Exploration vs Exploitation with Partially Observable Gaussian Autoregressive Arms

Julia Kuhn[•,⋆,∘], Michel Mandjes[⋆], Yoni Nazarathy[•,∘]

[•] *The University of Queensland*, [⋆] *University of Amsterdam*

[∘] Supported by the Australian Research Council grant DP130100156.

11 December 2014

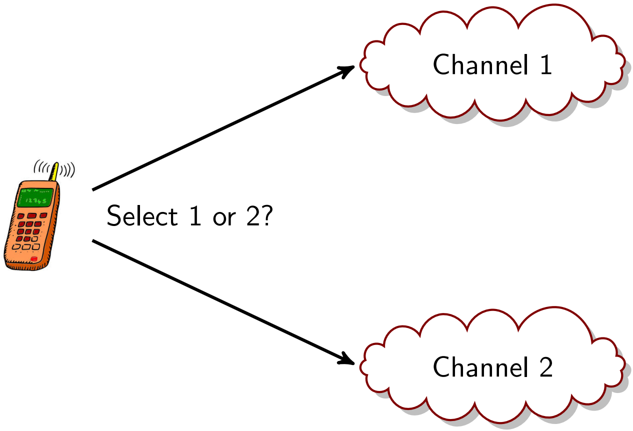
What is a bandit problem?

Classical Multi-armed Bandit Problem

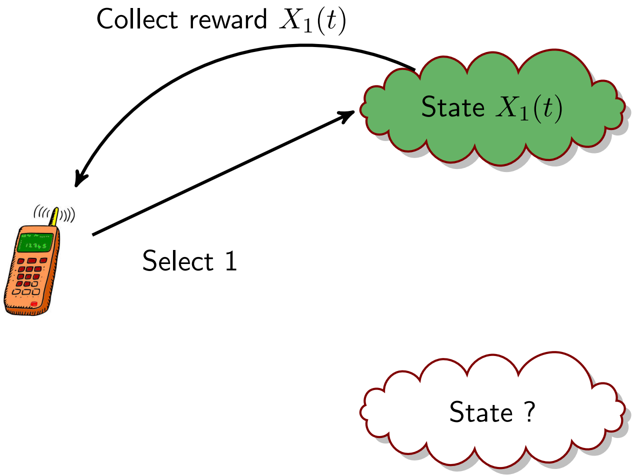
Classical Multi-armed Bandit Problem

Pick k out of d independent arms at every decision time.
States are *resting* unless the arm is played.
An optimal policy is known (Gittins index).

Channel Selection Problem



Channel Selection Problem



Restless Bandit Problems

P. Whittle (1988):
Restless Bandits: Activity Allocation in a Changing World.

Partially Observable

Exploration vs Exploitation:

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

States and Belief States

State processes are assumed to be AR(1),

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

where $\varphi \in (0,1)$, and $\varepsilon_i(t) \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$.

States and Belief States

State processes are assumed to be AR(1),

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

where $\varphi \in (0,1)$, and $\varepsilon_i(t) \sim_{\text{i.i.d.}} \mathcal{N}(0,\sigma^2)$.

Belief state of arm i at time t:

$$\mu_i(t) := \mathbb{E}\Big[X_i(t) \,\Big|\, X_i\big(t-\eta_i(t)\big), \eta_i(t)\Big] = \varphi^{\eta_i(t)} X_i\big(t-\eta_i(t)\big),$$

$$\nu_i(t) := \operatorname{Var}\Big(X_i(t) \,\Big|\, X_i\big(t-\eta_i(t)\big), \eta_i(t)\Big) = \sigma^2 \frac{1-\varphi^{2\eta_i(t)}}{1-\varphi^2},$$

where $\eta_i(t)$ is the number of time steps since arm i was last played.

Why is the Gaussian model special?

Why is the Gaussian model special?

- The belief states $\big(\mu_i(t), \nu_i(t)\big)$ contain all relevant information available at time t.

Why is the Gaussian model special?

- The belief states $\big(\mu_i(t), \nu_i(t)\big)$ contain all relevant information available at time t.

- $\mu_i(t)$: expected gain from exploiting an arm,
  $\nu_i(t)$: the need for exploring it.

Belief State Evolution

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

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

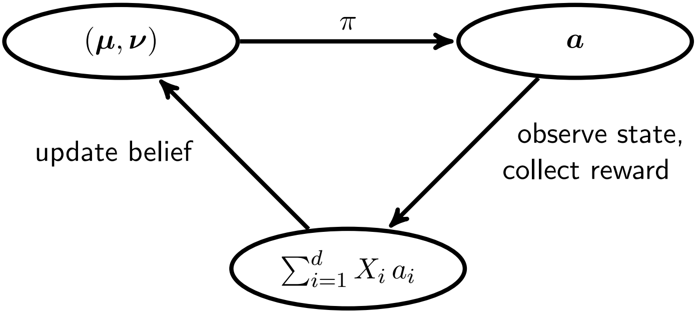
Belief State Evolution

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

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

$\Rightarrow$ **Markov Decision Process**

Chain of Actions



Index Policies

An index policy is of the form

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

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

Index Policies

An index policy is of the form

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

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

Example: Myopic Policy with $\gamma^M(\mu, \nu) = \mu$.

Definition

$$\boxed{\gamma^W(\mu,\nu) \;=\; \inf\left\{\lambda \,\middle|\, \pi^{\lambda}_{\mathsf{opt}}(\mu,\nu) = 0\right\}}$$

Here $\pi^{\lambda}_{\mathsf{opt}}$ is the optimal policy for a

one-armed bandit problem with subsidy,

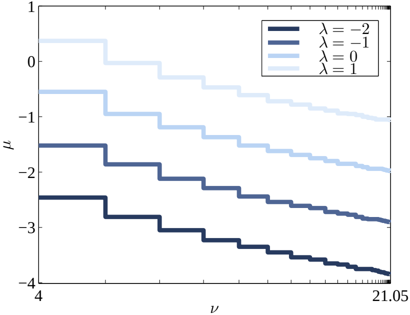
where the decision maker observes and collects the reward when playing, and obtains a subsidy λ otherwise.

Optimality Equation

$$
\begin{aligned}
V^{\lambda}(\mu,\nu) = \max\Big\{ & \lambda + \beta\, V^{\lambda}(\varphi\,\mu,\ \varphi^2\nu + \sigma^2)\,, \\
& \mu + \beta \int_{-\infty}^{\infty} V^{\lambda}\left(\varphi\, y,\ \sigma^2\right) \phi_{\mu,\nu}(y)\, \mathrm{d}y \Big\}
\end{aligned}
$$

Threshold Policy

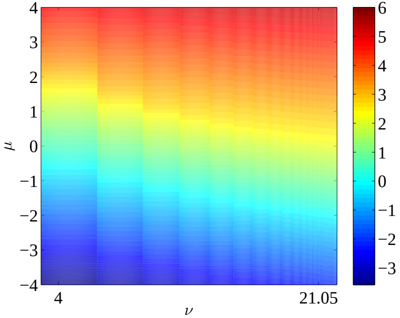
The optimal policy for the one-armed bandit problem with subsidy is a *threshold policy*.



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

Monotonicity of the Whittle Index

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



Whittle indices.
$\beta = 0.8$, $\varphi = 0.9$, $\sigma = 2$.

Parametric Index

$$\boxed{\gamma(\mu,\nu) = \mu + \theta\nu, \quad \text{where } \theta > 0.}$$

The correction term $\theta\nu$ allows to adjust the priority the decision maker wants to give to exploration.

Asymptotic Behaviour: Intuition

- Consider the system under stationarity. Let $d \to \infty$ while $k_d/d \to \rho$.

Asymptotic Behaviour: Intuition

- Consider the system under stationarity. Let $d \to \infty$ while $k_d/d \to \rho$.
- Note that the stochastic processes of indices are generally dependent.

Asymptotic Behaviour: Intuition

- Consider the system under stationarity. Let $d \to \infty$ while $k_d/d \to \rho$.
- Note that the stochastic processes of indices are generally dependent.
- As we add more arms to the system, it approaches an equilibrium state in which the proportion of arms associated with a certain belief state remains **fixed**.

Asymptotic Behaviour: Intuition

- Consider the system under stationarity. Let $d \to \infty$ while $k_d/d \to \rho$.
- Note that the stochastic processes of indices are generally dependent.
- As we add more arms to the system, it approaches an equilibrium state in which the proportion of arms associated with a certain belief state remains **fixed**.
- Thus, in the limit, the action chosen for a certain arm is independent of the current belief state of any other arm.

Conjecture: Many-Arms Behaviour

Assume that empirical distribution $M_h^d(x,0)$ converges weakly to non-random measure $m_h(B,0)$ for all $h \geq 0$,

$$M_h^d(B,0) \xrightarrow{w} m_h(B,0),$$

as $d \to \infty$ while $\lim_{d\to\infty} k_d/d = \rho$. Then, for all $t, h \geq 0$,

$$M_h^d(B,t) \xrightarrow{w} m_h(B,t).$$

State of System

$f_h(x,t):$ Mass of arms played $h+1$ time units ago with conditional mean in $[x, dx)$

State of System

$f_h(x,t)$: Mass of arms played $h+1$ time units ago with conditional mean in $[x, dx)$

State of the system at time t is described by

$$\{f_h(x,t),\ x \in \mathbb{R},\ h = 0, 1, 2, \ldots\},$$

where

$$\int_{-\infty}^{\infty} \sum_{h=0}^{\infty} f_h(x,t)\, dx = 1.$$

Threshold Process

$\ell_h^*(t) := \ell^*(t) - \theta\nu^{(h)}(t)$ such that

$$\int_{\ell_h^*(t)}^{\infty} \sum_{h=0}^{\infty} f_h(x,t)\, dx = \rho$$

defines the proportion of ρ "best" arms as determined by the parametric policy.

Many-Arms Asymptotic Behaviour

$$
f_h(x,t) = \begin{cases} \frac{1}{\varphi} f_{h-1}\left(\frac{x}{\varphi},\, t-1\right) \mathbb{1}\{x \le \varphi\, \ell_h^*(t-1)\}, & h \ge 1, \\ \frac{1}{\varphi} \sum_{h=0}^{\infty} \int_{\ell_h^*(t-1)}^{\infty} \phi_{z,\nu_h}\left(\frac{x}{\varphi}\right) f_h(z, t-1)\, dz, & h = 0. \end{cases}
$$

Many-Arms Asymptotic Behaviour

$$f_h(x,t) = \begin{cases} \frac{1}{\varphi} f_{h-1}\left(\frac{x}{\varphi}, t-1\right) \mathbb{1}\{x \le \varphi\, \ell_h^*(t-1)\}, & h \ge 1, \\ \frac{1}{\varphi} \sum_{h=0}^{\infty} \int_{\ell_h^*(t-1)}^{\infty} \phi_{z,\nu_h}\left(\frac{x}{\varphi}\right) f_h(z, t-1)\, dz, & h = 0. \end{cases}$$

Motivated by evolution of belief states:

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

Many-Arms Asymptotic Behaviour

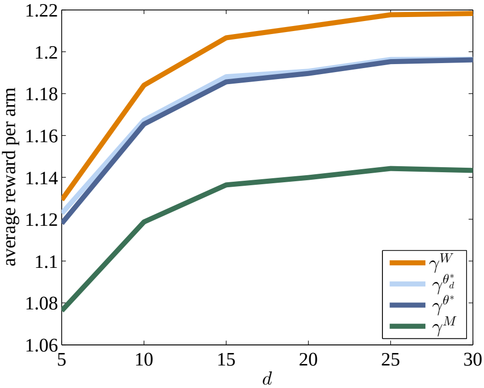
$$f_h(x,t) = \begin{cases} \frac{1}{\varphi} f_{h-1}\left(\frac{x}{\varphi},\, t-1\right) \mathbb{1}\{x \le \varphi\, \ell_h^*(t-1)\}, & h \ge 1, \\ \frac{1}{\varphi} \sum_{h=0}^{\infty} \int_{\ell_h^*(t-1)}^{\infty} \phi_{z,\nu_h}\left(\frac{x}{\varphi}\right) f_h(z,t-1)\, dz, & h = 0. \end{cases}$$

Motivated by evolution of belief states:

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

Conjecture: Equilibrium State

measure-valued dynamical system at equilibrium	$\sim$	one-armed process: active whenever index exceeds ℓ^*



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^*). $\varphi = 0.9$, $\sigma = 2$, $\rho = 0.4$, $T = 10^5$.

Some 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. Gittins, K. Glazebrook and R. Weber (2011). *Multi-armed Bandit Allocation Indices*, 2nd Ed., John Wiley & Sons.
3. K. Liu and Q. Zhao (2010). Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access. *IEEE Trans. Inf. Theory*, 56, pp. 5547–5567
4. R. Weber and G. Weiss (1990). On an Index Policy for Restless Bandits. *J. Appl. Probab.*, 27, pp. 37–648.
5. P. Whittle (1988). Restless Bandits: Activity Allocation in a Changing World. *J. Appl. Probab.*, 25, pp. 287–298.

Thank you!