Sequencing as an Instrument in Decentralized Bandits with Myopic Agents

Carlos Gonzalez
University of Oxford, Department of Economics

Univ Graduate Economists' Lunch Seminar

7th February

Outline

Canonical Bandit Problems

- Sequencing as an Instrument in Decentralized Bandits with Myopic Agents

Canonical Bandit Problems

- Sequencing as an Instrument in Decentralized Bandits with Myopic Agents
- Choosing a route from the Department to Univ

Canonical Bandit Problems

- Sequencing as an Instrument in Decentralized Bandits with Myopic Agents
- Choosing a route from the Department to Univ
- Exploitation vs Exploration

Canonical Bandit Problems

- Sequencing as an Instrument in Decentralized Bandits with Myopic Agents
- Choosing a route from the Department to Univ
- Exploitation vs Exploration
- A more interesting example: Restaurants in Google Maps

Canonical Bandit Problems

- Sequencing as an Instrument in Decentralized Bandits with Myopic Agents
- Choosing a route from the Department to Univ
- Exploitation vs Exploration
- A more interesting example: Restaurants in Google Maps
- Refined theory which prescribe optimal sequence of actions in the limit (even in adversarial scenarios)

Myopic Agents

- Wait a second... this is not how Google Maps recommendations work

Myopic Agents

- Wait a second... this is not how Google Maps recommendations work
- Carlos does not want to explore, he just wants to go to the best restaurant in town

- Wait a second... this is not how Google Maps recommendations work
- Carlos does not want to explore, he just wants to go to the best restaurant in town
- Agents do not internalize the returns to exploration. They would just rater exploit. We call them **myopic** (although they remain rational)

- Wait a second... this is not how Google Maps recommendations work
- Carlos does not want to explore, he just wants to go to the best restaurant in town
- Agents do not internalize the returns to exploration. They would just rater exploit. We call them **myopic** (although they remain rational)
- In a MT context, this problem can be depicted as an **Incentive Compatibility** problem between a long lived policy maker (the Principal) who would like to trade-off exploration and exploitation optimally, and a set of myopic pure exploitation agents

Myopic Agents

- Wait a second... this is not how Google Maps recommendations work
- Carlos does not want to explore, he just wants to go to the best restaurant in town
- Agents do not internalize the returns to exploration. They would just rater exploit. We call them **myopic** (although they remain rational)
- In a MT context, this problem can be depicted as an **Incentive Compatibility** problem between a long lived policy maker (the Principal) who would like to trade-off exploration and exploitation optimally, and a set of myopic pure exploitation agents
- These problems are referred as **Decentralized Bandits** (as exploration is delegated to agents) with Myopic Agents (as they are all about exploitation)

- Some Literature in CS including transfers [1], bayesian persuasion [2], [3]

- Some Literature in CS including transfers [1], bayesian persuasion [2], [3]
- I explore an alternative instrument for inducing optimal exploration: **Sequencing**

- Some Literature in CS including transfers [1], bayesian persuasion [2], [3]
- I explore an alternative instrument for inducing optimal exploration: **Sequencing**
- Intuition: Agents are more likely to accept early offers, than late offers (regardless their priors). Think of it as they having a discount factor δ

- Some Literature in CS including transfers [1], bayesian persuasion [2], [3]
- I explore an alternative instrument for inducing optimal exploration: **Sequencing**
- Intuition: Agents are more likely to accept early offers, than late offers (regardless their priors). Think of it as they having a discount factor δ
- Can the Principal create **ordering strategies** which induce optimal level of exploration?

- Some Literature in CS including transfers [1], bayesian persuasion [2], [3]
- I explore an alternative instrument for inducing optimal exploration: **Sequencing**
- Intuition: Agents are more likely to accept early offers, than late offers (regardless their priors). Think of it as they having a discount factor δ
- Can the Principal create **ordering strategies** which induce optimal level of exploration?
- Can she do so when δ is unknown or only behaviour and not outcomes are observed?

- Hot topic in Computational Econ, Econ Theory and Learning Theory

Motivation

- Hot topic in Computational Econ, Econ Theory and Learning Theory
- Cool economic applications (see operating example to come), especially in market (platforms) and experimental design

Motivation

- Hot topic in Computational Econ, Econ Theory and Learning Theory
- Cool economic applications (see operating example to come), especially in market (platforms) and experimental design
- Cool non-economic applications: Is there any economics in the way that Elon Musk orders the tweets that we see?

Notation and Operating Example

- Public Officer (Principal - she) wants to match workers and firms. Ideally, send workers to firms with the highest MPL, but MPL is not observed $\rightarrow$ Bandit Problem!

- Public Officer (Principal - she) wants to match workers and firms. Ideally, send workers to firms with the highest MPL, but MPL is not observed $\rightarrow$ Bandit Problem!
- Two firms j, h, hence two possible orderings $\{jh, hj\}$

- Public Officer (Principal - she) wants to match workers and firms. Ideally, send workers to firms with the highest MPL, but MPL is not observed $\rightarrow$ Bandit Problem!
- Two firms j, h, hence two possible orderings $\{jh, hj\}$
- Workers (Agents - he) hold invariant priors over the firms m_h^0, m_j^0. They have discount factor δ_i. Whenever worker i visits a firm j, a random *iid* reward $m_j^i \sim M_j \perp\!\!\!\perp \delta_i$ is realized. They can then decide whether to accept such offer T or continue C and get to see the next offer realization

Notation and Operating Example

- Public Officer (Principal - she) wants to match workers and firms. Ideally, send workers to firms with the highest MPL, but MPL is not observed $\rightarrow$ Bandit Problem!

- Two firms j, h, hence two possible orderings $\{jh, hj\}$

- Workers (Agents - he) hold invariant priors over the firms m_h^0, m_j^0. They have discount factor δ_i. Whenever worker i visits a firm j, a random *iid* reward $m_j^i \sim M_j \perp\!\!\!\perp \delta_i$ is realized. They can then decide whether to accept such offer T or continue C and get to see the next offer realization

- Workers only play once. As soon as they continue, the offer remains no longer available. Outside option normalized to 0. In case of indifference they will play in this period

Notation and Operating Example II

- Workers play optimally according to these principles (i.e. in case of order jh)

$$a^{jh} = \begin{cases} T & \text{if } m_j^i \geq \delta_i \cdot m_h^0 \\ \{C, T\} & \text{if } m_j^i < \delta_i \cdot m_h^0 \ \ \& \ \ m_h^i \geq 0 \\ \{C, C\} & \text{if } m_j^i < \delta_i \cdot m_h^0 \ \ \& \ \ m_h^i < 0 \end{cases} \tag{1}$$

Notation and Operating Example II

- Workers play optimally according to these principles (i.e. in case of order jh)

$$a^{jh} = \begin{cases} T & \text{if } \ m_j^i \geq \delta_i \cdot m_h^0 \\ \{C, T\} & \text{if } \ m_j^i < \delta_i \cdot m_h^0 \ \ \& \ \ m_h^i \geq 0 \\ \{C, C\} & \text{if } \ m_j^i < \delta_i \cdot m_h^0 \ \ \& \ \ m_h^i < 0 \end{cases} \tag{1}$$

- For simplicity we further assume that $m_J^i \sim Log(\mu_J, 1)$

Notation and Operating Example II

- Workers play optimally according to these principles (i.e. in case of order jh)

$$a^{jh} = \begin{cases} T & \text{if } m_j^i \geq \delta_i \cdot m_h^0 \\ \{C, T\} & \text{if } m_j^i < \delta_i \cdot m_h^0 \;\&\; m_h^i \geq 0 \\ \{C, C\} & \text{if } m_j^i < \delta_i \cdot m_h^0 \;\&\; m_h^i < 0 \end{cases} \tag{1}$$

- For simplicity we further assume that $m_J^i \sim Log(\mu_J, 1)$
- The Principal may select a sequence of orderings $p \in P_J$ based on the history of agents actions and outcome realizations. Define a policy $\pi : H_i \to P_J$. We characterize the problem of the Principal in terms of regret

$$\min_{\pi} N \cdot \sup_{p_J^P} \mathbb{E}\left[r^p\right] - \mathbb{E}\left[\sum_{i}^{N} r_i^{p^{\pi,i}}\right] \tag{2}$$

Expected Returns

- Under full information (i.e. $\mu, F_\delta \in H_0$), the Principal could simply play the sequence with the highest expected return

$$\mathbb{E}[r_i^{jh}] = q_{jh} \cdot \mu_j + (1 - q_{jh}) q_h^0 \cdot \delta_{jh} \cdot \mu_h \tag{3}$$

where

$$q_J^0 = \mathbb{P}(m_J^i \geq 0) = \mathbb{P}(\epsilon_i > -\mu_J) = \frac{\exp(\mu_J)}{1 + \exp(\mu_J)} \tag{4}$$

$$q_{hj} = \int_0^1 \frac{\exp(-\delta \cdot m_j^0 + \mu_h)}{1 + \exp(-\delta \cdot m_j^0 + \mu_h)} f_\delta d\delta \tag{5}$$

$$\delta_{hj} = \mathbb{E}\left[\delta \mid \delta > \frac{m_h^i}{m_j^0}\right] \tag{6}$$

- Under full information (i.e. $\mu, F_\delta \in H_0$), the Principal could simply play the sequence with the highest expected return

$$\mathbb{E}[r_i^{jh}] = q_{jh} \cdot \mu_j + (1 - q_{jh}) q_h^0 \cdot \delta_{jh} \cdot \mu_h \tag{3}$$

where

$$q_J^0 = \mathbb{P}(m_J^i \geq 0) = \mathbb{P}(\epsilon_i > -\mu_J) = \frac{\exp(\mu_J)}{1 + \exp(\mu_J)} \tag{4}$$

$$q_{hj} = \int_0^1 \frac{\exp(-\delta \cdot m_j^0 + \mu_h)}{1 + \exp(-\delta \cdot m_j^0 + \mu_h)} f_\delta d\delta \tag{5}$$

$$\delta_{hj} = \mathbb{E}\left[\delta \mid \delta > \frac{m_h^i}{m_j^0}\right] \tag{6}$$

- So now we have a model (under full information), can we start playing with it?

- What if we just ignore IC? Why not simply $p^\pi = jh \iff \mu_j \geq \mu_h$?

A Naive Policy

- What if we just ignore IC? Why not simply $p^\pi = jh \iff \mu_j \geq \mu_h$?
- **Claim 3.7:** Consider $\mu_h < \mu_j = 1$. There exist two tuples (δ_i, m^0) arbitrarily away under the maximum norm, such that $jh \succ hj$ in period i

A Naive Policy

- What if we just ignore IC? Why not simply $p^\pi = jh \iff \mu_j \geq \mu_h$?
- **Claim 3.7:** Consider $\mu_h < \mu_j = 1$. There exist two tuples (δ_i, m^0) arbitrarily away under the maximum norm, such that $jh \succ hj$ in period i
- **Interpretation:** *Tricking* vs *Conceeding*

UNIVERSITY OF OXFORD

A Naive Policy

- What if we just ignore IC? Why not simply $p^\pi = jh \iff \mu_j \geq \mu_h$?
- **Claim 3.7:** Consider $\mu_h < \mu_j = 1$. There exist two tuples (δ_i, m^0) arbitrarily away under the maximum norm, such that $jh \succ hj$ in period i
- **Interpretation:** *Tricking* vs *Conceeding*
- It follows that (i) naive policies can do poorly and (ii) it is a bandit problem over the orderings not over the firms! However, to learn about the expected return of an order, I will most likely have to induce the agent to play both firms (to suitably learn μ_h and μ_j)

A Naive Policy

- What if we just ignore IC? Why not simply $p^\pi = jh \iff \mu_j \geq \mu_h$?
- **Claim 3.7:** Consider $\mu_h < \mu_j = 1$. There exist two tuples (δ_i, m^0) arbitrarily away under the maximum norm, such that $jh \succ hj$ in period i
- **Interpretation:** *Tricking* vs *Conceeding*
- It follows that (i) naive policies can do poorly and (ii) it is a bandit problem over the orderings not over the firms! However, to learn about the expected return of an order, I will most likely have to induce the agent to play both firms (to suitably learn μ_h and μ_j)

- UCB algorithms suggest creating an UCB for each arm (ordering) of the following kind: $\text{UCB}_a = \widehat{a} + \text{exploration term}$ (where the ET depends negatively on the number of times an arm has been pulled)

Brief Intro to UCB Algorithms

- UCB algorithms suggest creating an UCB for each arm (ordering) of the following kind: $\mathrm{UCB}_a = \widehat{a} + \text{exploration term}$ (where the ET depends negatively on the number of times an arm has been pulled)

- How to adapt UCB logic to the ordering setting? Easy under known δ_i or F_δ AND observed returns. Just compute UCB as usual (featuring the stochastic properties of the problem). Proofs are messy, but results hold!

Brief Intro to UCB Algorithms

- UCB algorithms suggest creating an UCB for each arm (ordering) of the following kind: $\mathrm{UCB}_a = \widehat{a} + \text{exploration term}$ (where the ET depends negatively on the number of times an arm has been pulled)
- How to adapt UCB logic to the ordering setting? Easy under known δ_i or F_δ AND observed returns. Just compute UCB as usual (featuring the stochastic properties of the problem). Proofs are messy, but results hold!
- What if F_δ is unknown? Sadly no hope... The empty job offer strategy?

Brief Intro to UCB Algorithms

- UCB algorithms suggest creating an UCB for each arm (ordering) of the following kind: $\mathrm{UCB}_a = \widehat{a} + \text{exploration term}$ (where the ET depends negatively on the number of times an arm has been pulled)
- How to adapt UCB logic to the ordering setting? Easy under known δ_i or F_δ AND observed returns. Just compute UCB as usual (featuring the stochastic properties of the problem). Proofs are messy, but results hold!
- What if F_δ is unknown? Sadly no hope... The empty job offer strategy?
- What if only agent actions are observed? I think it's doable, but I'm working on it

Discussion Time!

Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg.
Incentivizing exploration.
In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 5–22, 2014.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis.
Bayesian incentive-compatible bandit exploration.
In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582, 2015.

Yiangos Papanastasiou, Kostas Bimpikis, and Nicos Savva.
Crowdsourcing exploration.
Management Science, 64(4):1727–1746, 2018.