A bandit model of bilateral trade with two-sided learning

Mitchell Watt *with Yunus Aybas*

Third Year Seminar

28 April 2021

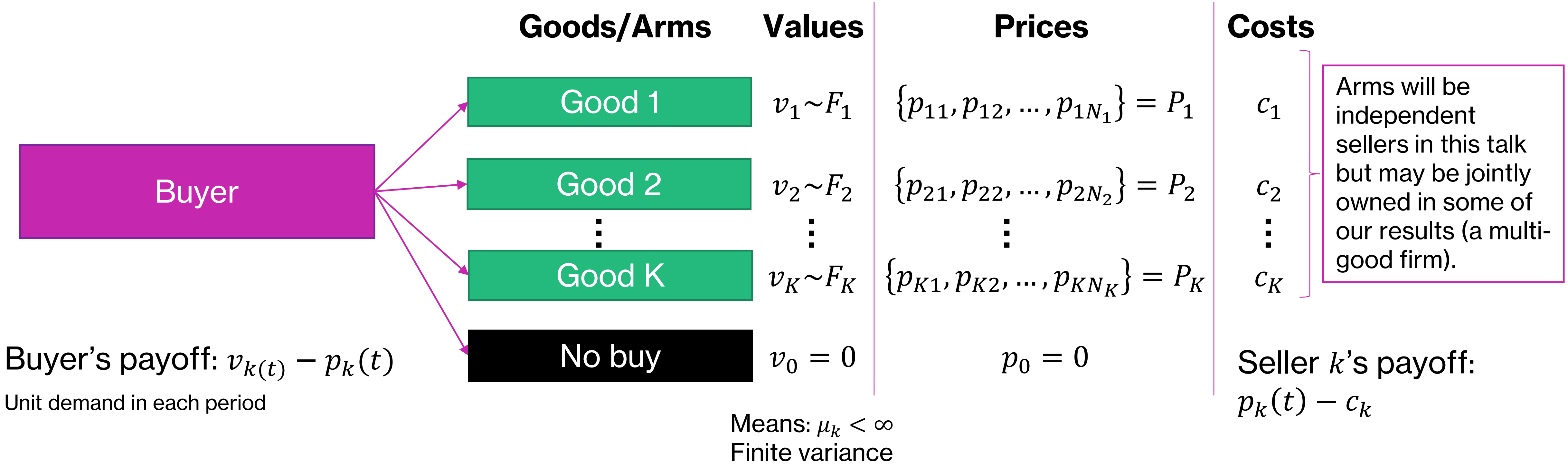
Introduction

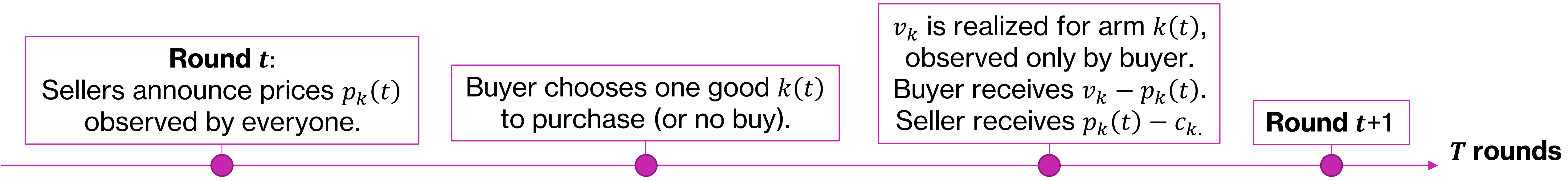
- We study a problem of trade in a setting with one buyer and many sellers with differentiated goods, repeated interaction and two-sided uncertainty about valuations.
  - Buyers and sellers engage in **experimentation** and seek to **learn** value distributions and costs, and **exploit** information learned.
  - Interpret as a 'strategic armed bandit' (as in Braverman et al. 2019).

- CS perspective: we seek **algorithms** for the buyer which provide payoff guarantees for all possible value distributions / cost profiles.
  - **'Negative' result:** classical bandit regret-minimizing algorithms may be exploited by sellers and result in very low payoffs for the buyer.
  - **'Positive' result:** we describe an algorithm for buyers with good payoff guarantees given optimal response by sellers.
- Economics perspective: algorithms act as a commitment device for the buyer

Agenda

1. Introduce model
2. Literature review
    a) Review of multi-armed bandit literature
    b) ‘Strategic-armed’ bandits: Braverman, Mao, Schneider and Weinberg (2019)
3. Non-strategic benchmark
4. Negative results
5. Positive results
6. Conclusion and next steps





Information structures:

- Mostly interested in **two-sided uncertainty:** neither buyer nor seller knows distributions F_i.
- Will also use one-sided uncertainty (seller knows F_i) as a benchmark.
- Will usually assume **all** sellers see which arm the buyer chooses.

Solution concept

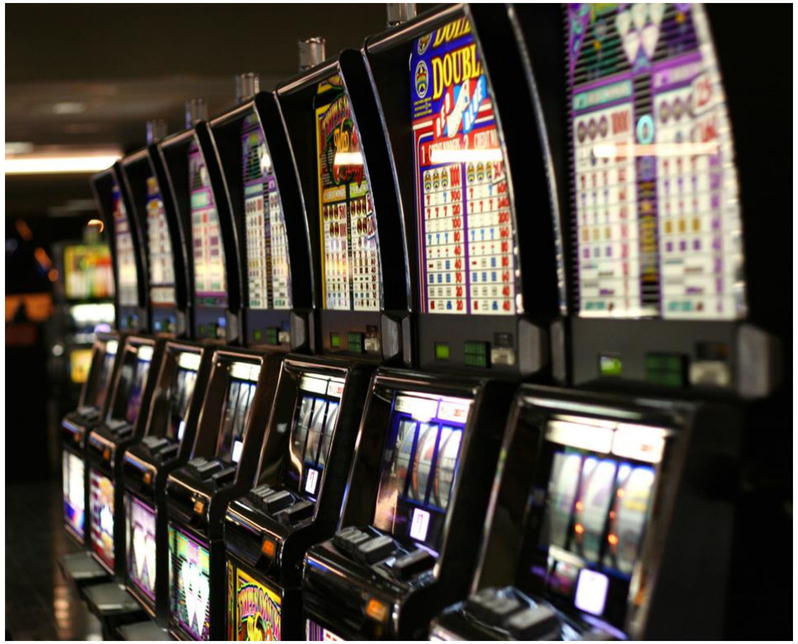
- Typical approach in economics: **Markov perfect equilibrium**
  - Not well-defined under ‘Knightian’ uncertainty about valuation distributions.
  - Difficult! Likely non-unique, complicated.

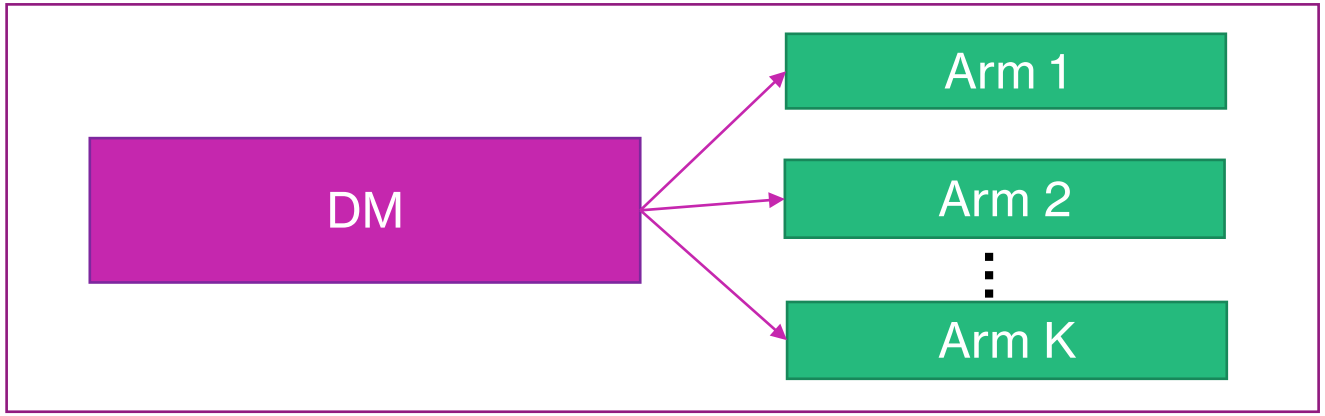
- We take a CS-inspired approach
  - Goal: An **algorithm** for the buyer with good **payoff guarantees,** assuming that sellers are behaving ‘reasonably’.
    - The algorithm should be robust to the distributions $F_1, \ldots, F_K$ and costs $c_1, \ldots, c_K$.
    - The algorithm will usually be random, in which case we seek payoff guarantees with high probability or in expectation.
    - The payoff guarantees might be relative to the maximal possible payoffs (‘regret’).
  - Sellers will be playing dominant strategies / approximate Nash equilibria / minimizing their own regret.

Agenda

1. Introduce model
2. Literature review
   a) Review of multi-armed bandit literature
   b) 'Strategic-armed' bandits: Braverman, Mao, Schneider and Weinberg (2019)
3. Non-strategic benchmark
4. Negative results
5. Positive results
6. Conclusion and next steps

Multi-armed bandits: review





- DM chooses one of K arms each round, over T rounds.
- On choosing arm $k(t)$, DM receives $v_{k(t),t}$.
- DM seeks to maximize $\mathrm{Rev} = \sum_{t=1}^{T} v_{k(t),t}$.
- Alternatively, DM minimizes $\mathrm{Regret} = \max_k \sum_{t=1}^{T} v_{k,t} - \mathrm{Rev}$

Bandit varieties

- **Stochastic bandit:** $v_{k,t} \sim F_k$ iid
- **Bayesian bandit:** learner assumes distribution $v_{k,t} \sim F_k(.|\theta)$ with prior $\pi(\theta)$ over θ.
- **Adversarial bandit:** $v_{k,t}$ is chosen by some (possibly adaptive) adversary to maximize regret.
- **Strategic bandit:** $w_{k,t} \sim F_k$ iid, if chosen arm k chooses $v_{k,t} < w_{k,t}$ to pass on, pocketing the residual for themselves (Braverman, Mao, Schneider and Weinberg 2019)

Bandit algorithms

- Typically, choosing randomly gives $\Theta(T)$ regret.
- We are interested in algorithms that result in sublinear regret.
- **Exploration** vs **exploitation** trade-off

Stochastic Bandit $v_{k,t} \sim F_k$	**UCB (Upper Confidence Bound)** • Choose arm at time t which maximizes $$\text{Sample mean of observed rewards} + \sqrt{\frac{c \log t}{\text{Number of times pulled}}}$$ • Expected regret is $O(\log T)$ with constant depending on $\mu^* - \mu^{(2)}$
Bayesian Bandit $v_{k,t} \sim F_k(\cdot \mid \theta)$ $\theta \sim \pi(\theta)$.	• **Gittins Index:** optimal for $T \to \infty$ • **Probability Matching / Thompson sampling**
Adversarial Bandit	**EXP3** • Given: $\gamma \in [0,1]$. Initialize: $w_k(t) = 1$. • In each round, choose k with probability $p_k = (1-\gamma)\frac{w_k(t)}{\sum w_j(t)} + \frac{\gamma}{K}$. • Update weight of chosen arm as $w_k(t+1) = w_k(t) \exp\left(\gamma \frac{v_{k,t}}{K p_k}\right)$. • Expected regret is $O(\sqrt{TK \log K})$

Strategic-armed bandits

Braverman, Mao, Schneider and Weinberg (2019)

- $w_{k,t} \sim F_k$ is drawn, and arm k (if chosen) determines how much of $w_{k,t}$ to pass on, $v_{k,t} < w_{k,t}$.
- Differences from our setting: existence of outside option, our sellers do not know F_k and learn from buyer behaviour, prices act as a signal to buyer.

Negative result

Given any low-regret algorithm for the adversarial multi-armed bandit problem, there exists an instance of the strategic multi-armed bandit problem and an $o(T)$ –Nash equilibrium for the arms where the principal earns at most $o(T)$ revenue. [As long as K is not too large]

- Arms collude via a market-sharing strategy – they calibrate their actions so that they each get played 1/K of the time, while passing on little utility to the principal.

Positive result

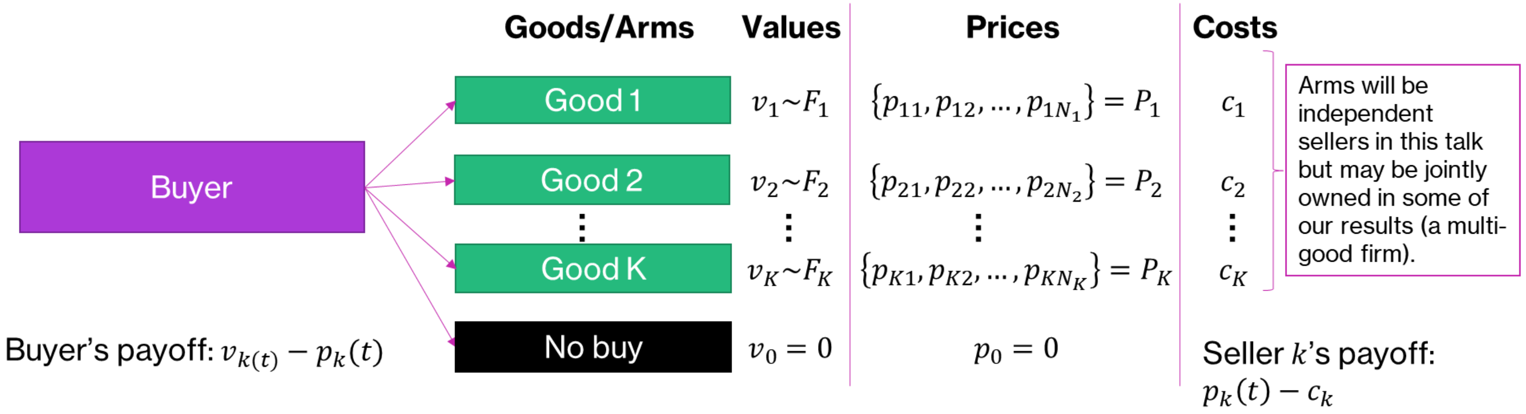
There exists an algorithm for the principal that guarantees revenue at least $\mu^{(2)}T - o(T)$ when the arms are playing according to an $o(T)$-Nash equilibrium. [As long as μ^* and $\mu^{(2)}$ not too different]

- Three phases: 1) arms report truthfully, 2) the most valuable arm pays the principal the second-largest mean each round, 3) arms are compensated for cooperating in stage 1.
- Defections are punished by never being picked again.

Agenda

1. Introduce model
2. Literature review
    a) Review of multi-armed bandit literature
    b) ‘Strategic-armed’ bandits: Braverman, Mao, Schneider and Weinberg (2019)
3. Non-strategic benchmark
4. Negative results
5. Positive results
6. Conclusion and next steps

Pricing bandit regret analysis



- Classic stochastic/adversarial bandit algorithms do not translate directly to this setting, due to prices ('contextual bandit').

- Adapted notion of regret similar to Arora et al. (2012) 'policy regret':
  - If faced with prices $(p_1(t), \dots, p_K(t))$, define **least-regret choice** as

$$k^*(t) = \max_k \sum_{s=1}^{t} v_{k,t} - p_k(t) \overset{\mathbb{E}}{=} \max_k \mu_k - p_k(t)$$

  - **Price-contextual regret** is $\text{PRegret} = \sum_t (v_{k^*(t),t} - p_k(t)) - \sum_t (v_{k(t),t} - p_k(t))$

Non-strategic no-regret algorithm

- Suppose that prices were chosen **randomly**, rather than strategically.

Claim

A modified UCB algorithm results in $O(\log t)$ expected $\mathrm{PRegret}$ for the buyer.

Algorithm

Initialize k-vectors $\hat{Q}(t) = (0,0,\ldots,0)$ and $N(t) = (1,1,\ldots.,1)$.

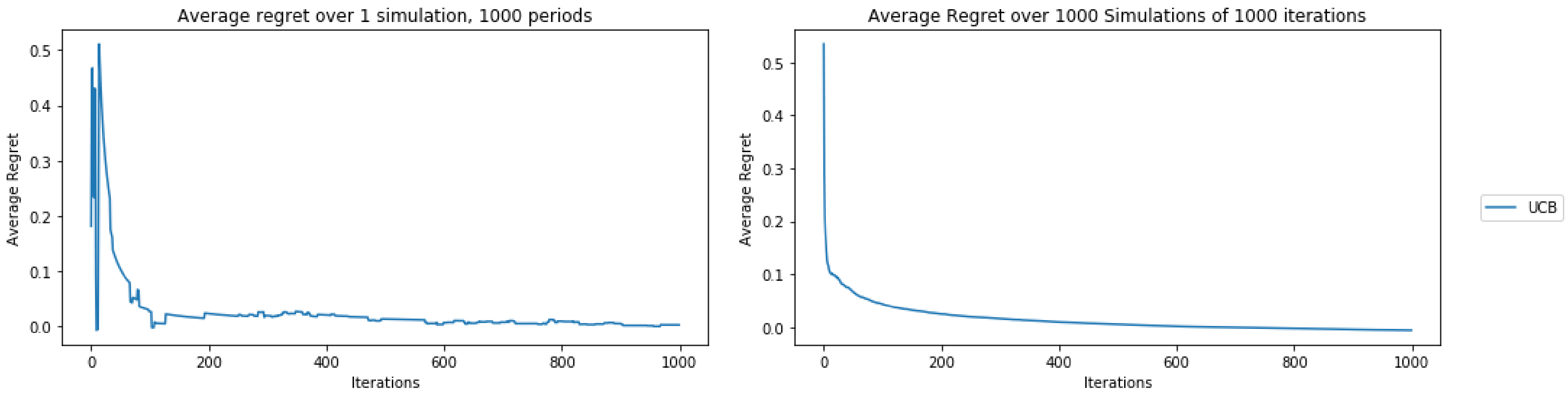
At time t, if $\max_{\mathrm{k}} \hat{Q}_k(t) + \sqrt{\frac{c \log t}{N_k(t)}} - p_k(t) > 0$, choose $k(t)$ as the argmax of this expression.

Otherwise, choose 'not buy'.

Observe utility $v_{k(t),t} - p_{k(t),t}$ and update $\hat{Q}_k(t) = \frac{N_k(t)\widehat{Q_k}(t)+v_{k(t),t}}{N_k(t)+1}$, increment $N_k(t)$ by 1.

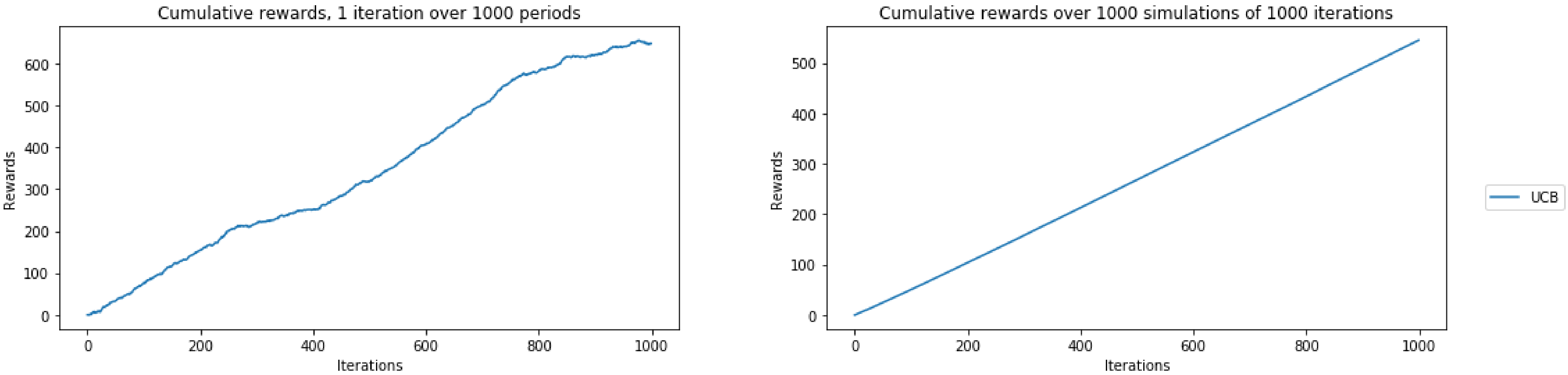
Numerical illustration of modified UCB

- Setting: 3 sellers $k = 1,2,3$ with $F_1 \sim N(1.2,1)$, $F_2 \sim N(1.6,1)$, $F_3 \sim N(1.4,1)$
- Costs zero, pricing strategy: random on $\{0.5, 0.7, 0.9, \ldots, 1.9\}$



Buyer identifies values of arm fairly rapidly, and chooses the best one given the price. Regret is $o(T)$.

Numerical illustration of modified UCB (2)



- Rewards are $\Omega(T)$.

Remarks

- Clearly not the only low-regret algorithm.
- We could also use the usual UCB algorithm or any adversarial algorithm where each $(\mathrm{arm}, \mathrm{price})$ pair is treated as a separate arm, and the agent is presented a subset of such arms in each round

Agenda

1. Introduce model
2. Literature review
   a) Review of multi-armed bandit literature
   b) ‘Strategic-armed’ bandits: Braverman, Mao, Schneider and Weinberg (2019)
3. Non-strategic benchmark
4. Negative results
5. Positive results
6. Conclusion and next steps

‘Negative’ result

Theorem

Suppose *A* is a δ-low $\mathrm{PRegret}$ algorithm for the stochastic pricing bandit problem (or the adversarial pricing bandit problem with $(\mathrm{seller}, \mathrm{price})$ arms), where $\delta < o(T)$.

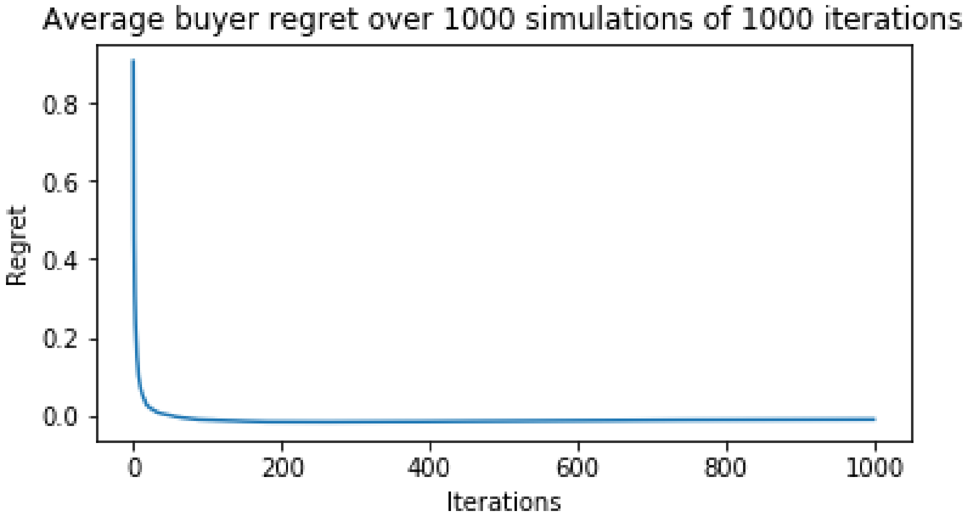
Then in the strategic bandit setting, where the buyer uses algorithm A, there exist distributions F_i and an $o(T)$-approximate Nash equilibrium for the sellers in which the buyer’s expected time-averaged utility per round is small (in particular, equal to the average difference between μ_k and $\max_{p_{kl} \leq \mu_k} p_{kl}$) and the sellers extract almost all surplus.

Intuition: single seller

- Because the buyer is using a low-regret algorithm, they should almost always (i.e. $\Omega(T)$ of the time) accept a price $p < \mu$.
- Therefore, the seller can use a low-regret algorithm to explore the price-space and estimate the demand at various prices.
- If the seller chooses a price just below the mean of F_1, then the buyer will accept this price most of the time, and the expected time-averaged utility for the buyer will be the difference between μ_1 and the price. The payoff for the seller is the price.
- Easily extends to the multi-good monopoly setting.

Illustration: single seller UCB

- Single seller with $F_1 \sim N(1.4, 1)$, zero costs, pricing set $\{0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7, 1.9\}$
- Seller uses UCB algorithm to determine price.
- Buyer is using the pricing-contextual UCB algorithm (similar results if they use $(arm, price)$ EXP3)



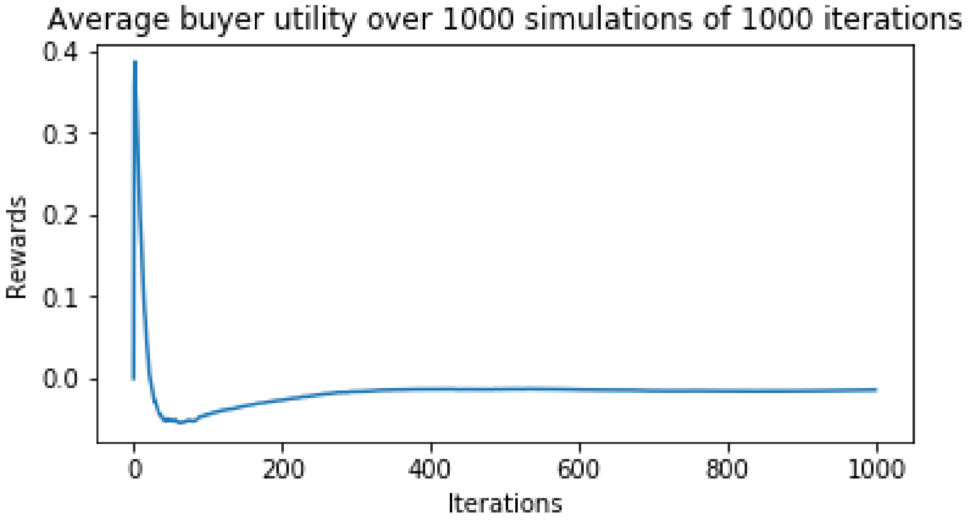
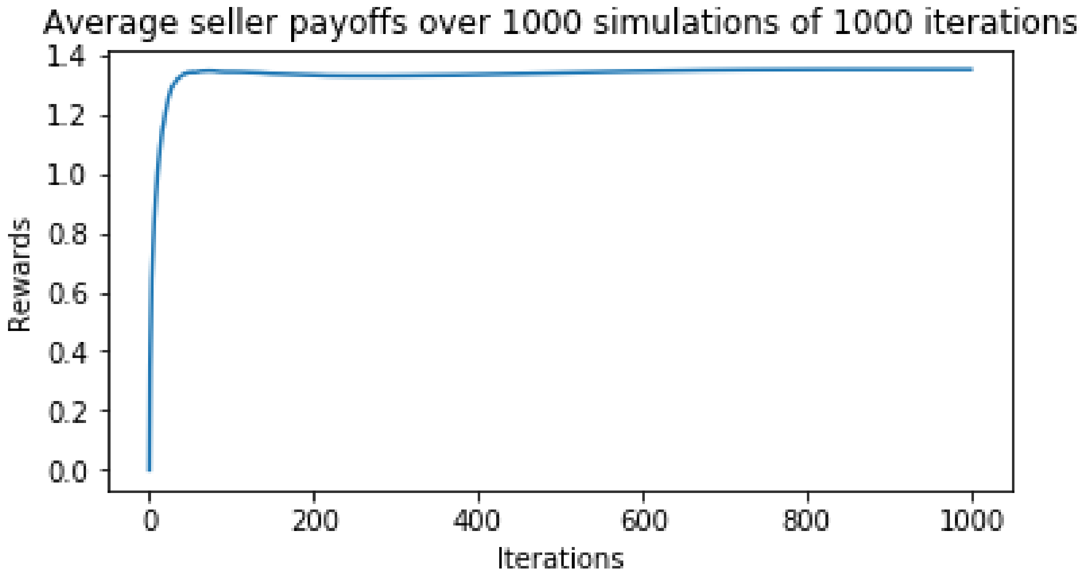
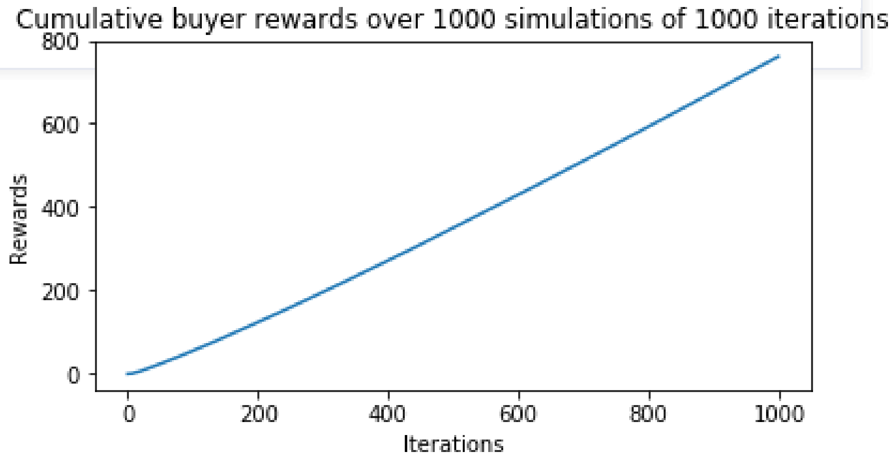


Illustration: single seller UCB (2)

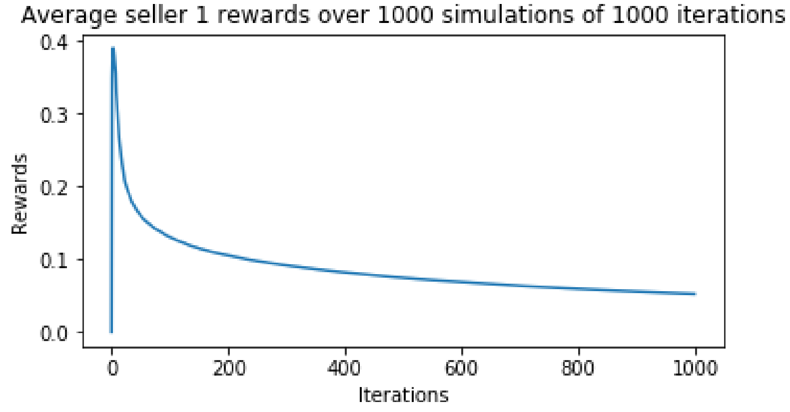


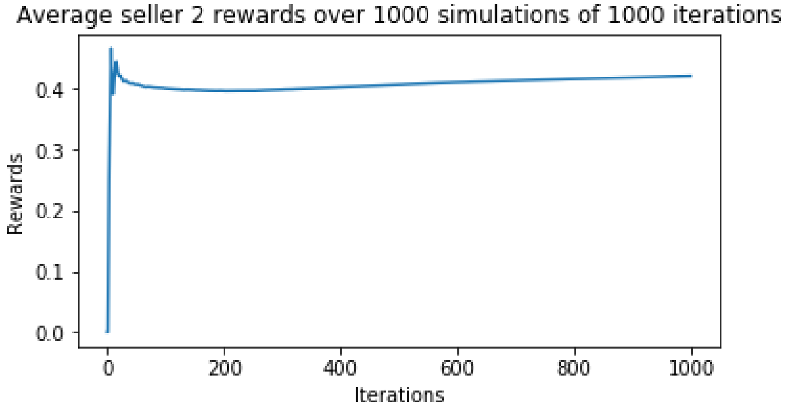
Price	Proportion of time offered by seller	Proportion of time accepted by buyer
0.5	1.35%	85%
0.7	2.27%	95.6%
0.9	4.16%	97.5%
1.1	10.27%	98.9%
1.3	56.20%	99.7%
1.5	21.65%	82.3%
1.7	3.60%	52.2%
1.9	1.31%	25.5%

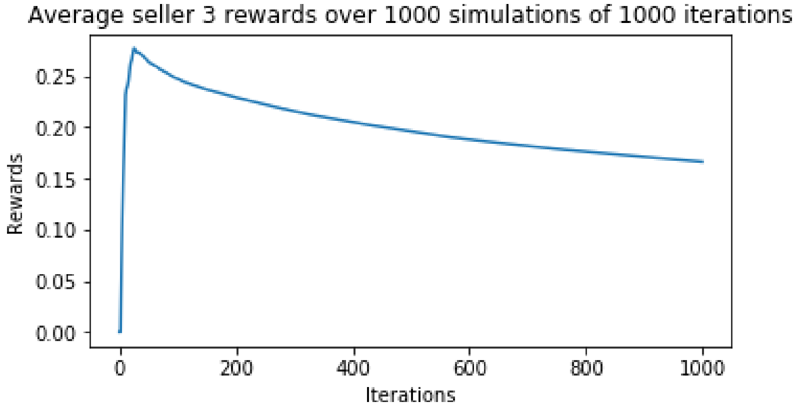
Many sellers: independent learning



- Under independent learning by sellers, no-regret learning by the buyer does quite well.
  - Example: $F_1 \sim N(1.2,1), F_2 \sim N(1.6,1), F_3 \sim N(1.4,1)$
  - High-value seller offers lower prices to be chosen more often.
  - c.f. Calvano et al. (2019)







0.5	0.7	0.9	1.1	1.3	1.5	1.7	1.9
21%	16%	13%	11%	10%	10%	9%	9%

0.5	0.7	0.9	1.1	1.3	1.5	1.7	1.9
30%	33%	13%	8%	6%	4%	3%	3%

0.5	0.7	0.9	1.1	1.3	1.5	1.7	1.9
32%	20%	13%	10%	8%	6%	6%	6%

Many sellers: market-sharing strategy

- If sellers know F_i, then the problem is similar to Braverman et al. (2019).
  - As long as means are not too different, seller can calibrate their actions so that they each get played $1/K$ of the time, while passing on little utility to the principal.

- Without knowledge of F_i, sellers need to estimate **demand** for their goods.
  - Intuitively, because the buyer is using a low-regret strategy, this should not be too difficult for the seller (the buyer need to be choosing optimally $\Omega(t)$ of the time).

Seller joint tâtonnement strategy

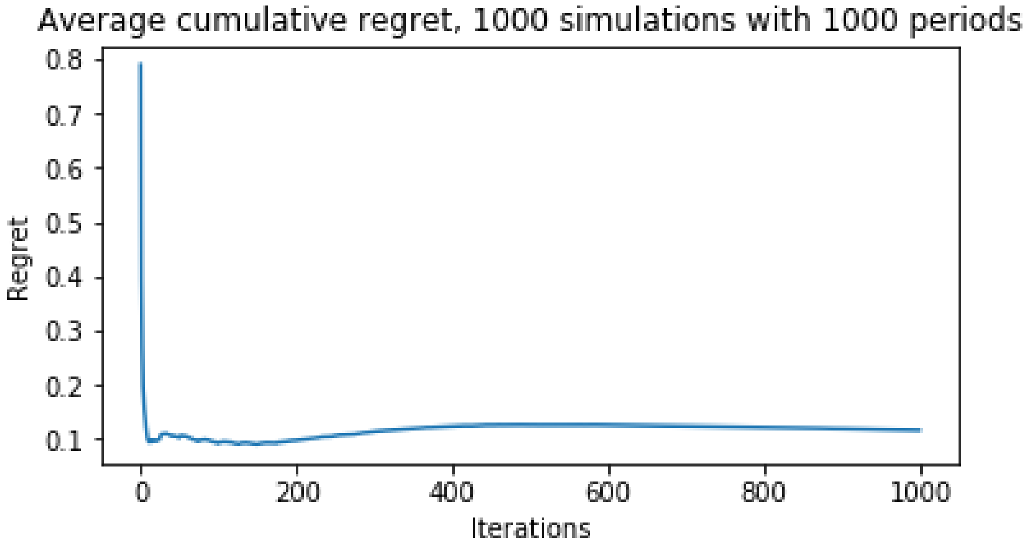
Strategy for seller k

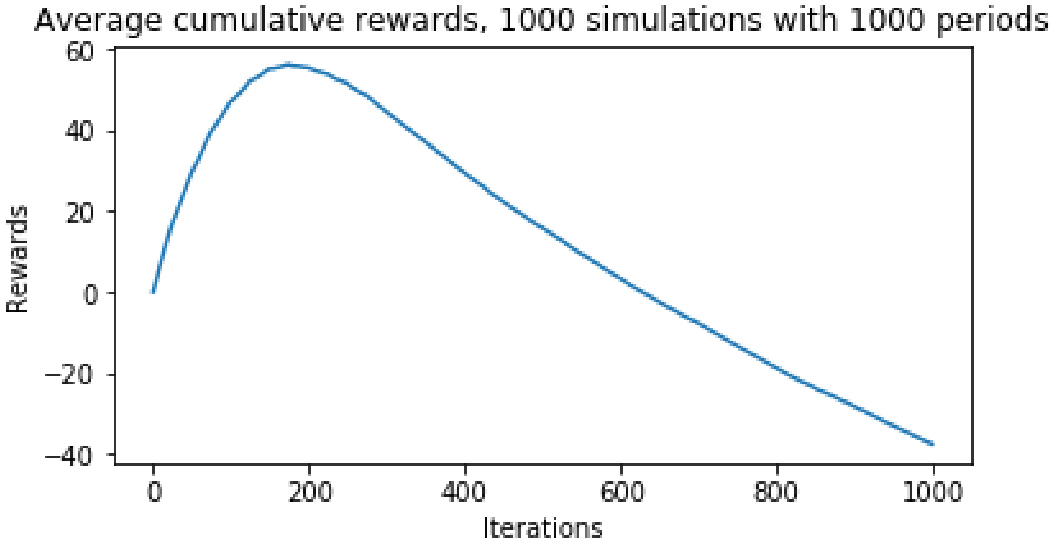
- Given parameters $\tau \sim O(\sqrt{\delta T})$ and β.
- Initialize: each seller selects a random price p_k in P_k.
- Each seller offers price p_k, observes counts N_k of sales by each arm.
- If $t = \tau n$ for $n > 1$, each seller examines sales data for last τ periods:
  - If over last τ periods, $N_k > \frac{\tau}{K} + \beta$, seller k increments price upwards.
  - If over last τ periods, $N_{\text{no buy}} > \frac{\tau}{K} + \beta$, each seller decrements their price downwards.
- If any seller deviates from the strategy, play the lowest price above cost forever.

Claim: if $\frac{\max_{p \in P_k : p \le \mu_k} p}{K} > \max_{p \in P_k : p \le \mu^* - (\mu^{(2)} - p_{min})} p$, all sellers playing the above strategy is an $o(T)$ –Nash equilibrium.

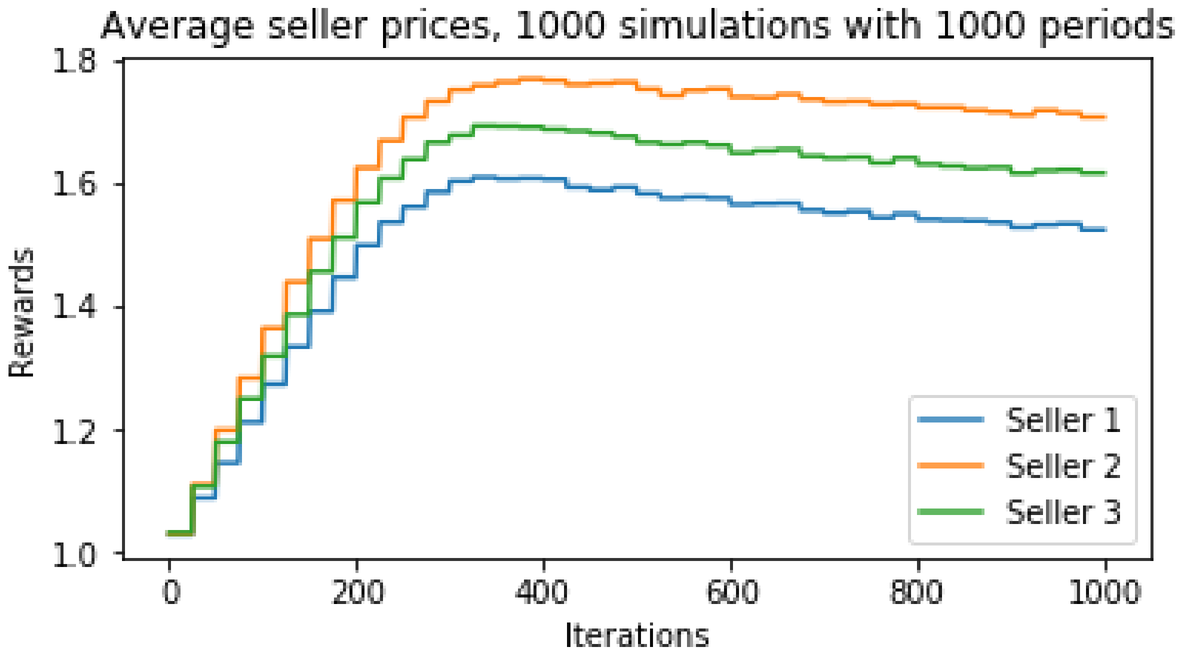
Numerical illustration (1)

- Three sellers $v_1 \sim N(1.3,1)$, $v_2 \sim N(1.5,1)$, $v_3 \sim N(1.4,1)$, zero costs.
- Buyer using modified UCB algorithm, sellers using joint tâtonnement strategy





Numerical illustration (2)



Agenda

1. Introduce model
2. Literature review
   a) Review of multi-armed bandit literature
   b) 'Strategic-armed' bandits: Braverman, Mao, Schneider and Weinberg (2019)
3. Non-strategic benchmark
4. Negative results
5. Positive results
6. Conclusion and next steps

One-sided uncertainty

- **Goal**: to identify an algorithm for the buyer which results in them capturing a large share of the potential gains from trade.
- If sellers know their distribution F_i, then we modify an algorithm from Braverman et al. (2019).

Buyer algorithm

Initialize primitive: confidence level t^*.

1. Observe first price vector and set $p^1 = (p_1^1, \dots, p_K^1)$. Purchase from a random seller in period 1.
2. In subsequent periods:
   a) Let p^t be the price vector offered by sellers. Purchase from remaining seller with largest 'gains from trade' $p_k^1 - p_k^2$, iff they offer a price no larger than $p_k^2 + \left(p_k^{1\,(2)} - p_k^{2\,(2)}\right)$.
   b) Track valuations of purchased goods. If average value $\bar{v}_k$ of goods purchased from seller k ever fails a t –test of the hypothesis that $H_0\colon \mu_k \geq p_k^1$ given confidence level t^*, then never buy from seller k again.
3. In final periods, play each remaining arm sufficiently often that their rewards are *larger* than the expected benefits of misreporting their value in the first period (given t^*).

Seller approximately dominant strategy

- In period 1, choose $p_k^1 = \mu_k$ (or the largest one smaller than it in the price set).
- In subsequent period, choose $p_k^2 = c_k$ (or minimum price above this).
- In subsequent periods in phase 2, seller with largest μ_k plays $\mu_k - \left(p_k^{1\,(2)} - p_k^{2\,(2)}\right)$ (or the nearest price below).
  - e.g. if all costs are zero, this is just $\mu^{(1)} - \mu^{(2)}$.
- In subsequent periods in phase 2, other sellers play c_k (or minimum price above this).
- In phase 3, all players play the maximum price.

Two-sided learning (at least 2 sellers)

- **Goal**: to identify an algorithm for the buyer which results in them capturing a large share of the potential gains from trade.
- Additional challenge of buyer needing to learn values from experimentation, seller needing to infer valuations from buyer behavior.

Buyer algorithm
Initialize primitive: experimentation time $\tau = O(1)$.

1. Buyer commits to purchase from each arm a fixed number τ of times and forms an estimate of the mean value of the arm $\bar{x}_{k,t}$.
2. In subsequent periods:
   a) Let p^t be the price vector offered by sellers. Purchase from remaining seller that offers price which maximizes $\bar{x}_{k,t} - p_k^t$, as long as this value is larger than zero (continuing to track mean value of arms pulled).
   b) If any seller ***ever*** raises their price, never purchase from that seller again.

Seller approximately dominant strategy

- In first $K\tau$ periods, always play highest price.
- In subsequent periods, play highest price.
  - If not chosen in some period, decrement price.

Analogous to a descending auction from the point of view of the buyer

- Somewhat unsatisfying: perhaps the sellers could commit to some strategy of their own to prevent the price from dropping to costs.

Agenda

1. Introduce model
2. Literature review
   a) Review of multi-armed bandit literature
   b) ‘Strategic-armed’ bandits: Braverman, Mao, Schneider and Weinberg (2019)
3. Non-strategic benchmark
4. Negative results
5. Positive results
6. Conclusion and next steps

Conclusion and next steps

Conclusions

- Strategic sellers can take advantage of buyers using bandit regret-minimization algorithms to learn values.
- Buyers can select algorithms to earn a large share of the surplus by exploiting competition between sellers.

Next steps

- Formalize preceding results.
- More to explore in this specific setting: Is there an algorithm for the buyer to capture surplus in single seller case? What about algorithms for the seller? Multiple buyers? A mixed population of strategic and non-strategic buyers? Bayesian strategic bandits?
- More general results on strategic bandits:
    - Other settings, e.g. repeated matching setting of Das and Kamenica (2005)
    - General theorems, characterization of algorithms.
    - Algorithms as an equilibrium selection? Robustness to Knightian uncertainty.