Bandit Problem

and its applications in Economics

Jinze LI

April 13, 2024

Content

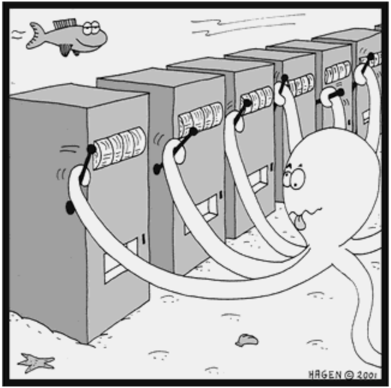
What is Bandit?

One-armed Bandit (Slot Machine 老虎机)

[1]

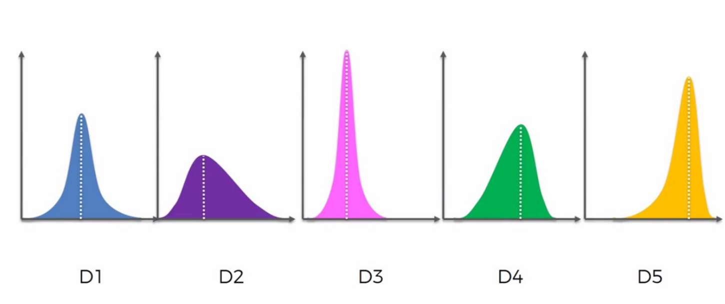
[1] Origin: Empty players' pockets and wallets as thieves

What is Bandit?



Multi-armed Bandit

Bandit Problem



- sequential decision making problem with uncertainty

Introduction

- MAB(Multi-armed bandit) is a set of real distributions, each distribution is associated with the rewards delivered by one of the levels.
- The gambler iteratively plays one lever per round and observes the associated reward.
- The objective is to maximize the sum of the collected rewards.

MAB Definition

- The Optimal Value: $v^* = \mu(a^*) = max_{a \in A} \mu(a)$, where $\mu(a) = E[R|a]$
- The Regret: opportunity loss for one step(Expectation using agent's strategy), $l_t = E[v^* - \mu(a_t)]$
- The total expected regret
$$L_T = E[T \cdot v^* - \sum_{t=1}^{T} \mu(a_t)]$$

Goal: Max the total reward $\Longleftrightarrow$ Min the total regret[1]

Exploration V.S. Exploitation

- Exploitation: Make the best decision given current information (short-term reward)
- Exploration: Gather more information (long-term reward)
- eg. A student learns to eat at AC1,AC2,AC...

Greedy Algorithm

- Select the arm that has the max average reward

Arm	Mean	Period1	Afterwards (w.h.p.)
arm 1	10	7,8,9	End
arm 2	9	9,12,15	Play this forever!

- Total Regret: $(10 - 9)T = T$

Epsilon-Greedy

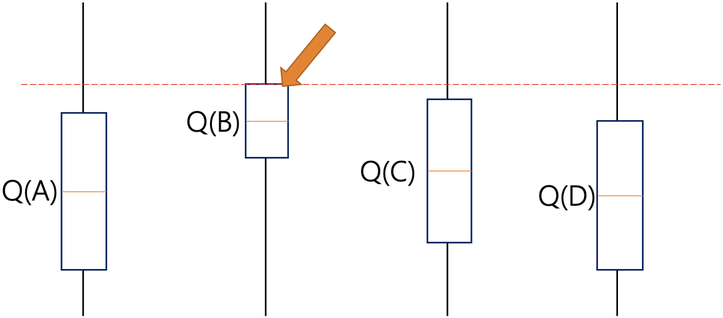
Select action

- (best option) $a_{t+1} = argmax_{a \in A} Q_t(a)$ with probability $1 - \varepsilon$
- (random action) with probability ε

Linear Total Regret porpotional to ε

UCB

- Upper Confidence Bound (UCB) algorithm



UCB

- Estimate an upper confidence $\hat{U}_t(a)$ for each action value, with high probability we have $Q(A) \leq \hat{Q}(A) + \hat{U}_t(a)$
- The upper confidence should decrease with pulls.
- "Optimism in the face of uncertainty."

Short Summary

The Exploitation/Exploration trade-off matters.

Comparision:

- Supervised Machine Learning (eg. Linear Regression)
  — Data are given and fixed.
- Reinforcement Learning (eg. Bandit)
  — Data generated from agents by the adaptive experiment.

Bandit & Exploration Benefit

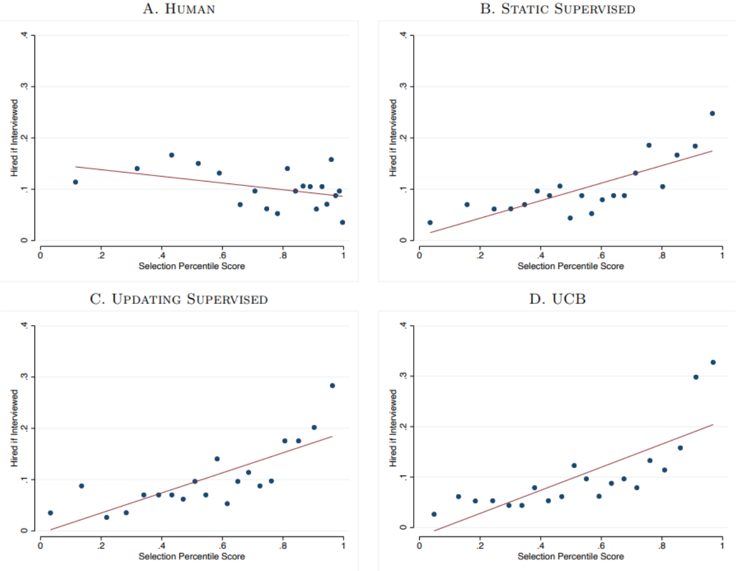
Hiring as exploration [2]

View hiring as a contextual bandit problem:
to find the best workers over time, firms must balance "exploitation" (selecting from groups with proven track records) with "exploration" (selecting from under-represented groups to learn about quality).

- Algorithms decide the candidate
- Hiring yield is a measure of whether a candidate meets the firm's internal hiring criteria

Bandit & Exploration Benefit

FIGURE 2: CORRELATIONS BETWEEN ALGORITHM SCORES AND HIRING LIKELIHOOD



Bandit & Exploration Benefit

Supervised learning may ignore the exploration.
Bandit algorithm gives more chance to women and minorities.

- UCB model increases the proportion of black or Latino applicants interviewed from 10% to 24%.
- All algorithms increases the proportion of women among selected applicants from 35% for human screening to 42% (static SL), 40% (updated SL) and 48% (UCB).
- Average hiring rates for selected applicants were 33%(UCB), 35%(updated SL), and 24%(static SL), compared to 10% for human-screened applicants.

Bandit & Game Theory

Strategic experimentation with bandits [3]

Settings:

- Players face identical two-armed bandit problems.
- The safe arm offers a known and constant flow payoff.
- The risky arm can be either good or bad. Bad one gives a negative payoff, while the good one yields reward by a Poisson process.

Bandit & Game Theory

- Each player is endowed with a stream of one unit of a perfectly divisible resource and, at each point in time, must decide how to split this resource between the two arms.
- Players' actions and outcomes are publicly observed, so there are perfect informational spillovers between players

Bandit & Game Theory

One Application:

Contests for Innovation Experimentation [4]
(Government Procurement)

- One principal and a set of homogeneous agents (contestants)
- Agents play bandits to allocate the innovation

Bandit & Game Theory

Contest Design:

Prize-sharing Scheme

- Equal-sharing V.S. Winner-takes-all

Disclosure Policy

- Hidden V.S. Public

Expected reward V.S. Innovation Feasibility

Bandit & Inference

After the sampling/treatment/adaptive data collection, we would like to know the treatment effect/policy evaluation!

Why not easy?

- Because the data collected strategies make the samples being dependent. (Non i.i.d)
- It is very likely to get biased estimators.

Bandit & Inference—Downward Bias

- Two-arm Bandits which are i.i.d Normal Dist

Arm	Period1	Period2
arm 1	evenly & lucky	oversampling
arm 2	evenly but unlucky	reduction in sampling

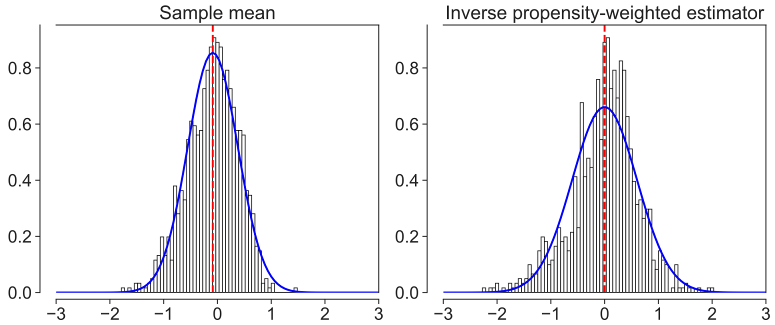
- Because we are greedy, we always get downward bias. [5]
- Asymmetric sampling makes inference difficult.

Bandit & Inference—Methods

- $\widehat{\mathrm{E}}_{\text{Avg}} = \frac{1}{T}\sum_{t=1, W_t=w}^{T} Y_t$
- $\widehat{\mathrm{E}}_{\text{IPW}} = \frac{1}{T}\sum_{t=1}^{T} \frac{I[W_t=w]\cdot Y_t}{\widehat{\pi}(w)}$
- $\widehat{\mathrm{E}}_{AIPW} = \frac{1}{T}\sum_{t=1}^{T}\{\frac{I[W_t=w]\cdot Y_t}{\widehat{\pi}(w)} + (1 - \frac{I[W_t=w]\cdot Y_t}{\widehat{\pi}(w)})\hat{m}_t(w)\}$

Issues: Biased, Heavy-tailed, Non-gaussian

Bandit & Inference—Distributions of Estimators



Summary

Intro to Bandit problem

Algorithms: Greedy/Epsilon-greedy/UCB

Applications:

- Exploration: Hiring as exploration
- Game: Contests for innovation
- Inference: Model for dynamic treatment

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[2] Danielle Li, Lindsey R Raymond, and Peter Bergman. *Hiring as exploration*. Tech. rep. National Bureau of Economic Research, 2020.

[3] Godfrey Keller, Sven Rady, and Martin Cripps. "Strategic experimentation with exponential bandits". In: *Econometrica* 73.1 (2005), pp. 39–68.

[4] Marina Halac, Navin Kartik, and Qingmin Liu. "Contests for experimentation". In: *Journal of Political Economy* 125.5 (2017), pp. 1523–1569.

[5] Xinkun Nie et al. "Why adaptively collected data have negative bias and how to correct for it". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2018, pp. 1261–1269.