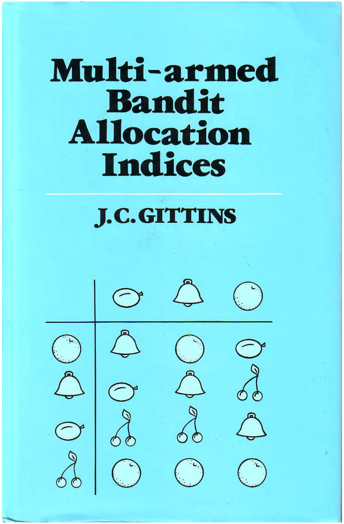
The Multi-Armed Bandit Problem: Index Theory Since Gittins

Richard Weber
Statistical Laboratory, University of Cambridge

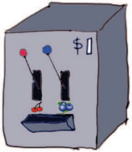
March 2, GOCPS Leipzig 2010

The Multi-armed Bandit Problem

The Multi-armed Bandit Problem



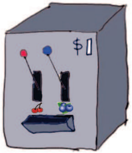
Two-armed Bandit



3, 10, 4, 9, 12, 1, ...

5, 6, 2, 15, 2, 7, ...

Two-armed Bandit

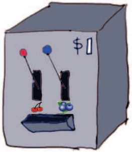


3, 10, 4, 9, 12, 1, ...

, 6, 2, 15, 2, 7, ...

$\longrightarrow$ 5

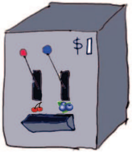
Two-armed Bandit



3, 10, 4, 9, 12, 1, ...

, , 2, 15, 2, 7, ... $\longrightarrow$ 5, 6

Two-armed Bandit

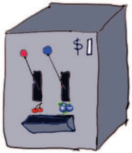


, 10, 4, 9, 12, 1, ...

,   , 2, 15, 2, 7, ...

$\longrightarrow$ 5, 6, 3

Two-armed Bandit

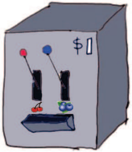


, , 4, 9, 12, 1, ...

, , 2, 15, 2, 7, ...

$\longrightarrow$ 5, 6, 3, 10,

Two-armed Bandit

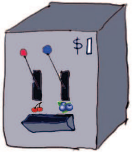


, , , 9, 12, 1, ...

, , 2, 15, 2, 7, ...

$\longrightarrow$ 5, 6, 3, 10, 4

Two-armed Bandit

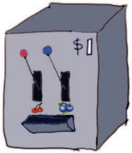


, , , , 12, 1, ...

, , 2, 15, 2, 7, ...

$\longrightarrow$ 5, 6, 3, 10, 4, 9

Two-armed Bandit

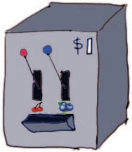


, , , , , 1, ...

, , 2, 15, 2, 7, ...

$\longrightarrow$ 5, 6, 3, 10, 4, 9, 12

Two-armed Bandit

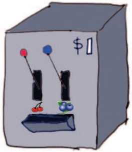


, , , , , 1, ...

, , , 15, 2, 7, ...

$\longrightarrow$ 5, 6, 3, 10, 4, 9, 12, 2

Two-armed Bandit

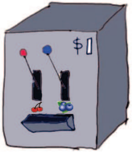


, , , , , 1, ...

, , , , 2, 7, ...

$\longrightarrow$ 5, 6, 3, 10, 4, 9, 12, 2, 15

Two-armed Bandit



, , , , , 1, ...

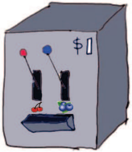
, , , , 2, 7, ...

$\longrightarrow$ 5, 6, 3, 10, 4, 9, 12, 2, 15

Reward = $5 + 6\,\beta + 3\,\beta^2 + 10\,\beta^3 + \cdots$

$0 < \beta < 1.$

Two-armed Bandit



, , , , , 1, ...

, , , , 2, 7, ...

$\longrightarrow$ 5, 6, 3, 10, 4, 9, 12, 2, 15

Reward = $5 + 6\,\beta + 3\,\beta^2 + 10\,\beta^3 + \cdots$

$0 < \beta < 1$. Of course, in practice we must choose which arms to pull without knowing the future sequences of rewards.

Dynamic Effort Allocation

- **Research projects**: how should I allocate my research time amongst my favorite open problems so as to maximize the value of my completed research?

Dynamic Effort Allocation

- **Research projects**: how should I allocate my research time amongst my favorite open problems so as to maximize the value of my completed research?
- **Job Scheduling**: in what order should I work on the tasks in my in-tray?

Dynamic Effort Allocation

- **Research projects**: how should I allocate my research time amongst my favorite open problems so as to maximize the value of my completed research?
- **Job Scheduling**: in what order should I work on the tasks in my in-tray?
- **Searching for information**: shall I spend more time browsing the web, or go to the library, or ask a friend?

Dynamic Effort Allocation

- **Research projects**: how should I allocate my research time amongst my favorite open problems so as to maximize the value of my completed research?
- **Job Scheduling**: in what order should I work on the tasks in my in-tray?
- **Searching for information**: shall I spend more time browsing the web, or go to the library, or ask a friend?
- **Dating strategy**: should I contact a new prospect, or try another date with someone I have dated before?

Information vs. Immediate Payoff

In all these problems one wishes to learn about the effectiveness of alternative strategies, while simultaneously wishing to use the best strategy in the short-term.

Information vs. Immediate Payoff

In all these problems one wishes to learn about the effectiveness of alternative strategies, while simultaneously wishing to use the best strategy in the short-term.

> **"Bandit problems embody in essential form a conflict evident in all human action: information versus immediate payoff."**
>
> — *Peter Whittle (1989)*

Information vs. Immediate Payoff

In all these problems one wishes to learn about the effectiveness of alternative strategies, while simultaneously wishing to use the best strategy in the short-term.

> **"Bandit problems embody in essential form a conflict evident in all human action: information versus immediate payoff."**
>
> — *Peter Whittle (1989)*

"Exploration versus exploitation"

Clinical Trials

Bernoulli Bandits

One of N drugs is to be administered at each of times $t = 0, 1, \ldots$

Bernoulli Bandits

One of N drugs is to be administered at each of times $t = 0, 1, \dots$

The sth time drug i is used it is successful, $X_i(s) = 1$,
or unsuccessful, $X_i(s) = 0$.

Bernoulli Bandits

One of N drugs is to be administered at each of times $t = 0, 1, \dots$

The sth time drug i is used it is successful, $X_i(s) = 1$, or unsuccessful, $X_i(s) = 0$.

$P(X_i(s) = 1) = \theta_i$.

Bernoulli Bandits

One of N drugs is to be administered at each of times $t = 0, 1, \ldots$

The sth time drug i is used it is successful, $X_i(s) = 1$,
or unsuccessful, $X_i(s) = 0$.

$P(X_i(s) = 1) = \theta_i$.

$X_i(1), X_i(2), \ldots$ are i.i.d. samples.

Bernoulli Bandits

One of N drugs is to be administered at each of times $t = 0, 1, \dots$

The sth time drug i is used it is successful, $X_i(s) = 1$,
or unsuccessful, $X_i(s) = 0$.

$P(X_i(s) = 1) = \theta_i$.

$X_i(1), X_i(2), \dots$ are i.i.d. samples.

θ_i is unknown, but has a *prior* distribution,
— perhaps uniform on $[0, 1]$

$$f(\theta_i) = 1\,, \quad 0 \le \theta_i \le 1\,.$$

Bernoulli Bandits

Having seen s_i successes and f_i are failures, the posterior is

$$f(\theta_i \mid s_i, f_i) = \tfrac{(s_i+f_i+1)!}{s_i!f_i!}\theta_i^{s_i}(1-\theta_i)^{f_i}\,, \quad 0 \le \theta_i \le 1\,,$$

with mean $(s_i + 1)/(s_i + f_i + 2)$.

Bernoulli Bandits

Having seen s_i successes and f_i are failures, the posterior is

$$f(\theta_i \mid s_i, f_i) = \tfrac{(s_i+f_i+1)!}{s_i! f_i!} \theta_i^{s_i} (1-\theta_i)^{f_i}\,, \quad 0 \le \theta_i \le 1\,,$$

with mean $(s_i + 1)/(s_i + f_i + 2)$.

We wish to maximize the expected total discounted sum of number of successes.

Multi-armed Bandit

N independent arms, with known states $x_1(t), \ldots, x_N(t)$.

Multi-armed Bandit

N independent arms, with known states $x_1(t), \dots, x_N(t)$.

At each time, $t \in \{0, 1, 2, \dots\}$,

- One arm is to be activated (pulled/continued)
  If arm i activated then it changes state:

$$x \to y \quad \text{with probability } P_i(x, y)$$

  and produces reward $r_i(x_i(t))$.

Multi-armed Bandit

N independent arms, with known states $x_1(t), \ldots, x_N(t)$.

At each time, $t \in \{0, 1, 2, \ldots\}$,

- One arm is to be activated (pulled/continued)
  If arm i activated then it changes state:

  $$x \to y \quad \text{with probability } P_i(x, y)$$

  and produces reward $r_i(x_i(t))$.
- Other arms are to be passive (not pulled/frozen).

Multi-armed Bandit

N independent arms, with known states $x_1(t), \dots, x_N(t)$.

At each time, $t \in \{0, 1, 2, \dots\}$,

- One arm is to be activated (pulled/continued)
  If arm i activated then it changes state:

  $$x \to y \quad \text{with probability } P_i(x, y)$$

  and produces reward $r_i(x_i(t))$.
- Other arms are to be passive (not pulled/frozen).

Objective: maximize the expected total β-discounted reward

$$E\left[\sum_{t=0}^{\infty} r_{i_t}(x_{i_t}(t))\,\beta^t\right],$$

where i_t is the arm pulled at time t, $(0 < \beta < 1)$.

Dynamic Programming Solution

The dynamic programming equation is

$$V(x_1, \ldots, x_N)$$

$$= \max_i \Big\{ r_i(x_i) + \beta \sum_y P_i(x_i, y) V(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_N) \Big\}$$

Dynamic Programming Solution

The dynamic programming equation is

$$V(x_1, \ldots, x_N)$$
$$= \max_i \Big\{ r_i(x_i) + \beta \sum_y P_i(x_i, y) V(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_N) \Big\}$$

If bandit i moves on a state space of size k_i, then $(x_1, \ldots, x_N)$ moves on a state space of size $\prod_i k_i$ (exponential in N).

Gittins Index Solution

Theorem [Gittins, '74, '79, '89]

Reward is maximized by always continuing the bandit having greatest value of 'dynamic allocation index'

$$G_i(x_i) = \sup_{\tau \geq 1} \frac{E\left[\sum_{t=0}^{\tau-1} r_i(x_i(t))\beta^t \;\middle|\; x_i(0) = x_i\right]}{E\left[\sum_{t=0}^{\tau-1} \beta^t \;\middle|\; x_i(0) = x_i\right]}$$

where τ is a (past-measurable) stopping-time.

Gittins Index Solution

Theorem [Gittins, '74, '79, '89]

Reward is maximized by always continuing the bandit having greatest value of 'dynamic allocation index'

$$G_i(x_i) = \sup_{\tau \geq 1} \frac{E\left[\sum_{t=0}^{\tau-1} r_i(x_i(t))\beta^t \;\middle|\; x_i(0) = x_i \right]}{E\left[\sum_{t=0}^{\tau-1} \beta^t \;\middle|\; x_i(0) = x_i \right]}$$

where τ is a (past-measurable) stopping-time.

One problem (on a state space size $\prod_i k_i$)
$\longrightarrow N$ problems (on state spaces sizes $k_1, \ldots, k_N$.)

Gittins Index Solution

Theorem [Gittins, '74, '79, '89]

Reward is maximized by always continuing the bandit having greatest value of 'dynamic allocation index'

$$G_i(x_i) = \sup_{\tau \geq 1} \frac{E\left[\sum_{t=0}^{\tau-1} r_i(x_i(t))\beta^t \;\middle|\; x_i(0) = x_i\right]}{E\left[\sum_{t=0}^{\tau-1} \beta^t \;\middle|\; x_i(0) = x_i\right]}$$

where τ is a (past-measurable) stopping-time.

One problem (on a state space size $\prod_i k_i$)
$\longrightarrow N$ problems (on state spaces sizes $k_1, \ldots, k_N$.)

$G_i(x_i)$ is called the **Gittins index**.

Gittins Index Solution

Theorem [Gittins, '74, '79, '89]

Reward is maximized by always continuing the bandit having greatest value of 'dynamic allocation index'

$$G_i(x_i) = \sup_{\tau \geq 1} \frac{E\Big[\sum_{t=0}^{\tau-1} r_i(x_i(t))\beta^t \;\Big|\; x_i(0) = x_i\Big]}{E\Big[\sum_{t=0}^{\tau-1} \beta^t \;\Big|\; x_i(0) = x_i\Big]}$$

where τ is a (past-measurable) stopping-time.

One problem (on a state space size $\prod_i k_i$)
$\longrightarrow N$ problems (on state spaces sizes $k_1, \dots, k_N$.)

$G_i(x_i)$ is called the **Gittins index**.

It can be computed in time $O(k_i^3)$.

Gittins Index

$$G_i(x_i) = \sup_{\tau \geq 1} \frac{E\left[\sum_{t=0}^{\tau-1} r_i(x_i(t))\beta^t \;\middle|\; x_i(0) = x_i \right]}{E\left[\sum_{t=0}^{\tau-1} \beta^t \;\middle|\; x_i(0) = x_i \right]}$$

Gittins Index

$$G_i(x_i) = \sup_{\tau \geq 1} \frac{E\left[\sum_{t=0}^{\tau-1} r_i(x_i(t))\beta^t \;\middle|\; x_i(0) = x_i\right]}{E\left[\sum_{t=0}^{\tau-1} \beta^t \;\middle|\; x_i(0) = x_i\right]}$$

Discounted reward up to τ.

Gittins Index

$$G_i(x_i) = \sup_{\tau \geq 1} \frac{E\left[\left. \sum_{t=0}^{\tau-1} r_i(x_i(t))\beta^t \,\right|\, x_i(0) = x_i\right]}{E\left[\left. \sum_{t=0}^{\tau-1} \beta^t \,\right|\, x_i(0) = x_i\right]}$$

Discounted reward up to τ.

Discounted time up to τ.

Gittins Indices for Bernoulli Bandits, $\beta = 0.9$

s / f	1	2	3	4	5	6	7	8
1	.7029	.8001	.8452	.8723	.8905	.9039	.9141	.9221
2	.5001	.6346	.7072	.7539	.7869	.8115	.8307	.8461
3	.3796	.5163	.6010	.6579	.6996	.7318	.7573	.7782
4	.3021	.4342	.5184	.5809	.6276	.6642	.6940	.7187
5	.2488	.3720	.4561	.5179	.5676	.6071	.6395	.6666
6	.2103	.3245	.4058	.4677	.5168	.5581	.5923	.6212
7	.1815	.2871	.3647	.4257	.4748	.5156	.5510	.5811
8	.1591	.2569	.3308	.3900	.4387	.4795	.5144	.5454

$(s_1, f_1) = (2, 3)$: posterior mean $= \frac{3}{7} = 0.4286$, index $= 0.5163$

Gittins Indices for Bernoulli Bandits, $\beta = 0.9$

s / f	1	2	3	4	5	6	7	8
1	.7029	.8001	.8452	.8723	.8905	.9039	.9141	.9221
2	.5001	.6346	.7072	.7539	.7869	.8115	.8307	.8461
3	.3796	.5163	.6010	.6579	.6996	.7318	.7573	.7782
4	.3021	.4342	.5184	.5809	.6276	.6642	.6940	.7187
5	.2488	.3720	.4561	.5179	.5676	.6071	.6395	.6666
6	.2103	.3245	.4058	.4677	.5168	.5581	.5923	.6212
7	.1815	.2871	.3647	.4257	.4748	.5156	.5510	.5811
8	.1591	.2569	.3308	.3900	.4387	.4795	.5144	.5454

$(s_1, f_1) = (2, 3)$: posterior mean $= \frac{3}{7} = 0.4286$, index $= 0.5163$

$(s_2, f_2) = (6, 7)$: posterior mean $= \frac{7}{15} = 0.4667$, index $= 0.5156$

Gittins Indices for Bernoulli Bandits, $\beta = 0.9$

s		2	3	4	5	6	7	8
f								
1	.7029	.8001	.8452	.8723	.8905	.9039	.9141	.9221
2	.5001	.6346	.7072	.7539	.7869	.8115	.8307	.8461
3	.3796	.5163	.6010	.6579	.6996	.7318	.7573	.7782
4	.3021	.4342	.5184	.5809	.6276	.6642	.6940	.7187
5	.2488	.3720	.4561	.5179	.5676	.6071	.6395	.6666
6	.2103	.3245	.4058	.4677	.5168	.5581	.5923	.6212
7	.1815	.2871	.3647	.4257	.4748	.5156	.5510	.5811
8	.1591	.2569	.3308	.3900	.4387	.4795	.5144	.5454

$(s_1, f_1) = (2, 3)$: posterior mean $= \frac{3}{7} = 0.4286$, index $= 0.5163$

$(s_2, f_2) = (6, 7)$: posterior mean $= \frac{7}{15} = 0.4667$, index $= 0.5156$

So we prefer to use drug 1 next, even though it has the smaller probability of success.

What has Happened Since 1989?

- Index theorem has become better known.

What has Happened Since 1989?

- Index theorem has become better known.
- Alternative proofs have been explored.

What has Happened Since 1989?

- Index theorem has become better known.
- Alternative proofs have been explored.

  Playing golf with N balls

What has Happened Since 1989?

- Index theorem has become better known.
- Alternative proofs have been explored.

  Playing golf with N balls

  Achievable Performance Region Approach

What has Happened Since 1989?

- Index theorem has become better known.
- Alternative proofs have been explored.

  Playing golf with N balls

  Achievable Performance Region Approach
- Many applications (economics, engineering, . . .).

What has Happened Since 1989?

- Index theorem has become better known.
- Alternative proofs have been explored.

  Playing golf with N balls

  Achievable Performance Region Approach
- Many applications (economics, engineering, . . .).
- Notions of indexation have been generalized.

What has Happened Since 1989?

- Index theorem has become better known.
- Alternative proofs have been explored.

  Playing golf with N balls

  Achievable Performance Region Approach
- Many applications (economics, engineering, ...).
- Notions of indexation have been generalized.

  Restless Bandits

Gittins Index Theorem has become Better Known

Peter Whittle tells the story:

"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?'*

Gittins Index Theorem has become Better Known

Peter Whittle tells the story:

> "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.'*

Proofs of the Index Theorem

Since Gittins (1974, 1979), many researchers have reproved, remodelled and resituated the index theorem.

Beale (1979)

Karatzas (1984)

Varaiya, Walrand, Buyukkoc (1985)

Chen, Katehakis (1986)

Kallenberg (1986)

Katehakis, Veinott (1986)

Eplett (1986)

Kertz (1986)

Tsitsiklis (1986)

Mandelbaum (1986, 1987)

Lai, Ying (1988)

Whittle (1988)

Proofs of the Index Theorem

Since Gittins (1974, 1979), many researchers have reproved, remodelled and resituated the index theorem.

Beale (1979)

Karatzas (1984)

Varaiya, Walrand, Buyukkoc (1985)

Chen, Katehakis (1986)

Kallenberg (1986)

Katehakis, Veinott (1986)

Eplett (1986)

Kertz (1986)

Tsitsiklis (1986)

Mandelbaum (1986, 1987)

Lai, Ying (1988)

Whittle (1988)

Weber (1992)

El Karoui, Karatzas (1993)

Ishikida and Varaiya (1994)

Tsitsiklis (1994)

Bertsimas, Niño-Mora (1996)

Glazebrook, Garbe (1996)

Kaspi, Mandelbaum (1998)

Bäuerle, Stidham (2001)

Dimitriu, Tetali, Winkler (2003)

What has Happened Since 1989?

- Index theorem has become better known.
- Alternative proofs have been explored.

  Playing golf with N balls

  Achievable Performance Region Approach
- Many applications (economics, engineering, . . .).
- Notions of indexation have been generalized.

  Restless Bandits

Golf with N Balls

[Dimitriu, Tetali, Winkler '03, W. '92]

N balls are strewn about a golf course at locations $x_1, \ldots, x_N$.

Golf with N Balls

[Dimitriu, Tetali, Winkler '03, W. '92]

N balls are strewn about a golf course at locations $x_1, \ldots, x_N$.

Hitting a ball i, that is in location x_i, costs $c(x_i)$,

$$x_i \to y \quad \text{with probability } P(x_i, y)$$

Ball goes in the hole with probability $P(x_i, 0)$.

Objective

Minimize the expected total cost incurred up to sinking a first ball.

Golf with 1 Ball

- Given the golfer's ball is in location x, let us offer him a prize of value $g(x)$ if he eventually sinks the ball.

Golf with 1 Ball

- Given the golfer's ball is in location x, let us offer him a prize of value $g(x)$ if he eventually sinks the ball.
- Let us set this prize just great enough so that he can break even, by playing one more stroke, and then quitting thereafter whenever he likes.

Golf with 1 Ball

- Given the golfer's ball is in location x, let us offer him a prize of value $g(x)$ if he eventually sinks the ball.
- Let us set this prize just great enough so that he can break even, by playing one more stroke, and then quitting thereafter whenever he likes.

  $g(x) =$ fair prize'.

Golf with 1 Ball

- Given the golfer's ball is in location x, let us offer him a prize of value $g(x)$ if he eventually sinks the ball.
- Let us set this prize just great enough so that he can break even, by playing one more stroke, and then quitting thereafter whenever he likes.

  $g(x) =$ fair prize'.
- If the ball arrives at a location y, from which $g(x)$ is no longer great enough to motivate the golfer to continue playing, then, — just as he is about to quit —, we increase the prize to $g(y)$, which becomes the new 'prevailing prize'.

Golf with 1 Ball

- Given the golfer's ball is in location x, let us offer him a prize of value $g(x)$ if he eventually sinks the ball.
- Let us set this prize just great enough so that he can break even, by playing one more stroke, and then quitting thereafter whenever he likes.

  $g(x) =$ fair prize'.
- If the ball arrives at a location y, from which $g(x)$ is no longer great enough to motivate the golfer to continue playing, then, — just as he is about to quit —, we increase the prize to $g(y)$, which becomes the new 'prevailing prize'.
- Continue doing this until the ball is sunk.

Golf with 1 Ball

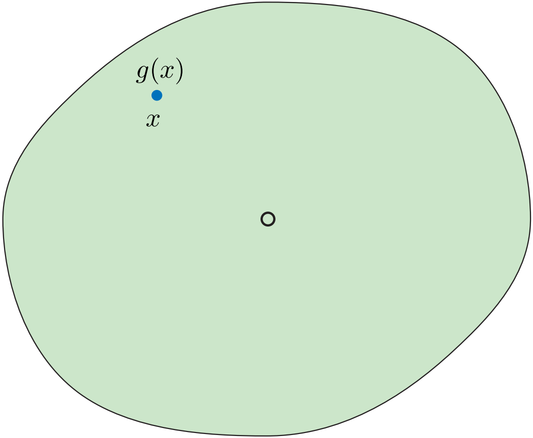
- Given the golfer's ball is in location x, let us offer him a prize of value $g(x)$ if he eventually sinks the ball.
- Let us set this prize just great enough so that he can break even, by playing one more stroke, and then quitting thereafter whenever he likes.

  $g(x) =$ fair prize'.
- If the ball arrives at a location y, from which $g(x)$ is no longer great enough to motivate the golfer to continue playing, then, — just as he is about to quit —, we increase the prize to $g(y)$, which becomes the new 'prevailing prize'.
- Continue doing this until the ball is sunk.
- This presents the golfer with a fair game, and it is optimal for him to keep playing until the ball is sunk.

$$E(\text{cost incurred}) = E(\text{prize won})$$

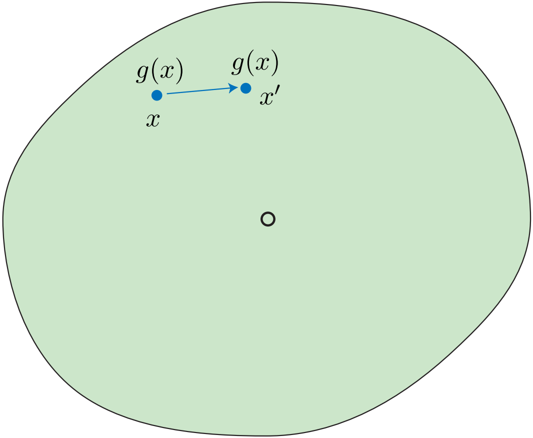
Golf with 1 Ball

$g(x) = 3.0$



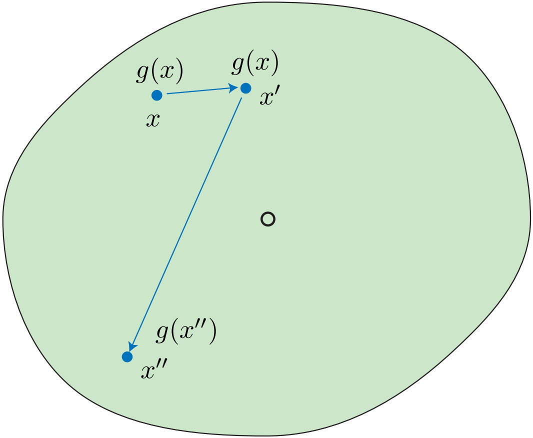
Golf with 1 Ball

$g(x) = 3.0$, $g(x') = 2.5$



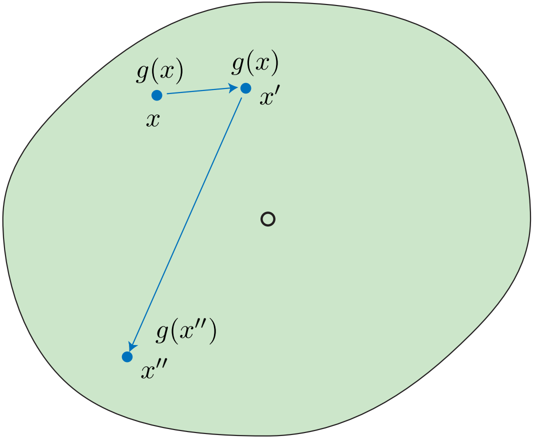
Golf with 1 Ball

$g(x) = 3.0,\ g(x') = 2.5,\ g(x'') = 4.0$



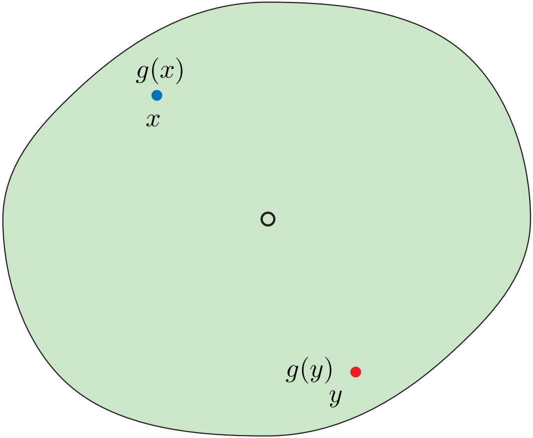
Golf with 1 Ball

$g(x) = 3.0$, $g(x') = 2.5$, $g(x'') = 4.0$
Prevailing prize sequence is 3.0, 3.0, 4.0, $\ldots$



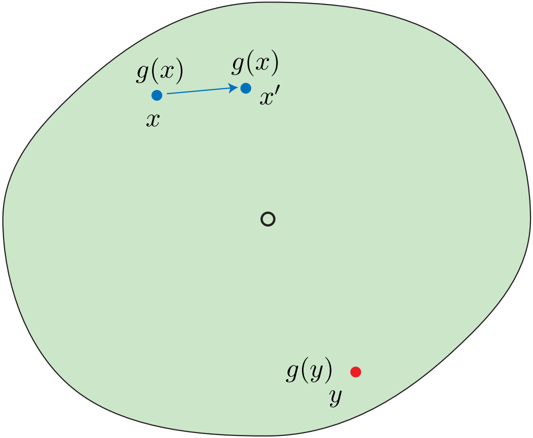
Golf with 2 Balls

$g(x) = \overline{3.0}$
$g(y) = \overline{3.2}$



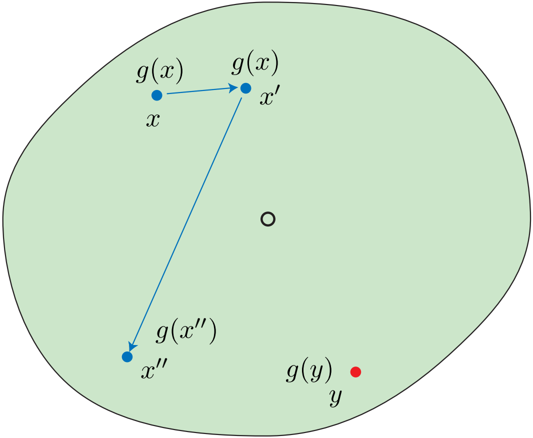
Golf with 2 Balls

$g(x) = \overline{3.0}$, $g(x') = 2.5$
$g(y) = \overline{3.2}$



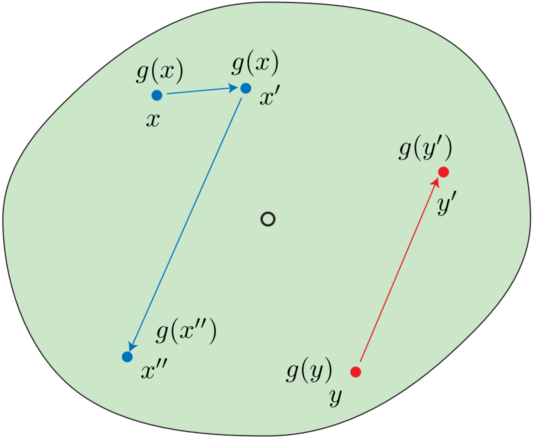
Golf with 2 Balls

$g(x) = 3.0,\ g(x') = 2.5,\ g(x'') = \overline{4.0}$
$g(y) = \overline{3.2}$



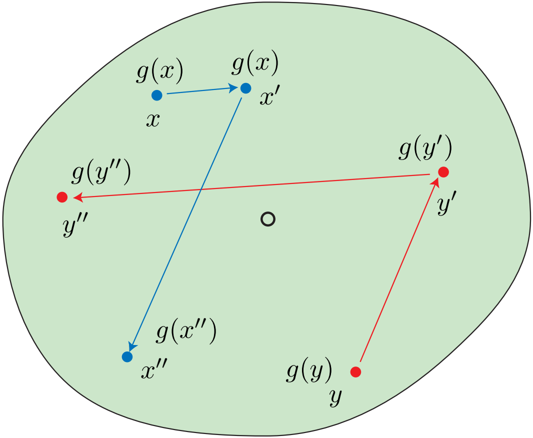
Golf with 2 Balls

$g(x) = 3.0,\ g(x') = 2.5,\ g(x'') = \overline{4.0}$
$g(y) = 3.2,\ g(y') = \overline{3.5}$



Golf with 2 Balls

$g(x) = 3.0$, $g(x') = 2.5$, $g(x'') = \overline{4.0}$
$g(y) = 3.2$, $g(y') = 3.5$, $g(y'') = \overline{4.2}$



Optimal Play with N Balls

- Each of the N balls has an initial 'prevailing prize', $\bar{g}_i(0)$, attached to it. $\bar{g}_i(0) = g(x_i(0))$.

Optimal Play with N Balls

- Each of the N balls has an initial 'prevailing prize', $\bar{g}_i(0)$, attached to it. $\bar{g}_i(0) = g(x_i(0))$.
- Prevailing prize, $\bar{g}_i(t)$, is increased whenever it is insufficient to motivate golfer to play that ball; so it is nondecreasing.

Optimal Play with N Balls

- Each of the N balls has an initial 'prevailing prize', $\bar{g}_i(0)$, attached to it. $\bar{g}_i(0) = g(x_i(0))$.
- Prevailing prize, $\bar{g}_i(t)$, is increased whenever it is insufficient to motivate golfer to play that ball; so it is nondecreasing.
- $\bar{g}_i(t)$ depends only on the ball's history, not what has happened to other balls. $\bar{g}_i(t) = \max_{s \le t} g(x_i(s))$.

Optimal Play with N Balls

- Each of the N balls has an initial 'prevailing prize', $\bar{g}_i(0)$, attached to it. $\bar{g}_i(0) = g(x_i(0))$.
- Prevailing prize, $\bar{g}_i(t)$, is increased whenever it is insufficient to motivate golfer to play that ball; so it is nondecreasing.
- $\bar{g}_i(t)$ depends only on the ball's history, not what has happened to other balls. $\bar{g}_i(t) = \max_{s \le t} g(x_i(s))$.
- Game is fair. It is impossible for golfer to make a strictly positive profit (since he would have to do so for some ball).

$$E(\text{cost incurred}) \geq E(\text{prize won})$$

Optimal Play with N Balls

- Each of the N balls has an initial 'prevailing prize', $\bar{g}_i(0)$, attached to it. $\bar{g}_i(0) = g(x_i(0))$.
- Prevailing prize, $\bar{g}_i(t)$, is increased whenever it is insufficient to motivate golfer to play that ball; so it is nondecreasing.
- $\bar{g}_i(t)$ depends only on the ball's history, not what has happened to other balls. $\bar{g}_i(t) = \max_{s \le t} g(x_i(s))$.
- Game is fair. It is impossible for golfer to make a strictly positive profit (since he would have to do so for some ball).

$$E(\text{cost incurred}) \ge E(\text{prize won})$$

- Equality is achieved provided golfer does not switch away from a ball unless its prevailing prize increases.

Optimal Play with N Balls

- Each of the N balls has an initial 'prevailing prize', $\bar{g}_i(0)$, attached to it. $\bar{g}_i(0) = g(x_i(0))$.
- Prevailing prize, $\bar{g}_i(t)$, is increased whenever it is insufficient to motivate golfer to play that ball; so it is nondecreasing.
- $\bar{g}_i(t)$ depends only on the ball's history, not what has happened to other balls. $\bar{g}_i(t) = \max_{s \leq t} g(x_i(s))$.
- Game is fair. It is impossible for golfer to make a strictly positive profit (since he would have to do so for some ball).

$$E(\text{cost incurred}) \geq E(\text{prize won})$$

- Equality is achieved provided golfer does not switch away from a ball unless its prevailing prize increases.
- Right hand side is minimized by always playing ball with least prevailing prize.

Golf and the Multi-armed Bandit

Having solved the golf problem, the solution to the multi-armed bandit problem follows. Just let $P(x, 0) = 1 - \beta$ for all x.

The expected cost incurred until a first ball is sunk equals the expected total β-discounted cost over the infinite horizon.

Golf and the Multi-armed Bandit

Having solved the golf problem, the solution to the multi-armed bandit problem follows. Just let $P(x, 0) = 1 - \beta$ for all x.

The expected cost incurred until a first ball is sunk equals the expected total β-discounted cost over the infinite horizon.

The fair prize, $g(x)$, is $1/(1 - \beta)$ times the Gittins index, $G(x)$.

Golf and the Multi-armed Bandit

Having solved the golf problem, the solution to the multi-armed bandit problem follows. Just let $P(x,0) = 1-\beta$ for all x.

The expected cost incurred until a first ball is sunk equals the expected total β-discounted cost over the infinite horizon.

The fair prize, $g(x)$, is $1/(1-\beta)$ times the Gittins index, $G(x)$.

$$g(x) = \inf\left\{ g \;:\; \sup_{\tau \geq 1} E\left[\sum_{t=0}^{\tau-1} -c(x(t))\beta^t \right.\right.$$
$$\left.\left. + (1-\beta)(1+\beta+\cdots+\beta^{\tau-1})g \,\middle|\, x(0)=x \right] \geq 0 \right\}$$

Golf and the Multi-armed Bandit

Having solved the golf problem, the solution to the multi-armed bandit problem follows. Just let $P(x,0) = 1 - \beta$ for all x.

The expected cost incurred until a first ball is sunk equals the expected total β-discounted cost over the infinite horizon.

The fair prize, $g(x)$, is $1/(1-\beta)$ times the Gittins index, $G(x)$.

$$g(x) = \inf\left\{ g \ : \ \sup_{\tau \geq 1} E\left[\sum_{t=0}^{\tau-1} -c(x(t))\beta^t \right.\right.$$

$$\left.\left. + (1-\beta)(1+\beta+\cdots+\beta^{\tau-1})g \,\middle|\, x(0) = x \right] \geq 0 \right\}$$

$$= \frac{1}{1-\beta} \inf_{\tau \geq 1} \frac{E\left[\sum_{t=0}^{\tau-1} c(x(t))\beta^t \,\middle|\, x(0) = x\right]}{E\left[\sum_{t=0}^{\tau-1} \beta^t \,\middle|\, x(0) = x\right]}$$

Golf with N Balls and a Set of Clubs

Suppose that a ball in location x can be played with a choice of shots, from a set $A(x)$. Choosing shot $a \in A(x)$,

$$x \to y \quad \text{with probability } P_a(x, y)$$

Now the golfer must choose, not only which ball to play, but with which shot to play it.

Golf with N Balls and a Set of Clubs

Suppose that a ball in location x can be played with a choice of shots, from a set $A(x)$. Choosing shot $a \in A(x)$,

$$x \to y \quad \text{with probability } P_a(x, y)$$

Now the golfer must choose, not only which ball to play, but with which shot to play it.

Under a condition, an index policy is again optimal.

He should play the ball with least prevailing prize, choosing the shot from A that is optimal if that ball were the only ball present.

What has Happened Since 1989?

- Index theorem has become better known.
- Alternative proofs have been explored.

  Playing golf with N balls

  Achievable Performance Region Approach
- Many applications (economics, engineering, . . .).
- Notions of indexation have been generalized.

  Restless Bandits

Achievable Performance Region Approach

Suppose all arms move on state space $E = \{1, \dots, N\}$.
Let $I_i(t)$ be an indicator for the event that at time t an arm is pulled that is in state i.
We wish to maximize (conditional on the starting states of arms)

$$E_\pi \left[\sum_{i \in E} r_i \sum_{t=0}^{\infty} I_i(t) \beta^t \right]$$

Achievable Performance Region Approach

Suppose all arms move on state space $E = \{1, \ldots, N\}$.
Let $I_i(t)$ be an indicator for the event that at time t an arm is pulled that is in state i.
We wish to maximize (conditional on the starting states of arms)

$$E_\pi \left[\sum_{i \in E} r_i \sum_{t=0}^{\infty} I_i(t) \beta^t \right]$$

Suppose that under policy π,

$$z_i^\pi = E_\pi \left[\sum_{t=0}^{\infty} I_i(t) \beta^t \right]$$

We wish to maximize $\sum_{i \in E} r_i z_i^\pi$.

Some Conservation Laws

Consider a MABP with $r_i = 1$ for all i. This shows that for all π.

$$\sum_{i \in E} z_i^\pi = 1 + \beta + \beta^2 + \cdots = \frac{1}{1-\beta}$$

Some Conservation Laws

Consider a MABP with $r_i = 1$ for all i. This shows that for all π.

$$\sum_{i \in E} A_i^E z_i^\pi = 1 + \beta + \beta^2 + \cdots = \frac{1}{1-\beta}$$

Some Conservation Laws

Consider a MABP with $r_i = 1$ for all i. This shows that for all π.

$$\sum_{i \in E} A_i^E z_i^\pi = 1 + \beta + \beta^2 + \cdots = \frac{1}{1-\beta}$$

Now pick a subset of states $S \subset E = \{1, \ldots, N\}$.

Some Conservation Laws

Consider a MABP with $r_i = 1$ for all i. This shows that for all π.

$$\sum_{i \in E} A_i^E z_i^\pi = 1 + \beta + \beta^2 + \cdots = \frac{1}{1-\beta}$$

Now pick a subset of states $S \subset E = \{1, \ldots, N\}$. Let

- $T_i^S =$ 'number of pulls needed for state to return to S from i'.

  $A_i^S = E\left[1 + \beta + \cdots + \beta^{T_i^S - 1}\right]$.

Some Conservation Laws

Consider a MABP with $r_i = 1$ for all i. This shows that for all π.

$$\sum_{i \in E} A_i^E z_i^\pi = 1 + \beta + \beta^2 + \cdots = \frac{1}{1-\beta}$$

Now pick a subset of states $S \subset E = \{1, \ldots, N\}$. Let

- $T_i^S =$ 'number of pulls needed for state to return to S from i'.

  $A_i^S = E\left[1 + \beta + \cdots + \beta^{T_i^S - 1}\right]$.

- $r_i^S = 0$, $i \notin S$, and $r_i^S = A_i^S$, $i \in S$.

Some Conservation Laws

Consider a MABP with $r_i = 1$ for all i. This shows that for all π.

$$\sum_{i \in E} A_i^E z_i^\pi = 1 + \beta + \beta^2 + \cdots = \frac{1}{1-\beta}$$

Now pick a subset of states $S \subset E = \{1, \ldots, N\}$. Let

- $T_i^S =$ 'number of pulls needed for state to return to S from i'. $A_i^S = E\left[1 + \beta + \cdots + \beta^{T_i^S - 1}\right]$.
- $r_i^S = 0$, $i \notin S$, and $r_i^S = A_i^S$, $i \in S$.

This is a near-trivial MABP. Easy to show $\sum_i r_i^S z_i^\pi$ minimized by any policy that gives priority to arms whose states are not in S. So

$$\sum_{i \in E} A_i^S z_i^\pi \geq \min_\pi \left\{ \sum_{i \in E} A_i^S z_i^\pi \right\}$$

Some Conservation Laws

Consider a MABP with $r_i = 1$ for all i. This shows that for all π.

$$\sum_{i \in E} A_i^E z_i^\pi = 1 + \beta + \beta^2 + \cdots = \frac{1}{1-\beta}$$

Now pick a subset of states $S \subset E = \{1, \ldots, N\}$. Let

- $T_i^S =$ 'number of pulls needed for state to return to S from i'.
  $A_i^S = E\left[1 + \beta + \cdots + \beta^{T_i^S - 1}\right]$.
- $r_i^S = 0$, $i \notin S$, and $r_i^S = A_i^S$, $i \in S$.

This is a near-trivial MABP. Easy to show $\sum_i r_i^S z_i^\pi$ minimized by any policy that gives priority to arms whose states are not in S. So

$$\sum_{i \in E} A_i^S z_i^\pi \geq \min_\pi \left\{ \sum_{i \in E} A_i^S z_i^\pi \right\}$$

Some Conservation Laws

Consider a MABP with $r_i = 1$ for all i. This shows that for all π.

$$\sum_{i \in E} A_i^E z_i^\pi = 1 + \beta + \beta^2 + \cdots = \frac{1}{1-\beta} = b(E)$$

Now pick a subset of states $S \subset E = \{1, \ldots, N\}$. Let

- $T_i^S =$ 'number of pulls needed for state to return to S from i'. $A_i^S = E\left[1 + \beta + \cdots + \beta^{T_i^S - 1}\right]$.
- $r_i^S = 0$, $i \notin S$, and $r_i^S = A_i^S$, $i \in S$.

This is a near-trivial MABP. Easy to show $\sum_i r_i^S z_i^\pi$ minimized by any policy that gives priority to arms whose states are not in S. So

$$\sum_{i \in E} A_i^S z_i^\pi \geq \min_\pi \left\{ \sum_{i \in E} A_i^S z_i^\pi \right\} = b(S)$$

Constraints on the Achieveable Region

Lemma

There exist positive A_i^S, as defined above, such that for any scheduling policy π,

$$\sum_{i \in S} A_i^S z_i^\pi \geq b(S)\,, \text{ for all } S \subset E, \tag{1}$$

$$\sum_{i \in E} A_i^E z_i^\pi = b(E)\,, \tag{2}$$

and such that equality holds in (1) if π gives priority to arms whose states are not in S over any arms whose states are in S.

A Linear Programming Relaxation

Primal

$$\underset{\{z_i\}}{\text{maximize}} \sum_{i \in E} r_i z_i$$

$$\sum_{i \in S} A_i^S z_i \geq b(S)\,, \text{ for all } S \subset E,$$

$$\sum_{i \in S} A_i^E z_i = b(E)\,,$$

$$z_i \geq 0\,, \text{ for all } i\,.$$

A Linear Programming Relaxation

Primal

$$\underset{\{z_i\}}{\text{maximize}} \sum_{i \in E} r_i z_i$$

$$\sum_{i \in S} A_i^S z_i \geq b(S)\,, \text{ for all } S \subset E,$$

$$\sum_{i \in S} A_i^E z_i = b(E)\,,$$

$$z_i \geq 0\,, \text{ for all } i\,.$$

The optimal value of this LP is an upper bound on the optimal value for our bandit problem.

A Linear Programming Relaxation

Primal

$$\underset{\{z_i\}}{\text{maximize}} \sum_{i \in E} r_i z_i$$

$$\sum_{i \in S} A_i^S z_i \geq b(S)\,, \text{ for all } S \subset E,$$

$$\sum_{i \in S} A_i^E z_i = b(E)\,,$$

$$z_i \geq 0\,, \text{ for all } i\,.$$

Dual

$$\underset{\{y_S\}}{\text{minimize}} \sum_{S} y_S b(S)$$

$$\sum_{S\,:\,i \in S} y_S A_i^S \geq r_i\,, \text{ for all } i,$$

$$y_S \leq 0\,, \text{ for all } S \subset E\,,$$

y_E unrestricted in sign.

The optimal value of this LP is an upper bound on the optimal value for our bandit problem.

A Linear Programming Relaxation

Primal

$$\underset{\{z_i\}}{\text{maximize}} \sum_{i \in E} r_i z_i$$

$$\sum_{i \in S} A_i^S z_i \geq b(S)\,, \text{ for all } S \subset E,$$

$$\sum_{i \in S} A_i^E z_i = b(E)\,,$$

$$z_i \geq 0\,, \text{ for all } i\,.$$

Dual

$$\underset{\{y_S\}}{\text{minimize}} \sum_{S} y_S b(S)$$

$$\sum_{S\,:\,i \in S} y_S A_i^S \geq r_i\,, \text{ for all } i,$$

$$y_S \leq 0\,, \text{ for all } S \subset E\,,$$

y_E unrestricted in sign.

The optimal value of this LP is an upper bound on the optimal value for our bandit problem.

A greedy algorithm computes dual vectors $\bar{y}_S$ that are dual feasible and complementary slack, and a primal solution $\bar{z}_i = z_i^\pi$ which is the performance vector of a priority policy.

A Linear Programming Relaxation

Primal

$$\underset{\{z_i\}}{\text{maximize}} \sum_{i \in E} r_i z_i$$

$$\sum_{i \in S} A_i^S z_i \geq b(S)\,, \text{ for all } S \subset E,$$

$$\sum_{i \in S} A_i^E z_i = b(E)\,,$$

$$z_i \geq 0\,, \text{ for all } i\,.$$

Dual

$$\underset{\{y_S\}}{\text{minimize}} \sum_{S} y_S b(S)$$

$$\sum_{S\,:\,i \in S} y_S A_i^S \geq r_i\,, \text{ for all } i,$$

$$y_S \leq 0\,, \text{ for all } S \subset E\,,$$

y_E unrestricted in sign.

The optimal value of this LP is an upper bound on the optimal value for our bandit problem.

A greedy algorithm computes dual vectors $\bar{y}_S$ that are dual feasible and complementary slack, and a primal solution $\bar{z}_i = z_i^{\pi}$ which is the performance vector of a priority policy.

A Linear Programming Relaxation

Primal

$$\underset{\{z_i\}}{\text{maximize}} \sum_{i \in E} r_i z_i$$

$$\sum_{i \in S} A_i^S z_i \geq b(S)\,, \text{ for all } S \subset E,$$

$$\sum_{i \in S} A_i^E z_i = b(E)\,,$$

$$z_i \geq 0\,, \text{ for all } i\,.$$

Dual

$$\underset{\{y_S\}}{\text{minimize}} \sum_{S} y_S b(S)$$

$$\sum_{S\,:\,i \in S} y_S A_i^S \geq r_i\,, \text{ for all } i,$$

$$y_S \leq 0\,, \text{ for all } S \subset E\,,$$

y_E unrestricted in sign.

The optimal value of this LP is an upper bound on the optimal value for our bandit problem.

A greedy algorithm computes dual vectors $\bar{y}_S$ that are dual feasible and complementary slack, and a primal solution $\bar{z}_i = z_i^\pi$ which is the performance vector of a priority policy.

Greedy Algorithm

Dual has $2^N - 1$ variables, y_S, but only N of them are non-zero.
They can be computed one by one: $\bar{y}_E, \bar{y}_{S_2}, \bar{y}_{S_3}, \ldots, \bar{y}_{S_N}$.

Greedy Algorithm

Dual has $2^N - 1$ variables, y_S, but only N of them are non-zero.
They can be computed one by one: $\bar{y}_E, \bar{y}_{S_2}, \bar{y}_{S_3}, \ldots, \bar{y}_{S_N}$.

Input: r_i, p_{ij}, $i, j \in E$.

Initialization:

let $\bar{y}_E := \max\{r_i / A_i^E \,:\, i \in E\}$.

Greedy Algorithm

Dual has $2^N - 1$ variables, y_S, but only N of them are non-zero.
They can be computed one by one: $\bar{y}_E, \bar{y}_{S_2}, \bar{y}_{S_3}, \ldots, \bar{y}_{S_N}$.

Input: r_i, p_{ij}, $i, j \in E$.

Initialization:

let $\bar{y}_E := \max\{r_i / A_i^E \,:\, i \in E\}$.

choose $i_1 \in E$ attaining the above maximum

Greedy Algorithm

Dual has $2^N - 1$ variables, y_S, but only N of them are non-zero.
They can be computed one by one: $\bar{y}_E, \bar{y}_{S_2}, \bar{y}_{S_3}, \ldots, \bar{y}_{S_N}$.

Input: r_i, p_{ij}, $i, j \in E$.

Initialization:

let $\bar{y}_E := \max\{r_i / A_i^E \,:\, i \in E\}$.
choose $i_1 \in E$ attaining the above maximum
set $\bar{y}_S = 0$ for all S, s.t. $i_1 \in S \subset E$.

So dual constraint, $\sum_{S\,:\, i_1 \in S} y_S A_i^S \geq r_{i_1}$, holds with equality.

Greedy Algorithm

Dual has $2^N - 1$ variables, y_S, but only N of them are non-zero.
They can be computed one by one: $\bar{y}_E, \bar{y}_{S_2}, \bar{y}_{S_3}, \ldots, \bar{y}_{S_N}$.

Input: r_i, p_{ij}, $i, j \in E$.

Initialization:

let $\bar{y}_E := \max\{r_i / A_i^E \, : \, i \in E\}$.
choose $i_1 \in E$ attaining the above maximum
set $\bar{y}_S = 0$ for all S, s.t. $i_1 \in S \subset E$.
let $g_{i_1} := \bar{y}_E$
let $S_1 := E$; let $k := 2$

Greedy Algorithm

Dual has $2^N - 1$ variables, y_S, but only N of them are non-zero. They can be computed one by one: $\bar{y}_E, \bar{y}_{S_2}, \bar{y}_{S_3}, \ldots, \bar{y}_{S_N}$.

Input: r_i, p_{ij}, $i, j \in E$.

Initialization:

let $\bar{y}_E := \max\{r_i / A_i^E \,:\, i \in E\}$.

choose $i_1 \in E$ attaining the above maximum

set $\bar{y}_S = 0$ for all S, s.t. $i_1 \in S \subset E$.

let $g_{i_1} := \bar{y}_E$

let $S_1 := E$; let $k := 2$

Loop: while $k \leq N$ do

let $S_k := S_{k-1} \setminus \{i_{k-1}\}$.

let $\bar{y}_{S_k} := \max\left\{ \left(r_i - \sum_{j=1}^{k-1} A_i^{S_j} \right) / A_i^{S_k} \,:\, i \in S_k \right\}$

choose $i_k \in S_k$ attaining the above maximum

Greedy Algorithm

Dual has $2^N - 1$ variables, y_S, but only N of them are non-zero. They can be computed one by one: $\bar{y}_E, \bar{y}_{S_2}, \bar{y}_{S_3}, \ldots, \bar{y}_{S_N}$.

Input: r_i, p_{ij}, $i, j \in E$.

Initialization:

let $\bar{y}_E := \max\{r_i / A_i^E \,:\, i \in E\}$.

choose $i_1 \in E$ attaining the above maximum

set $\bar{y}_S = 0$ for all S, s.t. $i_1 \in S \subset E$.

let $g_{i_1} := \bar{y}_E$

let $S_1 := E$; let $k := 2$

Loop: while $k \leq N$ do

let $S_k := S_{k-1} \setminus \{i_{k-1}\}$.

let $\bar{y}_{S_k} := \max\left\{ \left(r_i - \sum_{j=1}^{k-1} A_i^{S_j} \right) / A_i^{S_k} \,:\, i \in S_k \right\}$

choose $i_k \in S_k$ attaining the above maximum

set $\bar{y}_S = 0$ for all S, s.t. $i_k \in S \subset S_k$.

So dual constraint, $\sum_{S\,:\,i_k \in S} y_S A_i^S \geq r_{i_k}$, holds with equality.

Greedy Algorithm

Dual has $2^N - 1$ variables, y_S, but only N of them are non-zero. They can be computed one by one: $\bar{y}_E, \bar{y}_{S_2}, \bar{y}_{S_3}, \ldots, \bar{y}_{S_N}$.

Input: r_i, p_{ij}, $i, j \in E$.

Initialization:

let $\bar{y}_E := \max\{r_i / A_i^E \,:\, i \in E\}$.

choose $i_1 \in E$ attaining the above maximum

set $\bar{y}_S = 0$ for all S, s.t. $i_1 \in S \subset E$.

let $g_{i_1} := \bar{y}_E$

let $S_1 := E$; let $k := 2$

Loop: while $k \leq N$ do

let $S_k := S_{k-1} \setminus \{i_{k-1}\}$.

let $\bar{y}_{S_k} := \max\left\{ \left(r_i - \sum_{j=1}^{k-1} A_i^{S_j} \right) / A_i^{S_k} \,:\, i \in S_k \right\}$

choose $i_k \in S_k$ attaining the above maximum

set $\bar{y}_S = 0$ for all S, s.t. $i_k \in S \subset S_k$.

let $g_{i_k} := g_{i_{k-1}} + \bar{y}_{S_k}$

let $k := k + 1$

end {while}

What has Happened Since 1989?

- Index theorem has become better known.
- Alternative proofs have been explored.

  Playing golf with N balls

  Achievable Performance Region Approach
- Many applications (economics, engineering, . . .).
- Notions of indexation have been generalized.

  Restless Bandits

Restless Bandits

Spinning Plates

Restless Bandits

[Whittle '88]

- Two actions are available: **active** $(a = 1)$ or **passive** $(a = 0)$.

Restless Bandits

[Whittle '88]

- Two actions are available: **active** $(a = 1)$ or **passive** $(a = 0)$.
- Rewards, $r(x, a)$, and transitions, $P(y \mid x, a)$, depend on the state and the action taken.

Restless Bandits

[Whittle '88]

- Two actions are available: **active** $(a = 1)$ or **passive** $(a = 0)$.
- Rewards, $r(x, a)$, and transitions, $P(y \,|\, x, a)$, depend on the state and the action taken.
- **Objective**: Maximize time-average reward from n restless bandits under a constraint that only m $(m < n)$ of them receive the active action simultaneously.

Restless Bandits

[Whittle '88]

- Two actions are available: **active** $(a = 1)$ or **passive** $(a = 0)$.
- Rewards, $r(x, a)$, and transitions, $P(y \mid x, a)$, depend on the state and the action taken.
- **Objective**: Maximize time-average reward from n restless bandits under a constraint that only m $(m < n)$ of them receive the active action simultaneously.

active $a = 1$	passive $a = 0$
work, increasing fatigue	rest, recovery

Restless Bandits

[Whittle '88]

- Two actions are available: **active** ($a = 1$) or **passive** ($a = 0$).
- Rewards, $r(x, a)$, and transitions, $P(y \,|\, x, a)$, depend on the state and the action taken.
- **Objective**: Maximize time-average reward from n restless bandits under a constraint that only m ($m < n$) of them receive the active action simultaneously.

active $a = 1$	passive $a = 0$
work, increasing fatigue	rest, recovery
high speed	low speed

$$P(y \,|\, x, 0) = \epsilon P(y \,|\, x, 1)\,, \quad y \neq x$$

Restless Bandits

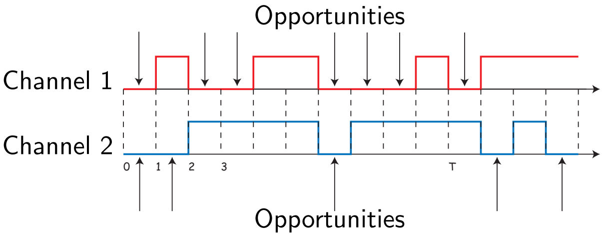
[Whittle '88]

- Two actions are available: **active** $(a = 1)$ or **passive** $(a = 0)$.
- Rewards, $r(x, a)$, and transitions, $P(y \mid x, a)$, depend on the state and the action taken.
- **Objective**: Maximize time-average reward from n restless bandits under a constraint that only m $(m < n)$ of them receive the active action simultaneously.

active $a = 1$	passive $a = 0$
work, increasing fatigue	rest, recovery
high speed	low speed
inspection	no inspection

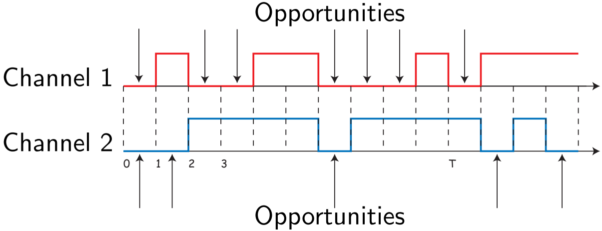
Opportunistic Spectrum Access

Communication channels may be busy or free.



Opportunistic Spectrum Access

Communication channels may be busy or free.



Aim is to ‘inspect’ m out of n channels, maximizing the number of these that are found to be free.

Opportunistic Spectrum Access

'Condition' of the channel (busy or free) evolves as Markov chain.

$x(t) = P(\text{channel is free at time } t)$.

Opportunistic Spectrum Access

'Condition' of the channel (busy or free) evolves as Markov chain.

$x(t) = P($channel is free at time $t)$.

$$a = 0: \quad x(t+1) = x(t)p_{11} + (1 - x(t))p_{01}$$

Opportunistic Spectrum Access

'Condition' of the channel (busy or free) evolves as Markov chain.

$x(t) = P(\text{channel is free at time } t)$.

$$a = 0: \quad x(t+1) = x(t)p_{11} + (1 - x(t))p_{01}$$

$$a = 1: \quad x(t+1) = \begin{cases} p_{01} \\ p_{11} \end{cases} \text{with probability} \quad \begin{matrix} 1 - x(t) \\ x(t) \end{matrix}$$

Dynamic Programming Equation

Action set is $\Omega = \{(a_1, \ldots, a_n) \,:\, a_i \in \{0, 1\},\ \sum_i a_i = m\}$.

For a state $x = (x_1, \ldots, x_n)$,

$$V(x) = \max_{a \in \Omega} \left\{ \sum_i r(x_i, a_i) + \beta \sum_{y_1, \ldots, y_n} V(y_1, \ldots, y_n) \prod_i P(y_i \,|\, x_i, a_i) \right\}$$

Relaxed Problem for a Single Restless Bandit

Let us consider a **relaxed problem**, posed for 1 bandit only.

The aim is to maximize average reward obtained from this bandit under a constraint that $a = 1$ for only a fraction m/n of the time.

LP for the Relaxed Problem

Let z_x^a be proportion of time that the bandit is in state x and action a is taken (under a stationary Markov policy).

An upper bound for our problem can found from a LP in variables $\{z_x^a : x \in E,\ a \in \{0,1\}\}$:

$$\text{maximize } \sum_{x,a} r(x,a) z_x^a$$

LP for the Relaxed Problem

Let z_x^a be proportion of time that the bandit is in state x and action a is taken (under a stationary Markov policy).

An upper bound for our problem can found from a LP in variables $\{z_x^a : x \in E,\ a \in \{0,1\}\}$:

$$\text{maximize } \sum_{x,a} r(x,a) z_x^a$$

$$\text{s.t. } z_x^a \geq 0\,,\ \text{for all } x, a$$

LP for the Relaxed Problem

Let z_x^a be proportion of time that the bandit is in state x and action a is taken (under a stationary Markov policy).

An upper bound for our problem can found from a LP in variables $\{z_x^a : x \in E,\ a \in \{0,1\}\}$:

$$\text{maximize } \sum_{x,a} r(x,a) z_x^a$$

$$\text{s.t. } z_x^a \geq 0\,,\ \text{for all } x, a\,;\ \sum_{x,a} z_x^a = 1$$

LP for the Relaxed Problem

Let z_x^a be proportion of time that the bandit is in state x and action a is taken (under a stationary Markov policy).

An upper bound for our problem can found from a LP in variables $\{z_x^a \,:\, x \in E,\ a \in \{0,1\}\}$:

$$\text{maximize } \sum_{x,a} r(x,a) z_x^a$$

$$\text{s.t. } z_x^a \geq 0\,,\ \text{for all } x, a\,;\ \sum_{x,a} z_x^a = 1\,;$$

$$\sum_a z_x^a = \sum_y z_y^a P(x \,|\, y, a(y))\,,\ \text{for all } x$$

LP for the Relaxed Problem

Let z_x^a be proportion of time that the bandit is in state x and action a is taken (under a stationary Markov policy).

An upper bound for our problem can found from a LP in variables $\{z_x^a \,:\, x \in E,\ a \in \{0,1\}\}$:

$$\text{maximize } \sum_{x,a} r(x,a) z_x^a$$

$$\text{s.t. } z_x^a \geq 0\,, \text{ for all } x, a\,; \ \sum_{x,a} z_x^a = 1\,;$$

$$\sum_a z_x^a = \sum_y z_y^a P(x \,|\, y, a(y))\,, \text{ for all } x\,; \quad \sum_x z_x^0 = 1 - m/n\,.$$

The Subsidy Problem

Optimal value of the dual LP problem is g, where this can be found from the average-cost dynamic programming equation

$$\phi(x) + g = \max_{a \in \{0,1\}} \left\{ r(x,a) + \lambda(1-a) + \sum_y \phi(y) P(y \mid x, a) \right\}.$$

λ and $\phi(x)$ are the Lagrange multipliers for constraints.

λ may be interpreted as a *subsidy* for taking $a = 0$.

The Subsidy Problem

Optimal value of the dual LP problem is g, where this can be found from the average-cost dynamic programming equation

$$\phi(x) + g = \max_{a \in \{0,1\}} \left\{ r(x,a) + \lambda(1-a) + \sum_y \phi(y) P(y \mid x, a) \right\}.$$

λ and $\phi(x)$ are the Lagrange multipliers for constraints.

λ may be interpreted as a *subsidy* for taking $a = 0$.

Solution partitions state space into sets: E_0 $(a = 0)$, E_1 $(a = 1)$ and E_{01} (randomization between $a = 0$ and $a = 1$).

Indexability

Reasonable that as the subsidy λ (for $a = 0$) increases from $-\infty$ to $+\infty$ the set of states E_0 (where $a = 0$ optimal) should increase monotonically.

If it does then we say the bandit is **indexable**.

Indexability

Reasonable that as the subsidy λ (for $a = 0$) increases from $-\infty$ to $+\infty$ the set of states E_0 (where $a = 0$ optimal) should increase monotonically.

If it does then we say the bandit is **indexable**.

Whittle index, $W(x)$, is the least subsidy for which it can be optimal to take $a = 0$ in state x.

Indexability

Reasonable that as the subsidy λ (for $a = 0$) increases from $-\infty$ to $+\infty$ the set of states E_0 (where $a = 0$ optimal) should increase monotonically.

If it does then we say the bandit is **indexable**.

Whittle index, $W(x)$, is the least subsidy for which it can be optimal to take $a = 0$ in state x.

This motivates a heuristic policy:
use active action on the m bandits with the greatest Whittle indices.

Indexability

Reasonable that as the subsidy λ (for $a = 0$) increases from $-\infty$ to $+\infty$ the set of states E_0 (where $a = 0$ optimal) should increase monotonically.

If it does then we say the bandit is **indexable**.

Whittle index, $W(x)$, is the least subsidy for which it can be optimal to take $a = 0$ in state x.

This motivates a heuristic policy:
use active action on the m bandits with the greatest Whittle indices.

Like Gittins indices for classical bandits, Whittle indices can be computed separately for each bandit.
Same as the Gittins index when $a = 0$ is freezing action.

Two Questions

- **Under what assumptions is a restless bandit indexable?**

Two Questions

- **Under what assumptions is a restless bandit indexable?**

  This is somewhat mysterious.

  Special classes of restless bandits are indexable: such as 'dual speed', Glazebrook, Niño-Mora, Ansell (2002), W. (2007).

Two Questions

- **Under what assumptions is a restless bandit indexable?**

  This is somewhat mysterious.

  Special classes of restless bandits are indexable: such as 'dual speed', Glazebrook, Niño-Mora, Ansell (2002), W. (2007).

  Indexability can be proved in some problems (such as the opportunistic spectrum access problem, Liu and Zhao (2009)).

Two Questions

- **Under what assumptions is a restless bandit indexable?**

  This is somewhat mysterious.

  Special classes of restless bandits are indexable: such as 'dual speed', Glazebrook, Niño-Mora, Ansell (2002), W. (2007).

  Indexability can be proved in some problems (such as the opportunistic spectrum access problem, Liu and Zhao (2009)).

- **How good is the heuristic policy using Whittle indices?**

Two Questions

- **Under what assumptions is a restless bandit indexable?**

  This is somewhat mysterious.

  Special classes of restless bandits are indexable: such as 'dual speed', Glazebrook, Niño-Mora, Ansell (2002), W. (2007).

  Indexability can be proved in some problems (such as the opportunistic spectrum access problem, Liu and Zhao (2009)).

- **How good is the heuristic policy using Whittle indices?**

  It may be optimal. (opportunistic spectrum access — identical channels, Ahmad, Liu, Javidi, and Zhao (2009)).

Two Questions

- **Under what assumptions is a restless bandit indexable?**

  This is somewhat mysterious.

  Special classes of restless bandits are indexable: such as 'dual speed', Glazebrook, Niño-Mora, Ansell (2002), W. (2007).

  Indexability can be proved in some problems (such as the opportunistic spectrum access problem, Liu and Zhao (2009)).

- **How good is the heuristic policy using Whittle indices?**

  It may be optimal. (opportunistic spectrum access — identical channels, Ahmad, Liu, Javidi, and Zhao (2009)).

  Lots of papers with numerical work.

Two Questions

- **Under what assumptions is a restless bandit indexable?**

  This is somewhat mysterious.

  Special classes of restless bandits are indexable: such as 'dual speed', Glazebrook, Niño-Mora, Ansell (2002), W. (2007).

  Indexability can be proved in some problems (such as the opportunistic spectrum access problem, Liu and Zhao (2009)).

- **How good is the heuristic policy using Whittle indices?**

  It may be optimal. (opportunistic spectrum access — identical channels, Ahmad, Liu, Javidi, and Zhao (2009)).

  Lots of papers with numerical work.

  It is often asymptotically optimal, W. and Weiss (1990).

Asymptotic Optimality

Suppose a priority policy orders the states $1, 2, \dots$.
At time t there are $(n_1, \dots, n_k)$ bandits in states $1, \dots, k$. Let $m = \rho n$.

Asymptotic Optimality

Suppose a priority policy orders the states $1, 2, \ldots$.
At time t there are $(n_1, \ldots, n_k)$ bandits in states $1, \ldots, k$. Let

$m = \rho n$.

$z_i = n_i/n$ be proportion in state i.

Asymptotic Optimality

Suppose a priority policy orders the states $1, 2, \dots$.
At time t there are $(n_1, \dots, n_k)$ bandits in states $1, \dots, k$. Let

$m = \rho n$.

$z_i = n_i/n$ be proportion in state i.

$n_i^a =$ number that receive action a.

Asymptotic Optimality

Suppose a priority policy orders the states $1, 2, \dots$.
At time t there are $(n_1, \dots, n_k)$ bandits in states $1, \dots, k$. Let

$m = \rho n$.

$z_i = n_i/n$ be proportion in state i.

$n_i^a =$ number that receive action a.

$u_i^a(z) = n_i^a/n_i$.

Asymptotic Optimality

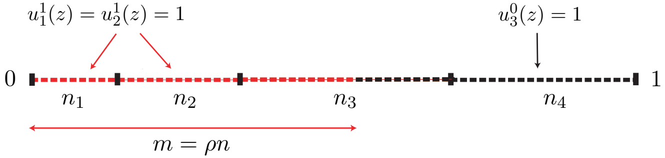
Suppose a priority policy orders the states $1, 2, \ldots$.
At time t there are $(n_1, \ldots, n_k)$ bandits in states $1, \ldots, k$. Let

$m = \rho n$.

$z_i = n_i/n$ be proportion in state i.

$n_i^a =$ number that receive action a.

$u_i^a(z) = n_i^a/n_i$.



Asymptotic Optimality

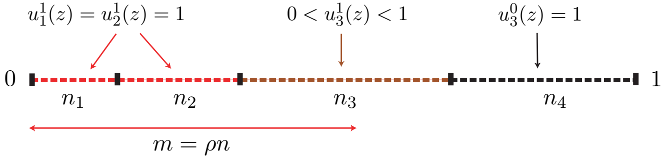
Suppose a priority policy orders the states $1, 2, \ldots$.
At time t there are $(n_1, \ldots, n_k)$ bandits in states $1, \ldots, k$. Let

$m = \rho n$.

$z_i = n_i/n$ be proportion in state i.

$n_i^a =$ number that receive action a.

$u_i^a(z) = n_i^a/n_i$.



Asymptotic Optimality

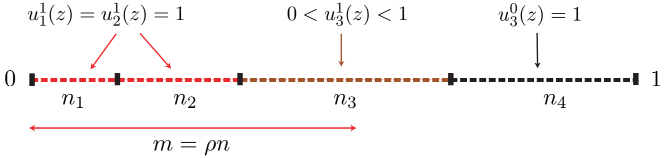
Suppose a priority policy orders the states $1, 2, \dots$.
At time t there are $(n_1, \dots, n_k)$ bandits in states $1, \dots, k$. Let

$m = \rho n$.

$z_i = n_i/n$ be proportion in state i.

$n_i^a =$ number that receive action a.

$u_i^a(z) = n_i^a/n_i$.



$q_{ij}^a =$ rate a bandit in state i jumps to state j under action a;

$$q_{ij}(z) = u_i^0(z) q_{ij}^0 + u_i^1(z) q_{ij}^1$$

Fluid Approximation

The 'fluid approximation' for large n is given by piecewise linear differential equations, of the form:

$$dz_i/dt = \sum_j q_{ji}(z) z_j - \sum_j q_{ij}(z) z_i$$

E.g., $k = 2$.

$$dz_1/dt = \begin{cases} -(q^0_{12} + q^0_{21})z_1 + (q^0_{12} - q^1_{12})\rho + q^0_{21}\,, & z_1 \geq \rho \\ -(q^1_{12} + q^1_{21})z_1 - (q^0_{21} - q^1_{21})\rho + q^0_{21}\,, & z_1 \leq \rho \end{cases}$$

Fluid Approximation

The 'fluid approximation' for large n is given by piecewise linear differential equations, of the form:

$$dz_i/dt = \sum_j q_{ji}(z)z_j - \sum_j q_{ij}(z)z_i$$

E.g., $k = 2$.

$$dz_1/dt = \begin{cases} -(q^0_{12} + q^0_{21})z_1 + (q^0_{12} - q^1_{12})\rho + q^0_{21}\,, & z_1 \geq \rho \\ -(q^1_{12} + q^1_{21})z_1 - (q^0_{21} - q^1_{21})\rho + q^0_{21}\,, & z_1 \leq \rho \end{cases}$$

$dz/dt = A(z)z + b(z)$, where $A(z)$ and $b(z)$ are constant within k polyhedral regions.

Asymptotic Optimality

Theorem [W. and Weiss '90]

If bandits are indexable, and the fluid model has an asymptotically stable equilibrium point, then the Whittle index heuristic is asymptotically optimal, — in the sense that the reward per bandit tends to the reward that is obtained under the relaxed policy.

(proof via a theorem about law of large numbers for sample paths.)

Heuristic May Not be Asymptotically Optimal

$$\left(q_{ij}^0\right) = \begin{pmatrix} -2 & 1 & 0 & 1 \\ 2 & -2 & 0 & 0 \\ 0 & 56 & -\frac{113}{2} & \frac{1}{2} \\ 1 & 1 & \frac{1}{2} & -\frac{5}{2} \end{pmatrix}, \quad \left(q_{ij}^1\right) = \begin{pmatrix} -2 & 1 & 0 & 1 \\ 2 & -2 & 0 & 0 \\ 0 & \frac{7}{25} & -\frac{113}{400} & \frac{1}{400} \\ 1 & 1 & \frac{1}{2} & -\frac{5}{2} \end{pmatrix}$$

$$r^0 = (0, 1, 10, 10)\,, \quad r^1 = (10, 10, 10, 0)\,, \quad \rho = 0.835$$

Heuristic May Not be Asymptotically Optimal

$$\left(q^0_{ij}\right) = \begin{pmatrix} -2 & 1 & 0 & 1 \\ 2 & -2 & 0 & 0 \\ 0 & 56 & -\frac{113}{2} & \frac{1}{2} \\ 1 & 1 & \frac{1}{2} & -\frac{5}{2} \end{pmatrix}, \quad \left(q^1_{ij}\right) = \begin{pmatrix} -2 & 1 & 0 & 1 \\ 2 & -2 & 0 & 0 \\ 0 & \frac{7}{25} & -\frac{113}{400} & \frac{1}{400} \\ 1 & 1 & \frac{1}{2} & -\frac{5}{2} \end{pmatrix}$$

$$r^0 = (0, 1, 10, 10)\,, \quad r^1 = (10, 10, 10, 0)\,, \quad \rho = 0.835$$

Bandit is indexable.

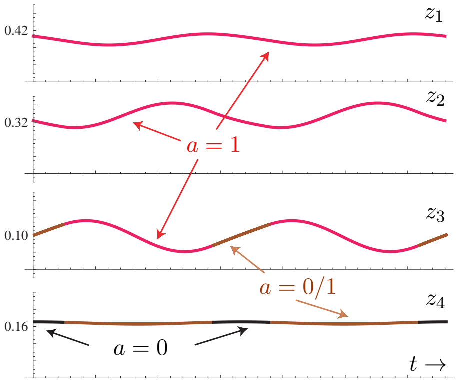
Equilibrium point is $(\bar{z}_1, \bar{z}_2, \bar{z}_3, \bar{z}_4) = (0.409, 0.327, 0.100, 0.164)$.

$\bar{z}_1 + \bar{z}_2 + \bar{z}_3 = 0.836$.

Relaxed policy obtains 10 per bandit per unit time.

Heuristic is Not Asymptotically Optimal

But equilibrium point $\bar{z}$ is not asymptotically stable.



Relaxed policy obtains 10 per bandit.
Heuristic obtains only 9.9993 per bandit.

Questions