Exploration vs. Exploitation with Partially Observable Gaussian Autoregressive Arms

Julia Kuhn,
The University of Queensland, University of Amsterdam
Yoni Nazarathy, *UQ*
Michel Mandjes, *UvA*

13 October 2014

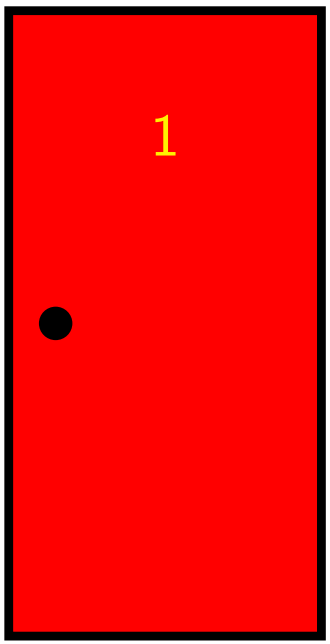
What is a bandit problem?

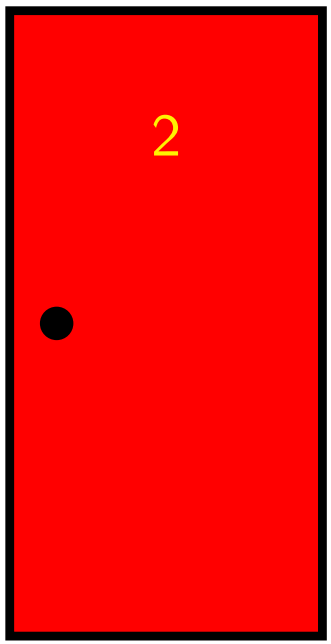
Multiarmed Bandit Problem

Multiarmed Bandit Problem

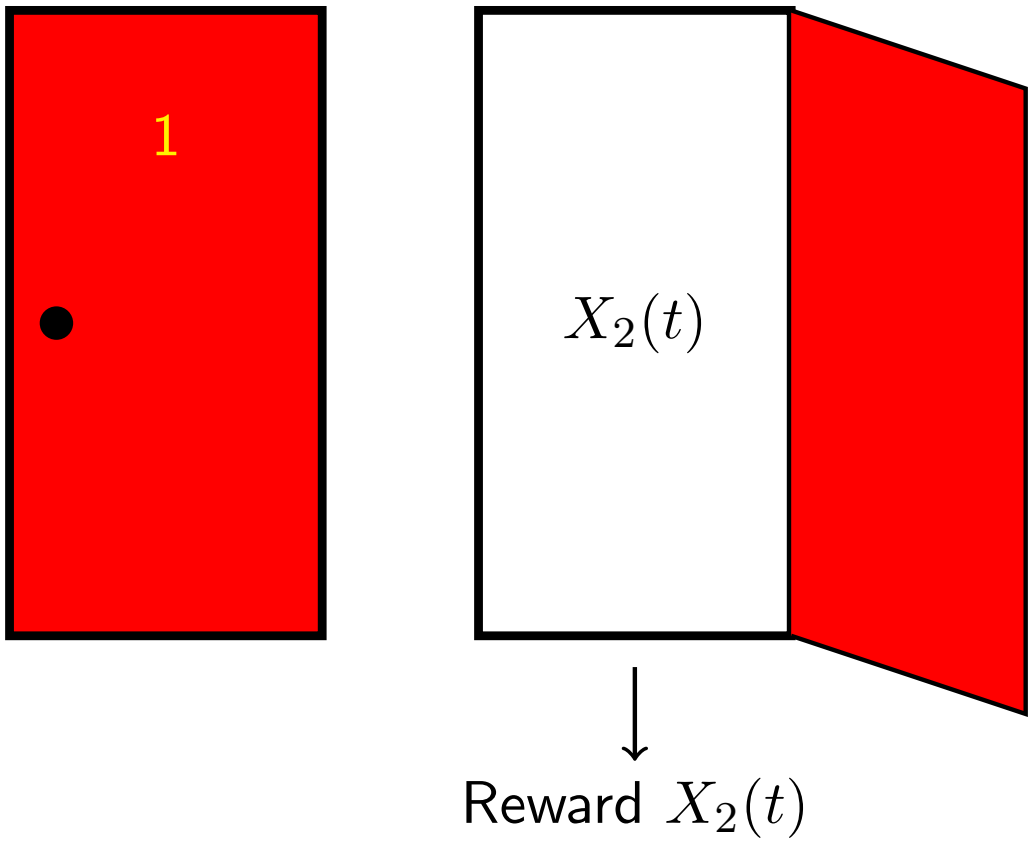
Pick k out of d arms at every decision time. States are *resting* unless the arm is played. These are the classical bandit problems. An optimal policy is the Gittins index.

Decision Problem





Decision Problem



Restless Bandit Problems

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

Partially Observable

Partially Observable

Exploration vs. Exploitation:

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

Partially Observable Restless Bandit Problems

K. Liu and Q. Zhao (2010):
Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access.

States vs. Belief States

State processes are assumed to be Gaussian autoregressions of order 1 ($\text{AR}(1)$),

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

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

States vs. Belief States

State processes are assumed to be Gaussian autoregressions of order 1 ($\text{AR}(1)$),

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

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

Belief state of arm i at time t:

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

$$\nu_i(t) := \text{Var}\left(X_i(t) \,\middle|\, X_i\left(t-\eta_i(t)\right), \eta_i(t)\right) = \sigma^2 \frac{1-\varphi^{2\eta_i(t)}}{1-\varphi^2},$$

where $\eta_i(t) := \min\{h \geq 1 | a_i(t-h) = 1\}$.

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.
- 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.

Belief State Evolution

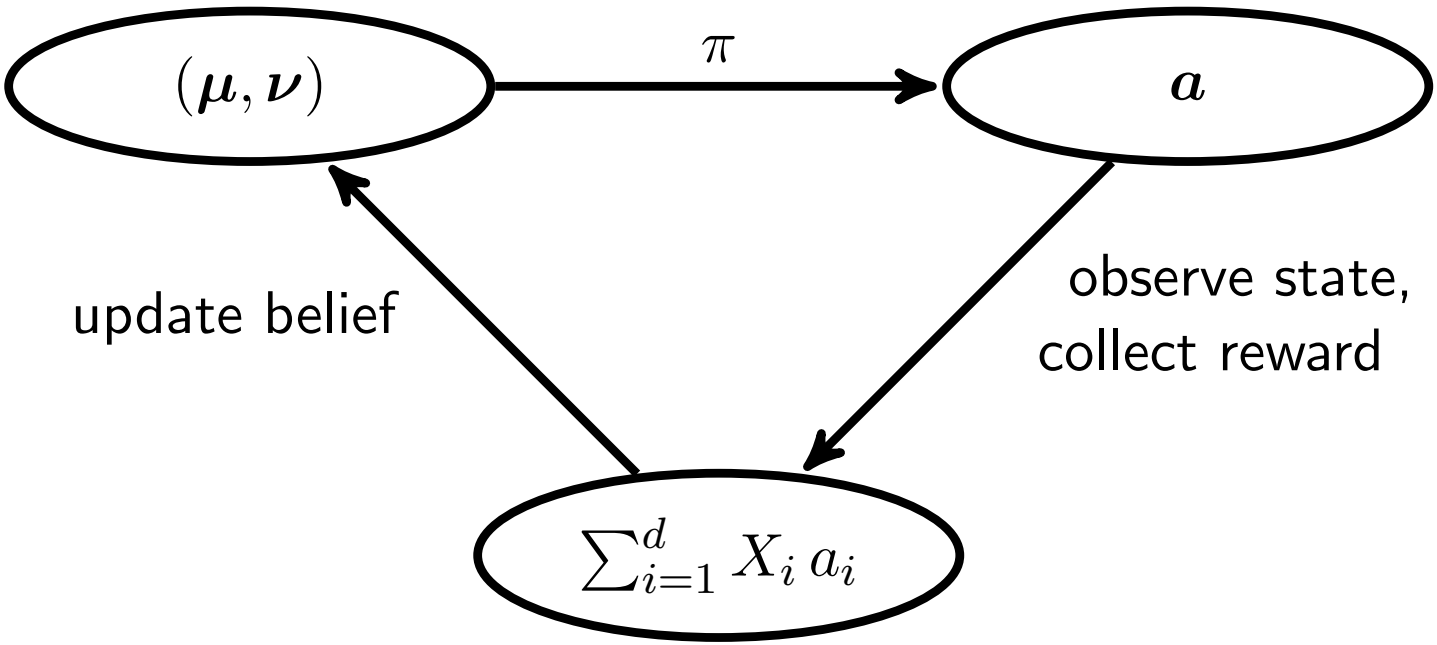
$$\big(\mu_i(t+1),\, \nu_i(t+1)\big) = \begin{cases} \big(\varphi\, \mu_i(t),\, \varphi^2\, \nu_i(t) + \sigma^2\big), & a_i(t) = 0, \\ \big(\varphi\, Y_{\mu_i(t),\, \nu_i(t)}\,,\, \sigma^2\big), & a_i(t) = 1. \end{cases}$$

Belief State Evolution

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

$\Rightarrow$ **Markov Decision Process**

Chain of Actions



Objective

Find a policy π so as to maximize the *total expected discounted reward criterion*,

$$V^{\pi}(\boldsymbol{\mu}, \boldsymbol{\nu}) := \lim_{T \to \infty} \mathbb{E}^{\pi}_{\boldsymbol{\mu}, \boldsymbol{\nu}} \left[\sum_{t=0}^{T} \beta^t \sum_{i=1}^{d} X_i(t)\, a_i(t) \right], \quad \beta \in (0, 1),$$

Objective

Find a policy π so as to maximize the *total expected discounted reward criterion*,

$$V^{\pi}(\boldsymbol{\mu}, \boldsymbol{\nu}) := \lim_{T \to \infty} \mathbb{E}^{\pi}_{\boldsymbol{\mu}, \boldsymbol{\nu}} \left[\sum_{t=0}^{T} \beta^t \sum_{i=1}^{d} X_i(t)\, a_i(t) \right], \quad \beta \in (0,1),$$

or the *average expected reward criterion*,

$$G^{\pi}(\boldsymbol{\mu}, \boldsymbol{\nu}) := \liminf_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi}_{\boldsymbol{\mu}, \boldsymbol{\nu}} \left[\sum_{t=0}^{T-1} \sum_{i=1}^{d} X_i(t)\, a_i(t) \right].$$

Objective

Find a policy π so as to maximize the *total expected discounted reward criterion*,

$$V^{\pi}(\boldsymbol{\mu}, \boldsymbol{\nu}) := \lim_{T \to \infty} \mathbb{E}^{\pi}_{\boldsymbol{\mu}, \boldsymbol{\nu}} \left[\sum_{t=0}^{T} \beta^t \sum_{i=1}^{d} X_i(t)\, a_i(t) \right], \quad \beta \in (0,1),$$

or the *average expected reward criterion*,

$$G^{\pi}(\boldsymbol{\mu}, \boldsymbol{\nu}) := \liminf_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi}_{\boldsymbol{\mu}, \boldsymbol{\nu}} \left[\sum_{t=0}^{T-1} \sum_{i=1}^{d} X_i(t)\, a_i(t) \right].$$

Note that we can replace $X_i(t)$ by $\mu_i(t)$!

How to find a good policy?

How to find a good policy?

Dynamic programming is typically intractable in practice.

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.

Examples for Index Functions

Myopic	$\gamma^M(\mu,\nu) = \mu$
Parametric	$\gamma^\theta(\mu,\nu) = \mu + \theta\nu, \quad \theta > 0$
Whittle	$\gamma^W(\mu,\nu) = \inf\left\{\lambda \,\middle|\, \pi^\lambda_{\mathsf{opt}}(\mu,\nu) = 0\right\}$

Definition

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

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.

Dynamic Programming Operator

Let $Tv := \max_{a\in\{0,1\}} T_a v$, where

$$T_a v\,(\mu,\nu) := \begin{cases} \lambda + \beta\, v(\varphi\,\mu,\ \varphi^2\nu + \sigma^2), & a = 0, \\ \mu + \beta \int_{-\infty}^{\infty} v\,(\varphi\, y,\ \sigma^2)\,\phi_{\mu,\nu}(y)\,dy, & a = 1, \end{cases}$$

with $\phi_{\mu,\nu}$ denoting the normal density with mean μ and variance ν.

Value iteration "works"

For $V_0^\lambda \equiv 0$ the iteration

$$V_n^\lambda = TV_{n-1}^\lambda$$

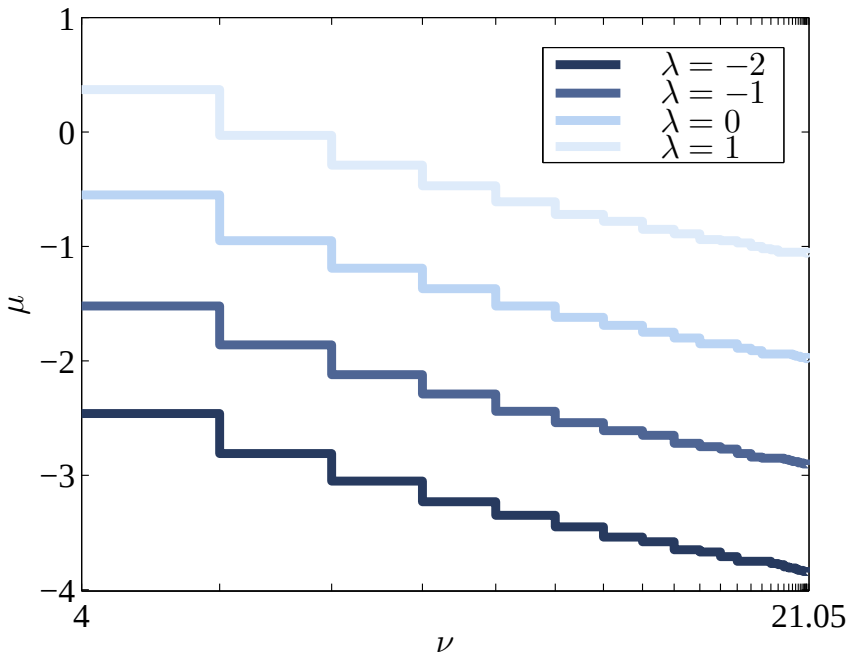
converges to a unique function $V^\lambda : \Psi \to \mathbb{R}$ as $n \to \infty$ that satisfies the Bellman equation,

$$V^\lambda = TV^\lambda.$$

This V^λ is the discount-optimal value function for the one-arm bandit problem with subsidy λ. An optimal policy for this problem maps (μ, ν) to action a if $V^\lambda(\mu, \nu) = T_a V^\lambda(\mu, \nu)$.

Threshold Policy

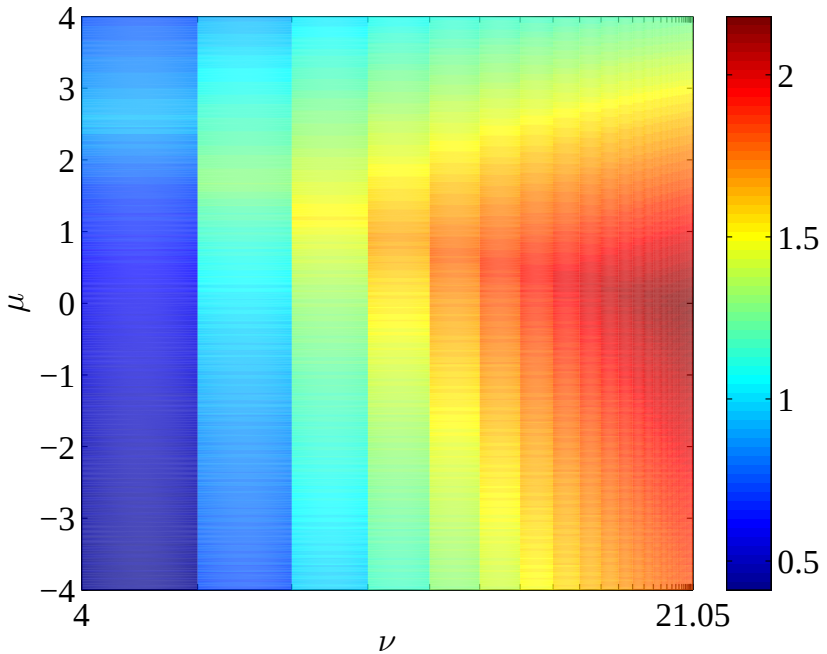
The optimal policy for the one-armed bandit problem with subsidy is a *threshold policy*.



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.



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

The Whittle index is a likely candidate for an asymptotically optimal policy, but no closed-form expression is known.

Parametric Index

$$\gamma(\mu, \nu) = \mu + \theta\nu,$$

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.

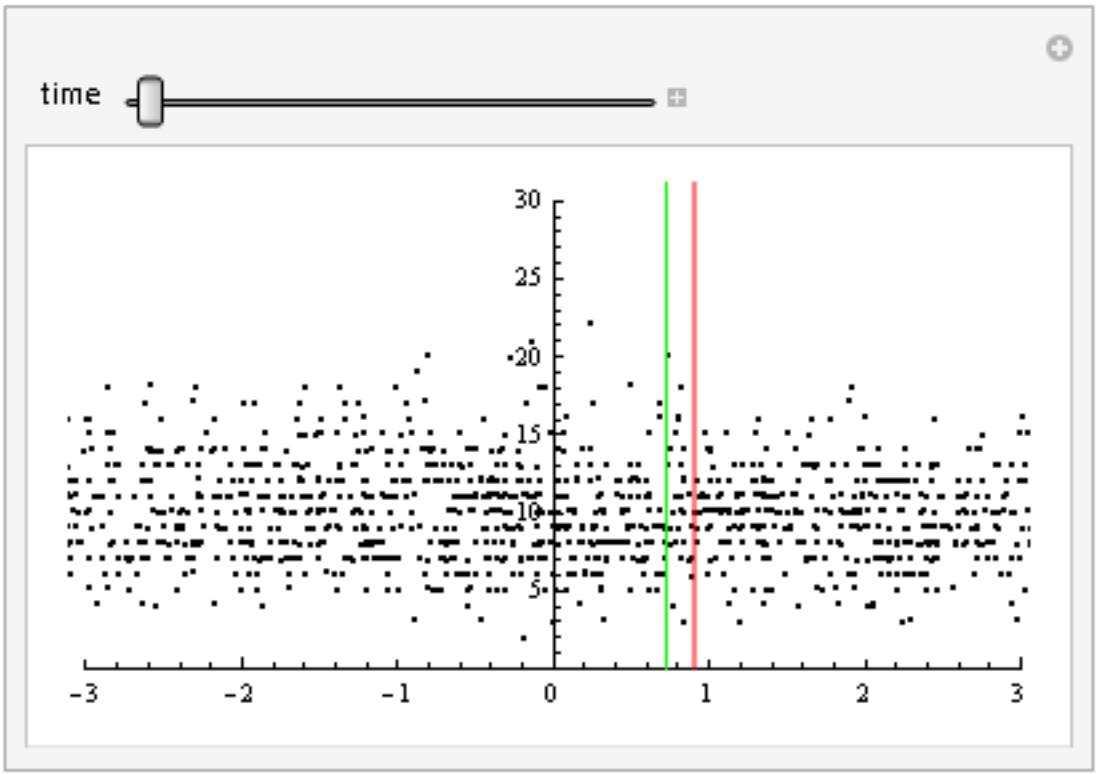
Asymptotic Behaviour: Intuition

- 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**.

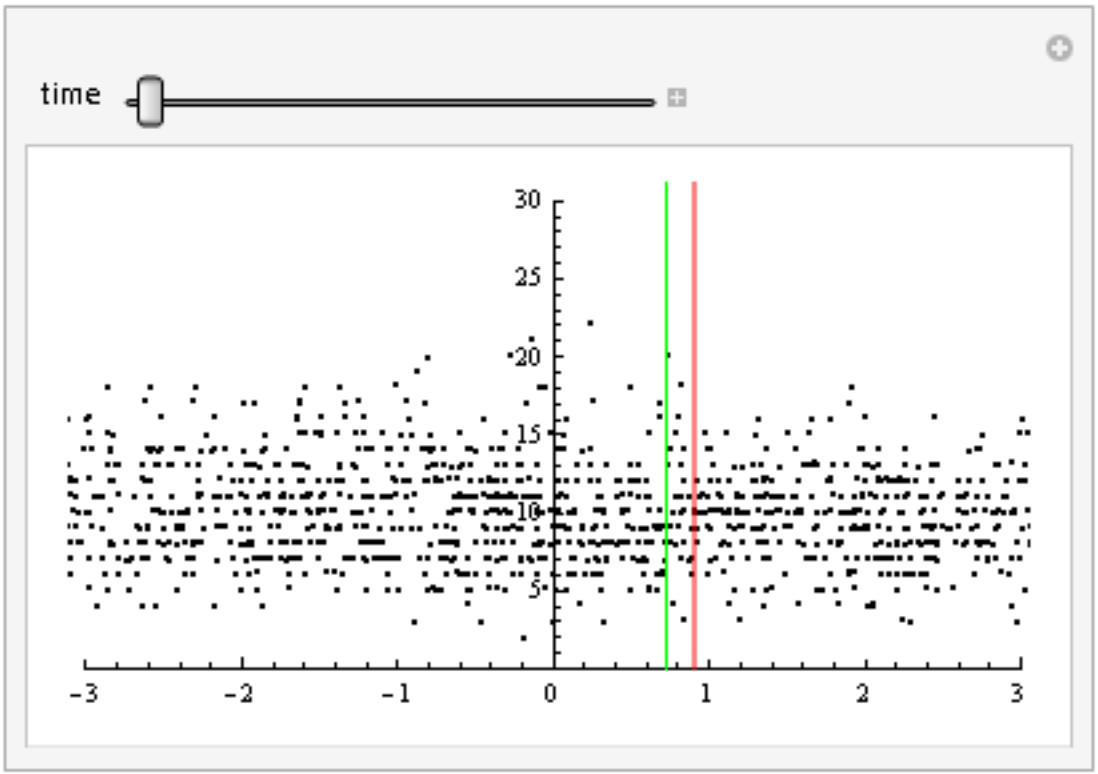
Asymptotic Behaviour: Intuition

- 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.

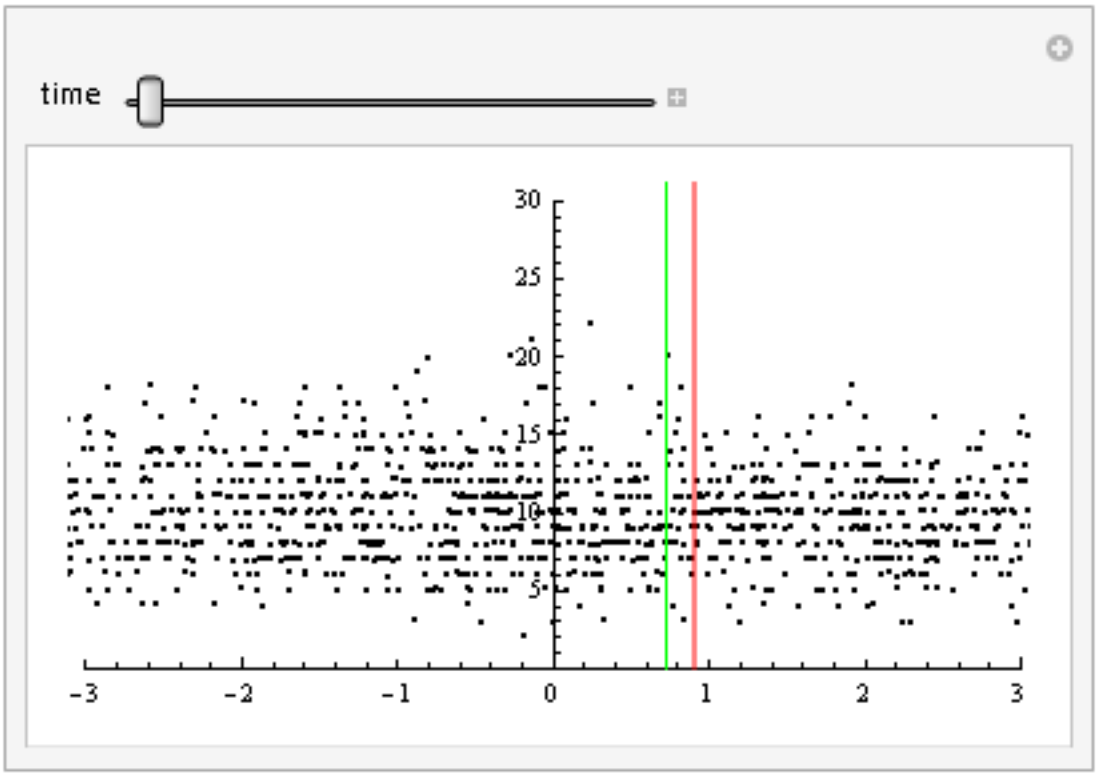
Example: Myopic (i.e. $\theta = 0$)



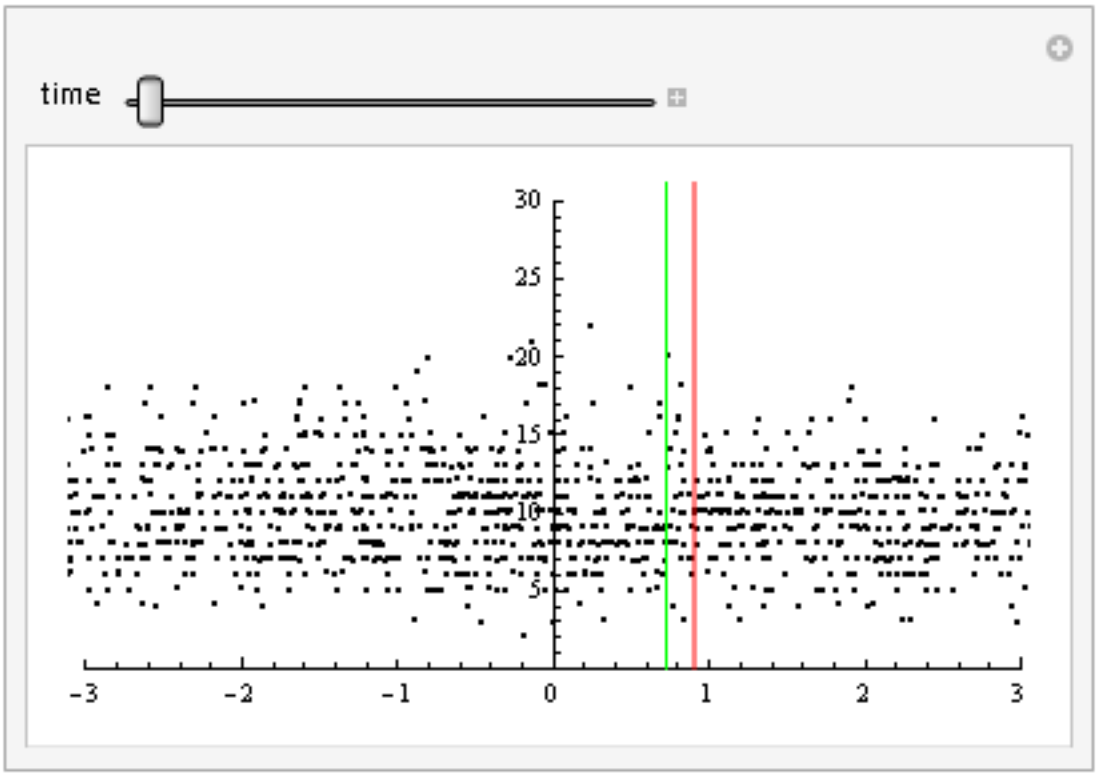
Example: Myopic (i.e. $\theta = 0$)



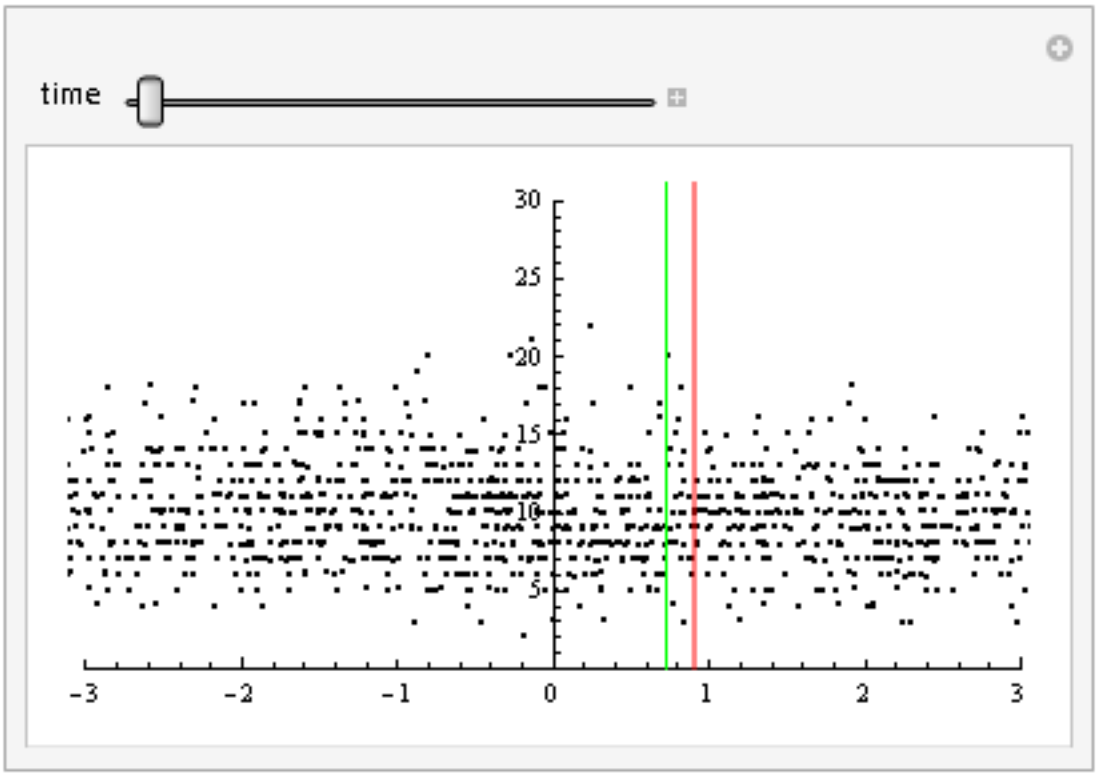
Example: Myopic (i.e. $\theta = 0$)



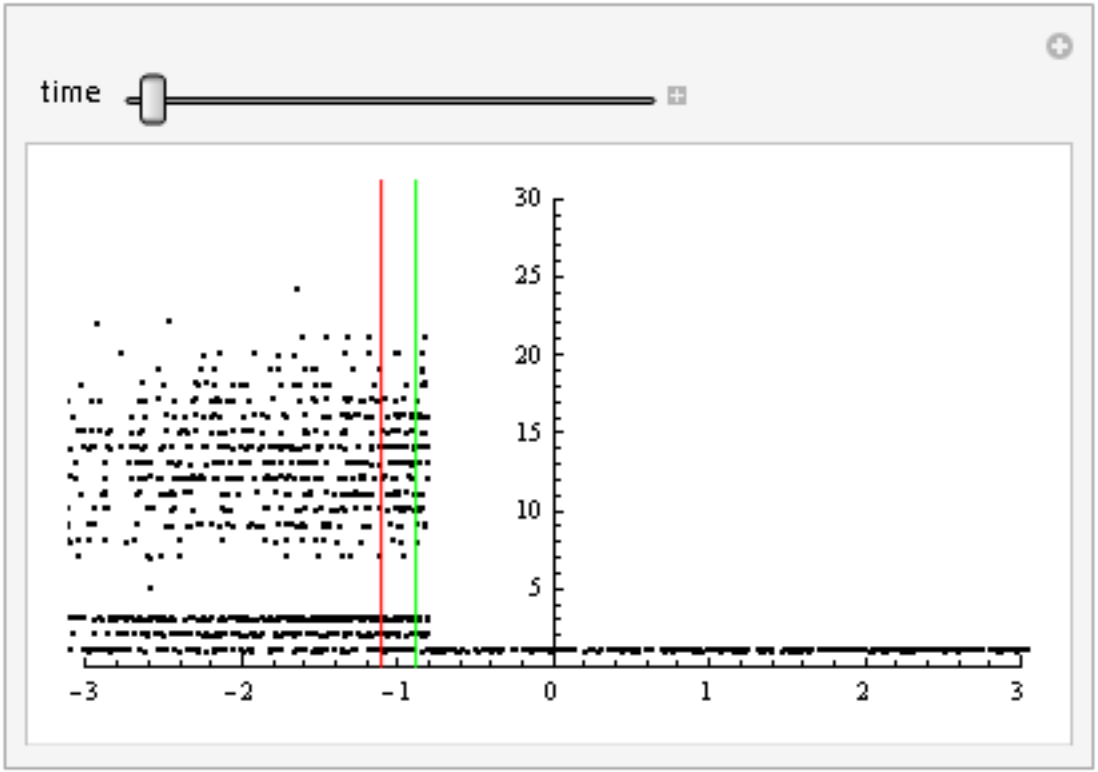
Example: Myopic (i.e. $\theta = 0$)



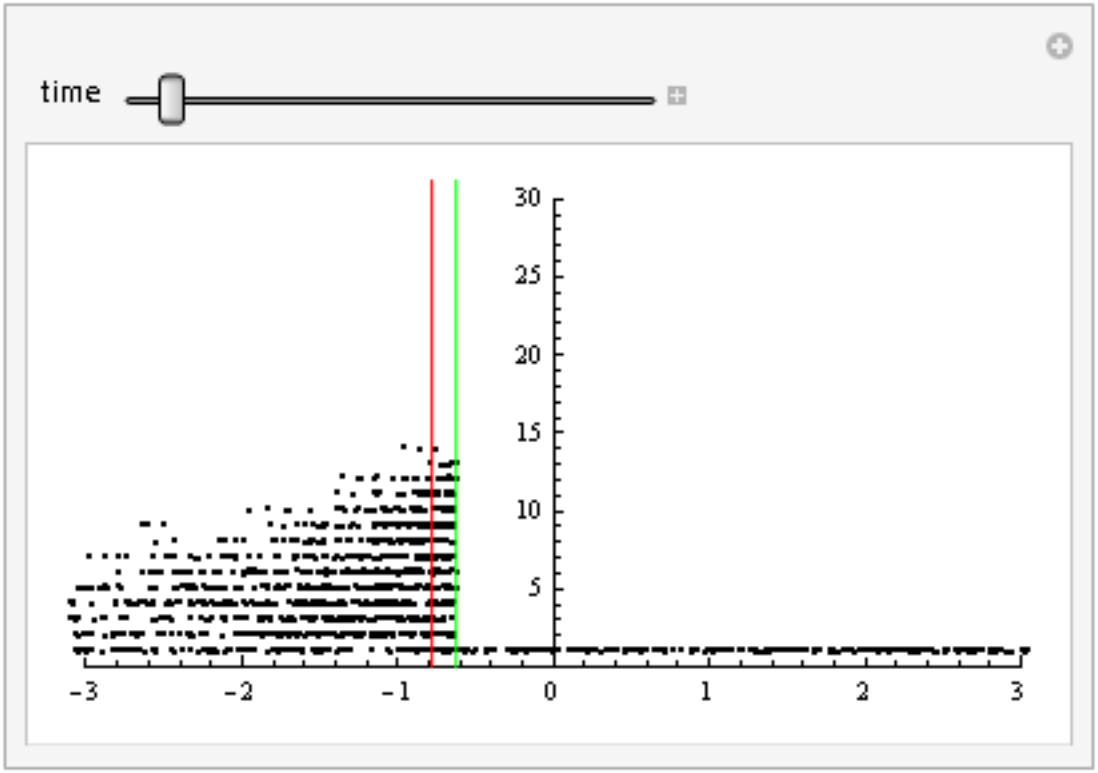
Example: Myopic (i.e. $\theta = 0$)



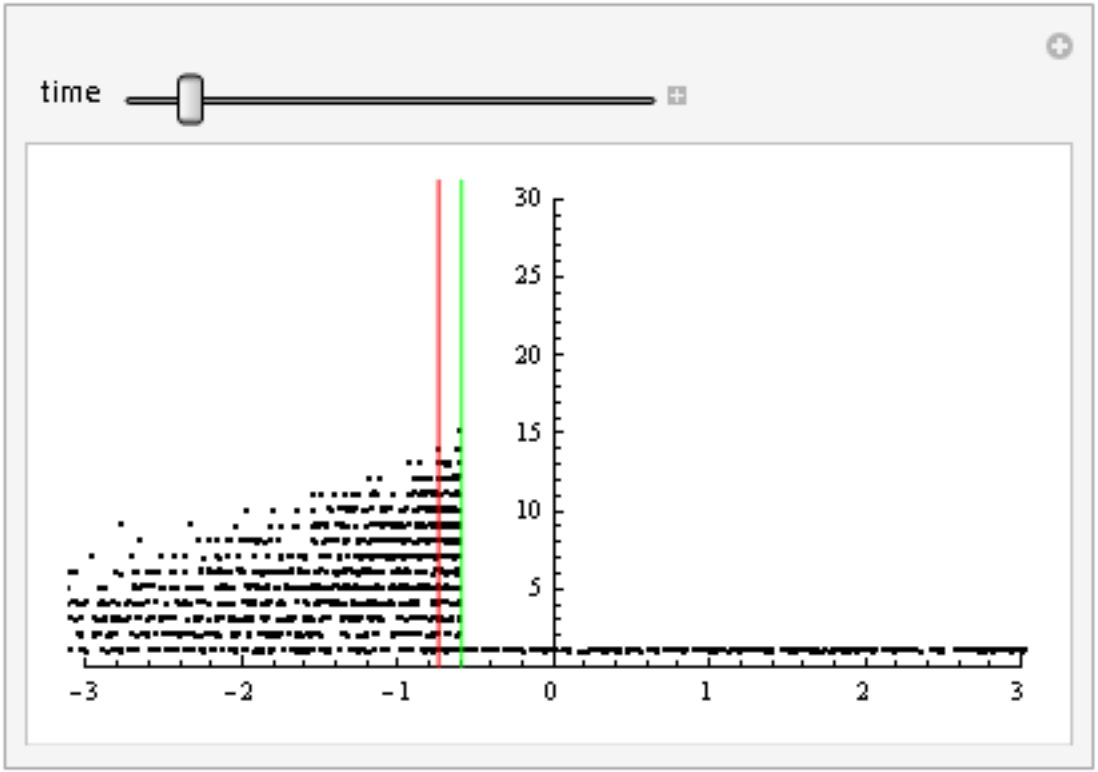
Example: Myopic (i.e. $\theta = 0$)



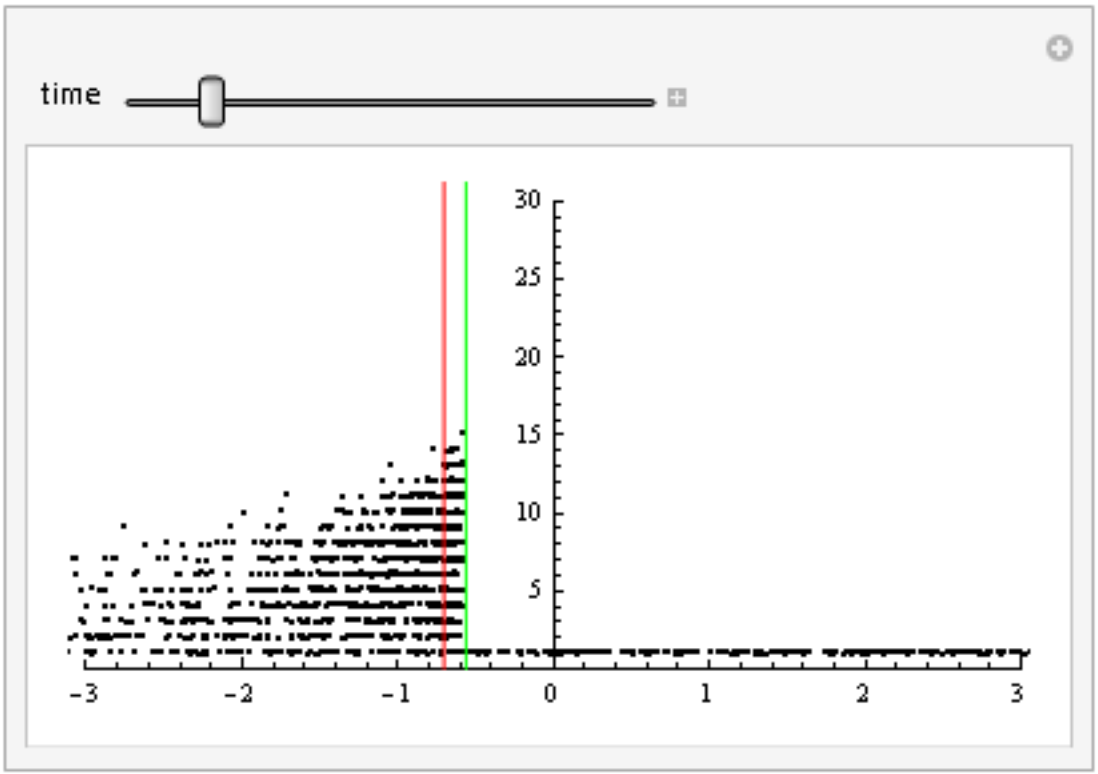
Example: Myopic (i.e. $\theta = 0$)



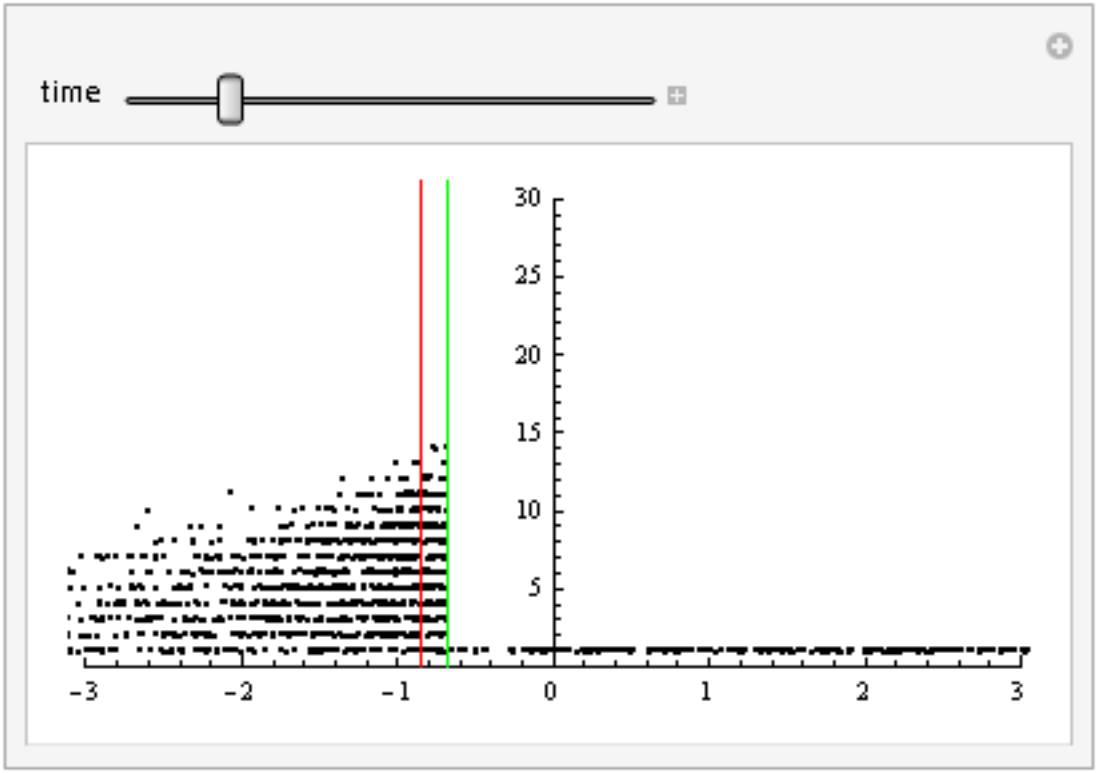
Example: Myopic (i.e. $\theta = 0$)



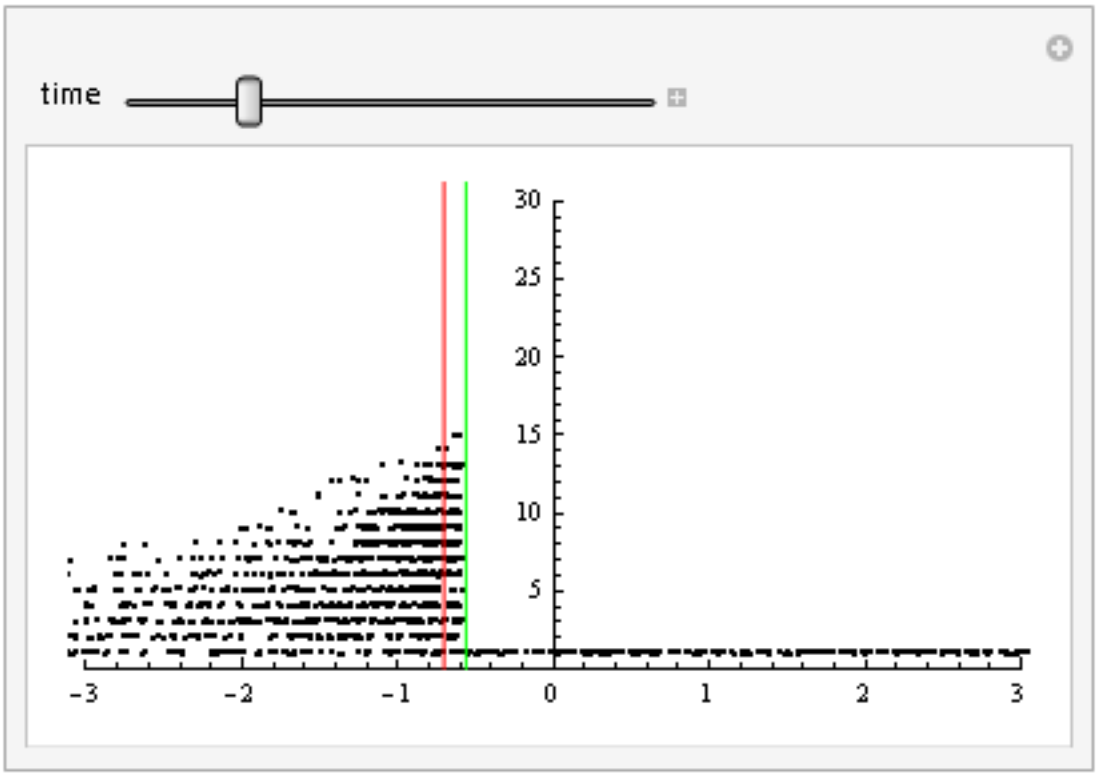
Example: Myopic (i.e. $\theta = 0$)



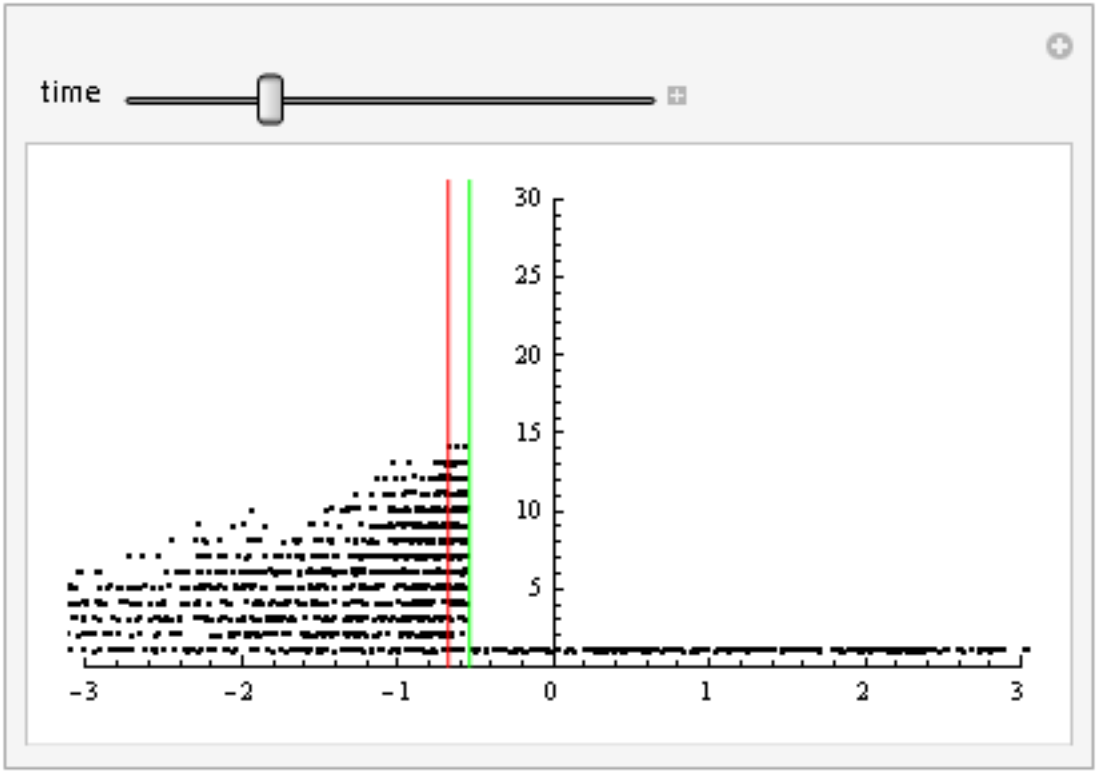
Example: Myopic (i.e. $\theta = 0$)



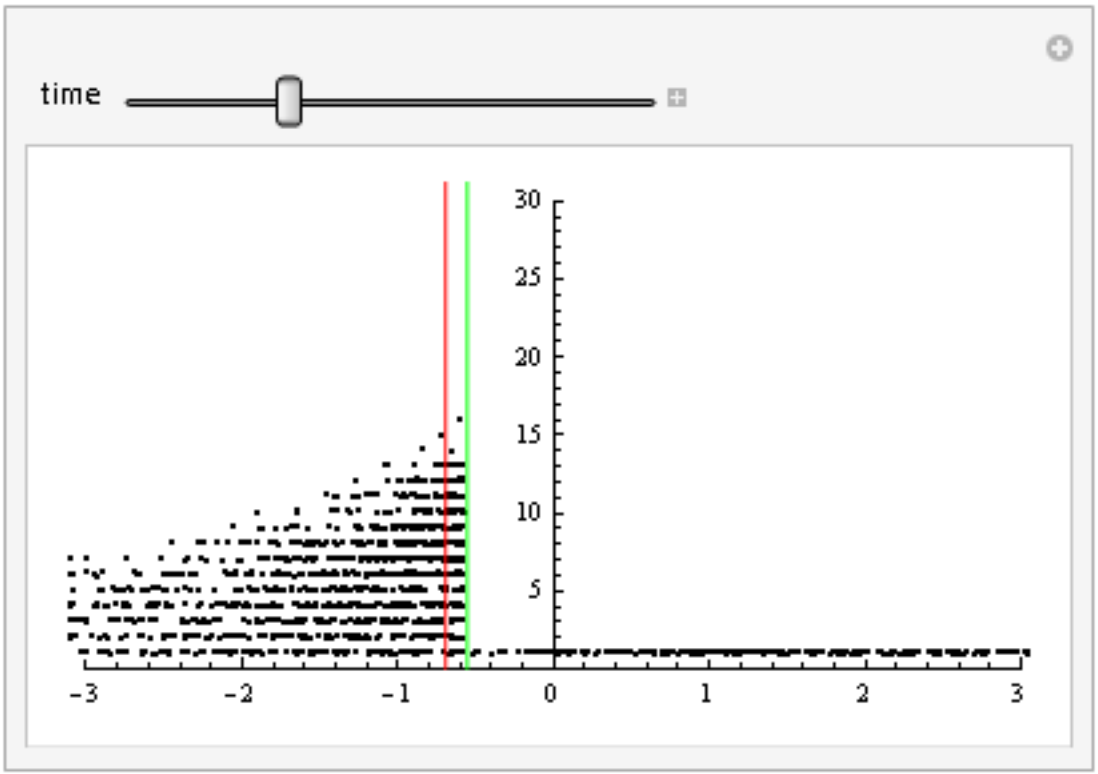
Example: Myopic (i.e. $\theta = 0$)



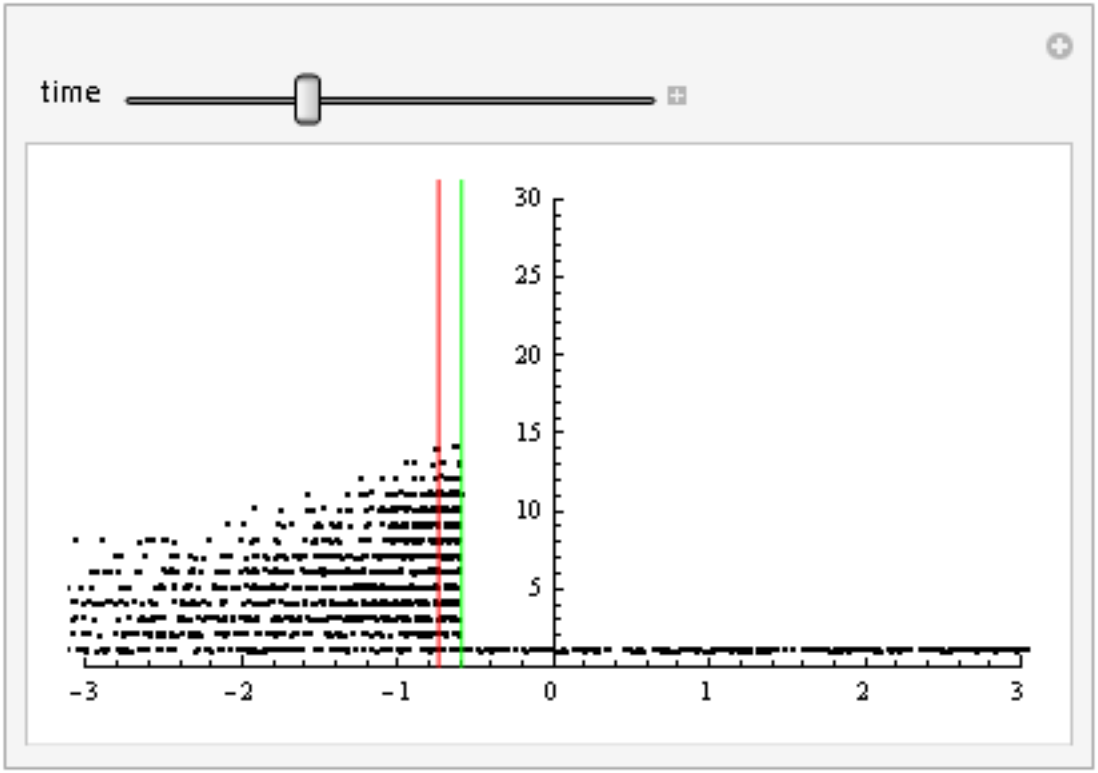
Example: Myopic (i.e. $\theta = 0$)



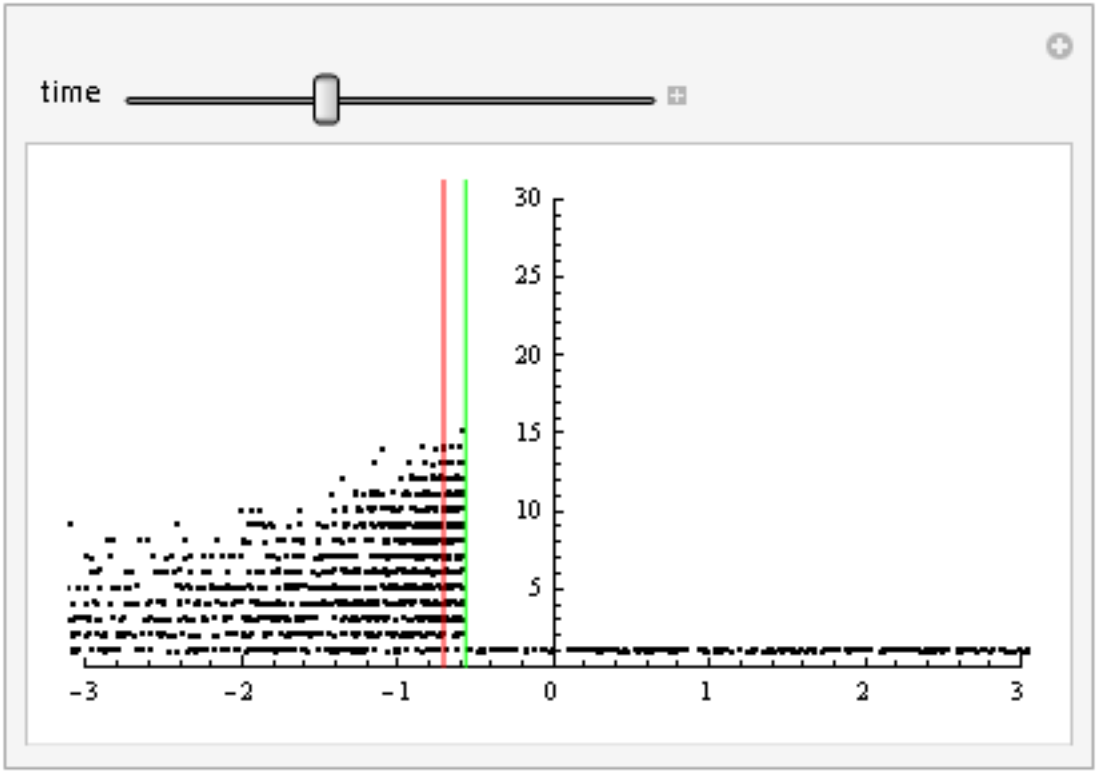
Example: Myopic (i.e. $\theta = 0$)



Example: Myopic (i.e. $\theta = 0$)



Example: Myopic (i.e. $\theta = 0$)



Proportion of arms with a certain belief state

Proportion of arms with a certain belief state

The empirical measure

$$M^d(C,t) := \frac{1}{d}\sum_{i=1}^{d} \mathbf{1}\Big\{\big(\mu_i(t),\, \nu_i(t)\big) \in C\Big\}$$

quantifies the proportion of arms in the d-dimensional system whose belief state falls into $C \in \mathcal{B}(\Psi)$ at time t.

Proportion of arms with a certain belief state

The empirical measure

$$M^d(C,t) := \frac{1}{d}\sum_{i=1}^{d} \mathbf{1}\Big\{\big(\mu_i(t),\, \nu_i(t)\big) \in C\Big\}$$

quantifies the proportion of arms in the d-dimensional system whose belief state falls into $C \in \mathcal{B}(\Psi)$ at time t.

It can be written as

$$\sum_{h=0}^{\infty} M_h^d(B,t) := \sum_{h=0}^{\infty} \frac{1}{d} \sum_{i:\, \nu_i(t)=\nu^{(h)}} \mathbf{1}\big\{\mu_i(t) \in B\big\},$$

where $B \in \mathcal{B}(\Psi_1)$, and $\nu^{(h)}$ is the h-th element in Ψ_2, the state space of ν.

Many-Arms Asymptotic Behaviour

Many-Arms Asymptotic Behaviour

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

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

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

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 $\mu(t) \geq \ell^*_h(t)$.

Many-Arms Behaviour

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

Many-Arms Behaviour

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

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: Many-Arms Behaviour

Assume that $M_h^d(B,0)$ converges weakly to $m_h(B,0)$ 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 $t, h \geq 0$,

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

Conjecture: Equilibrium State

Conjecture: Equilibrium State

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 $\overline{\ell}$, namely $\overline{\ell} = \ell^*$ from before.

Conjecture: Equilibrium State

Assume that the index is parametric, and that $\Gamma^{\ell}(t)$ is stationary. Then the equation

$$\mathbb{P}\left(\Gamma^{\ell}(t) \geq \ell\right) = \rho$$

has a unique solution ℓ^*, which satisfies

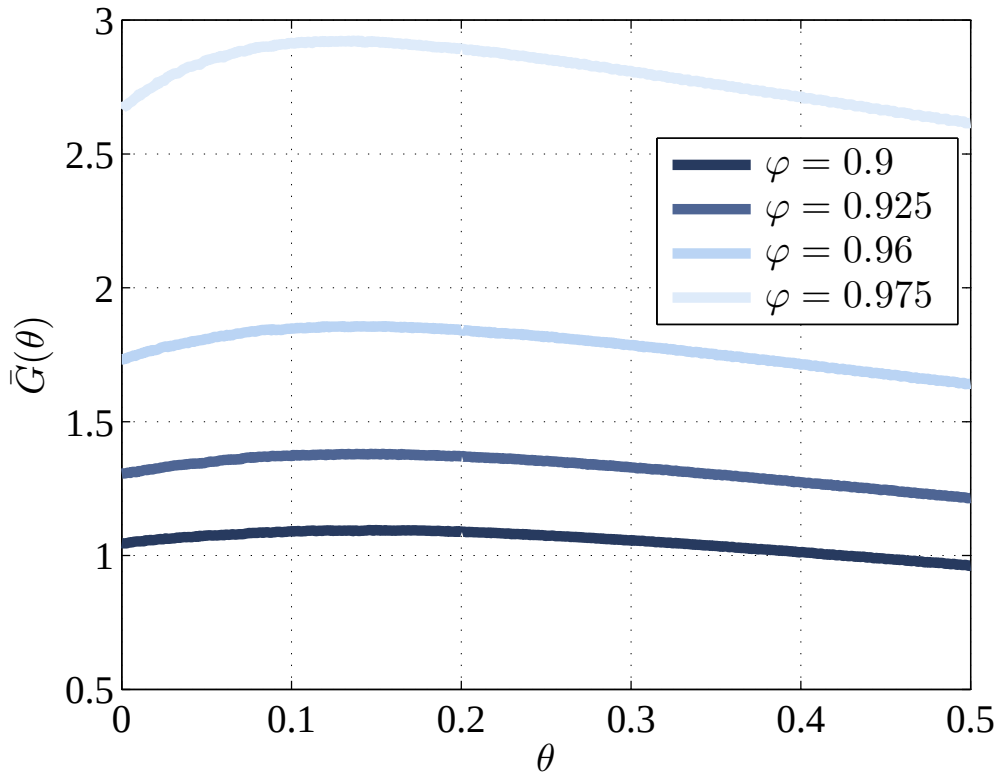
$$\ell^* = \sup\left\{\ell \,\middle|\, \sum_{h=0}^{\infty} \widetilde{m}_h^*\Big([\ell, \infty)\Big) = \rho\right\},$$

and

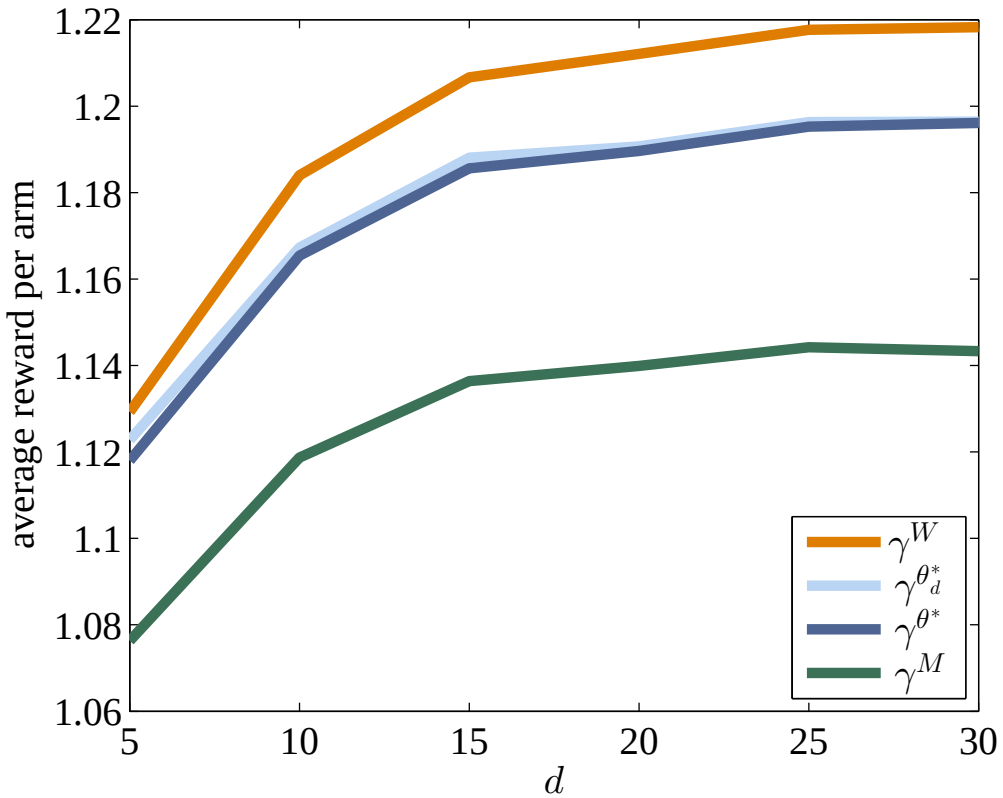
$$\mathbb{P}\left(\Gamma^{\ell^*}(t) \in B\right) = \sum_{h=0}^{\infty} \widetilde{m}_h^*(B), \, \forall \, B \in \mathcal{B}(\mathbb{R}).$$

Algorithm for Performance Evaluation

1. For large T determine $\overline{\ell}$ 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^*). $\varphi = 0.9$, $\sigma = 2$, $\rho = 0.4$, $T = 10^5$.

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. J. K., M. Mandjes and Y. Nazarathy (2014). Exploration vs. Exploitation with Partially Observable Gaussian Autoregressive Arms. *Submitted*.
4. K. Liu and Q. Zhao (2010). Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access. *IEEE Trans. o. Inf. Theory*, 56, pp. 5547–5567.
5. I. M. Verloop (2014). Asymptotic Optimal Control of Multi-Class Restless Bandits. *Submitted*.
6. P. Whittle (1988). Restless bandits: Activity Allocation in a Changing World. *Journal of Applied Probability*, 25, pp. 287–298.

Thank you!