MRAS Bandits: Control Engineering meets Slot Machines

Gayatri Ryali
cs16b106

Varun Sundar
ee16b068

Indian Institute of Technology Madras

1 Introduction

Model Reference Adaptive Search (MRAS) is a well-known algorithm in control engineering that allows for handling of Markov Decision Processes (MDPs) with either uncountable actions spaces or uncountable state spaces. To be accurate however, the MRAS algorithm is designed in the deterministic optimiziation context of a function $J(x)$ over variable x on space X.

The crux of the MRAS algorithm is to parametrise the sampling distribution in finding a set of candidate solutions. Each step, the parameter is updated so as to approach closer to a reference distribution. By doing so, the hope is that if the parameter updating rule is chosen appropriately, the future sampling process will be more and more concentrated on regions containing high-quality solutions. Convergence guarantees rely on the cornerstone that the reference distribution converges with its level set as the singleton optimal value. Hence, by showing that our parametric distribution closely approximates the reference distribution, we obtain the optimal value x^*. Reference [2] generalises this to situations of noisy observations of $J(x)$. We utilise this model based optimisation approach in formulating a novel bandit algorithm and studying its convergence properties.

We propose a MRAS inspired algorithm in section 3, and derive update rules for the case of Categorical, Dirichlet and Multivariate Gaussians. In section 4, we study the effect of hyperparameters including exploration rate λ, positive increasing function $H(x)$, simulation allocation M_k, sample sizes n_k as well as components of the algorithm such as elite sampling and exploration decay. In section 5, we compare our best performing categorical and dirichlet MRAS algorithms against Thomson Sampling (TS) [3], UCB-asymptotic [5] and UCB [1] (with tune-able confidence intervals). Finally, we conclude this report with a possible direction for attempting regret bounds.

2 Model Reference Adaptive Sampling (MRAS)

We outline the brief ideas of MRAS variants proposed in [2]- MRAS 0, 1 and 2. In all three cases, our objective is to find the global maximiser for deterministic $J(x)$, over a set X. We assume J has a unique global maximiser, and has compact neighbourhoods around that maximiser x^*. The key difference in the three algorithm arises from their utility in handling cases where either expectations cannot be computed as closed form integrals (one either does not have access to the underlying probability distribution, or if so, it may be intractable in nature), or J can only be inferred from noisy updates of $\tilde{J}$.

In figure 1, we show the steps of MRAS-0, as stated in [2]. Notice that the update for θ can easily be shown to correspond to θ which minimises $D(g_k, f(.,\theta)$. Here, we choose a simple reference for $g_k(x)$ as:

$$g_k(x) = \frac{J(x)g_{k-1}(x)}{\int_X J(x)g_{k-1}(x)dx}$$

Which clearly has increasing expectation (wrt to g_k) as k increases. Finally, to accommodate the case where $J(x)$ is not necessarily non-negative throughout, we replace it by $H(J(x))$, where $H : R \to R^+$ is a non-decreasing positive function.

The cornerstone result in the convergence proof of MRAS-0 is to note that:

$$\lim_{k\to\infty} m(\theta_k) \coloneqq \lim_{k\to\infty} E_{\theta_k}[Y(X)] = \lim_{k\to\infty} E_{\theta_k}[Y(X)] = Y(x^*) \tag{1}$$

where, $Y(x)$ is a sufficient statistic of the distribution $f(.;\theta)$. For the distributions we shall encounter, ie..., categorical, dirichlet and gaussian; we utilise the fact that all three belong to the exponential family. In particular, for the categorical case, we have that,

$$Y(x) = \sum_{i=1}^{K} p_i I(x = i)\hat{i} \quad \text{where } I \text{ is the discrete indicator function.} \tag{2}$$

Thus, $Y(x^*) = p_{a^*}\hat{a^*}$, ie... it identifies the optimal arm ia degenerate fashion.

The alogrithms MRAS-1,2 differ only slightly from MRAS-0, replacing expectations by summations, and varying both quantile measure ρ_k and threshold χ_k. The idea is to closely approximate MRAS-0, and this implies that the "gaps" incurred due to approximating expectations with summations vanishes w.p.1 for large k. In addition, MRAS-2 requires assumptions on concentration of $\tilde{J}$ with respect to J. This is satisfied for the bandits setting, and is clearly seen by application of either Chebyshev or Chernoff bounds. These assumptions allow for equivalent relations such as equation 1, albeit replacing θ_k with the approximated $\tilde{\theta}_k$. We shall make use of this in the penultimate section on regret bounds.

3 Proposed Bandit Algorithm

In Algorithm 1, we outline the proposed bandit setting and methodology. Given as inputs to the algorithm are the family of distributions $f(.;\theta)$, positive non-decreasing function H, initial population N_0 and simulation allocation M_k.

Observe that we need to arrive at tractable solutions to the maximisation step (or with close approximations). For instance, the case of the *Dirchlet* distribution entails derivatives of $log(\Gamma(x))$ which can only be approximately evaluated for non-integeral values of x. In the sub-sections that follow, we consider particular values of $f(.;\theta)$ and choices for λ and M_k.

3.1 The Categorical Case

In this case, $\theta = [p_1, p_2, ..., p_K]$ such that $\sum_{i=1}^{K} p_i = 1$. Here, $f(x = i, \theta) = p_i$. Or $f(x = a_j(k), \theta) = p_{a_j(k)}$

Algorithm MRAS$_0$

Input: $\rho \in (0, 1]$, $\varepsilon \geq 0$, strictly increasing function $\mathcal{H} : \Re \to \Re^+$, family of distributions $\{f(\cdot, \theta)\}$, with θ_0 s.t. $f(\mathbf{x}, \theta_0) > 0 \; \forall \mathbf{x} \in \mathcal{X}$.

Initialization: Set iteration count $k = 0$.

Loop until Stopping Rule is satisfied:

- Calculate the $(1 - \rho)$-quantile:

$$\chi_k = \sup_l \{l : P_{\theta_k}\big(J(\mathbf{X}) \geq l\big) \geq \rho\}.$$

- Update elite threshold:

$$\bar{\chi}_k = \begin{cases} \chi_k & \text{if } k = 0 \text{ or } \chi_k \geq \bar{\chi}_{k-1} + \varepsilon, \\ \bar{\chi}_{k-1} & \text{otherwise.} \end{cases}$$

- Update parameter vector:

$$\theta_{k+1} \in \arg\max_{\theta \in \Theta} E_{\theta_k}\left[\frac{[\mathcal{H}(J(\mathbf{X}))]^k}{f(\mathbf{X}, \theta_k)} I\{J(\mathbf{X}) \geq \bar{\chi}_k\} \ln f(\mathbf{X}, \theta)\right]. \tag{4.4}$$

- $k \leftarrow k + 1$.

Output: θ_k.

Figure 1: MRAS-0 Algorithm

Denote: $G_k(j,t) = \frac{e^{Ck\hat{\mu}_{j,t}}}{f(a_j(k),\theta_k)}$ at time t, phase k, sample j.

We use Lagrange multipliers for real p_i, constrained by $\sum_i p_i = 1$. (One can also use the full form of KKT, for positive p_i, but it is not hard to show that it simplifies to this case). Denoting the Lagrange multiplier as λ, we get:

$$p_i = \lambda \sum_{j \in \Lambda_k^*} G_k(j,t) I(a_j(k) = i)$$

Note that in the case $\Lambda_k^* = i^*$, p simplifies to $p_{i^*} = 1, 0$ otherwise.

3.2 The Dirichlet Case

The update rule for θ, here represented as $\boldsymbol{\alpha}$, is:

$$\begin{aligned} \alpha^+ &= \text{argmax} \sum_{j \in \Lambda_k} G_k(j,t) \ln\left(\frac{\Pi_{i=1}^K x_{i,j}^{\alpha_i - 1}}{B(\boldsymbol{\alpha})}\right) \\ &= \text{argmax} \sum_{j \in \Lambda_k} G_k(j,t) \left[\sum_{i=1}^K (\alpha_i - 1) \ln x_{i,j} - \sum_{i=1}^K \ln \Gamma(\alpha_i) + \ln \Gamma\left(\sum_{i=1}^K \alpha_i\right)\right] \end{aligned}$$

Algorithm 1 MRAS Bandits

Require: simulation budget: n

parameterized distribution family : $f(,\theta)$

initial parameters: θ_0

initial population size: N_0

strictly increasing function: $H : R \to R+$

simulation allocation rule: $M_k, k = 0, 1, ...p$, p partitions.

Denote : $\hat{\mu}_{j,t}$ as jth sample mean, upto time t.

procedure (MRAS Bandits) Loop until: $\sum_{k=0}^{k=p} n_k M_k \geq N$.

Here n_k is the no of samples to draw in stage k and M_k the number of times you pull each arm at stage k.

Sample Arm Λ_k: $a_j(k) \sim (1-\lambda)f(\mathbf{x}, \theta_k) + \lambda f(\mathbf{x}, \theta_0)$ for $j = 1, ..., n_k$.

Update sample means $\hat{\mu}_{j,t}$: of arm $a_j(k)$, pulling each arm M_k times.

Ordered Statistic χ_k: Set $\hat{\chi}_k = J_{[(1-\rho)N]}$ and $\chi_k = max(\hat{\chi}_k, \chi_{k-1})$.

Elite Subset: Set $\Lambda_k^* = \{a \in \Lambda_k | \hat{J}(a) \geq \chi_k\}$.

Update parameter: $\theta_{k+1} = argmax_{\theta \in \Theta} \sum_{j \in \Lambda_k^*} \frac{H(\hat{\mu}_{j,n})^k}{f(a_j(k), \theta_k)} lnf(a_j(k), \theta)$.

Decay Exploration: if $\hat{\chi}_k > \chi_{k-1}$, Set $\lambda = \lambda * 0.9$.

Update time: $t = t + n_k * M_k$

where, boldface represents vectors.

Now, $\psi(x) = \frac{d \ln \Gamma(x)}{dx} = \frac{\Gamma'(x)}{\Gamma(x)}$. We have the known properties that, $\psi(x+1) = 1/x + \psi(x)$. Furthermore, if x is a positive integer,

$$\psi(n+1) = -\gamma + \sum_{k=1}^{n} \frac{1}{k}$$

Here, γ is Euler-Mascheroni constant, with a value of 0.57721. We approximate the summations by its series-integral lower bound, ie... $ln(x)$ .Maximising in an unconstrained fashion, we get:

$$\begin{aligned} 0 &= \sum_{j \in \Lambda_k} G_k(j,t)[\ln x_{i,j} - \psi(\alpha_i^+) + \psi(\sum_{i=1}^{K} \alpha_i^+)] \\ &= \sum_{j \in \Lambda_k} G_k(j,t)[\ln x_{i,j} + \sum_{k=\alpha_i^+}^{\sum_i \alpha_i^+} \frac{1}{k}] \end{aligned}$$

, where we additionally assume that α_i (and the updates α_i^+) are integers (Assumption A1). This simplifies to:

$$\sum_{k=\alpha_i^+}^{\sum_i \alpha_i^+} \frac{1}{k} \approx \ln(\frac{\sum_i \alpha_i^+}{\alpha_i^+}) = -\frac{\sum_{j \in \Lambda_k} G_k(j,t) \ln x_{i,j}}{\sum_{j \in \Lambda_k} G_k(j,t)}$$

, making the integral lower bound approximation (A2). This yields:

$$\alpha_i^+ = (\sum_i \alpha_i^+) e^{\frac{\sum_{j\in\Lambda_k} G_k(j,t) \ln x_{i,j}}{\sum_{j\in\Lambda_k} G_k(j,t)}}$$

Which is homogeneous in $\boldsymbol{\alpha}$, we make the assumption that $\sum_i \alpha_i^+ = 1000$ (assumption A3), or some arbitrary number. We note these assumptions hold fairly well in practice, especially for $\alpha_i \geq 5$. This holds for most α_i. If the reader is insistent however, one may drop the lower bound and homogeneity approximation by using an optimisation method such as Newton's method.

3.3 The Multivariate Gaussian Case

the update rule for θ , which in this case is mean θ and covariance matrix Σ is as follows:
Denote $G_j(k,t) = G_k(j,t) = \frac{exp(Ck\hat{f}_{j,t})}{f(a_j(k),\theta_k)}$

then by differentiating the required term with respect to θ and Σ, we get:

$$\theta_{k+1} = \frac{\sum_{j\in\lambda_k} G_k(j,t) a_j(k)}{\sum_{j\in\lambda_k} G_k(j,t)}$$
$$\Sigma_{k+1} = \frac{\sum_{j\in\lambda_k} G_k(j,t)(a_j(k) - \theta_{k+1})(a_j(k) - \theta_{k+1})^T}{\sum_{j\in\lambda_k} G_k(j,t)}$$

We observed that the regret of this algorithm varies widely, thus making it difficult to tune hyper-parameters to obtain acceptable regret values. (A fair bit of difficulty associated is the consequent numerical stability; we are required to have Σ_k as positive definite in every phase). Hence, we choose not to stduy this algorithm in the scope of ablative and comparative studies in this paper.

4 Effect of Hyper-parameters

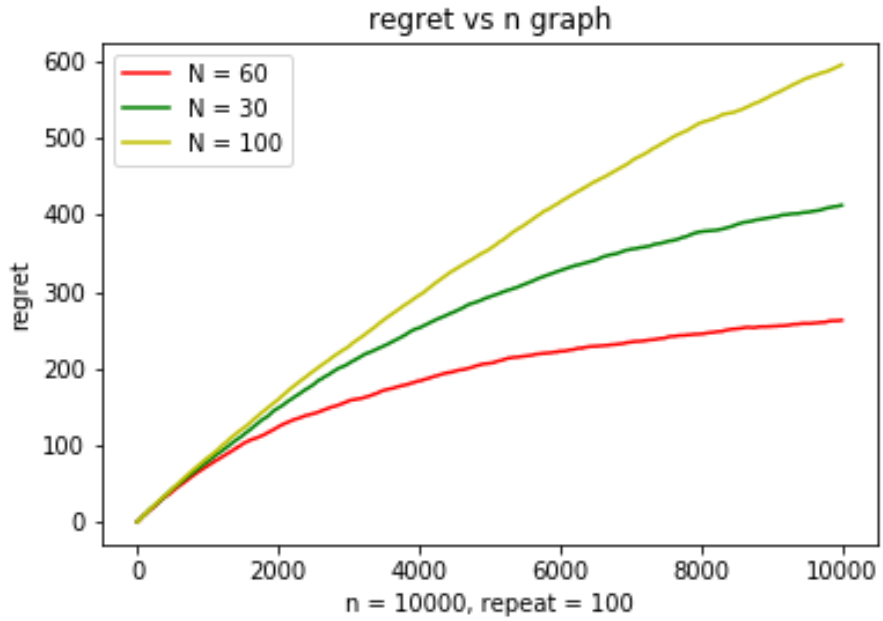
This section studies the effects of various hyper-parameters on regret,using the mentioned algorithms, on the following game:

$$game : [0.5, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4]$$

The values of n (simulation budget) and *repeat* (to estimate expected regret) are set to 10000 and 100 respectively.

4.1 Initial population: N_0

Figure 2 is a "regret vs n" plot, using MRAS-categorical algorithm , when varying N_0:
The algorithm performs better when $N_0 = 60$.



4.2 Positive Non-Decreasing function: H(x)

Figure 3 is a "regret vs n" plot,using the MRAS-dirichlet algorithm ,whilst varying H(x).
The functions used are logarithmic, square-root and exponential.

It is seen in the graph that when H(x) = log(x), the regret is fairly large, in fact it grows linearly. This is because the MRAS converges in cases, where H(x): R→R^+. As such convergence is not guaranteed for a non-positive monotone. And as from the graph, the regret is minimum when $H(x) = e^x$. We therefore use e^x for our remaining ablative and comparative studies.

4.3 Exploration-exploitation parameter: λ

Figure 4 is a "regret vs n" plot, using MRAS-dirichlet algorithm, while varying the value of λ. The parameter λ determines the fraction of exploration and exploitation done by the algorithm. The algorithm exploits by the factor of (1-λ) and explores by λ. As the number of iterations increase, the λ decreases (λ decays with every iteration), thus approaching zero. In that case, we no longer explore, but only exploit, which is the desired condition.

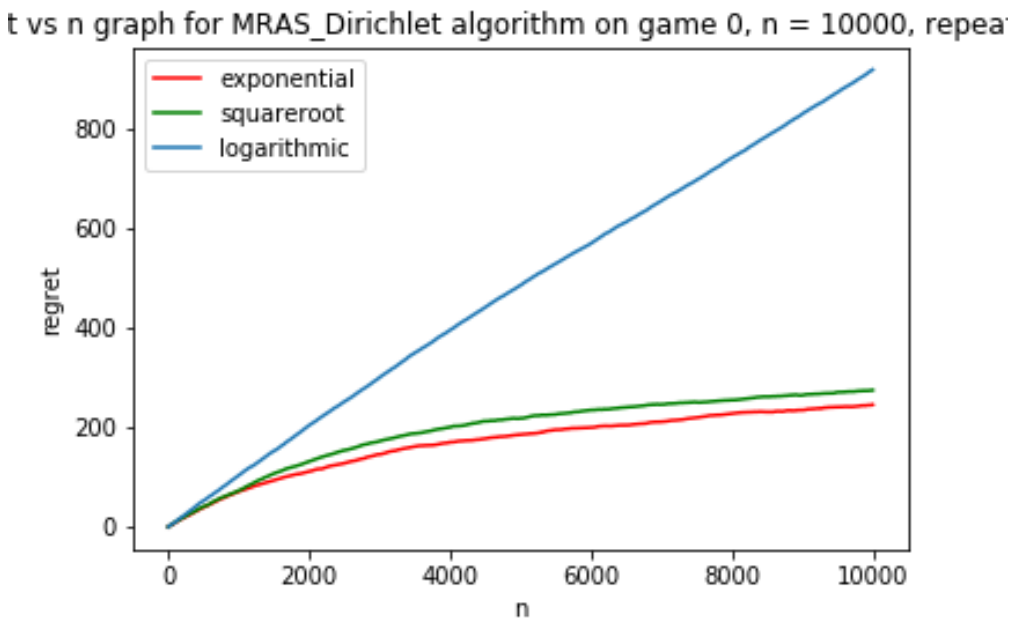


Figure 3: Regret vs n, while varying H(x), using MRAS-dirichlet

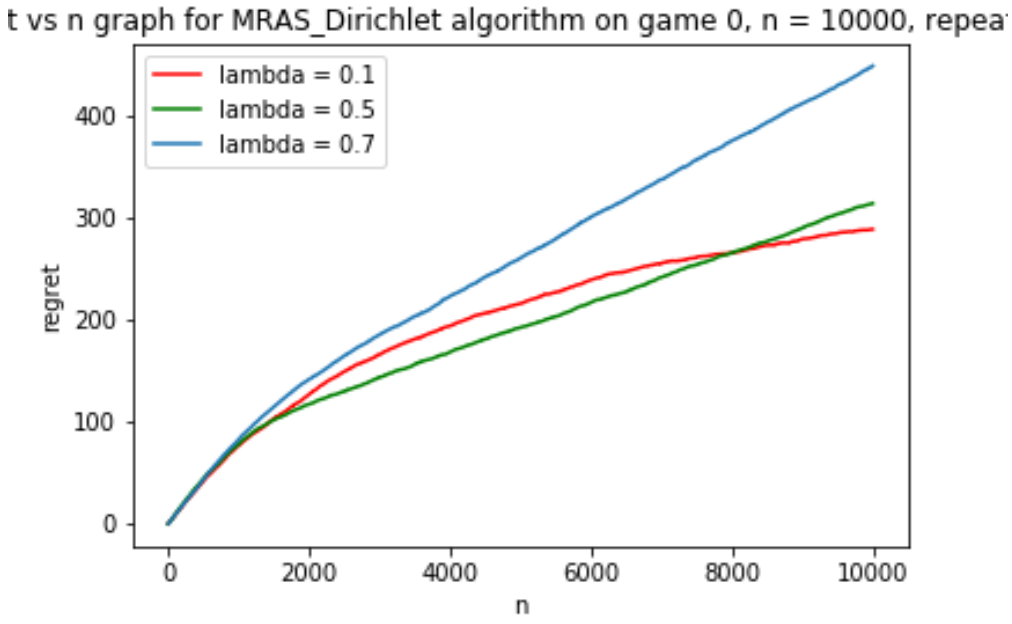


Figure 4: Regret vs n, while varying λ, using MRAS-dirichlet

A smaller λ would ensure more exploitation and ensures that aforementioned desired condition is obtained faster, thereby giving smaller regrets. As seen in the graph, the regret is the lowest when $\lambda = 0.3$. However, there is always the exploration-exploitation trade-off, and one cannot simply reduce λ to 0. For the rest of the experiments, we find $\lambda = 0.2$ to be optimal.

4.4 Simulation-allocation rule: $\mathbf{M}_k$

The following is "regret vs n graph" , using MRAS-dirichlet algorithm ,when varying the simulation-allocation rule.
Figure 5 contains the plots pertaining to the following cases:

- each arm is pulled equally during every phase, denoted by "equal".
- each arm is pulled in an increasing fashion (sequence $b, b+1, b+2, ...$) during every phase, denoted by "increasing".
- number of times each arm is pulled as according to the successive rejects phase rule, denoted by "logkbar".

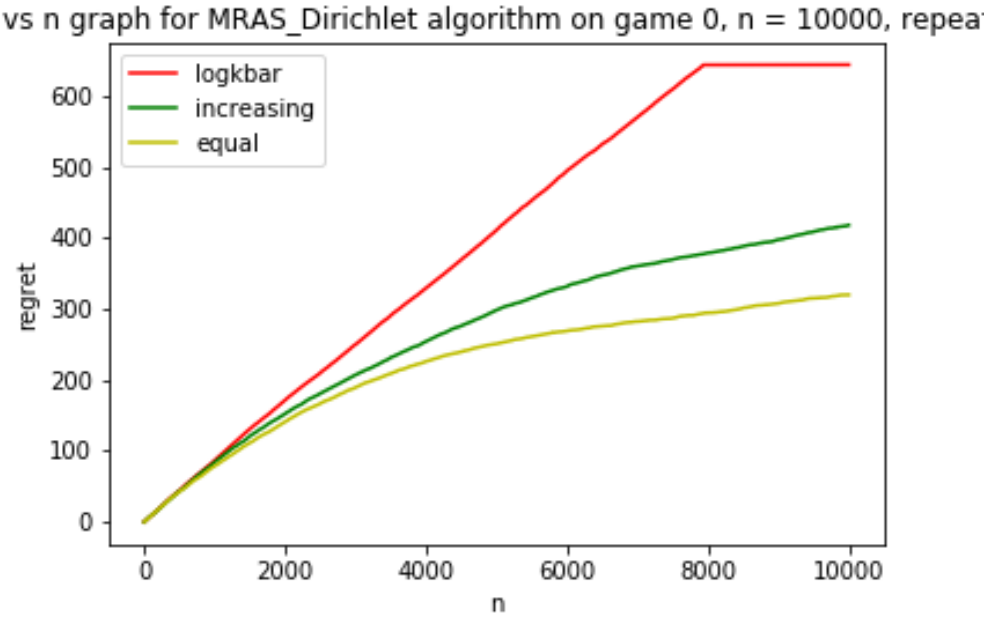


Figure 5: Regret vs n, while varying M_k, using MRAS-dirichlet

5 Comparative Studies

In this section,we compare tuned versions of MRAS categorical and dirichlet against Thomson Sampling (with two different beta priors), UCB (asymptotic), UCB and gap-dependent minimax regret with tune-able confidence widths on three different games.

- $game_1$:[0.5,0.4,0.4,0.4,0.4,0.4,0.4,0.4,0.4,0.4]
- $game_2$:[0.5,0.48,0.48,0.48,0.48,0.48,0.48,0.48,0.48,0.48]
- $game_3$:[0.5,0.2,0.1]

In all the cases we have fixed n = 10000, and repeats = 100.The initial population size N_0 is kept constant across all the games.

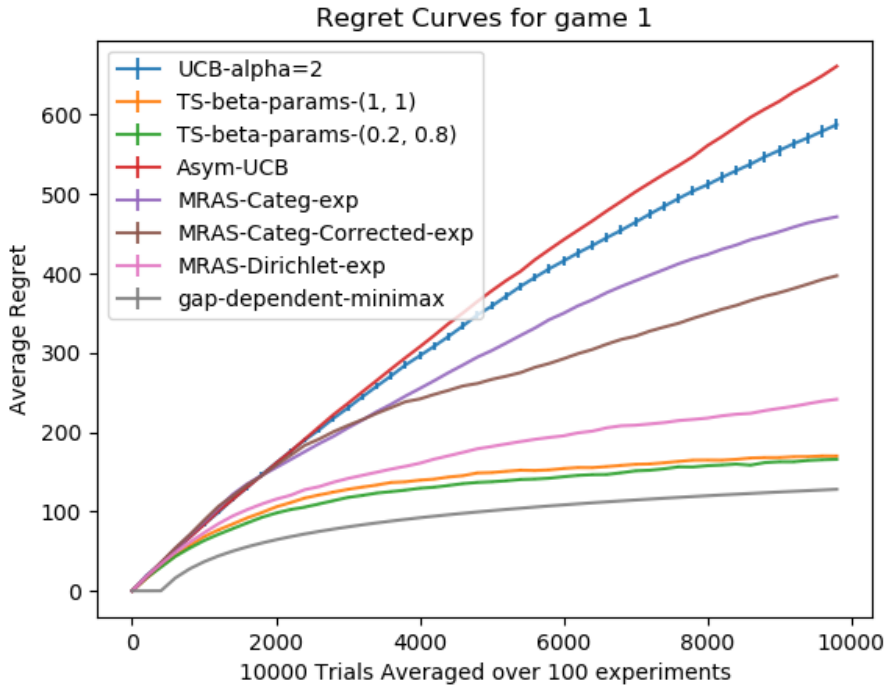


Figure 6: regret vs n, game1

As seen from the graphs, MRAS algorithm (both categorical, dirichlet) perform better than UCB, but worse than Thomson sampling.But in game2, MRAS-dirichlet performs better than the thomson sampling.

In case of game 3, it can be seen that the MRAS algorithms perform worse in comparison to both UCB and Thomson sampling. The most plausible reason for this is that the initial population size, N_0 has to be tuned as according to the number of arms in the game. But here we are using same value of N_0 for all the three games.Hence the regret obtained using all of the MRAS algorithms, is quite larger that the expected.

Amongst the three MRAS algorithms,we can empirically conclude that MRAS-dirichlet performs better than MRAS-categorical. In the categorical case, the categorical-corrected performs better than the simple categorical case. This can attributed to the elite subset sampling and the λ decay that are performed in the categorical-corrected case.
MRAS-dirichlet performs better than the categorical case, due to its similarity with the thomson sampling algorithm. This stems from the fact that the marginal of the dirichlet distribution follows beta distribution, and hence updates to θ_k act as a proxy for posterior belief.

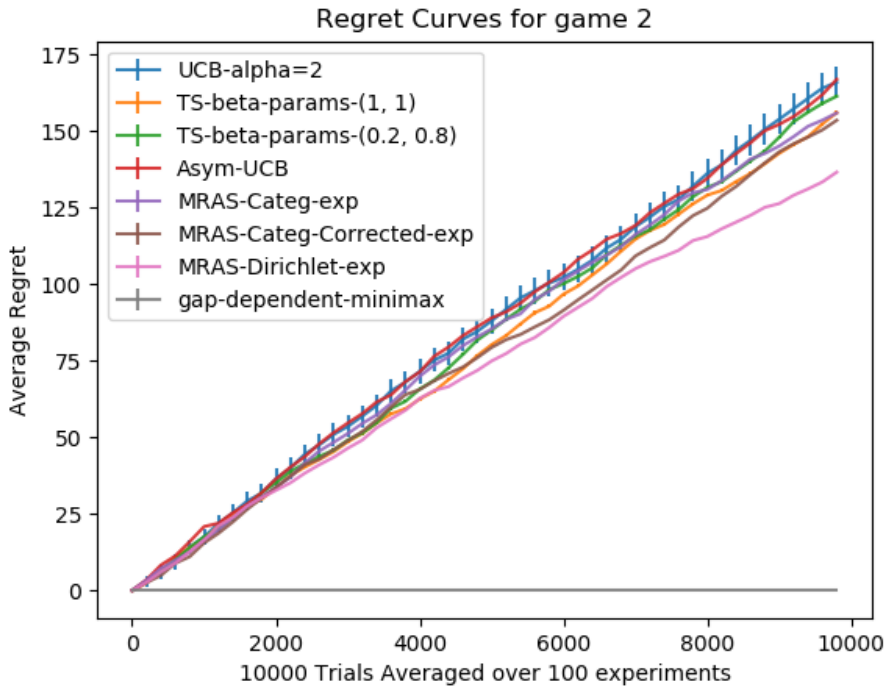


Figure 7: regret vs n,game2

6 Regret Bounds: Possible Approaches

We discuss regret bound strategies for the simplest case - MRAS Categorical. In the kth phase, probability of choosing a sub-optimal arm is:

$$(1-\lambda)(1-p_{1,k}) + \lambda\frac{K-1}{K}$$

, assuming arm 1 is optimal. Notice that the right hand term is usually troublesome, as it leads to linear (in n) regret. However, in MRAS bandits, we have non-decreasing allocation rule M_k and non-decreasing sample size n_k. Further, we decay λ with each phase and thus, the constant term in the expected regret vanishes for large k.

Regret incurred in phase k is: $\sum_{i=0}^{n_k} M_k\Delta_i$. Assuming equal gaps for all sub-optimal arm, overall regret simplifies to:

$$\sum_k M_k n_k \Delta[(1-\lambda_k)(1-p_{1,k}) + \lambda_k\frac{K-1}{K}]$$

Thus,it suffices to bound $p_{1,k}$ for finite k, assuming we are annealing λ_k with phase. The variant of equation (1) as proved in [2] is applicable to the setting of MRAS-2, providing asymptotic bounds for $p_{1,k}$. Equation (2) shows that $p1,k$ is the value of the sufficient statistic $Y(x)$ at $x = x^*$. A good starting point here would be to draw equivalent, finite time bounds for theorems 4.15 and 4.17 to bound the same.

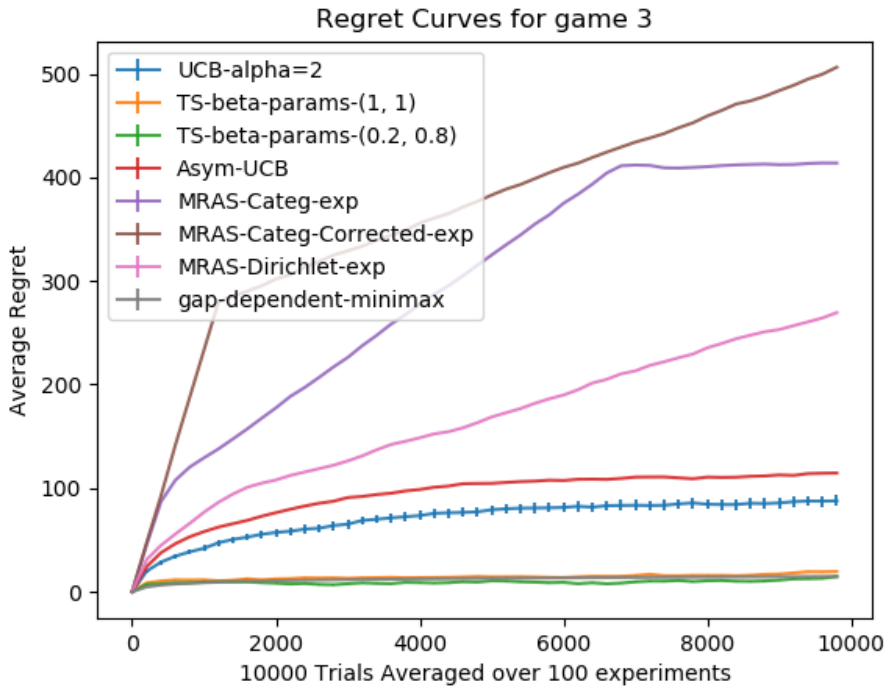


Figure 8: regret vs n,game3

Another alternate approach could be to consider the similarity between MRAS-Dirichlet and Thomson Sampling. Casting the update to θ_k as a form of $p(\theta|y)$ could be a starting point. This can be extended to the entire Exponential Family by considering concentration results and KL divergence expressions used by [4].

7 Conclusions and Future Work

We present a general framework for MRAS-Bandits, inspired by the MRAS algorithm as proposed in [2]. We study in detail two particular cases of MRAS Bandits - Categorical and Dirichlet; exploring the effect of hyper-parameters and compare them to other state-of-the-art bandit algorithms such as Thomson Sampling and UCB. We note the similarity of MRAS-Dirichlet and Thomson Sampling with beta priors. Finally, we conclude our work with possible approaches to arrive at regret bounds for the proposed algorithms.

Potential future work could begin with regret bounds for the two proposed algorithms. If we are able to determine the theoretical sensitivity and variability of θ_k with respect to its hyper-parameters, we can propose a variant of this algorithm for the Best Arm Identification (BAI) scenario. Considering that MRAS theorems hold for parametrised distributions in the Lebesgue measure as well, an extension to linear or structured bandits should also be feasible.

References

[1] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, March 2003. ISSN 1532-4435. URL

http://dl.acm.org/citation.cfm?id=944919.944941.

[2] Hyeong Soo Chang, Jiaqiao Hu, Michael C. Fu, and Steven I. Marcus. *Simulation-Based Algorithms for Markov Decision Processes*. Springer Publishing Company, Incorporated, 2nd edition, 2013. ISBN 144715021X, 9781447150213.

[3] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2249–2257. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4321-an-empirical-evaluation-of-thompson-sampling.pdf.

[4] Nathaniel Korda, Emilie Kaufmann, and Remi Munos. Thompson sampling for 1-dimensional exponential family bandits. *Advances in Neural Information Processing Systems*, 07 2013.

[5] Pierre Mnard and Aurlien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In Steve Hanneke and Lev Reyzin, editors, *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pages 223–237, Kyoto University, Kyoto, Japan, 15–17 Oct 2017. PMLR. URL