Decentralized and Uncoordinated Learning of Stable Matchings: A Game-Theoretic Approach*

S. Rasoul Etesami, R. Srikant

University of Illinois Urbana-Champaign
(etesami1, rsrikant)@illinois.edu

Abstract

We consider the problem of learning stable matchings with unknown preferences in a decentralized and uncoordinated manner, where "decentralized" means that players make decisions individually without the influence of a central platform, and "uncoordinated" means that players do not need to synchronize their decisions using pre-specified rules. First, we provide a game formulation for this problem with known preferences, where the set of pure Nash equilibria (NE) coincides with the set of stable matchings, and mixed NE can be rounded to a stable matching. Then, we show that for *hierarchical* markets, applying the exponential weight (EXP) learning algorithm to the stable matching game achieves logarithmic regret in a fully decentralized and uncoordinated fashion. Moreover, we show that EXP converges locally and exponentially fast to a stable matching in general matching markets. We complement our results by introducing another decentralized and uncoordinated learning algorithm that globally converges to a stable matching with arbitrarily high probability.

Introduction

Learning stable matchings is one of the fundamental problems in computer science, economics, and engineering that has received considerable attention over the past decades. Stable matchings provide a desirable notion of stability in two-sided matching markets where agents on each side of the market have preferences over the other side. A matching is called stable if no two agents prefer each other over their current matches. For instance, in college admissions, applicants have different preferences for colleges and vice versa. The goal is to match applicants to colleges in such a way that no applicant-college pair would prefer to break their current matches and instead be matched to each other. Similar situations arise in other applications, such as kidney exchange programs, job assignment to workers, matching in online dating platforms, and scheduling jobs on machines.

It is known that when all preferences are known, stable matchings always exist, and a simple decentralized and uncoordinated *Deferred-Acceptance* (DA) algorithm, first proposed by Gale and Shapley (1962), can find such stable matches in a polynomial number of iterations. However, when preferences are unknown, developing such an algorithm faces major challenges due to the lack of coordination. While the DA algorithm provides a satisfactory solution for many practical applications, there are scenarios where the information structure of the problem hinders the implementation of the DA algorithm. For instance, as argued in Maheshwari, Sastry, and Mazumdar (2022), there has been a recent emergence of online matching markets, such as online labor markets (e.g., TaskRabbit, Upwork), online dating markets (e.g., Tinder, Match.com), and online crowdsourcing platforms (e.g., Amazon Mechanical Turk), where users do not know their preferences a priori and can repeatedly interact with the market to improve their matching quality. The more a pair on both sides of the market gets matched, the more certain they become about their preferences. Thus, an important question is how agents should interact with the market so that, in the absence of any coordination, they can learn their preferences quickly and achieve a stable matching. The mathematical model in this work is an abstraction of such interactions and serves as a starting point for understanding these markets. Our main contribution is to show that such markets can be studied using a game-theoretic framework, which yields stronger results than prior work.

In this work, we consider a two-sided matching market framed as a marriage problem, where the agents on one side are referred to as men and the agents on the other side as women. We assume that both men and women have preferences regarding the other side, and that women are aware of their ordinal preferences for men. However, men do not know their preferences for women and only learn them if they propose and get matched. In this case, the matched men receive a noisy estimate of their preferences. We assume that agents cannot observe any other information (e.g., who is rejected or accepted) and cannot coordinate in any way. The goal for men is to follow a decentralized and uncoordinated proposal strategy such that the entire market converges to a stable matching over time. One of the major challenges in learning a stable matching in such markets is handling collisions. More precisely, when multiple men propose to the same woman, only one of them gets matched and receives useful information, while all others are rejected and receive no information about their preferences. Therefore, resolving such collisions without coordination is a significant issue.

*An extended version of this work can be found in Etesami and Srikant (2024).

Moreover, even if a man is matched with a woman, he receives a noisy utility drawn from an unknown distribution that may differ from his true preference. Consequently, a collision-avoidance process needs to be repeated many times before men can accurately estimate their true preferences. One of our goals in this work is to provide decentralized and uncoordinated algorithms that can effectively learn the true preferences while minimizing the number of collisions.

Related Work

Stable matching was first introduced by Gale and Shapley (1962) as a model for college admissions and to study the stability of marriages. Since then, there has been a tremendous effort to generalize and extend stable matchings to various market settings Roth (2008). In their seminal work, Gale and Shapley provided a simple Deferred Acceptance (DA) algorithm in which men propose to their most preferred women, and the women reject all proposals except the one from their most preferred man. They showed that such a decentralized algorithm converges to a stable matching in at most $O(n^2)$ steps, where n is the total number of men and women in the market. Unfortunately, when the agents' preferences are unknown, the DA algorithm is no longer guaranteed to converge to a stable matching without additional coordination. Therefore, to extend the DA algorithm to the case of unknown preferences, earlier works Pokharel and Das (2023); Maheshwari, Sastry, and Mazumdar (2022); Pagare and Ghosh (2023); Liu, Mania, and Jordan (2020); Kong, Wang, and Li (2024); Kong and Li (2023) have proposed various coordination-based DA algorithms. The main idea is to develop a phase-dependent process where, in some phases, the agents primarily explore, and in other phases, they implement the DA algorithm based on current estimates of their unknown preferences.

It is known that stable matchings with known preferences can be characterized using the extreme points of a fractional polytope that is totally dual integral Vohra (2012); Teo and Sethuraman (1998). Moreover, stable matchings exhibit other useful properties that allow their geometry to be characterized using so-called *rotations* Király and Pap (2008). It was shown in Roth and Vate (1990) that uncoordinated random better-response dynamics converge to a stable matching with probability one. Subsequently, Ackermann et al. (2008) provided an exponential lower bound for the worst-case convergence time of the uncoordinated random better/best response dynamics to a stable matching. However, all these results apply to the case when market preferences are fully known, allowing agents to compute their best or better responses and update their decisions accordingly.

More recently, there has been a significant interest in matching markets with unknown preferences. However, depending on the information structure and the feedback received by the agents, one might expect a wide range of performance guarantees, which also depends on the type of algorithms followed by the agents (e.g., centralized/decentralized or coordinated/uncoordinated). The work in Bei, Chen, and Zhang (2013) devised a randomized polynomial-time centralized algorithm for finding a stable matching with unknown deterministic preferences. The study in Wang et al. (2022) considers learning a specific stable matching with unknown stochastic preferences by adopting a suitable notion of stable regret, which measures the number of times that men propose to women other than their stable pair in that stable matching. The work in Maheshwari, Sastry, and Mazumdar (2022) addressed the problem of learning stable matchings in hierarchically structured markets, and developed a phase-coordinated decentralized algorithm that achieves logarithmic regret in time with respect to the market's unique stable matching. Additionally, Liu et al. (2021) considered learning stable matchings in an uncoordinated and decentralized fashion. However, they require stronger assumptions on information feedback (e.g., a player observes the actions of other players in the previous round) and use a different performance metric than the one we consider in this work. We refer to Jagadeesan et al. (2021); Liu, Mania, and Jordan (2020) and Basu, Sankararaman, and Sankararaman (2021) for other learning algorithms in two-sided matching markets with different performance guarantees. While all these works address the problem of learning a stable matching with unknown preferences, their algorithms differ substantially due to the information/feedback structure, the type of performance guarantee, and the level of coordination allowed among the agents.

Contributions

We consider the problem of learning stable matchings in a decentralized and uncoordinated manner. In particular, we focus on the weakest type of feedback that men can receive: a man observes a noisy version of his true preference only if his proposal is accepted, and receives no information otherwise. *Our main objective is to provide a novel and principled approach to designing decentralized and uncoordinated learning algorithms for stable matchings by examining them through the lens of NE learning in noncooperative games.* Our contributions can be summarized as follows:

- We first formulate the *stable matching game* and show that its set of pure NE coincides with the set of stable matchings. Additionally, its mixed NE points can be rounded in a decentralized way to obtain a stable matching. This connection provides an alternative approach for measuring the closeness of the market dynamics to stable matchings through the lens of NE computation in games.
- Leveraging this game-theoretic formulation, we present a simple decentralized and uncoordinated algorithm, EXP, that globally converges to a stable matching in hierarchical markets. This algorithm achieves logarithmic regret in time, thereby extending the existing phase-coordinated algorithms in the literature to an uncoordinated one.
- We then prove that EXP converges locally to a stable matching in general matching markets at an exponential rate. Moreover, if EXP converges globally with positive probability, it must converge to a stable matching.
- We complement our results by providing an alternative decentralized and uncoordinated algorithm that globally converges to a stable matching in general matching markets with arbitrarily high probability, a fact that was unknown previously.

Problem Formulation

Stable Matchings with Known Preferences

Here, we first introduce the stable matching problem with *known* preferences Gale and Shapley (1962). In this problem, there are a set M of men and a set W of women, where by introducing dummy agents with appropriate preferences, without loss of generality we may assume $|M| = |W| = n$ Vohra (2012) (we use the term "agents" to refer to either men or women). Each man $m \in M$ has a cardinal preference for each women $w \in W$, denoted by $\mu_{mw} \in (0, 1]$, such if $\mu_{mw} > \mu_{mw'}$, it means that m prefers w over w'. Moreover, each woman w has an ordinal preference over the men, and we write $m >_w m'$ if woman w strictly prefers m over m'. In this work we assume that no agent has ties in their preferences and we define

$$\Delta = \min_{m,w \neq w'} |\mu_{mw} - \mu_{mw'}|,$$

$$\mu_{\min} = \min_{m,w} \mu_{mw}, \quad \mu_{\max} = \max_{m,w} \mu_{mw}.$$

In particular, we note that $\Delta > 0$ and $\mu_{\min} > 0$.

Definition 1 *Given a matching and two matched pairs (m, w) and (m', w'), we say that (m, w') forms a* blocking pair *if $\mu_{mw'} > \mu_{mw}$ and $m >_{w'} m'$. In other words, (m, w') is a blocking pair if both m and w' prefer each other over their current matches. A perfect matching is called a* stable matching *if it does not contain any blocking pair.*

Stable Matchings with Unknown Preferences

In this work, we consider the problem of finding a stable matching with the main difference that men do not know their true preferences μ_{mw}, and they only get to learn them through interactions with the market. More precisely, for any $m \in M, w \in W$, we assume that the preference of man m about woman w is in the form of a $[0, 1]$-supported unknown distribution $\mathcal{D}_{mw}$ with unknown mean $\mu_{mw} > 0$. We assume men and women interact in this market through a discrete-time process, where at any time $t = 0, 1, 2, \ldots$ that a man m proposes to a woman w, he receives a feedback in the following form:

- If the proposal of m gets rejected at time t because woman w has received an offer from a more preferred man m', i.e., $m' >_w m$, then m receives no information as feedback other than the fact that he was rejected by woman w at time t.
- If the proposal of m gets accepted by w at time t, then m observes an i.i.d. realization of his sampled preference drawn from $\mathcal{D}_{mw}$, denoted by $\hat{\mu}^t_{mw}$.

In the stable matching problem with unknown preferences, the agents' goals are to interact with the market through the information feedback structure described above such that the emerging dynamics converge to a stable matching of the market with *known* preferences $\{\mu_{mw}\}$. We note that in the discrete-time process described above, men are the decision makers on whom to propose at each time t; women merely respond to men's decisions by accepting their most preferred proposal (if they received any) and rejecting all others.

A Game-Theoretic Formulation and Nash Equilibrium Characterization Results

In this section, we provide a complete information noncooperative game formulation for the stable matching problem with known preferences whose set of pure Nash equilibrium (NE) points coincides with the set of stable matchings. Later, we will show how to leverage this formulation to extend our results for learning stable matchings with unknown preferences. Such a formulation has three main advantages: i) it reduces the learning task in matching markets to learning NE in continuous-strategy concave games, ii) it captures the selfish behavior of men (players) and the feedback they receive through their payoff functions, and iii) it simplifies the combinatorial structure inherent in learning stable matchings to learning NE in noncooperative games.

Stable Matching Game

Consider a complete information game in which the action set of each man (player) is given by the set of women W, and the set of (mixed) strategies for each player is given by the probability simplex over the set of women W, i.e., the strategy set of player m is defined by $\mathcal{X}_m = \{x_m \in \mathbb{R}_+^{|W|} : \sum_{w \in W} x_{mw} = 1\}$. Let x_{-m} denote the strategy vector of all the players except the mth one. Given a strategy profile $x = (x_m, x_{-m})$, we define the payoff of player m by

$$u_m(x) = \sum_{w \in W} \Big(\mu_{mw} \prod_{k >_w m} (1 - x_{kw})\Big) x_{mw}, \qquad (1)$$

where $\mu_{mw} > 0$ is the utility (true preference) received by man m if he gets matched to women w.

Definition 2 *A strategy x_m for player m is called* pure *if it is zero in all coordinates except one. Otherwise, it is called a mixed strategy. A pure (mixed) strategy profile x^* is called a pure (mixed) NE if for any player m and any pure (mixed) strategy x_m, we have $u_m(x_m, x^*_{-m}) \leq u_m(x^*_m, x^*_{-m})$.*

One can interpret $u_m(x)$ as the expected utility that player m would receive by successfully getting matched if each man independently proposes to women according to his mixed strategy distribution x_m. The reason for defining the payoff functions as in (1) is that we want our devised algorithms to be implemented in a fully decentralized and uncoordinated manner among men by relying only on their received feedback (embedded into their payoff functions). While such payoff functions are highly nonlinear, they are essential to eliminate any degrees of coordination among the players. In other words, the cost of devising a fully coordination-free algorithm comes in analyzing more complex payoff functions with higher degrees of nonlinearity. The following are three properties of the payoffs (1).

- Player m's strategy does not have any impact on the coefficient terms $\mu_{mw} \prod_{k >_w m} (1 - x_{kw})$, $w \in W$. In particular, the gradient of $u_m(\cdot)$ with respect to x_{mw} equals

$$v_{mw}(x) := \nabla_{mw} u_m(x) = \mu_{mw} \prod_{k >_w m} (1 - x_{kw}).$$

- The payoff of each player m is linear with respect to his own strategy x_m.

- For any fixed strategy of other players x_{-m}, player m always has a pure strategy best response that is obtained by setting $x_{mw} = 1$ for the woman w that achieves the maximum value $v_{mw}(x)$, and $x_{mw} = 0$ otherwise.

In the remainder of this paper, we will refer to the above noncooperative game as the *stable matching game* and denote it by $\mathcal{G} = (M, \{u_m\}_{m\in M}, \{\mathcal{X}_m\}_{m\in M})$. It is important to note that although our interest is in obtaining a NE of the stable matching game, which is a *complete information* one-shot game, our goal is to learn such a NE by repeatedly playing the *incomplete information* game, where the true preferences μ_{mw} are not fully known.

Nash Equilibrium Characterization

In this part, we show that the set of pure and mixed NE points of the stable matching game has interesting connections with the set of stable matchings when preferences are *known*. The following theorem establishes one such result.

Theorem 1 *A pure strategy profile $x^* \in \{0,1\}^{n^2}$ corresponds to the characteristic vector of a stable matching if and only if it is a pure NE for the stable matching game.*

We note that computing a pure NE in the stable matching game with known preferences is not PPAD-complete. The reason is that the DA algorithm can always find a stable matching (and hence a pure NE) in polynomial time. However, the challenge arises from the restrictions on information feedback, non-coordination, and incomplete information with stochastic rewards, which we will address in the subsequent sections. While the above equilibrium characterization theorem concerns pure NE points, it is possible that the stable matching game admits a mixed NE. In fact, mixed NE points may exhibit unpredictable patterns, making it difficult to draw general conclusions about their structure. Nevertheless, the following theorem establishes an interesting connection between mixed and pure NE points, implying that obtaining a mixed NE with full support on the women's side is as satisfactory as obtaining a pure NE.

Theorem 2 *Let x be any mixed NE in which each woman receives at least one (fractional) proposal. For each man m, let w_m be the least preferred woman among those that he fractionally proposes to them, i.e., $w_m = \operatorname{argmin}_{w\in W}\{\mu_{mw} : x_{mw} > 0\}$. Then $\{(m, w_m), m \in M\}$ forms a stable matching for the stable matching game.*[1]

The rationale behind the rounding in Theorem 2 is that men aim to ensure that, even in the worst-case scenario, they are matched with their stable partner. Therefore, among all the women they are proposing to fractionally, they eventually tend to propose to the least preferred one on their list. In fact, Theorem 2 offers a simple decentralized method to convert mixed NE points into stable matchings. We can first design a decentralized algorithm (e.g., using Algorithm 1 in conjunction with a preference estimation oracle) to learn a mixed NE of the stable matching game and acquire estimated preferences. Subsequently, each man can convert his fractional mixed strategy into a pure strategy in a decentralized manner by proposing only to his least (estimated) preferred woman among those he fractionally proposes to. This process ensures that the resulting pure strategy profile will be a pure NE with high probability, and the probability of error vanishes over time. Importantly, each man has complete access to his own estimated preferences, and the rounding process outlined in Theorem 2 does not require knowledge of others' mixed strategies. Each man only needs to be aware of his own mixed strategy and estimated preferences.

[1] If the condition that each woman receives at least one proposal does not hold, the theorem still holds for the induced submarket in which each woman receives at least one fractional proposal.

Logarithmic Regret for Hierarchical Markets

In this section, we consider the stable matching game with unknown preferences and develop a decentralized and uncoordinated algorithm for learning its NE points. In particular, we focus on matching markets with specific structures because without imposing any assumption on the matching market, there are exponential lower bounds for the number of iterations to find a stable matching through better response dynamics Ackermann et al. (2008). One example of such a structured matching market is the *hierarchical* matching market, where it was shown in Maheshwari, Sastry, and Mazumdar (2022) that if men follow a certain decentralized (but coordinated) algorithm, the men's expected regret (see Definition 3) is at most logarithmic in time but exponential in terms of other parameters such as the number of men n. In this section, we show that a much simpler algorithm (Algorithm 1) also achieves logarithmic regret in a fully uncoordinated fashion, hence, removing the phase-dependent coordination required by the algorithm in Maheshwari, Sastry, and Mazumdar (2022). In particular, our analysis provides a clear and closed-form characterization of the other constants involved in the regret bound. Throughout this section, we impose the following hierarchical assumption on the matching market.

Assumption 1 *We assume that the sets of men and women can each be ordered as $M = \{m_1, \dots, m_n\}$ and $W = \{w_1, \dots, w_n\}$ such that any man and woman with the same rank prefer each other above any other partner with a lower rank; i.e., man m_k prefers woman w_k to women $w_{k+1}, w_{k+2}, \dots, w_n$, and woman w_k prefers man m_k to men $m_{k+1}, m_{k+2}, \dots, m_n$. It is easy to see that under this hierarchical assumption, the matching $\{(m_k, w_k), k = 1, \dots, n\}$ is a stable matching.*

Remark 1 *The condition imposed in Assumption 1 is also known as the* Sequential Preference Condition *(SPC) Clark (2006) and is weaker than the α-reducible condition considered in Maheshwari, Sastry, and Mazumdar (2022), which assumes that each submarket has a* fixed-pair*: A pair (m, w) is called a fixed pair if both m and w prefer each other the most among anyone else in that submarket. In fact, one can show that a matching market has a unique stable matching if and only if it satisfies α-reducible condition Maheshwari, Sastry, and Mazumdar (2022). Since Assumption 1 is weaker than the α-reducible condition, all our results immediately applies to α-reducible markets.*

Definition 3 *Let us assume that players follow a decentralized and uncoordinated algorithm $\mathcal{A}$ that results in a sequence of proposals $\alpha^t = (\alpha_m^t, m \in M), t = 1, \ldots, T$. The* expected regret *of algorithm $\mathcal{A}$ is defined by*

$$R(T) = \sum_{t=1}^{T} \mathbb{E}\big[\mathbf{1}_{\{\exists k \in [n]: \alpha^t_{m_k} \neq w_k\}}\big],$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. In other words, $R(T)$ is the expected number of times that the men's proposals do not form a stable matching.

Algorithm Design and Preliminaries

The algorithm that we propose is an adaptation of the online dual mirror descent for adversarial multi-arm bandit problem (see, e.g., EXP3 Algorithm in Lattimore and Szepesvári (2020, Chapter 11)) to the stable matching game, where for simplicity we refer to it as the EXP algorithm. In particular, it uses an entropy regularizer and adaptive stepsize for updating the players' mixed strategy distributions over time.

In the initial stage of EXP, when the men have no prior information about the women, they propose randomly. However, as time progresses, the men learn the women's preferences through their cumulative scores and gradually avoid collisions. If a man is repeatedly rejected by a woman, her cumulative score decreases, and as a result, the algorithm will propose to her with less probability. The stepsize in the EXP algorithm diminishes at a controlled sublinear rate. If proposing to a certain woman is very rewarding, her score increases roughly at a linear rate while the stepsize decreases at a sublinear rate. As a result, her cumulative score will increase over time, and the players will gradually learn the preferences. The normalization of the exponential scores ensures that the rate of increase/decrease among different candidates scales proportionally so that those with higher scores have a higher probability of receiving proposals. Moreover, the scores are designed to provide an unbiased estimator of the payoff gradients, allowing agents to eventually maximize their own payoffs and reach a NE (stable matching).

To describe the EXP algorithm, let us denote the gradient vector of player m by $v_m(x) = \nabla_m u_m(x)$ such that for any woman w, we have $v_{mw}(x) = \mu_{mw} \prod_{k >_w m}(1 - x_{kw})$. In particular, we note that $u_m(x) = \langle x_m, v_m(x) \rangle$. Therefore, when restricting to a pure strategy (action) profile $\alpha^t = (\alpha_1^t, \ldots, \alpha_n^t) \in W \times \cdots \times W$ that is played at time t, the w-th coordinate of the gradient vector is given by

$$v_{mw}^t := \mu_{mw} \mathbf{1}_{\{\alpha_k^t \neq w \ \forall k >_w m\}}, \tag{2}$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. On the other hand, by playing a pure strategy profile α^t, each man m receives a feedback only for the woman α_m^t that he has proposed to her at time t. In particular, if $\alpha_m^t = w$, player m observes the payoff $\mu_{mw}^t \mathbf{1}_{\{\alpha_k^t \neq w \ \forall k >_w m\}} \mathbf{1}_{\{\alpha_m^t = w\}}$ at time t, where we recall that μ_{mw}^t is the realization of a random reward drawn independently from an unknown distribution $\mathcal{D}_{mw}$ with unknown mean μ_{mw} and bounded variance $\sigma_{mw}^2 \in [0, 1]$.

During the course of the algorithm, at each time t, each player m holds a mixed strategy $X_m^t \in \mathcal{X}_m$ and updates it whenever he receives new information. He then computes his adjusted mixed strategy $\hat{X}_m^t = (1 - \gamma^t) X_m^t + \frac{\gamma^t}{n} \mathbf{1}$, where $\gamma^t > 0$ is a mixing parameter and $\mathbf{1}$ is the vector of all ones. Using ideas from importance sampling for bandit problems Bubeck, Cesa-Bianchi et al. (2012); Giannou, Vlatakis-Gkaragkounis, and Mertikopoulos (2021), and to obtain an unbiased estimator of the actual gradient vector, player m constructs an estimate vector $\hat{v}_m^t$ by normalizing the received feedback with his adjusted mixed strategy probabilities as

$$\hat{v}_{mw}^t = \frac{\mu_{mw}^t \mathbf{1}_{\{\alpha_k^t \neq w \ \forall k >_w m\}}}{\hat{X}_{mw}^t} \cdot \mathbf{1}_{\{\alpha_m^t = w\}}, \quad w \in W. \tag{5}$$

The reason why we use adjusted mixed strategies $\hat{X}^t$ with additional mixing parameter γ^t is to ensure that the mixed strategies remain bounded away from zero by a positive quantity. That makes the estimators (5) have bounded variance (see Lemma 1), which allows us to use martingale concentration results to bound the probability of various events.

Let us denote the (random) score vector of player m at time $t-1$ by $\hat{L}_m^{t-1} = (\hat{L}_{mw}, w \in W)$. Player m uses $\hat{L}_m^{t-1}$ to compute his mixed strategy vector X_m^t using the logit update rule (3). He then proposes to a woman α_m^t chosen independently at random according to his adjusted mixed strategy $\hat{X}_m^t$, and receives as feedback $\hat{v}_m^t$ with entries given by (5). Player m then updates his score vector by $\hat{L}_m^t = \hat{L}_m^{t-1} + \hat{v}_m^t$ and proceeds to the next round. The detailed description of the EXP algorithm is summarized in Algorithm 1.

Algorithm 1: EXP Algorithm for Player m

Input: A decreasing stepsize sequence $\{\eta^t > 0\}$, a mixing sequence $\{\gamma^t > 0\}$, and an initial vector $\hat{L}_m^0 = \mathbf{0}$.
For $t \geq 1$, player m independently performs the following:

- Player m computes his mixed-strategy vector X_m^t using

$$X_{mw}^t = \frac{\exp(\eta^t \hat{L}_{mw}^{t-1})}{\sum_{w'} \exp(\eta^t \hat{L}_{mw'}^{t-1})}, \ \ w \in W. \tag{3}$$

- Player m draws a pure strategy $\alpha_m^t \in W$ according to his adjusted mixed strategy defined by

$$\hat{X}_m^t = (1 - \gamma^t) X_m^t + \frac{\gamma^t}{n} \mathbf{1}, \ \ w \in W,$$

  where γ^t is a mixing parameter and η^t is the stepsize.
- Player m receives a feedback in terms of his observed payoff at time t and constructs an unbiased estimate of his actual payoff gradient vector:

$$\hat{v}_{mw}^t = \frac{\mu_{mw}^t \mathbf{1}_{\{\alpha_k^t \neq w \ \forall k >_w m\}}}{\hat{X}_{mw}^t} \cdot \mathbf{1}_{\{\alpha_m^t = w\}}, \ w \in W, \tag{4}$$

  where μ_{mw}^t is the realization of the random reward if m is matched to w at time t, and updates

$$\hat{L}_{mw}^t = \hat{L}_{mw}^{t-1} + \hat{v}_{mw}^t, \ \ w \in W.$$

In order to analyze the convergence behavior of Algorithm 1 to a stable matching, let $\{\mathcal{F}^{t-1}\}_{t=1}^{\infty}$ be the increasing filtration sequence that is adapted to the history of the random processes generated by Algorithm 1. More precisely,

$$\mathcal{F}^{t-1} = \{X^\tau, \hat{L}^\tau, \alpha^\tau, \hat{v}^\tau, \tau = 0, 1, \ldots, t-1\} \cup \{X^t\},$$

contains all the realized events up to time $t-1$ except the realization of pure strategies α^t and the rewards μ^t at time t. In particular, all relevant processes at time $t-1$ as well as X^t and $\hat{X}^t$ are $\mathcal{F}^{t-1}$-measurable, but α^t and $\hat{v}^t$ are not $\mathcal{F}^{t-1}$-measurable. In fact, it is easy to show that the feedback $\hat{v}_m^t$ received by player m at time t is conditionally an unbiased and bounded estimator of the actual gradient vector v_m^t, as stated in the following lemma.

Lemma 1 *The received feedback is a conditionally unbiassed and bounded estimate of the gradients, i.e., $\forall m, w$:*

$$\mathbb{E}[\hat{v}_m^t | \mathcal{F}^{t-1}] = v_m(\hat{X}^t), \quad \mathbb{E}[(\hat{v}_{mw}^t)^2 | \mathcal{F}^{t-1}] \leq \frac{n}{\gamma^t}.$$

Performance of EXP for Hierarchical Markets

In this section, we analyze the performance of Algorithm 1 for hierarchical matching markets. To that end, let us consider the stochastic dynamics of Algorithm 1, which for any pair (m, w) are described by

$$X_{mw}^t = \frac{\exp(\eta^t \hat{L}_{mw}^{t-1})}{\sum_{w'} \exp(\eta^t \hat{L}_{mw'}^{t-1})},$$
$$\hat{X}_{mw}^t = (1-\gamma^t) X_{mw}^t + \frac{\gamma^t}{n},$$
$$\hat{L}_{mw}^t = \hat{L}_{mw}^{t-1} + \hat{v}_{mw}^t.$$

Define $Y_{mw}^t = \sum_{\tau=1}^t v_{mw}(\hat{X}^\tau)$, where $v_{mw}(\hat{X}^\tau) = \mu_{mw} \prod_{k >_w m} (1 - \hat{X}_{mw}^\tau)$, and consider the event

$$\Omega = \Big\{ \big| (\hat{L}_{m\tilde{w}}^{t-1} - \hat{L}_{mw}^{t-1}) - (Y_{m\tilde{w}}^{t-1} - Y_{mw}^{t-1}) \big| \leq \frac{2c\sqrt{t}}{\eta^t} \ \forall t, m, w, \tilde{w} \Big\},$$

where $c = \frac{1}{8} \min_{k \in [n]} \{\Delta, \mu_{m_k w_k}\}$ is a constant. The "good" event Ω represents the set of circumstances where the difference between the accumulated realized rewards, $\eta^t(\hat{L}_{m\tilde{w}}^t - \hat{L}_{mw}^t)$, for any $m, w, \tilde{w}$, and t, remains close to its conditional expected value. The following lemma shows that the event Ω occurs with very high probability, which allows us to analyze the performance of Algorithm 1 under its conditional mean trajectory while incurring a small loss.

Lemma 2 *Fix any $\delta \in (0,1)$, and suppose each player follows Algorithm 1 with stepsize $\eta^t = \frac{1}{\sqrt{t}}$ and mixing parameter $\gamma^t = M \frac{\log t}{t}$, where $M = \frac{4n}{c} \log \frac{1}{\delta}$. Then, $\mathbb{P}\{\Omega\} \geq 1 - \delta$.*

Conditioned on the event Ω, in the following lemma we show that, due to the market structure and payoff functions, as time progresses, more pairs in the market are matched according to the underlying hierarchical order with high probability. Additionally, the size of such nested matchings grows with high probability after a constant number of steps.

Lemma 3 *Assume that each man follows Algorithm 1 with stepsize $\eta^t = \frac{1}{\sqrt{t}}$ and mixing parameter $\gamma^t = M \frac{\log t}{t}$, where $M = \frac{4n}{c} \log \frac{1}{\delta}$. Let $a_1 \geq \ldots \geq a_n$ be a sequence defined by $a_1 = \frac{\Delta}{2}$ and $a_k = \frac{1}{4} \min_{i \in [k]} \{\Delta, \mu_{m_i w_i}\}, \forall k = 2, \ldots, n$. Conditioned on the event Ω, there exists a sequence of time instances $t_1 \leq \ldots \leq t_n = O\big(\frac{nM}{c^{n+1}} \log^2(\frac{M}{c})\big)$, such that for any $k \in [n]$, $t \geq t_k$, and $w \neq w_k$, we have*

$$\hat{X}_{m_k w_k}^t \geq \frac{1-\gamma^t}{1 + (n-1)e^{-\eta^t a_k t}} + \frac{\gamma^t}{n},$$
$$\hat{X}_{m_k w}^t \leq \frac{(1-\gamma^t)e^{-\eta^t a_k t}}{1 + (n-1)e^{-\eta^t a_k t}} + \frac{\gamma^t}{n}. \tag{6}$$

Finally, in the following theorem, we combine the results of Lemma 2 and Lemma 3 to bound the expected regret of Algorithm 1 by conditioning on the event Ω.

Theorem 3 *Assume that each man follows Algorithm 1 with $\eta^t = \frac{1}{\sqrt{t}}$ and $\gamma^t = M \frac{\log t}{t}$, where $M = \frac{4n}{c} \log T$ and $c = \frac{1}{8} \min_{k \in [n]} \{\Delta, \mu_{m_k w_k}\}$. Then, the expected regret of Algorithm 1 in hierarchical matching markets is at most $R(T) = \tilde{O}\Big(\frac{n^3}{c^{n+2}} \log T + \frac{n^2}{c} \log^3 T\Big)$.*

Exponential Local Convergence for General Matching Markets

In this section, we consider learning a stable matching in general matching markets (i.e., without any hierarchical assumption) and show that Algorithm 1 converges locally at an exponential rate to a stable matching when players' strategies get sufficiently close to one of the stable matchings. To that end, we first consider the following technical lemma, which shows that if a strategy profile is sufficiently close to a stable matching, then the reward of choosing an action according to that stable matching is strictly the best decision.

Lemma 4 *Let X^* be a pure NE of the stable matching game and set $c = \frac{1}{8} \min\{\Delta, \mu_{\min}\}$. For any strategy profile x that satisfies $\|x - X^*\|_1 \leq \frac{c}{\mu_{\max}}$, we have $v_{mw^*}(x) - v_{mw}(x) > c \ \forall w \neq w^*$, where w^* is the woman that m is matched to her under the pure Nash equilibrium X^*.*

Next, we consider the following lemma that will be used to capture the local behavior of Algorithm 1 around a NE.

Lemma 5 *Suppose X^* is a pure NE and $\{X^t\}$ be the sequence of iterates generated by Algorithm 1. Moreover, let $A^{t-1} = \sum_{\tau=1}^{t-1} \big(v_{mw^*}(X^\tau) - v_{mw}(X^\tau)\big)$, where w^* is the woman that man m is matched to her under the pure Nash equilibrium X^*. Then, the following statements hold:*

- *i) If $\|X^t - X^*\|_1 \leq \frac{c}{250n^2}$, then $\eta^t \hat{L}_{mw^*}^{t-1} - \eta^t \hat{L}_{mw}^{t-1} \geq \ln(\frac{n^2}{2c}) + 6 \ \forall m, w \neq w^*$.*
- *ii) If $\eta^t \hat{L}_{mw^*}^{t-1} - \eta^t \hat{L}_{mw}^{t-1} \geq \ln(\frac{n^2}{2c}) \ \forall m, w \neq w^*$, then $\|X^t - X^*\|_1 \leq c$.*
- *iii) Let t_0 be the first time such that $1/\eta^{t_0+1} - 1/\eta^{t_0} \leq c/(\ln(\frac{n^2}{2c}) + 3)$. If for some $t \geq t_0$ we have $\eta^t A^{t-1} \geq \ln(\frac{n^2}{2c}) + 3$ and $v_{mw^*}(X^t) - v_{mw}(X^t) \geq c \ \forall m, w \neq w^*$, then $\eta^{t+1} A^t \geq \ln(\frac{n^2}{2c}) + 3$.*

Finally, using Lemmas 4 and 5 recursively, we can show that if dynamics of Algorithm 1 enter a small neighborhood of a pure NE, they remain there forever with arbitrarily high probability and converge to that NE exponentially fast.

Theorem 4 *Fix an arbitrary $\delta \in (0,1)$, and suppose each player follows Algorithm 1 with parameters that satisfy*

$$\eta^t \le \Big(8n\ln(\frac{n\pi}{\sqrt{\delta}}t)\sum_{\tau=1}^{t-1}\frac{1}{\gamma^\tau}\Big)^{-\frac{1}{2}},$$

$$\eta^t \le \Big(2n^2\sum_{\tau=1}^{t-1}\gamma^\tau\Big)^{-1}, \quad \eta^t \ge \Big(\sum_{\tau=1}^{t-1}\frac{\gamma^{t-1}}{\gamma^\tau}\Big)^{-1}. \quad (7)$$

Let t_0 be such that $1/\eta^{t_0+1} - 1/\eta^{t_0} \le c/(\ln(\frac{n^2}{2c})+3)$ and $\|X^{t_0} - X^*\| \le \frac{c}{250n^2}$, where X^* is some pure NE. Then

$$\mathbb{P}^*\Big\{\|X^{t+1} - X^*\|_1 \le 41n\exp(-ct\eta^{t+1})\ \forall t \ge t_0\Big\} \ge 1-\delta,$$

where $\mathbb{P}^*\{\cdot\} = \mathbb{P}\big\{\cdot \,\big|\, \|X^{t_0} - X^*\|_1 \le \frac{c}{250n^2}\big\}$.

Corollary 1 *For $\gamma^t = \frac{1}{t^{1/3}}, \eta^t = \frac{1}{t^{3/4}}$, the convergence rate of Theorem 4 becomes $\exp(-ct^{1/4})\ \forall t \ge t_0 = (\frac{\ln(n^2/c)+3}{c})^4$.*

Global Learning in General Markets

In this section, we address the question of whether there is an uncoordinated and decentralized algorithm capable of *globally* learning a stable matching in general markets, regardless of the convergence rate. We answer this question affirmatively by introducing an alternative algorithm that leverages the weak acyclicity of the stable matching game.

Definition 4 *In a pure strategy profile x, a matched man is the one whose strategy x_m has exactly one coordinate equal to 1 (say $x_{mw} = 1$), and no other man of higher preference proposes to w. A pure strategy profile x is a* good state *if all the men matched under x are at their best responses.*

Definition 5 (Marden, Arslan, and Shamma (2007)) *A better response path in a finite action noncoopertaive game is a sequence of pure strategy profiles $x^1, x^2, \ldots, x^L$ such that for each $\ell = 1, \ldots, L-1$, $x^{\ell+1}$ is obtained from x^ℓ by letting some player i_ℓ to play a better response. A game is called* weakly acyclic *if from any pure strategy profile x^0, there is a finite length better response path to a pure NE.*

Definition 6 *A pure NE $x^* = (x^*_m, x^*_{-m})$ of the stable matching game is called* strict *if each player has a unique best response at x^*, i.e., for every m and any pure strategy $x_m \ne x^*_m$, we have $u_m(x_m, x^*_{-m}) < u_m(x^*_m, x^*_{-m})$.*

The following lemma shows that if the stable matching game starts from a good state, then any sequence of best response dynamics by the players reaches a pure NE. We will use this lemma in the proof of our Theorem 5 to show that the stable matching game is a weakly acyclic game whose pure NE points are strict.

Lemma 6 *Consider the stable matching game that starts from a good initial state x^0. Consider any sequence of updates where at each time t, an arbitrary subset of players that includes at least one unsatisfied player updates their actions by playing their best responses. Then, the sequence of updates converges to a pure NE in no more than n^2 steps.*

Algorithm 2: Globally Convergent Algorithm for Player m

Input: An initial action $\alpha^0_m \in [W]$, initial baseline action $b^1_m = \alpha^0_m$, episode length τ_m, tolerance level $\delta > 0$, exploration probability $\epsilon \in (0,1)$, inertia probability $\omega \in (0,1)$.
For $s = 1, 2, \ldots$, do the following steps:

- During the sth episode $t \in \{(s-1)\tau_m, \ldots, s\tau_m - 1\}$, player m selects his baseline action obtained at the end of the previous episode with probability $1-\epsilon$ or explores a new uniformly sampled action with probability ϵ:

$$\alpha^t_m = \begin{cases} b^s_m & \text{w.p. } 1-\epsilon, \\ w \in \text{Unif}[W] & \text{w.p. } \frac{\epsilon}{n}. \end{cases}$$

- At the end of episode s, player m evaluates his average utility when playing each action $w \in W$ as

$$u^s_{mw} = \frac{1}{n^s_{mw}} \sum_{t=(s-1)\tau_m}^{s\tau_m - 1} \mu^t_{mw} \mathbf{1}_{\{\alpha^t_k \ne w\ \forall k >_w m\}} \mathbf{1}_{\{\alpha^t_m = w\}},$$

  where $n^s_{mw} = \sum_{t=(s-1)\tau_m}^{s\tau_m-1} \mathbf{1}_{\{\alpha^t_m = w\}}$, and computes the set $\mathcal{A}^s_m = \{w : u^s_{mw} \ge u^s_{mb^s_m} + \delta\}$.
- If $\mathcal{A}^s_m \ne \emptyset$, player m uniformly samples an action $w \in \mathcal{A}^s_m$, and updates his baseline action to $b^{s+1}_m = w$ with probability $1-\omega$. Otherwise, he sets $b^{s+1}_m = b^s_m$.

The proposed algorithm (Algorithm 2) is an adaptation of the sample experimentation algorithm for weakly acyclic games Marden et al. (2009) to the stable matching game. Intuitively, each player m goes through a sequence of phase-dependent exploration/exploitation. At the end of each episode, player m evaluates his payoff for each action he has taken during the last episode. He then updates his baseline action (an action that brings him the highest reward in the past episode up to some tolerance) and sticks to that baseline action most of the time in the next episode while still exploring new actions with a small probability.

Theorem 5 *Let $p, \omega \in (0,1)$ be arbitrary probabilities and $\epsilon \le \min\{\frac{1-p}{n}, \frac{\delta}{4n}, \frac{\Delta-\delta}{4n}\}$, where $\delta \in (0, \Delta)$. For each man m, there exists a sufficiently large episode length $\tau_m = \tau_m(p,\epsilon)$, such that if each player m follows Algorithm 2 with episode length τ_m, for all sufficiently large times $t > 0$, the pure strategy profile α^t will be a (fixed) stable matching with probability at least p.*

Conclusions

We considered the problem of learning stable matchings with unknown preferences in a fully decentralized and uncoordinated manner. Using a game-theoretic framework, we showed how to design principled learning algorithms to drive the matching market to its stable points and established several global and local convergence results in hierarchical and general markets. Our results provide new insights into learning stable matchings through NE learning in games and bridge the discrete problem of learning a stable matching with that of learning a NE in continuous-strategy games.

Acknowledgements

This research was supported by the AFOSR YIP FA9550-23-1-0107, AFOSR MURI FA9550-24-1-0002, NSF CAREER Award EPCN1944403, and NSF CCF 22-07547.

References

Ackermann, H.; Goldberg, P. W.; Mirrokni, V. S.; Röglin, H.; and Vöcking, B. 2008. Uncoordinated two-sided matching markets. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, 256–263.

Basu, S.; Sankararaman, K. A.; and Sankararaman, A. 2021. Beyond $\log^2(T)$ regret for decentralized bandits in matching markets. In *International Conference on Machine Learning*, 705–715. PMLR.

Bei, X.; Chen, N.; and Zhang, S. 2013. On the complexity of trial and error. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, 31–40.

Bubeck, S.; Cesa-Bianchi, N.; et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1): 1–122.

Clark, S. 2006. The uniqueness of stable matchings. *Contributions in Theoretical Economics*, 6(1).

Etesami, S. R.; and Srikant, R. 2024. Decentralized and uncoordinated learning of stable matchings: A game-theoretic approach. *arXiv preprint arXiv:2407.21294*.

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

Giannou, A.; Vlatakis-Gkaragkounis, E.-V.; and Mertikopoulos, P. 2021. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. *Advances in Neural Information Processing Systems*, 34: 22655–22666.

Jagadeesan, M.; Wei, A.; Wang, Y.; Jordan, M.; and Steinhardt, J. 2021. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34: 3323–3335.

Király, T.; and Pap, J. 2008. Total dual integrality of Rothblum's description of the stable-marriage polyhedron. *Mathematics of Operations Research*, 33(2): 283–290.

Kong, F.; and Li, S. 2023. Player-optimal Stable Regret for Bandit Learning in Matching Markets. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1512–1522. SIAM.

Kong, F.; Wang, Z.; and Li, S. 2024. Improved Analysis for Bandit Learning in Matching Markets. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit Algorithms*. Cambridge University Press.

Liu, L. T.; Mania, H.; and Jordan, M. 2020. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, 1618–1628. PMLR.

Liu, L. T.; Ruan, F.; Mania, H.; and Jordan, M. I. 2021. Bandit learning in decentralized matching markets. *The Journal of Machine Learning Research*, 22(1): 9612–9645.

Maheshwari, C.; Sastry, S.; and Mazumdar, E. 2022. Decentralized, communication-and coordination-free learning in structured matching markets. *Advances in Neural Information Processing Systems*, 35: 15081–15092.

Marden, J. R.; Arslan, G.; and Shamma, J. S. 2007. Regret based dynamics: Convergence in weakly acyclic games. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, 1–8.

Marden, J. R.; Young, H. P.; Arslan, G.; and Shamma, J. S. 2009. Payoff-based dynamics for multiplayer weakly acyclic games. *SIAM Journal on Control and Optimization*, 48(1): 373–396.

Pagare, T.; and Ghosh, A. 2023. Two-Sided Bandit Learning in Fully-Decentralized Matching Markets. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*.

Pokharel, G.; and Das, S. 2023. Converging to Stability in Two-Sided Bandits: The Case of Unknown Preferences on Both Sides of a Matching Market. *arXiv preprint arXiv:2302.06176*.

Roth, A. E. 2008. Deferred acceptance algorithms: History, theory, practice, and open questions. *International Journal of Game Theory*, 36: 537–569.

Roth, A. E.; and Vate, J. H. V. 1990. Random paths to stability in two-sided matching. *Econometrica: Journal of the Econometric Society*, 1475–1480.

Teo, C.-P.; and Sethuraman, J. 1998. The geometry of fractional stable matchings and its applications. *Mathematics of Operations Research*, 23(4): 874–891.

Vohra, R. V. 2012. Stable matchings and linear programming. *Current Science*, 1051–1055.

Wang, Z.; Guo, L.; Yin, J.; and Li, S. 2022. Bandit Learning in Many-to-One Matching Markets. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2088–2097.