

FOUR OPTIMALITY PROOFS
of the
GITTINS INDEX POLICY
for
MULTI-ARMED BANDITS

Gideon Weiss,
Esther Frostig,
Haifa University.

Presented at
Uri Yechiali retirement conference
Shefayim 2006

I was sharing office with Uri Yechiali some 30 years ago when I met my first bandit

We have since moved like branching bandits and restless bandits throughout our diverging and converging careers, and it is a great pleasure be here on the festive occasion of his retirement.

Retired he may be, but I know that Uri will continue for many years to take part in our searches for new ideas and better understanding of our stochastic world - I wish him lots of success and happiness

I chose to return to bandits today as I feel they may lead us to new directions in optimal control of queues. My survey talk today also aims to re-state that bandits and their Gittins solution are both elegant and simple!

Multiarmed Bandits

Arms: $n = 1, \dots, N$

Processes: $Z_n(t)$ i.i.d. countable states space E

Activate: $n = n(t)$

Transitions: $p(i, j) = P\{Z_n(t+1) = j \mid Z_n(t) = i\}$,

Reward: $R(i)$ bounded $-C < R(i) < C$

Other arms: frozen state, no reward.

$$\max_{\pi} E_{\pi} \sum_{t=0}^{\infty} \alpha^t R(Z_{n(t)}(t))$$

Theorem (Gittins 72-76): The problem is solved by a priority policy. Priority (Gittins) index is calculated for each arm separately.

Contributors:

Gittins, Klimov, Sevcik, Harrison, Tcha & Pliska, Meilijson & Weiss, Glazebrook, Whittle, Kelly, Weber, Mandelbaum, Kaspi, Veinott, Katehakis, Kharoubi, Karatzas, Lai, Berry, Fristedt, Varaiya, Walrand, Tsoucas, Tsitsiklis, Bertsimas, Niño-Mora, Garbe.

Gittins Index:

Max over all positive stopping times:

$$v(i) = \sup_{\sigma > 0} \frac{E \left\{ \sum_{t=0}^{\sigma-1} \alpha^t R(Z(t)) \mid Z(0) = i \right\}}{E \left\{ \sum_{t=0}^{\sigma-1} \alpha^t \mid Z(0) = i \right\}}$$

Gittins order:

Define complete order $j \prec i \Leftrightarrow v(j) < v(i)$

$$S_i^- = \{j : j \prec i\} \quad S_i = S_i^- \cup i$$

Lemma: $v(i)$ is achieved by:

$$\tau(i) = \min \{t : Z(t) \prec i \mid Z(0) = i\} = T_i^{S_i^-}$$

Proof: Direct proof

$$\frac{a}{c} < \frac{a+b}{c+d} \Leftrightarrow \frac{a+b}{c+d} < \frac{b}{d} \Leftrightarrow \frac{a}{c} < \frac{b}{d}$$

- (a) Do not stop when $v(Z(t)) > v(i)$
- (b) Do not continue when $v(Z(t)) < v(i)$
- (c) Assume $v(i)$ is not achieved.

Construct improving sequence $\sigma_n \uparrow \tau(i)$, Contradiction

- (d) All σ : $v(Z(\sigma)) \leq v(i)$, $\sigma \leq \tau(i)$ achieve supremum.

Single Arm Dynamic Programming:

The fixed arm problem (Gittins):

Play the bandit $Z(t)$ against a fixed arm with a reward γ .

$$V(i) = \max\{R(i) + \alpha \sum p(i, j)V(j), \gamma + \alpha V(i)\}$$

The fair charge problem (Weber):

Play the bandit $Z(t)$, for a charge of γ .

$$W(i) = \max\{R(i) - \gamma + \alpha \sum p(i, j)W(j), \alpha W(i)\}$$

- Once you do not play, you may as well stop forever

Retirement Problem (Whittle):

Play arm as long as you want, then retire and take M .

$$V(i, M) = \max\{R(i) + \alpha \sum p(i, j)V(j, M), M\}$$

$$\text{Penison: } \gamma = (1 - \alpha)M$$

Solution of single arm dynamic prog:m:

Stopping set: $S_M = \{i : V(i, M) < M\}$

Stopping time: $\tau(i, M)$ (can be 0 or $+\infty$)

$$V(i, M) = E \left\{ \sum_{t=0}^{\tau(i, M)-1} \alpha^t R(Z(t)) + \alpha^{\tau(i, M)} M \mid Z(0) = i \right\}$$

Connection to Gittins index:

$$M \uparrow \Rightarrow S_M \uparrow, \tau(i, M) \downarrow$$

$$M(i) = \inf \{M : i \in S_M\},$$

$$\gamma(i) = (1 - \alpha)M(i),$$

(standard arm, fixed charge, pension for i)

$$v(i) = \gamma(i),$$

$$\tau(i) = \tau(i, M(i)), \quad (> 0, \leq \infty)$$

Lagrangian approach (dual retirement award)

Fixed policy π , $V_\pi(i, M)$ = linear in M

$V(i, M)$ convex increasing in M

Slope increases from 0 to 1

$$E \alpha^{\tau(i, M)} = \frac{\partial}{\partial M} V(i, M) \text{ or } = \text{slope of subgradient}$$

The index sample path:

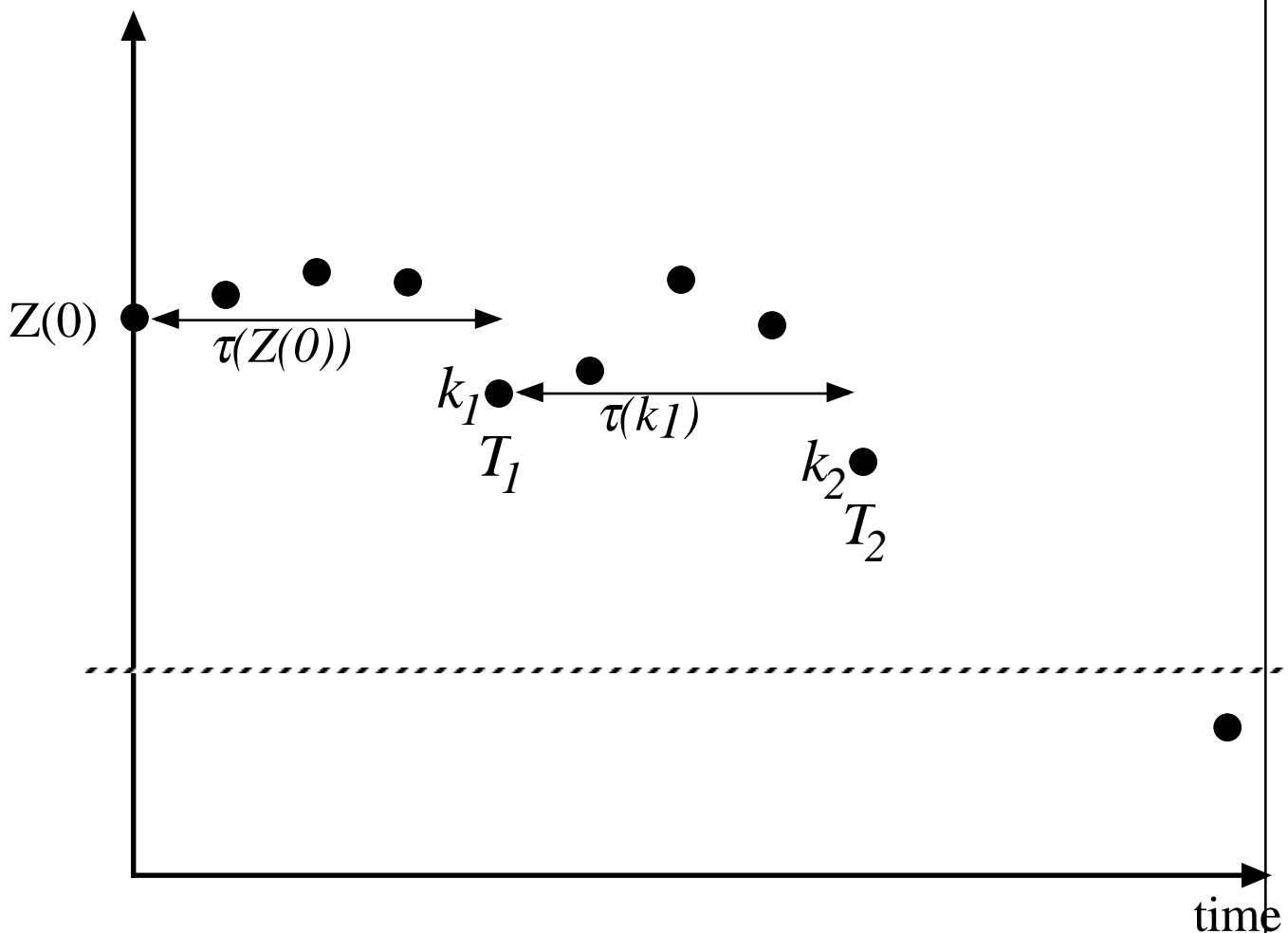
Fair charge: $g(t) = \gamma(Z(t))$

Prevailing charge: $\underline{g}(t) = \min_{s \leq t} g(s) = \min_{s \leq t} \gamma(Z(s))$

$$T_0 = 0, \quad k_0 = Z(0)$$

$$T_\ell = T_{\ell-1} + \tau(k_{\ell-1}), \quad k_\ell = Z(T_\ell)$$

index=fair charge



Klimov's Algorithm

i to S first passage time

$$T_i^S = \min\{t : t > 0, Z(t) \in S \mid Z(0) = i\}$$

i to S expected discounted first passage time

$$A_i^S = E \left\{ \sum_{t=0}^{T_i^S - 1} \alpha^t \mid Z(0) = i \right\}$$

$$A_j^{S_i^-} - A_j^{S_i} = E \left\{ \sum_{\ell=1}^{\infty} I(k_\ell = i) \alpha^{T_\ell} \sum_{t=0}^{\tau(i)-1} \alpha^t \mid Z(0) = j \right\}$$

Lemma:

$$v(i) = \frac{R(i) + \sum_{j \succ i} (A_i^{S_j^-} - A_i^{S_j}) v(j)}{A_i^{S_i}}$$

Corollary:

$$v(i) = \sup_{k \in S_i} \frac{R(k) + \sum_{j \succ k} (A_k^{S_j^-} - A_k^{S_j}) v(j)}{A_k^{S_k}}$$

Klimov's Algorithm: FINITE STATE SPACE:

$$\varphi(1) \succ \varphi(2) \succ \dots \succ \varphi(|E|)$$

$$\varphi(1) = \arg \max_k R(k)$$

$$\varphi(i) = \arg \max_{k \neq \varphi(1), \dots, \varphi(i-1)} \frac{R(k) + \sum_{j=1}^{i-1} (A_k^{S_{\varphi(j)}^-} - A_k^{S_{\varphi(j)}}) v(\varphi(j))}{A_k^{E \setminus \{\varphi(1), \dots, \varphi(i-1)\}}}$$

1st: Gittins Pairwise Interchange

$$\mathbf{Z}(0) = Z_1(0), \dots, Z_n(0)$$

$$Z_{n^*}(0) = i^*$$

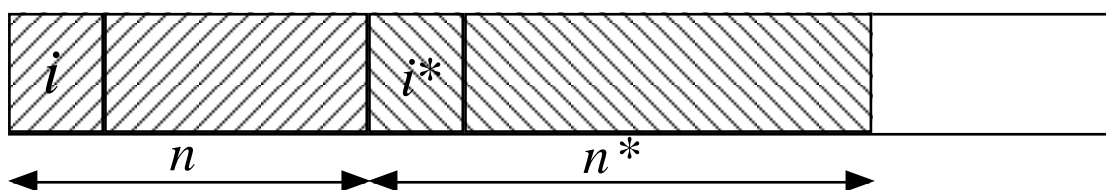
$$Z_n(0) = i \quad i < i^*$$

Not Gittins

Gittins

Gittins

Gittins



$$T_i^{S_{i^*}}$$

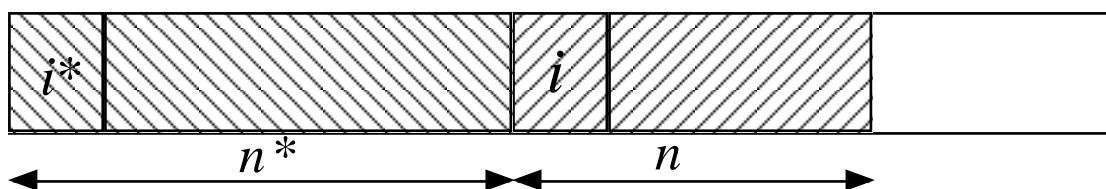
$$\tau(i^*)$$

? Gittins ??

Gittins

Gittins

Gittins



$$\tau(i^*)$$

$$T_i^{S_{i^*}}$$

$$E \left\{ \sum_{t=0}^{\tau^*-1} \alpha^t R(Z(t)) \mid Z(0) = i^* \right\} + E \alpha^{\tau^*} E \left\{ \sum_{t=0}^{\sigma-1} \alpha^t R(Z(t)) \mid Z(0) = j \right\} \geq$$

$$E \left\{ \sum_{t=0}^{\sigma-1} \alpha^t R(Z(t)) \mid Z(0) = j \right\} + E \alpha^{\sigma} E \left\{ \sum_{t=0}^{\tau^*-1} \alpha^t R(Z(t)) \mid Z(0) = i^* \right\}$$

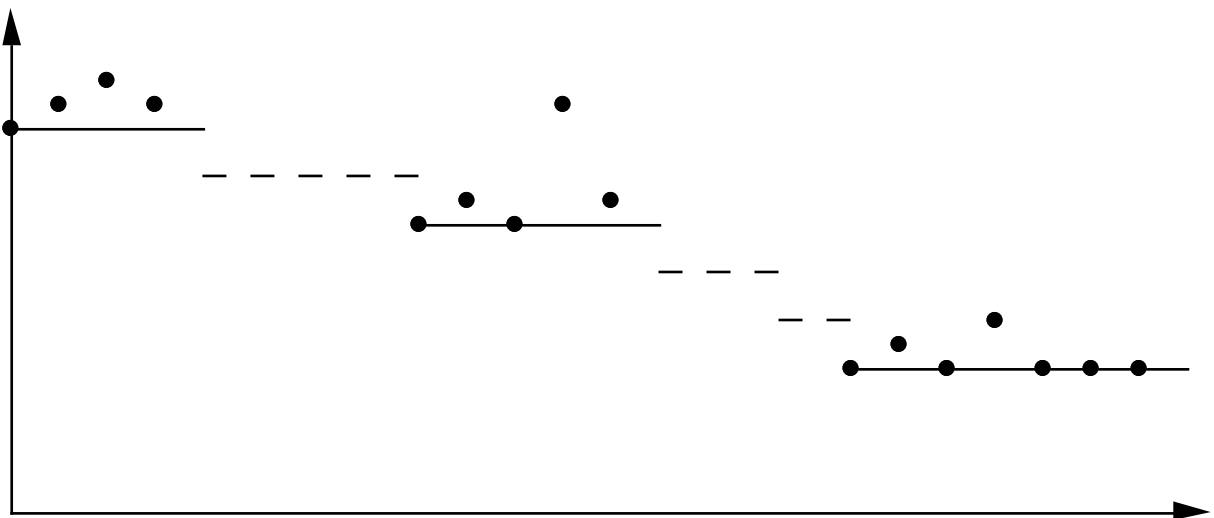
2nd: Prevailing Charge (Weber)

Fair charge g_n , prevailing charge \underline{g}_n

If arm n is played at the times $t^n(1) < t^n(2) < \dots < t^n(s) < \dots$:

Optimal

- (1) Never stop an arm if fair charge exceeds prevailing charge.
- (2) Interleave the arms by decreasing prevailing charge.



Let: $\underline{g}^*(t) =$ decreasing rearrangement of $\underline{g}_n(t^n(\cdot))$

$$\begin{aligned}
 E \left\{ \sum_{t=0}^{\infty} \alpha^t \underline{g}^*(t) \right\} &\geq E \left\{ \sum_{n=1}^N \sum_{s=0}^{\infty} \alpha^{t^n(s)} \underline{g}_n(t^n(s)) \right\} \\
 &\geq E \left\{ \sum_{n=1}^N \sum_{s=0}^{\infty} \alpha^{t^n(s)} R(Z_n(t^n(s))) \right\}
 \end{aligned}$$

Equality for Gittins policy

3rd: Lagrangean Approach (Whittle)

With retirement option, for $\mathbf{Z}(0) = \mathbf{i}$ optimality equation:

$$V(\mathbf{i}, M) = \max_n \{M, R(i_n) + \alpha \sum p(i_n, j) V(i_1, \dots, j, \dots, i_N, M)\}$$

Theorem (Whittle): With retirement option, optimal policy is:

1. Play arm with maximal $M(i_n)$.
2. Retire when $M > M(i_n)$.

Recall:
$$\frac{\partial}{\partial M} V(i, M) = E \alpha^{\tau(i, M)}.$$

Let T_M be the retirement stopping time.

GUESS:
$$T_M = \sum \tau(i_n, M)$$

$$\frac{\partial}{\partial M} V(\mathbf{i}, M) = E \alpha^{T_M} = E \alpha^{\sum \tau(i_n, M)} = \prod \frac{\partial}{\partial M} V(i_n, M)$$

Let
$$Q_n(\mathbf{i}, M) = \prod_{n' \neq n} \frac{\partial}{\partial M} V(i_{n'}, M)$$

then:
$$V(\mathbf{i}, M) = C - \int_M^C \frac{\partial}{\partial m} V(i_n, m) Q_n(\mathbf{i}, m) dm$$

To complete the proof:

show $V(\mathbf{i}, m)$ satisfies the optimality equation.

4th: Achievable Regions - GCL (Bertsimas & Nino-Mora)

For Initial state $\mathbf{Z}(0)$ and policy π

$I_i^\pi(t)$ indicator that π plays arm in state i at t .

$$x_i^\pi = E \left\{ \sum_{t=0}^{\infty} I_i^\pi(t) \alpha^t \right\}.$$

x_i^π is performance measure of π

Recall: $A_i^S = E \left\{ \sum_{t=0}^{T_i^S-1} \alpha^t \mid Z(0) = i \right\}$

For initial state $\mathbf{Z}(0)$ define

$$T_{\mathbf{Z}(0)}^S = \text{time to move all arms to } S$$

$$b(S) = E \left\{ \sum_{t=T_{\mathbf{Z}(0)}^S}^{\infty} \alpha^t \right\}$$

Theorem - Generalized Conservation law:

For every policy π ,

$$\sum_{i \in S} A_i^S x_i^\pi \geq b(S), \quad \text{all } S \subseteq E.$$

Equality $\Leftrightarrow \pi$ gives priority to \bar{S} over S

For $S = E$:
$$\sum_{i \in E} x_i^\pi = \sum_{i \in E} A_i^E x_i^\pi = b(E) = \frac{1}{1 - \alpha},$$

Proof Consider realization under π .

Time axis is divided into:

$s_o(1) < \dots < s_o(T_{Z(0)}^S)$ arms originally in S^c are played before they enter S

$s_i(1) < \dots < s_i(l) < \dots$ arms in state $i \in S$ are played.

$s_{i,l}(1) < \dots < s_{i,l}(T_i^S(i,l) - 1)$ the arm played at time $s_i(l)$ is subsequently played as long as it is in S^c

$$(1): \frac{1}{1-\alpha} = \sum_{r=1}^{T_{Z(0)}^S} \alpha^{s_o(r)} + \sum_{i \in S} \sum_{l=0}^{\infty} \left(\alpha^{s_i(l)} + \sum_{r=1}^{T_i^S(i,l)-1} \alpha^{s_{i,l}(r)} \right)$$

(2):

$$\sum_{i \in S} \sum_{l=0}^{\infty} \alpha^{s_i(l)} \left(1 + \alpha + \dots + \alpha^{T_i^S(i,l)-1} \right) \geq$$

$$\geq \sum_{i \in S} \sum_{l=0}^{\infty} \left(\alpha^{s_i(l)} + \sum_{r=1}^{T_i^S(i,l)-1} \alpha^{s_{i,l}(r)} \right) = \frac{1}{1-\alpha} - \sum_{r=1}^{T_{Z(0)}^S} \alpha^{s_o(r)} \geq$$

$$\geq \frac{1}{1-\alpha} - \left(1 + \alpha + \dots + \alpha^{T_{Z(0)}^S-1} \right) = \frac{\alpha^{T_{Z(0)}^S}}{1-\alpha}$$

Equality : holds (pathwise) $\Leftrightarrow \pi$ gives priority to S^c over S

Take expectations to complete the proof.

4th: LP Formulation

Finite State Space

For initial state $\mathbf{Z}(0)$ and policy π

$I_i^\pi(t)$ indicator that π plays arm in state i at t .

$$x_i = E\left(\sum_{t=0}^{\infty} I_i^\pi(t) \alpha^t\right)$$

Achievable region - relaxation of bandit problem:

$$\max \sum_{i \in E} R(i) x_i$$

$$\sum_{i \in S} A_i^S x_i \geq b(S), \quad S \subset E$$

$$\sum_{i \in E} x_i = b(E) = \frac{1}{1 - \alpha}$$

$$x_i \geq 0$$

This LP is an **extended Polymatroid**

Extreme points \Leftrightarrow Priority laws of all permutations

Optimal solution coincides with Gittins index priority policy.

The LP solution - finite state space

Primal LP

$$\begin{aligned} \max \quad & \sum_{i \in E} R(i)x_i \\ & \sum_{i \in S} A_i^S x_i \geq b(S), \quad S \subset E \\ & \sum_{i \in E} x_i = b(E) = \frac{1}{1-\alpha} \\ & x_i \geq 0 \end{aligned}$$

Dual LP*

$$\begin{aligned} \min \quad & \sum_{S \subseteq E} b(S)y^S \\ & \sum_{S: i \in S} A_i^S y^S \leq R(i), \quad i \in E \\ & y^S \leq 0, \quad S \subset E \end{aligned}$$

Gittins priority policy:

$$\begin{aligned} \sum_{j \succ i} A_j^{S_i} x_j &= b(S_i) & \sum_{j \in S} A_j^S x_j &> b(S), S \neq S_i \\ y^{S_i} &\neq 0, & y^S &= 0, \quad S \neq S_i \end{aligned}$$

Dual Solution

$$y^E = \max R(i), \quad y^{S_{\varphi(i)}} = v(\varphi(i)) - v(\varphi(i-1)) \leq 0$$

This is seen from Klimov's algorithm

$$v(\varphi(i)) = \max_{k \neq \varphi(1), \dots, \varphi(i-1)} \frac{R(k) + \sum_{j=1}^{i-1} (A_k^{S_{\varphi(j)}^-} - A_k^{S_{\varphi(j)}}) v(\varphi(j))}{A_k^{E \setminus \{\varphi(1), \dots, \varphi(i-1)\}}}$$

Extension to countable state space:

Infinite countable number of unknowns x_i $i \in E$ and

Infinite uncountable number of constraints $S \subset 2^E$

(1) Generalized conservation laws hold:

$$\begin{aligned} \max \quad & \sum_{i \in E} R(i)x_i \\ & \sum_{i \in S} A_i^S x_i \geq b(S), \quad S \subset E \\ & \sum_{i \in E} x_i = b(E) = \frac{1}{1-\alpha} \\ & x_i \geq 0 \end{aligned}$$

(2) Instead of Klimov's algorithm we use:

$$v(i) = \sup_{k \in S_i} \frac{R(k) + \sum_{j \succ k} (A_k^{S_j^-} - A_k^{S_j})v(j)}{A_k^{S_k}}$$

Generalized LP:

$$\begin{array}{ll} \max \langle \mathbf{x}, R \rangle & \min \langle \mathbf{b}, \mathbf{y}^* \rangle \\ \mathbf{Ax} - \mathbf{b} \in P_{\mathbf{Y}} & \mathbf{A}^* \mathbf{y}^* - R \in P_{\mathbf{X}^*} \\ \mathbf{x} \in P_{\mathbf{X}} & \mathbf{y}^* \in P_{\mathbf{Y}^*} \end{array}$$

\mathbf{x}, \mathbf{y}^* optimal if

(i) \mathbf{x} is feasible: $\mathbf{Ax} - \mathbf{b} \in P_{\mathbf{Y}}, \quad \mathbf{x} \in P_{\mathbf{X}}$

(ii) \mathbf{y}^* is feasible: $\mathbf{A}^* \mathbf{y}^* - R \in P_{\mathbf{X}^*}, \quad \mathbf{y}^* \in P_{\mathbf{Y}^*}$

(iii) \mathbf{x}, \mathbf{y}^* complementary slack: $\langle \mathbf{Ax} - \mathbf{b}, \mathbf{y}^* \rangle = 0, \quad \langle \mathbf{x}, \mathbf{A}^* \mathbf{y}^* - R \rangle = 0$

Generalized LP

Primal space

$$\mathbf{X} : \mathbf{x} = \{x_i\}_{i \in E} \quad \sum_{i \in E} |x_i| < \infty$$

$P_{\mathbf{X}}$ Positive cone $|x_i| \geq 0$

\mathbf{X}^* linear functionals on \mathbf{X} , $R \in \mathbf{X}^*$, $\langle \mathbf{x}, R \rangle = \sum_{i \in E} R(i)x_i$

$$P_{\mathbf{X}^*} = \{\mathbf{x}^* : \langle \mathbf{x}, \mathbf{x}^* \rangle \geq 0 \text{ all } \mathbf{x} \in P_{\mathbf{X}}\}$$

Transformation :

$$\mathbf{x} \rightarrow \mathbf{y} = \mathbf{A}\mathbf{x}, \quad \mathbf{y} : 2^E \rightarrow R, \quad y(S) = \sum_{i \in S} A_i^S x_i$$

Dual space

$$\mathbf{Y} : \text{contains } \mathbf{b} : 2^E \rightarrow R, \quad b(S) = \frac{E(\alpha^{T_{\mathbf{Z}}^S(0)})}{1 - \alpha}$$

$$\text{and } \mathbf{y} = \mathbf{A}\mathbf{x}, \quad \mathbf{x} \in \mathbf{X}$$

$$P_{\mathbf{Y}} : Y(S) \geq 0, \quad S \subset E, \quad Y(E) = 0$$

$$\mathbf{Y}^*, P_{\mathbf{Y}^*}$$

$$\mathbf{A}^* \mathbf{y}^* \in \mathbf{X}^* : \quad \langle \mathbf{x}, \mathbf{A}^* \mathbf{y}^* \rangle = \langle \mathbf{A}\mathbf{x}, \mathbf{y}^* \rangle$$
$$\max \langle \mathbf{x}, R \rangle \quad \min \langle \mathbf{b}, \mathbf{y}^* \rangle$$

$$\mathbf{A}\mathbf{x} - \mathbf{b} \in P_{\mathbf{Y}} \quad \mathbf{A}^* \mathbf{y}^* - R \in P_{\mathbf{X}^*}$$

$$\mathbf{x} \in P_{\mathbf{X}} \quad \mathbf{y}^* \in P_{\mathbf{Y}^*}$$

Lemma: \mathbf{Y} is a subspace of all functions from $2^E \rightarrow R$ with "bounded variation".

The generalized LP Solution:

$$\begin{array}{ll}
 \max \langle \mathbf{x}, R \rangle & \min \langle \mathbf{b}, \mathbf{y}^* \rangle \\
 \mathbf{Ax} - \mathbf{b} \in P_{\mathbf{Y}} & \mathbf{A}^* \mathbf{y}^* - R \in P_{\mathbf{X}^*} \\
 \mathbf{x} \in P_{\mathbf{X}} & \mathbf{y}^* \in P_{\mathbf{Y}^*}
 \end{array}$$

if \mathbf{x}, \mathbf{y}^* satisfy: feasible and complementary slack then:
 \mathbf{x}, \mathbf{y}^* are optimal.

$\mathbf{x}^G \in \mathbf{X}$: $x_i^G = \begin{cases} \text{expected discounted time an arm in state } i \\ \text{is activated under Gittins priority policy} \end{cases}$

$$\mathbf{v}^* \in \mathbf{Y}^* : \langle \mathbf{y}, \mathbf{v}^* \rangle = \sum_{i \in E} v(i) [y(S_i) - y(S_i^-)]$$

(a) $\mathbf{Ax}^G - \mathbf{b} \in P_{\mathbf{Y}}, \mathbf{x}^G \in P_{\mathbf{X}}$ by conservation laws (LP is a relaxation)

(b) $\mathbf{A}^* \mathbf{v}^* - R \in P_{\mathbf{X}^*}$, and $\langle \mathbf{x}^G, \mathbf{A}^* \mathbf{v}^* - R \rangle = 0$

follow by showing that $\mathbf{A}^* \mathbf{v}^* - R = 0$

(c) $\langle \mathbf{Ax}^G - \mathbf{b}, \mathbf{v}^* \rangle = 0$

follows directly from definition and conservation laws.

(d) $\mathbf{v}^* \in P_{\mathbf{Y}^*}$ follows from Klimov's algorithm.

Do we need countable state space?

(1) Bernoulli bandits: Arm wins with probability θ . Prior distribution for arm is $\theta \sim \text{Beta}(a, b)$.

$$Z(0) = (a, b) \quad Z(t+1) | Z(t) = (m, n) = \begin{cases} (m+1, n) & \text{w.p. } \theta \\ (m, n+1) & \text{w.p. } 1-\theta \end{cases}$$

we do not know how to calculate the index!

(2) A single server queue has countable but quite well ordered state space

Multi class queueing networks have countable not well ordered state spaces

Do we need countable achievable region proof?

Yes, because new theory for restless bandits, using Whittle index and Nino Mora's Marginal Productivity Index are based on achievable region and partial conservation laws.

Restless Bandits

Arms: $n = 1, \dots, N$

Processes: $Z_n(t)$ i.i.d. countable states space E

Activate: m out of N . remaining $N - m$ are passive

Transitions:

$$p^1(i, j) = P\{Z_n(t+1) = j \mid Z_n(t) = i, \text{ arm } n \text{ active}\},$$

$$p^2(i, j) = P\{Z_n(t+1) = j \mid Z_n(t) = i, \text{ arm } n \text{ passive}\}$$

Rewards:

$R^1(i)$ active reward

$R^2(i)$ passive reward

bounded $-C < R^1(i), R^2(i) < C$

$$\max_{\pi} E_{\pi} \sum_{t=0}^{\infty} \alpha^t \sum_{n=1}^N R^{a(n,t)}(Z_n(t))$$

Whittle index: Value of subsidy for being passive which makes the state indifferent between active or passive (defined for undiscounted long term average reward).

Generalized by Nino Mora as Marginal Productivity Index

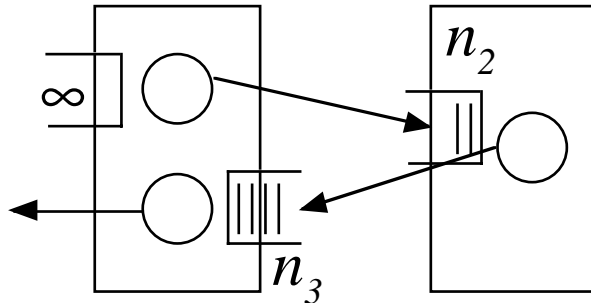
Index priority policy is not optimal

Contributors:

Whittle, Weber & Weiss, Glazebrook, Niño-Mora

Two models:

(1) The 2 node 3 classes network with infinite supply of work (Adan & Weiss):

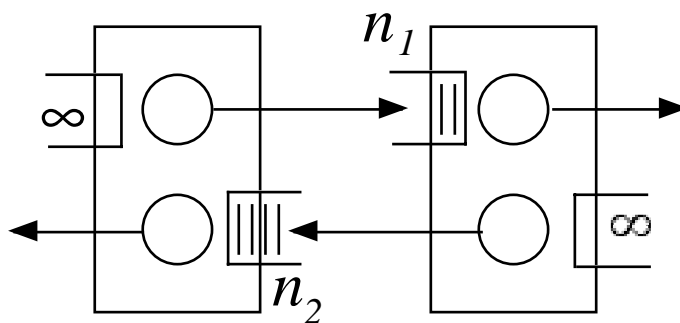


We know how to calculate steady state distribution under LBFS

We do not know how to calculate optimal policy. Should be threshold policy: Feed n_2 only when $n_2 < s(n_3)$

Can we calculate an index?

(2) The Rybko Stolyar network with infinite supply of work (Kopzon & Weiss)



This system will be stable with both nodes working all the time, under a whole class of threshold policies. We do not know which is optimal.

Can we calculate an index?