MRAS Bandits: Control Engineering meets Slot Machines

Gayatri Ryali and Varun Sundar

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 [1] generalises this to situations of noisy observations of $J(x)$. We hope to utilise this model based optimisation approach in formulating a novel bandit algorithm and studying its convergence properties.

2 Objective

To obtain finite time error bounds on the MRAS algorithm (Model reference adaptive search) and utilise the ideas in creating a bandit algorithm.

3 Proposed Steps

We make all references to theorems and lemmas with respect to reference [1].

1. Prove finite time bounds theorem 4.5 of MRAS-2.
2. Explore (1) using simulations to validate the regularity conditions assumed and understand convergence related issues when these regularity conditions donot hold.
3. Formulate a bandit (stochastic bandit) setting using MRAS algorithm.
4. Set up a simulation setting using this algorithm.
5. Provide bounds for regret minimisation, and study the effect of hyperparamters (including discount λ , arms distribution $f(., \theta)$, increasing function $H(x)$ and simulation allocation rule M_k).

6. Explore possibility of MRAS being extended to structured bandit problems.

4 Proposed Algorithm

Algorithm 1 MRAS Bandits

Require: simulation budget n

parameterized distribution family $f(,\theta)$

initial θ_0 strictly increasing function $H : R \to R+$

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

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

procedure M(R)AS Bandit. Loop until: $\sum_{k=0}^{k=p} n_k M_k \geq N$. Here n_k is the no of distinct arms to draw in stage k and M_k the number of times you pull each arm at stage k.

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

    Update sample means of arm $a_j(k)$, pulling each arm M_k times.

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

We shall analyse a particular case of this algorithm where $H(x) = e^x$, $f(.;\theta)$ is the categorical distribution, and all $M_k = 1$. This simplifies to:

Algorithm 2 Simplified MRAS Bandits

Require: simulation budget n

Number of arms K

initial $\theta_0 = 1/K$

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

procedure M(R)AS Bandit. Loop until: N.

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

    Update sample means of arm $a_j(k)$, pulling each arm once.

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

    $t = t + n_k$

However, note 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.

References

[1] 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.