Competing Bandits

-

Abstract

...

Introduction

...

Related Work

Mansour, Slivkins, and Wu 2017, Che and Hörner 2017, Kremer, Mansour, and Perry 2014, Mansour, Slivkins, and Syrgkanis 2015, etc.

Model

...

Simulation Details

Bandit Priors

We study the implications of our model via simulation. For computational tractability we focus our attention on Bernoulli bandits with beta priors. We consider the following bandit "priors" from which the bandit instances faced in simulation are drawn:

1. "HeavyTail": We took the mean rewards to be randomly drawn from Beta($\alpha = 0.6, \beta = 0.6$). With this distribution it was likely to have arms that were at the extremes (close to 1 and close to 0) but also some of the arms with intermediate value means.
2. Needle-in-haystack High - $K-1$ arms with mean 0.50, 1 arm with mean 0.70.

[We can't feasibly include all the priors. I think Heavy Tail is interesting because of the relative reputation plots, but I'm not sure what other one to include that would be meaningful. For now I've just used NIH]

In the appendix we consider the performance of additional priors, but qualitatively the results presented are unchanged.

Learning Algorithms

For our simulations we evaluate the performance of three algorithms, each endowed with a "fake prior" of $Beta(1, 1)$. In general, there are three different classes of learning algorithms and we select one representative algorithm from each class:

1. Adaptive exploration algorithms that engage in purposeful exploration in a "smart" way by adapting to the previous history of results. We consider *ThompsonSampling* (from hereon TS) from this class, which, in a given period, will pull an arm according to the probability that that arm is "optimal" in the sense of having the highest mean reward.
2. Non-adaptive exploration algorithms that engage in purposeful exploration without considering the past results. We consider *Dynamic ϵ-greedy* (from hereon DEG) from this class, which, in a given period, pulls the arms with the highest posterior mean for $1 - \epsilon$ probability and selects a random arm with ϵ probability. For our experiments we keep $\epsilon = 0.05$ fixed.
3. Greedy / Myopic algorithms that engage in no purposeful exploration and take the best short-sighted action. We consider *DynamicGreedy* (from hereon DG) from this class, which, in a given period, pulls the arm with the highest posterior mean.

Run in isolation we expect that the performance of the algorithms (according to cumulative regret) should be
Adaptive Exploration >= Non-adaptive exploration >= Greedy. We verify that, for the chosen algorithms and chosen priors, this is indeed the case (with some caveats, discussed in the next section).

The primary question that we are interested in is under what conditions in competition are the firms in our model incentivized to commit to adaptive exploration algorithms.

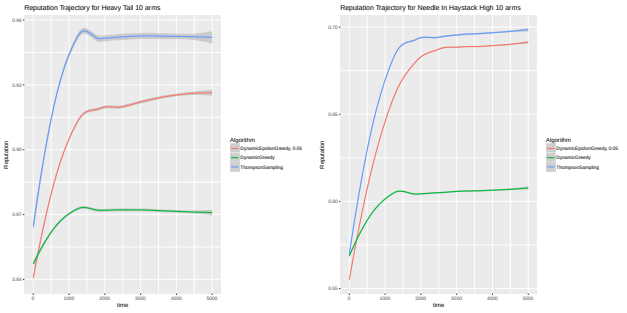
Simulation of Competition Game

Algorithm 1 (FIX TO BE FIGURE) describes the simulation procedure utilized to evaluate the results of the competition game. T is the finite time horizon of the competition game and K is the number of arms in the bandit instance that is considered.

Algorithm 1 Simulation Pseudo-Code

for Each prior p **do**
  Generate true distribution from p
  Generate $T \times K$ realizations for the arms
  for Each agent algorithm *agentalg* **do**
    for Each principal algorithm pair *principalalg*1, *principalalg*2 **do**
      for N simulations **do**
        Give principal 2 X free observations (the agents also get these observations)
        Give the agents k observations from each principal (warm start)
        Run simulation for T periods
      end for
    end for
  end for
end for

Figure 1: Reputation Trajectories in Isolation



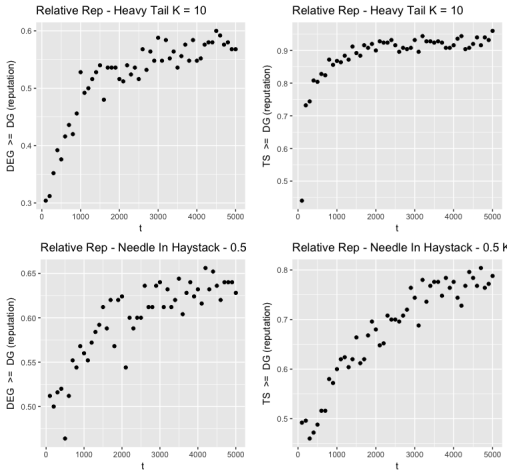
The plots contain the average reputation over 250 runs for a memory size of 100 where, for a given t, we record the reputation of a given algorithm on a given instance and then average this value across all the runs. TODO: get rid of smoothing on plot

Performance in Isolation

We look at the performance of each of our learning algorithms in isolation, meaning that we simply run the learning algorithm for T steps on a realization of the prior instances that we consider and record the realized rewards and the arm that is pulled. In our model the reputation score is an input to the decision rule of the agents and thus this statistic is an important property of the learning algorithm selected by the arm. Figure 1 displays the mean average reputation score of each of the algorithms across the different simulations.

As expected we see that according to the mean reputation plots $TS > DEG > DG$ for sufficiently large values of t. In the competition game we have defined, however, another relevant statistic is the relative comparison of reputation values across algorithms. Figure **??** displays the proportion of rounds where one algorithm has at least as high of a reputation as another. Interestingly we observe that, on the Heavy Tail prior, $DG > DEG$ according to this metric even though $DEG > DG$ according to the mean reputation. The intuition for this is simple. DG will either find the best arm or not. If

Figure 2: Relative Reputation Plots



The plots contain the proportion of simulations over 250 runs for a memory size of 100 where, for a given t, the reputation of one algorithm is at least as high as the reputation of another on the same instance.

it does then since DEG engages in non-adaptive exploration DG will, with ϵ probability, pick an arm uniformly at random which potentially hurts its reputation if the selected arm is not the best arm. DEG eventually converges on the best arm with a sufficiently large time horizon, but DG does not necessarily do so and so when DG does not find the best arm it may potentially have substantially worse reputation which drags down its mean reputation. However, if it is able to find the best arm in sufficiently many simulations then it may have slightly higher reputation than DEG in those simulations and thus do better according to the relative reputation proportion but, due to the simulations where it does not find the best arm it may then have worse mean reputation.

[Would be nice to get at what is the driving factor behind this. Right now, we simply observe it.]

Simultaneous Entry Experiment

In this section we present simulation results of the model discussed in Section where $X = 0$ so that both firms enter at the same time. We want to understand what algorithms the firms are incentivized to commit to in the competition game and so we take the algorithms as given and study the consequences of playing them in the competition game. For this experiment we fix the $HardMax$ agent decision rule. Under this decision rule we expect that one firm will end up taking the market in most of the simulations and so when reasoning about the decision rule we track the effective end of game. [Reword this]

Definition 0.1. *Effective End of Game (EEOG)* - the last round, t in a simulation where there is a difference between the decision of the agent alive at $t-1$ and the

agent alive at t (todo: state this formally using notation of model)

Table 1: Needle In Haystack Prior

Alg 1 vs 2	k = 20	k = 100	k = 200
TS vs DG	**0.66** ±0.05 eeog avg: 330 med: 31	**0.61** ±0.05 eeog avg: 590 med: 61.5	
TS vs DEG	**0.58** ±0.06 eeog avg: 180 med: 1	**0.54** ±0.05 eeog avg: 670 med: 60	
DG vs DEG	**0.41** ±0.05 eeog avg: 650 med: 182	**0.41** ±0.05 eeog avg: 1500 med: 672	

The first line in each cell contains the average market share received by the firm playing Alg 1 (and the market share of Alg 2 is 1 - Alg 1 Market Share) as well as a 95 % confidence band. For example, the cell in the top left indicates that TS gets on average 32% of the market when played against DG. The next line contain the average and median effective end of game for this set of simulations.

Table 2: Heavy Tail Prior

Alg 1 vs 2	k = 20	k = 100	k = 200
TS vs DG	**0.32** ±0.05 eeog avg: 150 med: 0	**0.55** ±0.05 eeog avg: 300 med: 0	**0.73** ±0.04 eeog avg: 1400 med: 226.5
TS vs DEG	**0.34** ±0.05 eeog avg: 36 med: 0	**0.65** ±0.05 eeog avg: 330 med: 0	**0.91** ±0.02 eeog avg: 1100 med: 107.5
DG vs DEG	**0.62** ±0.05 eeog avg: 800 med: 10	**0.6** ±0.05 eeog avg: 1500 med: 866.5	**0.62** ±0.05 eeog avg: 1900 med: 1627.5

Tables 1 and 2 display the results. Under Heavy Tail we see that for small warm start of $k = 20$, DG gives the firm more of the market than TS and DEG. As we increase the warm start to $k = 100, 200$ we observe that TS does better than DG and DEG. Note that, in particular for the simulations involving TS, the effective end of game had a median of 0. This means that the result of the simulation was *completely determined by the warm start period*. Thus, when reasoning about what algorithm would win in the competition game under the HardMax decision rule we need to think about the performance of the algorithm on k samples, which may be small. For low warm start, it is thus not necessarily the case that TS is the best algorithm. The intuition for this is that TS needs to engage in pure exploration early on in order to identify the best arm. A consequence of exploration in our model is that it may lead to a reputational cost relative to the greedy algorithm since it involves experimenting with potentially sub-optimal arms relative to the current belief of the best arm. However, a sufficiently large warm start allows TS to incur this reputational cost without the informational gain.

Another interesting result is that DG still does better than DEG even with a large warm start under the Heavy Tail prior. In general, we observe that even under other agent response functions besides HardMax that DG does better than DEG. Thus, even though the mean reputation plot in Figure 1 shows that DEG does better than DG in terms of average mean reputation, this does not lead to DEG doing better under competition. Rather, the relative reputation plot shown in Figure 2 is more predictive in this case of what algorithm wins under competition.

Looking at the results for Needle In a Haystack we observe that TS always does better than DEG, DG. T

[TODO: is there something more interesting / general that we can discuss here besides this? Why does TS win? Learning is easy? Something completely unrelated to learning hardness?]

[TODO: Can we say anything general about this using the "learning hardness" metric that we have started looking at?]

Incumbent Experiment

We now re-run the same simulations but let $X = 200$ so that firm 2 has an "incumbency" advantage. Recall that X controls the number of rounds the firm is in the market by itself before the other firm enters and so in these rounds agents arrive and must choose firm 2. After X rounds, firm 1 enters the market and each agent that enters in subsequent periods must select between the two firms.

Table 3: Incumbent Experiment for Heavy Tail

Entrant Algorithm \ Incumbent Algorithm	TS	DEG	DG
TS	**0.0067** ± 0.0092 Variance: 0.007 ES: 100 %	**0.023** ± 0.017 Variance: 0.02 ES: 99 %	**0.064** ± 0.027 Variance: 0.05 ES: 97 %
DEG	**0.024** ± 0.015 Variance: 0.02 ES: 98 %	**0.13** ± 0.034 Variance: 0.09 ES: 86 %	**0.14** ± 0.036 Variance: 0.1 ES: 89 %
DG	**0.063** ± 0.024 Variance: 0.04 ES: 93 %	**0.19** ± 0.041 Variance: 0.1 ES: 91 %	**0.15** ± 0.032 Variance: 0.08 ES: 77 %

TODO: fix table names and general table alignment. The first line in each cell contains the average market share for the entrant over $N = 250$ simulations as well as a 95% confidence interval. The second line contains the sample variance of the observed market shares and the third line contains the fraction of simulations that ended up with one principal getting $> 90\%$ of the market. Note that smaller values in the table are better for the incumbent. Market shares are calculated as the fraction of users selecting a particular firm *after* the entrant has already entered (i.e. the free rounds to firm 2 do not count towards the share)

Table **??** shows the results of the simulations for the Heavy Tail prior. TS is a dominant strategy for the incumbent when X is sufficiently high. In the appendix we include additional simulations where $X = 50$ and observe that TS is not always a dominant strategy in this case. These results are robust across the different priors we considered as well as across agent response models. Namely, for sufficiently large X, TS is a dominant strategy for the incumbent.

Whereas with simultaneous entry TS was a worse strategy than DG (for small warm start), by allowing one firm to have a first mover advantage and be a monopolist in the market in the early rounds we incentivize the monopolist to play the "best" algorithm. The intuition for this is that competition forces the firms to worry about their reputation which dissuades them from committing to algorithms that involve pure exploration. In some sense one can view allowing one firm to temporarily be a monopolist as temporarily relaxing the "incentive" component of exploration, exploitation, and incentives so that the incumbent firm only faces the classic tradeoff between exploration and exploitation. The incumbent only needs to worry about her reputation after X periods when the entrant comes into the market and again needs to worry about incentivizing agents to select them over their competition. As a result, the incumbent is incentivized to commit to an algorithm that does exploration in the early rounds since she no longer suffers the same reputational cost that she would suffer under competition as long as the X is sufficiently large that she can begin to recover the reputational costs of exploration.

[This section merits much more explanation]

Data and Reputation as Barriers to Entry

As the simulation results of the incumbent experiment in Table **??** and in the appendix show, the incumbent usually ends up with a substantial share of the market over the entrant. The rounds where the incumbent was a monopolist provided the incumbent with free data for learning (data advantage) as well as a chance to establish a reputational advantage over a potential entrant (reputational advantage). Which plays a bigger role in preventing the entrant from being able to establish market share? We run two additional experiments, modifying the previous incumbent experiment so that in one set of simulations the reputation of the incumbent is artificially erased and another in which the information gained by the incumbent is artificially erased so that the posterior is reset to the prior.

Tables 4 and 5 display the results of these experiments. Even with the data advantage or reputation advantage alone the incumbent ends up getting a substantial portion of the market, regardless of the algorithm played. However, fixing the same algorithm played by the incumbent and the entrant, the market share for the incumbent is always higher when the incumbent retains her reputation instead of her information. The intuition for this is that the reputational advantage protects the incumbent from "bad" decisions or bad luck on rewards that adversely affect reputation and thus, especially in the earlier rounds, agents in the earlier rounds will still choose the incumbent which will allow the incumbent to regain the information they lost. However, with reputation re-initialized the better information does not protect the incumbent from getting unlucky with her rewards. Looking at the variability in the shares in the two experiments indicates that not only does the incumbent get a smaller market share on average in the reputation erased experiment, but the variability in the shares is also higher.

Though the setup is purely experimental, it is nonetheless interesting to look at if TS still remains as a dominant strategy for the incumbent. In the information erased experiment, TS is still a dominant strategy while in the reputation erased experiment it is not a dominant strategy. The intuition for this is that in the reputation erased experiment, the incumbent still has the data that she accumulated during her time as a monopolist and as a result, there is less value to an algorithm that engages in pure exploration as she wants to rebuild her reputation and thus might be better off exploiting the information she has in order to rebuild reputation instead of engaging in exploration and potentially damage her reputation. However, in the information erased experiment, she has a reputational cushion that allows for exploration so that she can regain the information that

Table 4: Reputation Erased Experiment

	TS	DEG	DG
TS	**0.18** ± 0.042 Var: 0.1 ES: 97 %	**0.17** ± 0.042 Var: 0.1 ES: 98 %	**0.23** ± 0.047 Var: 0.2 ES: 98 %
DEG	**0.23** ± 0.045 Var: 0.2 ES: 92 %	**0.32** ± 0.049 Var: 0.2 ES: 86 %	**0.25** ± 0.046 Var: 0.2 ES: 90 %
DG	**0.28** ± 0.047 Var: 0.2 ES: 87 %	**0.33** ± 0.051 Var: 0.2 ES: 91 %	**0.3** ± 0.047 Var: 0.2 ES: 83 %

Table 5: Information Erased Experiment

	TS	DEG	DG
TS	**0.059** ± 0.025 Var: 0.05 ES: 97 %	**0.11** ± 0.033 Var: 0.09 ES: 94 %	**0.15** ± 0.038 Var: 0.1 ES: 92 %
DEG	**0.11** ± 0.031 Var: 0.08 ES: 89 %	**0.19** ± 0.038 Var: 0.1 ES: 82 %	**0.18** ± 0.038 Var: 0.1 ES: 82 %
DG	**0.17** ± 0.038 Var: 0.1 ES: 87 %	**0.22** ± 0.042 Var: 0.1 ES: 82 %	**0.26** ± 0.043 Var: 0.1 ES: 77 %

she lost and thus TS remains a dominant strategy.

In the appendix we include the results of the same experiment but with different agent response models. In these experiments we get that information and reputation serve as substitutes for each other

[TODO: leave this, expound on this, or throw it out?]

Thus, we see that both information and reputation advantages alone can serve as substantial barriers to entry in our model. However, under the HardMax agent response model, we see that reputation serves as more of a barrier to entry compared to information.

Varying the agent response model

Thus far we have largely presented simulation results utilizing the deterministic HardMax rule. This is a natural decision rule where agents simply use the reputation score as a strict decision-making rule. In this section we consider the implications of adding randomness to the agent rule. What are the implications on what algorithm the firms are incentivized to play? How does the distribution of market share change?

[Discuss that, under random agent models, firms that play better learning algorithms do better in the long-run since they get enough free agents to move to the part of the relative reputation curve where better algorithms win]

[Discuss that, under HMR, market share variance is lower and see less extreme market shares.]

[Discuss that, under SM, market share variance is very low and market shares are around 50/50].

Conclusion

...

References

Che, Y.-K., and Hörner, J. 2017. Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics* 133(2):871–925.

Kremer, I.; Mansour, Y.; and Perry, M. 2014. Implementing the "wisdom of the crowd". *Journal of Political Economy* 122(5):988–1012.

Mansour, Y.; Slivkins, A.; and Syrgkanis, V. 2015. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 565–582. ACM.

Mansour, Y.; Slivkins, A.; and Wu, Z. S. 2017. Competing bandits: Learning under competition. *arXiv preprint arXiv:1702.08533*.