Learning Equilibria in Matching Games with Bandit Feedback

Andreas Athanasopoulos
University of Neuchatel
andreas.athanasopoulos@unine.ch

Christos Dimitrakakis
University of Neuchatel
christos.dimitrakakis@gmail.com

Abstract

We investigate the problem of learning an equilibrium in a generalized two-sided matching market, where agents can adaptively choose their actions based on their assigned matches. Specifically, we consider a setting in which matched agents engage in a zero-sum game with initially unknown payoff matrices, and we explore whether a centralized procedure can learn an equilibrium from bandit feedback. We adopt the solution concept of *matching equilibrium* [24], where a pair consisting of a matching $\mathfrak{m}$ and a set of agent strategies X forms an equilibrium if no agent has the incentive to deviate from $(\mathfrak{m}, X)$. To measure the deviation of a given pair $(\mathfrak{m}, X)$ from the equilibrium pair $(\mathfrak{m}^\star, X^\star)$, we introduce *matching instability* that can serve as a regret measure for the corresponding learning problem. We then propose a UCB algorithm in which agents form preferences and select actions based on optimistic estimates of the game payoffs, and prove that it achieves sublinear, instance-independent regret over a time horizon T.

1 Introduction

Two-sided matching markets model situations in which two distinct sets of agents aim to match with each other based on individual preferences [40]. Classic examples include students applying to universities, workers seeking employment with firms, or large online platforms that match users with services (e.g., Uber, Amazon Mechanical Turk, Airbnb) [39, 17, 18]. A fundamental solution concept in this setting is *stable matching* [17], a notion of equilibrium where no pair of agents has an incentive to deviate from the proposed assignment. Stability has motivated extensive research in economics, while more recently, computer science has focused on learning such equilibria to analyze competition in online decision-making problems [12, 30, 25].

Prior work on learning to match has only considered scenarios where, after being matched, agents directly receive a random payoff with an unknown expected value. A more realistic setting is for matched agents to *interact*, and only receive payoffs as a result of their interaction. This reflects real world scenarios where the outcome of a matching depends on the actions taken by the individuals.

In particular, we consider two-sided markets where, after being matched, the agents select an action, as introduced in [24] in a *non-learning setting*. In our case, once agents are matched, they engage in a two-player game, where the pair of payoff matrices (A, B) is *unknown*. As a solution concept, Jackson and Watts [24] proposed *matching equilibrium* (Definition 3), where the equilibrium depends jointly on the matching $\mathfrak{m}$ and the set of strategies X selected by the agents. Informally, agents matched under $\mathfrak{m}$ must play a Nash equilibrium within each pair, and the matching $\mathfrak{m}$ must be stable. Intuitively, a pair of agents dissatisfied with their current outcome, they may have an incentive to deviate and form a new match in which they adopt a different Nash equilibrium.

We illustrate an example with *known* underlying zero-sum games (ZSGs), below.

Example 1. *Consider the following simplified example with one agent $\{p\}$ on the left side and two agents $\{a_1, a_2\}$ on the right side, where agents always prefer to match with each other from being*

18th European Workshop on Reinforcement Learning (EWRL 2025).

single. The agent can adaptively choose their action, with utilities represented by the following payoff matrices of the respective zero-sum games:

$$A_{p,a_1} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \qquad\qquad A_{p,a_2} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

Agents can naturally form their preferences based on the values of the Nash equilibrium in the respective ZSGs. Specifically, for the pair (p, a_1), the game value is $V^\star_{p,a_1} = 0$, with the Nash equiblirium $(x^\star_{p,a_1}, x^\star_{a_1})$ given by $x^\star_{p,a_1} = [0.5,\ 0.5]$ and $x^\star_{a_1} = [0.5,\ 0.5]$, while for the second pair (p, a_2), the value is $V^\star_{p,a_2} = 1$, with Nash strategies $x^\star_{p,a_2} = [1,\ 0]$ and $x^\star_{a_2} = [0,\ 1]$.

Thus, agent p prefers a_2 over a_1 under Nash strategies $(x^\star_{p,a_2}, x^\star_{p,a_1})$, and the solution consisting of the matching $M = \{(p, a_2)\}$ and the strategies $X = \{x^\star_{p,a_2}, x^\star_{a_2}\}$ corresponds to desired matching equilibrium *in the market.*

On the other hand, any deviation—either in the matching or in the strategies— can lead to an unstable outcome for the specific example [1].

In the case of known payoff matrices, a stable outcome can be trivially computed by running the algorithm proposed by Gale and Shapley [17] on preferences formed according to the value of each game [23]. In ZSG, the agents can efficiently compute the values of the game and select their strategies independently of their opponents.

Contributions. In practice, however, agents may be unaware of the underlying payoffs of the games (thus their preferences), and instead must learn through repeated interactions, observing the outcomes of their actions under specific assignments. This presents a complex learning scenario, as agents must simultaneously learn to form their preferences and choose actions. We ask the following question:

Is it possible to learn a matching equilibrium through repeated iterations of agents?

To address this question, we consider the learning setting where a central platform matches agents, and each agent receives a reward through *bandit feedback* based on their selected actions. In our model, agents select their strategies based on their match, and after each interaction, they observe the action chosen by their pair. This information allows them to update their estimates of the utilities and refine their preferences.

To characterize algorithms in our learning scenario, we propose an objective that measures the distance of the solution from equilibrium at each step. Specifically, we define *matching instability* (Definition 4), which can be interpreted as the minimum subsidy required to make the solution "stable", retaining participants in the market. Our notion extends the *Subset Instability* (Definition 2) introduced in [25, 26] for the two-sided matching problem. We also establish several properties of our metric that justify cumulative *matching instability* as a suitable measure of regret.

On the algorithmic side, we show that if agents follow the general principle of optimism in the face of uncertainty the equilibrium can be efficiently learned. Specifically, we propose a UCB-like algorithm, where each agent maintains upper confidence bounds for the payoff matrices of the relevant games. Based on these estimates, agents report their preferences to a central platform, which then computes a (potentially) stable matching. Given the assignment, each agent also selects a strategy for their match, samples an action, and receives stochastic feedback.

We prove that our algorithm achieves a regret of $\tilde{O}\left(\sqrt{T\mathbf{mkpa}}\right)$, where $\mathbf{m}$ and $\mathbf{k}$ are the number of actions available to each agent on the two sides, and $\mathbf{p}$ and $\mathbf{a}$ denote the number of agents on each side, respectively. This bound unifies and generalizes existing results in the literature, all of which are based on UCB-style algorithms. Specifically, when each agent has only one available action (i.e., $\mathbf{m} = \mathbf{k} = 1$), the setting reduces to a classical two-sided matching problem, and our regret matches the bound $\tilde{O}(\sqrt{T\mathbf{pa}})$ from Theorem 6.6 in [26]. On the other extreme, when there is only one agent on each side (i.e., $\mathbf{p} = \mathbf{a} = 1$), our regret becomes $\tilde{O}(\sqrt{T\mathbf{mk}})$, which corresponds exactly to the regret bound of the UCB algorithm proposed in [36] for learning in ZSGs with bandit feedback.

[1] Note that in general there is a set of possible stable matchings

2 Preliminaries

In this section, we formally introduce the concepts of two-player zero-sum games and the two-sided matching problem.

Notation. Throughout this paper, we use $\tilde{O}(\cdot)$ to hide logarithmic factors; formally, $f(x) = \tilde{O}(g(x))$ means that there exists a positive integer k such that $f(x) = O(g(x)\log^k g(x))$. We use $1 \vee k := \max(1, k)$ for any integer k. For a positive integer k, we let $[n] := \{1, 2, \ldots, k\}$. We also write Δ_k to denote the probability simplex of size k.

We consider an element $\mathrm{a} \in \mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$, where $\mathcal{S}_1$ and $\mathcal{S}_2$ are disjoint sets. We define $OS(\mathrm{a})$ as the opposite set of element a, i.e., $OS(\mathrm{a}) = \mathcal{S}_2$ if $\mathrm{a} \in \mathcal{S}_1$, and $OS(\mathrm{a}) = \mathcal{S}_1$ if $\mathrm{a} \in \mathcal{S}_2$.

2.1 Two-player zero-sum games

A two-player zero-sum game models an interaction between two competing agents, p and a, with $\mathbf{m}$ and $\mathbf{k}$ actions, respectively. The game is specified by a pair of payoff matrices (A_p, A_a) such that $A_p = -A_a^T = A$, where $A \in [-1, 1]^{\mathbf{m}\times\mathbf{k}}$ representing the utilities.

The agent selects an action simultaneously: p chooses $i \in [\mathbf{m}]$ and a chooses $j \in [\mathbf{k}]$. Following the action selection, the agents receive payoffs $A(i, j)$ and $-A(i, j)$, respectively.

Agents typically select their actions from mixed strategies $x \in \Delta_m$ and $y \in \Delta_n$ denoting probability distributions over the action sets of p and a respectively. The expected utilities are given by

$$U_p(x, y) = \mathbb{E}[A_p(i, j)] = x^T A y \quad \text{and} \quad U_a(y, x) = \mathbb{E}[A_a(j, i)] = -y^T A^T x.$$

A central solution concept in game theory is the *Nash equilibrium* (NE) [34], in which no agent can unilaterally improve their utility by deviating from their equilibrium strategies $(x^\star, y^\star)$ i.e.:

$$U_p(x^\star, y^\star) \geq U_p(x, y^\star) \quad \text{and} \quad U_a(y^\star, x^\star) \geq U_a(y, x^\star) \quad \forall\, y, x \in \Delta_m \times \Delta_n \tag{1}$$

where for two-player zero-sum games, this is equivalent to:

$$x^T A y^\star \leq x^{\star T} A y^\star \leq x^{\star T} A y \quad \forall\, x, y \in \Delta_m \times \Delta_n$$

The celebrated Minimax theorem [44] states that the NE $(x^\star, y^\star)$ is attained if and only if each agent selects a strategy that maximizes their payoff against the opponent's worst-case strategy, i.e.,

$$x^\star = \arg\max_{x\in\Delta_m} \min_{y\in\Delta_n} x^T A y \quad \textit{and} \quad y^\star = \arg\max_{y\in\Delta_n} \min_{x\in\Delta_m} -y^T A^T x = \arg\min_{y\in\Delta_n} \max_{x\in\Delta_m} y^T A^T x$$

where strategies can be computed efficiently via dual linear programs [45]. We denote the value of the game for each agent as $V_p^\star = U_p(x^\star, y^\star)$ and $V_a^\star = U_a(y^\star, x^\star)$, which satisfy $V_p^\star = -V_a^\star$.

2.2 Two-sided matching

The two-sided bipartite matching problem consists of two *nonempty* finite sets of agents, denoted by $\mathcal{P}$ and $\mathcal{A}$, that aims to be matched one-to-one according to a matching $\mathbb{M} \subseteq \mathcal{P} \times \mathcal{A}$, representing a set of matched agents that are pairwise disjoint. Let $\mathcal{M}_{\mathfrak{A}}$ denote the set of all matchings, while the $\mathfrak{A} = \mathcal{P} \cup \mathcal{A}$ represents the complete set of agents. By a slight abuse of notation, we also use the equivalent functional representation $\mathfrak{m} : \mathfrak{A} \to \mathfrak{A} \cup \{\perp\}$ of a matching M, where $\mathfrak{m}(p) = a$ and $\mathfrak{m}(a) = p$ for the pair $(p, a) \in \mathbb{M}$, and $\mathfrak{m}(\mathrm{a}) = \perp$ if $\mathrm{a} \in \mathfrak{A}$ is unmatched.

Each agent $\mathrm{a} \in \mathfrak{A}$ maintains a complete list of preferences π_{a} for being matched with an agent of the other side agent that is formed according to a utility function. We denote the global utility function with $U : \mathfrak{A} \times \mathfrak{A} \to [-1, 1]$, where $U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}$ denotes the utility of the agent a for being matched to agent $\mathfrak{m}(\mathrm{a})$. Note that utilities can be negative in the case of an undesirable match, while the utilities of unmatched agents are arbitrary i.e., $U_{\mathrm{a},\perp} \in [-1, 1]$, but typically not greater than 0.

Stability: To characterize a matching $\mathfrak{m}$ that aligns with agents' preferences, Gale and Shapley proposed stability [17] as a notion of *equilibrium* in the market, where no pair agents have the incentive to deviate from $\mathfrak{m}$. Bellow we define stability in the case of cardinal utility:

Definition 1 (Stability)**.** *A matching $\mathfrak{m}$ is stable under a utility function $U : \mathfrak{A} \times \mathfrak{A} \to \mathbb{R}$ if it is individually rational and has no blocking pairs, i.e.:*

1. ***Individual rationality:*** $U_{\mathrm{a},\mathfrak{m}(\mathrm{a})} \geq U_{\mathrm{a},\perp} \forall \mathrm{a} \in \mathfrak{A}$
2. ***No blocking pairs:*** *There is no pair* $(p, a) \in \mathcal{P} \times \mathcal{A}$ *such that*

$$U_{p,a} > U_{p,\mathfrak{m}(p)} \quad \textit{and} \quad U_{a,p} > U_{a,\mathfrak{m}(\mathrm{a})}$$

Gale and Shapley prove that the set of stable matching $\mathcal{S}$ is always non-empty using their Deferred Acceptance (DA) algorithm [17]. In the DA algorithm the agents of one side of the market sequentially propose to the other side, while the other side is temporarily matched with the most preferred agents so far until all agents are matched. The stable matching $\mathfrak{m}_s^\star$ produced by the algorithm is optimal for the proposing side of the market and pessimal for the side that received the proposals, in the sense that the agents of the sets are matched with their most/least preferred partner among any stable matching in $\mathcal{S}$.

To measure the distance of a matching $\mathfrak{m}$ from the equilibrium, Jagadeesan et al. [25, 26] propose Subset Instability (SI), which is formally defined below.

Definition 2 (Subset Instability, Definition 6.4 from [26])**.** *The subset instability* $\mathbf{SI}(\mathfrak{m}, U, \mathfrak{A})$ *of a matching* $\mathfrak{m}$ *according to a utility function* $U : \mathfrak{A} \times \mathfrak{A} \to \mathbb{R}$ *over a set of agents* $\mathfrak{A}$ *is defined by*

$$\min_{s \in \mathbb{R}} \sum_{\mathrm{a} \in \mathfrak{A}} s_{\mathrm{a}}$$

subject to:

$$\min\left(U_{p,a} - U_{p,\mathfrak{m}(p)} - s_p,\ U_{a,p} - U_{a,\mathfrak{m}(a)} - s_a\right) \leq 0, \quad \forall\ (p,a) \in \mathcal{P} \times \mathcal{A}, \tag{2}$$

$$U_{\mathrm{a},m(\mathrm{a})} - U_{\mathrm{a},\perp} + s_{\mathrm{a}} \geq 0, \quad \forall\ \mathrm{a} \in \mathfrak{A}, \tag{3}$$

$$s_{\mathrm{a}} \geq 0, \quad \forall\ \mathrm{a} \in \mathfrak{A}. \tag{4}$$

Note that this measure corresponds to the minimum subsidies s_{a} required to ensure the absence of blocking pairs and to satisfy individual rationality (Constraints 2 and 3). In their work, Jagadeesan et al. [25, 26] use cumulative SI as a regret measure in the learning problem with unknown utilities.

3 Setting

We study an intersection of two-sided matching and normal-form games, where matched agents engage in a game G with their assigned partners. Jackson and Watts formalized such interactions as *social games* [23, 24].

We consider the special case where an agent a is matched to $\mathfrak{m}(\mathrm{a})$ and they play a *distinct* **ZSG** described by the pair of payoff matrices $(A_{\mathrm{a},\mathfrak{m}(\mathrm{a})}, A_{\mathfrak{m}(\mathrm{a}),\mathrm{a}})$, with entities in range $[-1, 1]$. The agents $\mathrm{a} \in \mathfrak{A}$ then choose their strategies $x_{\mathrm{a},\mathfrak{m}(\mathrm{a})} \in \Delta_{k_{\mathrm{a},\mathfrak{m}(\mathrm{a})}}$ for their match, with $k_{\mathrm{a},\mathfrak{m}(\mathrm{a})}$ represents the number of actions in their respective game. When the match is clear from the context we refer to the strategy of the agent a with x_{a}. The expected utility of agent a when matched with $\mathfrak{m}(\mathrm{a})$, using strategies $(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})})$, is given by:

$$U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) = x_{\mathrm{a}}^T A_{\mathrm{a},\mathfrak{m}(\mathrm{a})} x_{\mathfrak{m}(\mathrm{a})}, \tag{5}$$

Stability: In this setting, equilibrium depends on both the matching $\mathfrak{m}$ and the set of strategies $X = \{x_{a,\mathfrak{m}(a)}\}_{a \in \mathfrak{A}}$ chosen by the agents. Intuitively, for the solution $(\mathfrak{m}, X)$ to be stable, three conditions must hold: (i) agents must be individually rational: they should prefer their assigned match over remaining unmatched; (ii) matched agents must play a Nash equilibrium with their partners; and (iii) there should be no blocking pair, i.e. no unmatched pair of agents who would both prefer to be matched with each other under the resulting equilibrium strategies. Below, we present a formal definition of a *matching equilibrium* [23, 24], for the case of cardinal utilities and ZSG:

Definition 3 (Matching Equilibrium)**.** *A* matching equilibrium *in social ZSG, is a pair* $(\mathfrak{m}, X)$*, consisting of a matching* $\mathfrak{m}$ *and a profile of strategies* $X = \{x_a\}_{a \in \mathfrak{A}}$*, such that the outcome satisfies the following conditions:*

1. ***Individual rationality:*** $U_{\mathrm{a},\perp} \leq U_{\mathrm{a},m(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})})$ *for all* $\mathrm{a} \in \mathfrak{A}$.

2. ***Nash rationality:*** $U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \geq U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x, x_{\mathfrak{m}(\mathrm{a})}) \forall x \in \Delta_{k_{\mathrm{a},\mathfrak{m}(\mathrm{a})}}$ *for all* $\mathrm{a} \in \mathfrak{A}$.

3. ***No blocking pairs:*** *There exists no pair* $(p, a) \in \mathcal{P} \times \mathcal{A}$ *such that*
$$V^{\star}_{p,a} > U_{p,\mathfrak{m}(p)}(x_p, x_{\mathfrak{m}(p)}) \quad \textit{and} \quad V^{\star}_{a,p} > U_{a,\mathfrak{m}(a)}(x_a, x_{m(a)}).$$

The first condition ensures that agents are individually rational: each receives at least as much utility as they would if unmatched. Nash rationality requires that each matched pair plays a Nash equilibrium: no agent can improve their utility by deviating from x_{a}, assuming their partner plays $x_{\mathfrak{m}(\mathrm{a})}$, as defined in Eq. 1. The third condition rules out blocking pairs—i.e., no pair of agents would both obtain strictly higher utility by deviating and playing a Nash equilibrium with each other.

Jackson and Watts [24] prove that *matching equilibrium* always exist, by extending the classical result of Gale and Shapley [17]. To see this, consider running the DA algorithm, where agents naturally form preferences over potential partners based on the value of the induced games at their NE. Specifically, assuming that agents know the payoffs of the games, each agent a can rank the agents on the opposite side $\mathrm{a}' \in OS(\mathrm{a})$ according to the value $V^{*}_{\mathrm{a},\mathrm{a}'}$ of the NE $(x^{*}_{\mathrm{a}}, x^{*}_{\mathrm{a}'})$ in their corresponding ZSG. Note that in two-player ZSG, a NE always exists and all equilibria yield the same value V^{*}.

3.1 Learning problem

While prior work on social games assumes that agents $\mathrm{a} \in \mathfrak{A}$ know the payoff structure $A_{\mathrm{a},\mathfrak{m}(\mathrm{a})}$ of the games they play with their match $\mathfrak{m}(\mathrm{a})$, we consider a setting where the game is initially *unknown*.

Specifically, we study a repeated interaction scenario in which agents are matched in pairs and engage in a ZSG that depends on their partner. Matched agentsx observe only each other's actions and receive a stochastic reward that reflects the utility of their actions. This extends the social games framework to settings with bandit feedback, where agents must learn to act strategically from partial observations.

More formally, we consider the sequential learning problem, where at each step $t = 1, \ldots, T$:

1. A central platform chooses a matching $\mathfrak{m}_t$, based on the estimated agent preferences.
2. Each agent $\mathrm{a} \in \mathfrak{A}$ observes their match $\mathfrak{m}_t(\mathrm{a})$, and selects a strategy x^{t}_{a}.
3. The agents $\mathrm{a} \in \mathfrak{A}$ draw their actions $i^{t}_{\mathrm{a}} \sim x^{t}_{\mathrm{a}}$ from their strategies.
4. Each agent $\mathrm{a} \in \mathfrak{A}$ receives a random reward $r^{t}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}$ with expectation $A_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(i^{t}_{\mathrm{a}}, i^{t}_{\mathfrak{m}(\mathrm{a})})$.
5. The agents $\mathrm{a} \in \mathfrak{A}$ observes the action of their match, $i^{t}_{\mathfrak{m}(\mathrm{a})}$, and updates their estimate of the reward $\hat{A}^{t}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(i^{t}_{\mathrm{a}}, i^{t}_{\mathfrak{m}(\mathrm{a})})$.
6. The agents update their preferences based on the estimates $\hat{A}^{t}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}$.

Assumptions 1. *Below, we enumerate the list the assumptions for our learning problem.*

A1. We assume that the utilities $U_{\mathrm{a},\perp} \in [-1, 1]$ *for remaining unmatched are known for all* $\mathrm{a} \in \mathfrak{A}$.

A2. For simplicity of the analysis, we assume that agents on the same side have the same number of available actions for each possible matching pair, i.e., $A_{p,a} \in [-1, 1]^{\mathbf{m} \times \mathbf{k}} \forall (p, a) \in \mathcal{P} \times \mathcal{A}$.

A3. We assume that the stochastic reward $r^{t}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}$ *is a 1-sub-Gaussian random variable with expected value* $\mathbb{E}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}[r^{t}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} \mid i^{t}_{\mathrm{a}}, i^{t}_{\mathfrak{m}(\mathrm{a})}] = A_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(i^{t}_{\mathrm{a}}, i^{t}_{\mathfrak{m}(\mathrm{a})})$.

Assumption A1. regarding knowledge of $U_{\mathrm{a},\perp}$ is standard in learning scenarios, where such utilities are typically set to zero [31]. Assumption A2. is made for notational convenience and is not restrictive, as the number of actions can always be bounded by the maximum across all agent pairs. Finally, Assumption A3., which requires the rewards to be 1-sub-Gaussian, is typical in the bandit literature.

3.2 Regret

We now introduce our incentive-aware learning objective, designed to analyze the cumulative behavior of learning algorithms. Specifically, we define the notion of *matching instability* to quantify the distance between a solution $(\mathfrak{m}, X)$ and the equilibrium $(\mathfrak{m}^{\star}, X^{\star})$ at each time step.

Definition 4 (Matching Instability)**.** *The matching instability* $\mathbf{MI}(m, X)$ *of a matching* $\mathfrak{m}$ *under mixed strategies* $X = \{x_{\mathrm{a}}\}_{\mathrm{a}\in\mathfrak{A}}$ *is defined as:*

$$\min_{s\in\mathbb{R}} \sum_{a\in\mathfrak{A}} s_a$$

subject to:

$$\min(V^{\star}_{p,a} - U_{p,\mathfrak{m}(p)}(x_p, x_{\mathfrak{m}(p)}) - s_p, V^{\star}_{a,p} - U_{a,\mathfrak{m}(a)}(x_a, x_{\mathfrak{m}(a)}) - s_a) \leq 0, \forall (p,a) \in \mathcal{P}\times\mathcal{A} \tag{C1}$$

$$U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) - U_{\mathrm{a},\perp} + s_{\mathrm{a}} \geq 0 \quad \forall \mathrm{a} \in \mathfrak{A} \tag{C2}$$

$$V^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{m(\mathrm{a})}) - s_{\mathrm{a}} \leq 0 \quad \forall \mathrm{a} \in \mathfrak{A} \tag{C3}$$

$$s_{\mathrm{a}} \geq 0 \quad \forall \mathrm{a} \in \mathfrak{A} \tag{C4}$$

Our metric extends the *subset instability* metric (Definition 2), originally proposed by [25, 26] for two-sided markets, by incorporating additional constraints. It can be interpreted as the minimal external payment that must be offered to retain users—or, equivalently, to stabilize the market by satisfying conditions similar to those in Definition 3.

More specifically, the subsidies are defined to satisfy constraint C2, which ensures individual rationality, and constraint C1, which eliminates blocking pairs under Nash strategies. In addition, constraint C3 guarantees that each agent, once appropriately subsidized, obtains a utility at least as high as the payoff they would receive by playing the NE of their matched game (see also Appendix A). We now describe properties of our instability measure that are important for learning.

Proposition 1. *The Matching Instability* $\mathbf{MI}(\mathfrak{m}, X)$ *of a solution* $(\mathfrak{m}, X)$ *satisfies the following properties:*

1. *It is always non-negative.*
2. *It is equal to zero if and only if both of the following hold:*
   - *A) For strategies* X*, each agent receives the same valuation as in the Nash equilibrium of their respective game, i.e.,* $U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) = V^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} \forall \mathrm{a} \in \mathfrak{A}$.
   - *B) The matching* $\mathfrak{m}$ *is stable under the strategies* X*.*
3. *In the case where each agent has a single action, i.e.,* $\mathbf{m} = \mathbf{k} = 1$*, the model reduces to the two-sided matching problem of Section 2.2, and the matching instability coincides with the subset instability of Definition 2.*
4. *In the case of a single agent on each side, i.e.,* $\mathcal{P} = \{p\}$ *and* $\mathcal{A} = \{a\}$*, where they are always matched, i.e.,* $m(p) = a$*, the model reduces to the ZSG setting. In this case, the matching instability* $\mathbf{MI}(\mathfrak{m}, X)$ *for strategies* $X = \{x_p, x_a\}$ *becomes:*

$$\mathbf{MI}(\mathfrak{m}, X) = \left| V^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \right| \forall \mathrm{a} \in \mathfrak{A}, \tag{6}$$

   which measures the deviation of (x_p, x_a) *from NE.*

Regret: We define the regret of an algorithm as the cumulative matching instability over a horizon T:

$$R_T = \sum_{t=1}^{T} \mathbf{MI}(\mathfrak{m}_t, X_t)$$

4 Algorithm

Now we discuss how the optimism-in-the-face-of-uncertainty principle can be applied to our setting to efficiently learn a matching equilibrium.

We introduce an Upper Confidence Bound (UCB) algorithm (Algorithm 1), where the agents both form their preferences and select their strategies based on the upper confidence estimates of the payoff matrix. Specifically, a central platform matches the agents $\mathrm{a} \in \mathfrak{A}$ according to their preferences. These

are derived from the upper confidence estimates of the payoff matrix for each agent. Subsequently, each matched pair selects their min-max strategies X based on the upper confidence payoffs associated with their respective assignment.

More formally, at each time step $t \in [T]$, each agent $\mathrm{a} \in \mathfrak{A}$ maintains an optimistic estimate of the payoff matrix $\bar{A}^t_{\mathrm{a},\mathrm{a}'}$ for playing a game with agent $\mathrm{a}' \in OS(\mathrm{a}) \cup \{\perp\}$. Based on these estimates, agents form their preferences $\bar{\pi}_\mathrm{a}$ according to the game values $\bar{V}^\star_{\mathrm{a},\mathrm{a}'}$ when matched with agent a', i.e., $\bar{\pi}_\mathrm{a} = \arg\mathrm{sort}_{\mathrm{a}'}\, \bar{V}^\star_{\mathrm{a},\mathrm{a}'}$[2]. Each agent then reports their preferences to the central platform, which produces a matching $\mathfrak{m}_t$ using the DA algorithm based on the estimated preferences.

Once matched, each agent $\mathrm{a} \in \mathfrak{A}$ observes their match $\mathfrak{m}_t(\mathrm{a})$ and samples actions i^t_a from their respective min-max strategies x^t_a according to the upper confidence payoff matrix $\bar{A}^t_{\mathrm{a},\mathfrak{m}_t(\mathrm{a})}$. After choosing their actions, the agents $\mathrm{a} \in \mathfrak{A}$ observe the actions of their matched partners $\mathfrak{m}_t(\mathrm{a})$, and receive a reward $r^t_{\mathrm{a},\mathfrak{m}_t(\mathrm{a})}$, with $r^t_{\mathfrak{m}_t(\mathrm{a}),\mathrm{a}} = -r^t_{\mathrm{a},\mathfrak{m}_t(\mathrm{a})}$. These observations are used to update the empirical mean of the payoff matrix $\hat{A}^t_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(i,j)$.

The upper confidence bounds of the payoff matrices for each agent pair is then computed as:

$$\bar{A}^t_{\mathrm{a},\mathrm{a}'}(i,j) = \hat{A}^t_{\mathrm{a},\mathrm{a}'}(i,j) + \sqrt{\frac{2\log(1/\delta)}{1 \vee n^t_{\mathrm{a},\mathrm{a}'}(i,j)}} \quad \forall\, \mathrm{a} \in \mathfrak{A},\ \mathrm{a}' \in OS(\mathrm{a}),$$

where $n^t_{\mathrm{a},\mathrm{a}'}(i,j)$ is the number of times agents (a,a') played actions (i,j) up to time t.

The complete algorithm, with explicit reference to the agent on each side, is presented in Algorithm 1.

Algorithm 1 UCB-MG: UCB-Based Algorithm for Matching with Adaptive Agents in Games

Input: T, δ, $\mathcal{P}$, $\mathcal{A}$, $\mathcal{I}$, $\mathcal{J}$,
$n_{p,a}(i,j) = 0\ \forall\ (p,a,i,j) \in \mathcal{P} \times \mathcal{A} \times \mathcal{I} \times \mathcal{J}$
$\hat{A}_{p,a} = 0\ \forall\ (p,a) \in \mathcal{P} \times \mathcal{A}$
for $t \in 1, \cdots, T$ **do**
  $\bar{A}_{p,a}(i,j) = \hat{A}^t_{p,a}(i,j) + \sqrt{\frac{2\log(1/\delta)}{1 \vee n_{p,a}(i,j)}}\ \forall\ (p,a,i,j) \in \mathcal{P} \times \mathcal{A} \times \mathcal{I} \times \mathcal{J}$
  $\bar{A}_{a,p}(j,i) = -\hat{A}^t_{p,a}(i,j) + \sqrt{\frac{2\log(1/\delta)}{1 \vee n_{p,a}(i,j)}}\ \forall\ (p,a,i,j) \in \mathcal{P} \times \mathcal{A} \times \mathcal{I} \times \mathcal{J}$
  $\bar{V}_{p,a} = max_{x \in \Delta_x} min_{y \in \Delta_y} x^T \bar{A}_{p,a} y\ \forall\ (p,a) \in \mathcal{P} \times \mathcal{A}$
  $\bar{V}_{a,p} = max_{y \in \Delta_y} min_{x \in \Delta_x} y^T \bar{A}_{a,p} x\ \forall\ (p,a) \in \mathcal{P} \times \mathcal{A}$
  $\hat{\pi}_\mathrm{a} = \arg\mathrm{sort}_i\, \bar{V}_{\mathrm{a},i}\ \forall \mathrm{a} \in \mathfrak{A}$
  $\mathfrak{m}_t = GS(\{\hat{\pi}\}_{\mathrm{a} \in \mathfrak{A}})$
  for $(p,a) \in \mathfrak{m}_t$ **do**
    $x = arg\max_{x \in \Delta_x} min_{y \in \Delta_y} x^T \bar{A}_{p,a} y$
    $y = arg\max_{y \in \Delta_y} min_{x \in \Delta_x} y^T \bar{A}_{a,p} x$
    sample actions $i \sim x$ and $j \sim y$
    sample reward $r^t_{p,a}(i,j)$ and update $\hat{A}^t_{p,a}(i,j)$
    $n_{p,a}(i,j) = n_{p,a}(i,j) + 1$
  end for
end for

5 Results

We now show that Algorithm 1 gradually leads the agents to converge toward matchings equilibrium. In particular, we prove that the cumulative matching instability incurred over time grows sublinearly with the horizon T.

Theorem 1. *Let* $|\ \mathcal{P}\ | = \mathbf{p}$, $|\ \mathcal{A}\ | = \mathbf{a}$, $|\ \mathcal{I}\ | = \mathbf{m}$ *and* $|\ \mathcal{J}\ | = \mathbf{k}$. *When Algorithm 1 is run with parameter* $\delta = 1/(4T^2\mathbf{p}^2\mathbf{a}^2\mathbf{mk})$*, the expected regret* R_T *under horizon* T *is bounded by:*

$$\mathbb{E}[R_T] = \mathbb{E}[\sum_{t=1}^{T} \mathbf{MI}(m_t, X_t)] \leq \tilde{O}\left(\sqrt{T\mathbf{kmpa}}\right).$$

[2]Note that the values of the game $\bar{V}^\star_{\mathrm{a},\mathrm{a}'}$ can be computed without knowledge of the opponent's strategy

This regret bound recovers and generalizes several existing results in the literature. In particular, when each agent has access to only a single action (i.e., $\mathbf{m} = \mathbf{k} = 1$), the problem reduces to a classical two-sided matching scenario, and the bound simplifies to $\tilde{O}(\sqrt{T\mathbf{pa}})$, similar to the bound of Theorem 6.6 of [26]. Conversely, when there is only one agent on each side (i.e., $\mathbf{p} = \mathbf{a} = 1$), the regret becomes $\tilde{O}(\sqrt{T\mathbf{mk}})$, which corresponds to the bound established in [36] for two-player zero-sum games under bandit feedback.

5.1 Simulations

We now present a set of simulations that complement our theoretical results, in which we empirically compare three different scenarios. For each scenarios, we generate 50 random instances by independently sampling the entities payoff matrices of the ZSG from a normal distribution, i.e., $A_{p,a}(i,j) \sim \mathcal{N}(0,1)\ \forall (p,a,i,j) \in \mathcal{P} \times \mathcal{A} \times \mathcal{I} \times \mathcal{J}$. The code for reproducing all the experiments is provided in the supplementary material[3] while additional details are provided in Appendix D.

The baseline setting is the **Self-play** scenario, in which all agents act according to Algorithm 1. We compare this with two cases where the agents in $\mathcal{A}$ have additional information. In **Nash-response**, $\mathcal{P}$ use Algorithm 1, but agents in $\mathcal{A}$ from their prefereces and select strategies according the NE of the true payoff matrices A. The third is the **Best-response** scenario, where agents $a \in \mathcal{A}$ know both the true payoff matrices and the strategies each agent in $\mathcal{P}$ selects, i.e., $x_{p,a} \forall p \in \mathcal{P}$. Here, agents $a \in \mathcal{A}$ form their preferences and choose strategies to exploit their match based on best responses. Specifically, each agent computes their best-response strategy as $x_{a,p} = \arg\max_x U_{a,p}(x, x_{p,a})$, and form their preferences according to the resulting payoffs: $\pi_a = \arg\text{sort}_{p \in \mathcal{P}} U_{a,p}(x_{a,p}, x_{p,a})$.

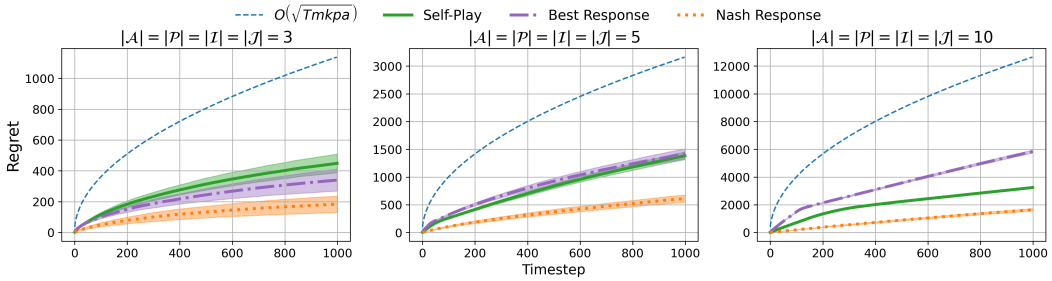

Figure 1: Regret of the different settings, including the theoretical bound. Lines represent the average performance of the algorithm, and shaded areas indicate the standard deviation across runs.

Figure 1 compares the regret for the different settings for varying numbers of agents and actions, along with the theoretical bound of $O\left(\sqrt{T\mathbf{apmk}}\right)$ (dashed blue), where we omit the logarithmic term for clarity. In all experimental settings, we observe sublinear regret growth, indicating that the system progressively learns a *matching equilibrium*. The **Nash-response** (dotted orange) scenario yields the lowest regret, as one side of the market has full knowledge of the payoff matrices and acts according to NE. In addition, the **Best-response** (dash-dotted purple) scenario initially outperforms **Self-play** (solid green), but as the number of agents and actions increases, its performance deteriorates. This is due to agents in $\mathcal{A}$ exploiting their matched partners in $\mathcal{P}$, which introduces greater complexity into the learning process.

6 Related work

Learning in games. There is a substantial body of research on learning in games, where agents operate with limited information in uncertain environments [35, 10]. A foundational result shows that no-regret adversarial algorithms—whether under full-information (expert advice) or partial-information (bandit feedback)— converge toward certain equilibrium [19, 29, 16, 20]. Further research has focused on improving these results. In the *full-information* setting, tighter bounds have been obtained by leveraging structural properties of the game, such as in two-player zero-sum games [13, 37] and multiplayer general-sum games [43, 11, 14, 15, 1, 2]. In the *bandit feedback* setting,

[3]To be published at `www.github.com/to/be/published`

progress has been made by deriving instance-dependent bounds [32, 22], or by considering alternative feedback models [37, 46, 36, 8].

Our work is closely related to O'Donoghue et al. [36], who study a bandit setting where agents observe their opponents' actions. They propose UCB-style algorithms that empirically outperform adversarial methods like EXP3 [4] by leveraging this feedback. We adopt a similar information setting (Section 3.1), and design a UCB-based algorithm with sublinear regret tailored to our setting. Finally, the setting of social games bears some similarities to zero-sum polymatrix games [7], where each agent is engaged in a game with a fixed number of agents.

Learning in two-sided matching. Learning in two-sided matching has recently gained attention, with the initial work of Das and Kamenica [12] framing it as a multi-armed bandit (MAB) problem and providing empirical evidence on the performance of proposed algorithms. The follow-up work of Liu et al. [30] introduced the notions of Player Stable Optimal and Pessimal regret for agents in the market, and highlighted fundamental challenges in achieving sublinear regret using UCB-like algorithms in a centralized setting, where a platform matches agents at each time step. Subsequently, Jagadeesan et al. [25, 26] established sublinear regret by proposing an alternative regret notion of subset instability, defined as the minimum subsidies required to stabilize the market. In our work, we extend this regret in the setting where matched agents engaged in ZSG. Related concepts to subset instability have been explored in coalitional games, notably the *Cost of Stability* [5], which is used to characterize games with an empty core, as well as the notion of the ϵ-core [42].

Later studies have also investigated the sample complexity of learning stable matchings within the probably approximately correct (PAC) framework [3, 21]. Another line of research explores decentralized settings, where agents act independently in the market [31, 27, 41, 6], as well as alternative matching models with transferable utilities [25, 9]. A conceptually related line of work to ours is [33], which considers markets with time-varying preferences, motivating the concept of adaptive agents. In contrast, our setting involves agents that learn through repeated interactions in ZSGs—that to the best of our knowledge, has not been previously explored.

7 Discussion

We studied the problem of learning equilibria from bandit feedback in a generalized two-sided matching market, where interactions between matched agents follow a zero-sum game. We show that an algorithm based on the principle of optimism in the face of uncertainty achieves sublinear regret, demonstrating that *matching equilibria* can be learned efficiently in this setting.

Our study opens several interesting research directions. First, we would like to highlight that our learning objective captures the process of learning an equilibrium, which is generally considered a *fair* (or desirable) outcome, as it aligns the interests of the agents in the market. However, this objective does not necessarily ensure equitable outcomes of the agent, since the learning process may disproportionately benefit specific subsets of agents or a side of the market. Therefore, careful consideration is required when applying this framework. In certain contexts, a more desirable metric could involve a notion of regret that is bounded for each individual agent, similar to [30], where regret is defined by comparing the cumulative rewards to those achieved under equilibrium. Further investigations utilizing real-world data could provide valuable insights into how our approach and the associated metrics translate to practical settings beyond controlled simulations.

Additional research directions include exploring alternative models and information settings that more accurately reflect real-world constraints. These include scenarios with uncoupled dynamics [13], where agents cannot observe the actions of their match, as well as decentralized settings [31], in which agents act independently without coordination through a central platform. Another important direction involves learning equilibria in adversarial environments, such as adversarially changing games [8]. Extending the framework to general-sum games or matching models involving more than two players also introduces further challenges for learning equilibrium; in such settings, designing approximate notions of matching equilibrium and analyzing their integration into the learning process remain promising avenues for future work.

References

[1] Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 736–749, 2022.

[2] Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. Uncoupled learning dynamics with o(\log t) swap regret in multiplayer games. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 3292–3304. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/15d45097f9806983f0629a77e93ee60f-Paper-Conference.pdf`.

[3] Andreas Athanasopoulos, Anne-Marie George, and Christos Dimitrakakis. Probably correct optimal stable matching for two-sided markets under uncertainty. *arXiv preprint arXiv:2501.03018*, 2025.

[4] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th annual foundations of computer science*, pages 322–331. IEEE, 1995.

[5] Yoram Bachrach, Edith Elkind, Reshef Meir, Dmitrii Pasechnik, Michael Zuckerman, Jörg Rothe, and Jeffrey S Rosenschein. The cost of stability in coalitional games. In *Algorithmic Game Theory: Second International Symposium, SAGT 2009, Paphos, Cyprus, October 18-20, 2009. Proceedings 2*, pages 122–134. Springer, 2009.

[6] Soumya Basu, Karthik Abinav Sankararaman, and Abishek Sankararaman. Beyond $log^2(t)$ regret for decentralized bandits in matching markets. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 705–715, Virtual, 18–24 Jul 2021. PMLR. URL `https://proceedings.mlr.press/v139/basu21a.html`.

[7] Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos Papadimitriou. Zero-sum polymatrix games: A generalization of minmax. *Mathematics of Operations Research*, 41(2): 648–655, 2016.

[8] Adrian Rivera Cardoso, Jacob Abernethy, He Wang, and Huan Xu. Competing against Nash equilibria in adversarially changing zero-sum games. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 921–930. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/cardoso19a.html`.

[9] Sarah H. Cen and Devavrat Shah. Regret, stability & fairness in matching markets with bandit learners. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 8938–8968, Virtual, 28–30 Mar 2022. PMLR. URL `https://proceedings.mlr.press/v151/cen22a.html`.

[10] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[11] Xi Chen and Binghui Peng. Hedging in games: Faster convergence of external and swap regrets. *Advances in Neural Information Processing Systems*, 33:18990–18999, 2020.

[12] Sanmay Das and Emir Kamenica. Two-sided bandits and the dating market. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, page 947–952, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

[13] Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 235–254. SIAM, 2011.

[14] Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems*, 34:27604–27616, 2021.

[15] Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-optimal no-regret learning dynamics for general convex games. *Advances in Neural Information Processing Systems*, 35:39076–39089, 2022.

[16] Dean P Foster and Rakesh V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

[17] D. Gale and L. S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2312726.

[18] Guillaume Haeringer and Myrna Wooders. Decentralized job matching. *International Journal of Game Theory*, 40:1–28, 02 2011. doi: 10.1007/s00182-009-0218-x.

[19] James Hannan. Approximation to bayes risk in repeated play. *Contributions to the Theory of Games*, 3(2):97–139, 1957.

[20] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

[21] Hadi Hosseini, Sanjukta Roy, and Duohan Zhang. Putting gale & shapley to work: Guaranteeing stability through learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 69043–69068. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7fc0aac0d700873ac520af9b05e3f2b6-Paper-Conference.pdf.

[22] Shinji Ito, Haipeng Luo, Taira Tsuchiya, and Yue Wu. Instance-dependent regret bounds for learning two-player zero-sum games with bandit feedback. *arXiv preprint arXiv:2502.17625*, 2025.

[23] Matthew O Jackson and Alison Watts. Equilibrium existence in bipartite social games: A generalization of stable matchings. *Economics Bulletin*, 3(12):1–8, 2008.

[24] Matthew O Jackson and Alison Watts. Social games: Matching and the play of finitely repeated games. *Games and Economic Behavior*, 70(1):170–191, 2010.

[25] Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael Jordan, and Jacob Steinhardt. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34:3323–3335, 2021.

[26] Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael I. Jordan, and Jacob Steinhardt. Learning equilibria in matching markets from bandit feedback. *CoRR*, abs/2108.08843, 2021. URL https://arxiv.org/abs/2108.08843.

[27] Fang Kong, Junming Yin, and Shuai Li. Thompson sampling for bandit learning in matching markets. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3164–3170, Vienna, Austria, 7 2022. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2022/439. URL https://doi.org/10.24963/ijcai.2022/439. Main Track.

[28] Tor Lattimore and Csaba Szepesvari. Bandit algorithms. 2017. URL https://tor-lattimore.com/downloads/book/book.pdf.

[29] Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

[30] Lydia T. Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1618–1628. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/liu20c.html.

[31] Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22(211):1–34, 2021.

[32] Arnab Maiti, Kevin Jamieson, and Lillian Ratliff. Instance-dependent sample complexity bounds for zero-sum matrix games. In *International Conference on Artificial Intelligence and Statistics*, pages 9429–9469. PMLR, 2023.

[33] Deepan Muthirayan, Chinmay Maheshwari, Pramod Khargonekar, and Shankar Sastry. Competing bandits in time varying matching markets. In Nikolai Matni, Manfred Morari, and George J. Pappas, editors, *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, volume 211 of *Proceedings of Machine Learning Research*, pages 1020–1031. PMLR, 15–16 Jun 2023. URL https://proceedings.mlr.press/v211/muthirayan23a.html.

[34] John F Nash. Non-cooperative games. In *The Foundations of Price Theory Vol 4*, pages 329–340. Routledge, 2024.

[35] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, USA, 2007. ISBN 0521872820.

[36] Brendan O'Donoghue, Tor Lattimore, and Ian Osband. Matrix games with bandit feedback. In *Uncertainty in Artificial Intelligence*, pages 279–289. PMLR, 2021.

[37] Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013.

[38] Julia Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54(2):296–301, 1951. ISSN 0003486X, 19398980. URL http://www.jstor.org/stable/1969530.

[39] Alvin E Roth. The evolution of the labor market for medical interns and residents: a case study in game theory. *Journal of political Economy*, 92(6):991–1016, 1984.

[40] Alvin E. Roth and Marilda A. Oliveira Sotomayor. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Econometric Society Monographs. Cambridge University Press, 1990.

[41] Abishek Sankararaman, Soumya Basu, and Karthik Abinav Sankararaman. Dominate or delete: Decentralized competing bandits in serial dictatorship. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1252–1260, Virtual, 13–15 Apr 2021. PMLR. URL https://proceedings.mlr.press/v130/sankararaman21a.html.

[42] Lloyd S Shapley and Martin Shubik. Quasi-cores in a monetary economy with nonconvex preferences. *Econometrica: Journal of the Econometric Society*, pages 805–827, 1966.

[43] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems*, 28, 2015.

[44] J v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

[45] John Von Neumann. Morgenstern, 0. *Theory of games and economic behavior*, 3, 1944.

[46] Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

A Remark on constrain C3 of Definition 4

In this section, we provide a remark supporting constraint C3 in Definition 4 of matching instability.

Specifically, we focus on the case of a single pair of agents and demonstrate that if the condition holds for both agents in the pair, their utilities coincide with the value of the game, $V^\star$.

Remark 1. *Consider a pair of agents (p, a) matched under a matching $\mathfrak{m}$, who select strategies x and y, respectively. If the condition C3 in Definition 4 holds for both agents ($s_p = s_a = 0$), i.e.,*

$$V^\star_{p,a} \leq U_{p,a}(x, y) \quad \textit{and} \quad V^\star_{a,p} \leq U_{a,p}(y, x),$$

then the utility that the agents receives is equal to the value of the game i.e $U_{p,a}(x, y) = V^\star$.

Proof. Since the interaction is a ZSG, we have the properties:

$$V^\star_{a,p} = -V^\star_{p,a} \quad \text{and} \quad U_{a,p}(y, x) = -U_{p,a}(x, y).$$

So we can expand the condition for agent a, as follows:

$$\begin{aligned} V^\star_{a,p} &\leq U_{a,p}(y, x) \\ -V^\star_{p,a} &\leq -U_{p,a}(x, y) \\ U_{p,a}(x, y) &\leq V^\star_{p,a}. \end{aligned}$$

Combined with the condition for agent p, $V^\star_{p,a} \leq U_{p,a}(x, y)$, we have that

$$V^\star_{p,a} \leq U_{p,a}(x, y) \leq V^\star_{p,a},$$

so the equality holds:

$$U_{p,a}(x, y) = V^\star_{p,a}.$$

□

We also highlight that having the same value as a Nash equilibrium does not guarantee that the strategy profile is a Nash equilibrium. However, the convergence of iterative procedures to the value of the game is commonly used to analyze algorithms in the game theory literature [38, 36, 13].

Consider for example the game:

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

Here, the top-left action profile is a Nash equilibrium (with value 0), but the bottom-right action profile is not a NE.

B Proof of Proposition 1

For completeness and ease of reading, we restate Proposition 1 along with its proof.

Proposition 1. *The Matching Instability $\mathbf{MI}(\mathfrak{m}, X)$ of a solution $(\mathfrak{m}, X)$ satisfies the following properties:*

1. *It is always non-negative.*
2. *It is equal to zero if and only if both of the following hold:*
   - *A) For strategies X, each agent receives the same valuation as in the Nash equilibrium of their respective game, i.e., $U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_\mathrm{a}, x_{\mathfrak{m}(\mathrm{a})}) = V^\star_{\mathrm{a},\mathfrak{m}(\mathrm{a})} \forall \mathrm{a} \in \mathfrak{A}$.*
   - *B) The matching $\mathfrak{m}$ is stable under the strategies X.*
3. *In the case where each agent has a single action, i.e., $\mathbf{m} = \mathbf{k} = 1$, the model reduces to the two-sided matching problem of Section 2.2, and the matching instability coincides with the subset instability of Definition 2.*
4. *In the case of a single agent on each side, i.e., $\mathcal{P} = \{p\}$ and $\mathcal{A} = \{a\}$, where they are always matched, i.e., $m(p) = a$, the model reduces to the ZSG setting. In this case, the matching instability $\mathbf{MI}(\mathfrak{m}, X)$ for strategies $X = \{x_p, x_a\}$ becomes:*

$$\mathbf{MI}(\mathfrak{m}, X) = \left| V^\star_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_\mathrm{a}, x_{\mathfrak{m}(\mathrm{a})}) \right| \forall \mathrm{a} \in \mathfrak{A}, \tag{7}$$

   which measures the deviation of (x_p, x_a) from NE.

Proof.
•*Case 1: Non-negativity.*
By definition, the subset matching instability is always non-negative due to constraint C4.

•*Case 2:.*
We begin by showing that matching instability $\mathbf{MI}$ equals zero i.e., $s_{\mathrm{a}} = 0$ for all $\mathrm{a} \in \mathfrak{A}$, then the matching $\mathfrak{m}$ is stable under the stategies X, and the utlities of the agents are the same with NE.

We start by showing that the satisfaction of the constraint C3 ensures $(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})})$ has the same valuation with a Nash equilibrium for an arbitrary pair of matched agents a and $\mathfrak{m}(\mathrm{a})$. To see this consider the constrain C3 for the agents a where we obtain that:

$$V^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} \leq U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{m(\mathrm{a})}) \tag{8}$$

while the constrain for the agent $\mathfrak{m}(\mathrm{a})$ is:

$$V^{\star}_{\mathfrak{m}(\mathrm{a}),\mathrm{a}} \leq U_{\mathfrak{m}(\mathrm{a}),\mathrm{a}}(x_{\mathfrak{m}(\mathrm{a})}, x_{\mathrm{a}}) \tag{9}$$

Expanding the inequality from 8 using the bilinear form we have that:

$$x^{\star T}_{\mathrm{a}} A_{\mathrm{a},\mathfrak{m}(\mathrm{a})} x^{\star}_{\mathfrak{m}(\mathrm{a})} \leq x^{T}_{\mathrm{a}} A_{\mathrm{a},\mathfrak{m}(\mathrm{a})} x_{\mathfrak{m}(\mathrm{a})} \tag{E1}$$

Similarly, for the second inequality 9 :

$$\begin{aligned}
x^{\star T}_{\mathfrak{m}(\mathrm{a})} A_{\mathfrak{m}(\mathrm{a}),\mathrm{a}} x^{\star}_{\mathrm{a}} &\leq x^{T}_{\mathfrak{m}(\mathrm{a})} A_{\mathfrak{m}(\mathrm{a}),\mathrm{a}} x_{\mathrm{a}} \\
x^{\star T}_{\mathfrak{m}(\mathrm{a})} (-A_{\mathrm{a},m(\mathrm{a})})^{T} x^{\star}_{\mathrm{a}} &\leq x^{T}_{\mathfrak{m}(\mathrm{a})} (-A_{\mathrm{a},\mathfrak{m}(\mathrm{a})})^{T} x_{\mathrm{a}} \\
x^{\star T}_{\mathfrak{m}(\mathrm{a})} A^{T}_{\mathrm{a},m(\mathrm{a})} x^{\star}_{\mathrm{a}} &\geq x^{T}_{\mathfrak{m}(\mathrm{a})} A^{T}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} x_{\mathrm{a}} \\
x^{\star T}_{\mathrm{a}} A_{\mathrm{a},m(\mathrm{a})} x^{\star}_{\mathfrak{m}(\mathrm{a})} &\geq x^{T}_{\mathrm{a}} A_{\mathrm{a},m(\mathrm{a})} x_{\mathfrak{m}(\mathrm{a})}
\end{aligned} \tag{E2}$$

Thus, from E1 and E2 we obtain:

$$\begin{aligned}
x^{\star T}_{\mathrm{a}} A_{\mathrm{a},m(\mathrm{a})} x^{\star}_{\mathfrak{m}(\mathrm{a})} \leq & x^{T}_{\mathrm{a}} A_{\mathrm{a},\mathfrak{m}(\mathrm{a})} x_{\mathfrak{m}(\mathrm{a})} \leq x^{\star T}_{\mathrm{a}} A_{\mathrm{a},m(\mathrm{a})} x^{\star}_{\mathfrak{m}(\mathrm{a})} c \Rightarrow \\
& U_{\mathrm{a},m(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) = V^{\star}_{\mathrm{a},m(\mathrm{a})}
\end{aligned}$$

which implies that $(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})})$ has the same valuation with the Nash equilibrium for the pair $(\mathrm{a}, \mathfrak{m}(\mathrm{a}))$.

Now, the constraints C1 and C2 imply that the matching $\mathfrak{m}$ is stable under the strategies X, as it has no blocking pair and is individually rational.

Conversely, if we have a matching equilibrium $(\mathfrak{m}, X)$, then $s_{\mathrm{a}} = 0 \; \forall \mathrm{a} \in \mathfrak{A}$ is a feasible solution to the optimization constraints, and we have $\mathbf{MI}(\mathfrak{m}, X) = 0$, since the metric is always non-negative, that completing the proof.

•*Case 3: Single Action for Each Agent.*
In the case where each agent has a single action, the constraint C3 of Definition 4 is trivially satisfied, while the constraints C1 and C2 reduce to those in Definition 2 by omitting the strategy indices.

•*Case 4: Single pair of agents.*
We now consider the case of a single pair of agents p and a who always match, i.e., $\mathfrak{m}(p) = a$. This situation can occur in our matching model when agents strictly prefer matching over staying single, that is, when $U_{p,\perp} = U_{a,\perp} = -1$.

In this setting, we obtain a zero-sum game where we denote the strategies $X = \{x_p, x_a\}$ for the agents, where the matching instability becomes:

$$\min_{s \in \mathbb{R}} \sum_{a \in \mathfrak{A}} s_a$$

subject to:

$$V^{\star}_{p,a} - U_{p,a}(x_p, x_a) - s_p \leq 0 \tag{10}$$

$$V^{\star}_{a,p} - U_{a,p}(x_a, x_p) - s_a \leq 0 \tag{11}$$

$$s_{\mathrm{a}} \geq 0 \quad \forall \mathrm{a} \in \mathfrak{A}. \tag{12}$$

as constraints C1 and C2 of Definition 4 are trivially satisfied in this case. More specifically, the blocking pairs constraint C1 holds, as there is only a single pair (p, a) of agents. Constraint C2, which enforces individual rationality, is also satisfied because both agents always prefer to be matched rather than remain single: $U_{p,\perp} = U_{a,\perp} = -1 \leq A$.

Using the symmetric properties of the ZSG, $V^\star_{p,a} = -V^\star_{a,p}$ and $U_{p,a}(x_p, x_a) = -U_{a,p}(x_a, x_p)$, the constraints in (10) and (11) can be written as:

$$-s_a \leq V^\star_{p,a} - U_{p,a}(x_p, x_a) \leq s_p \quad \textit{and} \quad -s_p \leq V^\star_{a,p} - U_{a,p}(x_a, x_p) \leq s_a$$

or equivalently:

$$\mathbf{MI}(\mathfrak{m}, X) = \min_{s \in \mathbb{R}^{\mathcal{A}}_{\geq 0}} \left\{ \sum_{\mathrm{a} \in \mathfrak{A}} s_{\mathrm{a}} \;\middle|\; \left| V^\star_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \right| \leq \max_{\mathrm{a} \in \mathfrak{A}}(s_{\mathrm{a}}) \quad \forall \mathrm{a} \in \mathfrak{A} \right\}$$

In addition, note that only one of the values s_p and s_a can be positive, because of properties of the zero-sum game. So we let, $\mathfrak{s} \in \mathbb{R}_{\geq \mathfrak{o}}$ be the positive subsidy we have that:

$$\mathbf{MI}(\mathfrak{m}, X) = \min_{\mathfrak{s} \in \mathbb{R}_{\geq 0}} \left\{ \mathfrak{s} \;\middle|\; \left| V^\star_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \right| \leq \mathfrak{s} \quad \forall \mathrm{a} \in \mathfrak{A} \right\}$$

and because the values $\left| V^\star_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \right|$ are equal for every agent $\mathrm{a} \in \mathfrak{A}$ we have:

$$\mathbf{MI}(\mathfrak{m}, X) = \left| V^\star_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \right| \quad \forall \mathrm{a} \in \mathfrak{A} \tag{13}$$

□

C Proof of Theorem 1

In this section, we provide an upper bound for Algorithm 1 in the self-play scenario, where all agents follow the principle of optimism in the face of uncertainty, as described in Section 4.

More specifically, each agent $\mathrm{a} \in \mathfrak{A}$ maintains an upper confidence estimate $\bar{A}_{\mathrm{a},\mathrm{a}'}$ for matching with an agent $\mathrm{a}' \in OS(\mathrm{a})$ of the other side of the market. At each step t, agents estimate their preferences $\hat{\pi}_{\mathrm{a}}$ by sorting the estimated values of the different games, i.e., $\hat{\pi}_{\mathrm{a}} = \arg\mathrm{sort}_{\mathrm{a}' \in OS(\mathrm{a})} \bar{V}^\star_{\mathrm{a},\mathrm{a}'}$, and report them to the platform. The platform selects matchings $\mathfrak{m}_t$ using the DA algorithm based on the agents' estimated preferences. Then, the matched agents select the min-max strategies with respect to the upper confidence payoff matrix of the game $\bar{A}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}$.

Next, we present Proposition 2, which describes key properties of our Algorithm 1 that we will use in the proof of our main theorem.

Proposition 2. *Let $\mathfrak{m}_t$ be the matching and $X = \{x_{\mathrm{a}}\}_{\mathrm{a} \in \mathfrak{A}}$ the strategies of the agents under Algorithm 1 at time step t. Then the following statements hold:*

(S1) $\bar{V}^\star_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - \bar{U}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \leq 0 \; \forall \mathrm{a} \in \mathfrak{A}$

(S2) $\min(\bar{V}^\star_{p,a} - \bar{U}_{p,\mathfrak{m}(p)}(x_p, x_{\mathfrak{m}(p)}), \bar{V}^\star_{a,p} - \bar{U}_{a,\mathfrak{m}(a)}(x_a, x_{\mathfrak{m}(a)})) \leq 0 \quad \forall (p,a) \in \mathcal{P} \times \mathcal{A}$

where $\bar{V}^\star_{\mathrm{a},\mathrm{a}'} = \bar{U}_{\mathrm{a},\mathrm{a}'}(x^\star_{\mathrm{a}}, x^\star_{\mathrm{a}'}) = x^{\star T}_{\mathrm{a}} \bar{A}_{\mathrm{a},\mathrm{a}'} x^\star_{\mathrm{a}'}$, is the value of the game for the pair of agents $(\mathrm{a}, \mathrm{a}')$, with respect to the upper confidence payoff matrix $\bar{A}_{\mathrm{a},\mathrm{a}'}$.

Proof. At each step t, the platform selects a matching $\mathfrak{m}_t$ and the matching pairs $(p,a) \in \mathfrak{m}_t$ select strategies x^t_p, x^t_a. In the following, we omit the indexing for the time step t for the convenience of notation.

•*Proof of statement S1:.*
More specifically, each agent selects their min-max strategy with respect to their upper confidence payoff matrix. For a pair (p,a), this results in:

1. Agent p selects $x_p = x^\star_p = \arg\max_x \min_y x^T \bar{A}_{p,a} y$. We denote by $(x^\star_p, y^\star_a)$ the Nash equilibrium of the game with payoff matrix $\bar{A}_{p,a}$, and by $\bar{V}^\star_{p,a} = x^{\star T}_p \bar{A}_{p,a} y^\star_a$ the value of the game.

2. Accordingly, agent a selects $x_a = x_a^\star = \arg\max_x \min_y x^T \bar{A}_{a,p} y$, with $(x_a^\star, y_p^\star)$ being the Nash equilibrium of the game with payoff matrix $\bar{A}_{a,p}$. The value of the game for agent a is $\bar{V}_{a,p}^\star = x_a^{\star T} \bar{A}_{a,p} y_p^\star$

Note that the pair of strategies $(x_p = x_p^\star, x_a = x_a^\star)$ selected by the agents is not necessarily equal to the Nash equilibrium strategy profiles $(x_p^\star, y_a^\star)$ and $(x_a^\star, y_p^\star)$ of their respective games.

Instead, using the property of the ZSG, where $U(x^\star, y) \leq U(x^\star, y^\star) \forall y \in \Delta_y$, we have that:

$$\bar{V}_{p,a}^\star - \bar{U}_{p,a}(x_p, x_a) = \bar{U}_{p,a}(x_p^\star, y_a^\star) - \bar{U}_{p,a}(x_p, x_a) = \bar{U}_{p,a}(x_p^\star, y_a^\star) - \bar{U}_{p,a}(x_p^\star, x_a) \leq 0$$

as $x_p = x_p^\star$, and similarly for agent a, we have:

$$\bar{V}_{a,p}^\star - \bar{U}_{a,p}(x_a, x_p) = \bar{U}_{a,p}(x_a^\star, y_p^\star) - \bar{U}_{a,p}(x_a^\star, x_p) \leq 0$$

So, for every agent $\mathrm{a} \in \mathfrak{A}$, using the notation of the matching $\mathfrak{m}$, we have:

$$\bar{V}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}^\star - \bar{U}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \leq 0 \quad \forall \mathrm{a} \in \mathfrak{A}.$$

•*proof of statement S2.*
Now we prove that the solution $(\mathfrak{m}, X)$ has no blocking pair, with respect to the upper confidence estimate of the payoff matrices $\bar{A}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}$, i.e.:

$$\min(\bar{V}_{p,a}^\star - \bar{U}_{p,\mathfrak{m}(p)}(x_p, x_{\mathfrak{m}(p)}), \bar{V}_{a,p}^\star - \bar{U}_{a,\mathfrak{m}(\mathrm{a})}(x_a, x_{\mathfrak{m}(\mathrm{a})})) \leq 0 \quad \forall (p,a) \in \mathcal{P} \times \mathcal{A}$$

More specifically, we have that:

$$\begin{aligned}&\min(\bar{V}_{p,a}^\star - \bar{U}_{p,\mathfrak{m}(p)}(x_p, x_{\mathfrak{m}(p)}), \bar{V}_{a,p}^\star - \bar{U}_{a,\mathfrak{m}(\mathrm{a})}(x_a, x_{\mathfrak{m}(\mathrm{a})})) \leq \\ &\min(\bar{V}_{p,a}^\star - \bar{V}_{p,\mathfrak{m}(p)}^\star, \bar{V}_{a,p}^\star - \bar{V}_{a,\mathfrak{m}(\mathrm{a})}^\star) \leq 0 \quad \forall (p,a) \in \mathcal{P} \times \mathcal{A}\end{aligned}$$

The first inequality follows from the statement S1, where $-\bar{U}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \leq -\bar{V}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}^\star \ \forall \mathrm{a} \in \mathfrak{A}$. The second inequality holds, from the fact that in each step, the central platform selects a stable matching $\mathfrak{m}$, by running the DA algorithm using preferences that are formed according to the upper confidence values of the games $\bar{V}^\star$, so there are no blocking pairs. □

C.1 Per step regret

We now proceed with the regret analysis of the algorithm. We begin by examining the regret at a single time step t under the event that true utilities lies between the confidence intervals of the estimated values, as formalized in the following lemma.

First, let $\bar{A}_{\mathrm{a},\mathrm{a}'}^t(i,j)$ and $\underline{\mathrm{A}}_{\mathrm{a},\mathrm{a}'}^t(i,j)$ denote the upper and lower confidence bounds, respectively, for the entries of the true payoff matrix $A_{\mathrm{a},\mathrm{a}'}$, as defined below:

$$\bar{A}_{\mathrm{a},\mathrm{a}'}^t(i,j) = \hat{A}_{\mathrm{a},\mathrm{a}'}^t(i,j) + \sqrt{\frac{2\log(1/\delta)}{1 \vee n_{\mathrm{a},\mathrm{a}'}^t(i,j)}} \quad \forall \mathrm{a} \in \mathfrak{A},\ \mathrm{a}' \in OS(\mathrm{a}),$$

$$\underline{\mathrm{A}}_{\mathrm{a},\mathrm{a}'}^t(i,j) = \hat{A}_{\mathrm{a},\mathrm{a}'}^t(i,j) - \sqrt{\frac{2\log(1/\delta)}{1 \vee n_{\mathrm{a},\mathrm{a}'}^t(i,j)}} \quad \forall \mathrm{a} \in \mathfrak{A},\ \mathrm{a}' \in OS(\mathrm{a}).$$

Lemma 1 (Per-step Regret)**.** *Consider the event where the entries of the true payoff matrices lie within the confidence intervals $C_{\mathrm{a},\mathrm{a}'}(i,j) = [\underline{A}_{\mathrm{a},\mathrm{a}'}(i,j), \bar{A}_{\mathrm{a},\mathrm{a}'}(i,j)]$, i.e.:*

$$\mathcal{E}_t = \left\{ A_{\mathrm{a},\mathrm{a}'}(i,j) \in C_{\mathrm{a},\mathrm{a}'}(i,j)\ \forall \mathrm{a}, \mathrm{a}' \in \mathfrak{A} \times \mathfrak{A},\ \forall i,j \in \mathcal{I} \times \mathcal{J} \right\}. \tag{14}$$

Assuming that the event $\mathcal{E}_t$ holds the matching instability of the solution $(\mathfrak{m}_t, X_t)$ of Algorithm 1 at time t is bounded by

$$\mathbf{MI}(\mathfrak{m}_t, X_t) \leq \sum_{\mathrm{a} \in \mathfrak{A}} \mathbb{E}\left[\sqrt{\frac{2}{1 \vee n_{\mathrm{a},\mathfrak{m}_t(\mathrm{a})}^t(i_{\mathrm{a}}^t, i_{\mathfrak{m}_t(\mathrm{a})}^t)} \log\left(\frac{2}{\delta}\right)} \mid i_{\mathrm{a}}^t \sim x_{\mathrm{a}}^t, i_{\mathfrak{m}_t(\mathrm{a})}^t \sim x_{\mathfrak{m}_t(\mathrm{a})}^t \right].$$

Proof. In the following, we omit the time index t from the strategies x and the matching $\mathfrak{m}$ for convenience.

•*Step 1. Per-step Regret.*
We bound the per-step regret of the algorithm under the $\mathcal{E}_t$, by constructing subsidies s_{a}, which we prove to be feasible.

More specifically we define s_{a} as follows :

$$s_{\mathrm{a}} = \bar{U}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \ \forall \mathrm{a} \in \mathfrak{A} \tag{15}$$

where x_{a} are the strategies chosen by the agents, and $\bar{U}$ denotes the expected utility of the selected strategies with respect to the upper confidence matrix $\bar{A}$.

To prove that $s = \{s_{\mathrm{a}}\}_{\mathrm{a}\in\mathfrak{A}}$ is a feasible solution to the linear program we show that the constrains C1, C2, C3 and C4 are satisfied.

•*Satisfaction of constrain C4.*
Due to the optimism we have that $\bar{U}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} \geq U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}$, and thus

$$s_{\mathrm{a}} = \bar{U}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \geq 0 \ \forall \mathrm{a} \in \mathfrak{A} \tag{16}$$

satisfying constrain C4.

•*Satisfaction of constrain C3.*
We continue with the Nash rationality constrains C3 for the agents $\mathrm{a} \in \mathfrak{A}$ were we have that:

$$V^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) - s_{\mathrm{a}} = \tag{17}$$

$$V^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) - \bar{U}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) + U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) = \tag{18}$$

$$V^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - \bar{U}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \leq \tag{19}$$

$$\bar{V}^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - \bar{U}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \leq 0 \tag{20}$$

In the first equality, we replace s_{a} according to our definition in equation (15). The first inequality follows from the fact the $V^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} \leq \bar{V}^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}$ due to the optimism, and the second from the previews argument S1 of Proposition 2. So our subsidies s_{a} satisfy the Nash rationality constraint (C3).

•*Satisfaction of constrain* (C2).
For the individual rationality of constrains (C2) we have that:

$$\begin{aligned}
&U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) - U_{\mathrm{a},\perp} + s_{\mathrm{a}} = \\
&U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) - U_{\mathrm{a},\perp} + \bar{U}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) = \\
&\bar{U}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) - U_{\mathrm{a},\perp} \geq \\
&\bar{V}^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} - U_{\mathrm{a},\perp} \geq 0
\end{aligned}$$

The first inequality follows from the argument S1 of Proposition 2. The last inequality follows from the fact that the matching $\mathfrak{m}$ is stable under x_{a} according to the upper confidence bounds $\bar{V}^{\star}$, and therefore satisfies individual rationality.

•*Satisfaction of constrain C1.*
We now continue with the blocking pair constraint, so for (C1), we have:

$$\min\left(V^{\star}_{p,a} - U_{p,\mathfrak{m}(p)}(x_p, x_{\mathfrak{m}(p)}) - s_p,\ V^{\star}_{a,p} - U_{a,\mathfrak{m}(a)}(x_a, x_{\mathfrak{m}(a)}) - s_a\right) = \tag{21}$$

$$\min\left(V^{\star}_{p,a} - \bar{U}_{p,\mathfrak{m}(p)}(x_p, x_{\mathfrak{m}(p)}),\ V^{\star}_{a,p} - \bar{U}_{a,\mathfrak{m}(a)}(x_a, x_{\mathfrak{m}(a)})\right) \leq \tag{22}$$

$$\min\left(\bar{V}^{\star}_{p,a} - \bar{U}_{p,\mathfrak{m}(p)}(x_p, x_{\mathfrak{m}(p)}),\ V^{\star}_{a,p} - \bar{U}_{a,\mathfrak{m}(a)}(x_a, x_{\mathfrak{m}(a)})\right) \leq 0 \tag{23}$$

Where the first inequality follows from the optimism $V^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} \leq \bar{V}^{\star}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}$, and the second from the previews argument S2 of Proposition 2.

•*Bounding per step regret.*
Therefore, since s_{a} is a feasible solution to the problem, the instability of the market can be bounded

as follows:

$$\mathbf{MI}(\mathfrak{m}_t, X_t) \leq \sum_{\mathrm{a}\in\mathfrak{A}} s_{\mathrm{a}} \tag{24}$$

$$= \sum_{\mathrm{a}\in\mathfrak{A}} \bar{U}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) - U_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(x_{\mathrm{a}}, x_{\mathfrak{m}(\mathrm{a})}) \tag{25}$$

$$= \sum_{\mathrm{a}\in\mathfrak{A}} x_{\mathrm{a}}^T \bar{A}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} x_{\mathfrak{m}(\mathrm{a})} - x_{\mathrm{a}}^T A_{\mathrm{a},\mathfrak{m}(\mathrm{a})} x_{\mathfrak{m}(\mathrm{a})} \tag{26}$$

$$\leq \sum_{\mathrm{a}\in\mathfrak{A}} x_{\mathrm{a}}^T \bar{A}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} x_{\mathfrak{m}(\mathrm{a})} - x_{\mathrm{a}}^T \underline{\mathrm{A}}_{\mathrm{a},\mathfrak{m}(\mathrm{a})} x_{\mathfrak{m}(\mathrm{a})} \tag{27}$$

$$= \sum_{\mathrm{a}\in\mathfrak{A}} \mathbb{E}\left[2\sqrt{\frac{2}{1 \vee n_{\mathrm{a},\mathfrak{m}(\mathrm{a})}(i_{\mathrm{a}}, i_{\mathfrak{m}(\mathrm{a})})} \log\left(\frac{1}{\delta}\right)} \mid i_{\mathrm{a}} \sim x_{\mathrm{a}}, i_{\mathfrak{m}(\mathrm{a})} \sim x_{\mathfrak{m}(\mathrm{a})} \right] \tag{28}$$

where for the inequality, we use the fact that under the event $\mathcal{E}_t$, it holds that $A_{\mathrm{a},\mathfrak{m}(\mathrm{a})} \geq \underline{A}_{\mathrm{a},\mathfrak{m}(\mathrm{a})}$. □

C.2 Main theorem

Now we prove our main theorem, that we restate bellow:

Theorem 1 (Restated)**.** *Let* $|\mathcal{P}| = \mathbf{p}$, $|\mathcal{A}| = \mathbf{a}$, $|\mathcal{I}| = \mathbf{m}$ *and* $|\mathcal{J}| = \mathbf{k}$*. When Algorithm 1 is run with parameter* $\delta = 1/(4T^2\mathbf{p}^2\mathbf{a}^2\mathbf{mk})$*, the expected regret* R_T *under horizon* T *is bounded:*

$$\mathbb{E}[R_T] \leq O\left(2\sqrt{4T\mathbf{mkpa}\log\left(4T^2\mathbf{mkp}^2\mathbf{a}^2\right)}\right)$$

Proof. To bound the expected regret, we consider the event where the entries of the true payoff matrices lie within the confidence intervals $C^t_{\mathrm{a},\mathrm{a}'}(i,j) = [\underline{\mathrm{A}}^t_{\mathrm{a},\mathrm{a}'}(i,j), \bar{A}^t_{\mathrm{a},\mathrm{a}'}(i,j)]$ for all time steps, i.e.:

$$\mathcal{E} = \{\mathcal{E}_t \;\forall\, t \in [T]\} = \left\{A_{\mathrm{a},\mathrm{a}'}(i,j) \in C^t_{\mathrm{a},\mathrm{a}'}(i,j)\;\forall \mathrm{a},\mathrm{a}' \in \mathfrak{A}\times\mathfrak{A},\; \forall i,j \in \mathcal{I}\times\mathcal{J}, \forall t \in [T]\right\}. \tag{29}$$

So for the expected regret, we have that:

$$\begin{aligned}
\mathbb{E}[R_T] &= \mathbb{E}[R_T \mid \mathcal{E}]\Pr(\mathcal{E}) + \mathbb{E}[R_T \mid \neg\mathcal{E}]\Pr(\neg\mathcal{E}) \\
&\leq \mathbb{E}[R_T \mid \mathcal{E}] + \mathbb{E}[R_T \mid \neg\mathcal{E}]\Pr(\neg\mathcal{E}) \\
&\leq \sum_{t=1}^{T} \underbrace{\mathbf{MI}(m_t, X_t)}_{(A)} + \underbrace{2T(\mathbf{a}+\mathbf{p})\Pr(\neg\mathcal{E})}_{(B)}
\end{aligned}$$

As the regret $\mathbb{E}[R_T \mid \neg\mathcal{E}]$ under $\neg\mathcal{E}$ can be bounded by, $2T(\mathbf{a}+\mathbf{p})$, in the worst-case.

The event $\mathcal{E}$ involves $O(T\times\mathbf{p}\times\mathbf{a}\times\mathbf{m}\times\mathbf{k})$ independent 1-sub-Gaussian random variables. Therefore, by applying the union bound and the concentration bounds for sub-Gaussian (see also Appendix C.3) random variables, we can bound $\Pr(\neg\mathcal{E})$ as follows:

$$\Pr(\neg\mathcal{E}) \leq 2\delta T\mathbf{pamk} = \frac{1}{2T\mathbf{pa}},$$

by setting $\delta = \frac{1}{4T^2\mathbf{p}^2\mathbf{a}^2\mathbf{mk}}$.

For the second term (B) we have that:

$$(B) \leq 2T(\mathbf{p}+\mathbf{a})\Pr(\neg\mathcal{E}) \leq 4T\mathbf{pa}\Pr(\neg\mathcal{E}) \leq 2 \tag{30}$$

using that $(a+b) \leq 2ab$ for $a, b \geq 1$.

For the first term (A) we have that:

$$(A) = \sum_{t=1}^{T} \mathbf{MI}(\mathfrak{m}_t, X_t) \leq \sum_{t=1}^{T} \sum_{\mathrm{a} \in \mathfrak{A}} \mathbb{E}\left[\sqrt{\frac{2}{1 \vee n^t_{\mathrm{a},\mathfrak{m}_t(\mathrm{a})}(i^t_{\mathrm{a}}, i^t_{\mathfrak{m}_t(\mathrm{a})})} \log\left(\frac{1}{\delta}\right)}\right] \tag{31}$$

$$\leq 2 \sum_{t=1}^{T} \sum_{(p,a) \in \mathfrak{m}_t} \mathbb{E}\left[\sqrt{\frac{2}{1 \vee n^t_{p,a}(i^t_p, i^t_a)} \log\left(\frac{1}{\delta}\right)}\right] \tag{32}$$

$$= 2 \sum_{i,j} \sum_{p,a} \mathbb{E}\left[\sum_{\substack{t=1: \\ i^t_p = i, i^t_a = j \\ (p,a) \in m_t}}^{T} \sqrt{\frac{2}{1 \vee n^t_{p,a}(i,j)} \log\left(\frac{1}{\delta}\right)}\right] \tag{33}$$

$$\leq 2 \sum_{i,j} \sum_{p,a} \mathbb{E}\left[\sqrt{4 n^T_{p,a}(i,j) \log\left(\frac{1}{\delta}\right)}\right] \tag{34}$$

$$\leq 2 \sqrt{4T\mathbf{mkpa} \log\left(\frac{1}{\delta}\right)} \tag{35}$$

In (32), we rewrite the summation by explicitly considering the matched agents. In (33), we restructure the sum to account for all occurrences where the matched agents (p,a) select actions (i,j). The transition from (33) to (34) follows from an integral bound, which is a consequence of the Fundamental Theorem of Calculus. Finally, in (35), we obtain an upper bound using the Cauchy Schwarz inequality.

Thus the overall regret by setting $\delta = 1/(4T^2\mathbf{p}^2\mathbf{a}^2\mathbf{mk})$ is bounded by:

$$R_T \leq 2\sqrt{4T\mathbf{mkpa} \log\left(4T^2\mathbf{mkp}^2\mathbf{a}^2\right)} + 2 \tag{36}$$

$$\leq \tilde{O}\left(\sqrt{T\mathbf{mkpa}}\right) \tag{37}$$

□

C.3 Concentration bound for sub-Gaussian random variables

For completeness, we present a lemma on the concentration bounds for sub-Gaussian random variables. More details can be found in book of Lattimore and Szepesvari [28].

Lemma 2. *Let $X_1, \ldots, X_n$ be i.i.d. σ-sub-Gaussian random variables with $\mathbb{E}[X_i] = \mu$. Then, for all $\epsilon \geq 0$, it holds that:*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right).$$

Applying Lemma 2 for $\sigma = 1$ and setting $\epsilon = \sqrt{\frac{2\ln(1/\delta)}{n}}$, we obtain:

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \sqrt{\frac{2\ln(1/\delta)}{n}}\right) \leq 2\delta.$$

Finally, using the union bound, the probability of the event $\neg\mathcal{E}$ in Equation 29 is bounded as follows:

$$\Pr(\neg\mathcal{E}) \leq 2\delta T\mathbf{pamk}.$$

D Simulation details

We now present a set of simulations that complement our theoretical results. The code for reproducing all the experiments is provided in the supplementary material[4].

[4]To be published at `www.github.com/to/be/published`

We perform simulations under three distinct settings to evaluate the role of available information and the resulting agent behavior. To avoid trivial equilibria involving unmatched agents, we restrict our simulations to the case where all agents strictly prefer being matched over remaining unmatched, i.e., $U_{\mathrm{a},\perp} = -1 \; \forall \mathrm{a} \in \mathfrak{A}$.

The first setting is the **Self-play** scenario, in which all agents act according to our proposed Algorithm 1.

The second setting is the **Nash-response** scenario, where one side of the market has access to additional information about the underlying games. Specifically, we assume that each agent $a \in \mathcal{A}$ knows the true payoff matrices, i.e., $A_{\mathrm{a},p} \; \forall p \in \mathcal{P}$. These agents select strategies and report their preferences based on the Nash Equilibrium (NE) of the true underlying games. The agents $p \in \mathcal{P}$ act according to Algorithm 1.

The third setting is the **Best-response** scenario, where agents $a \in \mathcal{A}$ know both the true payoff matrices (as in the previous setting) and the strategies that agents $p \in \mathcal{P}$ will play when matched with them, i.e., $x_{p,a} \; \forall p \in \mathcal{P}$. In this case, agents $a \in \mathcal{A}$ compute their preferences and strategies according to the best response. Specifically, preferences are defined as $\pi_a = \arg\mathrm{sort}_{p \in \mathcal{P}} U_{a,p}(x_{a,p}, x_{p,a})$, where $x_{a,p} = \arg\max_x U_{a,p}(x, x_{p,a})$. This setup resembles the one studied in [36] for two-player zero-sum games, where one player knows the true payoff matrix and the opponent's strategies, and selects the best response accordingly.

For each setting, we generate 50 random instances by independently sampling the entities payoff matrices of the ZSG from a normal distribution, i.e., $A_{p,a}(i,j) \sim \mathcal{N}(0,1)$ for all $(p,a,i,j) \in \mathcal{P} \times \mathcal{A} \times \mathcal{I} \times \mathcal{J}$. We compares the regret for the different settings for varying numbers of agents and actions, along with the theoretical bound of $O\left(\sqrt{T\mathbf{apmk}}\right)$ (dashed blue), where we omit the logarithmic term for clarity.

Figure 1 presents all our experiments for the different configuration of agents and actions. In all experimental settings, we observe sublinear regret growth, indicating that the system progressively learns a *matching equilibrium*. The **Nash-response** (dotted orange) scenario yields the lowest regret, as one side of the market has full knowledge of the payoff matrices and acts according to NE. In addition, the **Best-response** (dash-dotted purple) scenario initially outperforms **Self-play** (solid green), but as the number of agents and actions increases, its performance deteriorates. This is expected as agents in $\mathcal{A}$ exploit their matched partners in $\mathcal{P}$, which introduces greater complexity into the learning process.

D.1 System specifications

The approximate time required to reproduce all experiments using our implementation (including all algorithms and agent/action settings) described in Figure 2 is about 70 hours, with the following system specifications:

System specifications: **CPU Model**: Intel(R) Xeon(R) Gold 6326 @ 2.90GHz, 32 (16 cores, 2 threads per core), **RAM**: 62 GiB.

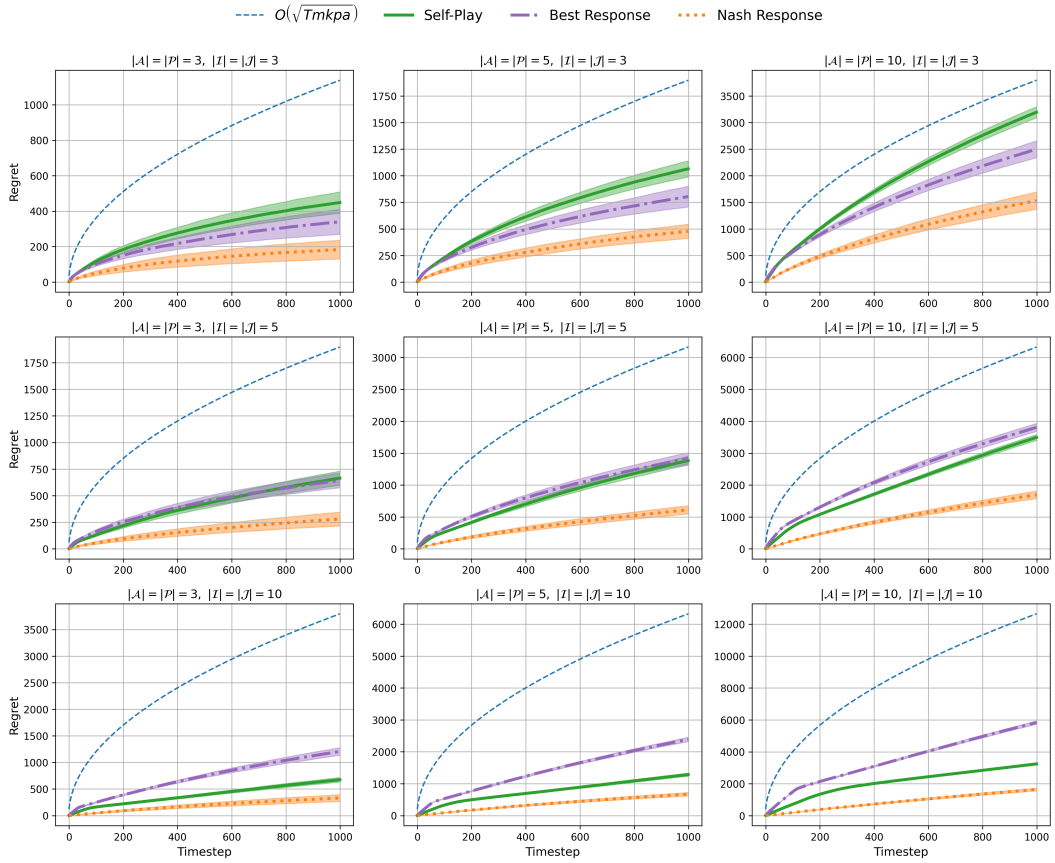

Figure 2: Regret of the different settings, including the theoretical bound for varying numbers of agents and actions. Lines represent the average performance of the algorithm over 50 runs and shaded areas indicate the standard deviation.