Optimal Ordering (and Information) Strategies in Sequential Search Problems

Carlos Gonzalez

University of Oxford - Department of Economics

Student Research Workshop in Micro Theory

10th May

Outline

Outline

Brief Oversight of my Work

- Machine Learning Theory and Micro Theory
- A different language, but similar interests on Learning and Dynamic Games
- **ML:** Result oriented. Heuristic approach to learning. Refined theory developed ex-post. Algorithms are very powerful, but usually a black box
- **MT:** Economically founded learning rules (positive and normative). Predom of Bayesian learning. Deliberately simple and analytically limited
- My research reconciles both notions of learning to understand strategic interactions of economic agents in complicated/empirically relevant settings

Work in Progress

- Establish connections between Machine Learning and Econ Theory
  - Equivalence of Hannan Consistency and Convergence to Best Reply in Repeated Games [Gonzalez, 2023a]

- Expand Economic Theory leveraging ML heuristics and Algorithms
  - A Prior-Free Theory of Adverse Selection and Monopsony Markets [Gonzalez, 2023a]
  - Firm Theory through Knapsack Bandits [Gonzalez, 2023b]
  - **Ordering Strategies in Sequential Search Problems**

- Economic interpretation of ML heuristics
  - Rationalizing Upper Confidence Bound Algorithms [Gonzalez, 2024]

Outline

Motivation

- **Sequential Search** is ubiquitous in Economics: Online shopping, Job search, Medical testing, Investment decisions, Public transportation, etc.

- Order within the sequence (ordering/sequencing) is often **poorly characterized**. Some computational work in OR. Exogenous arrival processes in Economics (labor markets [Pissarides, 2000], strategic experimentation [Keller and Rady, 2010], political economy [Myerson, 2008], firm dynamics [Klette and Kortum, 2004], etc.)

- **Sequencing as PA problem**: Amazon, LinkedIn, medical testing procedures, financial outlets, Google Maps, etc.

This Paper

- We consider a special (but hopefully relevant) case
- Principal is long-lived **social welfare maximizer**
- Agents are **short-lived** expected utility maximizers (myopic). Interest misalignment in repeated games (**exploration vs exploitation**)
- Focus on **incomplete information repeated games** in **restricted feedback** scenarios. Incomplete information meaning that there is partial knowledge on the expected welfare of the elements in the sequence

Outline

Simple model

- **Public Officer** (Principal - she) who wants to match **workers** (Agents - he/they $i = 1, \ldots, N$) and **firms** $J \in \{j, h\}$, where the quality of the firms is unknown

A	B	C	D	E
i arrives	**P**: $p_i, \mathscr{I}_i$	i updates priors Plays the Game	$r_i^{p_i, \mathscr{I}_i}$ and ψ_i	**P** updates strategy

Workers' Problem

$$a^{jh,\mathcal{I}_i} = \begin{cases} T & \text{if } \; m_i^j \geq m_{0i}^h \\ \{C, T\} & \text{if } \; m_i^j < m_{0i}^h \;\; \& \;\; m_i^h \geq 0 \\ \{C, C\} & \text{if } \; m_i^j < m_{0i}^h \;\; \& \;\; m_i^h < 0 \end{cases} \tag{1}$$

- where $m_J^i = \mu^J + \varepsilon_i^J, \varepsilon_i^J \sim M^J$, $m_{0i}^J = \mathbb{E}_{0i}[M^J \mid M^J \geq 0]$

$$r_i^{jh,\mathcal{I}_i} = \mathbb{1}(m_i^j \geq m_{0i}^h) \cdot m_i^j + \mathbb{1}(m_i^j < m_{0i}^h, m_i^h \geq 0) \cdot m_i^h \tag{2}$$

- (Many) **implicit assumptions**: Workers only update priors through $\mathcal{I}_i$, no participation cost (no IR), no discounting, workers are risk neutral, they can't go back, they only get to play once, outside option is normalized to 0, present bias if indifferent

Principal's Problem

- Define policy/algorithm $\pi : H_i(\psi) \rightarrow \{\Delta(P), \Delta(\mathscr{P}(H_i))\}$
  Today $\pi : H_i \rightarrow \Delta(P)$

- Assume wlog $m_{0i}^J = m_0^J$. **Unknown** to the Principal

- Define $\mathbb{E}[r^p] = \mathfrak{r}^p$. Let $\pi^* = H_i \rightarrow p^*$, where $p^* = \arg\max_p \mathfrak{r}^p$

- **Principal's Problem**

$$\begin{aligned} \arg\max_{\pi} \mathbb{E}\left[\sum_{i}^{N} r_i^{\pi(i)}\right] &= \arg\min_{\pi} N \cdot \mathfrak{r}^{p^*} - \mathbb{E}\left[\sum_{i}^{N} r_i^{\pi(i)}\right] \\ &= \arg\min_{\pi} \mathscr{R}_N(\pi) \end{aligned} \tag{3}$$

Limitations of Bayesian Learning

- Under reasonable ψ some learning is possible
- **Optimal learning policy** is prescribed by the solution to the **dynamic optimization** problem: **Bayesian Learning Policy** π^B ($\mathscr{R}(\pi^B) > \mathscr{R}(\pi^*)$)
- π^B is **intractable** and **computationally infeasible** even for small N! What is the value of exploration? (Simple characterization of π^B is an exception/miracle)
- Instead, **near-optimal policies:** (i) Not much worse than π^* (hence π^B), (ii) not trivial $\lim_{N\to\infty} \mathscr{R}_N(\pi)/N \leq C < \infty$ (sublinear regret)

Outline

Main Results

- Under **observed rewards** of the selected firm: (i) Full learning is possible in non-param, (ii) we characterize a **near-optimal policy**

- When only **workers' actions** are observable: (i) **Full learning is possible under param assumptions**, (ii) **innovative near-optimal policy**

- **Additional regret** coming from feedback reduction (given parametric assumptions) **is minimal**

- **Three ordering regimes**: Alignment, Tricking and Conceding

- **PE is not enough hence constant information provision is not enough** for sublinear regret. Non-monotonicity of Information strategies!

Outline

An Heuristic Approach under Full Feedback: UCB

- **Full feedback**: $\psi_i^* = m_i^{J_i} = r_i^{p_i}$ (as opposed to $\psi_i^{**} = m_i^J$)

- **UCB logic**. Optimism in face of uncertainty. Every period select $p^i = \arg\max \mathrm{UCB}_i^p$, where

$$\mathrm{UCB}_i^p = \hat{r}_i^p + B_i^p(I^p(i)) \tag{4}$$

- Explotation term vs Exploration term

- **Proof intuition:** To select p^i at least one of the following must be true
  - $\hat{r}_i^{p^*} + B_i^{p^*} \leq \mathfrak{r}^{p^*}$,
  - $\hat{r}_i^p - B_i^p \geq \mathfrak{r}^p$,
  - $B_i^p \geq 2 \cdot (\mathfrak{r}^{p^*} - \mathfrak{r}^p) = 2\Delta^p$

- For "well-behaved" (subgaussian) rv, and carefully designed B_i^p, the probability of the first two events cannot be very big. Moreover, $B^p(I^p)$ is decreasing in I^p, so third condition can only be true for small i

Proposition 1: Near-Optimality under Full Feedback

Let M^J be σ-subgaussian for all J. Then UCB with $B_i^p = \sqrt{\frac{2\ln f(i)}{I^p(i)}}$, where $f(i) = 1 + i\ln^2(i)$ yields

$$\mathscr{R}_N \leq C_1(\Delta^p + \frac{\ln(N)}{\Delta^p}) \tag{5}$$

UNIVERSITY OF OXFORD

Interpretation and Limitations

- **Learning is possible** under full feedback
  - in a non-parametric setting (subgaussian assumption)
  - for any non-degenerate prior on M^J
  - without knowledge of workers' priors

- UCB is asymptotically **not worse than** π^* (and of course π^B)

- **Nothing too new** from an ML perspective

- **Leaving lots of information in the table:** J^i, a^{p_i}, m_0. It is unclear how much it can buy us in terms of regret (TBC)

Outline

Breaking UCB under Partial Feedback

- UCB is a powerful workehorse, but **relies strongly on feedback**. In many relevant applications, the principal will fail to recover $m_i^{J_i}$ from agents
- The **missing review problem**
- What can be obtained under **weaker feedback** structures like $\psi = a_i^{p_i} \subset \psi^*$?
- $\hat{r}_i^p$ (and its convenient statistical properties) are simply **not available** under ψ

A Surpring Result

Definition 2: Identifiability

Let $Q^o = \{q = \mathbb{E}[\hat{q}] > 0\}$, $\mathfrak{r} = \mathfrak{r}^p(Q^p \subseteq Q^o) = \{\mathfrak{r}^p\}_{p \in P}$ is Q^o-**identified** if $\mathfrak{r}^p = f^p(Q^p)$, with f^p well behaved around Q^p for all p

A Surpring Result

Definition 2: Identifiability

Let $Q^o = \{q = \mathbb{E}[\hat{q}] > 0\}$, $\mathfrak{r} = \mathfrak{r}^p(Q^p \subseteq Q^o) = \{\mathfrak{r}^p\}_{p \in P}$ is Q^o-**identified** if $\mathfrak{r}^p = f^p(Q^p)$, with f^p well behaved around Q^p for all p

Proposition 3: Near Optimality under Partial Feedback

Let $\mathfrak{r}$ be Q^o-identified. Let $k = \max_p \mid Q^p \mid$, then a version of UCB yields

$$\mathcal{R}_N \leq C_2 \cdot 2^k \left(\Delta^p + \frac{\ln(N)}{\Delta^p} \right) \tag{6}$$

Interpretation

- Virtually **no loss in performance** despite the sharp information decrease (is k that bad? In our setting $k = 3$)
- Keeping up with performance comes at the expense of **parametric assumptions**. In particular Q^p must be sufficient to recover $\imath$
- We can recover at most $|Q^o| = 4$ independent parameters. Still **great flexibility**:
  - Reward and Prior locations with known variances
  - Reward location and scale with known priors
  - Virtually any two-parametric well behaved distribution can be identified (TBC)
- **Today**: $M^J \sim Log(0, \sigma)$, with σ known and unknown m_0^J (LKVUP)

Algorithm

Algorithm Cross UCB for LKVUP

Input N, $P = \{jh, hj\}, g(\cdot)$
Initialize $I^p(0) = 0$
while $P \neq \emptyset$
  Select $p^i = P_1$
  if $I^p(i) = 0$ **Update** $\hat{q}^{p_1}(m_0^{p_2}) = \mathbb{1}(a_1^{pi} = T)$, $I^p(i) = 1$
  else continue
  if $a_1^{pi} = C$ **Update** $\hat{q}^{p_2}(0) = \mathbb{1}(a_2^{pi} = T)$, $P = P \setminus p^i$
while $i \leq N$
  Define $B_i^p(\hat{q}^p) = \left\{q : d(\hat{q}, q) \leq \sqrt{\frac{2\ln f(i)}{I^p(i)}}\right\}$, $q_0^p = \arg\max_{q \in B_i^p(\hat{q}^p)} \mathcal{r}^p(q)$
  Let $\tilde{p} = \arg\max_p \mathcal{r}^p(q_0^p)$
    Select $p^i = \tilde{p}$ wp $1 - g(I^p)$, $p^i = \tilde{p}'$ otherwise
  Update $I^{p^i}(i)$, $\hat{q}^p$ (for all p)

q-Space and Cross-Exploration

- We work in q-space of apposed to r-space
  This forces us to be optimistic in k dimensions (and 2^k directions)

- $\mathfrak{z}$ must be Q^o-identified. Under logit,

$$\mathfrak{z}^{jh}(q) = q^j(m_0^h)\ln\left(\frac{q^j(0)}{1-q^j(0)}\cdot\frac{1-q^j(m_0^h)}{q^j(m_0^h)}\cdot(1-q^h(0))\right)-\ln\Big((1-q^j(m_0^h))\cdot(1-q^h(0))\Big)$$

- Interestingly, $\mathfrak{z}^p$ is a function of q^k which can only be inferred when **playing the alternative order**. **Need for cross-exploration!**

- Surprisingly, cross-exploration does **not** entail a **significant performance loss** for fined tuned g

Algorithm Interpretation

- Clever **initialization** to get initial unbiased estimates of $\hat{q}$
- $\arg\max_p \mathrm{UCB}_i^p$ is replaced by best point in a ball
- **Cross exploration** is guaranteed via fine-tuned g.
  $B_i^p \to 0$ only if $I^p(i) \to \infty$ for all p
- Technical note: $\mathfrak{z}^p(q_0^p)$ is very much **not well behaved** when q_0^p is near $\{0, 1\}$.
  Fortunately, small probability of bad behavior provided
  $q \in [1/(1+e), e/(1+e)]$
- Well-behaviour is needed to (i) establish mappings between q and $\mathfrak{z}$ spaces (lipschitz condition), (ii) guarantee a sufficient sample size of $\hat{q}^3$

Outline

Better Always First

- Why not **ignoring workers' priors**? $p^i = jh \iff \mu^j \geq \mu^h$
  Equivalent to Cross-UCB under **alignment**

- This policy is **dominated** under two different sets of priors
  - **Conceding** $\mu^h = \mu^j - \epsilon, m_0^h = 1, m_0^j = 0$. Principal rather let worker pick h safely (in first stage) than letting him move to second stage (**exiting risk**)
  - **Tricking** $\mu^j = 1, \mu^h = 0, m_0^h = 1, m_0^j = 1$. Unconditional higher acceptance probability of second firm. Risk of worker accepting h in period 1 is offset by high transition probability. The **exploratory worker**

Pure Exploitation Policies (PE)

- Doomed to fail in standard bandits, but here...?

- Under ψ and param **PE is not enough**. Cross-exploration is necessary

- Under ψ^* and non-param **PE is not enough** even with known priors! **Intuition:** Let M^j being fully characterized right to $m_0^h > 0$ (but not right to 0), and M^h being fully characterized right to 0. Let $\widehat{\tau}^{jh} > \tau^{jh} > \widehat{\tau}^{hj} + \delta$, but $\tau^{hj} > \tau^{hj}$. No observation of $p_i = p$ can update $\widehat{\tau}_i^{hj}$. Moreover, with high prob $\widehat{\tau}^{jh}$ does not fall below $\widehat{\tau}^{hj}$

- **Conclusion:** Either ψ^* under param, or ψ^{**} under non-param, but **not as bad** as in standard bandits

Information Provision

- **Binding orders is a big restriction**. Let workers pick p based on order-priors
- With **no learning**, this can be a **disaster** (no requirement on priors), What if they could learn?
- **Full-communication $\mathcal{I}_i = H_i$ cannot be optimal**. Same intuition than PE (firm priors and order priors can get stuck with positive prob in suboptimal orders which do not deliver enough information about the contrary order)
- **Communication can ease exploration**. Literature in IC communication in bandit problems [Papanastasiou et al., 2018], [Che and Hörner, 2018], [Mansour et al., 2015]

Information Provision II

- **Challenge 1** (technical): Characterize **optimal information provision** in searching games **without sequencing**

- **Challenge 2** (conceptual): Understand the **interplay between communication and sequencing**. Priors are part of the game!
  - In classic bandits, incentive to induce the correct expected posterior in workers. This remains correct in the limit $m_0^J \to \mu_0^J$
  - **Fact: High posteriors hinder exploration** in sequential search!
  - Implication: Optimal communication strategy might be **non-monotonic**
  - Implication: What is the **competing class**?
  - Implication: If $m_0^J = 0$ can be induced, then Explore-Then-Commit (ETC $\approx$ **PE**) **policies can beat UCB**

Outline

Moving Forward

- Incentive Compatible Sequencing
- Extend analysis to $J > 2$ arms (some initial inefficient results)
- Refine bounds
- Data Application: Forgiven welfare of incorrect sequencing strategies
- Interplay between Information and Sequencing strategies

Bibliography I

- Che, Y.-K. and Hörner, J. (2018).
  Recommender systems as mechanisms for social learning.
  The Quarterly Journal of Economics, 133(2):871–925.

- Gonzalez, C. (2023a).
  Adaptive wage setting: A prior-free theory of adverse selection and monopsony markets.

- Gonzalez, C. (2023b).
  Hiring decisions with knapsack bandits.

- Gonzalez, C. (2024).
  Rationalizing upper confidence bound algorithms.

- Keller, G. and Rady, S. (2010).
  Strategic experimentation with poisson bandits.
  Theoretical Economics, 5(2):275–311.

Bibliography II

- Klette, T. J. and Kortum, S. (2004).
  Innovating firms and aggregate innovation.
  Journal of political economy, 112(5):986–1018.
- Mansour, Y., Slivkins, A., and Syrgkanis, V. (2015).
  Bayesian incentive-compatible bandit exploration.
  In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 565–582.
- Myerson, R. B. (2008).
  The autocrat's credibility problem and foundations of the constitutional state.
  American Political Science Review, 102(1):125–139.
- Papanastasiou, Y., Bimpikis, K., and Savva, N. (2018).
  Crowdsourcing exploration.
  Management Science, 64(4):1727–1746.

Bibliography III

Pissarides, C. A. (2000).
Equilibrium unemployment theory.
MIT press.