Multi Armed Bandits

DOROZHKO Anton

Novosibirsk State University

May 30, 2020

Novosibirsk State University
*THE REAL SCIENCE

Outline

*THE REAL SCIENCE

Framework

MAB is one of the frameworks for **algorithms that make decisions over time under uncertainty**

Examples

in MAB

agent? ~~S?~~ A? R?

1. **News website** : a new user arrives, website picks an article to show, observes user clicks. Goal: maximize the total number of clicks

agent action R cumulative reward

2. **Dynamic pricing** : an app store, customer arrives, the store chooses the price, the customer buys or leaves forever. Goal: maximize the total profit

agent A R{0,1} R{0,p} p_1

3. **Investment** : each morning choose one stock to invest \$. In the end of the day, observe the change in value for each stock. Goal: maximize the total wealth

$p_2 - p_1$

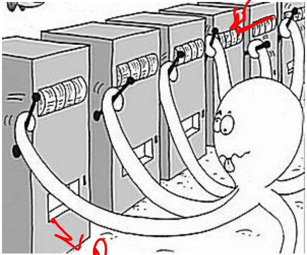
agent? A? R?
trader choice of stock Δ

Novosibirsk State University
*THE REAL SCIENCE

Framework

MAB unifies these examples.
Basic version:

- K possible actions, a.k.a **arms** at each time
- T rounds



One-armed Bandit

k

Connection to MDP

Definition

A Markov Decision Process is a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$

- ~~$\mathcal{S}$ is a space of states~~ no states /or only 1 MAB
- $\mathcal{A}$ is a space of actions
- ~~$\mathcal{P}$ is a transition probability~~

$$\cancel{\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]}$$

- $\mathcal{R}$ is a reward function :

$$\mathcal{R}^a_{\cancel{s}} = \mathbb{E}[R_{t+1} | \cancel{S_t = s}, A_t = a]$$

- γ is a discount factor, $\gamma \in [0, 1]$

THE REAL SCIENCE

Examples MABs

Example	Action	Reward
News website	an article to display	1 if clicked, 0 otherwise
Dynamic pricing	a price to offer	p is sale, 0 otherwise
Ivestment	a stock to invest	change in value during the day

Novosibirsk State University
*THE REAL SCIENCE

Exploration / Exploitation

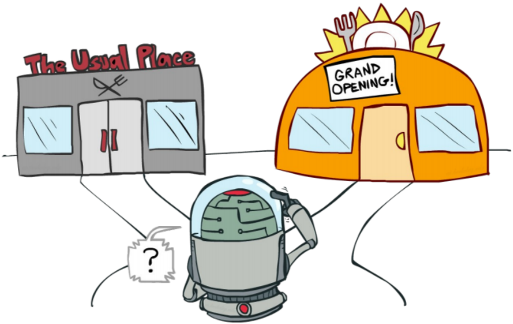
- observe reward only for chosen arm, not for all
- needs to **explore**
- **explore** = try different arms to get new information
- make optimal nea~~t~~r-term decisions based on available info - **exploitation**

Novosibirsk State University
*THE REAL SCIENCE

Exploration vs Exploitation

TRADEOFF

learn which arm is the best, but not spend much time learning



[1] CS188 http://ai.berkeley.edu/home.html

More complex MABs **Feedback**

Auxiliary feedback : other than the reward for chosen arm

Example	Auxiliary feedback	Reward for any other arm?
1) News website	N/A	no
2) Dynamic pricing	sale =>sale at any lower price no sale =>no sale for higher price	yes, for some arms
3) Ivestment	change in value for all stocks	yes, for all arms

- *bandit feedback* : reward for only the chosen arm
- *full feedback* : reward for all arms, that can be chosen
- *partial feedback* : only for some arms

Novosibirsk State University
*THE REAL SCIENCE

More complex MABs **Reward**

$P(1 \text{ from } a_1, 1 \text{ from } a_2) = p_1 \cdot p_2$ p_3 p_4 p_4

a_1 a_2 a_3 a_4 a_5

identical independent distributions

- *IID rewards* : the reward for each arm is drawn from fixed distribution that depends on the arm, but not on the round t
- *Adversarial rewards*: rewards can be arbitrary, as if they are chosen by "adversary" to fool the agent
- *Constrained adversary*: as Adversarial rewards + some constraints. (e.g. cannot change much from one round to another, ...)
- *Stochastic rewards* : rewards evolves over time as random process, e.g. random walk.

Novosibirsk State University
*THE REAL SCIENCE

More complex MAB **Contexts**

RL: $\pi : S \to a$

Contextual bandits $\pi : X \to a$

Contextual MABs

each round, agent can observe some **context** for each action

goal: learn the best **policy** which maps context to arms, while not spending much time learning

$\boxed{s} \to \underset{a}{\bullet} \to \boxed{s'}$

Example	Context
News website	user location and demographics
Dynamic pricing	customer's device, location, ...
Investments	earning multipliers, state of the company, ...

More examples

Application domain	Action	Reward
medical trials	which drug to prescribe	health outcome
web design	font color or page layout	#clicks A/B
content optimization	which item/article to emphasize	#clicks
recommender systems	which movie to watch	1 if follows recommendation
datacenter design	which server to route the job to	job completion time
robot control	a "strategy" for a given task	job completion time
radio networks	which radio frequency to use ?	1 if successful transmission
crowdsourcing	which task to give to which workers, at which price	1 if task completed at sufficient quality

Novosibirsk State University
*THE REAL SCIENCE

Stochastic Bandits

Bernoulli bandits

- Given: K arms, T rounds
- at each round $t \in [T]$
  1. agent picks arm a_t
  2. agent observes reawrd $r_t \in [0,1]$ for the chosen arm
- Goal: maximize total reward over T rounds

$r_t \sim$ Bernolli distr

p_{a_i} $r=1$

$1-p_{a_i}$ $r=0$

$$\max \sum_{t=1}^{T} r_t$$

Novosibirsk State University
*THE REAL SCIENCE

Notation

action

- arms a, rounds t
- mean reward of arm a : $\mu(a) = \mathbb{E}[D_a]$
- best mean reward $\mu^* = max_a \mu(a)$
- difference / gap of arm a : $\Delta(a) = \mu^* - \mu(a)$

Novosibirsk State University
*THE REAL SCIENCE

Regret : Motivation

- How do we argue if agent is doing a good job ?

Regret : Motivation

- How do we argue if agent is doing a good job ?
- Different tasks will have different rewards

Regret : Motivation

- How do we argue if agent is doing a good job ?
- Different tasks will have different rewards
- Some problems have inherently higher rewards

Novosibirsk State University
*THE REAL SCIENCE

Regret : Motivation

- How do we argue if agent is doing a good job ?
- Different tasks will have different rewards
- Some problems have inherently higher rewards
- **Standard approach** - compare to the best-arm benchmark

$\mu^* \cdot T$

$$\mu^* = \max_a \mu(a)$$

NOVOSIBIRSK STATE UNIVERSITY
*THE REAL SCIENCE

Regret : Definition

Definition

Regret at round T is a difference between the expected reward of always playing and optimal arm and the algorithm's cumulative reward:

$$R(T) = \mu^* \cdot T - \sum_{t=1}^{T} \mu(a_t)$$

realized

expected → $\mathbb{E}[R(T)]$

Novosibirsk State University

*THE REAL SCIENCE

Explore First

for K·N steps
first

1. Exploration phase: try each arm N times
2. Select the arm $\hat{a}$ with the highest average reward (break ties arbitrarily)
3. Exploitation phase: play arm $\hat{a}$ in all the remaining rounds

N is parameter

Novosibirsk State University
*THE REAL SCIENCE

Hoeffding's inequality

$X \sim [a_i, b_i]$

$S_n = \frac{X_1 + X_2 \dots X_N}{N}$

$$P(|S_n - \mathrm{E}[S_n]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

n

for Bernoulli $a_i = 0$

$b_i = 1$

$P(|\bar{X} - \mathbb{E}[X]| \geq t) \leq 2e^{-2nt^2}$

Novosibirsk State University
*THE REAL SCIENCE

Explore First regret bounds

$\mu(a)$, $\mu(a^*)$

$P(|X - E[X]| \geq t) \leq 2e^{-2nt^2}$

$p_1 \quad p_2 \quad \dots \quad p_k$

$\mu(a_1) = p_1$ for $B(p)$

compute after exploration

$$P\{|\bar{\mu}(a) - \mu(a)| \leq r(a)\} \geq 1 - 2e^{-2Tr^2(a)}$$

$$r(a) = \sqrt{\frac{2\log T}{N}} \qquad \geq 1 - \frac{2}{T^4}$$

suppose "clean event"

$$\mu(a) - r(a) \leq \bar{\mu}(a) \leq \mu(a) + r(a)$$

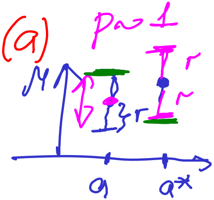
Explore First regret bounds 2

$K = 2$

$$R(T) = \mu^* \cdot T - \sum \mu(a)$$

exploration — exploitation

best a $\max \bar{\mu}$



KN — a — T

chosen after KN steps

$\underline{a \neq a^*}$

$$\mu(a) > \mu(a^*) \quad \underline{\mu(a) + r(a)} \geq \underbrace{\bar{\mu}(a)} \geq \underline{\mu(a^*) - r(a^*)}$$

$$\mu(a^*) - \mu(a) \leq r(a) + r(a^*) = O\left(\sqrt{\frac{\log T}{N}}\right) \text{ after } KN$$

$$R(T) = \sum_{t=1}^{T} (\mu^* - \mu(a)) \leq N + O\left(\sqrt{\frac{\log T}{N}} \cdot (T - 2N)\right) \leftarrow ?$$

$\downarrow 0$ $\downarrow 1$ explore — exploit

$2N$ $\quad N \quad N$

Novosibirsk State University
*THE REAL SCIENCE

Explore First regret bounds conclusion

$N << T$

$R(T) \le N + O\left(\sqrt{\frac{\log T}{N}} \times T\right) \leftarrow \min R(T)$ w.r.t N

$N = \sqrt{\frac{\log T}{N}}\, T \rightarrow N^3 = \log T \cdot T^2$

$N = T^{2/3} (\log T)^{1/3}$

Explore-first achieves regret

$$\mathbb{E}[R(T)] \le O\left(T^{2/3} \times \cancel{O}(K \log T)^{1/3}\right)$$

Observations:

- Performance of exploration phase is terrible

Explore First regret bounds conclusion

Explore-first achieves regret

$$\mathbb{E}[R(T)] \leq T^{2/3} \times O(K \log T)^{1/3}$$

Observations:

- Performance of exploration phase is terrible
- It's better to spread exploration more uniformly over time.

Novosibirsk State University
*THE REAL SCIENCE

Explore First regret bounds conclusion

played all actions N times each $\to a^*$

Explore-first achieves regret

$$\mathbb{E}[R(T)] \leq T^{2/3} \times O(K \log T)^{1/3}$$

Observations:

- Performance of exploration phase is terrible
- It's better to spread exploration more uniformly over time.
- E.g. with ϵ-Greedy exploration

$1-\varepsilon$ a_t^*

ε random a

Novosibirsk State University
*THE REAL SCIENCE

ϵ-Greedy exploration

for *each round* $t = 1, 2, \ldots$ **do**
  Toss a coin with success probability ϵ_t;
  if *success* **then**
    explore: choose an arm uniformly at random
  else
    exploit: choose the arm with the highest average reward so far
end

Algorithm 1.2: Epsilon-Greedy with exploration probabilities $(\epsilon_1, \epsilon_2, \ldots)$.

Novosibirsk State University
*THE REAL SCIENCE

ϵ-Greedy exploration regret

\# all rounds

Explore-first achieves regret

$$\mathbb{E}[R(T)] \leq T^{2/3} \times O(K \log T)^{1/3}$$

T

ϵ-Greedy exploration regret with $\epsilon = t^{-1/3} \cdot (K \log t)^{1/3}$ $\leq$

$$\mathbb{E}[R(t)] \leq t^{2/3} \times O(K \log t)^{1/3}$$

t

for each round t

- ϵ-greedy regret grows linearly $T^{2/3}$ O
- UCB and Thompson sampling grows with $log(T)$

Novosibirsk State University
*THE REAL SCIENCE

upper confidence bound

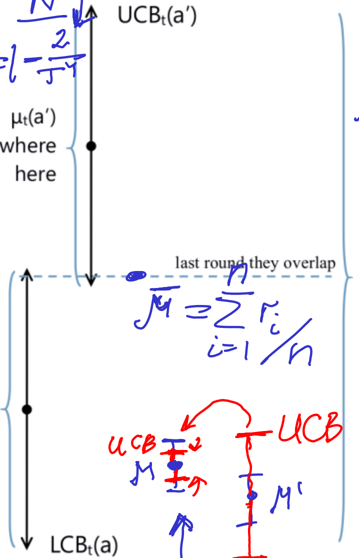
Hoeffding's ineq

N

$$P(|\mu - \mu^*| \le r) = 1 - \frac{2}{T^4}$$

$$r = \sqrt{\frac{2\log T}{n_1}}$$

$$r = \sqrt{\frac{2\log T}{N}}$$



$$\bar{\mu} = \sum_{i=1}^{n} r_i / n$$

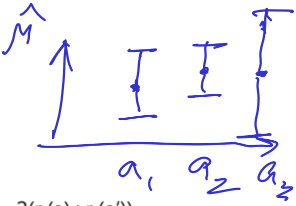
UCB

μ μ'

t $\quad$ $t \nearrow$

n' - const

$n_1, n_2, n_3 \dots$



$t \nearrow \quad r \nearrow$

$n_1 \nearrow \quad r \searrow$

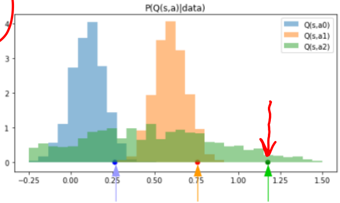
Optimism in face of uncertainty

UCB

- Policy:
  - Compute 95% upper confidence bound *for each a*
  - Take action with highest confidence bound
  - Adjust: change 95% to more/less

$\underset{a}{\operatorname{argmax}}\, Q(s,a)$

ε greedy



points = 95% percentiles

$Q(s,a) \to$ scalar

$\to \mu$, σ

Q(s,a)

Novosibirsk State University
*THE REAL SCIENCE