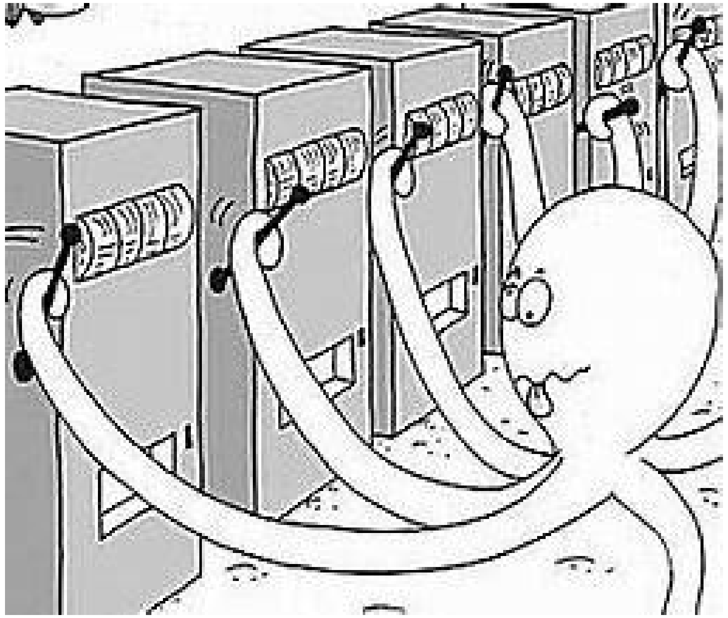
Dueling Bandit Review

Bowen Xu

Sep 2022

Outline

- Review of Multi Bandits Algorithm
- Introduction of Dueling Bandits
- Self-Sparring Algorithm
    - Independent Self Sparring Algorithm
    - Kernel Self Sparring Algorithm



Multi-Armed Bandit (MAB) Problem

- K arms (actions)
- Each arms has an average reward: μ
  - Unknown to agent
  - Assume μ_1 (reward of arms 1) is the largest among K arms
- Procedure: For t = 1,……,T:
  - Algorithm chooses arm a(t) = i
  - Receive random reward y(t) from the chosen arm i
    - Expectation reward $\mu_{a(t)}$
- Objective: minimize total regret
  - Regret: $T\mu_1 - (\mu_{a(1)} + \ldots\ldots + \mu_{a(T)})$

Example of MAB Problem

Time	1	2	3	4	5	6	7	8	9	10
Left arm	\$1	\$0			\$1	\$1	\$0			
Right arm			\$1	\$0						

Example of MAB Problem

Time	1	2	3	4	5	6	7	8	9	10
Left arm	\$1	\$0			\$1	\$1	\$0			
Right arm			\$1	\$0						

- Average reward in first 7 slots:
  - Left arm: 4/7
  - Right arm: 1/2
  - Conclustion: $\mu_{left} > \mu_{right}$ in first 7 slots
- Regret: $R(7) = 7 * \mu_{left} - (4 * \mu_{left} + 3 * \mu_{right})$

Example of MAB Problem

Time	1	2	3	4	5	6	7	8	9	10
Left arm	\$1	\$0			\$1	\$1	\$0			
Right arm			\$1	\$0						

- Exploit and Exploration Trade-off:
  - At 8th slots:
    - Exploration: pull righ arms (less reward but less chosen times)
    - Exploitation: pull left arms (more reward in former slots)

Thompson Sampling

- θ_k : an action's success probability or mean reward
  - Prior of each θ_k satisfied Beta distribution Beta(α_k, β_k)
- x_t: the actions selected at time t
  - $x_t \leftarrow argmax_k \, \theta_k$
- r_t: the corresponding reward of action x_t, r_t satisfies Bern(θ_k), if x=k
- Each action's posterior distribution is also Beta with parameters updated as follows:
- $(\alpha_k, \beta_k) \leftarrow \begin{cases} (\alpha_k, \beta_k) & if \; x_t \neq k \\ (\alpha_k, \beta_k) + (r_t, 1 - r_t) & if \; x_t = k \end{cases}$

Drawback of Conventional MAB problem:

- When payoff is a **relative comparison** result rather than an absolute value, it is difficult to apply conventional MAB Algorithm.

Introduction of Dueling Bandits

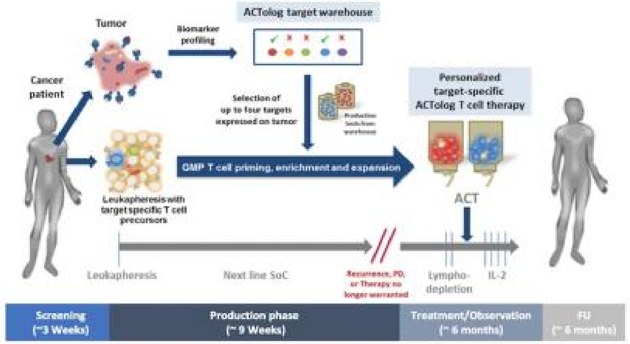
- Motivation
  - Solve the problem with only **binary feedback** about the **relative reward** of two chosen strategies is available
- Suitable application scenarios:
  - Search engine
  - Online advertising

Applications of Dueling Bandits

(a) search engine

(b) online advertisement

(c) recommend system



(d) personalized clinical treatment

Introduction of Dueling Bandits

- K arms (actions)
  - For each pair of arms A and B, they have a probability to beat each other
    - i.e. $P(A > B)$ means the probability of A beating B. $P(B > A)$ means the probability of B beating A.
    - $P(A > B) - 0.5 = 0.5 - P(B > A)$
    - $P(A > B) - 0.5$ generally written as Δ_{AB}(distinguishability), so we have $\Delta_{AB} = -\Delta_{BA}$.
  - Suppose there exists an optimal arm b^*which can beats all other arms (Condorcet winner)

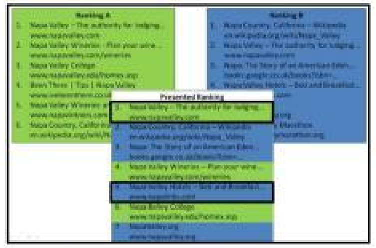
Introduction of Dueling Bandits

- Procedure: for For t = 1,……,T:
  - Choose two arms b and b’ and compare
  - Observe the outcome
    - e.g. arm b_t beats b'_t at slot t.
- Objective: minimize total regret and find the Condorcet winner
  - Regret: $R_T = \sum_{t=1}^{T} (P(b^* > b_t) + P(b^* > b'_t)) - 1$

Example of Dueling Bandit

- Suppose we have 3 page lists: A,B,C
  - We need to find the optimal one for user
  - Interleave the two lists and let user find their favourite page.
  - If the favourite page ranks highest in lists A, then A beats the other lists.

Example of Dueling Bandit



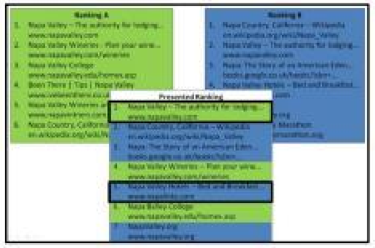
Interleave A vs B

. . .

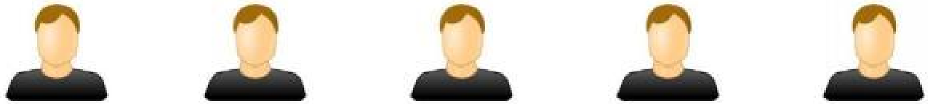
	Left wins	Right wins
A vs B	0	1
A vs C	0	0
B vs C	0	0

[From Yisong Yue]

Example of Dueling Bandit



Interleave A vs C



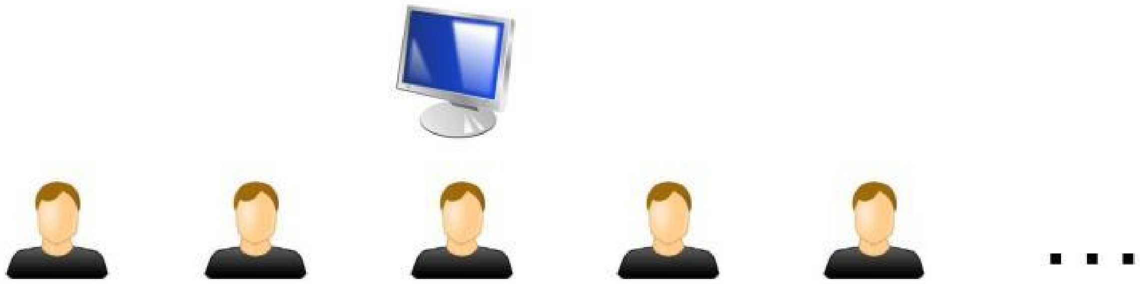
. . .

	Left wins	Right wins
A vs B	0	1
A vs C	0	1
B vs C	0	0

[From Yisong Yue]

Example of Dueling Bandit

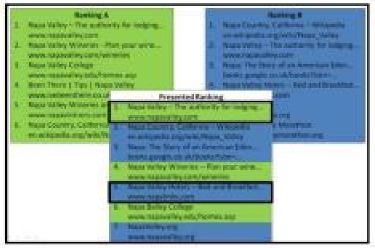
Interleave B vs C



	Left wins	Right wins
A vs B	0	1
A vs C	0	1
B vs C	0	1

[From Yisong Yue]

Example of Dueling Bandit



Interleave A vs C

...

	Left wins	Right wins
A vs B	0	1
A vs C	1	1
B vs C	0	1

[From Yisong Yue]

Example of Dueling Bandit

- From the first 4 users:
  - lists C wins more times than A and B
  - lists C is the optimal for the time being
- Trade-off for the 5th user:
  - Exploitation: interleave C with itself (C wins the most times)
  - Explore: compare B vs A (B and A compare the fewer times than C)

More Complex Dueling Bandit Algorithms

- How to choose the two arms to compare at each slot?
- What if the arms are dependent?

Self-Sparring

- Idea:
  - Applying the conventional MAB algorithms to solve dueling bandit problem
  - View the dueling bandit as the dueling of two arms with different MAB strategies

Self-Sparring

- Instantiate 2 conventional MAB algorithms: P_1 & P_2
- For t = 1, ……
  - P_1 chooses a_1
  - P_2 chooses a_2
  - Duel a_1 vs a_2
  - Provide feedback

Ind-Self-Sparring

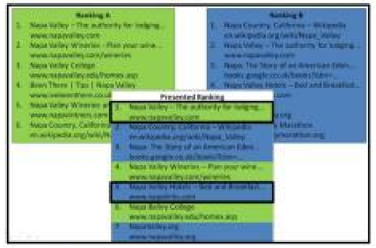
- For independent arms cases, we can choose some conventional MAP algorithms as P_1 and P_2. e.g. Thompson Sampling, UCB.
- Generally, we use Thompson Sampling in Self-Sparring
  - choose arms:
    - $\theta_k \sim Beta(\alpha_k, \beta_k)$
    - $x_t \leftarrow argmax_k\, \theta_k$
  - Provide feedback:
    - pairwise feedback matrix: $R = \{r_{ij} \in \{0,1,\emptyset\}\}_{K\times K}$
    - $(\alpha_k, \beta_k) \leftarrow \begin{cases} (\alpha_k, \beta_k) & if\ x_t \neq k \\ (\alpha_k, \beta_k) + (r_t, 1 - r_t) & if\ x_t = k \end{cases}$

Example of Ind-Self-Sparring

- Initialization:

	α	β
A	5	5
B	5	5
C	5	5

Example of Ind-Self-Sparring



Interleave A vs B

. . .

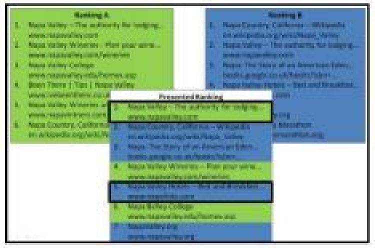
	α	β
A	5	6
B	6	5
C	5	5

$$\theta_A \sim Beta(5,6)$$
$$\theta_B \sim Beta(6,5)$$
$$\theta_C \sim Beta(5,5)$$

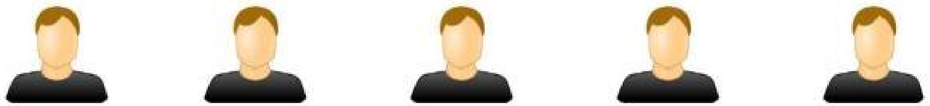
	Left wins	Right wins
A vs B	0	1
A vs C	0	0
B vs C	0	0

[From Yisong Yue]

Example of Ind-Self-Sparring



Interleave A vs C



. . .

	α	β
A	5	7
B	6	5
C	6	5

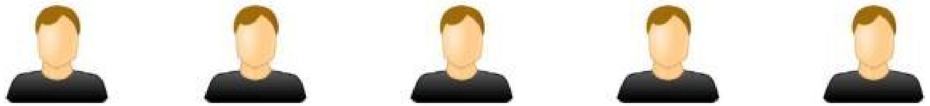
$$\theta_A \sim Beta(5,7)$$
$$\theta_B \sim Beta(6,5)$$
$$\theta_C \sim Beta(6,5)$$

	Left wins	Right wins
A vs B	0	1
A vs C	0	1
B vs C	0	0

[From Yisong Yue]

Example of Ind-Self-Sparring

Interleave B vs C



...

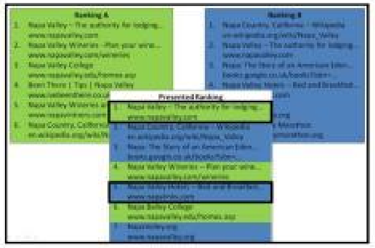
	Left wins	Right wins
A vs B	0	1
A vs C	0	1
B vs C	0	1

	α	β
A	5	7
B	6	6
C	7	5

$$\theta_A \sim Beta(5,7)$$
$$\theta_B \sim Beta(6,6)$$
$$\theta_C \sim Beta(7,5)$$

[From Yisong Yue]

Example of Ind-Self-Sparring



Interleave A vs C

. . .

	Left wins	Right wins
A vs B	0	1
A vs C	1	1
B vs C	0	1

	α	β
A	6	7
B	6	6
C	7	6

$$\theta_A \sim Beta(6,7)$$
$$\theta_B \sim Beta(6,6)$$
$$\theta_C \sim Beta(7,6)$$

[From Yisong Yue]

What about Dependent Arms Cases?

- Generally, we use **covariance** to describe the dependency
- K arms can be modeled to a collection of r.v. with characteristics below:
  - **multivariate Gaussian distribution**
  - **reward mean**
  - **covariance function**

Gaussian Process

Kernel-Self-Sparring

- For dependent arms cases, we use Gaussian Process to describe the dependency
- Gaussian Process:
  - use covariance between arms to model the dependency
    - Covariance Matrix $C = (c_{ij})_{K \times K}$
    - posterior inference updates the mean reward vector μ and the covariance matrix σ

Kernel-Self-Sparring

- Operation in Kernel-Self-Sparring
  - choose arms:
    - $\theta_k \sim GP(\mu_{t-1}, \sigma_{t-1})$
      - (sample from Gaussian Process: by marginal distribution
      - https://peterroelants.github.io/posts/gaussian-process-tutorial/
      - https://blog.csdn.net/shenxiaolu1984/article/details/50386518)
    - $x_t \leftarrow argmax_k\, \theta_k$
  - Provide feedback:
    - pairwise feedback matrix: $R = \{r_{ij} \in \{0,1,\emptyset\}\}_{K\times K}$
    - Beyesian update using R to obtain (μ_t, σ_t)

Example of Kernel-Self-Sparring

- Initialization:
  - mean reward and covariance matrix

	μ
A	$(\mu_A)_0$
B	$(\mu_B)_0$
C	$(\mu_C)_0$

Cov	A	B	C
A	$(\sigma_A)_0$	$(\sigma_{AB})_0$	$(\sigma_{AC})_0$
B	$(\sigma_{BA})_0$	$(\sigma_B)_0$	$(\sigma_{BC})_0$
C	$(\sigma_{CA})_0$	$(\sigma_{CB})_0$	$(\sigma_C)_0$

Example of Kernel-Self-Sparring

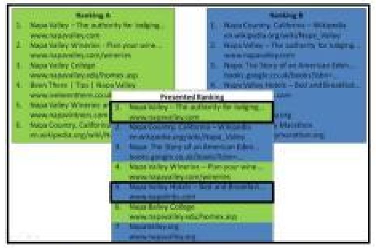
- Bayesian Update:
  - Prior distribution $(\mu_{t-1}, \sigma_{t-1})$ to Posterior distribution (μ_t, σ_t)
  - **Conjugate distribution** of Multivariate Gaussian distribution

	μ
A	$(\mu_\mathrm{A})_0$
B	$(\mu_\mathrm{B})_0$
C	$(\mu_\mathrm{C})_0$

Cov	A	B	C
A	$(\sigma_A)_0$	$(\sigma_{AB})_0$	$(\sigma_{AC})_0$
B	$(\sigma_{BA})_0$	$(\sigma_B)_0$	$(\sigma_{BC})_0$
C	$(\sigma_{CA})_0$	$(\sigma_{CB})_0$	$(\sigma_C)_0$

Example of Kernel-Self-Sparring

	μ
A	$(\mu_A)_1$
B	$(\mu_B)_1$
C	$(\mu_C)_1$



Interleave A vs B

...

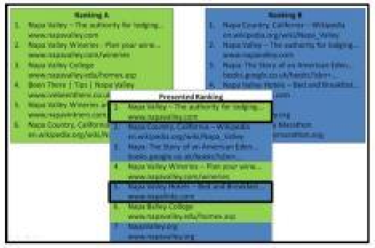
Cov	A	B	C
A	$(\sigma_A)_1$	$(\sigma_{AB})_1$	$(\sigma_{AC})_1$
B	$(\sigma_{BA})_1$	$(\sigma_B)_1$	$(\sigma_{BC})_1$
C	$(\sigma_{CA})_1$	$(\sigma_{CB})_1$	$(\sigma_C)_1$

	Left wins	Right wins
A vs B	0	1
A vs C	0	0
B vs C	0	0

[From Yisong Yue]

Example of Kernel-Self-Sparring

	μ
A	$(\mu_A)_2$
B	$(\mu_B)_2$
C	$(\mu_C)_2$



Interleave A vs C

. . .

Cov	A	B	C
A	$(\sigma_A)_2$	$(\sigma_{AB})_2$	$(\sigma_{AC})_2$
B	$(\sigma_{BA})_2$	$(\sigma_B)_2$	$(\sigma_{BC})_2$
C	$(\sigma_{CA})_2$	$(\sigma_{CB})_2$	$(\sigma_C)_2$

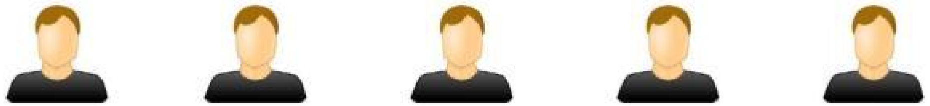
	Left wins	Right wins
A vs B	0	1
A vs C	0	1
B vs C	0	0

[From Yisong Yue]

Example of Kernel-Self-Sparring

	μ
A	$(\mu_A)_3$
B	$(\mu_B)_3$
C	$(\mu_C)_3$

Interleave B vs C



. . .

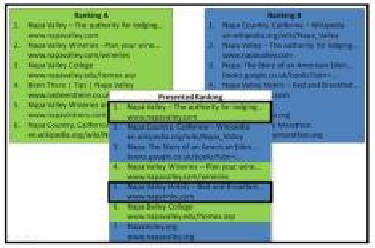
Cov	A	B	C
A	$(\sigma_A)_3$	$(\sigma_{AB})_3$	$(\sigma_{AC})_3$
B	$(\sigma_{BA})_3$	$(\sigma_B)_3$	$(\sigma_{BC})_3$
C	$(\sigma_{CA})_3$	$(\sigma_{CB})_3$	$(\sigma_C)_3$

	Left wins	Right wins
A vs B	0	1
A vs C	0	1
B vs C	0	**1**

[From Yisong Yue]

Example of Kernel-Self-Sparring

	μ
A	$(\mu_A)_4$
B	$(\mu_B)_4$
C	$(\mu_C)_4$



Interleave A vs C

Cov	A	B	C
A	$(\sigma_A)_4$	$(\sigma_{AB})_4$	$(\sigma_{AC})_4$
B	$(\sigma_{BA})_4$	$(\sigma_B)_4$	$(\sigma_{BC})_4$
C	$(\sigma_{CA})_4$	$(\sigma_{CB})_4$	$(\sigma_C)_4$

...

	Left wins	Right wins
A vs B	0	1
A vs C	1	1
B vs C	0	1

[From Yisong Yue]

Theoretical Analysis

- Regret bound: $O(K/\epsilon\ logT)$
  - K : # of Arms
  - T: time horizon
  - Distinguishability between the best 2 arms: Δ_{12}