Multi-armed Bandits and the Gittins Index Theorem

Richard Weber

Statistical Laboratory, University of Cambridge

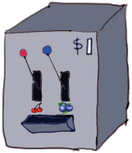
A talk to accompany Lecture 7

Two-armed Bandit

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

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

Two-armed Bandit

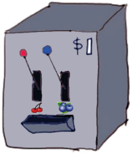


3, 10, 4, 9, 12, 1, ...

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

$\longrightarrow$ 5

Two-armed Bandit

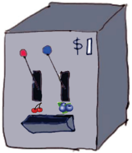


3, 10, 4, 9, 12, 1, ...

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

⟶ 5, 6

Two-armed Bandit



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

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

→ 5, 6, 3

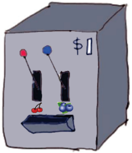
Two-armed Bandit

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

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

⟶ 5, 6, 3, 10,

Two-armed Bandit

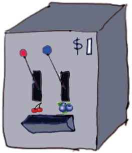


, , , 9, 12, 1, ...

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

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

Two-armed Bandit

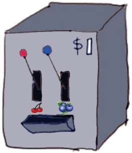


, , , , 12, 1, ...

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

⟶ 5, 6, 3, 10, 4, 9

Two-armed Bandit

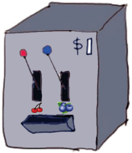


, , , , , 1, ...

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

⟶ 5, 6, 3, 10, 4, 9, 12

Two-armed Bandit

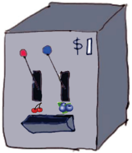


, , , , , 1, ...

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

→ 5, 6, 3, 10, 4, 9, 12, 2

Two-armed Bandit

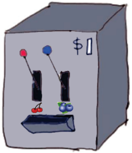


, , , , , 1, ...

, , , , 2, 7, ...

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

Two-armed Bandit



, , , , , 1, ...

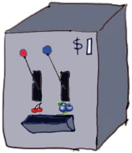
, , , , 2, 7, ...

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

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

$0 < \beta < 1.$

Two-armed Bandit



, , , , , 1, ...

, , , , 2, 7, ...

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

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

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

Bandit Processes

A **bandit process** is a special type of Markov Decision Process in which there are just two possible actions:

- $u = 1$ (**continue**)
  produces reward $r(x_t)$ and the state changes, to x_{t+1}, according to Markov dynamics $P_i(x_t, x_{t+1})$.

- $u = 0$ (**freeze**)
  produces no reward and the state does not change (hence the term 'freeze').

Bandit Processes

A **bandit process** is a special type of Markov Decision Process in which there are just two possible actions:

- $u = 1$ (**continue**)
  produces reward $r(x_t)$ and the state changes, to x_{t+1}, according to Markov dynamics $P_i(x_t, x_{t+1})$.

- $u = 0$ (**freeze**)
  produces no reward and the state does not change (hence the term 'freeze').

A **simple family of alternative bandit processes** (SFABP)

- is a collection of n such bandit processes.

- states are $x_1(t), \ldots, x_n(t)$.

SFABP

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

- One bandit process 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 $r_i(x_i(t))$.

SFABP

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

- One bandit process 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 $r_i(x_i(t))$.
- All other bandit processes remain passive (not pulled/**frozen**).

SFABP

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

- One bandit process 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 $r_i(x_i(t))$.
- All other bandit processes remain 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, $(0 < \beta < 1)$.

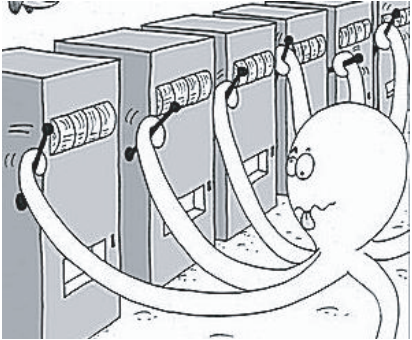
Dynamic Programming Solution

The dynamic programming equation is

$$F(x_1, \ldots, x_n)$$

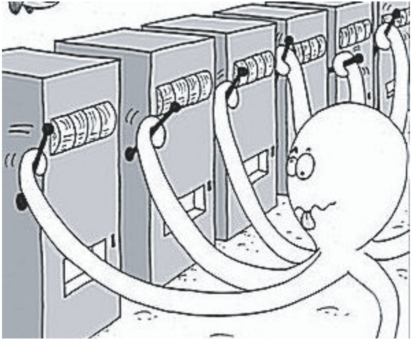
$$= \max_i \Big\{ r_i(x_i) + \beta \sum_y P_i(x_i, y) F(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n) \Big\}$$

Dynamic Effort Allocation



- **Job Scheduling**: in what order should I work on the tasks in my in-tray?
- **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



- **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?

Single Machine Scheduling

- n jobs are to be processed successively on one machine.

Single Machine Scheduling

- n jobs are to be processed successively on one machine.
- Job i has a known processing times t_i, a positive integer.

Single Machine Scheduling

- n jobs are to be processed successively on one machine.
- Job i has a known processing times t_i, a positive integer.
- On completion of job i a reward r_i is obtained.

Single Machine Scheduling

- n jobs are to be processed successively on one machine.
- Job i has a known processing times t_i, a positive integer.
- On completion of job i a reward r_i is obtained.
- If job 1 is processed immediately before job 2 the sum of discounted rewards from the two jobs is $r_1\beta^{t_1} + r_2\beta^{t_1+t_2}$.

$$r_1\beta^{t_1} + r_2\beta^{t_1+t_2} > r_2\beta^{t_2} + r_1\beta^{t_2+t_1}$$

$$\iff G_1 = (1-\beta)\frac{r_1\beta^{t_1}}{1-\beta^{t_1}} > (1-\beta)\frac{r_2\beta^{t_2}}{1-\beta^{t_2}} = G_2.$$

Single Machine Scheduling

- n jobs are to be processed successively on one machine.
- Job i has a known processing times t_i, a positive integer.
- On completion of job i a reward r_i is obtained.
- If job 1 is processed immediately before job 2 the sum of discounted rewards from the two jobs is $r_1\beta^{t_1} + r_2\beta^{t_1+t_2}$.

$$r_1\beta^{t_1} + r_2\beta^{t_1+t_2} > r_2\beta^{t_2} + r_1\beta^{t_2+t_1}$$

$$\iff G_1 = (1-\beta)\frac{r_1\beta^{t_1}}{1-\beta^{t_1}} > (1-\beta)\frac{r_2\beta^{t_2}}{1-\beta^{t_2}} = G_2.$$

- So total discounted reward is maximized by the **index policy** which processes jobs in decreasing order of **indices**, G_i.

Gittins Index Theorem

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

The expected discounted reward obtained from a simple family of alternative bandit processes is maximized by always continuing the bandit having greatest Gittins index

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

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

Gittins Index Theorem

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

The expected discounted reward obtained from a simple family of alternative bandit processes is maximized by always continuing the bandit having greatest Gittins index

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

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

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

Gittins Index Theorem

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

The expected discounted reward obtained from a simple family of alternative bandit processes is maximized by always continuing the bandit having greatest Gittins index

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

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

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

Gittins and Jones (1974). A dynamic allocation index for the sequential design of experiments. In Gani, J., editor, Progress in Statistics, pages 241–66. North-Holland, Amsterdam, NL. Read at the 1972 European Meeting of Statisticians, Budapest.

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

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

Discounted reward up to τ.

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.

Discounted time up to τ.

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.

Discounted time up to τ.

Note the role of the **stopping time** τ.
Stopping times are times recognisable when they occur.

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.

Discounted time up to τ.

Note the role of the **stopping time** τ.
Stopping times are times recognisable when they occur.
How do you make perfect toast?

There is a rule for timing toast,
One never has to guess,
Just wait until it starts to smoke,
then 7 seconds less. (David Kendall)

Calibration

Alternatively,

$$G_i(x_i) = \sup\Bigg\{\lambda :$$

$$\sum_{t=0}^{\infty} \beta^t \lambda \le \sup_{\tau \ge 1} E\left[\sum_{t=0}^{\tau-1} \beta^t r_i(x_i(t)) + \sum_{t=\tau}^{\infty} \beta^t \lambda \,\Bigg|\, x_i(0) = x_i\right]\Bigg\}.$$

Interpretation is a problem with two bandit processes:

- bandit process B_i and
- a **calibrating bandit process**, say Λ, paying known reward λ at each step it is continued.

Gittins index of B_i is the value of λ for which we are indifferent as to which of B_i and Λ to continue initially.

Calibration

Alternatively,

$$G_i(x_i) = \sup\Bigg\{\lambda :$$

$$\sum_{t=0}^{\infty} \beta^t \lambda \le \sup_{\tau \ge 1} E\left[\sum_{t=0}^{\tau-1} \beta^t r_i(x_i(t)) + \sum_{t=\tau}^{\infty} \beta^t \lambda \,\middle|\, x_i(0) = x_i\right]\Bigg\}.$$

Interpretation is a problem with two bandit processes:

- bandit process B_i and
- a **calibrating bandit process**, say Λ, paying known reward λ at each step it is continued.

Gittins index of B_i is the value of λ for which we are indifferent as to which of B_i and Λ to continue initially.

Notice that once we decide, at time τ, to switch from continuing B_i to continuing Λ then information about B_i does not change and so it must be optimal to stick with continuing Λ ever after.

Fair Charge

$$G_i(x_i) = \sup\Bigg\{ \lambda : \sum_{t=0}^{\infty} \beta^t \lambda \le \sup_{\tau \ge 1} E\left[\sum_{t=0}^{\tau-1} \beta^t r_i(x_i(t)) + \sum_{t=\tau}^{\infty} \beta^t \lambda \,\Bigg|\, x_i(0) = x_i \right] \Bigg\}$$

Alternatively,

$$G_i(x_i) \equiv \sup\left\{ \lambda : 0 \le \sup_{\tau \ge 1} E\left[\sum_{t=0}^{\tau-1} \beta^t \Big(r_i(x_i(t)) - \lambda \Big) \,\Bigg|\, x_i(0) = x_i \right] \right\}.$$

Example: Single Machine Scheduling

Problem in which n jobs are to be scheduled on one machine.

Job i has a known processing times t_i, a positive integer.

On completion of job i a positive reward r_i is obtained.

Interchange argument showed discounted sum of rewards maximized by processing jobs in decreasing order of index $r_i \beta^{t_1} / (1 - \beta^{t_1})$.

Example: Single Machine Scheduling

Problem in which n jobs are to be scheduled on one machine.

Job i has a known processing times t_i, a positive integer.

On completion of job i a positive reward r_i is obtained.

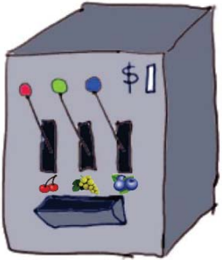
Interchange argument showed discounted sum of rewards maximized by processing jobs in decreasing order of index $r_i\beta^{t_1}/(1-\beta^{t_1})$.

Now we do this using Gittins index.

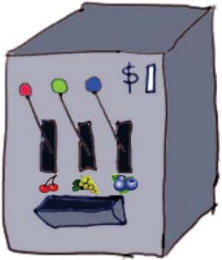
$$G_i = \sup_{\tau \geq 1} \frac{E\Big[\sum_{t=0}^{\tau-1} r_i(x_i(t))\beta^t \;\Big|\; x_i(0)=x_i\Big]}{E\Big[\sum_{t=0}^{\tau-1} \beta^t \;\Big|\; x_i(0)=x_i\Big]} = \frac{r_i\beta^{t_i}}{1+\beta+\cdots+\beta^{t_i-1}}$$

Optimal stopping time is $\tau = t_i$ and $G_i = \dfrac{r_i\beta^{t_i}(1-\beta)}{(1-\beta^{t_i})}$.

A Short History of Gittins Index Theorem

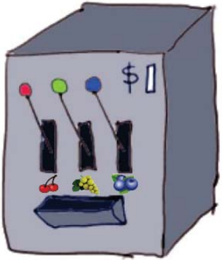


A Short History of Gittins Index Theorem



Many applications to clinical trials, job scheduling, search, etc.

A Short History of Gittins Index Theorem

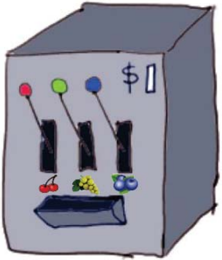


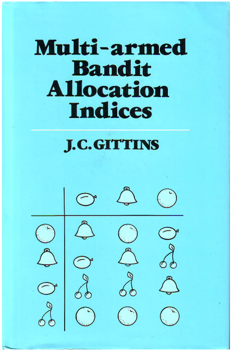
Exploration vs Exploitation

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

Many applications to clinical trials, job scheduling, search, etc.

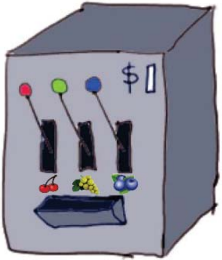
A Short History of Gittins Index Theorem

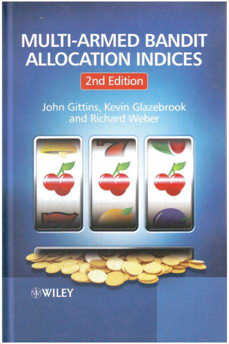




Many applications to clinical trials, job scheduling, search, etc.

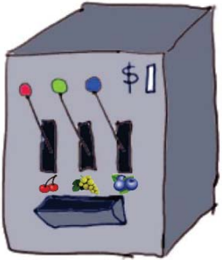
A Short History of Gittins Index Theorem





Many applications to clinical trials, job scheduling, search, etc.

A Short History of Gittins Index Theorem



Many applications to clinical trials, job scheduling, search, etc.

Clinical Trials

Clinical Trials

On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples

William R. Thompson

Biometrika

Vol. 25, No. 3/4 (Dec., 1933), pp. 285-294

Clinical Trials

On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples

William R. Thompson
Biometrika
Vol. 25, No. 3/4 (Dec., 1933), pp. 285-294

Robbins, H. (1952). "Some aspects of the sequential design of experiments".

Bernoulli Bandits

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

Bernoulli Bandits

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

- The sth time drug i is administered 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 $t = 0, 1, \ldots$

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

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

Bernoulli Bandits

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

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

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

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

Bernoulli Bandits

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

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

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

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

- θ_i is unknown, but has a *prior* distribution,

Bernoulli Bandits

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

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

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

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

- θ_i is unknown, but has a *prior* distribution, say 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.

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.

Gittins Index Theorem is Surprising

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 is Surprising

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)

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)

Proof of the Index Theorem

Start with a problem in which only bandit process B_i is available.

Proof of the Index Theorem

Start with a problem in which only bandit process B_i is available.

Define the **fair charge**, $\gamma_i(x_i)$, as the maximum amount that a gambler would be willing to pay **per step** to be permitted to continue B_i for at least one more step, and with option to stop continuing it whenever he likes thereafter.

Proof of the Index Theorem

Start with a problem in which only bandit process B_i is available.

Define the **fair charge**, $\gamma_i(x_i)$, as the maximum amount that a gambler would be willing to pay **per step** to be permitted to continue B_i for at least one more step, and with option to stop continuing it whenever he likes thereafter.

$$\gamma_i(x_i) = \sup\left\{\lambda : 0 \le \sup_{\tau \ge 1} E\left[\sum_{t=0}^{\tau-1} \beta^t \Big(r_i(x_i(t)) - \lambda\Big) \,\Big|\, x_i(0) = x_i\right]\right\}$$

$\gamma_i(x_i) = G_i(x_i)$, as defined previously.

Proof of the Index Theorem

Start with a problem in which only bandit process B_i is available.

Define the **fair charge**, $\gamma_i(x_i)$, as the maximum amount that a gambler would be willing to pay **per step** to be permitted to continue B_i for at least one more step, and with option to stop continuing it whenever he likes thereafter.

$$\gamma_i(x_i) = \sup \left\{ \lambda : 0 \leq \sup_{\tau \geq 1} E \left[\sum_{t=0}^{\tau-1} \beta^t \Big(r_i(x_i(t)) - \lambda \Big) \,\Big|\, x_i(0) = x_i \right] \right\}$$

$\gamma_i(x_i) = G_i(x_i)$, as defined previously.

The **stopping time** τ is the first time that $G_i(x_i(\tau)) < G_i(x_i(0))$, i.e. the first time that the charge is looking too expensive.

Gambler would rather stop than continue while paying this charge.

Prevailing Charges

When $G_i(x_i(\tau)) < G_i(x_i(0))$ the gambler will stop playing.

But suppose at this point the charge is reduced to $G_i(x_i(\tau))$; then it remains just-profitable for the gambler to keep on playing.

This defines a **prevailing charge**, say $g_i(t) = \min_{s \leq t} G_i(x_i(s))$.

Prevailing Charges

When $G_i(x_i(\tau)) < G_i(x_i(0))$ the gambler will stop playing.

But suppose at this point the charge is reduced to $G_i(x_i(\tau))$; then it remains just-profitable for the gambler to keep on playing.

This defines a **prevailing charge**, say $g_i(t) = \min_{s \le t} G_i(x_i(s))$.

$g_i(t)$ is a nonincreasing function of t and its value depends only on the states through which bandit i evolves.

Prevailing Charges

When $G_i(x_i(\tau)) < G_i(x_i(0))$ the gambler will stop playing.

But suppose at this point the charge is reduced to $G_i(x_i(\tau))$; then it remains just-profitable for the gambler to keep on playing.

This defines a **prevailing charge**, say $g_i(t) = \min_{s \leq t} G_i(x_i(s))$.

$g_i(t)$ is a nonincreasing function of t and its value depends only on the states through which bandit i evolves.

Observation 1. Suppose that in the problem with n alternative bandit processes, $B_1, \ldots, B_n$, the gambler not only collects $r_{i_t}(x_{i_t}(t))$, but must also pays the prevailing charge $g_{i_t}(x_{i_t}(t))$ of the bandit B_{i_t} that he chooses to continue at time t. Then he cannot do better than just break even (i.e. expected profit 0).
— *This is because he could only make a strictly positive profit (in expected value) if this were to happen for at least one bandit. Yet the prevailing charge has been defined so that if he pays the prevailing charges he can only just break even.*

Observation 2. He maximizes the expected discounted sum of the prevailing charges that he pays by always continuing the bandit with the greatest prevailing charge.
— *This is because he thereby interleaves the* n *nonincreasing sequences of prevailing charges* g_i *into one nonincreasing sequence of prevailing charges. This way of interleaving them maximizes their discounted sum.*

Observation 2. He maximizes the expected discounted sum of the prevailing charges that he pays by always continuing the bandit with the greatest prevailing charge.
— *This is because he thereby interleaves the n nonincreasing sequences of prevailing charges g_i into one nonincreasing sequence of prevailing charges. This way of interleaving them maximizes their discounted sum.*

For example, prevailing charges of

$$g_1 : 10, 10, 9, 5, 5, 3, \ldots$$
$$g_2 : 20, 15, 7, 4, 2, 2, \ldots$$

are best interleaved (so as to maximize discounted charge paid) as

$$20, 15, 10, 10, 9, 7, 5, 5, 4, 3, 2, 2, \ldots$$

$$\text{sum of discounted charges paid} = 20 + 15\beta + 10\beta^2 + 10\beta^3 + \cdots$$

Observation 3. Consider the Gittins index policy π^* of always continuing the bandit with the greatest $G_i(x_i)$ (which is also the one having greatest $g_i(x_i)$).
Using π^* he just breaks even (because by continuing B_i until its prevailing charge decreases is the way to break even).

Observation 3. Consider the Gittins index policy π^* of always continuing the bandit with the greatest $G_i(x_i)$ (which is also the one having greatest $g_i(x_i)$).

Using π^* he just breaks even (because by continuing B_i until its prevailing charge decreases is the way to break even).

Observation 1 is that for **any** policy π,

$$E_\pi \left[\sum_{t=0}^{\infty} \beta^t \Big(r_{i_t}(x_{i_t}(t)) - g_{i_t}(x_{i_{t(t)}}) \Big) \,\Big|\, x(0) \right] \leq 0$$

Observation 3. Consider the Gittins index policy π^* of always continuing the bandit with the greatest $G_i(x_i)$ (which is also the one having greatest $g_i(x_i)$).
Using π^* he just breaks even (because by continuing B_i until its prevailing charge decreases is the way to break even).

Observation 1 is that for **any** policy π,

$$E_\pi \left[\sum_{t=0}^{\infty} \beta^t \Big(r_{i_t}(x_{i_t}(t)) - g_{i_t}(x_{i_t(t)}) \Big) \,\Big|\, x(0) \right] \leq 0$$

$$\implies E_\pi \left[\sum_{t=0}^{\infty} \beta^t r_{i_t}(x_{i_t}) \,\Big|\, x(0) \right] \leq E_\pi \left[\sum_{t=0}^{\infty} \beta^t g_{i_t}(x_{i_t}) \,\Big|\, x(0) \right].$$

Observation 3. Consider the Gittins index policy π^* of always continuing the bandit with the greatest $G_i(x_i)$ (which is also the one having greatest $g_i(x_i)$).
Using π^* he just breaks even (because by continuing B_i until its prevailing charge decreases is the way to break even).

Observation 1 is that for **any** policy π,

$$E_\pi\left[\sum_{t=0}^{\infty}\beta^t\Big(r_{i_t}(x_{i_t}(t)) - g_{i_t}(x_{i_t(t)})\Big)\,\Big|\,x(0)\right] \leq 0$$

$$\implies E_\pi\left[\sum_{t=0}^{\infty}\beta^t r_{i_t}(x_{i_t})\,\Big|\,x(0)\right] \leq E_\pi\left[\sum_{t=0}^{\infty}\beta^t g_{i_t}(x_{i_t})\,\Big|\,x(0)\right].$$

Observation 2 is that the right hand side is maximized by π^*.

Observation 3. Consider the Gittins index policy π^* of always continuing the bandit with the greatest $G_i(x_i)$ (which is also the one having greatest $g_i(x_i)$).
Using π^* he just breaks even (because by continuing B_i until its prevailing charge decreases is the way to break even).

Observation 1 is that for **any** policy π,

$$E_\pi\left[\sum_{t=0}^{\infty}\beta^t\Big(r_{i_t}(x_{i_t}(t)) - g_{i_t}(x_{i_t(t)})\Big)\,\Big|\, x(0)\right] \leq 0$$

$$\implies E_\pi\left[\sum_{t=0}^{\infty}\beta^t r_{i_t}(x_{i_t})\,\Big|\, x(0)\right] \leq E_\pi\left[\sum_{t=0}^{\infty}\beta^t g_{i_t}(x_{i_t})\,\Big|\, x(0)\right].$$

Observation 2 is that the right hand side is maximized by π^*.

Observation 3 is that under π^* the inequality is an equality.

So the left hand side is maximized by π^*. □

Pandora's Boxes Problem

Econometrica, Vol. 47, No. 3 (May, 1979)

OPTIMAL SEARCH FOR THE BEST ALTERNATIVE

BY MARTIN L. WEITZMAN[1]

This paper completely characterizes the solution to the problem of searching for the best outcome from alternative sources with different properties. The optimal strategy is an elementary reservation price rule, where the reservation prices are easy to calculate and have an intuitive economic interpretation.

Pandora's Boxes Problem

Pandora's Boxes Problem

- Pandora has n boxes.

Pandora's Boxes Problem

- Pandora has n boxes.
- Box i contains a prize, of unknown value x_i, distributed with known c.d.f. F_i.

Pandora's Boxes Problem

- Pandora has n boxes.
- Box i contains a prize, of unknown value x_i, distributed with known c.d.f. F_i.
- At known cost c_i she can open box i and discover x_i.

Pandora's Boxes Problem

- Pandora has n boxes.
- Box i contains a prize, of unknown value x_i, distributed with known c.d.f. F_i.
- At known cost c_i she can open box i and discover x_i.
- Pandora may open boxes in any order, and stop at will.

  She then takes the greatest prize she has found.

Pandora's Boxes Problem

- Pandora has n boxes.
- Box i contains a prize, of unknown value x_i, distributed with known c.d.f. F_i.
- At known cost c_i she can open box i and discover x_i.
- Pandora may open boxes in any order, and stop at will.

  She then takes the greatest prize she has found.
- She opens a subset of boxes $S \subseteq \{1, \ldots, n\}$ and then stops, seeking to maximize the expected value of

$$R = -\sum_{i \in S} c_i + \max_{i \in S} x_i.$$

Pandora's Problem Recast as a Bandit Problem

- Box i is associated with bandit B_i, which starts in state 0.

Pandora's Problem Recast as a Bandit Problem

- Box i is associated with bandit B_i, which starts in state 0.

- First time B_i is continued reward is $-c_i$, and the state becomes x_i, chosen by the distribution F_i.

Pandora's Problem Recast as a Bandit Problem

- Box i is associated with bandit B_i, which starts in state 0.

- First time B_i is continued reward is $-c_i$, and the state becomes x_i, chosen by the distribution F_i.

- At all subsequent times B_i is continued the reward is $r(x_i) = (1 - \beta)x_i$, and the state remains x_i.

Pandora's Problem Recast as a Bandit Problem

- Box i is associated with bandit B_i, which starts in state 0.

- First time B_i is continued reward is $-c_i$, and the state becomes x_i, chosen by the distribution F_i.

- At all subsequent times B_i is continued the reward is $r(x_i) = (1-\beta)x_i$, and the state remains x_i.

Suppose we wish to maximize the expected value of

$$-\sum_{t=1}^{\tau}\beta^{t-1}c_{i_t} + \max\{r(x_{i_1}),\ldots,r(x_{i_\tau})\}\sum_{t=\tau}^{\infty}\beta^t$$

$$= -\sum_{t=1}^{\tau}\beta^{t-1}c_{i_t} + \beta^\tau \max\{x_{i_1},\ldots,x_{i_\tau}\}.$$

Pandora's Problem Recast as a Bandit Problem

Suppose we wish to maximize the expected value of

$$-\sum_{t=1}^{\tau}\beta^{t-1}c_{i_t} + \max\{r(x_{i_1}),\ldots,r(x_{i_\tau})\}\sum_{t=\tau}^{\infty}\beta^t$$
$$= -\sum_{t=1}^{\tau}\beta^{t-1}c_{i_t} + \beta^{\tau}\max\{x_{i_1},\ldots,x_{i_\tau}\}.$$

Gittins index of an opened box is $r(x_i)/(1-\beta) = x_i$.

Pandora's Problem Recast as a Bandit Problem

Suppose we wish to maximize the expected value of

$$
\begin{aligned}
&-\sum_{t=1}^{\tau}\beta^{t-1}c_{i_t} + \max\{r(x_{i_1}),\ldots,r(x_{i_\tau})\}\sum_{t=\tau}^{\infty}\beta^t \\
&\qquad = -\sum_{t=1}^{\tau}\beta^{t-1}c_{i_t} + \beta^\tau \max\{x_{i_1},\ldots,x_{i_\tau}\}.
\end{aligned}
$$

Gittins index of an opened box is $r(x_i)/(1-\beta) = x_i$.

Gittins index of an unopened box i is the solution to

$$
\frac{G_i}{1-\beta} = -c_i + \frac{\beta}{1-\beta} E\max\{r(x_i), G_i\}.
$$

Pandora's optimal strategy is thus:

Open boxes in decreasing order of G_i until first reaching a point that a revealed prize is greater than all G_i of unopened boxes.