Bandit problems, learning vs. earning goals and the role of the problem's horizon

Sofía S. Villar †

† MRC Biostatistics Unit, Cambridge

Multi-armed Bandit Workshop
STOR-i, Lancaster University, January 11-12, 2016

Outline

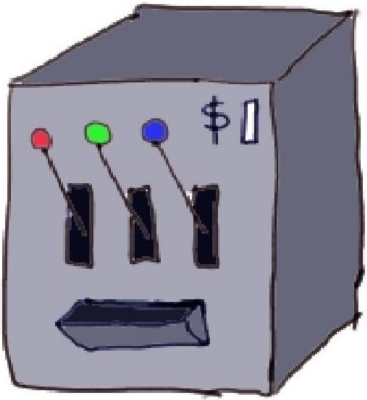
Outline

What is a multi-armed bandit problem?

Motivation

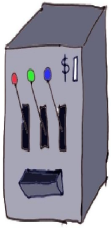
What is a multi-armed bandit problem?

Motivation



What is a multi-armed bandit problem?

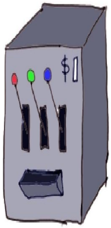
Motivation



A gambler's dilemma:

- correctly identifying the most rewarding arm (learning) requires playing the worse arms a large number of times
- making the highest expected profit of the game (earning) requires making a (possibly wrong) choice of a best arm to play thereafter

How does the gambler optimally balance these two goals?



A gambler's dilemma:

- correctly identifying the most rewarding arm (learning) requires playing the worse arms a large number of times
- making the highest expected profit of the game (earning) requires making a (possibly wrong) choice of a best arm to play thereafter

How does the gambler optimally balance these two goals?

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

What is the multi-armed bandit problem?

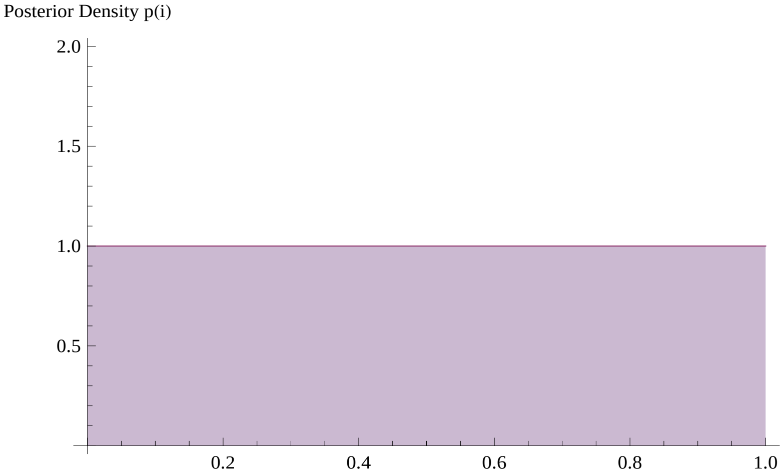
Problem definition: Bayesian Bernoulli K-armed bandit problem

- K independent arms (each is a draw from a Bernoulli population $Y_{k,t}$ with unknown parameter p_k)
- A each discrete time point t, only one arm can be played (earning the observed value $y_{k,t}$)
- A Bayesian approach to learning about p_k's defines the state space and dynamics over it.
- Each p_k has a Beta prior before arm k has been played ($Be(s_{k,0}, f_{k,0})$) that is sequentially converted into Beta posteriors as observations of that arm are collected ($Be(s_{k,0} + S_{k,t}, f_{k,0} + F_{k,t})$).
- $(S_{k,t}, F_{k,t})$ represents the random number of successes and failures observed for arm k after having pulled t arms.

The Bayesian approach to the bandit problem in pictures

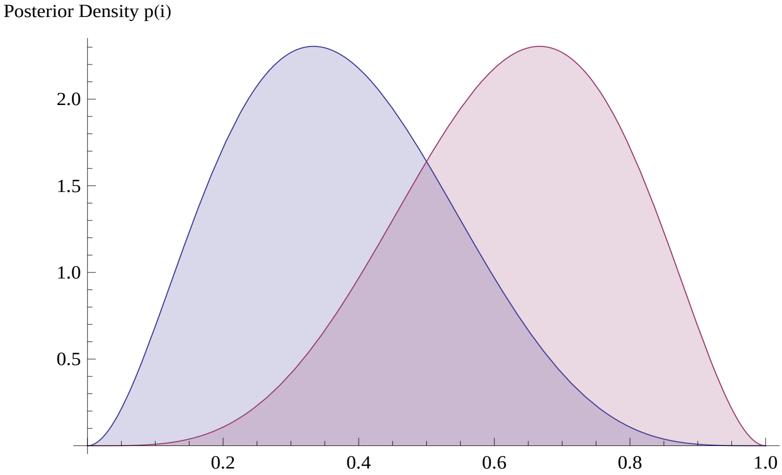
The Bayesian approach to the bandit problem in pictures

$$n_A = 0 s_A = 0 f_A = 0, n_B = 0 s_B = 0 f_B = 0 \longrightarrow \text{Uniform Priors on } p_i$$



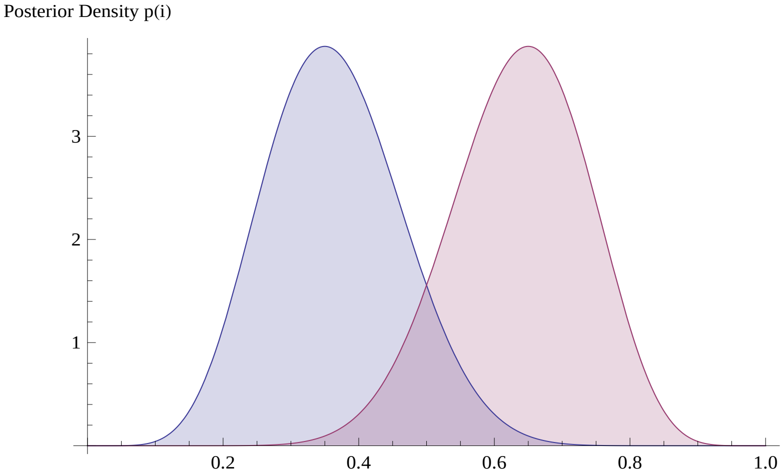
The Bayesian approach to the bandit problem in pictures

$$n_A = 6 s_A = 4 f_A = 2, n_B = 6 s_B = 2 f_B = 4 \longrightarrow \text{Beta Posteriors on } p_i$$



The Bayesian approach to the bandit problem in pictures

$n_A = 20 s_A = 13 f_A = 7, n_B = 20 s_B = 7 f_B = 13 \longrightarrow$ Beta Posteriors on p_i



- The state space:
  $\mathbb{X}_{k,t} = \{(s_{k,0}+S_{k,t}, f_{k,0}+F_{k,t}) \in \mathbb{N}_+^2 : S_{k,t}+F_{k,t} \leq t, \text{ for } t = 0, 1, \ldots, T\}$
- Denote the available information on arm k at time t as
  $\mathbf{x}_{k,t} = (s_{k,0} + s_{k,t}, f_{k,0} + s_{k,t})$
- The state (Markovian) dynamics:

$$\mathbf{x}_{k,t+1} = \begin{cases} (s_{k,0} + s_{k,t} + 1, f_{k,0} + f_{k,t}), & \text{if } a_{k,t} = 1 \text{ w.p } \frac{s_{k,0}+s_{k,t}}{s_{k,0}+f_{k,0}+s_{k,t}+f_{k,t}}, \\ (s_{k,0} + s_{k,t}, f_{k,0} + f_{k,t} + 1), & \text{if } a_{k,t} = 1 \text{ w.p } \frac{f_{k,0}+f_{k,t}}{s_{k,0}+f_{k,0}+s_{k,t}+f_{k,t}}, \\ \\ \mathbf{x}_{k,t}, & \text{if } a_{k,t} = 0 \text{ w.p } 1, \end{cases} \tag{1}$$

for any $\mathbf{x}_{k,t} \in \mathbb{X}_{k,t}$.

Outline

- To complete the specification of a multi-armed bandit model the problem's objective function must be selected.
- Given an objective function and a time horizon, a multi-armed bandit optimal control problem is the problem of finding a feasible playing policy/strategy/rule, π that optimizes the selected performance objective.
- The set of all the feasible policies/strategies/rules Π are those that fulfill the resource constraint (e.g. one arm at a time)

Solving the learning-earning dilemma

An earning oriented objective

- Let $a_{k,t}$ be a binary variable representing the selected action for arm k at time t. $a_{k,t} = 1$ represents that arm k is pulled at time t.

- The resource constrain is in this case $\sum_{k=1}^{K} a_{k,t} \leq 1$ for all t.

- Denote the expected reward function and playing horizon respectively as $\mathcal{R}(\mathbf{x}_{k,t}, a_{k,t}) = \frac{s_{k,0}+s_{k,t}}{s_{k,0}+f_{k,0}+s_{k,t}+f_{k,t}} a_{k,t}$ and T

- Then one optimisation criterion is to maximise the expected total discounted (ETD) number of successes after T observations, where $0 < d < 1$.

$$V_D^*(\tilde{\mathbf{x}}_\mathbf{0}) = \max_{\pi \in \Pi} \mathrm{E}^{\pi} \left[\sum_{t=0}^{T-1} \sum_{k=1}^{K} d^t R(\mathbf{x}_{k,t}, a_{k,t}) \middle| \tilde{\mathbf{x}}_\mathbf{0} = (\mathbf{x}_{k,0})_{k=1}^{K} \right] \tag{2}$$

Solving the learning-earning dilemma

A learning oriented objective

- Robbins (1952) proposed an alternative objective for the Bayesian Bernoulli bandit problem.
- He considered the average *regret* after playing T times (for a large T and for any given and unknown $(p_k)_{k=1}^K$).
- For the Bayesian Bernoulli K-armed bandit problem, the total regret ρ is defined as

$$\rho(T) = T \max_k \{p_k\} - \mathrm{E}^\pi \left[\sum_{t=0}^{T-1} \sum_{k=1}^{K} a_{k,t} Y_{k,t} \right] \text{ for some } (p_k)_{k=1}^K. \tag{3}$$

- A form of asymptotic optimality can be defined for sampling rules π in terms of (3) if it holds that for any $(p_k)_{k=1}^K$: $\lim_{T\to\infty} \frac{\rho(T)}{T} = 0$.

Outline

Outline

The classic multi-armed bandit problem

An earning oriented objective with an infinite horizon

- If we set $T = \infty$ and we consider the ETD objective then the resulting bandit is the *classic* bandit problem (OR).
- It attained this status because of the long standing challenge it posed.
- The problem can be solved via a dynamic programming (DP) approach but suffers from a severe computational burden.
- Before the alternative solution first obtained by Gittins and Jones (1974) the realistic scenarios of the problem (e.g. $K > 3$) were computationally unfeasible.

Classic Multi-armed Bandit: divide and conquer!

Infinite horizon case

Theorem ('74, '79, '89): The ETD reward is maximised by pulling at time t the arm having the greatest value of a dynamic allocation index:

$$G_k(\mathbf{x}_{k,t}) = \sup_{\tau \geq 1} \frac{\mathsf{E}_{\mathbf{x}_{k,t}} \sum_{s=0}^{\tau-1} R(\mathbf{x}_{k,t}, a_{k,t}) d^s}{\mathsf{E}_{\mathbf{x}_{k,t}} \sum_{s=0}^{\tau-1} d^s} \tag{4}$$

with τ is a (past-measurable) random stopping time.

Classic Multi-armed Bandit: divide and conquer!

Infinite horizon case

Theorem ('74, '79, '89): The ETD reward is maximised by pulling at time t the arm having the greatest value of a dynamic allocation index:

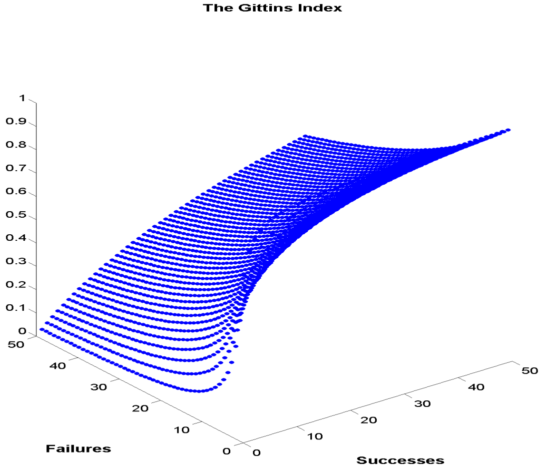
$$G_k(\mathbf{x}_{k,t}) = \sup_{\tau \geq 1} \frac{\mathsf{E}_{\mathbf{x}_{k,t}} \sum_{s=0}^{\tau-1} R(\mathbf{x}_{k,t}, a_{k,t}) d^s}{\mathsf{E}_{\mathbf{x}_{k,t}} \sum_{s=0}^{\tau-1} d^s} \tag{4}$$

with τ is a (past-measurable) random stopping time.

Huge computational gains! The index can be computed as the solution to a 1 armed bandit problem

The Gittins index

How does it look like? (d=0.99 T*=750)



The classic multi-armed bandit problem

Why is the ETD criteria not a learning oriented goal?

- The Gittins index (or equivalently the DP solution) chooses an arm at some $t < \infty$ and plays it thereafter.
- This "chosen arm" has a positive probability of being suboptimal (Rothschild, 1974).
- This is known as incomplete learning.
- A necessary condition to have complete learning is to have a strictly positive probability of playing every arm for every t.

The classic multi-armed bandit problem

Can we use the Gittins index to achieve a learning-oriented goal?

- Bather (1981) proposed to add random perturbations to an index rule based on the observed data at each stage.

- The (deterministic) part captures the importance of the *exploitation* or earning based on the accumulated information and the (random) perturbation part, captures the *exploration* or learning element.

- Glazebrook (1980)

$$I(\mathbf{x}_{k,t}) = G(\mathbf{x}_{k,t}) + Z_t * \lambda(s_{k,0} + s_{k,t} + f_{k,t} + f_{k,0}), \qquad (5)$$

  Z_t is an i.i.d. non-negative and unbounded random variable and $\lambda(s_{k,0} + s_{k,t} + f_{k,t} + f_{k,0}) \to 0$ as $s_{k,0} + s_{k,t} + f_{k,t} + f_{k,0} \to \infty$ and is a sequence of non-negative constants.

- Example: $Z_t(K) \sim exp(\frac{1}{K})$; $\lambda(s_{k,0} + s_{k,t} + f_{k,t} + f_{k,0}) = \frac{K}{s_{k,0}+s_{k,t}+f_{k,t}+f_{k,0}}$

Outline

The classic K-armed bandit problem with a finite horizon

Reformulated Problem's state space and dynamics

- The state space:
  $\tilde{\mathbb{X}}_{k,t} = \{(s_{k,0} + S_{k,t}, f_{k,0} + F_{k,t}, T - t) \in \mathbb{N}_+^2 : S_{k,t} + F_{k,t} \leq t, \text{ for } t = 0, 1, \ldots, T\}$ and absorbing state E for $t > T$

- The available information $\tilde{\mathbf{x}}_{k,t} = (s_{k,0} + s_{k,t}, f_{k,0} + s_{k,t}, T - t)$

- The State dynamics:

$$\tilde{\mathbf{x}}_{k,t+1} = \begin{cases} \text{if } a_{k,t} = 1 : & \\ (s_{k,0} + s_{k,t} + 1, f_{k,0} + f_{k,t}, T - (t+1)), & \text{w.p } \frac{s_{k,t} + s_{k,0}}{s_{k,t} + f_{k,t} + s_{k,0+} f_{k,0}}, \\ (s_{k,0} + s_{k,t}, f_{k,0} + f_{k,t} + 1, T - (t+1)), & \text{w.p } \frac{f_{k,t} + f_{k,0}}{s_{k,t} + f_{k,t} + s_{k,0} + f_{k,0}}, \\ & \\ \text{if } a_{k,t} = 0 \quad (\mathbf{x}_{k,t}, T - (t+1)), & \text{w.p } 1, \end{cases}$$

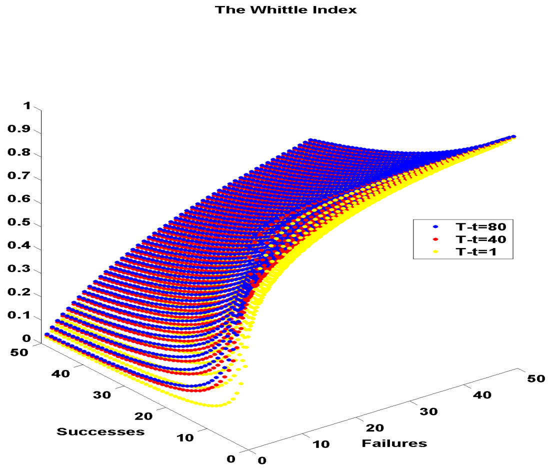
$\forall \tilde{\mathbf{x}}_{k,t}$ such that $0 \leq t \leq T - 1$. $\tilde{\mathbf{x}}_{k,T}$ and E, $\forall a$, lead to E w.p. 1

The classic K-armed bandit problem with a finite horizon

- In the previous reformulation $T = \infty$ yet the Gittins index theorem does not apply.
- Unplayed arms continue to evolve $\rightarrow$ bandits are Restless Whittle (1980)
- Again the problem can be solved via a dynamic programming approach at the expense of a severe computational burden.
- Whittle proposed an index that generalises Gittins index, but its existence is not guaranteed for every *restless* MABP. And, if it exists, is not necessarily optimal, being thus a heuristic rule.
- The reformulated restless problem is indexable and the Whittle index can be computed as a modified version of the Gittins index, in which the search of the optimal stopping time in (4) is truncated to be $\leq T - t$ (at each t)

The Whittle index

How does it look like? (d=1 T=180)



Outline

Comparison of rules in the context of clinical trials

Earning vs. learning dilemma solved differently

	Crit. Value	$H_0 : p_0 = p_k = 0.3$ for $k = 1, \ldots, 3$		
		α	p^* (s.e.)	ENS (s.e.)
Fixed Equal	*2.128*	0.047	0.250 (0.02)	126.86 (9.41)
Thompson Sampling	*2.128*	0.056	0.251 (0.07)	126.93 (9.47)
UCB	*2.128*	0.055	0.251 (0.06)	126.97 (9.41)
RBI	*2.128*	0.049	0.250 (0.03)	126.77 (9.40)
RGI	*2.128*	0.046	0.250 (0.03)	126.80 (9.36)
Current Belief	F_a	0.047	0.269 (0.39)	126.89 (9.61)
GI	F_a	0.048	0.248 (0.18)	126.68 (9.40)
CGI	*2.128*	0.034	0.250 (0.02)	127.16 (9.46)
Upper Bound				126.90 (0.00)

Table: Comparison of different four-arm trial designs of size $T = 423$. F_a: Fisher's adjusted test; α: family wise type I error; $1 - \beta$: power; p^*: expected proportion of patients in the trial assigned to the best treatment; ENS: expected number of patient successes.

Comparison of rules in the context of clinical trials

Earning vs. learning dilemma solved differently

	Crit. Value	$H_1 : p_0 = p_k = 0.3\ k = 1, 2:,\ p_3 = 0.5$ $(1-\beta)$	p^* (s.e.)	ENS (s.e.)
Fixed Equal	*2.128*	0.814	0.250 (0.02)	148.03 (9.77)
Thompson Sampling	*2.128*	0.884	0.529 (0.09)	172.15 (13.0)
UCB	*2.128*	0.877	0.526 (0.07)	171.70 (11.9)
RBI	*2.128*	0.846	0.368 (0.04)	158.34 (10.4)
RGI	*2.128*	0.847	0.358 (0.03)	157.26 (10.3)
Current Belief	F_a	0.213	0.677 (0.41)	184.87 (36.8)
GI	F_a	0.428	0.831 (0.10)	198.25 (13.7)
CGI	*2.128*	0.925	0.640 (0.08)	182.10 (12.3)
Upper Bound			1	211.50 (0.00)

Table: Comparison of different four-arm trial designs of size $T = 423$. F_a: Fisher's adjusted test; α: family wise type I error; $1-\beta$: power; p^*: expected proportion of patients in the trial assigned to the best treatment; ENS: expected number of patient successes.

Comparison of rules in the context of clinical trials

Earning vs. learning dilemma when patients are scarce

	Crit. Value	$H_1 : p_k = 0.3 + 0.1 \times k \;\; k = 0, 1, 2, 3$ $(1 - \beta)$	p^* (s.e.)	ENS (s.e.)
Fixd Equal	F	0.300	0.250 (0.04)	35.99 (4.41)
Thompson Sampling	F	0.246	0.338 (0.08)	38.34 (4.68)
UCB	F	0.218	0.362 (0.08)	38.84 (4.71)
RBI	F	0.295	0.268 (0.03)	36.52 (4.41)
RGI	F	0.298	0.265 (0.03)	36.45 (4.36)
Current Belief	F_a	0.056	0.419 (0.38)	40.92 (6.89)
WI	F_a	0.001	0.537 (0.31)	42.65 (6.02)
GI	F_a	0.002	0.492 (0.21)	41.60 (5.44)
CGI	F_a	0.349	0.393 (0.16)	38.29 (4.82)
Upper Bound			1	48.00 (0.00)

Table: Comparison of different four-arm trial designs of size $T = 80$. F: Fisher; α: type I error; $1 - \beta$: power; p^*: expected proportion of patients in the trial assigned to the best treatment; ENS: expected number of patient successes.

Outline

Solving the learn-earn dilemma, goals and horizon

Key takeaways

- The optimal learning-earning balance depends crucially on:
  (a) What are we concerned the most? Correct selection or expected total rewards over time
  (b) How many plays do we have to achieve this? infinite/cheap or few/expensive

- The trade-off between earnings and information is unavoidable. However, in many contexts current solutions can be improved.

- When the number of plays is limited and few, the problem gets harder and solutions need to get smarter by introducing the problem's horizon into them.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In Gani, J., Sarkadi, K., and Vincze, I., editors, *Progress in Statistics* (*European Meeting of Statisticians, Budapest, 1972*), pages 241–266. North-Holland, Amsterdam, The Netherlands.

Villar, S., Bowden, J. and Wason, J. (2015) Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. *Statistical Science Vol. 30, No. 2, 199-215*.

Villar, S., Wason, J. and Bowden, J. (2015) Response-adaptive Randomization for Multi-arm Clinical Trials using the Forward Looking Gittins Index rule *Vol. 71 (4) 696-978*.

John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. Wiley, 2011.

Glazebrook, K. (1980). On randomized dynamic allocation indices for the sequential design of experiments. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 342–346.

Bather, J. (1981). Randomized allocation of treatments in sequential experiments. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 265–292.

Whittle, P. (1980). Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 143–149.

Rothschild, M. (1974). A Two Armed Bandit Theory of Market Pricing. Journal of Economic Theory 9, 185202.

Further Discussion

Questions & Comments

Do you have a month? My thesis was on bandit problems. - Don Berry

Thanks for the attention! ☺