Exploration vs. Exploitation with Partially Observable AR(1) Arms

Julia Kuhn

Supervisors: Yoni Nazarathy (UQ) & Michel Mandjes (Universiteit van Amsterdam)

The University Of Queensland
AUSTRALIA

I. Model and Framework

A dynamic decision problem under uncertainty: We select k out of d restless *reward observing* one-armed bandits to play on, such as to maximize the expected total discounted or average reward. Rewards are collected and states are observed ONLY if an arm is played.

Should we collect new information or opt for the immediate payoff?

State processes are Gaussian AR(1),

$$X_i(t) = \varphi\, X_i(t-1) + \varepsilon_i(t),$$

where $\varphi \in (0,1)$ and $\varepsilon_i \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$. An application is channel selection in wireless networks.

Why is the Gaussian model so special?

- The *belief states* $\big(\mu_i(t),\, \nu_i(t)\big)$, i.e. the means and variances conditioned on the available information, contain all relevant information available at time t.
- At the same time, $\mu_i(t)$ and $\nu_i(t)$ quantify the expected gain from exploiting an arm vs. the need for exploring it.

Updating the Belief States

A *policy* π maps the information available to actions $a_i(t) = 1$ ("play") or $a_i = 0$ ("do not play"), such that in total k out of d are played at every time t. With $Y_{\mu,\nu} \sim \mathcal{N}(\mu, \nu)$,

$$\big(\mu_i(t+1),\, \nu_i(t+1)\big) = \begin{cases} \big(\varphi\, \mu_i(t),\ \varphi^2\, \nu_i(t) + \sigma^2\big), & a_i(t) = 0, \\ \big(\varphi\, Y_{\mu_i(t),\, \nu_i(t)}\,,\ \sigma^2\big), & a_i(t) = 1. \end{cases}$$

Chain of Actions

$$\begin{tikzcd} (\boldsymbol{\mu}, \boldsymbol{\nu}) \arrow[rr, "\pi"] & & \boldsymbol{a} \arrow[dl, "\text{observe state, collect reward}"] \\ & \sum_{i=1}^{d} X_i\, a_i \arrow[ul, "\text{update belief}"] & \end{tikzcd}$$

References

1. K. Avrachenkov, L. Cottatellucci, and L. Maggi (2012). Slow Fading Channel Selection: A Restless Multi-armed Bandit Formulation. *ISWCS*, pp. 1083–1087.
2. J. Kuhn, M. Mandjes and Y. Nazarathy (2014). Exploration vs. Exploitation with Partially Observable Gaussian Autoregressive Arms. *Submitted.*
3. J. Gittins, K. Glazebrook and R. Weber (2011). *Multi-armed Bandit Allocation Indices*, 2nd Ed., John Wiley & Sons.
4. P. Whittle (1988). Restless bandits: Activity Allocation in a Changing World. *Journal of Applied Probability*, 25, pp. 287–298.

II. Index Policies

An index policy is of the form

$$\pi_\gamma(\boldsymbol{\mu}, \boldsymbol{\nu}) = \underset{\boldsymbol{a}: \sum_{i=1}^{d} a_i = k}{\arg\max} \left\{ \sum_{i=1}^{d} \gamma\,(\mu_i, \nu_i)\, a_i \right\}$$

The *index function* γ maps the belief state of each arm to some priority index.

Myopic	$\gamma^M(\mu,\nu) = \mu$
Parametric	$\gamma^\theta(\mu,\nu) = \mu + \theta\nu, \quad \theta > 0$
Whittle	$\gamma^W(\mu,\nu) = \inf\{\lambda \mid \pi_{\text{opt}}(\mu,\nu) = 0\}$

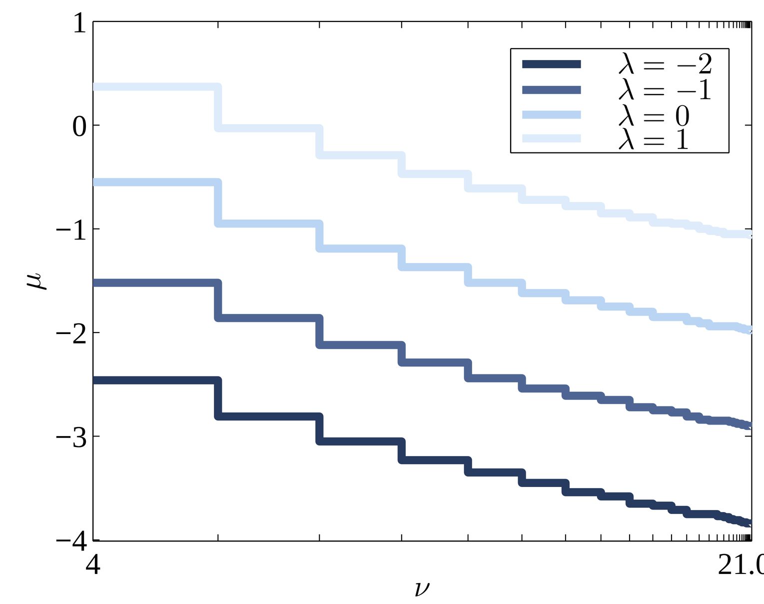
Here π_{opt} is the optimal policy for a *one-armed bandit problem with subsidy*, where the decision maker observes and collects the reward when playing, and obtains a subsidy λ otherwise.

Motivated by the many-arms asymptotic behaviour of the system.

III. Whittle Index: Structural Results

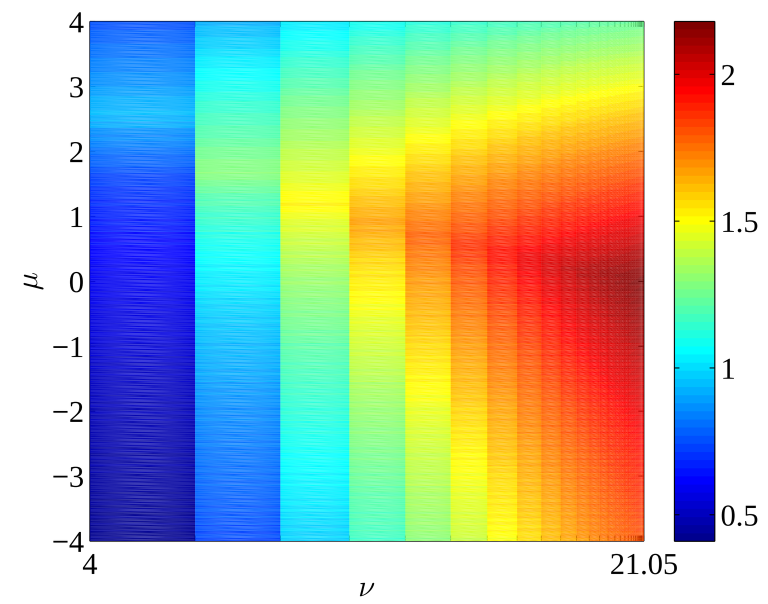
The Whittle index policy has been found to be asymptotically optimal in many cases (although no such result is known for our model) but no closed-form expression is known. The associated optimal value function can in principle be found using dynamic programming techniques. We can further prove the following.

The optimal policy for the one-armed bandit problem with subsidy is a *threshold policy*.



Switching curves: above the curve the optimal action is "play", below "do not play".
$\beta = 0.8,\ \varphi = 0.9,\ \sigma = 2.$

The Whittle index $\gamma^W(\mu,\nu)$ is monotone non-decreasing in μ and ν, and generally not constant.



Difference of Whittle and myopic index:
$\gamma^W(\mu,\nu) - \mu.$
$\beta = 0.8,\ \varphi = 0.9,\ \sigma = 2.$

IV. Parametric Index: Many-Arms Asymptotic Behaviour

1. Intuitively, as $d \to \infty$, $k_d/d \to \rho$, in the long-run the system approaches an equilibrium at which the proportion of arms associated with a certain belief state remains fixed. Then the action chosen for a certain arm is independent of the current belief state of any other arm, as there is always the same proportion of arms associated with a certain belief state in the system.

2. We explicitly identify a measure-valued recursion that describes the many-arms behaviour of the system at equilibrium. Namely, the limiting proportion of arms that have been observed h time steps ago and whose conditional mean falls in $(-\infty, x]$ can be modeled as

$$m_h\big(x,\, t+1\big) = \begin{cases} \sum_{h=0}^{\infty} \int_{\ell_h^*(t)}^{\infty} \Phi_{z,\, \nu^{(h)}}\left(\frac{x}{\varphi}\right) m_h(dz, t), & h = 0, \\ m_{h-1}\Big(\min\left\{\frac{x}{\varphi},\, \ell_{h-1}^*(t)\right\},\, t\Big), & h \geq 1, \end{cases}$$

where $\ell_h^*(t) := \ell^*(t) - \theta\nu^{(h)}(t)$ with $\ell^*(t)$ defined by

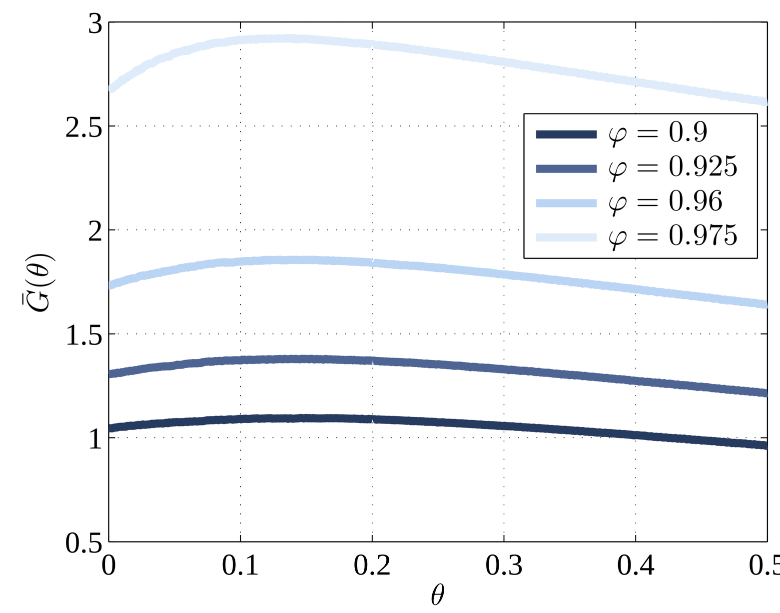
$$\ell^*(t) = \sup\left\{ \ell \,\middle|\, \sum_{h=0}^{\infty} m_h\Big(\{\mu \mid \mu + \theta\nu^{(h)} \in [\ell, \infty)\},\, t\Big) = \rho \right\}.$$

Thus, $\ell_h^*(t)$ is a threshold such that at time t the parametric policy activates all arms that are of age h and have conditional mean $\mu(t) \geq \ell_h^*(t)$.

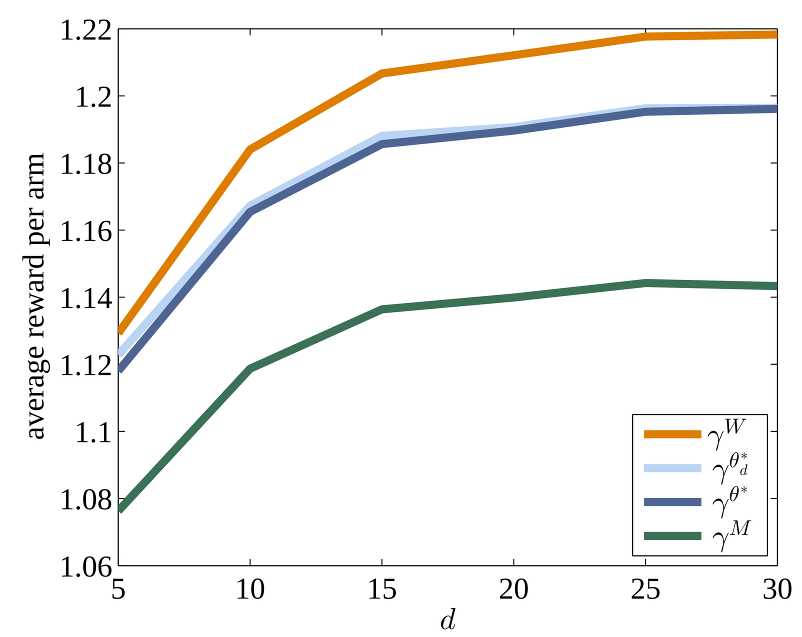
3. Based on **1.** and **2.** we conjecture that the measure-valued dynamical system at equilibrium is directly related to a one-armed process where the arm is activated whenever the index exceeds a particular threshold $\bar{\ell}$, namely $\bar{\ell} = \ell^*$.

Algorithm for Performance Evaluation

1. For large T determine $\bar{\ell}$ such that $T^{-1} \sum_{t=0}^{T} a_i(t) = \rho$ is achieved for a parametric index policy applied to the one-armed process.
2. Use the sample path of Step 1 to obtain an estimate $\overline{G}$ for the expected average reward of the one-armed system.
3. Output $\overline{G}_d := d\, \overline{G}$ as an approximation of the expected average reward of the multiarmed system with d arms.



Expected average reward $\overline{G}(\theta)$ computed by the algorithm as a function of θ. $\sigma = 2$, $\varphi \in \{0.9, 0.925, 0.95, 0.975\}$, $\rho = 0.4$, $T = 2 \times 10^6$.



Comparison of average rewards achieved per arm. θ is found by optimizing (i) the problem with d arms (θ_d^*), and (ii) the one-armed problem (θ^*). $\varphi = 0.9$, $\sigma = 2$, $\rho = 0.4$, $T = 10^5$.

LaTeX TikZposter