67)$ & $\Delta = 0.01$; (c) Exponential bandits with $(\theta_i)_{i=1}^{10} = (0.46, 0.45, 0.5, 0.48, 0.51, 0.4, 0.43, 0.42, 0.45, 0.44)$ & $\Delta = 0.01$.



(a) (b) (c)

Figure 4: Averaged cumulative regret: (a) Bernoulli bandits with $(\theta_i)_{i=1}^{10} = (0.755, 0.76, 0.765, 0.77, 0.775, 0.78, 0.785, 0.79, 0.795, 0.8)$ & $\Delta = 0.005$; (b) Gaussian bandits with $(\theta_i)_{i=1}^{10} = (0.65, 0.35, 0.66, 0.4, 0.65, 0.64, 0.55, 0.4, 0.57, 0.54)$ & $\Delta = 0.01$; (c) Exponential bandits with $(\theta_i)_{i=1}^{10} = (0.25, 0.28, 0.27, 0.3, 0.29, 0.22, 0.21, 0.24, 0.23, 0.26)$ & $\Delta = 0.01$.

Table 2: Statistics of the final cumulative regret in Figure 1(a). The best in each row is highlighted.

Algorithm	**RBMLE**	**IDS**	**VIDS**	**KLUCB**	**UCBT**	**TS**	**UCB**	**MOSS**	**BUCB**	**KG**	**KGMin**	**KGMN**
Mean Regret	**263.5**	406.3	449.6	730.4	474.7	426.9	1809.5	464.5	580.9	2379.5	2384.2	1814.3
Std. Dev.	233.5	466.7	618.2	109.3	176.3	149.3	113.0	**93.1**	105.8	2163.2	355.4	344.0
Quantile .10	142.4	74.7	54.2	584.4	309.9	283.2	1647.3	355.8	452.9	**3.7**	1899.1	1344.8
Quantile .25	161.4	113.5	**90.0**	661.0	362.8	326.6	1750.4	394.2	519.2	1000.9	2103.2	1555.3
Quantile .50	190.6	184.3	**134.5**	717.7	448.7	404.0	1815.5	458.9	563.77	2001.4	2411.6	1844.2
Quantile .75	**237.8**	461.4	1043.2	804.0	543.5	489.2	1874.3	520.8	638.0	3999.8	2620.3	2057.7
Quantile .90	**430.0**	1138.3	1116.4	860.9	655.0	595.1	1963.1	570.8	713.8	5013.8	2809.4	2254.1
Quantile .95	993.1	1247.9	1247.9	926.3	759.7	647.6	1996.6	**615.6**	766.0	6992.2	2911.4	2326.5

Table 3: Statistics of the final cumulative regret in Figure 3(a). The best in each row is highlighted.

Algorithm	**RBMLE**	**IDS**	**VIDS**	**KLUCB**	**UCBT**	**TS**	**UCB**	**MOSS**	**BUCB**	**KG**	**KGMin**	**KGMN**
Mean Regret	**361.5**	371.3	416.7	831.9	616.9	505.8	1437.9	582.9	712.5	1637.9	1309.0	991.1
Std. Dev.	247.6	285.9	342.8	131.9	130.7	156.3	**78.5**	169.9	120.3	1592.7	214.1	198.4
Quantile .10	133.0	116.3	77.2	650.3	440.6	334.4	1335.8	411.6	564.3	**2.3**	1013.3	741.9
Quantile .25	165.1	164.4	**147.0**	732.4	532.3	385.0	1388.3	461.1	641.2	501.5	1184.1	876.2
Quantile .50	**223.5**	262.8	248.3	823.1	598.0	477.9	1436.7	532.7	715.1	1002.6	1338.8	987.8
Quantile .75	608.4	**568.9**	593.1	930.3	693.4	575.3	1495.6	654.8	782.6	2996.5	1447.2	1119.8
Quantile .90	**661.2**	681.7	1003.2	1020.6	779.5	698.3	1540.4	816.1	865.3	3499.3	1538.3	1228.4
Quantile .95	**722.9**	835.1	1060.9	1062.0	857.9	793.4	1561.2	943.2	906.4	4497.6	1620.9	1343.0

Table 4: Statistics of the final cumulative regret in Figure 4(a). The best in each row is highlighted.

Algorithm	**RBMLE**	**IDS**	**VIDS**	**KLUCB**	**UCBT**	**TS**	**UCB**	**MOSS**	**BUCB**	**KG**	**KGMin**	**KGMN**
Mean Regret	**313.2**	355.7	425.5	740.5	669.5	493.0	1445.6	572.1	638.9	2131.2	1301.0	757.2
Std. Dev.	228.1	386.5	474.7	126.9	120.3	171.9	**69.1**	132.7	127.8	1336.5	209.8	178.3
Quantile .10	142.2	95.1	**70.2**	581.4	506.7	328.9	1347.8	433.6	485.6	451.3	1024.6	531.2
Quantile .25	169.4	133.2	**102.5**	651.6	573.5	374.0	1396.6	463.5	542.7	1001.1	1174.2	628.7
Quantile .50	203.7	179.0	**173.9**	725.9	680.5	463.3	1446.4	543.0	623.3	2001.2	1322.2	754.0
Quantile .75	**369.7**	543.2	589.7	806.5	751.9	534.0	1497.7	648.6	724.3	3000.2	1442.1	897.2
Quantile .90	**680.2**	695.5	1067.4	886.3	833.6	726.2	1541.1	760.8	813.5	3999.8	1563.4	963.8
Quantile .95	**720.2**	891.4	1739.8	999.6	867.1	774.1	1554.6	830.9	867.4	4498.3	1590.7	1039.2

Table 5: Statistics of the final cumulative regret in Figure 1(b). The best in each row is highlighted.

Algorithm	**RBMLE**	**VIDS**	**KLUCB&UCB**	**GPUCB**	**GPUCBT**	**TS**	**BUCB**	**KG**	**KG***
Mean Regret	**730.6**	775.0	1412.2	2640.3	848.5	932.7	1222.3	1684.3	1046.0
Std. Dev.	827.4	678.7	**219.2**	227.0	314.2	282.1	231.4	2056.8	238.9
Quantile .10	135.3	233.9	1147.2	2382.8	529.3	657.8	960.8	**20.4**	788.0
Quantile .25	160.2	336.0	1272.1	2500.0	608.0	706.6	1036.5	**59.9**	891.6
Quantile .50	**263.1**	544.1	1395.9	2600.4	814.7	876.0	1205.9	1035.8	1000.6
Quantile .75	1140.8	1137.7	1545.9	2787.1	**1001.1**	1125.3	1390.6	2028.0	1171.1
Quantile .90	2107.9	1516.5	1674.6	2916.1	**1228.6**	1304.8	1512.9	4028.8	1314.1
Quantile .95	2157.6	1862.0	1724.6	3024.4	1578.7	1472.7	1565.5	7818.3	**1413.7**

Table 6: Statistics of the final cumulative regret in Figure 3(b). The best in each row is highlighted.

Algorithm	**RBMLE**	**VIDS**	**KLUCB&UCB**	**GPUCB**	**GPUCBT**	**TS**	**BUCB**	**KG**	**KG***
Mean Regret	**531.1**	638.5	1102.7	2464.2	607.7	684.3	923.6	1995.0	760.2
Std. Dev.	469.5	1117.0	196.9	210.8	234.1	250.1	178.7	3541.8	**163.8**
Quantile .10	145.5	143.7	859.7	2200.1	361.4	411.1	724.5	**21.1**	568.4
Quantile .25	167.4	206.6	937.4	2320.7	444.3	501.7	792.9	**30.2**	664.5
Quantile .50	**207.7**	314.1	1093.2	2466.4	544.8	623.1	927.2	1014.4	752.5
Quantile .75	1131.8	889.0	1232.0	2605.0	**714.6**	792.2	1042.0	1044.3	851.4
Quantile .90	1188.1	1183.3	1346.8	2726.0	**926.2**	1058.9	1174.1	8121.5	930.0
Quantile .95	1204.2	1248.6	1439.0	2804.9	1041.8	1209.2	1193.5	9023.5	**959.5**

Table 7: Statistics of the final cumulative regret in Figure 4(b). The best in each row is highlighted.

Algorithm	**RBMLE**	**VIDS**	**KLUCB&UCB**	**GPUCB**	**GPUCBT**	**TS**	**BUCB**	**KG**	**KG***
Mean Regret	**652.0**	694.7	1302.0	2281.0	856.5	903.4	1149.5	1233.6	1001.7
Std. Dev.	581.8	776.1	**164.5**	169.5	255.8	268.2	201.0	1659.2	234.8
Quantile .10	127.3	193.6	1100.0	2062.5	561.1	574.8	897.0	**24.5**	747.2
Quantile .25	155.7	322.9	1173.4	2156.6	665.7	715.8	1000.4	**72.0**	827.9
Quantile .50	**265.4**	471.9	1295.7	2262.7	814.3	849.3	1130.5	1021.1	944.4
Quantile .75	1116.2	**861.0**	1428.3	2397.7	1007.8	1085.6	1294.0	1987.0	1128.1
Quantile .90	1202.8	1236.1	1492.8	2506.3	**1164.6**	1283.0	1404.6	2028.1	1346.7
Quantile .95	2021.8	1467.5	1549.4	2545.1	**1334.9**	1394.5	1511.5	2055.5	1467.2

Table 8: Statistics of the final cumulative regret in Figure 1(c). The best in each row is highlighted.

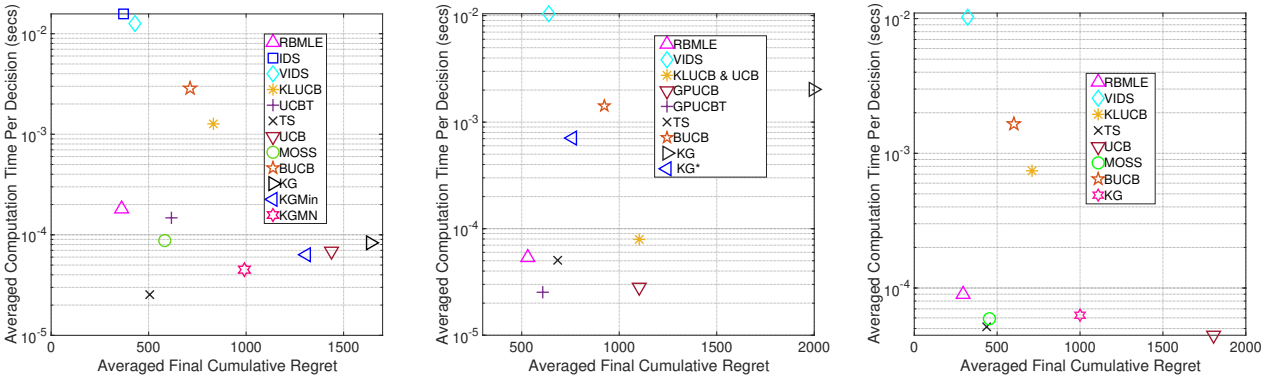
Algorithm	**RBMLE**	**VIDS**	**KLUCB**	**TS**	**UCB**	**MOSS**	**BUCB**	**KG**
Mean Regret	**179.6**	243.3	322.7	208.6	1504.6	379.9	288.2	961.6
Std. Dev.	119.4	463.1	63.9	61.3	66.1	**44.5**	71.9	1063.3
Quantile .10	128.7	37.6	239.4	132.8	1430.9	329.4	196.7	**26.5**
Quantile .25	139.7	47.9	271.3	157.7	1452.0	345.8	238.3	**37.2**
Quantile .50	155.2	**70.5**	331.7	202.3	1505.4	380.1	275.1	387.2
Quantile .75	173.4	**103.7**	367.2	243.4	1550.6	405.9	330.6	2450.7
Quantile .90	**195.4**	1039.9	407.0	303.1	1586.5	435.0	377.3	2509.9
Quantile .95	**291.7**	1074.1	423.2	320.1	1617.6	457.8	405.3	2522.7

Table 9: Statistics of the final cumulative regret in Figure 3(c). The best in each row is highlighted.

Algorithm	**RBMLE**	**VIDS**	**KLUCB**	**TS**	**UCB**	**MOSS**	**BUCB**	**KG**
Mean Regret	**294.6**	322.4	710.6	436.7	1805.6	453.5	600.8	1000.0
Std. Dev.	301.3	352.5	**118.0**	168.7	126.6	147.8	126.3	1637.9
Quantile .10	139.8	93.3	565.4	288.1	1653.3	342.8	464.1	**34.9**
Quantile .25	148.7	116.4	609.8	335.9	1713.9	374.8	792.9	**30.2**
Quantile .50	176.9	166.1	695.1	411.0	1789.0	419.6	592.2	**77.4**
Quantile .75	**237.4**	273.8	784.9	468.3	1898.1	483.9	662.3	1050.0
Quantile .90	919.0	1064.9	875.6	610.0	1970.0	**578.0**	739.0	4920.6
Quantile .95	1183.3	1112.1	916.6	682.5	2035.3	**644.9**	789.5	5042.0

Table 10: Statistics of the final cumulative regret in Figure 4(c). The best in each row is highlighted.

Algorithm	**RBMLE**	**VIDS**	**KLUCB**	**TS**	**UCB**	**MOSS**	**BUCB**	**KG**
Mean Regret	**195.2**	215.9	339.1	221.3	1815.8	462.0	298.8	1460.3
Std. Dev.	140.2	425.2	53.6	60.3	69.2	53.3	**45.9**	2035.8
Quantile .10	140.9	43.5	264.9	159.6	1729.3	402.8	247.8	**26.7**
Quantile .25	153.1	55.9	301.4	176.9	1776.9	428.6	263.5	**33.7**
Quantile .50	166.2	70.5	335.7	211.3	1818.0	456.7	297.7	**58.3**
Quantile .75	188.0	**94.1**	373.1	248.6	1863.7	480.2	326.5	3249.7
Quantile .90	**225.8**	1037.1	408.4	296.9	1897.8	532.3	365.8	4955.2
Quantile .95	**291.7**	1064.6	433.2	319.3	1934.1	563.1	383.3	4966.9

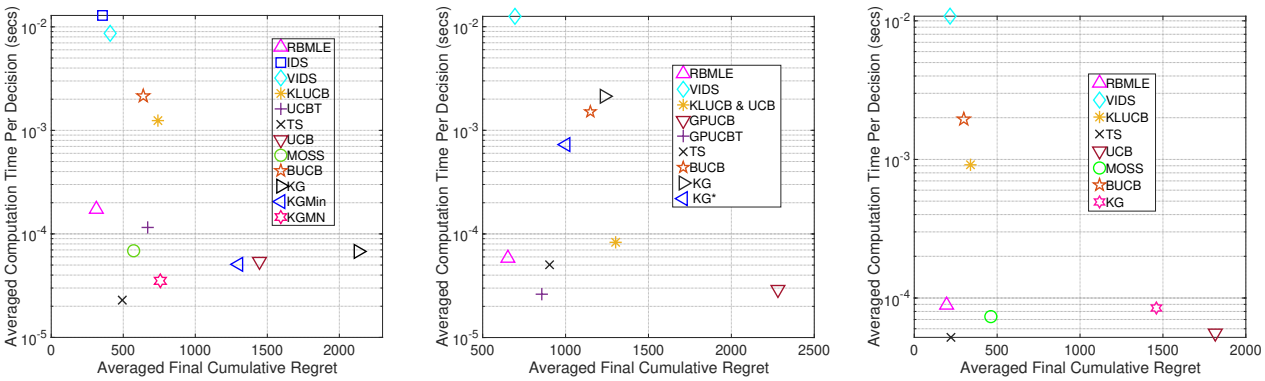


(a) (b) (c)

Figure 5: Averaged computation time per decision vs. averaged final cumulative regret: (a) Figure 3(a); (b) Figure 3(b); (a) Figure 3(c).

Table 11: Average computation time per decision for Bernoulli bandits, under different numbers of arms. All numbers are averaged over 100 trials with $T = 10^4$ and in 10^{-4} seconds. The best in each row is highlighted.

# Arms (Statistics)	**RBMLE**	**IDS**	**VIDS**	**KLUCB**	**UCBT**	**TS**	**UCB**	**MOSS**	**BUCB**	**KG**	**KGMin**	**KGMN**
10 (Mean)	1.36	175	123	12.8	1.53	0.225	0.712	0.895	0.855	28.7	0.649	0.453
30 (Mean)	3.61	1260	788	49.7	4.96	0.628	2.19	2.83	2.58	97.6	1.89	1.36
50 (Mean)	4.58	3630	1930	80.3	7.85	0.628	3.42	4.40	4.11	159	2.95	2.14
70 (Mean)	7.56	6660	3590	113	10.3	0.628	4.49	5.87	5.43	209	3.97	2.86
10 (Std. Err.)	0.236	54.8	33.1	1.53	0.586	0.0380	0.268	0.333	0.351	10.9	0.284	0.172
30 (Std. Err.)	1.30	458	232	17.3	1.52	0.106	0.646	0.844	0.714	29.2	0.557	0.408
50 (Std. Err.)	2.04	972	536	29.4	2.59	0.106	1.11	1.40	1.25	49.5	0.931	0.678
70 (Std. Err.)	2.70	1330	883	36.6	3.63	0.106	1.53	2.00	1.76	69.3	1.34	0.962



(a) (b) (c)

Figure 6: Averaged computation time per decision vs. averaged final cumulative regret: (a) Figure 4(a); (b) Figure 4(b); (a) Figure 4(c).

Table 12: Average computation time per decision for Gaussian bandits, under different numbers of arms. All numbers are averaged over 100 trials with $T = 10^4$ and in 10^{-4} seconds. The best in each row is highlighted.

# Arms (Statistics)	**RBMLE**	**VIDS**	**KLUCB&UCB**	**GPUCB**	**GPUCBT**	**TS**	**BUCB**	**KG**	**KG***
10 (Mean)	0.617	135	0.341	0.346	0.318	0.451	17.9	25.1	10.9
30 (Mean)	1.07	1410	1.12	1.10	1.08	1.33	75.2	103	21.2
50 (Mean)	1.49	3580	1.22	1.79	1.76	2.44	121	168	33.9
70 (Mean)	1.95	6610	1.67	2.24	2.22	3.16	162	226	45.9
10 (Std. Err.)	0.284	53.9	0.417	0.136	0.160	0.0425	6.98	9.37	2.77
30 (Std. Err.)	0.484	409	1.28	0.370	0.370	0.321	26.2	35	5.61
50 (Std. Err.)	0.686	866	2.14	0.563	0.563	0.562	42.1	56.1	9.77
70 (Std. Err.)	0.871	1290	2.95	0.755	0.773	0.774	58.5	77.6	15.7

Table 13: Average computation time per decision for Exponential bandits, under different numbers of arms. All numbers are averaged over 100 trials with $T = 10^4$ and in 10^{-4} seconds. The best in each row is highlighted.

# Arms (Statistics)	**RBMLE**	**VIDS**	**KLUCB**	**TS**	**UCB**	**MOSS**	**BUCB**	**KG**
10 (Mean)	1.01	133	7.26	1.38	0.420	0.548	14.9	0.519
30 (Mean)	1.93	1160	22.8	3.97	1.20	1.61	42.6	1.36
50 (Mean)	2.97	3170	36.5	6.64	1.92	2.53	75.5	2.23
70 (Mean)	3.79	6430	53.7	9.30	2.67	3.59	102	3.06
10 (Std. Err.)	0.435	13.6	0.884	0.316	0.0980	0.112	1.55	0.101
30 (Std. Err.)	0.890	187	2.79	0.777	0.263	0.340	5.02	0.265
50 (Std. Err.)	1.24	447	5.47	1.20	0.397	0.498	10.2	0.456
70 (Std. Err.)	1.56	788	6.92	1.96	0.531	0.688	12.3	0.605