Whittle Index Policies Become Optimal Exponentially Fast*

Nicolas Gast[1], Bruno Gaujal[1], Chen Yan[1]

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
{nicolas.gast,bruno.gaujal,chen.yan}@inria.fr

Mots-clés : *Multi-armed Bandits, Whittle Index, Asymptotic Optimality.*

Résumé : *Restless bandits are stochastic optimization problems that are known to be PSPACE-hard [6]. In [8], Whittle provides a computationally efficient heuristic, known as Whittle index policy (WIP). WIP is known to be asymptotically optimal as the number of bandits goes to infinity, and performs very well for many practical tasks (such as scheduling [4] or resource allocation). In this talk, we will review this result and answer a simple question: how fast does this heuristic become optimal?*

1 Restless Bandits and Whittle Indexes

Consider the following restless multi-armed Markovian bandit model in discrete time: A population N of bandits each evolves in an identical finite state space $\{1 \dots d\}$. At each decision epoch we choose to activate a proportion $0 < \alpha < 1$ of bandits (assume αN is an integer for simplicity). An active (action 1) bandit in state i gives an immediate reward R_i^1 and changes its state following $d \times d$ transition matrix $\mathbf{P}^1$; likewise a passive (action 0) bandit gives reward R_i^0 and has transition matrix $\mathbf{P}^0$ (in the literature, the term *restless* means that $\mathbf{P}^0$ is not necessarily the identity matrix. When $\mathbf{P}^0 = I_d$ the problem is easily solved by the Gittins index policy [3]). Our goal is to compute a decision rule that maximizes the infinite-horizon time average reward.

The above optimization problem has been proven to be PSPACE-hard in [6]. To overcome this difficulty, Whittle introduces in [8] the concept of indexability and the Whittle index, where the index value ν_i can be computed efficiently (algorithm with time complexity $\mathcal{O}(d^3)$ for computing the indices can be found in [5]) for each state i. The heuristic Whittle index policy (WIP) is then to activate at each decision epoch the αN bandits having currently the highest indices. There exist numerous applications of WIP in practical problems that perform surprisingly well. In [7] it is proven that under indexability as well as a further assumption that the ODE corresponding to the drift of the Markov system under WIP has a globally stable fixed point, then WIP is asymptotically optimal as $N \to \infty$. Formally, denote by $V_{\text{WIP}}^{(N)}(\alpha)$ the value of WIP and by $V_*^{(N)}(\alpha)$ the value of the optimal policy, [7] shows that $\lim_{N\to\infty} |V_*^{(N)}(\alpha) - V_{\text{WIP}}^{(N)}(\alpha)| = 0$ under the aforementioned conditions.

2 Exponential convergence rate

Our work goes a step further from the previous results: we show that the asymptotic convergence almost always occurs at exponential rate. More precisely, we claim that

Theorem 1 *Assume that bandits are indexable and unichain. Moreover, assume that the drift $\phi(\cdot)$ of the Markov system under WIP has a uniform global attractor $\mathbf{m}^*$ that is not singular. Then there exist constants $b, c > 0$ and a value $V_{\text{rel}}^{(1)}(\alpha)$ independent of N such that*

$$V_{\text{rel}}^{(1)}(\alpha) - b \cdot e^{-cN} \leq V_{\text{WIP}}^{(N)}(\alpha) \leq V_*^{(N)}(\alpha) \leq V_{\text{rel}}^{(1)}(\alpha). \tag{1}$$

*This talk is based on [1].

The word *singular* is explained as follows: we can show that the drift function $\phi(\cdot)$ is continuous and piecewise-affine with d affine pieces inside the simplex of probability measures Δ^d. A point in Δ^d is then called *singular* if it is on the boundary of two affine pieces.

Here is the idea for the proof of our result: The piecewise-affine property comes as a mixed blessing. On the one hand the dynamics is not differentiable on the boundary of two affine pieces, so previous mean field approach based on the smoothness of the drift such as [2] collapse here; on the other hand when $\mathbf{m}^*$ is non-singular, the dynamics in a small neighbourhood $\mathcal{N}$ of $\mathbf{m}^*$ is affine and the expected behavior of the system is relatively simple to analyze. We then divide the analysis of the stochastic system into two parts: before it enters $\mathcal{N}$ and after it does. The stein's method is used to compare its behavior with its mean field approximation inside $\mathcal{N}$; Hoeffding's inequality is used to control its behavior outside $\mathcal{N}$.

Our proof actually gives more. Although we do not aim at finding the optimal constant c in (1), we believe that the further $\mathbf{m}^*$ is away from being singular, the faster is the convergence rate (i.e., the larger is the constant c). We illustrate this in Figure 1. Notably, when $\mathbf{m}^*$ is singular, the rate can be as slow as $\Theta(\frac{1}{\sqrt{N}})$, in big contrast with the exponential rate.

For this numerical example, the fixed point $\mathbf{m}^*$ is a global attractor that is singular when α=0.4. In the range $0.3<\alpha<0.5$, $\mathbf{m}^*$ (that depends on α) is further away from any boundaries when α is further away from 0.4. We plot the quantity $V_{\text{WIP}}^{(N)}(\alpha)/V_{\text{rel}}^{(1)}(\alpha)$ for various N (note that $V_{\text{WIP}}^{(N)}(\alpha)/V_{*}^{(N)}(\alpha)$ is closer to 1 than $V_{\text{WIP}}^{(N)}(\alpha)/V_{\text{rel}}^{(1)}(\alpha)$ because $V_{\text{rel}}^{(1)}(\alpha) > V_{*}^{(N)}(\alpha)$). WIP is closer to optimal when α moves further away from the singular value 0.4.

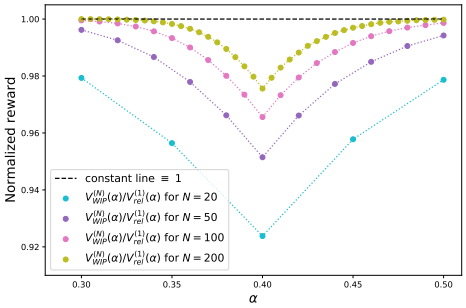

FIG. 1: Lower bound on $V_{\text{WIP}}^{(N)}(\alpha)/V_{*}^{(N)}(\alpha)$ as a function of α for $N \in \{20, 50, 100, 200\}$. When α is far from 0.4 (the singular value), WIP is very close to being optimal ($V_{\text{WIP}}^{(N)}(\alpha)/V_{*}^{(N)}(\alpha) \approx 1$).

References

[1] Nicolas Gast, Bruno Gaujal, and Chen Yan. Exponential convergence rate for the asymptotic optimality of whittle index policy, 2020.

[2] Nicolas Gast and Benny Van Houdt. A Refined Mean Field Approximation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* , 1(28), 2017.

[3] J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, pages 148–177, 1979.

[4] Bo Ji, Gagan R Gupta, Manu Sharma, Xiaojun Lin, and Ness B Shroff. Achieving optimal throughput and near-optimal asymptotic delay performance in multichannel wireless networks with low complexity: a practical greedy scheduling policy. *IEEE/ACM Transactions on Networking*, 23(3):880–893, 2014.

[5] José Niño-Mora. A fast-pivoting algorithm for whittle's restless bandit index. *Mathematics*, 8(12), 2020.

[6] Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of optimal queuing network control. *Math. Oper. Res*, pages 293–305, 1999.

[7] Richard R. Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.

[8] P. Whittle. Restless bandits: activity allocation in a changing world. *Journal of Applied Probability*, 25A:287–298, 1988.