Review: Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges

Ken Tanaka Hernández

The Abdus Salam International Centre for Theoretical Physics

Project for Reinforcement Learning:
Prof. Antonio Celani

July 12, 2023

Introduction
Multi-armed Bandit

$$p(s'|s,a) = \mathbb{I}(s'=s)$$
$$\mathrm{Ber}(\rho), \rho \in (0,1)$$
$$p(\rho|s=(q)_j, a=i) = q_i^\rho (1-q_i)^{1-\rho}/A$$

$$s = (q_1, q_2) \to b(s) = P(q_1, q_2)$$
$$p(\rho|b(s),a) = \int ds b(s) p(r|s,a)$$
$$p(\rho|b(q_1,q_2),1) = \int dq_1 dq_2 P(q_1,q_2) q_1^\rho (1-q_1)^{1-\rho}$$
$$b'(q|r) = q^r (1-q)^{(1-r)} \;\; b(q) \;\; / \int dq b(q)$$
$$V_\pi(b) = \mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t \, r_t \middle| b_0 = b \right]$$
$$V^*(b) = \max_{a\in\{1,2\}} \sum_{b'} p(b'|b,a) \left[r(b',b) + \gamma V^*(b') \right]$$

Procedure 1 Explore Then Commit (ETC)

Input: m

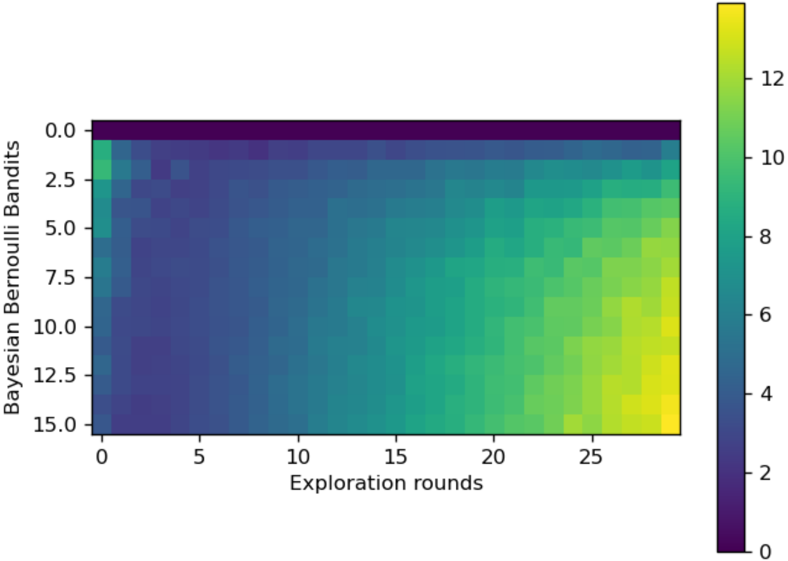
1: $\forall t$

$$A_t = \begin{cases} (t \mod k), \text{if } t \leq mk; \\ \text{argmax}_i \hat{\mu}_i(mk), t > mk \end{cases}$$

Naive Algorithms
Explore Then Commit (ETC)

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)]$$

$$R_n \leq m \sum_{i=1}^{k} \Delta_i + (n - mk) \sum_{i=1}^{k} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right)$$



Bayesian Bernoulli Bandits
0.0
2.5
5.0
7.5
10.0
12.5
15.0
0
5
10
15
20
25
Exploration rounds
12
10
8
6
4
2
0

Procedure 2 Explore Then Eliminate (ETE)

Input: k and $(m_l)_l$

1: $A_1 = 1, \ldots, k$
2: **for** $l = 1, \ldots$ **do**
3: Choose each arm $i \in A_l$ exactly m_l times
4: Let $\hat{\mu}_{i,l}$ be the average reward for arm i from this phase only
5: Update active set:

$$A_{l+1} = \left\{ i : \hat{\mu}_{i,l} + 2^{-l} \geq \max_{j \in A_l} \hat{\mu}_{j,l} \right\}$$

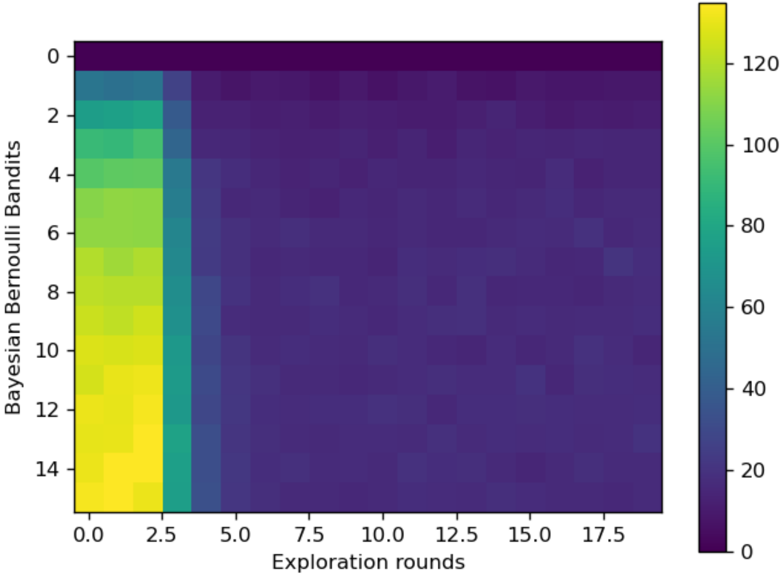
6: **end for**

Naive Algorithms

Explore Then Eliminate (ETE)

$$\mathbb{P}(A_l \ni 1 \notin A_{l+1}) \leq k \exp\left(-\frac{m_l 2^{-2l}}{4}\right)$$

$$R_n \leq C \sum_{i:\Delta_i > 0} \left(\Delta_i + \frac{1}{\Delta_i}\log(n)\right), C > 0$$



Bayesian Bernoulli Bandits
Exploration rounds
0
2
4
6
8
10
12
14
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
120
100
80
60
40
20
0

Motivation

Goals

- Identify the "Best" treatment : *Exploration or Learning*
- Treat patients as "Effectively" as possible during the trial: *Exploitation or Earnings*

Let $y_{k,t} \in \mathbf{Y}_{k,t} \sim \text{Ber}(\rho_k)$, where:

Treatment $\equiv k \in \{1, \ldots, K\} \leftarrow$ Arm

Patient $\equiv t \in \{1, \ldots, N\} \leftarrow$ Time

The *Bayesian* feature: $\rho_k \in \mathbf{P}_k \sim \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} x^{\beta-1}$;

Conjugate Prior distribution

$$(\alpha, \beta) = \begin{cases} (S_{k,0}, F_{k,0}) \in \mathbb{N}_+^2 & \\ (S_{k,0} + S_{k,t}, F_{k,0} + F_{k,t}) & : t \geq 1 \end{cases}$$

(Succesful, Failure) $\equiv (S_{k,t}, F_{k,t}) \in \mathbb{N}_0^2$

Action space:

$$\mathbb{A}_k \ni a_{k,t} = \{0, 1\}$$

Markovian *transition probability rule*:

$$P_k\{\mathbf{s}_{k,t+1} | \mathbf{s}_{k,t}, a_{k,t}\} \sim$$

$$\mathbf{s}_{k,t+1} = \begin{cases} \begin{cases} (S_{k,0} + S_{k,t} + 1, F_{k,0} + F_{k,t}) : (S_{k,0} + S_{k,t})/c_t \\ (S_{k,0} + S_{k,t}, F_{k,0} + F_{k,t} + 1) : (F_{k,0} + F_{k,t})/c_t \end{cases} & : a_{k,t} = 1 \\ \mathbf{s}_{k,t} & : a_{k,t} = 0 \end{cases}$$

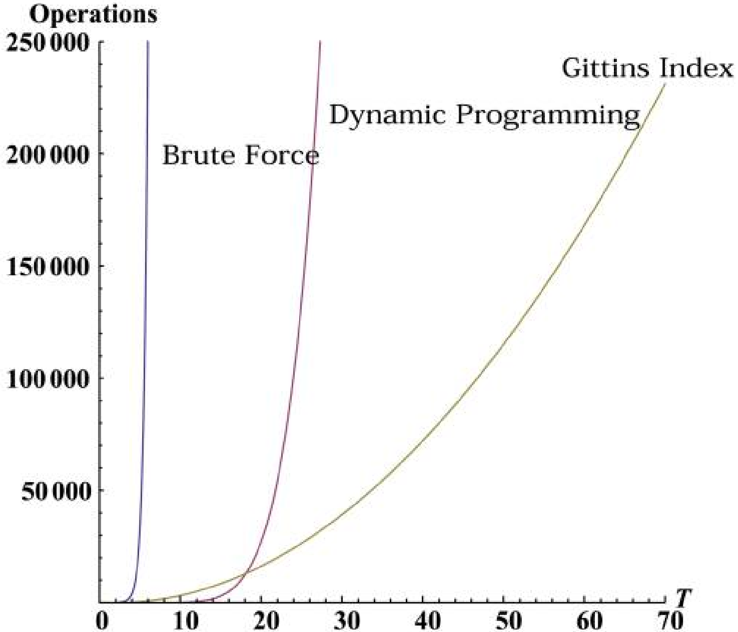
$$c_t = S_{k,0} + S_{k,t} + F_{k,0} + F_{k,t}$$

$$R(\mathbf{s}_{k,t}, a_{k,t}) = \frac{S_{k,0} + S_{k,t}}{c_t} a_{k,t}$$

$$V_\pi^*(\mathbf{s}) = \max_\pi \mathbb{E}_\pi \left[\sum_{t=1}^{N} \gamma^t \sum_{k=1}^{K} R(\mathbf{s}_{k,t}, a_{k,t}) \middle| \mathbf{s}_0 = \mathbf{s} \right]$$

Bayesian regret

$$R_{-1} = N \max_k(\rho_k) - \mathbb{E}_\pi \left[\sum_{k=0}^{K} \sum_{t=1}^{N} a_{k,t} y_{k,text} \right]$$


Operations
250 000
200 000
150 000
100 000
50 000
0
10
20
30
40
50
60
70
T
Brute Force
Dynamic Programming
Gittins Index