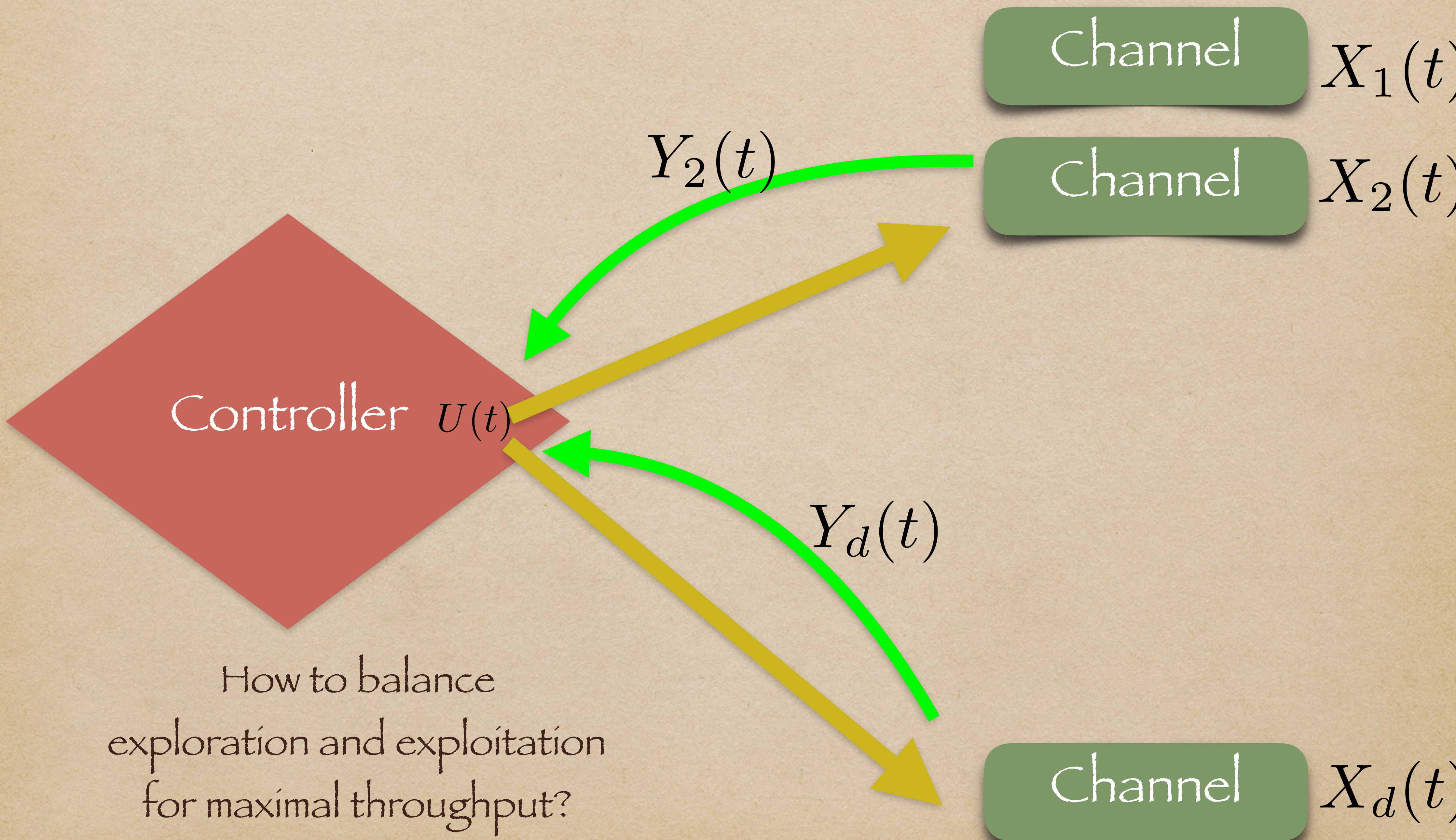
Reward Observing Restless Multi Armed Bandits

Yoni Nazarathy,
The University of Queensland

Controller Chooses Channels



How to balance
exploration and exploitation
for maximal throughput?

{State, Control, Observation} Model

$$X_i(t) \in \mathbb{R} \qquad U(t) \subset \{1, \ldots, d\} \qquad Y_i(t) \in \mathbb{R}$$

Instantaneous Reward: $\sum_{i \in U(t)} r\big(X_i(t)\big) - c\big|U(t) \setminus U(t^-)\big|$

Constraint: $\big|U(t)\big| = k < d$

$\epsilon(t) \sim \mathcal{N}(0, 1)$

$M(t)$ is Markov Chain

Channel: $X(t+1) = \alpha'_{M(t)}\big(X(t), X(t-1), \ldots, X(t-p+1)\big) + \sigma_{M(t)}\epsilon(t) + c_{M(t)}$

Observation: $Y_i(t) \sim \begin{cases} p_i\big(\,\cdot \mid X(t)\big), & i \in U(t), \\ q_i\big(\,\cdot \mid X(t)\big), & i \notin U(t). \end{cases}$

Control Policy: $U(t) = \pi\big(\{Y(t),\, t \in (-\infty, t)\}\big)$

Objective: Maximal Infinite Horizon Average Reward

MDP/POMDP State

Option 1: $\eta_i = \big(X_i(t-\tau_i),\, \tau_i\big)$

Option 2: $F_i(x) = \mathbb{P}(X_i(t) \leq x \mid \text{observed history})$

Option 2*: Find sufficient statistics, ω_i, for $F_i(\cdot)$

We use option 2* hence: $U(t) = \pi\big(\omega_1(t), \ldots, \omega_d(t)\big)$

Restless Bandits

Restless Bandits: Activity Allocation in a Changing World
Author(s): P. Whittle
Source: *Journal of Applied Probability*, Vol. 25, A Celebration of Applied Probability (1988), pp. 287-298

Example: 1 Mother, Triplets to Feed

Can feed at most 2 at a time

Triplets evolve between “sleeping”,”playing”,”crying”

Cost: Num Being Fed + Num Crying

Reward Observing Restless Bandit

Belief State Update - not considering $Y(t)$:

$$\omega_i(t+1) = \begin{cases} \mathcal{O}_i\big(X_i(t)\big)\,, & \text{if } i \in U(t)\,, \quad X_i(t) \sim \text{according to } \omega_i(t) \\ \mathcal{T}_i\big(\omega_i(t)\big)\,, & \text{if } i \notin U(t)\,. \end{cases}$$

Observation update:

("active" in RMAB language)

$$\mathcal{O}_i \sim \omega_i(t)$$

Belief propagation operator:

("passive" in RMAB language)

Deterministic $\mathcal{T}_i(\cdot)$

GE and AR Channels

Gilbert Elliot
(2 state MC)

$$\mathcal{O}_i(x) = \begin{cases} p_i^{01}, & \text{if } x = 0\,, \\ p_i^{11}, & \text{if } x = 1\,, \end{cases}$$

$$\mathcal{T}_i(\omega) = \omega\, p_i^{11} + \bar{\omega}\, p_i^{01}$$

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 56, NO. 11, NOVEMBER 2010

Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access

Keqin Liu and Qing Zhao

Auto Regressive Gaussian Process of Order 1

$$\mathcal{O}_i(x) = \left(\varphi_i\, x,\, \sigma_i^2\right)$$

$$\mathcal{T}_i(\mu_i,\, \nu_i) = \left(\varphi_i\, \mu_i,\, \varphi_i^2\, \nu_i + \sigma_i^2\right)$$

Slow Fading Channel Selection: A Restless Multi-Armed Bandit Formulation

Konstantin Avrachenkov
INRIA, Maestro Team
BP95, 06902 Sophia Antipolis, France
Email: k.avrachenkov@sophia.inria.fr

Laura Cottatellucci, Lorenzo Maggi
Eurecom
Mobile Communications Department
BP193, F-06560 Sophia Antipolis, France
Email: {laura.cottatellucci,lorenzo.maggi}@eurecom.fr

Exploration vs. Exploitation with Partially Observable Gaussian Autoregressive Arms

Julia Kuhn
The University of Queensland,
University of Amsterdam
j.kuhn@uq.edu.au

Michel Mandjes
University of Amsterdam
m.r.h.mandjes@uva.nl

Yoni Nazarathy
The University of Queensland
y.nazarathy@uq.edu.au

Index Policies

(because solving the POMDP is often hard)

$$I_i(t) = f_i\big(\omega_i(t)\big) \qquad U(t) = \arg\overset{(k)}{\max}\{I_1(t), \ldots, I_d(t)\}$$

Myopic Index: $f_i(\omega) = \mathbb{E}_\omega r_i(X_i)$

Index Considering Variance:

$f_i(\omega) = \mathbb{E}_\omega\, r_i(X_i) + \theta_i \,\mathrm{Var}(X_i)$ What is the best θ_i?

Whittle Index:

$f_i(\omega)$ is the minimal subsidy you pay to not select the channel

To calculate it - solve a family of "one armed subsidy problems"

Numerical Examples

Wireless Channel Selection with Restless Bandits

Julia Kuhn and Yoni Nazarathy

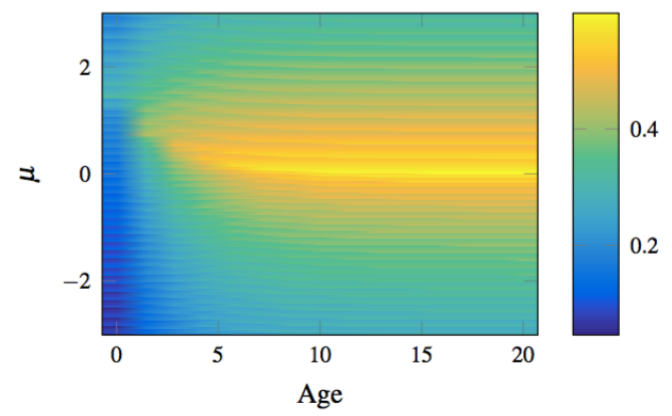


Fig. 6 Contour plot of $\gamma_W(\mu,\nu) - r(\mu)$, the difference of Whittle and myopic indices, for an AR channel with $\varphi = 0.8$, $\sigma = 2$.

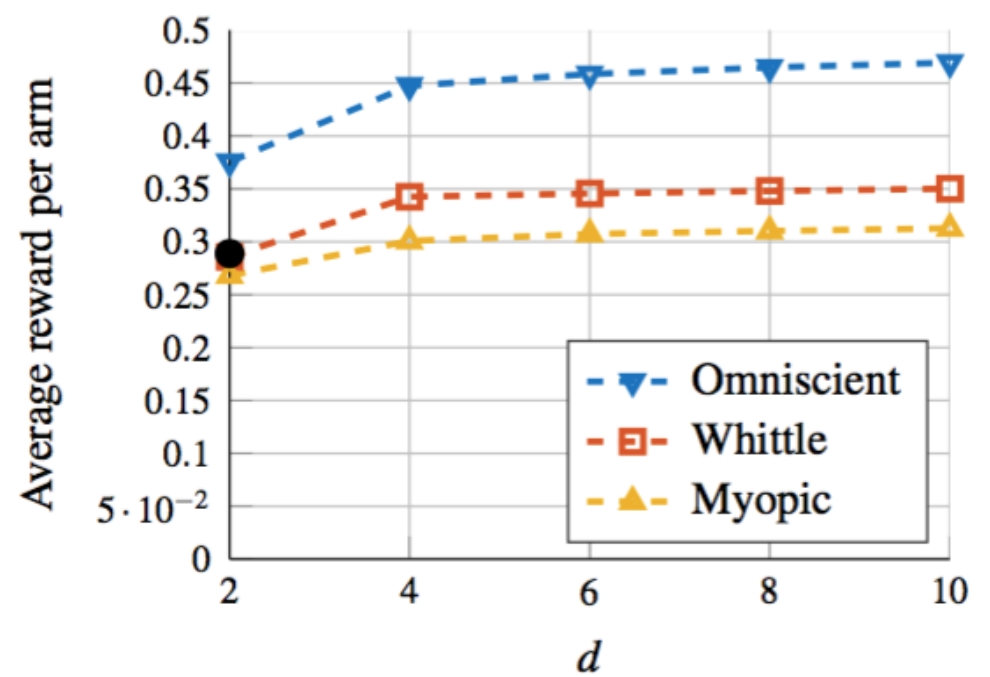


Fig. 3 Comparison of Whittle and myopic index policies for increasing number of channels d when half of the channels are GE and the other half is AR. For $d = 2$, the average reward obtained under the optimal policy is indicated by a black dot. We compare to the average reward that could be obtained if both arms were observed at each time point (that is in the fully observable or "omniscient" setting).

Themes and Methods

- Structural Properties of Optimal or Index Based Policies
- Queue Stability and Observation Error
- Switching Costs
- Use of regenerative structure
- Measure Valued Asymptotics for belief states
- In Progress: A computational framework and Unknown Parameters (regret)

Structural Properties

Exploration vs. Exploitation with Partially Observable Gaussian Autoregressive Arms

Julia Kuhn
The University of Queensland, University of Amsterdam
j.kuhn@uq.edu.au

Michel Mandjes
University of Amsterdam
m.r.h.mandjes@uva.nl

Yoni Nazarathy
The University of Queensland
y.nazarathy@uq.edu.au

The Role of Information in System Stability with Partially Observable Servers

Azam Asanjarani*, Yoni Nazarathy†

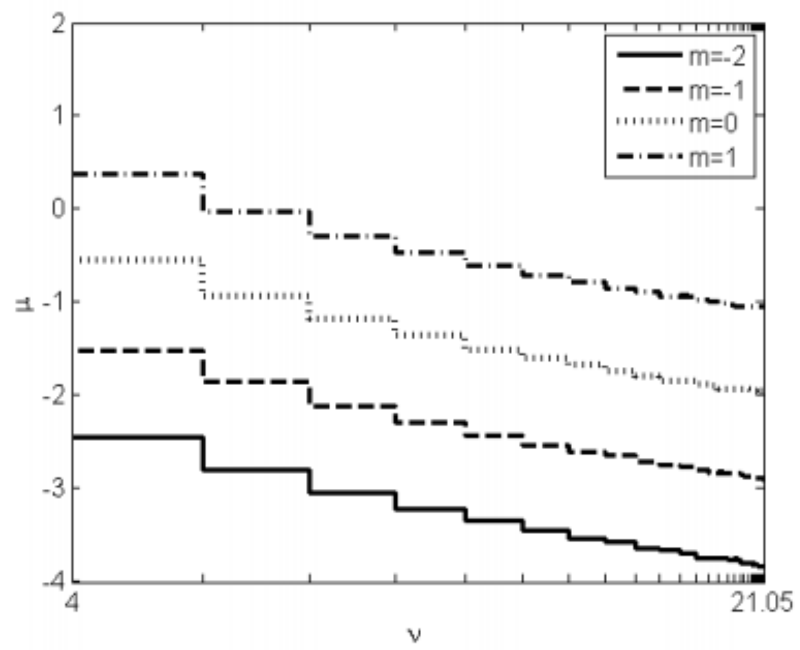
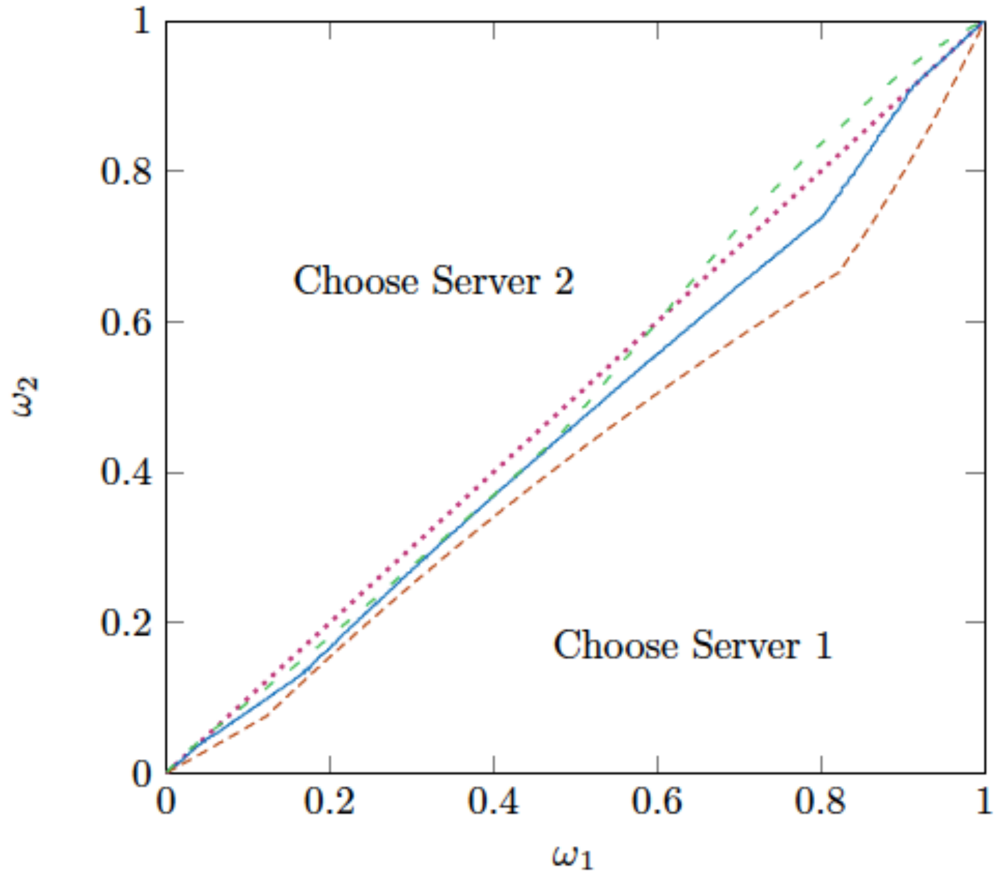


Figure 1: Switching curves: below the curve the optimal action is passive, above it is active. $\beta = 0.8$, $\varphi = 0.9$, $\sigma = 2$.



Queue Stability and Observation Error

The Challenge of Stabilizing Control for Queueing Systems with Unobservable Server States

Yoni Nazarathy*‡, Thomas Taimre*, Azam Asanjarani*, Julia Kuhn*†, Brendan Patch*†, and Aapeli Vuorinen*.
*School of Mathematics and Physics, The University of Queensland, Australia.
†Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands.
‡Email: y.nazarathy@uq.edu.au

$$P^j_{\mathrm{GE}} = \begin{bmatrix} \overline{p} & p \\ q & \overline{q} \end{bmatrix} = \begin{bmatrix} 1-\gamma\overline{\rho} & \gamma\overline{\rho} \\ \overline{\gamma}\overline{\rho} & 1-\overline{\gamma}\overline{\rho} \end{bmatrix}$$

$$\tau(\omega) = \overline{q}\omega + p\overline{\omega} = \omega\rho + \gamma(1-\rho),$$

$$\tau_0(\omega) = \frac{\overline{q}\overline{\mu}_2\omega + p\overline{\mu}_1\overline{\omega}}{\overline{r}(\omega)}, \qquad \tau_1(\omega) = \frac{\overline{q}\mu_2\omega + p\mu_1\overline{\omega}}{r(\omega)}$$

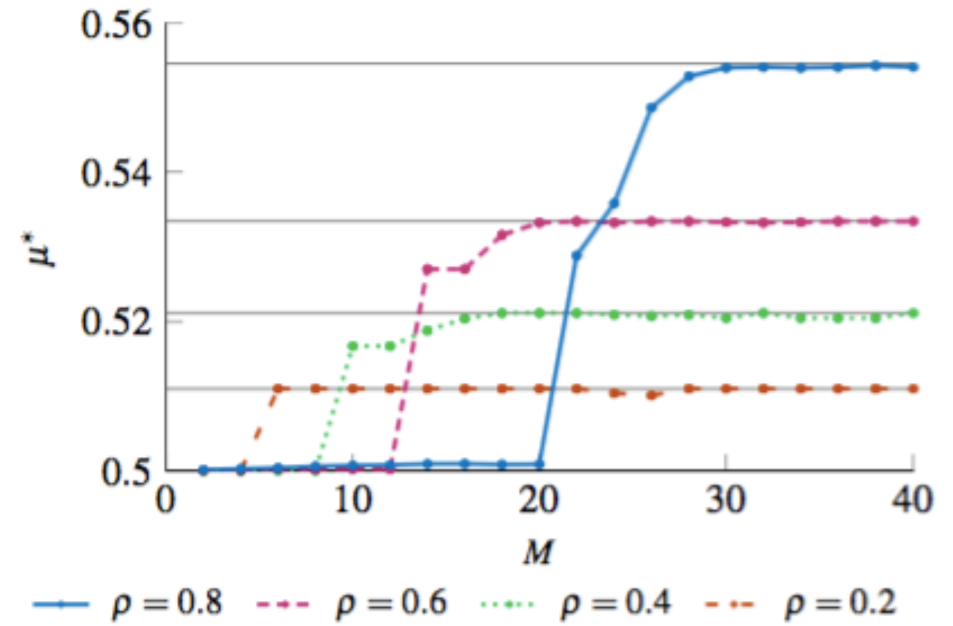


Fig. 6. Stability region achieved by finite state controllers for increasing M. The limiting horizontal lines are at μ^* as computed by means of relative valuate iteration of Section IV.

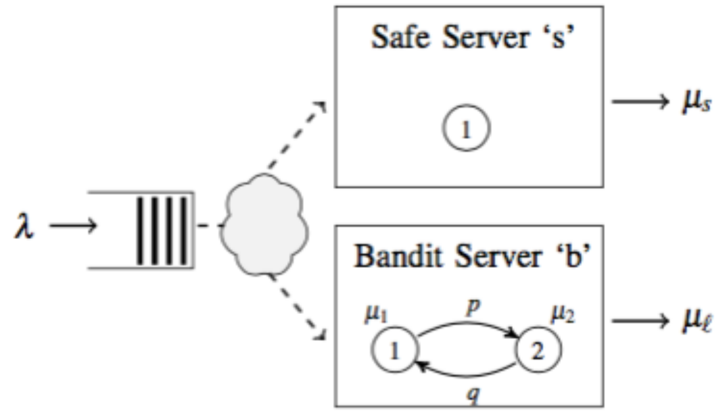
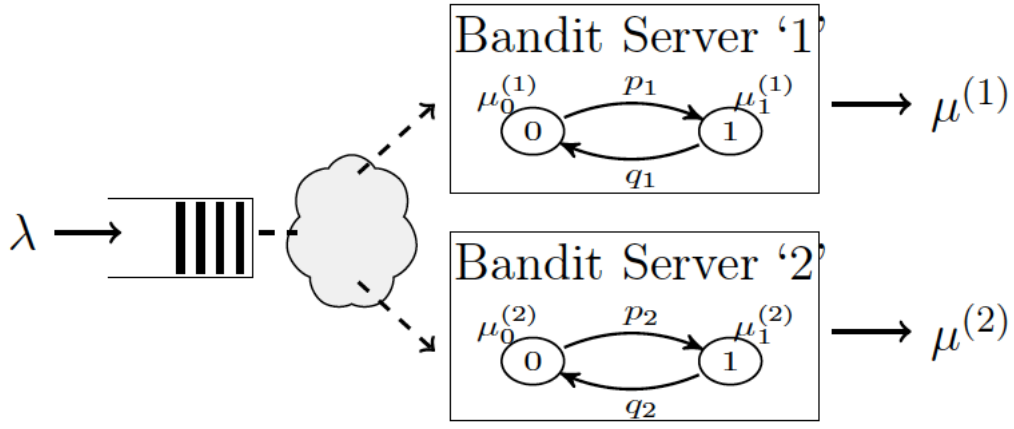
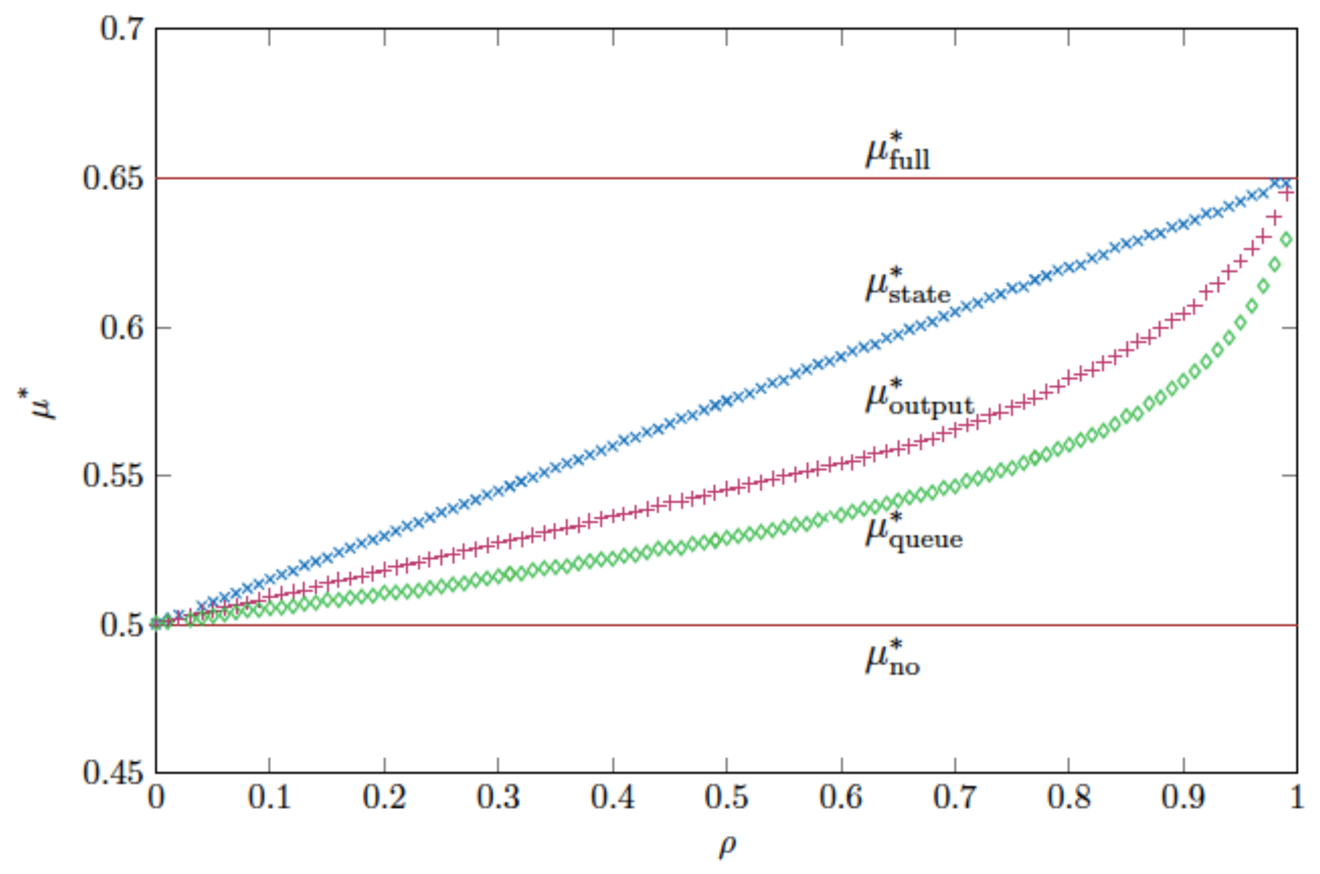


Fig. 2. The simplest (specialized) queueing system analyzed throughout this paper with the exception of Section III.

The Role of Information in System Stability with Partially Observable Servers

Azam Asanjarani[*], Yoni Nazarathy[†]



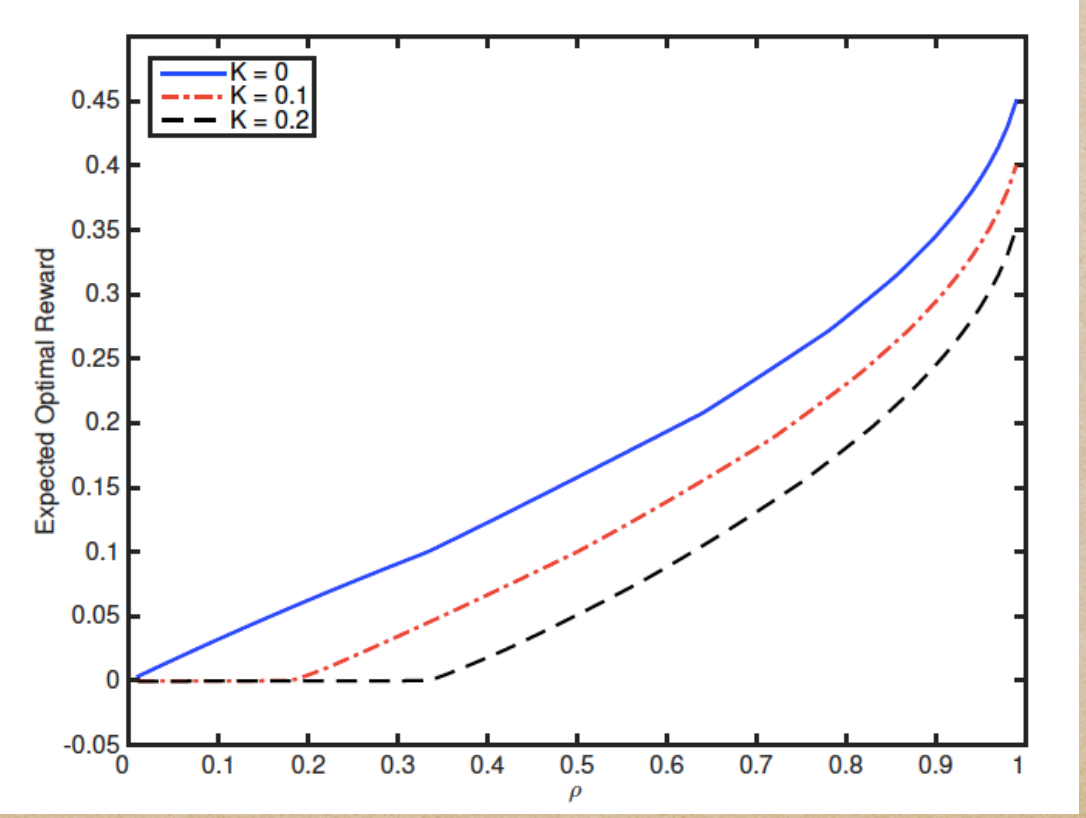


Use of Regenerative Structure

To Fish or Cut Bait?

Jiahao Diao, Yoni Nazarathy, Thomas Taimre, and Jerzy A. Filar.
School of Mathematics and Physics,
The University of Queensland.

$$R(t) = A(t)\big(X(t) - (1 - X(t)) - K\big)$$



Switching Costs

The Value of Information and Efficient Switching in Channel Selection

Jiesen Wang, Yoni Nazarathy, Thomas Taimre

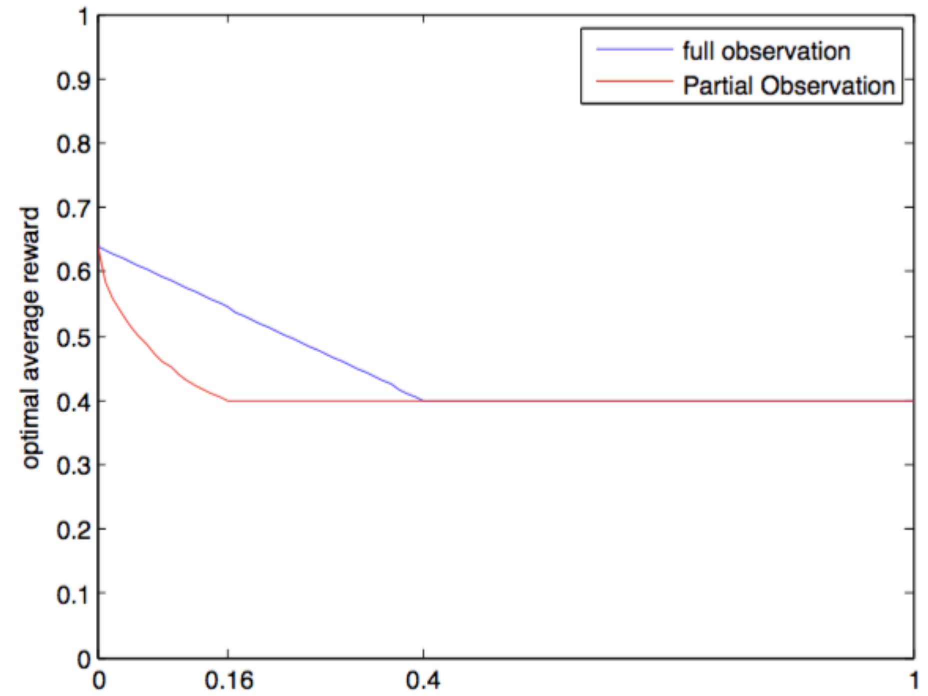


Fig. 1: The Optimal average reward for a system with $\gamma = 0.4$ as a function of cost.

Using Regenerative Calculations - An explicit equation for optimal call-gapping:

If $c < \gamma^2$ switch not before τ^*, solution of:

$$e^{\frac{2\tau^*}{\gamma}}(\gamma - 2)(c - \gamma^2) + 2\gamma e^{\frac{\tau^*}{\gamma}}(\gamma + \tau^*(1-\gamma)) + \gamma(c - \gamma^2) = 0$$

Measure Valued Asymptotics

For AR Channels (a bit complicated)

Exploration vs. Exploitation with Partially Observable Gaussian Autoregressive Arms

Julia Kuhn
The University of Queensland, University of Amsterdam
j.kuhn@uq.edu.au

Michel Mandjes
University of Amsterdam
m.r.h.mandjes@uva.nl

Yoni Nazarathy
The University of Queensland
y.nazarathy@uq.edu.au

$$m_h(x, t+1) = \begin{cases} \sum_{h=0}^{\infty} \int_{\ell_h(t)}^{\infty} \Phi_{z,\nu^{(h)}}\left(\frac{x}{\varphi}\right) m_h(dz, t), & h = 0, \\ m_{h-1}\left(\min\left\{\frac{x}{\varphi}, \ell_{h-1}(t)\right\}, t\right), & h \geq 1, \end{cases} \tag{14}$$

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

$$\ell(t) = \sup\left\{\ell \,\middle|\, \sum_{h=0}^{\infty} \widetilde{m}_h\big([\ell, \infty), t\big) = \rho\right\}. \tag{15}$$

Here, $\widetilde{m}_h$ denotes the measure on indices, i.e.

$$\widetilde{m}_h(B, t) = m_h\Big(\{\mu \in \mathbb{R} \,|\, \mu + \theta\nu^{(h)} \in B\}, t\Big), \tag{16}$$

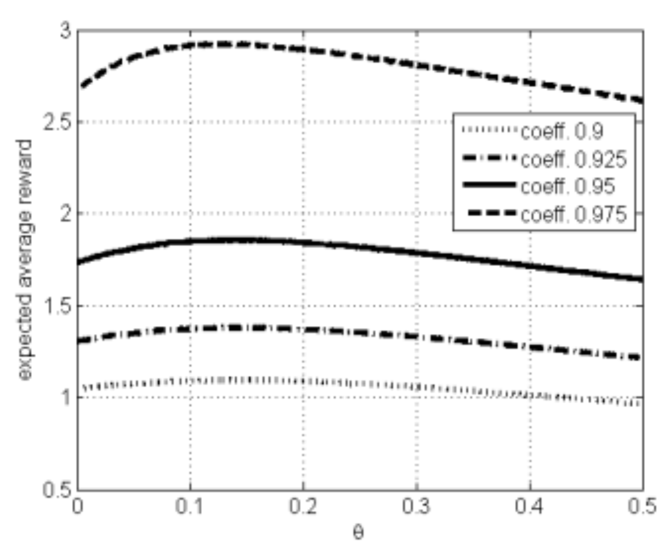


Figure 3: Expected average reward $\bar{G}(\theta)$ computed by the algorithm as a function of θ. $\sigma = 2$, $\varphi \in \{0.9, 0.925, 0.95, 0.975\}$, $\rho = 0.4$, $T = 2 \times 10^6$.

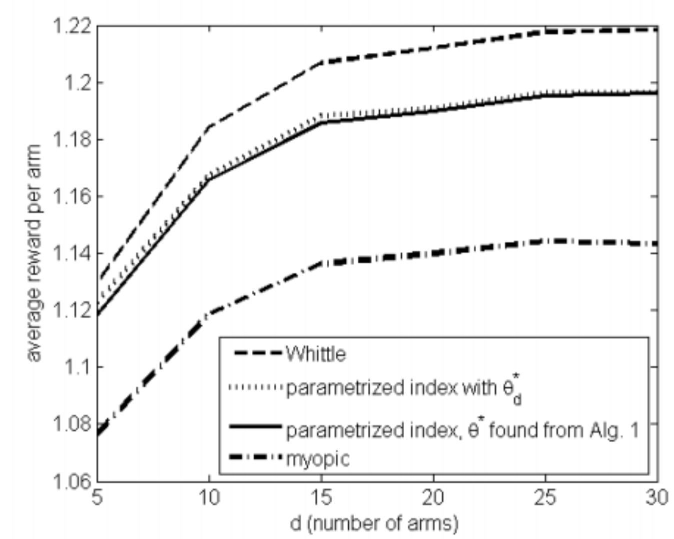
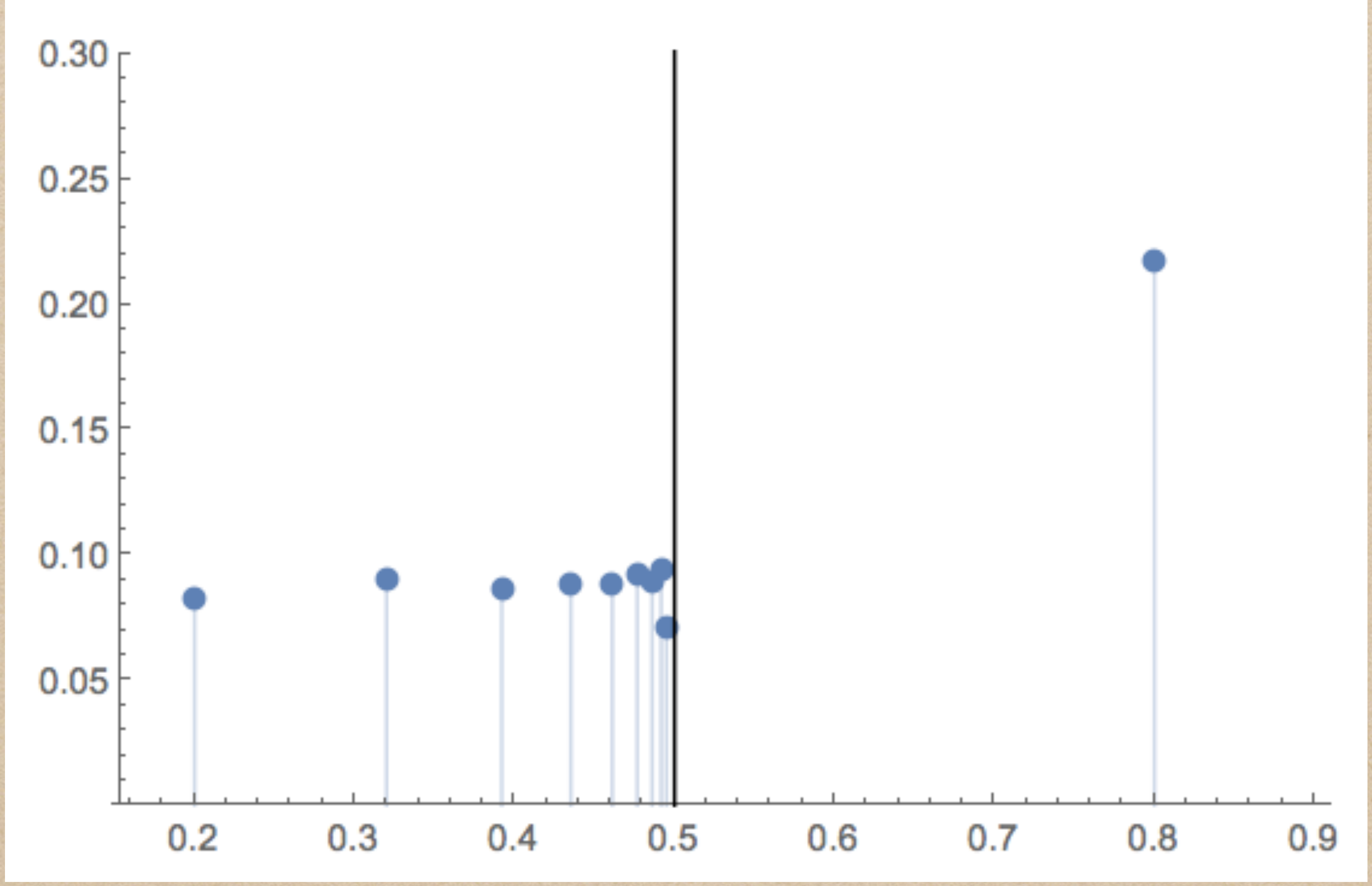


Figure 4: Comparison of average rewards achieved per arm under the Whittle, the parametric index (9) and the myopic policy. The parameter θ is found by optimizing (i) the problem with d arms (dotted), and (ii) the one-armed problem. $\varphi = 0.9$, $\sigma = 2$, $\rho = 0.4$, $T = 100,000$.

Measure Valued Asymptotics

For GE Channels - Promising



$$p, q = 0.2, r = 0.3, d = 10^4$$