Two-Sided Bandits and the Dating Market

Sanmay Das
Center for Biological and Computational Learning
and Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology
Cambridge, MA 02139
sanmay@mit.edu

Emir Kamenica
Department of Economics
Harvard University
Cambridge, MA 02138
kamenica@fas.harvard.edu

Abstract

We study the decision problems facing agents in repeated matching environments with learning, or *two-sided bandit problems*, and examine the dating market, in which men and women repeatedly go out on dates and learn about each other, as an example. We consider three natural matching mechanisms and empirically examine properties of these mechanisms, focusing on the asymptotic stability of the resulting matchings when the agents use a simple learning rule coupled with an ϵ-greedy exploration policy. Matchings tend to be more stable when agents are patient in two different ways — if they are more likely to explore early or if they are more optimistic. However, the two forms of patience do not interact well in terms of increasing the probability of stable outcomes. We also define a notion of regret for the two-sided problem and study the distribution of regrets under the different matching mechanisms.

1 Introduction

This paper analyzes the learning and decision problems of agents in a model of one-to-one two-sided matching, focusing on the role of the matching mechanism and the exploration-exploitation tradeoff. We consider a repeated game in which agents gain an uncertain payoff from being matched with a particular person on the other side of the market in each time period. A natural example of such a situation is the dating market, in which men and women repeatedly go out on dates and learn about each other. Another example is a spot labor market, in which employers and employees are matched for particular job contracts. A matching mechanism is used to pair the agents. For example, we can consider a mechanism in which all the women decide which man to "ask out," and then each man selects a woman from his set of offers, with the rejected women left unmatched for that period.

Standard models of matching in economics [Roth and Sotomayor, 1990] almost always assume that each agent knows his or her preferences over the individuals on the other side of the market. This assumption is too restrictive for many markets, including the market for romantic partnerships. Our model is driven by the need to relax this assumption. The existing literature on two-sided search with nontransferable utility (for example [Burdett and Wright, 1998]) assumes matching is exogenous and random. Our problem is more deeply related to bandit problems [Berry and Fristedt, 1985], in which an agent must choose which arm of an n-armed bandit to pull in order to maximize long-term expected reward, taking into account the tradeoff between *exploring*, that is learning more about the reward distribution for each arm, and *exploiting*, pulling the arm with the maximum expected reward. However, in our model the arms themselves have agency — they can decide whether to be pulled or not, or whom to be pulled by, and they themselves receive rewards based on who the puller is. This motivates our formulation of the problem as a "two-sided" bandit problem.

In principle, we would like to examine the equilibrium behavior of agents in two-sided bandit problems. However, perfect Bayesian equilibria of the game we formulate are prohibitively hard to compute. Because of this difficulty, the approach we use is closely related to the theory of learning in games [Fudenberg and Levine, 1998], which considers more generally how individual learning rules affect outcomes in games and whether agents reach static equilibria.

In this paper we formally define the two-sided bandit problem and describe three important matching mechanisms. We define regret as the difference between the actual reward received and the reward under the stable matching, i.e. a matching such that there is no pair that would rather be with each other than with whom they are matched. We experimentally analyze the asymptotic stability and regret properties when agents use ϵ-greedy learning algorithms adapted to the different matching mechanisms.

The Gale-Shapley mechanism [Gale and Shapley, 1962], which yields stability when information is complete and preferences are truthfully revealed, converges quickly to stable matchings, whereas mechanisms that are more realistic for the dating example, in which women make single offers to the men, do not always converge to stable matchings. Asymptotically stable matches are more likely when agents explore more early on. They are also more likely when agents are optimistic (again, early on) — that is, they assume a higher probability of their offer being accepted or an offer being made to them than is justified by the past empirical frequencies of these events. However, increased optimism does not

interact well with increased exploration in terms of stability; the probability of stability is actually higher for lower exploration probabilities when optimism is greater.

2 The Model

There are M men and W women, who interact for T time periods. v_{ij}^m is the value of woman j to man i, and v_{ij}^w is the value of man j to woman i. These values are constant through time. In each period, men and women are matched to each other through a *matching mechanism*. A matching is a pairing between men and women in which each woman is paired with one or zero men and each man is paired with one or zero women. Formally, a matching mechanism is a mapping from agents' actions to a matching. If man i is matched with woman j in period t, he receives $v_{ij}^m + \epsilon_{ijt}^m$, and she receives $v_{ji}^w + \epsilon_{jit}^w$. If unmatched, individual i receives some constant value K_i.

For our empirical analysis we put some structure on the reward processes and the matching mechanism. First, we make the strong assumption of sex-wide homogeneity of preferences. That is, every man is equally "good" for each woman and vice versa — there are no idiosyncratic preferences and there are no couples who "get along" better than others. Formally, $v_{ij}^m = v_j^m \ \forall i$ and $v_{ij}^w = v_j^w \ \forall i$. We also assume that people dislike being single: $\forall i, K_i \ll \min_j v_{ij}^z$ $\forall i, z \in \{m, w\}$, and that the noise terms ϵ are independently and identically distributed.[1] Extensions to more general preferences are straightforward.[2]

We consider three matching mechanisms. Without loss of generality, we assume that women always ask men out.

Gale-Shapley matching Each agent submits a list of preferences and a centralized matching procedure produces a matching based on these lists. The Gale-Shapley algorithm [Gale and Shapley, 1962] guarantees a matching that is stable under the submitted preferences. The man-optimal variant yields the stable matching that is optimal for men, and the woman-optimal variant the stable matching that is optimal for women. We use the woman-optimal variant.

Simultaneous offers Each woman independently chooses one man to make an offer to. Each man selects one of the offers he receives. Women who are rejected are unmatched for the period, as are the men who receive no offers.

Sequential offers Each woman independently chooses one man to make an offer to. The offers are randomly ordered and men must decide on these "exploding" offers without knowing what other offers are coming. Men see all the offers they receive, including ones that arrive after they accept. If an offer is rejected the woman making the offer is unmatched in that period. A man is unmatched if he rejects all offers he receives.

[1] Another plausible and interesting structure to consider is pairwise homogeneity of preferences with $v_{ij}^m = v_{ji}^w \ \forall i, j$.

[2] There is always a unique stable matching under the assumed preference structure. With multiple stable matches, we would need to use a different notion of regret, as discussed later.

Intuitively, it is useful to think of the simultaneous choice mechanism as capturing a situation in which women ask men out over e-mail and each man can review all his offers before making a decision, while the sequential choice mechanism captures the situation where women ask men out over the telephone. We are particularly interested in these two matching mechanisms because they are more plausible descriptions of reality than a centralized matchmaker and do not require agents to reveal their preferences to a third party.

3 The Decision and Learning Problems

We first describe the decision problems agents face at each time step if they want to optimize their myopic reward in that time step. After this we discuss the exploration-exploitation issues under the different matching mechanisms and describe specific forms of the ϵ-greedy algorithm.

Let $Q_{ij}^{\{m,w\}}$ denote man (woman) i's estimate of the value of going out with woman (man) j, p_{ij}^w denote woman i's estimate of the probability that man j will go out with her if she asks him out and p_{ij}^m denote man i's estimate of the probability that woman j will ask him out under the sequential choice mechanism.

3.1 Women's Decision Problem

Under Gale-Shapley matching, women's action space is the set of rankings of men. Under the other two mechanisms, a woman chooses which man to make an offer to. She must base her decision on any prior beliefs and the history of rewards she has received in the past. She has to take into account both the expected value of going on a date with each man and (for the non-Gale-Shapley mechanisms) the probability that he will accept her offer.

Under the woman-optimal variant of the Gale-Shapley mechanism, the dominant myopic strategy, and thus the greedy action, is for woman i to rank the men according to the expected value of going out with each of them, Q_{ij}^w. For the other two mechanisms, the greedy action is to ask out man $j = \arg\max_j (p_{ij}^w Q_{ij}^w)$.

3.2 Arms With Agency: Men's Decision Problem

The action space of men, the arms of the bandit, may be constrained by women's actions. The decision problem faced by a man depends on the matching mechanism used. Under the woman-optimal Gale-Shapley mechanism, men may have an incentive to misrepresent their preferences, but since the sex-wide homogeneity of preferences ensures a unique stable matching [Roth and Sotomayor, 1990], this is less likely to be a problem.[3] So, the greedy action for man i under Gale-Shapley is to rank women based on their Q_{ij}'s.

With the simultaneous choice mechanism, in each time period a man receives a list of women who have made him an offer. He must decide which one to accept. This is a bandit problem with a different subset of the arms available at each time period. The greedy action is to accept the woman $j = \arg\max_j Q_{ij}^m$.

[3] Specifically, if the *submitted* rankings satisfy sex-wide homogeneity, man- and woman-optimal algorithms yield the same matching so truthtelling is the dominant myopic strategy for men.

Under the sequential choice mechanism, a man might receive multiple offers within a time period, and each time he receives an offer he has to decide immediately whether to accept or reject it, and he may not renege on an accepted offer. The information set he has is the list of women who have asked him out so far. For each woman who has not asked him out, it could either be that she chose not to make him an offer, or that her turn in the ordering has not arrived yet. We can formulate the man's value function heuristically. Let i be the index of the man, let S be the set of women who have asked him out so far, and let h be the woman currently asking him out ($h \in S$).

$$V(S,h) = \max\{Q^m_{ih}, \sum_{k \notin S} \Pr(k \text{ next woman to ask } i \text{ out}) V(S \cup \{k\}, k)\}$$

The base cases are $V(\mathcal{W}, h) = Q^w_{ih}$ where $\mathcal{W}$ is the set of all women. The greedy action is to accept an offer when

$$Q^m_{ih} > \sum_{k \notin S} \Pr(k \text{ next woman to ask } i \text{ out}) V(S \cup \{k\}, k)$$

The relevant probabilities are:

$$\Pr(k \text{ next woman to ask } i \text{ out}) = \sum_{T \in \text{Perm}(S')} \left[\frac{1}{|S'|} \left(\prod_{j \text{ preceding } k \text{ in } T} (1 - p^m_{ij}) \right) p^m_{ik} \right]$$

where $S' = \mathcal{W} \setminus S$. This is a variant of the classic secretary problem [Gilbert and Mosteller, 1966]. We are not sure at the moment if this particular form can be simplified to yield a closed form solution for the value or decision function.

3.3 The Exploration-Exploitation Tradeoff

Women and men both have to consider the exploration-exploitation tradeoff (summarized in [Sutton and Barto, 1998]). *Exploitation* means maximizing expected reward in the current period (also called the greedy choice), and is solved as above. *Exploration* happens when an agent does not select the greedy action, but instead selects an action that has lower expected value in the current period in order to learn more and increase future rewards.

The one-sided version of the exploration-exploitation problem is central to n-armed bandit problems [Berry and Fristedt, 1985; Gittins and Jones, 1974, *inter alia*]. An n-armed bandit is defined by random variables $X_{i,t}$ where $1 \leq i \leq n$ is the index of the arm of the bandit, and $X_{i,t}$ specifies the payoff received from pulling arm i at time t. The distribution of some or all of the $X_{i,t}$ is unknown so there is value to exploring. The agent pulls the arms sequentially and wishes to maximize the discounted sum of payoffs. In our model, if there is a single woman and n men, the woman faces a standard n-armed bandit problem.

One of the simplest techniques used for bandit problems is the so-called ϵ-greedy algorithm. This algorithm selects the arm with highest expected value with probability $1 - \epsilon$ and otherwise selects a random arm. Although simple, the algorithm is very successful in most empirical problems, and we therefore use it in our experiments. We have also experimented with alternatives like softmax-action selection with Boltzmann distributions [Sutton and Barto, 1998; Luce, 1959] and the Exp3 algorithm [Auer *et al.*, 2002]. These do not improve upon the empirical performance of ϵ-greedy in our simulations.

Under each matching mechanism the exploratory action is to randomly select an action, other than the greedy one, from the available action space. Since the value of exploration decreases as learning progresses, we let ϵ decay exponentially over time which also ensures that the matchings converge.

At this stage we cannot solve for the perfect Bayesian equilibrium set. We believe, however, that if the agents are sufficiently patient and the horizon is sufficiently long, the matchings will converge to stability on the equilibrium path. Solving for the equilibrium set would enable us to explicitly characterize the differences between equilibrium behavior and behavior under the ϵ-greedy algorithm.

The two-sided nature of the learning problem leads to nonstationarities. Under the sequential and simultaneous mechanisms, women need to consider the reward of *asking out* a particular man, not the reward of going out with him. The reward of asking out a particular man depends on the probability that he will accept the offer. Thus, the reward distribution changes based on what the men are learning, introducing an externality to the search process. The same applies to men under the sequential mechanism since the probability that a particular woman will ask a man out changes over time. This is a problem of coordinated learning that is related to the literature on learning in games [Fudenberg and Levine, 1998] and to reinforcement learning of nonstationary distributions in multiagent environments [Bowling and Veloso, 2002]. Some recent work by Auer *et al*. on "adversarial" bandit problems, which makes no distributional assumptions in deriving regret bounds, is relevant in this context [Auer *et al.*, 2002].

Since the underlying v_{ij}'s are constant we define Q_{ij} as person i's sample mean of the payoff of *going out* with person j. In order to deal with the nonstationarity of p_{ij}'s, on the other hand, we use a fixed learning rate for updating the probabilities which allows agents to forget the past more quickly:

$$p_{ij}[t] = (1 - \eta) p_{ij}[t-1] + \eta I[\text{offer made / accepted}]$$

where η is a constant and I is an indicator function indicating whether a man accepted an offer (for the woman's update, applied only if woman i made an offer to man j at time t) or whether a woman made an offer to a man (for the man's update, applied at each time period t).

4 Empirical Results

Our simulations involve a market with 5 women and 5 men. The agents use η of 0.05 for updating their probability estimates and the probability of exploration evolves with time as $\epsilon_t = \epsilon^{t/1000}$. Agents have true values $v^m_0 = v^w_0 = 10, v^m_1 = v^w_1 = 9, v^m_2 = v^w_2 = 8, v^m_3 = v^w_3 = 7, v^m_4 = v^w_4 = 6$. We assume $K_i = 0 \ \forall i$. The noise signals $\epsilon^{\{m,w\}}_{ijt}$ are i.i.d. and

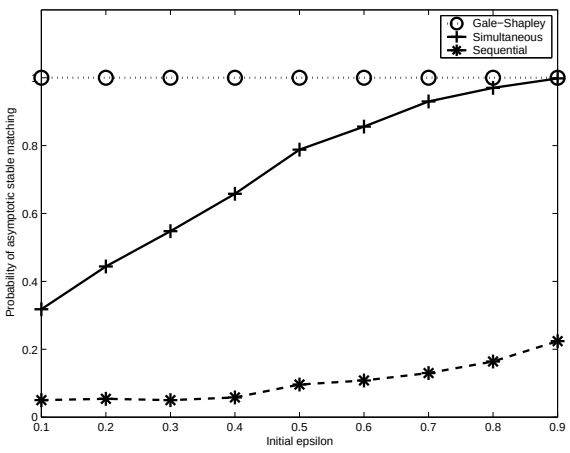

Figure 1: Probability of a stable (asymptotic) matching as a function of initial value of ϵ

drawn from a normal distribution with mean 0. Unless otherwise specified, the standard deviation of the noise distribution is 0.5. Reported results are averages from 500 simulations, each lasting a total of 30,000 time steps. Initial values of Q_{ij} are sampled from a uniform $[6, 10]$ distribution and initial values of p_{ij} are sampled from a uniform $[0, 1]$ distribution.

Our experiments show that settings in which agents are matched using the Gale-Shapley mechanism always result in asymptotically stable matchings, even for very small initial values of ϵ such as 0.1. After a period of exploration, where the agents match up with many different partners and learn their preferences, agents start pairing up regularly with just one partner, and this is always the agent with the same ranking on the other side. This indicates that agents are generally successful at learning their preferences. Interestingly, even if only one side explores (that is, either men or women always pick the greedy action), populations almost always converge to stable matchings, with a slight decline in the probability of stability when only men explore (under the woman-optimal matching algorithm, women's rankings can have a greater effect on the matching than men's rankings).

The probabilities of convergence under the simultaneous and sequential choice mechanisms are significantly lower, although they increase with larger initial values of ϵ. We can see this behavior in Figure 1, which also reveals that the probability of convergence to a stable matching is much higher with the simultaneous choice mechanism. Table 1 shows these probabilities as well as the score, which is a measure of how large the deviation from the stable matching is. If men and women are indexed in order of their true value ranking, the score for a matching is defined as $\frac{1}{W}\sum_{i\in\mathcal{W}}|i - \text{Partner}(i)|$ where $\text{Partner}(i)$ is the true value ranking of the man woman i is matched with, and $\mathcal{W}$ is the set of all women.

It is also interesting to look at who benefits from the instabilities. In order to do this, we define a notion of regret for an agent as the (per unit time) difference between the reward under the stable matching and the actual reward received (a negative value of regret indicates that the agent did better than

ID	Simultaneous Regret		Sequential Regret	
	Woman's	Man's	Woman's	Man's
0	0.126	0.126	0.578	0.552
1	0.090	0.278	-0.023	0.009
2	0.236	0.136	-0.153	-0.148
3	0.238	-0.126	-0.005	-0.024
4	-0.690	-0.414	-0.171	-0.187

Table 2: Distribution of regret under simultaneous choice ($\epsilon = 0.1$) and sequential choice ($\epsilon = 0.9$) mechanisms

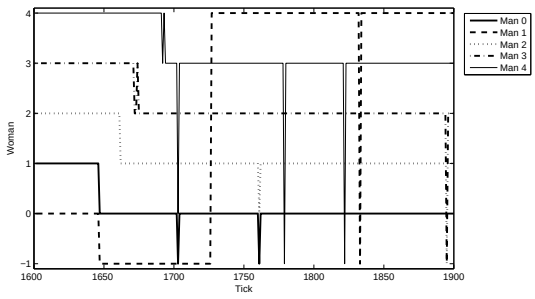

Figure 2: A "phase transition": men and women are ranked from 0 (highest) to 4 (lowest) with -1 representing the unmatched state. The graph shows the transition to a situation where the second highest ranked man ends up paired with the lowest ranked woman

(s)he would have done under the stable matching). This definition is unambiguous with sex-wide homogeneity of preferences since there is only one stable matching, but could be problematic in other contexts, when there may be multiple stable matchings. In such cases it might make sense to analyze individual agent performance in terms of the difference between average achieved reward and expected reward under one of the stable matchings depending on context.

In the case of sex-wide homogeneity of preferences we of course expect that regret will be greater for more desirable agents since they have more to lose when their value is not known. Table 2 shows the distribution of regrets for simultaneous and sequential choice. The regrets are averaged over the last 10,000 periods of the simulation. Under simultaneous choice, the worst woman benefits at the expense of all other women while the worst two men benefit at the expense of the top three. Under sequential choice, other agents benefit at the expense of the best ones. Further research is needed to better understand this apparent difference in the distribution of regrets under the two mechanisms.

Figure 2 shows interesting dynamic behavior in one particular simulation where the second best man ends up with the worst woman. The graph shows which man is matched with which woman at each time period. The lines represent the men, and the numbers on the Y axis represent the women. The value -1 represents the state of being unmatched in that period for a man. The men and women are ranked from 0 (best) to 4 (worst). Initially, the second best man is paired with the best woman so he keeps rejecting offers from all the

ϵ	Simultaneous Choice		Sequential Choice		Gale-Shapley	
	Pr (stability)	Score	Pr (stability)	Score	Pr (stability)	Score
.1	0.318	0.4296	0.050	0.9688	1.000	0.0000
.2	0.444	0.3832	0.054	0.9280	1.000	0.0000
.3	0.548	0.2920	0.050	0.8560	1.000	0.0000
.4	0.658	0.1880	0.058	0.8080	1.000	0.0000
.5	0.788	0.0992	0.096	0.7448	1.000	0.0000
.6	0.856	0.0672	0.108	0.7064	1.000	0.0000
.7	0.930	0.0296	0.130	0.6640	1.000	0.0000
.8	0.970	0.0120	0.164	0.5848	1.000	0.0000
.9	0.998	0.0008	0.224	0.4912	1.000	0.0000

Table 1: Convergence to stability as a function of ϵ

σ	Pr (stability)	Score
0.5	0.658	0.1880
1.0	0.636	0.1952
1.5	0.624	0.2120
2.0	0.600	0.2328

Table 3: Convergence to stability as a function of σ with simultaneous choice and initial $\epsilon = 0.4$

other women. These women thus learn that he is extremely particular about whom he dates and there is no point in asking him out. When the best woman finally learns that she can get a better man this triggers a chain of events in which all the men sequentially move to the woman ranked one higher than the one they were seeing. However, all the women have such a low opinion of the second best man that he ends up getting matched with the very worst woman. The matching shown at the end of the graph is the final asymptotic matching in this simulation. Note that the gender asymmetry (women ask men out) precludes this from happening to a woman.

Finally, we analyze how the noise distribution affects the probability of stability. We expect that there will be less convergence to stability when the signals are less precise. We ran experiments in which the standard deviation of the noise distribution was changed while holding other factors constant. We used an initial ϵ of 0.4 and the same underlying values as above. Table 3 shows the results using the simultaneous choice mechanism. We vary the standard deviation from one half of the distance between the two adjacent true values (0.5) to twice that distance (2.0), and the probability of stability falls by less than 10%. This suggests that the instabilities arise mostly from the structure of the problem and the non-stationarity of probability estimates rather than from the noise in the signals of value.

5 Optimism and Exploration

The insight that instabilities arise mostly from the structure of the problem suggests an alternative method for engineering asymptotic stability into the system. Suppose agents are initially optimistic and their level of optimism declines over time. This is another form of patience — a willingness to wait for the best — and it should lead to more stable outcomes.

Optimism can be represented by a systematic overestimation of the probability that your offer will be accepted or that

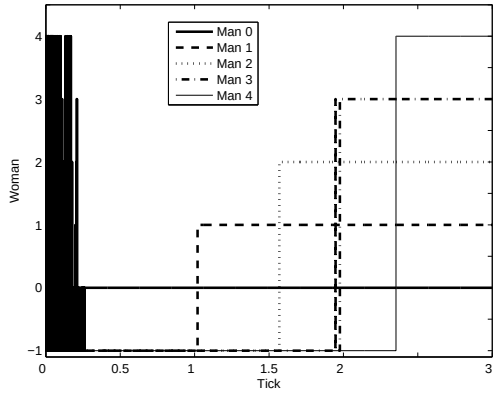

Figure 3: The mechanism of stability with optimism: agents keep trying better ranked agents on the other side until they finally "fall" to their own level

an offer will be made to you. We explore this empirically with the sequential choice mechanism. Instead of using the learned values of p_{ij} as previously defined, agents instead use an optimistic version. At time t, both men and women use the optimistic probability estimate:

$$p'_{ij} = \alpha_t + (1 - \alpha_t)p_{ij}$$

in decision making (the actual p_{ij} is maintained and updated as before). α_t should decline with time. In our simulations $\alpha_0 = 1, \alpha_T = 0$ (where T is the length of the simulation) and α declines linearly with t. There are no other changes to any of the decision-making or learning procedures.

Figure 3 shows the process by which agents converge to asymptotic matchings (in this case a stable one) with the optimistic estimates. The structure of the graph is the same as that in Figure 2. Essentially, each agent keeps trying for the best agent (s)he can match with until the optimism parameter has declined sufficiently so (s)he "falls" to the equivalently ranked agent on the other side of the market. Figure 4 shows that agents are considerably more likely to converge asymptotically to stable matchings using this algorithm for any value of the initial exploration probability. Of course, this convergence comes at the expense of utility achieved in the period before the agents settle down.

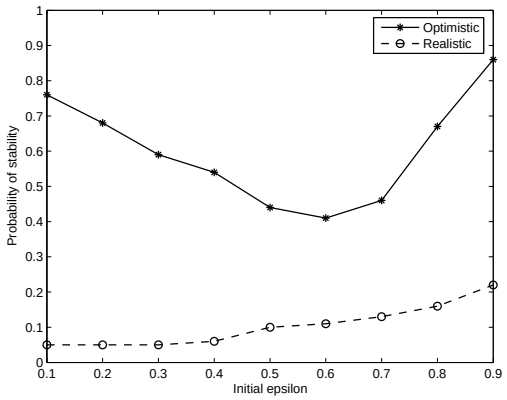

Figure 4: Probability of convergence to stability for different initial values of epsilon with all agents using the optimistic algorithm versus all agents using the realistic algorithm

The surprising feature in Figure 4 is that stable matchings are more likely with smaller initial exploration probabilities. The V-shape of the graph shows that the probability of stability declines with increasing exploration up to an initial ϵ value of 0.6, before starting to increase again, in contrast to the monotonically increasing probability of stability without optimism. This can be explained in terms of the fact that a small level of exploration is sufficient for agents to learn their preferences. Beyond that, additional exploration becomes counterproductive because the probability estimates at the key stages become less reliable.

6 Conclusions and Future Work

We have defined two-sided bandit problems, a new class of problems in multi-agent learning and described the properties of three important matching mechanisms with ϵ-greedy learning rules. Two-sided bandit problems are of great relevance for social science in general and the search for marriage partners in particular. The social norms governing exploration before marriage have been changing rapidly over the last few decades and until now we have had no formal structure within which to study the sources and consequences of these changes. Our model is also more generally applicable to two-sided markets in which agents have to learn about each other.

This paper only scratches the surface of a large and potentially fruitful set of theoretical and empirical questions. We are exploring learning algorithms that would allow agents to perform well[4] across a broad range of environments without having to make assumptions about the decision-making algorithms or learning processes of other agents. Another direction of research is to explicitly characterize equilibria in simpler settings. We are also interested in more complex versions of the problem that allow for a greater diversity of preferences and a larger number of agents.

[4]In the sense of regret minimization.

Acknowledgements

We would like to thank David Laibson, Sayan Mukherjee, Tommy Poggio, Al Roth and the referees for useful comments. SD acknowledges grants to CBCL from Merrill-Lynch, the National Science Foundation, the Center for e-Business at MIT, the Eastman Kodak Company, Honda R&D Co, and Siemens Corporate Research, Inc. EK acknowledges support from the National Science Foundation Graduate Research Fellowship, the National Institute on Aging (Grant # T32-AG00186), the Institute for Humane Studies, and the Chiles Foundation.

References

[Auer *et al.*, 2002] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[Berry and Fristedt, 1985] D. A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments*. Monographs on Statistics and Applied Probability. Chapman and Hall, London, UK, 1985.

[Bowling and Veloso, 2002] M. Bowling and M. M. Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136:215–250, 2002.

[Burdett and Wright, 1998] K. Burdett and R. Wright. Two-sided search with nontransferable utility. *Review of Economic Dynamics*, 1:220–245, 1998.

[Fudenberg and Levine, 1998] D. Fudenberg and D. Levine. *The Theory of Learning in Games*. MIT Press, Cambridge, MA, 1998.

[Gale and Shapley, 1962] D. Gale and L. S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

[Gilbert and Mosteller, 1966] J. Gilbert and F. Mosteller. Recognizing the maximum of a sequence. *Journal of the American Statistical Association*, 61:35–73, 1966.

[Gittins and Jones, 1974] J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gani, K. Sakadi, and I. Vinczo, editors, *Progress in Statistics*, pages 241–266. North Holland, Amsterdam, 1974.

[Luce, 1959] D. Luce. *Invidual Choice Behavior*. Wiley, New York, 1959.

[Roth and Sotomayor, 1990] A. E. Roth and M. Sotomayor. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Econometric Society Monograph Series. Cambridge University Press, Cambridge, UK, 1990.

[Sutton and Barto, 1998] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.