

A Class of Restless Bandit Problems: Indexability and Optimality of Whittle's Index

Qing Zhao

Department of Electrical and Computer Engineering
University of California at Davis




Supported by NSF, ARL, ARO.

History

Clinical Trial (Thompson '33)



Web Search (*Radlinski&etal'08*)


[Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more](#)

[Advanced Search](#) [Source Preferences](#) [Source Info](#)

Scholar All articles - Recent articles Results 1 - 10 of about 2,260 for [multi-armed bandit](#). (0.24 seconds)

Multi-armed bandit allocation indices - UC-eLinks
 Z Ying - Technometrics, 1991 - JSTOR
 Cited by 409 - Related articles - Web Search - Library Search

Why Imitate, and If So, How? A Bounded Rational Approach to Multi Armed Bandits
 KH Soltau - 1996 - all.repec.org
 ... multi-armed bandit setting. Individuals must repeatedly choose an action ... rule) whenever she is faced with a multi-armed bandit with the same set of actions. ...
 Cited by 279 - Related articles - View as HTML - Web Search - UC-eLinks - Library Search - BL Direct - All 15 versions

Multi-armed bandits and the Gittins index
 P Whittle - Journal of the Royal Statistical Society, Series B (...), 1980 - jstor.org
 ... bandit processes; dynamic allocation indices; two-armed bandit problem; optimal resource allocation 1. Introduction The multi-armed bandit problem: as it has ...
 Cited by 165 - Related articles - Web Search - UC-eLinks - All 4 versions

Extensions of the multiarmed bandit problem: The discounted case - UC-eLinks
 P Varaiya, J Walrand, C Buysukkoo - IEEE Transactions on Automatic Control, 1985 - ieeexplore.ieee.org
 ... Extensions of the multiarmed bandit problem: The discounted case Varaiya, P, Walrand, J, Buysukkoo, C, University of California, Berkeley, CA, USA; ...
 Cited by 141 - Related articles - Web Search - Library Search - All 4 versions

The nonstochastic multiarmed bandit problem - UC-eLinks
 P Auer, N Cesa-Bianchi, Y Freund, RE Schapire - SIAM Journal on Computing, 2003 - homes.cs.uminn.edu
 Page 1. THE NONSTOCHASTIC MULTIARMED BANDIT PROBLEM * PETER ... att.com), 48 Page 2. THE NONSTOCHASTIC MULTIARMED BANDIT PROBLEM 49 trying ...
 Cited by 148 - Related articles - View as HTML - Web Search - BL Direct - All 28 versions

Conservation laws, extended polymatroids and multiarmed bandit problems: a polyhedral approach to ... - mit.edu
 D Bertsimas, J Nino-Mora - Mathematics of Operations Research, 1995 - jstor.org
 ... CONSERVATION LAWS, EXTENDED POLYMATROIDS AND MULTIARMED BANDIT PROBLEMS; A POLYHEDRAL APPROACH TO INDEXABLE SYSTEMS ... The multiarmed bandit problem ...
 Cited by 109 - Related articles - Web Search - UC-eLinks - Library Search - BL Direct - All 8 versions

On the Gittins index for multiarmed bandits
 R Weber - The Annals of Applied Probability, 1992 - jstor.org
 ... 2, No. 4, 1024-1033 ON THE GITTINS INDEX FOR MULTIARMED BANDITS By Richard Weber
 University of Cambridge This paper considers the multiarmed bandit problem and ...
 Cited by 58 - Related articles - Web Search - UC-eLinks - All 4 versions

The multi-armed bandit problem: decomposition and computation
 MN Katehakis, AF Veinott Jr - Mathematics of Operations Research, 1987 - jstor.org
 ... THE MULTI-ARMED BANDIT PROBLEM: DECOMPOSITION AND COMPUTATION ... The multi-armed bandit problem arises in sequentially allocating effort to one of N projects ...
 Cited by 80 - Related articles - Web Search - UC-eLinks - All 7 versions

Adaptive treatment allocation and the multi-armed bandit problem
 TL Lai - The Annals of Statistics, 1987 - jstor.org
 ... 3, 1091-1114 ADAPTIVE TREATMENT ALLOCATION AND THE MULTI-ARMED BANDIT PROBLEM By Tze Leung Lai Columbia University A class of simple adaptive allocation rules ...
 Cited by 80 - Related articles - Web Search - UC-eLinks - All 6 versions

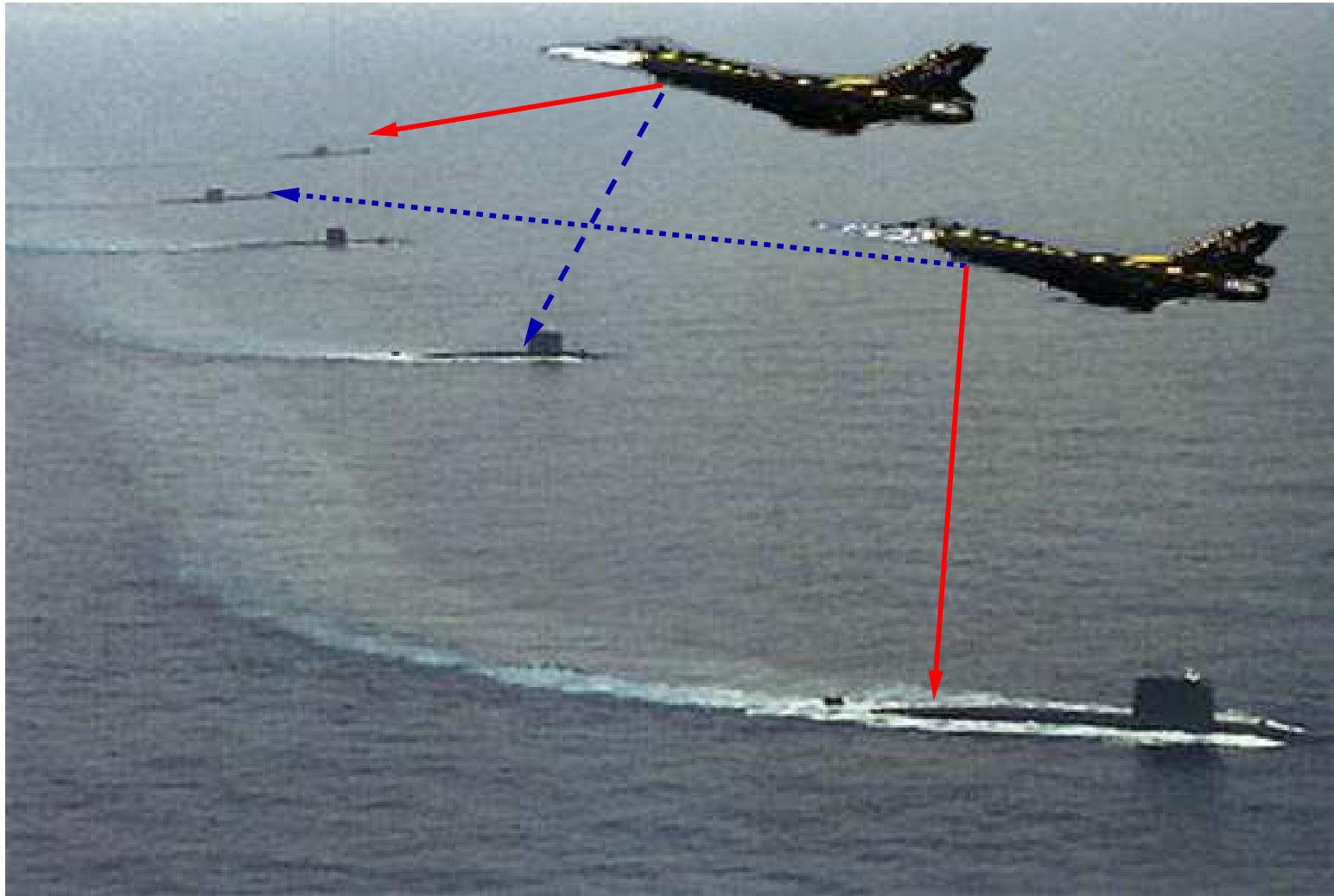
Restless bandits: Activity allocation in a changing world
 P Whittle - Journal of applied probability, 1988 - jstor.org
 ... GITTINS INDEX; MULTIARMED BANDITS; SEQUENTIAL SCHEDULING; STIMULATING PRICES; ... The multi-armed bandit problem is a classic version of the problem of ...
 Cited by 165 - Related articles - Web Search - UC-eLinks - All 2 versions

Key authors: [J Gittins](#) - [P Auer](#) - [M Indices](#) - [Z Ying](#) - [P Whittle](#)

Playing Golf with Multiple Balls *(Dumitriu&etal'03)*

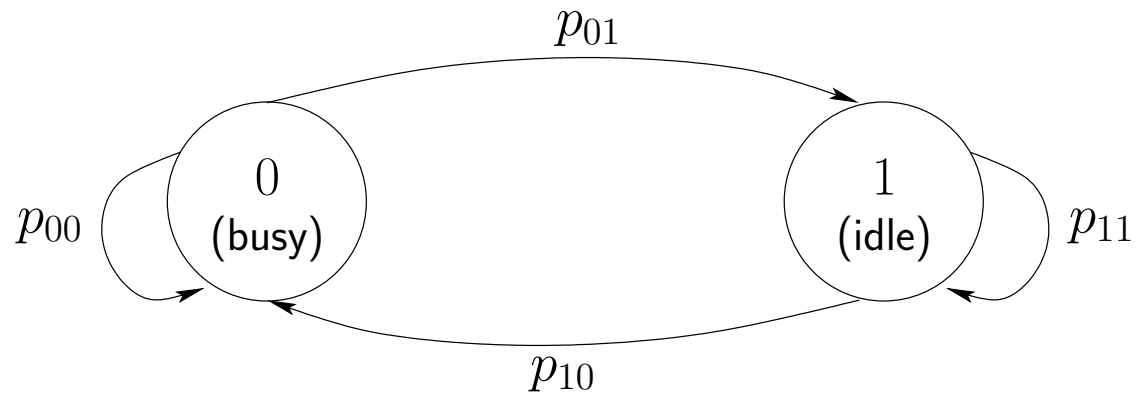
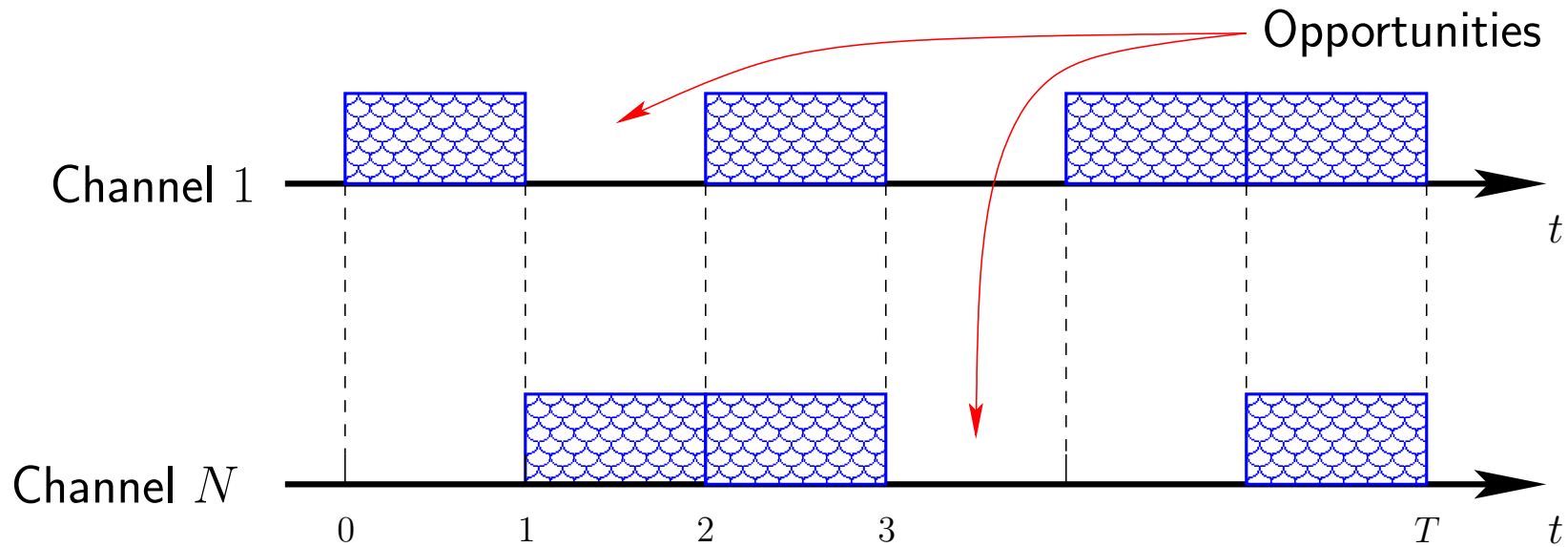


Multi-Agent System (*Whittle'88, LeNy&etal'08*)



Cognitive Radio (*Zhao&etal'05*)

- ▶ N independent Gilbert-Elliot channels:



Information vs. Immediate Payoff

“Bandit problems embody in essential form a conflict evident in all human action: information versus immediate payoff.”

— *P. Whittle (1989)*

Information vs. Immediate Payoff

A Two-Armed Bandit:

- ▶ Two coins with unknown bias θ_1, θ_2 .
- ▶ Head: reward = 1; Tail: reward = 0.
- ▶ Objective: maximize total reward over n flips.

An Example (Berry&Fristedt'85):

- $\theta_1 = \frac{1}{2}, \theta_2 = \begin{cases} 1, & \text{with probability } \frac{1}{4} \\ 0, & \text{with probability } \frac{3}{4} \end{cases}$
- To gain immediate payoff: flip Coin 1 indefinitely.
- To gain information: flip Coin 2 initially.

Non-Bayesian Formulation

- ▶ (θ_1, θ_2) are treated as unknown deterministic parameters.
- ▶ $V_n^\pi(\theta_1, \theta_2)$: total reward of policy π .
- ▶ $n \underbrace{\max\{\theta_1, \theta_2\}}_{\theta_{max}}$: total reward if (θ_1, θ_2) were known.
- ▶ The cost of learning (regret):

$$C_n^\pi(\theta_1, \theta_2) \triangleq n\theta_{max} - V_n^\pi(\theta_1, \theta_2) = (\theta_{max} - \theta_{min})\mathbb{E}[\text{time spent on } \theta_{min}]$$

- ▶ Objective: $\min_{\pi} C_n^\pi(\theta_1, \theta_2)$.

Classic Results under Non-Bayesian Formulation

▶ Lai&Robbins'85:

$$C_n^*(\theta_1, \theta_2) \sim \frac{\theta_{max} - \theta_{min}}{\underbrace{I(\theta_{min}, \theta_{max})}_{KL \text{ distance}}} \log n \quad \text{as } n \rightarrow \infty$$

▶ Anantharam&Varaiya&Walrand'87:

- extension from single play to multiple plays.
- extension from i.i.d to Markovian reward processes.

Classic and Recent Results under Non-Bayesian Formulation

▶ Lai&Robbins'85:

$$C_n^*(\theta_1, \theta_2) \sim \frac{\theta_{max} - \theta_{min}}{\underbrace{I(\theta_{min}, \theta_{max})}_{KL \text{ distance}}} \log n \quad \text{as } n \rightarrow \infty$$

▶ Anantharam&Varaiya&Walrand'87:

- extension from single play to multiple plays.
- extension from i.i.d to Markovian reward processes.

▶ Liu&Zhao'09:

- extension to *distributed* multiple players
(*distributed decision-making using only local observations*).
- decentralized policy achieving *the same* $\log n$ order of the regret.
- fairness among players.

Bayesian Formulation

- ▶ (θ_1, θ_2) are random variables with *prior* distributions $(f_{\theta_1}, f_{\theta_2})$.
- ▶ Policy π : choose an arm based on $(f_{\theta_1}, f_{\theta_2})$ and the observation history.
- ▶ $R_\pi(t)$: the reward obtained at time t .
- ▶ The total discounted reward over an infinite horizon:

$$V_\pi(f_{\theta_1}, f_{\theta_2}) \triangleq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \beta^t R_\pi(t) \mid (f_{\theta_1}, f_{\theta_2}) \right], \quad (0 < \beta < 1)$$

- ▶ Objective: $\max_\pi V_\pi(f_{\theta_1}, f_{\theta_2})$.

Bandit and MDP

Multi-Armed Bandit as A Class of MDP: (Bellman'56)

- ▶ N *independent* arms with *fully observable* states $[Z_1(t), \dots, Z_N(t)]$.
- ▶ *One* arm is activated at each time.
- ▶ Active arm changes state (*known Markov process*); offers reward $R_i(Z_i(t))$.
- ▶ Passive arms are frozen and generate no reward.

Bandit and MDP

Multi-Armed Bandit as A Class of MDP: (Bellman'56)

- ▶ N *independent* arms with *fully observable* states $[Z_1(t), \dots, Z_N(t)]$.
- ▶ *One* arm is activated at each time.
- ▶ Active arm changes state (*known Markov process*); offers reward $R_i(Z_i(t))$.
- ▶ Passive arms are frozen and generate no reward.

Why is sampling processes with unknown distributions an MDP?

- The state of each arm is the *posterior* distribution $f_{\theta_i}(t)$ (*information state*).
- For an active arm, $f_{\theta_i}(t+1)$ is updated from $f_{\theta_i}(t)$ and the new observation.
- For a passive arm, $f_{\theta_i}(t+1) = f_{\theta_i}(t)$.

Bandit and MDP

Multi-Armed Bandit as A Class of MDP: (Bellman'56)

- ▶ N *independent* arms with *fully observable* states $[Z_1(t), \dots, Z_N(t)]$.
- ▶ *One* arm is activated at each time.
- ▶ Active arm changes state (*known Markov process*); offers reward $R_i(Z_i(t))$.
- ▶ Passive arms are frozen and generate no reward.

Solving Multi-Armed Bandit using Dynamic Programming:

- ▶ Exponential complexity with respect to N .

The Slow Propagation of the Breaking News

“A colleague of high repute asked an equally well-known colleague:

- *‘What would you say if you were told that the multi-armed bandit problem had been solved?’*
- *‘Sir, the multi-armed bandit problem is not of such a nature that it can be solved.’*

— *P. Whittle (1989)*

Gittins's Index

The Index Structure of the Optimal Policy: (*Gittins:1960's*)

- ▶ Assign each state of each arm a *priority* index.
- ▶ Activate the arm with highest current index value.

Complexity:

- ▶ Arm are strongly decomposable (1 N -dim to N 1-dim problems).
- ▶ *Linear* complexity with N .
- ▶ Polynomial (cubic) with the state space size of an *individual* arm (*Varaiya&Walrand&Buyukkoc'85, Katta&Sethuraman'04*).

Extensions: (*Varaiya&Walrand&Buyukkoc'85*)

- ▶ From Markovian to non-Markovian dynamics.

Restless Bandit

Restless Multi-Armed Bandit: (*Whittle'88*)

- ▶ Activate K arms simultaneously.
- ▶ Passive arms also change state and offer reward.

Structure of the Optimal Policy:

- ▶ Not yet found.

Complexity:

- ▶ PSPACE-hard (*Papadimitriou&Tsitsiklis'99*).

Whittle's Index

Whittle's Index: (*Whittle'88*)

- Provide a subsidy m for passivity whenever the arm is made passive.
- Whittle's index: the subsidy m that makes active and passive actions equally attractive at the current state.

Whittle's Index

Whittle's Index: (*Whittle'88*)

- Provide a subsidy m for passivity whenever the arm is made passive.
- Whittle's index: the subsidy m that makes active and passive actions equally attractive at the current state.

Performance:

- Optimal under relaxed constraint on the average number of active arms.
- Asymptotically optimal under certain conditions (*Weber&Weiss'90*).
- Near optimal performance observed from extensive numerical examples.

Whittle's Index

Whittle's Index: (*Whittle'88*)

- Provide a subsidy m for passivity whenever the arm is made passive.
- Whittle's index: the subsidy m that makes active and passive actions equally attractive at the current state.

Performance:

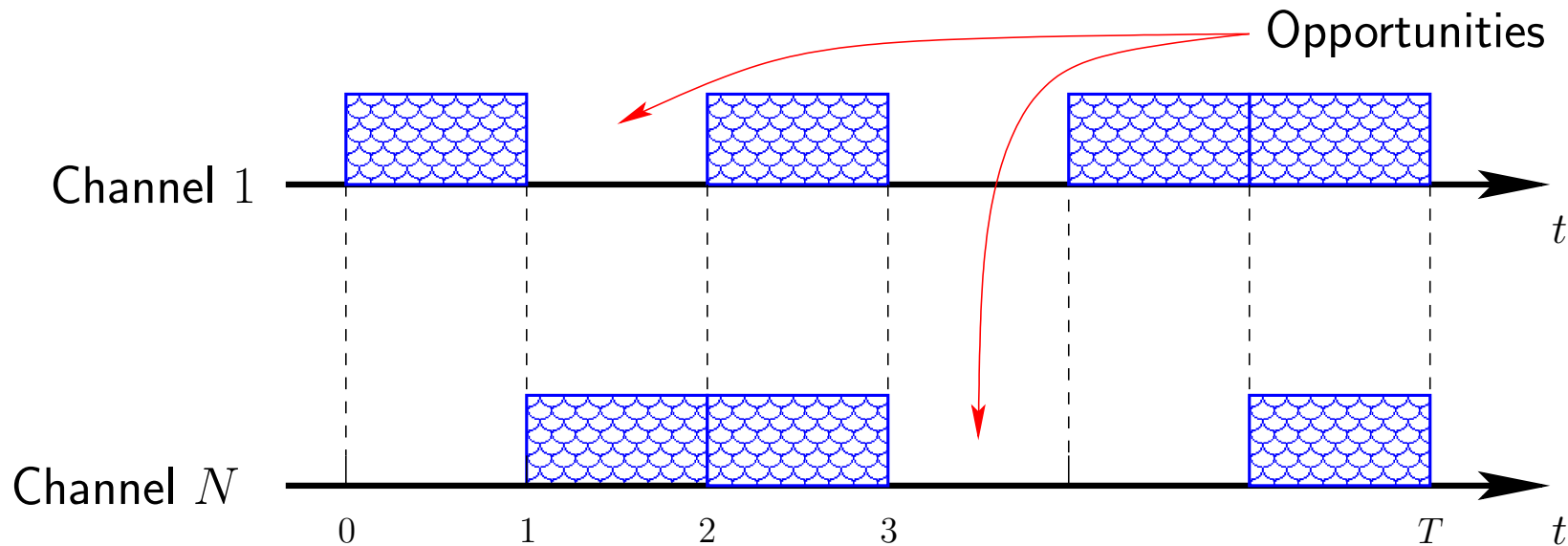
- Optimal under relaxed constraint on the average number of active arms.
- Asymptotically optimal under certain conditions (*Weber&Weiss'90*).
- Near optimal performance observed from extensive numerical examples.

Difficulties:

- Existence (indexability) not guaranteed and difficult to check.
- Numerical index computation infeasible for uncountable state space.
- Optimality in finite regime difficult to establish.

Main Results

Multichannel Dynamic Access

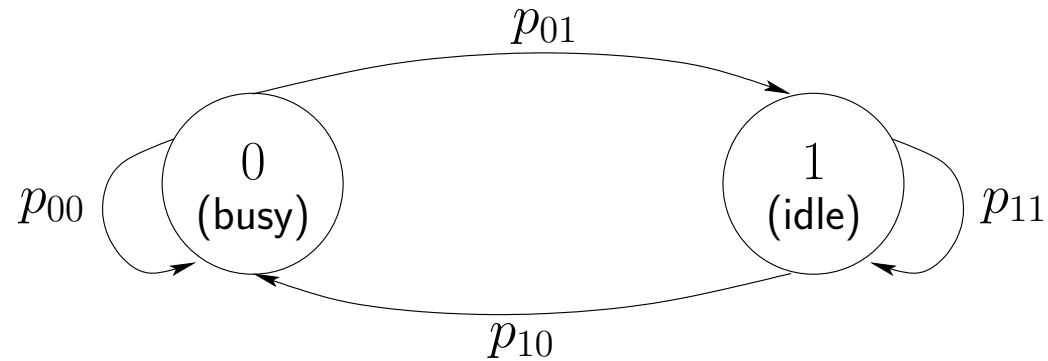


- ▶ Each channel is considered as an arm.
- ▶ State of arm i : *posterior* probability that channel i is idle.

$$\omega_i(t) = \Pr[\text{channel } i \text{ is idle in slot } t \mid \underbrace{O(1), \dots, O(t-1)}_{\text{observations}}]$$

- ▶ The expected immediate reward for activating arm i is $\omega_i(t)$

Markovian State Transition



- ▶ If channel i is activated in slot t :

$$\omega_i(t+1) = \begin{cases} p_{11}, & \text{if } O_i(t) = 1 \\ p_{01}, & \text{if } O_i(t) = 0 \end{cases}.$$

- ▶ If channel i is made passive in slot t :

$$\omega_i(t+1) = \omega_i(t)p_{11} + (1 - \omega_i(t))p_{01}.$$

Indexability

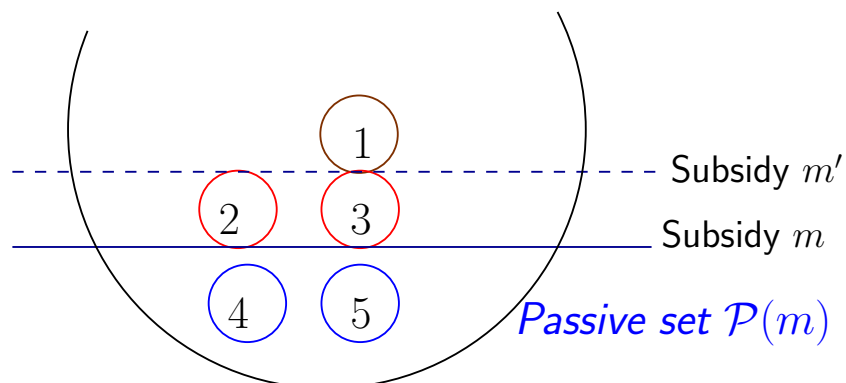
A Single-Armed Bandit with Subsidy:

- ▶ Being active at state ω : reward = ω .
- ▶ Being passive at any state: reward = m .

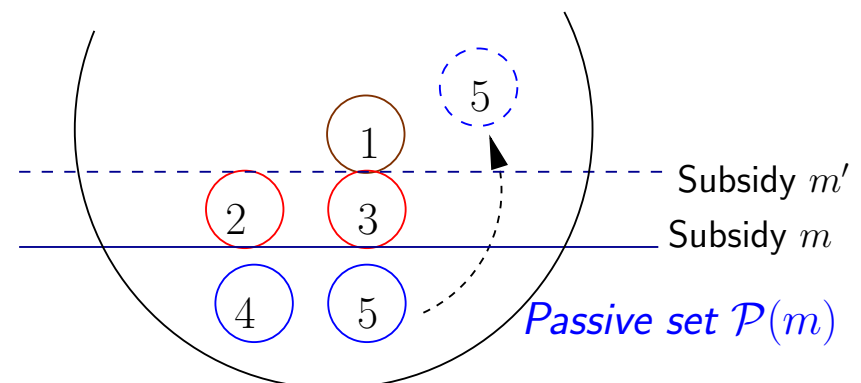
Indexability:

- ▶ $\mathcal{P}(m)$: the passive set under subsidy m .
- ▶ Indexability: $\mathcal{P}(m)$ increases monotonically as m increases from $-\infty$ to ∞ .

Indexable



Not Indexable



Proof for Indexability: Value Functions

A Single-Armed Bandit with Subsidy:

- ▶ Being active at state ω : reward = ω .
- ▶ Being passive at any state: reward = m .

Value Functions:

- ▶ $V_m(\omega)$: max total discounted reward starting from state ω

$$V_m(\omega) = \max\{V_m(\omega; \text{active}), V_m(\omega; \text{passive})\}$$

- ▶ $V_m(\omega; \text{active})$: max total discounted reward if active at ω

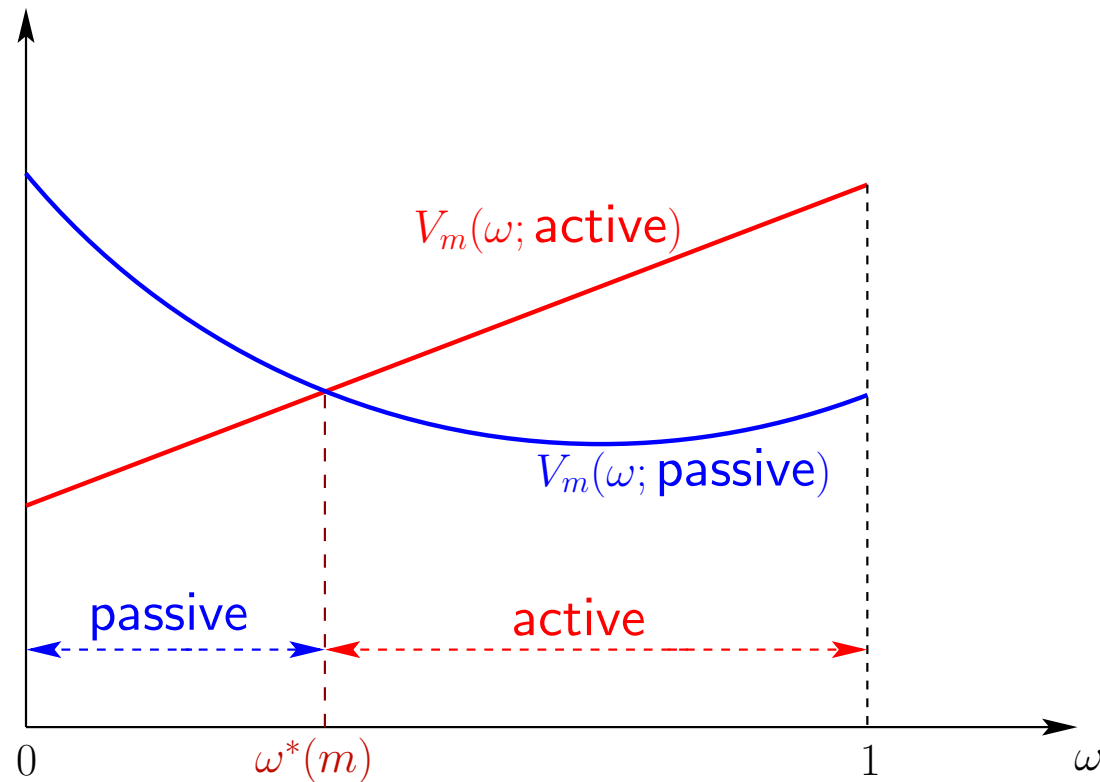
$$V_m(\omega; \text{active}) = \omega + \beta(\omega V_m(p_{11}) + (1 - \omega)V_m(p_{01})) \quad (\text{linear in } \omega)$$

- ▶ $V_m(\omega; \text{passive})$: max total discounted reward if passive at ω

$$V_m(\omega; \text{passive}) = m + \beta V_m(\underbrace{\omega p_{11} + (1 - \omega)p_{01}}_{\text{updated state}}) \quad (\text{convex in } \omega)$$

Proof for Indexability: Optimality of Threshold Policy

- ▶ A threshold policy is optimal for the single-armed bandit with subsidy.



- ▶ It thus suffices to prove the threshold $\omega^*(m) \nearrow$ with the subsidy m .

Proof for Indexability: Total Passive Time

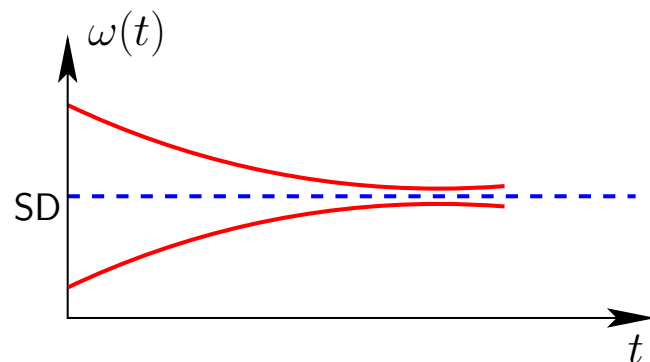
- A sufficient condition for $\omega^*(m) \nearrow$ with m :

$$\frac{d(V_m(\omega; \text{passive}))}{dm} \Big|_{\omega=\omega^*(m)} \geq \frac{d(V_m(\omega; \text{active}))}{dm} \Big|_{\omega=\omega^*(m)} \quad (\diamond)$$

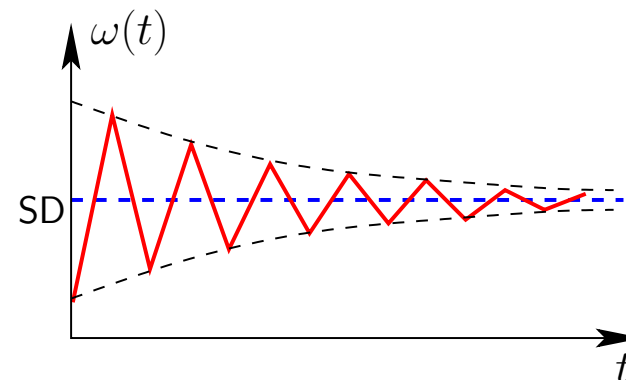
- (\diamond) is shown based on

- $\frac{d(V_m(\omega))}{dm} = \mathbb{E}[\text{total passive time}]$.
- properties of the *first crossing time*:

Positive Correlation:



Negative Correlation:

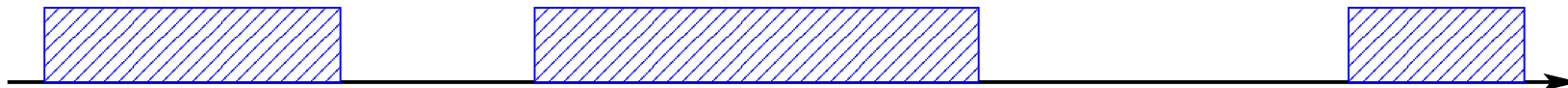


Structure of Whittle's Index Policy

The Semi-Universal Structure of Whittle's Index Policy:

- ▶ No need to compute the index.
- ▶ No need to know $\{p_{01}, p_{11}\}$ other than their order.

$p_{11} \geq p_{01}$ (*positive correlation*):

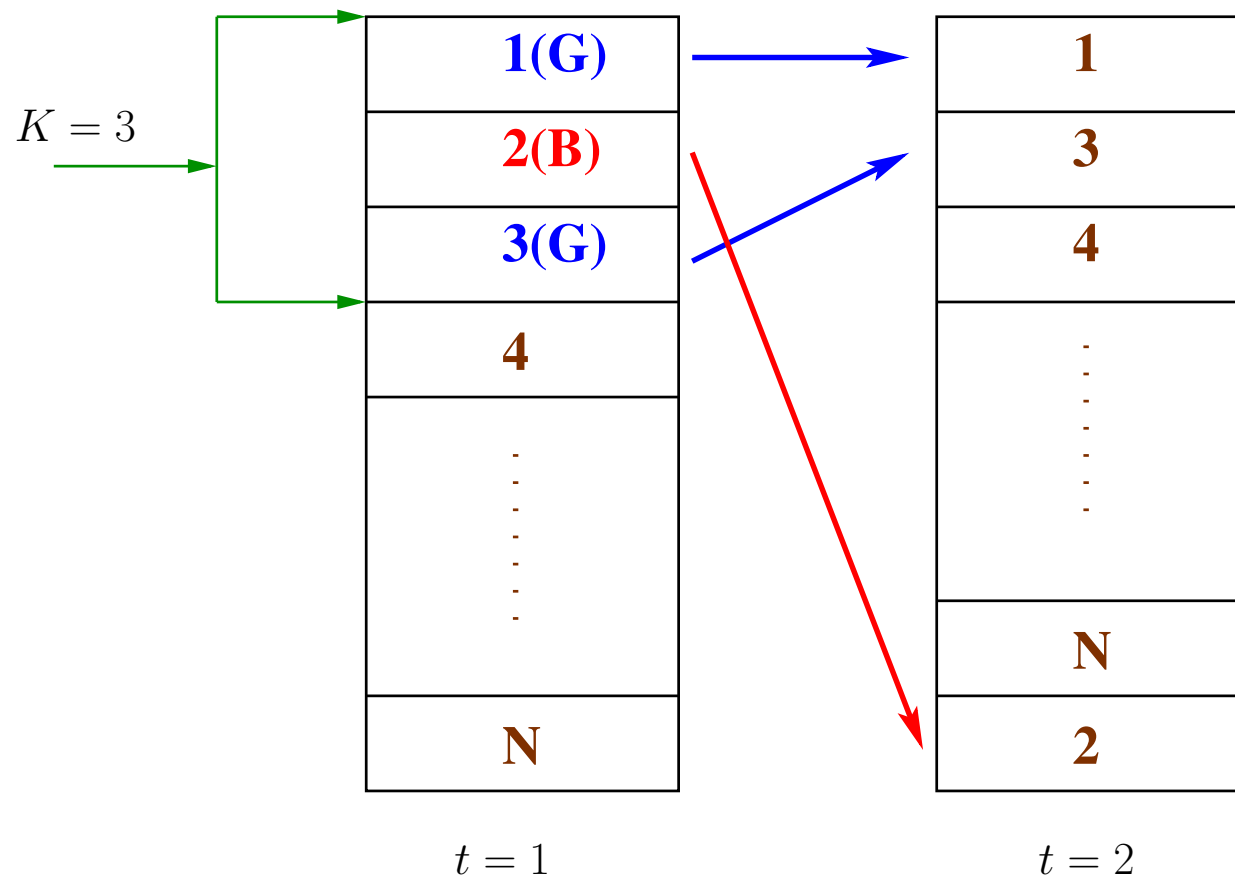


$p_{11} < p_{01}$ (*negative correlation*):



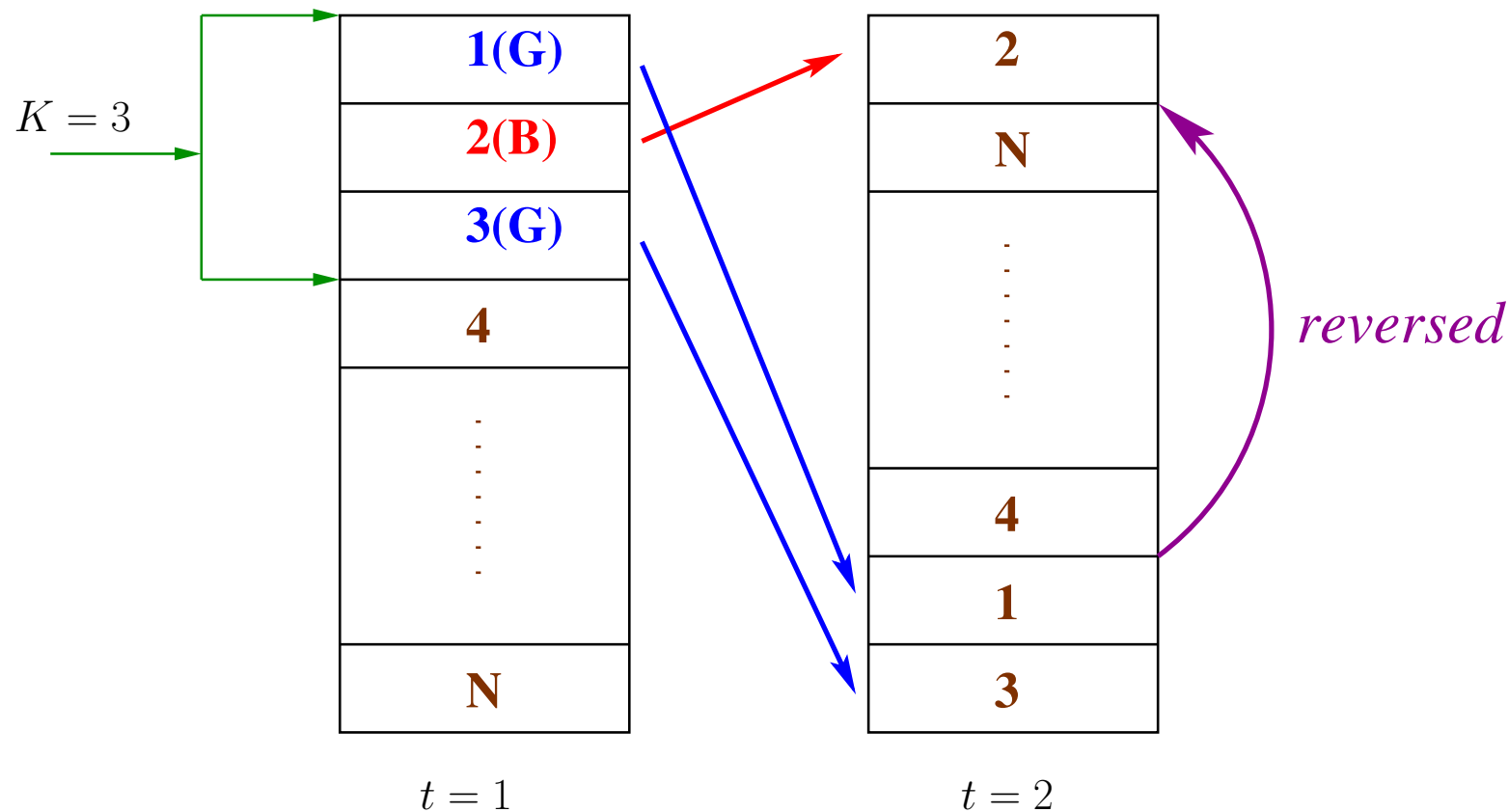
Structure of Whittle's Index Policy: Positive Correlation

- Stay with good (**G**) channels and leave bad (**B**) ones to the end of the queue.



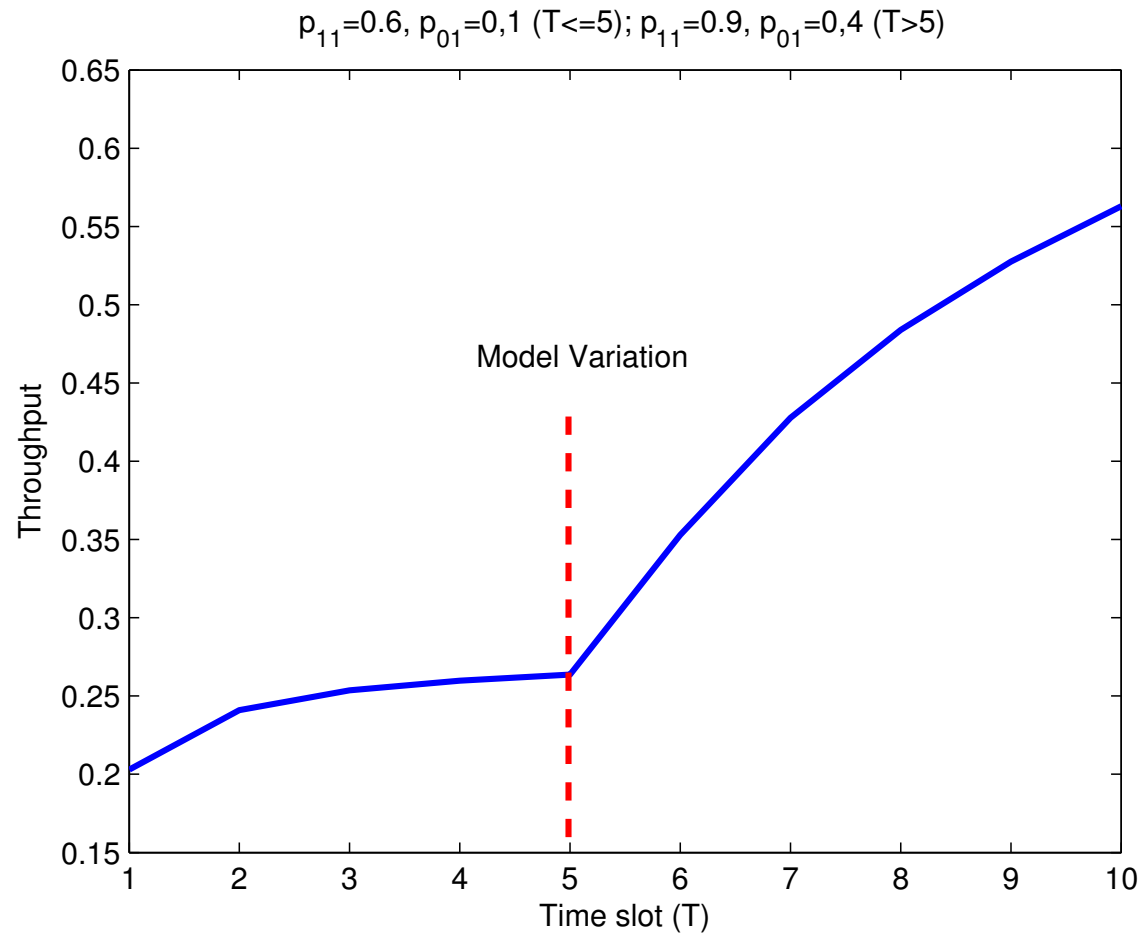
Structure of Whittle's Index Policy: Negative Correlation

- ▶ Stay with bad (**B**) channels and leave good (**G**) ones to the end of the queue.
- ▶ *Reverse* the order of unobserved channels.



Robustness of Whittle's Index Policy

- ▶ Automatically tracks model variations:



Optimality of Whittle's Index Policy

Optimality for positively correlated channels:

- ▶ holds for general N and K .
- ▶ holds for both finite and infinite horizon (discounted/average reward).

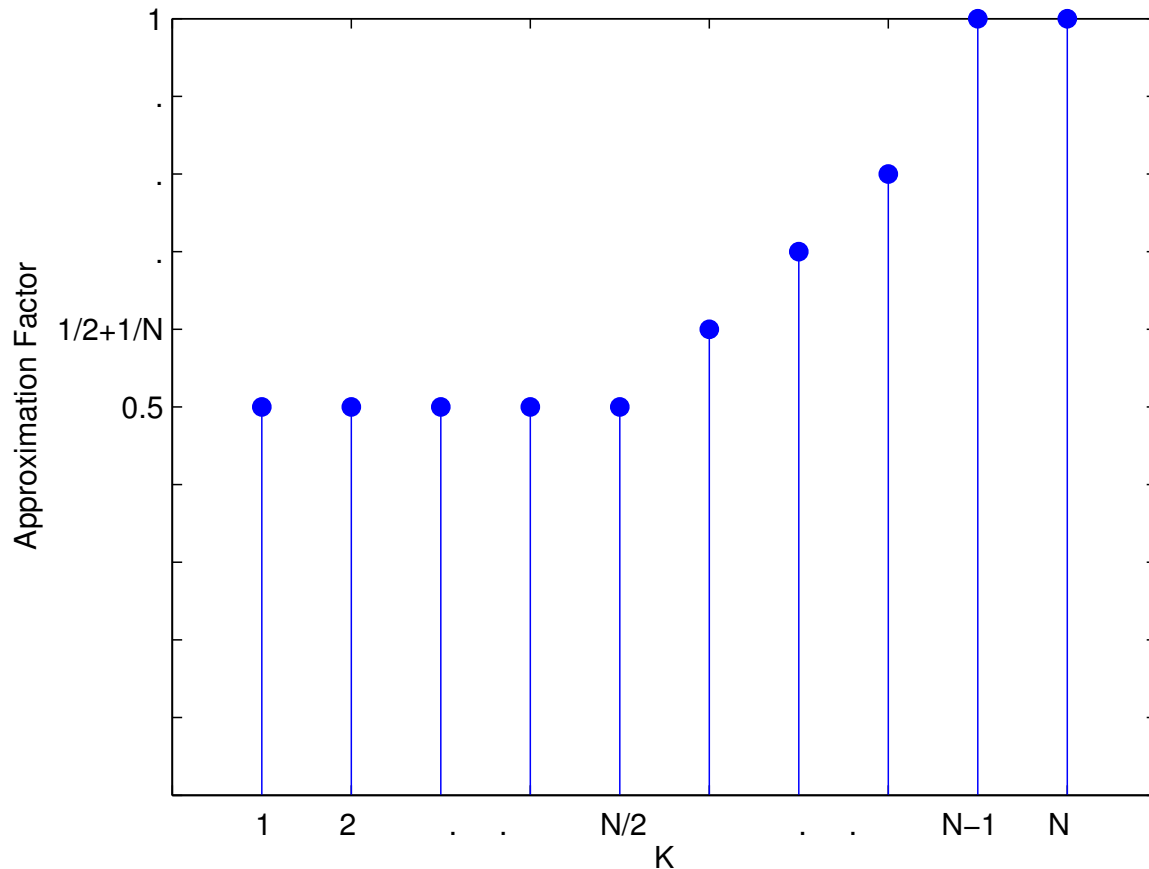
Optimality for negatively correlated channels:

- ▶ holds for all N with $K = N - 1$.
- ▶ holds for $N = 2, 3$.

Performance of Whittle's Index Policy w.r.t. K

Constant approximation factor $\eta = \frac{\text{Performance of Whittle's index policy}}{\text{Optimal performance}}$

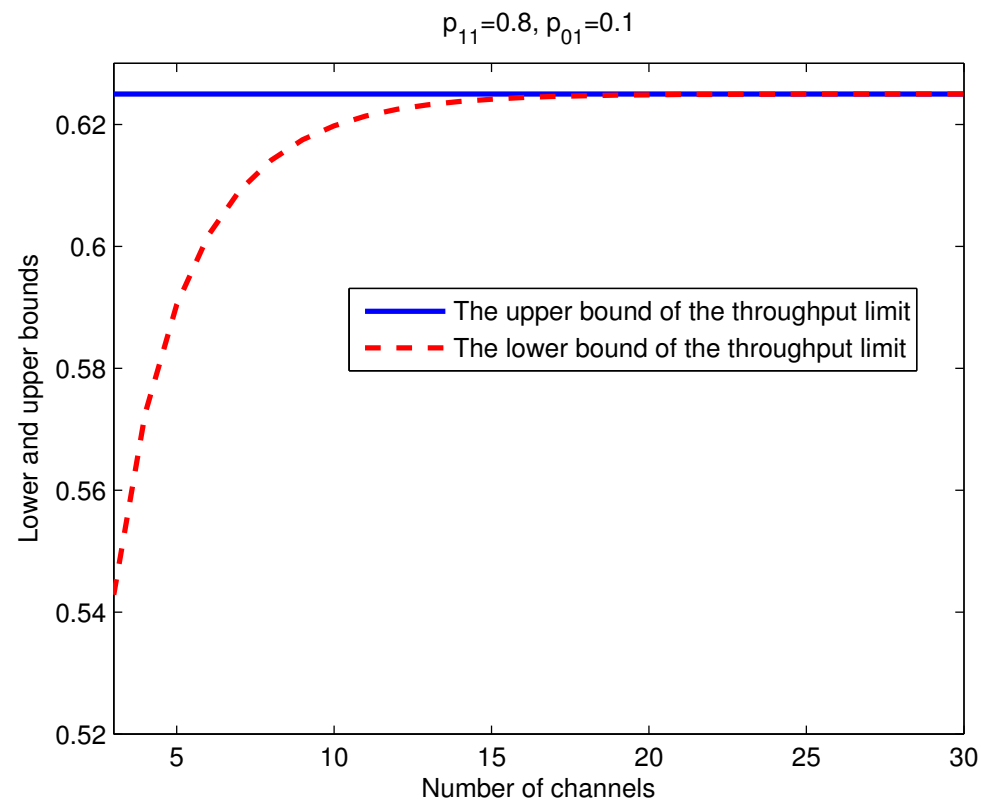
$p_{11} < p_{01}$ (Negative Correlation)



$$\begin{cases} \eta = 1, & K = N - 1, N \\ \eta \geq \max\{\frac{1}{2}, \frac{K}{N}\}, & \text{otherwise} \end{cases}$$

Performance of Whittle's Index Policy w.r.t. N

- ▶ $V(N)$: the average reward achieved by Whittle's index policy ($K = 1$).
- ▶ For $p_{11} \geq p_{01}$, $V(N)$ converges to a constant $\frac{\omega_o}{1-p_{11}+\omega_o}$ at geometric rate $p_{11} - p_{01}$.
- ▶ For $p_{11} < p_{01}$, $V(N)$ approaches a constant $\frac{p_{10}^{(2)}}{E-p_{01}G}$ at geometric rate $(p_{11} - p_{01})^2$.



Inhomogeneous Channels

Whittle's Index in Closed Form:

- ▶ Positive correlation ($p_{11} \geq p_{01}$):

$$I(\omega) = \begin{cases} \omega, & \omega \leq p_{01} \text{ or } \omega \geq p_{11} \\ \frac{\omega}{1-p_{11}+\omega}, & \omega_o \leq \omega < p_{11} \\ \frac{(\omega - \mathcal{T}^1(\omega))(L+2) + \mathcal{T}^{L+1}(p_{01})}{1-p_{11} + (\omega - \mathcal{T}^1(\omega))(L+1) + \mathcal{T}^{L+1}(p_{01})}, & p_{01} < \omega < \omega_o \end{cases}$$

- ▶ Negative correlation ($p_{11} < p_{01}$):

$$I(\omega) = \begin{cases} \omega, & \omega \leq p_{11} \text{ or } \omega \geq p_{01} \\ \frac{p_{01}}{1+p_{01}-\omega}, & \mathcal{T}^1(p_{11}) \leq \omega < p_{01} \\ \frac{p_{01}}{1+p_{01}-\mathcal{T}^1(p_{11})}, & \omega_o \leq \omega < \mathcal{T}^1(p_{11}) \\ \frac{\omega + p_{01} - \mathcal{T}^1(\omega)}{1+p_{01} - \mathcal{T}^1(p_{11}) + \mathcal{T}^1(\omega) - \omega}, & p_{11} < \omega < \omega_o \end{cases}$$

Solving for Whittle's Index in Closed-Form

- ▶ Whittle's index $I(\omega)$:

$$I(\omega) = \{m : V_m(\omega; \text{active}) = V_m(\omega; \text{passive})\}$$

- ▶ Need to obtain $V_m(\omega; \text{active})$ and $V_m(\omega; \text{passive})$ in closed-form

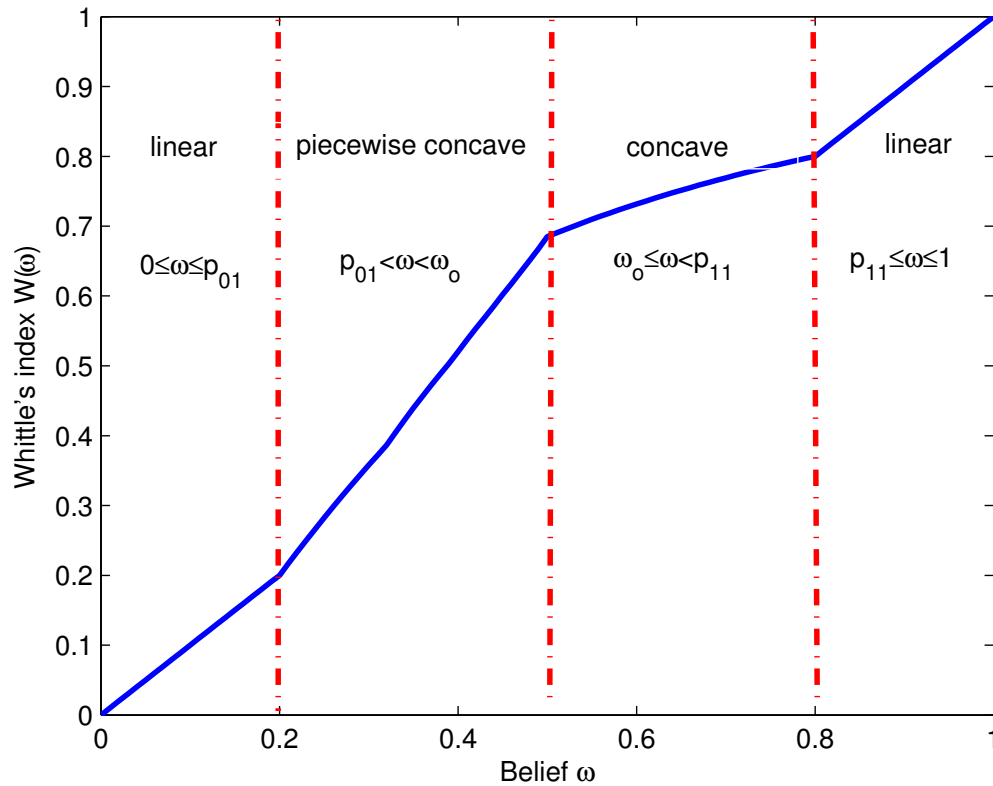
$$V_m(\omega; \text{active}) = \omega + \beta(\omega V_m(p_{11}) + (1 - \omega)V_m(p_{01}))$$

$$V_m(\omega; \text{passive}) = m + \beta V_m(\underbrace{\omega p_{11} + (1 - \omega)p_{01}}_{\text{updated state}})$$

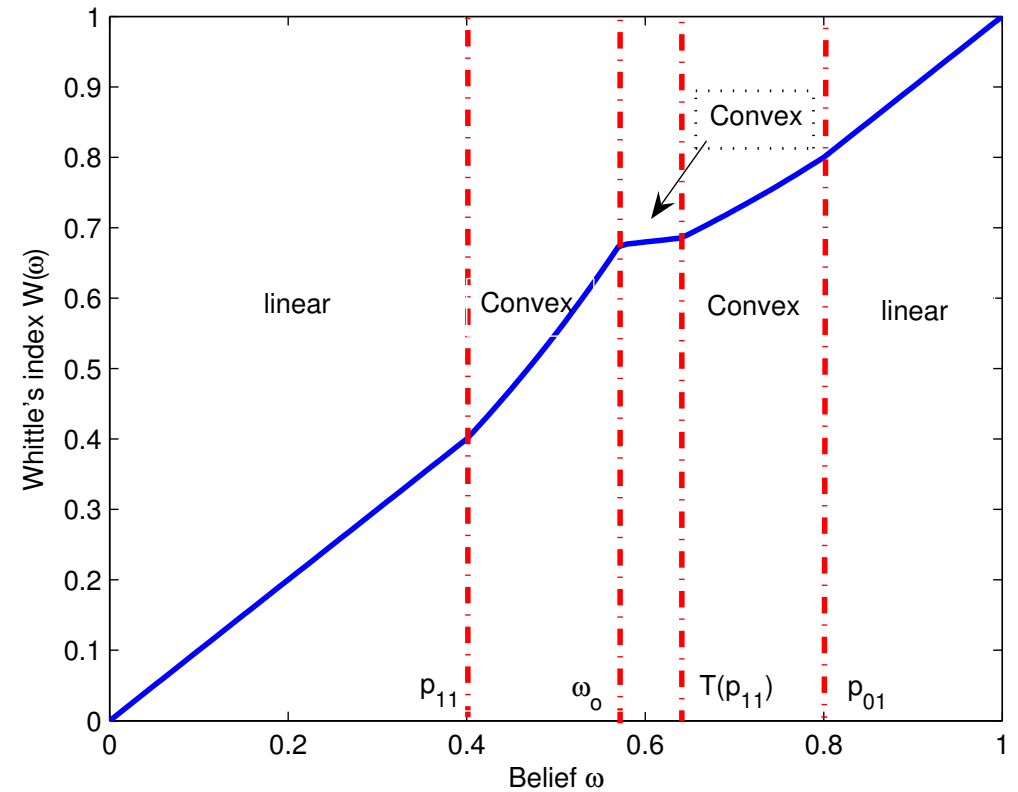
- ▶ $V_m(\omega; \text{active})$ and $V_m(\omega; \text{passive})$ can be written as functions of $\{V_m(p_{11}), V_m(p_{01})\}$ based on the threshold optimal policy and properties of the *first crossing time*.
- ▶ $\{V_m(p_{11}), V_m(p_{01})\}$ can be obtained in closed-form.

Properties of Whittle's Index

Positive correlation ($p_{11} \geq p_{01}$):



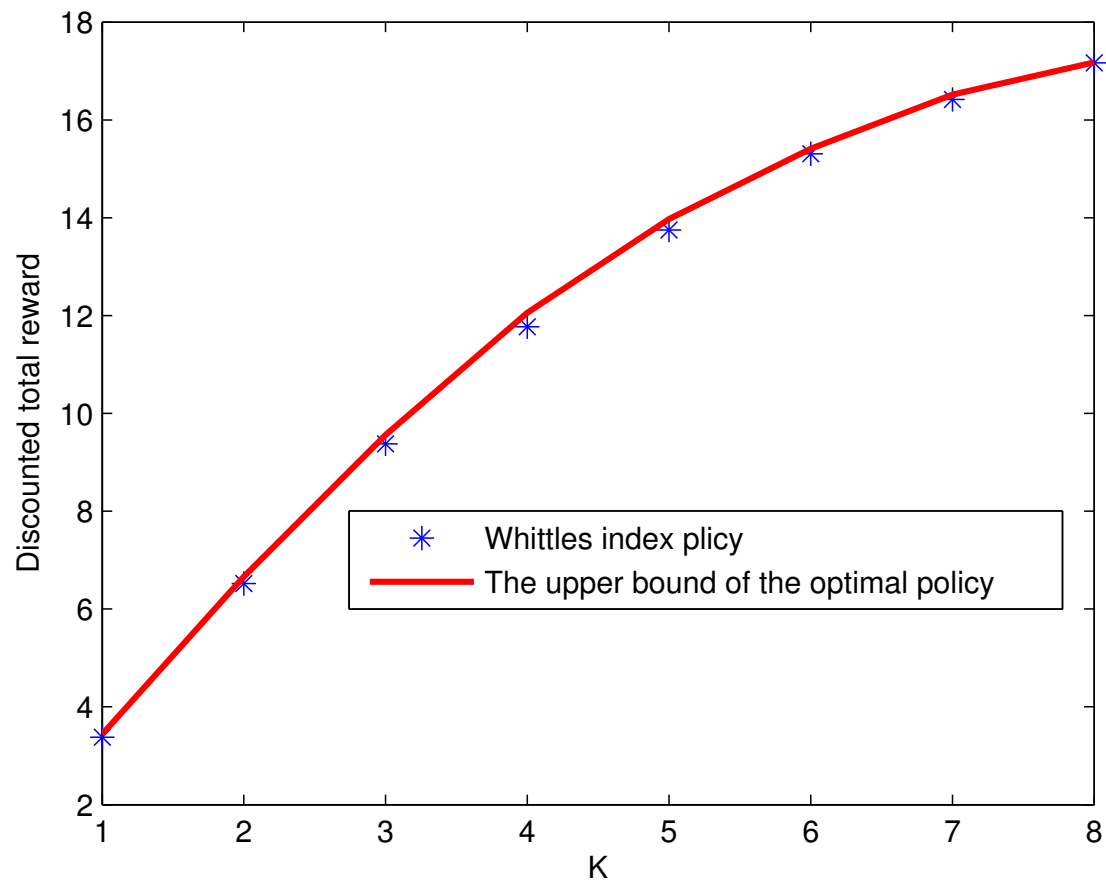
Negative correlation ($p_{11} < p_{01}$):



Monotonicity \Rightarrow equivalence with myopic policy \Rightarrow structure and optimality.

Performance for Inhomogeneous Channels

- ▶ The tightness of the performance upper bound ($\mathcal{O}(N(\log N)^2)$ running time).
- ▶ The near-optimal performance of Whittle's index policy



Conclusion and Acknowledgement

Conclusion and Acknowledgement

- ▶ Indexability and Whittle's index in closed-form: *K.Liu&Q.Zhao:08*.
- ▶ The semi-universal structure of Whittle's index policy:
 - Equivalence to the myopic policy: *K.Liu&Q.Zhao:08*.
 - Structure of the myopic policy: *Q.Zhao&B.Krishnamachari:07*.
- ▶ The optimality of Whittle's index policy:
 - Equivalence to the myopic policy: *K.Liu&Q.Zhao:08*.
 - Optimality of the myopic policy for $N = 2$: *Q.Zhao&B.Krishnamachari:07*.
 - Optimality of the myopic policy for $N > 2$:
S.Ahmad&M.Liu&T.Javidi&Q.Zhao&B.Krishnamachari:08, S.Ahmad&M.Liu:09.
- ▶ The performance of Whittle's index policy for non-identical arms:
 - Constant approximation factor w.r.t. K : *K.Liu&Q.Zhao:08*.
 - Scaling behavior w.r.t. N : *K.Liu&Q.Zhao:08*.
 - $\mathcal{O}(N(\log N)^2)$ -time algorithm for computing an upper bound: *K.Liu&Q.Zhao:08*.