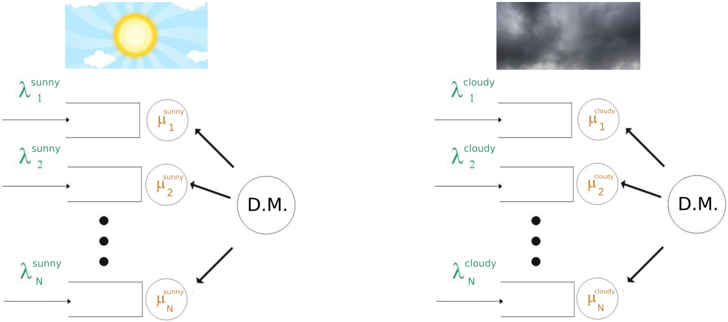
On the Whittle index of Markov Modulated Restless Bandits

Urtzi AYESTA

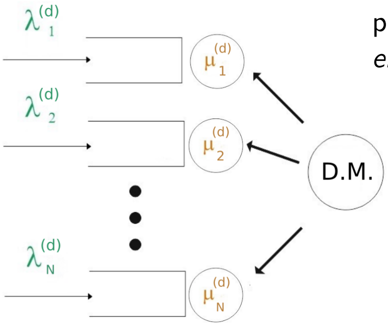
Grenoble, Novembre 21 2023

Resource allocation with time fluctuations



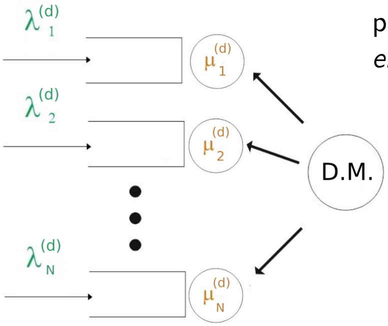
Motivating examples: cloud computing with varying arrivals, wireless downlink channels with changing quality, etc.

Control under changing conditions



The transition rates of the processes depend on *an environment* $D(t) = d$.

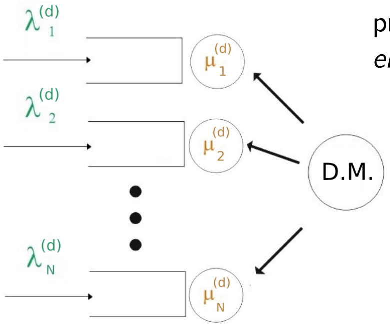
Control under changing conditions



The transition rates of the processes depend on *an environment* $D(t) = d$.

Goal: find control to optimise performance.

Control under changing conditions



The transition rates of the processes depend on *an environment* $D(t) = d$.

- **Problem 1**: unobservable environments.
- **Problem 2**: observable environments.

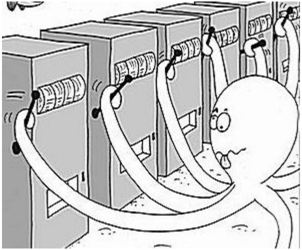
Goal: find control to optimise performance.

Outline

- MARB Problems
  - Classical multi-armed bandit
  - MARBP with environments
- Observable environments
  - Algorithm
  - Abandonment queue
  - Simulations
- Unobservable environment
  - Asymptotic optimality
  - Averaged Whittle's index

Classical multi-armed bandits

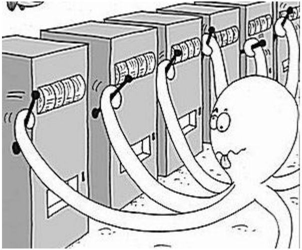
N arms or bandits, $R < N$ can be played.



Classical multi-armed bandits

N arms or bandits, $R < N$ can be played.

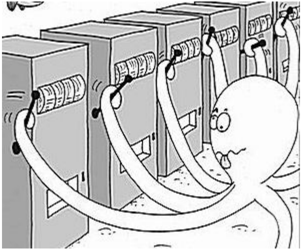
- State of bandit k: $M_k(t)$.
- Actions: $A_k(t) = 1$ (active) or $A_k(t) = 0$ (passive).
- Exponential rates $q_k(m'|m, a)$.



Classical multi-armed bandits

N arms or bandits, $R < N$ can be played.

- State of bandit k: $M_k(t)$.
- Actions: $A_k(t) = 1$ (active) or $A_k(t) = 0$ (passive).
- Exponential rates $q_k(m'|m, a)$. If $q_k(m'|m, 0) > 0$: *Restless* model.



Classical multi-armed bandits

For policy φ, the expected cost is given by:

$$V^{N,\varphi} = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^{N} C(M_k^{\varphi}(t), A_k^{\varphi}(t))dt\right),$$

where $C(m, a)$ is unit cost in state m under action a.

Classical multi-armed bandits

For policy φ, the expected cost is given by:

$$V^{N,\varphi} = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^{N} C(M_k^{\varphi}(t), A_k^{\varphi}(t))dt\right),$$

where $C(m,a)$ is unit cost in state m under action a.

Objective: find policy that minimises $V^{N,\varphi}$, subject to

$$\sum_{k=1}^{N} A_k^{\varphi}(t) = R \quad \forall t.$$

Relaxed version

Whittle ('88): relaxed constraint

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^{N} A_k^{\varphi}(t)dt\right) = R.$$

Relaxed version

Whittle ('88): relaxed constraint

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^{N} A_k^{\varphi}(t)dt\right) = R.$$

Lagrangians Multipliers approach $\implies$ find φ that minimises

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^{N} C(M_k^{\varphi}(t), A_k^{\varphi}(t)) - W\sum_{k=1}^{N} A_k^{\varphi}(t)dt\right).$$

Relaxed version

Whittle ('88): relaxed constraint

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^{N} A_k^{\varphi}(t)dt\right) = R.$$

Lagrangians Multipliers approach $\implies$ find φ that minimises

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^{N} C(M_k^{\varphi}(t), A_k^{\varphi}(t)) - W\sum_{k=1}^{N} A_k^{\varphi}(t)dt\right).$$

Reduces to solving N 1-dim subproblems

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T C(M_k^{\varphi}(t), A_k^{\varphi}(t)) - WA_k^{\varphi}(t)dt\right).$$

Whittle's Index

Definition

A bandit is *indexable* if for each m there exists a $W(m)$ such that

- If $W \leq W(m)$ active is optimal.
- If $W \geq W(m)$ passive is optimal.

$W(m)$ is the Whittle index for state m.

Whittle's Index

Definition

A bandit is *indexable* if for each m there exists a $W(m)$ such that

- If $W \leq W(m)$ active is optimal.
- If $W \geq W(m)$ passive is optimal.

$W(m)$ is the Whittle index for state m.

- Relaxed problem $\Longrightarrow$ optimal solution.
  Activate all bandits in state m such that $W \leq W(m)$.

Whittle's Index

Definition

A bandit is *indexable* if for each m there exists a $W(m)$ such that

- If $W \leq W(m)$ active is optimal.
- If $W \geq W(m)$ passive is optimal.

$W(m)$ is the Whittle index for state m.

- Relaxed problem $\implies$ optimal solution.
  Activate all bandits in state m such that $W \leq W(m)$.
- Original problem
  - with N fixed $\implies$ heuristic with high performance.
  - with $N \to \infty$ $\implies$ *asymptotically optimal.*

Summary

		Speed of the environment		
		Slow	Normal	Fast
Unob.	Gen	*Belief states*	*Belief States*	Averaged Whittle's index
	Aban.			$\overline{\mu}_k/\overline{\theta}_k$
Ob.	Gen.	$W_k^{(d)}(m)$	Algorithm	Algorithm
	Aban.	$\mu_k^{(d)}/\theta_k^{(d)}$	WI	$\mu_k^{(d)}/\overline{\theta}_k$

MARBP with environments

Bandit k has 2 processes:

$M_k^{\varphi}(t)$ controllable process $\Longrightarrow$ controlled by decision maker.

$D_k(t)$ environment process $\Longrightarrow$ exogenous and ergodic.

$\phi_k(d)$ stationary measure of $D_k(t)$.

$(D_k(t))_{k=1}^{N}$ may be correlated or not.

MARBP with environments

Bandit k has 2 processes:

$M_k^{\varphi}(t)$ controllable process $\implies$ controlled by decision maker.

$D_k(t)$ environment process $\implies$ exogenous and ergodic.

$\phi_k(d)$ stationary measure of $D_k(t)$.

$(D_k(t))_{k=1}^{N}$ may be correlated or not.

When $D_k(t) = d$,

- transition rates of controllable process : $q_k^{(d)}(m'|m, a)$.
- cost : $C_k^{(d)}(m, a)$.

MARBP with observable environments

The **decision maker** sees the current state of the bandit:

$$(M^{\varphi}(t), D(t)) = (m, d).$$

Definition (Threshold policies)

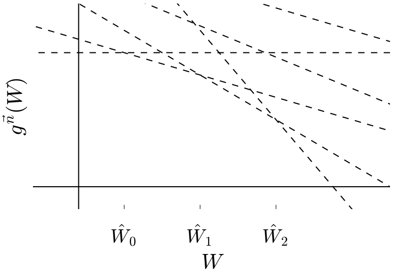
Threshold policy $\vec{n} = (n_1, n_2, \ldots)$ serves bandit iff current state (m, d) satisfies $m > n_d$.

We assume optimality of threshold policies, whose cost is given by

$$g^{\vec{n}}(W) := \sum_{d \in \mathcal{Z}} \sum_{m=0}^{\infty} C^{(d)}(m, a)\, \pi^{\vec{n}}(m, d) - W \sum_{d \in \mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m, d).$$

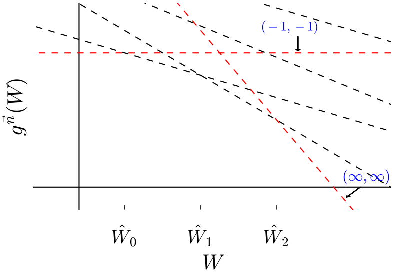
Example

$$g^{\vec{n}}(W) := \sum_{d \in \mathcal{Z}} \sum_{m=0}^{\infty} C^{(d)}(m, a)\, \pi^{\vec{n}}(m, d) - W \sum_{d \in \mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m, d).$$



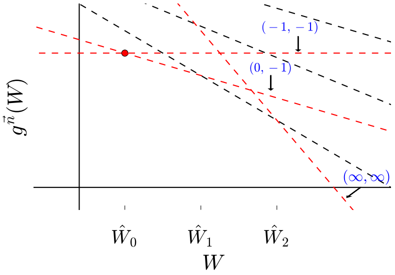
Example

$$g^{\vec{n}}(W) := \sum_{d\in\mathcal{Z}} \sum_{m=0}^{\infty} C^{(d)}(m,a)\,\pi^{\vec{n}}(m,d) - W \sum_{d\in\mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d).$$



Example

$$g^{\vec{n}}(W) := \sum_{d\in\mathcal{Z}} \sum_{m=0}^{\infty} C^{(d)}\,(m,a)\,\pi^{\vec{n}}(m,d) - W \sum_{d\in\mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d).$$

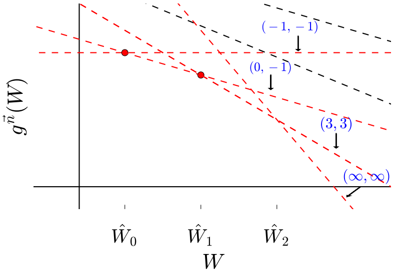


If $W \leq \hat{W}_0$, active is optimal in state $(m,d) = (0,1)$.
If $W \geq \hat{W}_0$, passive is optimal in state $(m,d) = (0,1)$.

$$\implies W(0,1) = \hat{W}_0$$

Example

$$g^{\vec{n}}(W) := \sum_{d \in \mathcal{Z}} \sum_{m=0}^{\infty} C^{(d)}(m, a)\, \pi^{\vec{n}}(m, d) - W \sum_{d \in \mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m, d).$$

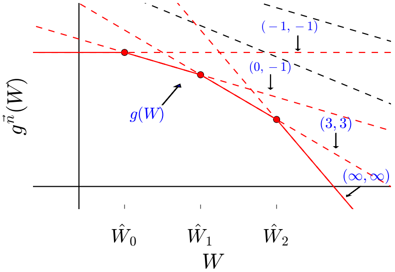


If $W \leq \hat{W}_1$, active is optimal in state $(m, d) = (3, 2)$.
If $W \geq \hat{W}_1$, passive is optimal in state $(m, d) = (3, 2)$.

$$\implies W(3, 2) = \hat{W}_1$$

Example

$$g^{\vec{n}}(W) := \sum_{d\in\mathcal{Z}} \sum_{m=0}^{\infty} C^{(d)}(m,a)\,\pi^{\vec{n}}(m,d) - W \sum_{d\in\mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d).$$



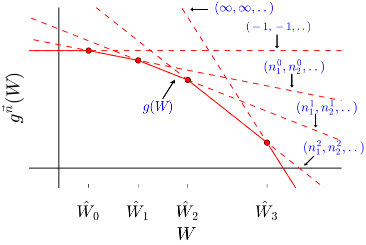
If $W \leq \hat{W}_1$, active is optimal in state $(m,d) = (3,2)$.
If $W \geq \hat{W}_1$, passive is optimal in state $(m,d) = (3,2)$.

$$\implies W(3,2) = \hat{W}_1$$

Algorithm

$\overline{W}(\vec{n}, \vec{n}')$: crossing point between $g^{\vec{n}}(W)$ and $g^{\vec{n}'}(W)$.

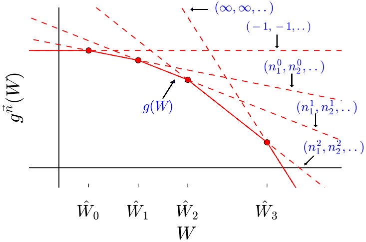
$$\overline{W}(\vec{n}, \vec{n}') = \frac{\sum_{d \in \mathcal{Z}} \sum_{m=0}^{\infty} C(m,d,a)\, \pi^{\vec{n}}(m,d) - \sum_{d \in \mathcal{Z}} \sum_{m=0}^{\infty} C(m,d,a)\, \pi^{\vec{n}'}(m,d)}{\sum_{d \in \mathcal{Z}} \sum_{m=0}^{n_d} \pi^{\vec{n}}(m,d) - \sum_{d \in \mathcal{Z}} \sum_{m=0}^{n'_d} \pi^{\vec{n}'}(m,d)}.$$



Algorithm

$\overline{W}(\vec{n},\vec{n}')$: crossing point between $g^{\vec{n}}(W)$ and $g^{\vec{n}'}(W)$.

$$\overline{W}(\vec{n},\vec{n}') = \frac{\sum_{d\in\mathcal{Z}}\sum_{m=0}^{\infty} C(m,d,a)\,\pi^{\vec{n}}(m,d) - \sum_{d\in\mathcal{Z}}\sum_{m=0}^{\infty} C(m,d,a)\,\pi^{\vec{n}'}(m,d)}{\sum_{d\in\mathcal{Z}}\sum_{m=0}^{n_d}\pi^{\vec{n}}(m,d) - \sum_{d\in\mathcal{Z}}\sum_{m=0}^{n'_d}\pi^{\vec{n}'}(m,d)}.$$

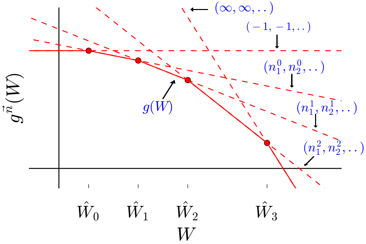


$\vec{n}^{-1} := (-1,-1,\ldots)$. Then, for $j \geq 0$,

Algorithm

$\overline{W}(\vec{n},\vec{n}')$: crossing point between $g^{\vec{n}}(W)$ and $g^{\vec{n}'}(W)$.

$$\overline{W}(\vec{n},\vec{n}') = \frac{\sum_{d\in\mathcal{Z}}\sum_{m=0}^{\infty} C(m,d,a)\,\pi^{\vec{n}}(m,d) - \sum_{d\in\mathcal{Z}}\sum_{m=0}^{\infty} C(m,d,a)\,\pi^{\vec{n}'}(m,d)}{\sum_{d\in\mathcal{Z}}\sum_{m=0}^{n_d}\pi^{\vec{n}}(m,d) - \sum_{d\in\mathcal{Z}}\sum_{m=0}^{n'_d}\pi^{\vec{n}'}(m,d)}.$$



$\vec{n}^{-1} := (-1,-1,\ldots)$. Then, for $j \geq 0$,

Step j $\hat{W}_j = \inf_{n_d \geq n_d^{j-1}\ \forall d} \overline{W}(\vec{n},\vec{n}^{j-1})$.

$\vec{n}^j$: minimiser. $W(m,d) := \hat{W}_j$ for $n_d^{j-1} < m \leq n_d^j$, $\forall d$.

Go to step $j+1$.

Slowly changing environment

Let the transitions of the environment be $\beta r^{dd'}$. Then it holds that:

$$\lim_{\beta \to 0} \pi^{\vec{n}}(m, d) = \phi(d) p^{n_d,(d)}(m)$$

Proposition

$$\lim_{\beta \to 0} W(m, d) = W^d(m)$$

Queue with abandonments

Assume $|\mathcal{Z}| = 2$ *and* $C^{(d)}(m, a) = cm$.

- $(m, d) \to (m + 1, d)$ at rate $\lambda^{(d)}$.
- $(m, d) \to (m - 1, d)$ at rate $m\theta^{(d)} + a\mu^{(d)}$.
- $(m, d) \to (m, 3 - d)$ at rate $r^{(d)}$.

Threshold policies

Proposition

For each W, there exists an $\vec{n}(W) = (n_1(W), n_2(W))$ such that $\vec{n}(W)$ is an optimal solution

Truncate at L and smooth arrivals, an invoke S. Bhulai et al, QUESTA 2014

An auxiliary result:

$$\lambda^{(d)}\phi(d) + r^{(3-d)}\mathbb{E}\left(M^{\vec{n}}\mathbf{1}_{(D=3-d)}\right)$$
$$= \left(\theta^{(d)} + r^{(d)}\right)\mathbb{E}\left(M^{\vec{n}}\mathbf{1}_{(D=d)}\right) + \mu^{(d)}\sum_{m=n_d+1}^{\infty}\pi^{\vec{n}}(m,d),$$

Indexability and Main results

$$W^{(d)} := c\mu^{(d)} \frac{\theta^{(3-d)} + r^{(1)} + r^{(2)}}{\theta^{(1)}\theta^{(2)} + r^{(1)}\theta^{(2)} + r^{(2)}\theta^{(1)}}.$$

Proposition (Whittle's index)

Assume $W^{(1)} < W^{(2)}$.

$$W(m,d) = \begin{cases} \overline{W}((m-1,0),(m,0)) & \text{for } d = 1 \\ W^{(2)} & \text{for } d = 2, m \geq 1. \end{cases}$$

Moreover, $W(m,1) \leq W^{(1)} \leq W^{(2)}$ for all m.

Slow and Fast environments

Proposition

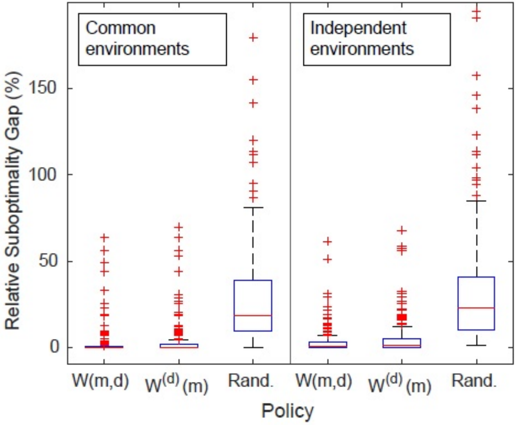
Scale the rates of environment as $\beta r^{(d)}$.
It holds that

$$\lim_{\beta \to 0} W(m,d) = c\frac{\mu^{(d)}}{\theta^{(d)}}, \quad \forall m, d.$$

$$\lim_{\beta \to \infty} W(m,d) = c\frac{\mu^{(2)}}{\overline{\theta}} \quad \text{for} \quad d = 2,$$

where $\overline{\theta} := \sum_{d=1}^{2} \phi(d)\theta^{(d)}$.

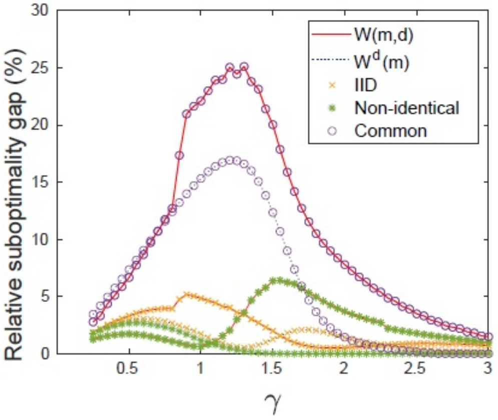
Numerical evaluation



$W(m, d)$: observable WI, $\overline{W}(m)$: averaged WI.

Numerical evaluation

$W_2^{(1)} << W_1^{(1)}$ and $W_1^{(2)} < W_2^{(1)}$



MARBP with unobservable environments

The **decision maker** does not see the current state of $D(t)$.

MARBP with unobservable environments

The **decision maker** does not see the current state of $D(t)$.

Transitions of bandits : $\frac{1}{N} q^{(d)}(m'|m, a)$.
$N \to \infty \implies$ **Both number of bandits and speed of the environment are scaled**.

Definition

Policy φ^* asymptotically optimal: for any other policy φ,

$$\liminf_{N\to\infty} V^{N,\varphi} \geq \liminf_{N\to\infty} V^{N,\varphi^*}.$$

Objective: show asymptotic optimality of a set of policies.

Averaged Whittle's index policy

- Modulated process: parameters $C^{(d)}$ and $q^{(d)}$.
- Unmodulated process with averaged parameters:

$$\overline{C}(m,a) = \sum_d \phi(d) C^{(d)}(m,a)$$

$$\overline{q}\,(m'|m,a) = \sum_d \phi(d) q^{(d)}\,(m'|m,a)$$

Averaged Whittle's index policy

- Modulated process: parameters $C^{(d)}$ and $q^{(d)}$.
- Unmodulated process with averaged parameters:

$$\overline{C}(m,a) = \sum_d \phi(d) C^{(d)}(m,a)$$

$$\overline{q}\,(m'|m,a) = \sum_d \phi(d) q^{(d)}\,(m'|m,a)$$

Definition (Averaged Whittle Index)

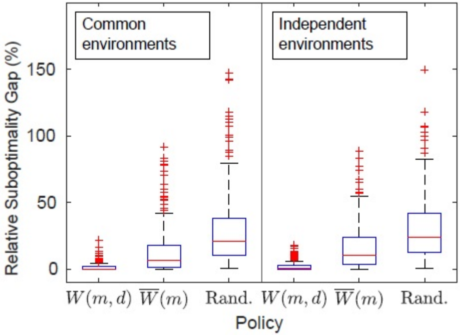
$\overline{W}(m)$ is the Whittle Index obtained for the restless bandit model with parameters $\overline{C}$ and $\overline{q}$.

Theorem

$\overline{W}(m)$ *policy is included in the set of asymptotically optimal policies* Φ^* *defined before.*

Performance of Averaged WI

Averaged index: $\bar{W}_k(m) = c_k \frac{\bar{\theta}_k}{\bar{\mu}_k}$



Unobservable: finite speed

Modulation ON,OFF
Properties of belief states
Non-preemptive:

Proposition

Serving the class i with $\max_k \frac{1}{E(S_i(1-p_i))} = \mu_i \frac{q_i}{p_i+q_i}$ maximizes the throughput with positive correlation.

Preemptive:

Proposition

Serving the class with $\max_k \mu_k \pi_k$ maximizes the throughput with positive correlation.

Characterization of performance loss due to unobservability.

Summary

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="3">Speed of the environment</td></tr>
<tr><td>Slow</td><td>Normal</td><td>Fast</td></tr>
<tr><td rowspan="2">Unob.</td><td>Gen</td><td rowspan="2">Belief states</td><td rowspan="2">Belief States</td><td>Averaged Whittle's index</td></tr>
<tr><td>Aban.</td><td>$\overline{\mu}_k/\overline{\theta}_k$</td></tr>
<tr><td rowspan="2">Ob.</td><td>Gen.</td><td>$W_k^{(d)}(m)$</td><td>Algorithm</td><td>Algorithm</td></tr>
<tr><td>Aban.</td><td>$\mu_k^{(d)}/\theta_k^{(d)}$</td><td>WI</td><td>$\mu_k^{(d)}/\overline{\theta}_k$</td></tr>
</table>

Conclusions and open problems

- Observable environments
  - Common environment: Indices in slow and fast environments, asymptotic optimality etc.
- Unobservable environments
  - Calculation of WI on important classes of problems
- Stability of WI

Thank you