Decentralized, Communication- and Coordination-free Learning in Structured Matching Markets

Chinmay Maheshwari, Eric Mazumdar, and Shankar Sastry *

Abstract

We study the problem of online learning in competitive settings in the context of two-sided matching markets. In particular, one side of the market, the agents, must learn about their preferences over the other side, the firms, through repeated interaction while competing with other agents for successful matches. We propose a class of decentralized, communication- and coordination-free algorithms that agents can use to reach to their stable match in structured matching markets. In contrast to prior works, the proposed algorithms make decisions based solely on an agent's own history of play and requires no foreknowledge of the firms' preferences. Our algorithms are constructed by splitting up the statistical problem of learning one's preferences, from noisy observations, from the problem of competing for firms. We show that under realistic structural assumptions on the underlying preferences of the agents and firms, the proposed algorithms incur a regret which grows at most logarithmically in the time horizon. Our results show that, in the case of matching markets, competition need not drastically affect the performance of decentralized, communication and coordination free online learning algorithms.

1 Introduction

Online decision-making under uncertainty is one of the central problems in modern machine learning, reflecting the need for efficient and high performing algorithms for real-time learning in real-world settings. Despite being such a well-researched area, there is a broad lack of understanding of how to deploy online learning algorithms into settings in which they must compete with each other for resources or information. Indeed, while classic problems of online learning deal with trading off the exploration of possible choices and the exploitation of current knowledge (i.e., the exploration-exploitation tradeoff [LS20, Sli19]), the addition of competition adds a new axis upon which algorithms must operate [MSW17, AMSW20]—namely that of competing (perhaps unsuccessfully) for highly desired outcomes or settling for less desired (but also less competitive) outcomes. Broadly, speaking, the dominant approach to dealing with competition in machine learning remains to treat opponents as adversarial[CBL06], despite a long literature in economics and game theory [Lit94, FDLL98] showing how agents who understand the competitive structure of problems can sometimes vastly outperform solutions based upon worst-case modeling.

In this paper, we address the problem of online learning in competitive settings in the context of *two-sided matching markets*. Two-sided matching markets *match* users on one side of the market to those on the other to facilitate the exchange of goods or services. In such settings, each user on one side of the market has an inherent preference ordering

*C. Maheshwari(chinmay_maheshwari@berkeley.edu) and S. Sastry (shankar_sastry@berkeley.edu) are with EECS department at University of California Berkeley. E. Mazumdar (mazumdar@caltech.edu) is with CMS And Economics department at Caltech.

for the users on the other side of the market. Since each user seeks to find their most desired match, this results in a game in which a natural equilibrium is that of a *stable matching* wherein no two users would prefer switching from their current match to each other given their preferences. In seminal work, [GS62] proposed a simple and effective algorithm— the *Deferred Acceptance (DA) Algorithm*— that users on one side of the market can implement to find such a solution when every user knows their own preferences. The algorithm has been widely used in examples ranging from kidney exchanges to medical resident matching where preferences can be assigned or reported to a central authority which does the matching. However, recent years have seen the emergence of a new form of *online* matching markets like online labor markets (e.g. TaskRabbit, Upwork), online dating markets (e.g. Tinder, Match.com), online crowdsourcing platforms (e.g. Amazon mechanical turk) where the users do not know their preferences apriori, and can repeatedly interact with the market to improve their match quality.

Motivated by these applications we consider a generalization of the problem studied in the seminal paper [GS62] wherein one side of the market— the agents— do not know their own preferences, but are able to interact repeatedly with the market. In particular, we analyze a repeated game in which, at each round, agents can request to match with a user or firm on the other side of the market. If, at a given round, multiple agents request the same firm, the firm— assumed to be a myopic utility maximizer— accepts the request of its most preferred agent (who receives a noisy measurement of their utility of the match from which they can learn their preferences) and rejects the others (who receive no information about their preferences). This setup serves has been studied in a line of recent works on online matching markets [LMJ20, LRMJ21, SBS21, BSS21].

Successful algorithms for this framework must simultaneously solve a statistical learning problem (that of learning about their own preferences) and a competitive problem (ensuring that agents get their most desired match despite the presence of other self-interested agents in the market). Previous works for addressing this problem propose algorithms that are centralized [LMJ20] (whereby agents send their current beliefs over their preferences to a central platform which does the matching), require coordination between agents (i.e., a choreographed set of strategies to minimize rejections) [SBS21, BSS21], or require agents to fully observe the market outcomes of other agents [LRMJ21]. In contrast, the DA algorithm— which we take to be the full-information benchmark to which we compare algorithms— is (i) fully decentralized, (ii) coordination-free, and (iii) requires agents to make decisions only based upon their own history of rejections and successful matchings. Designing learning algorithms that operate under conditions (i)-(iii) ensures scalability and privacy in large-scale systems where it is unrealistic to assume that agents can keep track of all other agents' matchings. Thus in this work we focus on the question:

> **Does there exist decentralized and coordination-free algorithms that are based only on local history of interactions which provably converges to stable matching?**

Contributions. In this work we design algorithms for learning while matching in a class of structured matching markets known as $\alpha-$reducible matching markets. This condition ensures that there exists an unique stable matching and encompasses many realistic preference structures including serial dictatorship and no crossing conditions [Cla06]. We show that the proposed algorithms incur a stable regret with respect to the unique stable matching that grows at most logarithmically in the time horizon. The particular contributions of this paper are:

1. We present a general framework for the construction of decentralized, communication, and coordination-free algorithms for learning while matching. In particular, we combine index-based stochastic bandit algorithms (in particular the Upper Confidence Bounds algorithm and Thompson Sampling) [Aue02], [LS20, Sli19] for solving the statistical problem of learning an agent's preferences with a path-length adversarial bandit algorithm [BLLW19, WL18] for dealing with the competitive problem. The resulting algorithms make are fully decentralized, and communication and coordination-free since they make use of only an agent's history of collisions, matches, and rewards to choose which firm to request at a given time. Furthermore the algorithms are "any-time" algorithms, in that they do not require knowledge of time horizon and do not require any specific parameters of the bandit instance beyond the sub-gaussian parameter of the noise.

2. We show that when the agents' and firms' preferences satisfy the $\alpha-$reducibility condition and *every* agent uses the algorithm, the regret accumulated by any agent a against the stable match is $O\left(\frac{C_a|\mathcal{A}||\mathcal{F}|log(T)}{\Delta^2}\right)$ where $\mathcal{A}$ is the set of agents, $\mathcal{F}$ is the set of firms, Δ is the minimum sub-optimality gap of any agent in the market, and C_a is a constant that depends on the $\alpha-$reducible structure of the market.

Organization The paper is organized as follows: In Section 2 we discuss and compare the prior literature related to the focus of this paper. In Section 3 we introduce the general problem setup, introduce matching markets and discuss the structural assumptions on the preferences of agents and firms. In Section 4 we present the algorithmic design paradigm along with a specific algorithm, based on Upper Confidence Bound. In Section 5 we show that the algorithm incurs $O(\log(T))$ regret along with a brief sketch of the proof. In Section 6 we study the performance of the algorithm in simulation. We conclude the paper in Section 7 and also provide some future research directions. The proofs of our results are relegated to the Appendix. Moreover, we introduce another important variant of algorithm based on Thompson Sampling with similar results in the Appendix.

2 Related works

Sequential decision-making under uncertainty has been extensively studied in machine learning under the guise of multi-armed bandit (MAB) problems. In general, MAB problems can be split into two distinct flavors, which differ in the type of feedback agents receive. Crucially, in both problems the key is trading off exploration of actions and exploiting ones current knowledge.

In the first class of MAB problems, the stochastic MAB, playing an action results in an unbiased estimate of the utility of playing that action. Solutions to the problem can be split among two dominant algorithmic paradigms. The first, based on principle of optimism in the face of uncertainty encompasses the well known upper confidence bounds (UCB) algorithm [LS20, LR85] and its variants, while the second, based on Thompson sampling takes a Bayesian approach [RRKO17, Tho33] Each of these approaches are known to have optimal performance measured in terms of *regret:* the expected cumulative utility generated from the algorithm's chosen actions compared to the expected utility that could have been generated from always choosing the best possible action (i.e., the best action that one would choose with full information) [LS20, AG12]. In particular, these algorithms are known to incur *logarithmic* regret, i.e., regret that grows at most logarithmically over time— which is known to be optimal for this class of problems up to constant factors. In our paper we present an algorithmic framework for learning in matching markets that works with

either class of algorithm, and further incurs logarithmic regret *even* while dealing with competition.

The second class of multi-armed bandit problems, coming from the literature on learning in games, seeks algorithms that can perform against arbitrary feedback sequences [CBL06]. Solutions to this class of problems, known as adversarial bandit algorithms, are an active research topic. While it is well known that using simple strategies like multiplicative weights can guarantee regret against the best fixed action in hindsight on the order of $\sqrt{T}$ against worst-case adversaries [CBL06], designing algorithms that can improve upon this when adversaries are *not* worst case remains an open research problem. In this paper we leverage advances on the development of *path-length* adversarial regret algorithms that address this problem and guarantee regret that directly depends on the amount of variation an adversary presents [BLLW19, WL18].

We briefly remark that there exists several lines of research on multi-agent bandits. One of them is on multi-agent bandits with collisions (with applications primarily in the area of spectrum sharing in wireless networks[LZ10, KNJ14, RSS16, LM21, BBS20]). In such models the arms do not have preferences and if more than one agents collide at any arm then no one receives any utility or attains maximum possible loss. However, these models differ from us since we consider that both sides of markets have preference over one another and when there is a collision only one agents gets matched. Another line of research deals with the problem of collaboratively learning an instance of multi-armed bandit [BTZ15, CCDJ17, SGS19] where agents can communicate. Note that in these settings there is no competition that is more than one agents apply at same arm at same time.

The particular intersection of MABs and two-sided matching markets that we analyze has seen a flurry of recent works [LMJ20, LRMJ21, BSS21, SBS21]. To the best of our knowledge, [DK05], presented the first numerical study on effectively using MAB algorithms to learn preferences in matching markets. However, it was only recently that [LMJ20] rigorously formulated the bandit learning problem in the matching markets, and generalized the notion of *regret* from the MAB literature to matching markets in terms of *stable regret*— i.e., the expected cumulative utility benchmarked against the expected cumulative reward that would have been received if everyone in the market requested their match in a certain stable match[1]. Moreover, they proposed a *centralized* UCB-based algorithm that facilitates the matching between agents and firms given each agents' current beliefs over their preferences and history of play, while ensuring that $\mathcal{O}(|\mathcal{A}||\mathcal{F}|\log(T))$ regret for a UCB based algorithm, where $\mathcal{A}$ is the set of agents, $\mathcal{F}$ is the set of firms, and T is the time horizon of the problem. In follow up work [LRMJ21] proposed a *decentralized* bandit learning algorithm based on UCB that allows each user to take its decision in a decentralized manner and still "converge" to stable matching while incurring $O(\exp(|\mathcal{F}|^4)\log^2(T))$ regret. More recently [KYL22] proposed a thompson sampling based variant of [LRMJ21]. However, these algorithms requires the knowledge of outcomes at other firms at every round, leaving algorithms that are based solely on agents' own history of play as an open problem. Concurrently, [SBS21] proposed an algorithm that works in phases and makes use of communication between agents to coordinate agents' actions. Under this information structure the algorithm achieves $\mathcal{O}\left(|\mathcal{F}|^2|\mathcal{A}|^2\log(T)\right)$ regret. Moreover their guarantees require that firms have homogeneous preference over the agents (also referred as *serial dictatorship*). Follow-up work, [BSS21] improved the regret for serial dictatorship to $\mathcal{O}\left(|\mathcal{F}||\mathcal{A}|\log(T)\right)$ by proposing a new algorithm. Additionally, they also showed that if the assumption of serial dictatorship

[1]Note that the stable matching need not be unique in general. Thus the stable regret has to be always specified with respect to which stable matching is being used. Typically, in literature two main stable matchings are considered namely *agent optimal stable matching* and *firm optimal stable matching*.

is relaxed to a weaker structural condition then they obtain $O(poly(|\mathcal{A}|, |\mathcal{F}|)\log(T))$ regret. Even though the proposed algorithm in [BSS21] has decentralization it is a phase based algorithm, the agents act according to a coordinated protocol at some rounds. In this paper we propose a simple, decentralized, communication and coordination free algorithm in which agents make use of their own local information to learn while matching. Unlike previous works [LMJ20, LRMJ21, SBS21, BSS21] where the algorithms are constructed using a UCB subroutine, we also show that our algorithmic design paradigm can be also seamlessly extended to Thompson sampling variant.

We would also like to remark about another line of research at the intersection of multiarmed bandits and matching markets [JWW+21], [JKK16] ,[CS21] which consider the problem of learning preferences from the perspective of a platform.

3 Setting

We define a two-sided market $\mathcal{M}$ as collection of agents $\mathcal{A}$ and firms $\mathcal{F}$. In the setting under consideration, we assume that every agent $a \in \mathcal{A}$ has *unknown* preferences over firms $f \in \mathcal{F}$ which are captured by utilities $u_a(f) \in \mathbb{R}$. Moreover, no two firms give the same utility to a given agent, i.e. $u_a(f) \neq u_a(f')$ if $f \neq f'$. We assume that every agent seeks to be matched to their most preferred firm, and that firms have preferences over all the agents which are also captured by utilities $u_f(a)$ for each a and f such that no two agents give same utility to firms i.e. $u_f(a) \neq u_f(a')$. Importantly, we assume that firms know their own preference orderings over agents and that there are more firms than agents, i.e. $|\mathcal{A}| \leq |\mathcal{F}|$. The interaction between agents and firms happens as follows: In each time step $t = 1, \ldots, T$ every agent $a \in \mathcal{A}$ independently *requests* a firm $f_a(t) \in \mathcal{F}$. As the agents request independently, it is possible that more than one agent requests the same firm f. For $f \in \mathcal{F}$, let $\mathbb{A}_f(t) \coloneqq \{a \in \mathcal{A} : f_a(t) = f\}$ denote the set of agents that request firm f at time step t. At each time step t, we assume that the firm f accepts the request of their most preferred agent in $\mathbb{A}_f(t)$ denoted by $a_f(t) \coloneqq \arg\max_{a \in \mathbb{A}_f(t)} u_f(a)$, and rejects the request of all other agents. That agent $a_f(t)$ is said to be the agent who got *matched* with firm f at time t. Moreover every matched agent receives a noisy measurement of their utility, denoted $U_{\mathbf{a},f}$ such that

$$U_{\mathbf{a},f} = u_{\mathbf{a}}(f) + \zeta_{\mathbf{a},f}, \tag{3.1}$$

where $\zeta_{\mathbf{a},f}$ is a zero-mean, one-sub-Gaussian random variable. Meanwhile, all the agents that are rejected are said to have *collided* on firm f, for which they receive no utility i.e. $U_{a,f}(t) = 0$.

We restrict that agents *only* receive the following information at any time step t:

1. $Y_a(t) = \mathbb{1}\left(a \text{ is matched to } f_a(t)\right)$. which captures if agent a gets matched at time t
2. if they get matched, the noisy measurement of their utility, $U_{a,f}(t)$.

Remark 1. *We note that in this setup an agent does not know* anything *about how other agents are performing in the market. Agents do* not *observe who gets successfully matched on firms that they have requested and do not observe who they have collided with. We remark that this is* the same *information structure as that assumed by the DA algorithm and is the key assumption that differentiates our work from prior work on this problem [LMJ20, LRMJ21, BSS21, SBS21].*

In the following subsection, we recall some important results from matching market literature that are crucial to further exposition.

3.1 Preliminaries on matching markets

To analyze the matching market defined in the previous section we recall key concepts from the literature on matching markets. A matching $\mathbb{M} : \mathcal{A} \longrightarrow \mathcal{F}$ is an injective function such that $\mathbb{M}(a) = f$ denotes that a and firm f are matched. We call a matching *unstable* if there is an agent-firm tuple $(a, f) \in \mathcal{A} \times \mathcal{F}$ such that $u_a(\mathbb{M}(a)) < u_a(f)$ and $u_f(a) > u_f(\mathbb{M}^{-1}(f))$. In words, there is a pair (a, f) who both prefer each other over their current match, such pair is called a *blocking pair*. A matching is *stable* if it is not unstable. It is usually the case that a market may have multiple stable matchings. However, for the purpose of this paper we focus on markets which are $\alpha-$reducible, first introduced in [Alc94] and further analyzed in [Cla06], that ensures there is a unique stable matching. Before formally describing this property we introduce the notion of a submarket and fixed pair.

A sub-market of $\mathcal{M}$ is a market $\mathcal{M}'$ such that $\mathcal{M}' = \mathcal{A}' \cup \mathcal{F}'$ where $\mathcal{A}' \subseteq \mathcal{A}$, $\mathcal{F}' \subseteq \mathcal{F}$, and $|\mathcal{A}'| \leq |\mathcal{F}'|$. Meanwhile, a pair $(a, f) \in \mathcal{A} \times \mathcal{F}$ is a *fixed pair* of market $\mathcal{M}$ if $u_a(f) \geq u_a(f')$ for all $f' \in \mathcal{F}$ and $u_f(a) \geq u_f(a')$ for all $a' \in \mathcal{A}$. In words, a fixed pair is any agent-firm pair that prefer each other over any other options in the market. We now define the notion of $\alpha-$reducibility.

Definition 2 (α-reducibility)**.** *A market $\mathcal{M} = \mathcal{A} \cup \mathcal{F}$ is α-reducible if every sub-market of $\mathcal{M}$ has a fixed pair.*

The notion of α-reducibility is weaker than the *no crossing condition* and serial dictatorship [Cla06]. These conditions have been introduced in the effort to characterize the existence and uniqueness of a stable matching. In [Cla06] the authors show that every sub-market of of $\mathcal{M}$ has a unique stable matching if $\mathcal{M}$ is α-reducible.

The preceding property of $\alpha-$reducible markets will be crucial to obtain regret guarantees for the proposed algorithm in this paper. Thus, we assume that $\mathcal{M}$ is α-reducible.

Remark 3. *An important property of $\alpha-$reducibility assumption that is central to the subsequent analysis is that it allows us to partition the market into various sub-markets by sequentially eliminating fixed pairs. More formally, lets define $\mathcal{A}_0 = \mathcal{F}_0 = \varnothing$ and $\mathcal{M}_0 = \mathcal{M}$. Now for $i \geq 1$ lets define inductively $\mathcal{A}_i \subseteq \mathcal{A} \backslash \{\cup_{j=1}^{i} \mathcal{A}_{j-1}\}, \mathcal{F}_i \subseteq \mathcal{F} \backslash \{\cup_{j=1}^{i} \mathcal{F}_{j-1}\}$ be the set of agents and set of firms that constitute fixed pair in market $\mathcal{M}_{i-1}$. That is, for every agent $a \in \mathcal{A}_i$ there exists a unique $f \in \mathcal{F}_i$ such that (a, f) is a fixed pair of market $\mathcal{M}_{i-1}$. The iteration evolves as $\mathcal{M}_i := \{\mathcal{A} \backslash \{\cup_{j=0}^{i} \mathcal{A}_j\}\} \cup \{\mathcal{F} \backslash \{\cup_{j=0}^{i} \mathcal{F}_j\}\}$. Let K be the total number of such sub-markets $\{\mathcal{M}_i\}$. Moreover such decomposition of market is unique.*

For any agent $a \in \mathcal{A}$ we denote by f_a^* its match in the unique stable matching. Furthermore, let $\overline{\mathbb{F}}_a := \{f \in \mathcal{F} : u_a(f) > u_a(f_a^*)\}$ be the set of firms that agent a prefers over its stable match. We call such firms *super-optimal* firms for a. Similarly, let $\underline{\mathbb{F}}_a := \{f \in \mathcal{F} : u_a(f) < u_a(f_a^*)\}$ be the set of firms which are less preferred than the stable match by agent a. We call such firms *sub-optimal* firms for a. Note that we have following lemma which states a crucial property of super-optimal firms for $\alpha-$reducible markets.

Lemma 4. *For any $i \in [K]$ and agent $a \in \mathcal{A}_i$ the set of super-optimal firms are contained in $\cup_{j=1}^{i-1} \mathcal{F}_j$.*

An immediate conclusion of Lemma 4 is that it creates a hierarchy in the market. That is, an agent $a \in \mathcal{A}_i$, for some $i \in [K]$, is in a sense "higher ranked" than a agent $a' \in \mathcal{A}_j$ for $j > i$ as the former's stable match can be super-optimal for the latter. This sort of hierarchy naturally manifests itself in the learning process where learning of agent a creates *externality* for agent a'.

For ease of reference, all key notations used in paper are presented in a table in the Appendix.

4 Description of the Algorithm

In this section we present a novel algorithm design principle for agents to learn about the preferences while ensuring that they perform competitively against the match that they could have achieved if they knew their preferences and used the DA algorithm. Throughout this section, we assume that every agent $a \in \mathcal{A}$ uses these algorithms in order to decide which firm to choose at time any time t. The proposed algorithms—by design— make use of only the feedback information outlined in (1)-(2) in Section 1, and have no implicit or explicit communication and coordination strategies like e.g., phase based strategies with coordinated actions [BSS21] or partial observation of actions of other agents [LRMJ21] etc. Thus, the algorithms operate in the same regime as the DA algorithm, but without the assumption that agents know their preferences. Key to our approach, is the blending stochastic bandit (SB) algorithms with an adversarial bandit (AB) algorithms. In the subsequent exposition we will formally describe our approach and show its desirable properties in terms of regret and convergence.

Before doing so, however, we comment on the difficulties of the problem at hand, and what makes the analysis of these algorithms highly non-trivial. The key challenge in designing algorithms for matching while learning is understanding when to stop requesting *super-optimal* firms (i.e. firms that they prefer more than their stable match) without any foreknowledge of the market structure. The crux of this problem is having an agent learn that certain firms are unattainable due to competition despite the non-stationarity in the environment stemming from fact that other agents are learning simultaneously and not knowing who they collide with and who is successfully getting matched at each round. Furthermore, due to a lack of communication or coordination, agents cannot learn about which firms are super-optimal without risking many collisions.

A sketch of the algorithm is described in words in Algorithm 1, and the exact algorithm for the setting in which agents use the UCB algorithm as a subroutine is presented in Algorithm 2. As per Algorithm 2, each agent is equipped with a stochastic bandit (SB)

Algorithm 1: High-level algorithmic description

Each agent $a \in \mathcal{A}$ at every time $t \in [T]$:

1. Keeps a ordering of firms as per an index-based stochastic bandit subroutine
2. Agent a goes over the firms as per the ordering one by one
3. Using an adversarial bandit subroutine decides whether to *request* the firm or to *prune it*
   (a) If a firm is requested then agent either gets matched or gets collided
   (b) If pruned then then the agent moves to next firm as per the ordering
4. Updates the stochastic and adversarial bandit subroutine based on the feedback received

subroutine. At every time step $t \in [T]$, the SB subroutine of every agent a maintains ordering of firms in decreasing order of preferences according to an index (e.g. UCB). We denote this index of firm f as maintained by agent a as $\mathsf{UCB}_{a,f}(t)$. Next, at that time step, every agent *considers* each firm one by one in decreasing order of $\mathsf{UCB}_{a,f}(t)$. For any firm f considered by agent a at time t, the agent makes a decision to either *request* f

or to *prune*[2] it (that is, to reject that firm). In particular, agent a requests firm f with probability $p_{a,f}(t)$. Let $P_{a,f}(t) \sim \text{Bernoulli}(p_{a,f}(t))$. If a firm is pruned (i.e. $P_{a,f}(t) = 0$) then the next best firm from the sorted list is chosen and the process continues until a firm is requested (i.e. $P_{a,f}(t) = 1$). However, if all of the firms are pruned then at that time instant the agent simply requests the firm $\arg\max_f \mathsf{UCB}_{a,f}(t)$. Once an agent decides which firm to request, it obtains a noisy utility if it gets successfully matched. This feedback is used by the agent to update its UCB-index. Based on whether an agent a decides to prune or request a particular firm f, it updates $p_{a,f}$ using an AB subroutine. The details about this are stated in Section[3] 4.2 We note that all firms are not considered by agent a at every time t. Once an agent decides to request a firm f, it does not consider firms in the set $\{f' \in \mathcal{F} : \mathcal{I}_{a,f'}(t) < \mathcal{I}_{a,f}(t)\}$. Formally, for any agent-firm tuple $(a, f) \in \mathcal{A} \times \mathcal{F}$ let the event that the agent a considers the firm f at time t, to decide whether to request it or prune it, be denoted by $E^{(\mathsf{c})}_{a,f}(t) = \mathbb{1}\left(P_{a,f'}(t) = 0, \quad \forall\, f' : \mathcal{I}_{a,f}(t) \leq \mathcal{I}_{a,f'}(t)\right)$. If a firm f is considered by agent a then the event when agent a requests f is denoted by $E^{(\mathsf{r})}_{a,f}(t) = \mathbb{1}\left(P_{a,f}(t) = 1, E^{(\mathsf{c})}_{a,f}(t) = 1\right)$.

Algorithm 2: UCB based Decentralized Matching Algorithm (UCB-DMA)

Initialize: $\hat{\mu}_{a,f} = 0, M_{a,f} = 0, p_{a,f} = 0.5, x_{a,f} = 0.5, L_{a,f} = 0, \quad \forall a \in \mathcal{A}, f \in \mathcal{F}$

for $t = 1, \ldots, T$ **do**
  for $f \in \mathcal{F}$ **do**
    Set $\mathsf{UCB}_{a,f} = \hat{\mu}_{a,f} + \sqrt{\frac{2\log(1+(\bar{M}_a+1)\log^2(\bar{M}_a+1))}{M_{a,f}}}$, where $\bar{M}_a = \sum_{f\in\mathcal{F}} M_{a,f}$
  end
  Set $\mathsf{ArgUCB}_a = \mathsf{ArgDescendingSort}(\{\mathsf{UCB}_{a,f}\}_{f\in\mathcal{F}})$ and $i = 1$
  while $i \leq |\mathcal{F}|$ **do**
    Set $f = \mathsf{ArgUCB}_a^{[i]}$ and sample $P_{a,f} \sim \mathsf{Bernoulli}(p_{a,f})$
    if $P_{a,f} = 0$ **then**
      Update $(x_{a,f}, p_{a,f}, L_{a,f}) \longleftarrow \mathsf{AB\_Subroutine}(P_{a,f}, x_{a,f}, p_{a,f}, L_{a,f}, Y_a)$
    end
    if $P_{a,f} = 1$ **then**
      Request firm f and receive (U_a, Y_a)
      Update $\hat{\mu}_{a,f} \longleftarrow Y_a \frac{\hat{\mu}_{a,f} M_{a,f} + U_a}{M_{a,f}+1} + (1 - Y_a)\hat{\mu}_{a,f}, \quad M_{a,f} \longleftarrow M_{a,f} + Y_a,$
      Update $(x_{a,f}, p_{a,f}, L_{a,f}) \longleftarrow \mathsf{AB\_Subroutine}(P_{a,f}, x_{a,f}, p_{a,f}, L_{a,f}, Y_a)$
      break while;
    end
    $i \longleftarrow i + 1$
  end
  if $i = |\mathcal{F}| + 1$ **then**
    Request firm $\mathsf{ArgUCB}_a^{[1]}$ and receive (U_a, Y_a)
    Update $\hat{\mu}_{a,f} \longleftarrow Y_a \frac{\hat{\mu}_{a,f} M_{a,f} + U_a}{M_{a,f}+1} + (1 - Y_a)\hat{\mu}_{a,f}, \;\; M_{a,f} \longleftarrow M_{a,f} + Y_a$
  end
end

In the Section 4.1 we describe the UCB computation method for the SB subroutine.

[2]Note that by pruning here we do not mean permanent pruning, it is used to describe that a particular firm is not consider at that time step

[3]The corresponding algorithmic subroutine AB_Subroutine is presented in the Appendix.

Finally, in Section 4.2, we illustrate how the matchings and collisions are used to update the probability $p_{a,f}(t)$ as per an AB subroutine.

4.1 Stochastic Bandit Subroutine

The stochastic bandit subroutine is used to efficiently deal with inherent uncertainty in the payoff obtained upon successful matching. In this section we develop the theory for the setting in which agents use a UCB based SB subroutine. Similar results for Thompson Sampling are supplied in the Appendix.

To being, we denote the number of times agent a gets successfully matched with firm f till time t as $M_{a,f}(t)$. Similarly, the number of times agent a gets collided with firm f till time t be $C_{a,f}(t)$. Given this notation, the UCB [Aue02] estimate of agent a for every f at time t is given by

$$\mathsf{UCB}_{a,f}(t) = \hat{\mu}_{a,f}(t-1) + \sqrt{\frac{2\log(1+\bar{M}_a \log^2(\bar{M}_a))}{M_{a,f}(t)}},$$

where $\bar{M}_a(t) = \sum_{f\in\mathcal{F}} M_{a,f}(t)$ and $\hat{\mu}_{a,f}(t-1)$ is the empirical average of the payoffs received from successfully matching to firm f until time t. The UCB estimate is composed of two parts: (i) the empirical mean which captures the exploitation aspect; and (ii) exploration bonus that decreases as $M_{a,f}(t)$ increases. We remark that it does not depend on the number of collisions $C_{a,f}(t)$.

4.2 Adversarial Bandit Subroutine

A key component of the proposed methodology is to use an adversarial bandit subroutine to deal with the competitive aspect of the problem. In particular, the AB subroutine updates the request probability $(p_{a,f})_{f\in\mathcal{F}}$ such that agent stops requesting firm on which the collisions are high (but ensures that it does not miss out on the firm if it is achievable). Intuitively, by construction, the adversarial bandit algorithm learns to prune arms on which collisions would happen frequently, and request firms where it is possible to successfully match very often. We show this by analyzing its regret and showing that high regret is incurred if the algorithm either prunes too often when successfully matching is possible or requesting a firm that is unachievable due to the frequent presence of higher ranked agents. By bounding the regret of the AB subroutine we immediately get a bound on the number of collisions.

We now describe the update scheme for $p_{a,f}(t)$ for any (a,f) at any time $t \in [T]$. In this work we consider an optimistic mirror descent based AB subroutine specialized from [BLLW19]. Interestingly such AB algorithms have data dependent regret bounds [WL18], [BLLW19] unlike other AB algorithms like Exp3 [LS20, Sli19]. Since the competition in the matching market is not actually adversarial such data-dependent regret bounds enables us characterize the competition more effectively in the analysis than just treating competition as adversarial[4]. We note that the proof techniques developed here can also be used to analyze an Exp3 based AB subroutine but the regret bounds of such an approach will not be as sharp.

For a given agent a, our algorithm associates a separate AB subroutine to every firm $f \in \mathcal{F}$. Each AB algorithm has *two arms* which correspond to the action of requesting

[4]We review the required background on optimistic mirror descent based AB algorithms in the Appendix along with a result which captures the characterizes the corresponding data-dependent regret bounds in the setting of matching markets.

the firm f or pruning it, each of which incurs different losses depending. In particular, if $P_{a,f}(t) = 0$ then it receives a fixed loss of 0; if $P_{a,f}(t) = 1$ the loss received is $+1$ or -1 if it collides or matches respectively. If we denote the loss received by the AB subroutine associated with (a, f) at time t by $L_{a,f}(t)$, we note that $L_{a,f}(t) = P_{a,f}(t)\,(1 - 2Y_a(t))$. Note that $Y_a(t)$ is unknown to any agent before requesting any firm as it also depends on the requests made by other agents.

We note that the request probability $p_{a,f}$ is not updated at every time t, but only when $E^{(\mathsf{c})}_{a,f}(t) = 1$ (i.e., if all firms with a higher UCB index have been pruned). As such the adversarial bandit algorithms can be seen as operating on a randomized timescale $\tau_{a,f}(T) = \{t \in [T] : E^{(\mathsf{c})}_{a,f}(t) = 1\}$ which are the time steps on which agent a considers firm f. We note that $p_{a,f}(t + 1) = p_{a,f}(t)$ if $t \notin \tau_{a,f}(T)$.

For the specific AB algorithm we analyze (which is a version of optimistic mirror descent with a log-barrier regularizer first analyzed in [WL18]), the simple setup of the losses leads to a closed form update for the probability of requesting or pruning a firm. In particular, for every $(a, f) \in \mathcal{A} \times \mathcal{F}$ and $t \in \tau_{a,f}(T)$, the optimistic mirror descent AB subroutine creates an unbiased estimate of the loss due to pruning and requesting as $\hat{L}^{(\mathsf{prune})}_{a,f}(t)$ and $\hat{L}^{(\mathsf{pull})}_{a,f}(t)$ respectively. In particular, if $P_{a,f}(t) = 1$

$$\hat{L}^{(\mathsf{prune})}_{a,f}(t) = \frac{1 + L_{a,f}(t-1)}{2}, \quad \hat{L}^{(\mathsf{pull})}_{a,f}(t) = \frac{1 - 2Y_a(t) - L_{a,f}(t-1)}{2p_{a,f}(t)} + \frac{1 + L_{a,f}(t-1)}{2}.$$

On the other hand, if $P_{a,f}(t) = 0$ then

$$\hat{L}^{(\mathsf{prune})}_{a,f}(t) = \frac{-L_{a,f}(t-1)}{2(1 - p_{a,f}(t))} + \frac{1 + L_{a,f}(t-1)}{2}, \quad \hat{L}^{(\mathsf{pull})}_{a,f}(t) = \frac{1 + L_{a,f}(t-1)}{2}$$

The term $\frac{1+L_{a,f}(t-1)}{2}$ is an optimistic prediction of the losses based on the last round of interaction [BLLW19]. Given these estimators the probability of requesting a firm is updated as:

$$p_{a,f}(t + 1) = (1 - \Lambda_{a,f}(t))x_{a,f}(t) + \Lambda_{a,f}(t)P_{a,f}(t),$$

where:

$$x_{a,f}(t) = \left(2 + \xi(t) - \sqrt{4 + \xi(t)^2}\right)(2\xi(t))^{-1}$$

for $\xi(t) = \eta\left(\hat{L}^{(\mathsf{pull})}_{a,f}(t) - \hat{L}^{(\mathsf{prune})}_{a,f}(t)\right) + \frac{1}{x_{a,f}(t-1)} - \frac{1}{1-x_{a,f}(t-1)}$, is the result of a step of mirror descent with the log-barrier regularizer, and $\Lambda_{a,f}(t) = \frac{\lambda(1-L_{a,f}(t))}{2+\lambda(1-L_{a,f}(t))}$, for $\lambda > 0$, promotes exploration. The algorithmic description of this process is stated in Algorithm 3.

5 Bounds on the regret of proposed algorithm

To capture the performance of the algorithm we use the natural notion of *stable regret* as introduced in [LMJ20]. More formally, the stable regret accrued by any agent $a \in \mathcal{A}$ is

$$\mathbb{E}[\mathcal{R}_a(T)] = \mathbb{E}\left[\sum_{t=1}^{T} u_{a,f^*_a} - \sum_{t=1}^{T} u_{a,f_a(t)}\right] \leq \sum_{f \in \mathbb{F}_a} \Delta_a(f)\mathbb{E}[M_{a,f}(T)] + u_a(f^*_a) \sum_{f \in \mathcal{F}} \mathbb{E}[C_{a,f}(T)], \tag{5.1}$$

Algorithm 3: AB_Subroutine

Input : $P_{a,f}, x_{a,f}, p_{a,f}, L_{a,f}, Y_a$

Parameters: $\eta \leq \frac{1}{50}, \lambda = 8\eta$

if $P_{a,f} = 0$ **then**

  Set $\hat{L}_{a,f}^{(\mathsf{prune})} = \frac{-L_{a,f}}{2(1-p_{a,f})} + \frac{L_{a,f}+1}{2}, \quad \hat{L}_{a,f}^{(\mathsf{pull})} = \frac{1+L_{a,f}}{2}$

  Update $L_{a,f} \longleftarrow 0$

end

if $P_{a,f} = 1$ **then**

  Set $\hat{L}_{a,f}^{(\mathsf{prune})} = \frac{1+L_{a,f}}{2}, \quad \hat{L}_{a,f}^{(\mathsf{pull})} = \frac{1-2Y_a-L_{a,f}}{2p_{a,f}} + \frac{1+L_{a,f}}{2}$

  Update $L_{a,f} \longleftarrow 1 - 2Y_a$

end

Set $\xi = \eta\left(\hat{L}_{a,f}^{(\mathsf{pull})} - \hat{L}_{a,f}^{(\mathsf{prune})}\right) + \frac{1}{x_{a,f}} - \frac{1}{1-x_{a,f}}$

Update $x_{a,f} \longleftarrow \frac{2+\xi-\sqrt{4+\xi^2}}{2\xi}$ and set $\Lambda_{a,f} = \frac{\lambda(1-L_{a,f})}{2+\lambda(1-L_{a,f})}$ Update $p_{a,f} \longleftarrow (1-\Lambda_{a,f})x_{a,f} + \Lambda_{a,f}P_{a,f}$

Output : $L_{a,f}, x_{a,f}, p_{a,f}$

where $\Delta_a(f) = u_a(f_a^*) - u_a(f)$ is the gap between the mean that agent a gets upon successfully matching with its stable match as compared firm f. If there are no collisions, then this regret definition is same as that used in stochastic bandits literature ([LS20]). In the following theorem, we present the regret of any agent using Algorithm 2:

Theorem 5. *Suppose every agent $a \in \mathcal{A}$ uses Algorithm 2. Then for any $i \in [K]$:*

$$\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \mathbb{E}[\mathcal{R}_a(T)] = \mathcal{O}\left(C_i |\mathcal{F}||\mathcal{A}| \log(T)\left(1 + \frac{1}{\Delta^2}\right)\right)$$

where $\Delta = \min_{a,f} \Delta_{a,f}$ and C_i is a constant dependent on market $\mathcal{M}_i$ and $C_1 < C_2 < ... < C_K$.

We see that the regret of any agent $a \in \mathcal{A}$ is logarithmic in horizon T, which matches the lower bound for single player stochastic bandit algorithms [LR85]. As such, perhaps surprisingly, we observe that in α-reducible markets, it is possible for agents to learn while competing without incurring drastically worse regret in the long run. It is interesting to note that the learning of agent depends on its position in the market as per preferences (Remark 3). An agent low in the hierarchy incurs more regret during the learning process due to the agents higher up in the hierarchy driven mainly by the larger number of collisions incurred while waiting for agents higher in the hierarchy to stop exploring. We note that in the worst case the constant C_i can grow exponentially in the number of agents in the market. We note that this is a consequence of the proof technique and not fundamental limitation of the algorithmic design paradigm as we show through numerical studies in next section. We leave this as a future work to establish tighter regret bounds in terms of number of agents. In the Appendix we also show that in Algorithm 2 if we use a SB subroutine based on Thompson Sampling then a similar regret guarantee can be obtained. We now present a sketch of the proof of Theorem 5.

Sketch of the proof. Before presenting the sketch, we first define few notations that would make the exposition clear. Let $M_{a,\underline{\mathbb{F}}_a}(T) = \sum_{f \in \underline{\mathbb{F}}_a} M_{a,f}(T), M_{a,\bar{\mathbb{F}}_a}(T) = \sum_{f \in \bar{\mathbb{F}}_a} M_{a,f}(T)$.

Moreover, for any $a \in \mathcal{A}$ define $H_{a,f_a^*}(t) = \{\exists a' \in \mathcal{A} \text{ s.t. } u_{f_a^*}(a') \geq u_{f_a^*}(a), f_{a'}(t) = f\}$ which is an event that characterizes if any other more preferred agent has requested the stable match of agent a at time t. Against the preceding backdrop, we now present the following crucial lemma:

Lemma 6. *Suppose every agent uses Algorithm 2 then the following holds:*

(L1) *For any $i \in [K]$, the cumulative regret can be decomposed as*

$$\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \mathbb{E}[\mathcal{R}_a(T)] = \mathcal{O}\Big(\sum_{i=1}^{k} \sum_{a \in \mathcal{A}_i} (\mathbb{E}[M_{a,\underline{\mathbb{F}}_a}(T)] + \sum_{\substack{f \in F \\ f \neq \{f_a^*\}}} \mathbb{E}[C_{a,f}(T)] + \mathbb{E}[\sum_{t=1}^{T} H_{a,f_a^*}(t)]) \Big);$$

(L2) *For any $i \in [K]$, the expected matches with suboptimal firm satisfies*

$$\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \mathbb{E}[M_{a,\underline{\mathbb{F}}_a}(T)] = \mathcal{O}\left(\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \left(|\underline{\mathbb{F}}_a| \log(T) \left(1 + \frac{1}{\Delta^2}\right) + \mathbb{E}\left[\sum_{t=1}^{T} H_{a,f_a^*}(t) \right] \right) \right)$$

(L3) *The expected number of collisions between for any agent $a \in \mathcal{A}$ satisfies*

$$\sum_{f \in \mathcal{F}} \mathbb{E}[C_{a,f}(T)] = \mathcal{O}\left(|\mathcal{F}| \log(T) + \mathbb{E}\left[M_{a,\underline{\mathbb{F}}_a}(T) + M_{a,\bar{\mathbb{F}}_a}(T) + \sum_{t=1}^{T} \mathbb{1}\left(H_{a,f_a^*}(t)\right) \right] \right);$$

(L4) *For any $i \in [K]$ we have*

$$\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f_a^*}(t)\right) \right] = \mathcal{O}\left(C_i \left(\sum_{j=1}^{i} |\mathcal{A}_j| \right) \log(T) \left(1 + \frac{1}{\Delta^2}\right) \right),$$

where C_i is a constant dependent on market $\mathcal{M}_i$ such that $C_1 < C_2 < ... < C_K$.

(L5) *For any $i \in [K]$ we have*

$$\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \sum_{f \in \bar{\mathbb{F}}_a} \mathbb{E}[M_{a,f}(T)] \leq \mathcal{O}\left(C_i \left(\sum_{j=1}^{i} |\mathcal{A}_j| \right) |\mathcal{F}| \log(T) \left(1 + \frac{1}{\Delta^2}\right) \right)$$

Theorem 5 is proved using **(L1)**-**(L5)** from Lemma 6. Note that **(L1)** follows from (5.1) and the definition of $H_{a,f_a^*}(t)$. From **(L1)** we see that to bound the regret we need to consider three components: (i) expected number of matchings with suboptimal firms, (ii) expected number of collisions with any firm other than stable match, (iii) the *potential collisions* at the stable match[5]. **(L2)** bounds the expected number of matchings with suboptimal firms. Note that the total matchings between agent a and firm f is $M_{a,f}(T) = \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f\right)$. Thus, we present the following lemma which plays a key role in the proof of **(L2)**:

Lemma 7. *The event that agent a chooses the firm $f \in \underline{\mathbb{F}}_a$ and successfully matches at time $t \in [T]$ satisfies*

$$\{Y_a(t) = 1, f_a(t) = f\} \subset \left\{Y_a(t) = 1, \mathsf{UCB}_{a,f_a^*}(t) \leq \mathsf{UCB}_{a,f}(t)\right\} \cup \{E_{a,f}^{(r)}(t) = 1, E_{a,f_a^*}^{(r)}(t) = 0\}$$

[5]by potential collision at stable match we mean total number of collision that would have been faced by an agent at its stable firm had it always requested the stable firm

Lemma 7 separates the challenge associated with uncertainty and that of competition. Note that the first event on the right hand side is the one which is standard to the analysis of UCB algorithm ([LS20]). Meanwhile, the other event corresponds to the case when the stable firm is pruned by agent a in order to avoid potential collisions. To bound latter event we use the regret bounds for the adversarial bandit subroutine (refer to Appendix).

To bound **(L3)** we use the path length based regret bounds from [BLLW19], [WL18] for the adversarial bandit subroutine. Meanwhile to bound **(L4)** we use the $\alpha-$reduciblity assumption and **(L2)**. In particular, the $\alpha-$reduciblity assumption induces a hierarchy in the market as per Remark 3. This decomposition reduces the bound in **(L4)** to appropriate accounting of number of matches with suboptimal firms via an induction argument. Finally, **(L5)** follows again due to hierarchy induced by $\alpha-$reducibility and using **(L2)-(L4)**.

6 Experimental Study

In this section we present the numerical experiments that demonstrates and validates the results presented in this paper. Moreover, we also observe that our algorithm performs surprisingly well in general market structure, that is in markets which are not $\alpha-$reducible. We leave this as a future work to establish the regret bounds for the proposed algorithms in general markets.

In both sets of experiments, we consider a market comprising of 5 agents and 5 firms. We consider the following two settings:

(S-I). randomly initialized preference for agents and randomly initialized (but uniform) preference for firms. This setting ensures that market is $\alpha-$reducible

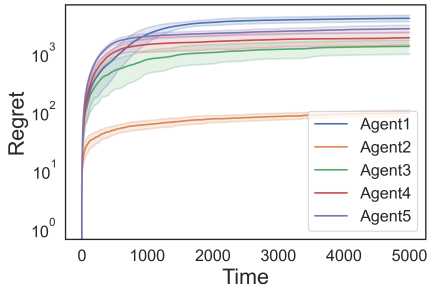
(S-II). randomly initialized preference for agents and firms. In this part we specifically consider setting where $\alpha-$reducibility does *not* hold. This would provide directions for future research in this area.

In our simulations for every agent we randomly sample the preference ordering of firms and assign a mean reward in $[0, 5]$ such that the successful match with most preferred firm gives mean reward 5 and the least preferred firm gives the mean reward 0 and the mean rewards from other firms are equally spaced between $[0, 5]$. The rewards follow a normal distribution with variance 1. We run both Algorithm 2 and Algorithm 5 for 25 times for two randomly sampled preference ordering for each of **(S-I)-(S-II)**.

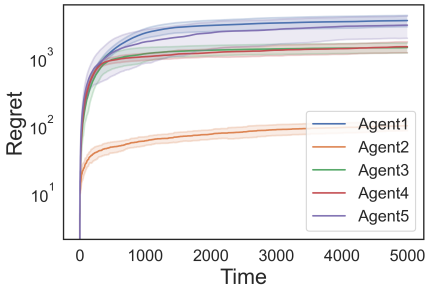
In Figure 1 we consider **(S-I)** and observe the performance of algorithms. We observe that the mean regret (taken over 25 runs) accumulated by the algorithms saturate very quickly and agents identify their stable match. In Figure 2 we consider **(S-II)** and observe the performance of algorithm. Surprisingly, even without the $\alpha-$reducibility structure, the mean regret[6] (taken over 25 runs) accumulated by the algorithms saturate very quickly and agents identify their stable match. This presents an opportunity to further explore the algorithm presented in this paper for general markets.

Furthermore, in both **(S-I)-(S-II)** we observe that the TS-DMA has higher variance but is faster than UCB-DMA. This is because, compared to UCB-DMA, we observe empirically that TS-DMA very rarely encounters the scenario where all of the firms gets pruned by the adversarial bandit module. We would also like to point that in some cases the regret can be negative (which is desirable) as is shown in Figure 1(c) for the red agent.

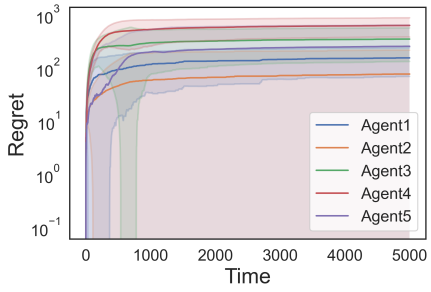
[6] mean regret here refers to the agent-optimal stable regret[LRMJ21]



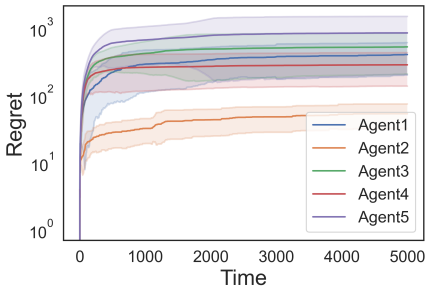
(a) UCB-DMA(Algorithm 2)



(b) UCB-DMA(Algorithm 2)



(c) TS-DMA(Algorithm 5)

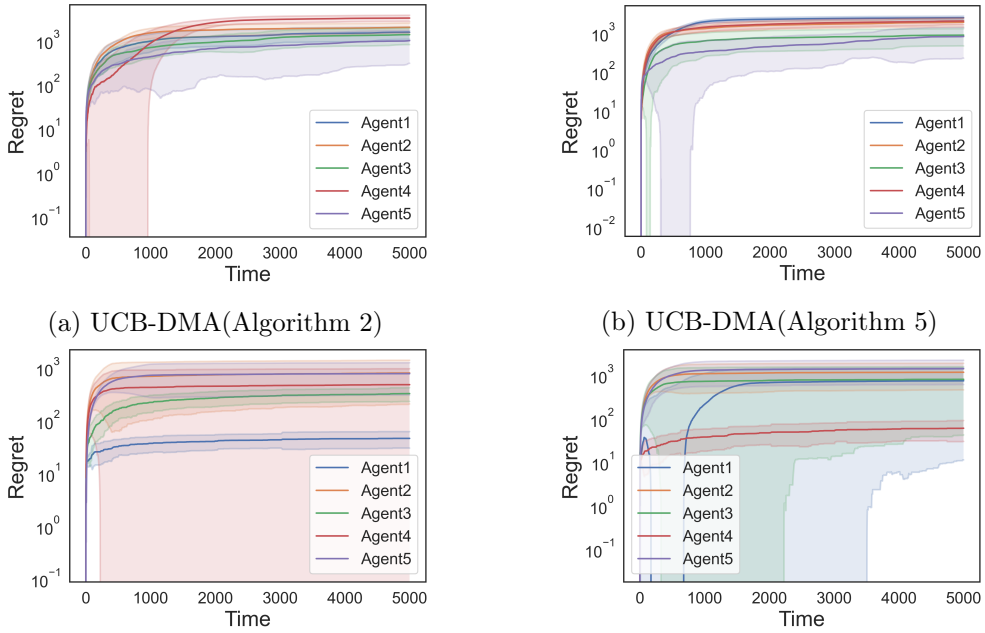


(d) TS-DMA(Algorithm 5)

Figure 1: Performance of UCB-DMA (Algorithm 2) and TS-DMA(Algorithm 5) where $\alpha-$reducibilty condition is satisfied. We simulated the algorithm for two randomly generated preference orderings which satisfy the α-reducibility condition. The simulation results of one of the preference ordering are presented in left column and for the other in right column. The bold lines and the corresponding shaded region denotes the mean regret and the variance of regret for the agents over 25 runs of the algorithm.

7 Conclusions

We consider a problem of bandit learning in two-sided matching markets comprising of agents and firms. We consider the setting where agents have unknown preferences over the firms. In this paper we present simple design principle for decentralized, communication and coordination free algorithm for learning in two-sided matching markets. The primary challenge in learning in two-sided matching market is to balance exploration, exploitation and collision avoidance. We embed the aforementioned properties in the algorithm by a novel idea of blending a stochastic bandit subroutine with an adversarial bandit subroutine . The stochastic bandit subroutine is required for balancing the exploration-exploitation trade-off while the adversarial bandit subroutine limits the collisions. As an instance of this design principle, we present an algorithm which has the stochastic bandit subroutine based on UCB and the adversarial bandit subroutine based on Optimistic Mirror Descent algorithm. We show that if the preferences of agents satisfy certain structure known as α-reducibility, then these algorithms incur a regret which is logarithmic in the time horizon. Two immediate directions of future work include: (i) extension of theoretical guarantees to general markets, and (ii) improving the dependence of regret bound on the number of agents.



(a) UCB-DMA(Algorithm 2)

(b) UCB-DMA(Algorithm 5)

(c) TS-DMA(Algorithm 5)

(d) TS-DMA(Algorithm 5)

Figure 2: Performance of UCB-DMA (Algorithm 2) and TS-DMA(Algorithm 5) where $\alpha-$reducibilty condition is NOT satisfied. We simulated the algorithm for two randomly generated preference orderings which satisfy the α-reducibility condition. The simulation results of one of the preference ordering are presented in left column and for the other in right column. The bold lines and the corresponding shaded region denotes the mean regret and the variance of regret for the agents over 25 runs of the algorithm.

Acknowledgements

Research was partially supported by NSF under grant DMS 2013985 THEORINet: Transferable, Hierarchical, Expressive, Optimal, Robust and Interpretable Networks and U.S. Office of Naval Research MURI grant N00014-16-1- 2710.

References

[AG12] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

[Alc94] José Alcalde. Exchange-proofness or divorce-proofness? stability in one-sided matching markets. *Review of Economic Design*, 1:275–287, 02 1994.

[AMSW20] Guy Aridor, Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Competing bandits: The perils of exploration under competition. *arXiv preprint arXiv:2007.10144*, 2020.

[Aue02] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[BBS20] Sébastien Bubeck, Thomas Budzinski, and Mark Sellke. Cooperative and stochastic multi-player multi-armed bandit: Optimal regret with neither communication nor collisions. *CoRR*, abs/2011.03896, 2020.

[BLLW19] Sébastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved path-length regret bounds for bandits. In *Conference On Learning Theory*, pages 508–528. PMLR, 2019.

[BSS21] Soumya Basu, Karthik Abinav Sankararaman, and Abishek Sankararaman. Beyond log-squared regret for decentralized bandits in matching markets. *arXiv preprint arXiv:2103.07501*, 2021.

[BTZ15] Swapna Buccapatnam, Jian Tan, and Li Zhang. Information sharing in distributed stochastic bandits. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 2605–2613. IEEE, 2015.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[CCDJ17] Mithun Chakraborty, Kai Yee Phoebe Chua, Sanmay Das, and Brendan Juba. Coordinated versus decentralized exploration in multi-agent multi-armed bandits. In *IJCAI*, pages 164–170, 2017.

[Cla06] Simon Clark. The uniqueness of stable matchings. *Contributions to Theoretical Economics*, 6:1283–1283, 02 2006.

[CS21] Sarah H Cen and Devavrat Shah. Regret, stability, and fairness in matching markets with bandit learners. *arXiv preprint arXiv:2102.06246*, 2021.

[DK05] Sanmay Das and Emir Kamenica. Two-sided bandits and the dating market. In *IJCAI*, volume 5, page 19. Citeseer, 2005.

[FDLL98] Drew Fudenberg, Fudenberg Drew, David K Levine, and David K Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

[GS62] David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

[JKK16] Ramesh Johari, Vijay Kamble, and Yash Kanoria. Matching while learning. *arXiv preprint arXiv:1603.04549*, 2016.

[JWW+21] 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, 2021.

[KNJ14] Dileep Kalathil, Naumaan Nayyar, and Rahul Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

[KYL22] Fang Kong, Junming Yin, and Shuai Li. Thompson sampling for bandit learning in matching markets. *arXiv preprint arXiv:2204.12048*, 2022.

[Lit94] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[LM21] Gábor Lugosi and Abbas Mehrabian. Multiplayer bandits without observing collision information. *Mathematics of Operations Research*, 2021.

[LMJ20] Lydia T Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pages 1618–1628. PMLR, 2020.

[LR85] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[LRMJ21] 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.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[LZ10] Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE transactions on signal processing*, 58(11):5667–5681, 2010.

[MSW17] Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Competing bandits: Learning under competition. *arXiv preprint arXiv:1702.08533*, 2017.

[RRKO17] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, and Ian Osband. A tutorial on thompson sampling. abs/1707.02038, 2017.

[RSS16] Jonathan Rosenski, Ohad Shamir, and Liran Szlak. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*, pages 155–163. PMLR, 2016.

[SBS21] Abishek Sankararaman, Soumya Basu, and Karthik Abinav Sankararaman. Dominate or delete: Decentralized competing bandits in serial dictatorship. In *International Conference on Artificial Intelligence and Statistics*, pages 1252–1260. PMLR, 2021.

[SGS19] Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–35, 2019.

[Sli19] Aleksandrs Slivkins. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.

[Tho33] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[WL18] Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291. PMLR, 2018.

Appendix

In Section A, we review the adaptive adversarial algorithms proposed in [BLLW19] and specialize the regret bounds in the setup of this paper. In Section B we provide the proof

of lemmas stated in Section 5. In Section C we provide proof of the main theorem of this paper stated in Section 5. In Section E we provide the Thompson sampling based variant of the Algorithm 2 and provide the analogous result as in Section 5. In Section F we provide a table of notations for ease of reference.

A Adaptive Adversarial Algorithms

In this work we deploy the optimistic mirror descent based adversarial bandit module. We adapt algorithms from [BLLW19], who improve the algorithm originally proposed in [WL18]. In this section we recap the results from [BLLW19]. For the sake of completeness we restate the problem formulation and algorithm here. Towards the end we will specialize their results in the setting of this paper and state an useful result which presents the regret of such algorithms, in the context of the bandit structure described in Sec 4.2, in terms of the number of matchings and collisions.

A.1 Problem formulation from [BLLW19]

In this section we review algorithm described in [BLLW19] which is an improvement over the one described in [WL18]. Consider a multi-armed bandit problem that proceeds in τ time steps with $A \leq \tau$ fixed actions. In each round t, the algorithm selects one arm $i(t) \in [A]$ and simultaneously an adversary decides the loss vector $\ell(t) = (\ell_i(t))_{i\in[A]} \in [-1, 1]^A$. Note that the adversary can be an adaptive one in that it can base its actions on the past rounds of algorithm's actions. The goal of the algorithm is to minimize the gap between total accumulated loss and the loss of best fixed arm in hindsight:

$$\text{Regret}^{(\mathsf{adv})}(\tau) = \max_{i^\star\in[A]} \mathbb{E}\left[\sum_{t=1}^{\tau} \ell_{i(t)}(t) - \sum_{t=1}^{\tau} \ell_{i^\star}(t)\right].$$

The algorithm is based on the optimistic mirror descend framework. At any time t, the algorithm samples an arm $i(t) \in [A]$ with probability $p(t) \in \Delta([A])$. The algorithm only receives the loss for the action taken and not other actions. Therfore, upon receiving the loss $\ell_{i(t)}(t)$ the algorithm creates an unbiased estimator of losses for other actions. The estimator is

$$\hat{L}_i(t) = \frac{\ell_i(t) - L(t-1)}{2p_i(t)} \mathbb{1}\left(i(t) = i\right) + \frac{1 + L(t-1)}{2}, \quad \forall\ i$$

The unbiased loss estimate $\hat{L}(t)$ is used to update the an auxiliary probability distribution $x(t+1) \in \Delta([A])$ through an optimistic mirror descend update with learning rate η. The optimistic mirror descend update is constructed from the Bregman divergence[7] associated with a log-barrier regularizer $\mathbb{R}^A \ni x \mapsto \psi(x) = \frac{1}{\eta}\sum_{i=1}^{A} \ln \frac{1}{x_i}$ as follows

$$x(t+1) = \arg\min_{z\in\Delta([A])} \langle z, \hat{L}(t)\rangle + D_\psi(z, x(t)).$$

The distribution $x(t+1)$ is used to update the arm sampling distribution $p(t+1)$ after mixing a small bias towards most recently picked arm as follows

$$p(t+1) = (1 - \lambda(t+1))x(t+1) + \lambda(t+1)\mathbf{e}_{i(t)}$$

Algorithm 4: Optimistic Mirror Descend based Adversarial Bandit Algorithm

Parameters: $\eta, \lambda \in (0,1), p(1), x(1) = \mathsf{Unif}([A])$, $\psi(x) = \frac{1}{\eta}\sum_{i=1}^{A} \ln \frac{1}{x_i}$

1 **for** $t = 1, 2, .., \tau$ **do**

2 Play $i(t) \sim p(t)$ and observe $L(t) = \ell_{i(t)}(t)$

3 Construct an unbiased estimator $\hat{L}_i(t) = \frac{\ell_i(t) - L(t-1)}{2p_i(t)} \mathbb{1}\left(i(t) = i\right) + \frac{1+L(t-1)}{2}$ for all $i \in [A]$

4 Update $x(t+1) = \arg\min_{z \in \Delta([A])} \langle z, \hat{L}(t) \rangle + D_\psi(z, x(t))$

5 $p(t+1) = (1 - \lambda(t+1))x(t+1) + \lambda(t+1)\mathbf{e}_{i(t)}$ where $\lambda(t+1) = \frac{\lambda(1-L(t))}{2+\lambda(1-L(t))}$

6 **end**

where $\mathbf{e}_{i^t} \in \mathbb{R}^A$ is an element of standard basis in $\mathbb{R}^A$ with $i(t)$ element as 1 and all others as zero and $\lambda(t+1) = \frac{\lambda(1-L(t))}{2+\lambda(1-L(t))}$ for some $\lambda > 0$.

Against the preceding backdrop, we restate Theorem 2 from [BLLW19] below:

Theorem 8. *Algorithm 4 with $\eta \leq \frac{1}{50}$, $\lambda = 8\eta$ ensures that*

$$\mathit{Regret}^{(\mathsf{adv})}(\tau) = \mathcal{O}\left(\frac{A\ln(T)}{\eta}\right) + 8\eta\mathbb{E}\left[V(T)\right]$$

where $V(T) \coloneqq \sum_{t=2}^{T} |\ell_{i(t-1)}(t) - \ell_{i(t-1)}(t-1)|$ is commonly referred as "path-length".

Remark 9. *Note that Theorem 2 in [BLLW19] requires*[8]*. But in fact the proof goes through for $\eta \leq 1/50$. $\eta \leq 1/162$ and $\lambda = 8\eta$. This is because in [BLLW19] for the proof of Theorem 2, they directly lift [WL18, Theorem 7] where $\eta \leq 1/162$ which is not tuned efficiently.*

A.2 Adaptive Adversarial Module

In this section we describe AB_Subroutine in Algorithm 2 which is based on the algorithm presented in Sec A.1.

For any $(a, f) \in \mathcal{A} \times \mathcal{F}$, the adversarial bandit module associated with (a, f) (as described in Algorithm 3) is a version of Algorithm 4 for case when there are two actions: *request the firm f* or *prune the firm f*. In addition, the loss incurred due to pruning the firm f is always 0 while the loss incurred due to pulling an firm f depends on whether the agent a got matched with it or collided with it. In this special case of two actions, the optimistic mirror descent update (line 4 in Algorithm 4) can be obtained in closed form (see Lemma 11). Note that the adversarial bandit module associated with any agent-firm tuple (a, f) is only used when $t \in \tau_{a,f}(T) \subset [T]$.

Lemma 10. *Given a scalar $\eta \leq \frac{1}{50}$, for any agent-firm pair $(a, f) \in \mathcal{A} \times \mathcal{F}$, the regret of the adversarial bandit algorithm is bounded as*

$$\mathbb{E}[\mathit{Regret}_{a,f}^{(\mathsf{adv})}(\tau_{a,f}(T))] \leq \mathcal{O}\left(\frac{\log(T)}{\eta}\right) + 32\eta\mathbb{E}\left[\min\left\{M_{a,f}^{\star}(T), C_{a,f}^{\star}(T), M_{a,f}(T) + C_{a,f}(T)\right\}\right],$$

where $M_{a,f}^{\star}(T) = \sum_{t=1}^{T} \mathbb{1}\left(H_{a,f}^{c}(t)\right)$ and $C_{a,f}^{\star}(T) = \sum_{t=1}^{T} \mathbb{1}\left(H_{a,f}(t)\right)$.

[7] Bregman divergence between two point x, y with respect to a convex regularizer ψ is given as $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle$.

[8] Moreover, it is an algebraic exercise to establish that $\eta < \frac{1}{24}$ and $\lambda = \frac{1 - 12\eta - c\cdot\sqrt{1-24\eta}}{24}$ also works for some $c \in (0,1)$. But we don't go in this direction to retain simplicity of algorithmic description.

Proof. To prove this lemma we only need to bound the path length $V_{a,f}(T)$ in Theorem 8. We claim that the path length $V_{a,f}(T) \leq \min\{C^{\star}_{a,f}(T), M^{\star}_{a,f}(T)\}$. Recall $\tau_{a,f}(T) = \{t \in T : E^{(\mathsf{c})}_{a,f}(t) = 1\}$. For the remaining proof for any $t \in \tau_{a,f}(T)$ by $t-1$ we mean $\max\{\mathfrak{t} < t : \mathfrak{t} \in \tau_{a,f}(T)\}$. For any $t \in \tau_{a,f}(T)$, let's denote the loss due to pruning at time t by $\ell^{(prune)}_{a,f}(t)$ and similarly let the loss due to pulling at time t by $\ell^{(pull)}_{a,f}(t)$. Note that by construction, the loss due to the pruning operation is deterministic and zero. That is, for any $t \in \tau_{a,f}(T)$, $\ell^{(prune)}_{a,f}(t) = 0$ and $\ell^{(pull)}_{a,f}(t) = 1 - 2Y_a(t)$. Furthermore, note that

$$
\begin{aligned}
V_{a,f}(T) &\leq \sum_{t \in \tau_{a,f}(T)} |\ell^{(pull)}_{a,f}(t) - \ell^{(pull)}_{a,f}(t-1)| \\
&\underset{(a)}{\leq} 2 \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(H_{a,f}(t-1), H^{\mathsf{c}}_{a,f}(t)\right) + \mathbb{1}\left(H^{\mathsf{c}}_{a,f}(t-1), H_{a,f}(t)\right) \\
&\leq 4 \min\left\{ \sum_{t=1}^{T} \mathbb{1}\left(H^{\mathsf{c}}_{a,f}(t)\right), \sum_{t=1}^{T} \mathbb{1}\left(H_{a,f}(t)\right) \right\} \\
&= 4 \min\left\{ M^{\star}_{a,f}(T), C^{\star}_{a,f}(T) \right\}
\end{aligned}
$$

where the factor of 2 in is by the fact that a path length change in going from matching to potential collision or collision to potential matching is 2. The remaining inequalities follow from algebra.

Furthermore, we have

$$
\begin{aligned}
V_{a,f}(T) &= \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 1, P_{a,f}(t-1) = 1\right) |\ell^{(pull)}_{a,f}(t) - \ell^{(pull)}_{a,f}(t-1)| \\
&\quad + \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 0, P_{a,f}(t-1) = 1\right) |\ell^{(pull)}_{a,f}(t) - \ell^{(pull)}_{a,f}(t-1)| \\
&\leq \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 1, P_{a,f}(t-1) = 1\right) |\ell^{(pull)}_{a,f}(t) - \ell^{(pull)}_{a,f}(t-1)| \\
&\quad + 2 \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 0, P_{a,f}(t-1) = 1\right) \\
&= \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 1, P_{a,f}(t-1) = 1\right) |\ell^{(pull)}_{a,f}(t) - \ell^{(pull)}_{a,f}(t-1)| \\
&\quad + 2 \sum_{t=2}^{T} \mathbb{1}\left(P_{a,f}(t) = 0, P_{a,f}(t-1) = 1\right)
\end{aligned}
$$

$$
\begin{aligned}
&= 2 \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 1, P_{a,f}(t-1) = 1, Y_a(t) = 0, Y_a(t-1) = 1\right) \\
&\quad + 2 \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 1, P_{a,f}(t-1) = 1, Y_a(t) = 1, Y_a(t-1) = 0\right) \\
&\quad + 2 \sum_{t \in \tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t) = 0, P_{a,f}(t-1) = 1\right)
\end{aligned}
$$

$$\leq 2\left(\sum_{t\in\tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t)=1, Y_a(t)=0\right) + \mathbb{1}\left(P_{a,f}(t-1)=1, Y_a(t-1)=1\right)\right)$$
$$+ 2\sum_{t\in\tau_{a,f}(T)} \mathbb{1}\left(P_{a,f}(t)=0, P_{a,f}(t-1)=1\right)$$
$$\leq 4\left(M_{a,f}(T) + C_{a,f}(T)\right)$$

□

A.3 Technical Lemma

Lemma 11. *For any $L \in \mathbb{R}^2$ and $X \in \Delta(\mathbb{R}^2)$ the update $X_+ = \arg\min_{Z\in\Delta(\mathbb{R}^2)}\langle Z, L\rangle + D_\psi(Z, X)$ can be analytically solved to be $X_+ = [x_+, 1-x_+]$ where*

$$x_+ = \frac{2+\xi-\sqrt{4+\xi^2}}{2\xi} \tag{A.1}$$

where $\xi = \eta(L_1 - L_2) + \frac{1}{X_1} - \frac{1}{X_2}$. For better interpretation we provide the graph for update (A.1) *in the Figure 3.*

Proof. For any $X, Z \in \Delta(\mathbb{R}^2)$ we represent $X = [x, 1-x]$ and $Z = [z, 1-z]$ for $x, z \in [0,1]$. Under this notation we can write $D_\psi(Z, X) = \frac{1}{\eta}\left(\log\left(\frac{x}{z}\right) + \log\left(\frac{1-x}{1-z}\right) + \frac{z-x}{x} + \frac{x-z}{1-x}\right)$. Thus the optimization problem becomes

$$\begin{aligned} x_+ &= \arg\min_{z\in[0,1]} \langle z, L\rangle + D_\psi(z, X) \\ &= \arg\min_{z\in[0,1]} zL_1 + (1-z)L_2 + \frac{1}{\eta}\left(\log\left(\frac{x}{z}\right) + \log\left(\frac{1-x}{1-z}\right) + \frac{z-x}{x} + \frac{x-z}{1-x}\right) \\ &= \arg\min_{z\in[0,1]} zL_1 + (1-z)L_2 + \frac{1}{\eta}\left(-\log(z) - \log(1-z) + \frac{z}{x} - \frac{z}{1-x}\right) \end{aligned}$$

Let $f(z) = zL_1 + (1-z)L_2 + \frac{1}{\eta}\left(-\log(z) - \log(1-z) + \frac{z}{x} - \frac{z}{1-x}\right)$. Note that $f(0) = +\infty$, and $f(1) = +\infty$ so the minimizer of $f(z)$ lies stricly inside $[0,1]$. Therefore $\nabla f(x_+) = 0$. We compute

$$\nabla f(z) = L_1 - L_2 + \frac{1}{\eta(1-z)} - \frac{1}{\eta z} + \frac{1}{\eta x} - \frac{1}{\eta(1-x)} = L_1 - L_2 + \frac{2z-1}{\eta z(1-z)} + \frac{1}{\eta x} - \frac{1}{\eta(1-x)}$$

Imposing the condition $\nabla f(x_+) = 0$ implies that

$$\xi x_+^2 - (2+\xi)x_+ + 1 = 0$$

where $\xi = \eta(L_1 - L_2) + \frac{1}{x} - \frac{1}{1-x}$. Thus there are two possibilities

$$x_+ = \frac{2+\xi+\sqrt{4+\xi^2}}{2\xi}, \quad \text{or} \quad x_+ = \frac{2+\xi-\sqrt{4+\xi^2}}{2\xi},$$

However the first possibility implies that $x_+ > 1$, thus the only solution which lies in $(0,1)$ is the latter. This completes the proof. □

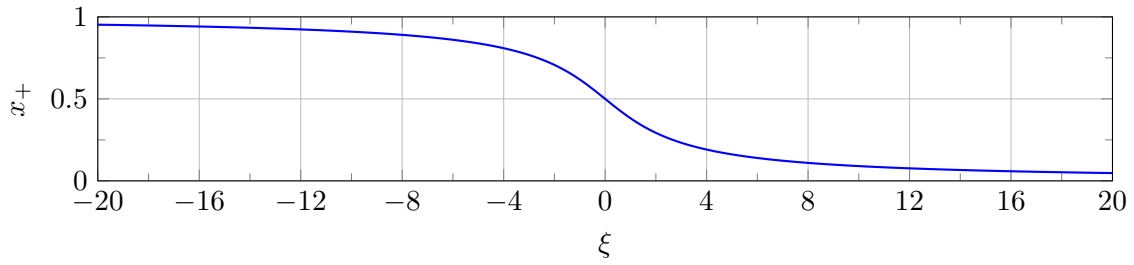


Figure 3: Update function of pulling probability based on line 10 in Algorithm 3

B Proofs of main Lemmas

We introduce the following notation for every $a \in \mathcal{A}, f \in \mathcal{F}$

$$H_{a,f}(t) = \mathbb{1}\left(\exists a' \in \mathcal{A} : f_{a'}(t) = f, u_f(a') > u_f(a)\right),$$

which characterizes an event some agent more preferred than a by firm f has requested firm f. We now present the proofs of Lemmas in main paper in the following subsections.

B.1 Proof of Lemma 7

Proof of Lemma 7 follows directly from the following Lemma.

Lemma 12. *The event that agent a chooses a firm $f \in \mathcal{F}$ at time $t \in [T]$ satisfies*

$$\{Y_a(t) = 1, f_a(t) = f\} \subset \left\{Y_a(t) = 1, \mathit{UCB}_{a,f_a^*}(t) \leq \mathit{UCB}_{a,f}(t)\right\} \bigcup \left\{E_{a,f}^{(r)}(t) = 1, E_{a,f_a^*}^{(r)}(t) = 0\right\}. \tag{B.1}$$

Proof. For any agent a fix some f. Recall that $f_a(t) = f$ implies that agent a has chosen to pull arm f. Based on design of Algorithm 2 there are two possibilities: either all the firms with higher UCB than firm f got pruned and the firm f was requested; or all of the firms in $\mathcal{F}$ got pruned and the firm f got selected as it was having highest UCB. Thus,

$$\begin{aligned}
\{f_a(t) = f\} &= \left\{E_{a,f}^{(\mathsf{r})}(t) = 1\right\} \bigcup \left\{E_{a,f}^{(\mathsf{r})}(t) = 0 \ \forall\ f \in \mathcal{F}, \mathsf{UCB}_{a,f} \geq \mathsf{UCB}_{a,f'} \ \forall\ f' \in \mathcal{F}\right\} \\
&\underset{(i)}{=} \left\{E_{a,f}^{(\mathsf{r})}(t) = 1, \mathsf{UCB}_{a,f_a^*}(t) \geq \mathsf{UCB}_{a,f}(t)\right\} \bigcup \left\{E_{a,f}^{(\mathsf{r})}(t) = 1, \mathsf{UCB}_{a,f_a^*}(t) \leq \mathsf{UCB}_{a,f}(t)\right\} \\
&\quad \bigcup \left\{E_{a,f}^{(\mathsf{r})}(t) = 0 \ \forall\ f \in \mathcal{F}, \mathsf{UCB}_{a,f} \geq \mathsf{UCB}_{a,f'} \ \forall\ f' \in \mathcal{F}\right\} \\
&\underset{(ii)}{\subset} \left\{E_{a,f}^{(\mathsf{r})}(t) = 1, \mathsf{UCB}_{a,f_a^*}(t) \geq \mathsf{UCB}_{a,f}(t)\right\} \bigcup \left\{E_{a,f}^{(\mathsf{r})}(t) = 1, \mathsf{UCB}_{a,f_a^*}(t) \leq \mathsf{UCB}_{a,f}(t)\right\} \\
&\quad \bigcup \left\{\mathsf{UCB}_{a,f_a^*}(t) \leq \mathsf{UCB}_{a,f}(t)\right\} \\
&\underset{(iii)}{\subset} \left\{E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f_a^*}^{(\mathsf{r})}(t) = 0, \mathsf{UCB}_{a,f_a^*}(t) \geq \mathsf{UCB}_{a,f}(t)\right\} \\
&\quad \bigcup \left\{E_{a,f}^{(\mathsf{r})}(t) = 1, \mathsf{UCB}_{a,f_a^*}(t) \leq \mathsf{UCB}_{a,f}(t)\right\} \bigcup \left\{\mathsf{UCB}_{a,f_a^*}(t) \leq \mathsf{UCB}_{a,f}(t)\right\} \\
&\underset{(iv)}{\subset} \left\{E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f_a^*}^{(\mathsf{r})}(t) = 0, \mathsf{UCB}_{a,f_a^*}(t) \geq \mathsf{UCB}_{a,f}(t)\right\} \bigcup \left\{\mathsf{UCB}_{a,f_a^*}(t) \leq \mathsf{UCB}_{a,f}(t)\right\} \\
&\underset{(v)}{\subset} \left\{E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f_a^*}^{(\mathsf{r})}(t) = 0\right\} \bigcup \left\{\mathsf{UCB}_{a,f_a^*}(t) \leq \mathsf{UCB}_{a,f}(t)\right\}
\end{aligned}$$

where in (i) we introduced two complementary events $\{\mathsf{UCB}_{a,f_a^*}(t) \geq \mathsf{UCB}_{a,f}(t)\}$ and $\{\mathsf{UCB}_{a,f_a^*}(t) \leq \mathsf{UCB}_{a,f}(t)\}$. Note that (ii) holds due to the fact that $\{\mathsf{UCB}_{a,f_a(t)} \geq \mathsf{UCB}_{a,f} \ \forall \ f \in \mathcal{F}\}$ implies $\{\mathsf{UCB}_{a,f_a(t)} \geq \mathsf{UCB}_{a,f_a^*}\}$. Furthermore, (iii) holds due to the fact that a firm with lower UCB will be pulled only if all the firms with higher UCB are pruned. Finally, $(iv), (v)$ holds by dropping appropriate events.

The result follows by noting that

$$\begin{aligned}
&\mathbb{1}\left(Y_a(t)=1, f_a(t)=f\right) \\
&\subset \left(\left\{E^{(\mathsf{r})}_{a,f}(t)=1, E^{(\mathsf{r})}_{a,f_a^*}(t)=0\right\} \bigcup \left\{\mathsf{UCB}_{a,f_a^*}(t) \leq \mathsf{UCB}_{a,f}(t)\right\}\right) \bigcap \mathbb{1}\left(Y_a(t)=1\right) \\
&\subset \left\{Y_a(t)=1, \mathsf{UCB}_{a,f_a^*}(t) \leq \mathsf{UCB}_{a,f}(t)\right\} \bigcup \left\{E^{(\mathsf{r})}_{a,f}(t)=1, E^{(\mathsf{r})}_{a,f_a^*}(t)=0\right\}
\end{aligned}$$

□

Remark 13. *The results in Lemma 12 holds even if we replace UCB subroutine in Algorithm 2 with any other index based stochastic bandit subroutine, e.g. Thompson sampling.*

B.2 Proof of Lemma 6

We present the proof of each result **(L1)-(L5)** in Lemma 6 individually in the following subsubsections. Before that we define an important notation as follows:

$$H_{a,f}(t) = \mathbb{1}\left(\exists\ a' \in \mathcal{A} : f_{a'}(t) = f, u_f(a') \geq u_f(a)\right) \tag{B.2}$$

B.2.1 Proof of (L1) in Lemma 6

From (5.1) we get

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i} R_a &\leq \bar{\Delta}\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\sum_{f\in\underline{\mathbb{F}}_a}\mathbb{E}[M_{a,f}(T)] + u\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\sum_{f\in F\setminus\{f_a^*\}}\mathbb{E}[C_{a,f}(T)] \\
&\qquad + \bar{u}\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[C_{a,f_a^*}(T)], \\
&\leq \bar{C}\Bigg(\sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\sum_{f\in\underline{\mathbb{F}}_a}\mathbb{E}[M_{a,f}(T)] + \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\sum_{f\in F\setminus\{f_a^*\}}\mathbb{E}[C_{a,f}(T)] \\
&\qquad + \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[\sum_{t=1}^{T} H_{a,f_a^*}(t)]\Bigg)
\end{aligned}$$

where $\bar{\Delta} = \max_{a,f}\Delta_a(f)$ and $\bar{u} = \max_a u_a(f_a^*)$. This completes the proof

B.2.2 Proof of (L2) in Lemma 6

Proof of **(L2)** in Lemma 6 follows immediately from the following more general result.

Lemma 14. *For any agent $a \in \mathcal{A}$ using Algorithm 2 the expected number of matches with any set $\tilde{\mathcal{F}} \subseteq \underline{\mathbb{F}}_a$ can be bounded as*

$$\mathbb{E}[M_{a,\tilde{\mathcal{F}}}(T)] \leq \mathcal{O}\left(|\tilde{\mathcal{F}}|\left(\log(T) + \frac{\log(T)}{\Delta^2}\right) + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)\right]\right)$$

where $\Delta = \min_{a,f}\Delta_a(f)$.

Proof. Note that we call an agent a matches with firm f at time t if $Y_a(t) = 1$ and $f_a(t) = f$. Therefore the total number of matchings between a and f till time T is $M_{a,f}(T) = \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f\right)$. Therefore from Lemma 7 the following holds for every $f \in \tilde{\mathcal{F}}$:

$$
\begin{aligned}
M_{a,\tilde{\mathcal{F}}}(T) &= \sum_{f\in\tilde{\mathcal{F}}} \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f\right) \\
&\leq \sum_{f\in\tilde{\mathcal{F}}} \sum_{t=1}^{T} \left(\mathbb{1}\left(Y_a(t) = 1, f_a(t) = f, \mathsf{UCB}_{a,f}(t) \geq \mathsf{UCB}_{a,f_a^*}(t)\right) + \mathbb{1}\left(E^{(\mathsf{r})}_{a,f}(t) = 1, E^{(\mathsf{r})}_{a,f_a^*} = 0\right)\right) \\
&\leq \sum_{f\in\tilde{\mathcal{F}}} \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f, \mathsf{UCB}_{a,f}(t) \geq \mathsf{UCB}_{a,f_a^*}(t)\right) \\
&\qquad + \sum_{t=1}^{T} \sum_{f\in\tilde{\mathcal{F}}} \mathbb{1}\left(E^{(\mathsf{r})}_{a,f}(t) = 1, E^{(\mathsf{r})}_{a,f_a^*} = 0\right) \\
&\leq \sum_{f\in\tilde{\mathcal{F}}} \underbrace{\sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f, \mathsf{UCB}_{a,f}(t) \geq \mathsf{UCB}_{a,f_a^*}(t)\right)}_{\text{Term A}} + \underbrace{\sum_{t=1}^{T} \mathbb{1}\left(E^{(\mathsf{r})}_{a,f_a^*} = 0\right)}_{\text{Term B}}
\end{aligned}
$$

For any fixed firm $f \in \tilde{\mathcal{F}}$ we now bound Term A. For that purpose, define an event

$$
\mathcal{Z}_{a,f}(t) \coloneqq \{\mathsf{UCB}_{a,f}(t) \geq u_a(f_a^*) - \epsilon\} = \left\{ \hat{\mu}_{a,f}(t-1) + \sqrt{\frac{2\log(B_a(t))}{M_{a,f}(t-1)}} \geq u_a(f_a^*) - \epsilon \right\},
$$

where $B_a(t) \coloneqq 1 + \bar{M}_a(t)\log^2\left(\bar{M}_a(t)\right) \leq 1 + t\log^2(t) =: \bar{B}(t)$,[9].

Using this notation, we have

$$
\begin{aligned}
\text{Term A} = &\underbrace{\sum_{t=1}^{T} \mathbb{1}(Y_a(t) = 1, f_a(t) = f, \mathsf{UCB}_{a,f}(t) \geq \mathsf{UCB}_{a,f_a^*}(t), \mathcal{Z}_{a,f}(t))}_{\text{Term C}} \\
&+ \underbrace{\sum_{t=1}^{T} \mathbb{1}(Y_a(t) = 1, f_a(t) = f, \mathsf{UCB}_{a,f}(t) \geq \mathsf{UCB}_{a,f_a^*}(t), \mathcal{Z}^{\mathsf{c}}_{a,f}(t))}_{\text{Term D}}
\end{aligned}
$$

[9]The inequality holds due to the fact that $\bar{M}_a(t) \leq t$ and monotonicity of the mapping $x \mapsto 1 + x\log^2(x)$.

We shall first bound $\mathbb{E}[\text{Term C}]$ below:

$$\begin{aligned}
\text{Term C} &= \sum_{t=1}^{T} \mathbb{1}(Y_a(t)=1, f_a(t)=f, \mathsf{UCB}_{a,f}(t) \geq \mathsf{UCB}_{a,f_a^*}(t), \mathcal{Z}_{a,f}(t)) \\
&\leq \sum_{t=1}^{T} \mathbb{1}(Y_a(t)=1, f_a(t)=f, \mathcal{Z}_{a,f}(t)) \\
&= \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t)=1, f_a(t)=f, \hat{\mu}_{a,f}(t-1) + \sqrt{\frac{2\log(B_a(t))}{M_{a,f}(t-1)}} \geq u_a(f_a^*) - \epsilon\right) \\
&\leq \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t)=1, f_a(t)=f, \hat{\mu}_{a,f}(t-1) + \sqrt{\frac{2\log(B_a(T))}{M_{a,f}(t-1)}} \geq u_a(f_a^*) - \epsilon\right) \\
&= \sum_{t=1}^{T}\sum_{s=0}^{t-1} \mathbb{1}\left(Y_a(t)=1, f_a(t)=f, \hat{\mu}_{a,f}^{(s)} + \sqrt{\frac{2\log(B_a(T))}{s}} \geq u_a(f_a^*) - \epsilon, M_{a,f}(t-1)=s\right) \\
&\leq \sum_{s=0}^{T-1}\sum_{t=s+1}^{T} \mathbb{1}\left(f_a(t)=f, \hat{\mu}_{a,f}^{(s)} + \sqrt{\frac{2\log(B_a(T))}{s}} \geq u_a(f_a^*) - \epsilon, M_{a,f}(t-1)=s, M_{a,f}(t)=s+1\right) \\
&\leq \sum_{s=0}^{T-1} \mathbb{1}\left(\hat{\mu}_{a,f}^{(s)} + \sqrt{\frac{2\log(B_a(T))}{s}} \geq u_a(f_a^*) - \epsilon\right) \\
&\leq \sum_{s=0}^{T-1} \mathbb{1}\left(\hat{\mu}_{a,f}^{(s)} - u_a(f) + \sqrt{\frac{2\log(\bar{B}(T))}{s}} \geq \underbrace{u_a(f_a^*) - u_a(f)}_{\Delta_a(f)} - \epsilon\right),
\end{aligned}$$

where $\mu_{a,f}^{(s)}$ is defined to be the empirical utility that agent a obtains on s independent successful pulls of arm f. Using Lemma 18 to further bound $\mathbb{E}[\text{Term C}]$ we get

$$\mathbb{E}[\text{Term C}] \leq 1 + \frac{2}{(\Delta_a(f)-\epsilon)^2}\left(\log(\bar{B}(T) + \sqrt{\pi\log(\bar{B}(T))} + 1)\right)$$

Next, we bound $\mathbb{E}[\text{Term D}]$ below:

$$\begin{aligned}
\mathbb{E}[\text{Term D}] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(Y_a(t)=1, f_a(t)=f, \mathsf{UCB}_{a,f}(t) \geq \mathsf{UCB}_{a,f_a^*}(t), \mathsf{UCB}_{a,f}(t) \leq u_a(f_a^*) - \epsilon\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(Y_a(t)=1, \hat{\mu}_{a,f_a^*}(t-1) + \sqrt{\frac{2\log(B_a(t))}{M_{a,f_a^*}(t-1)}} \leq u_a(f_a^*) - \epsilon\right)\right] \\
&\leq \sum_{t=1}^{T}\sum_{s=0}^{T-1} \mathsf{Pr}\left(\hat{\mu}_{a,f_a^*}^{(s)} + \sqrt{\frac{2\log(\bar{B}(t))}{s}} \leq u_a(f_a^*) - \epsilon\right) \\
&\leq \sum_{t=1}^{T}\sum_{s=0}^{T-1} \exp\left(-\frac{s\left(\sqrt{\frac{2\log(\bar{B}(t))}{s}} + \epsilon\right)^2}{2}\right) \\
&\leq \sum_{t=1}^{T} \frac{1}{\bar{B}(t)} \sum_{s=1}^{T} \exp\left(-\frac{s\epsilon^2}{2}\right) \\
&\leq \frac{\epsilon^2}{2} \sum_{t=0}^{T-1} \frac{1}{\bar{B}(t)}
\end{aligned}$$

which can further be bounded as $\mathbb{E}[\text{Term D}] \leq \frac{5}{\epsilon^2}$ in [LS20, Exercise 8.1]. For simplicity we choose $\epsilon = \Delta_a(f)/2$ which ensures that $\mathbb{E}[\text{Term A}] \leq \mathcal{O}\left(\frac{\log(T)}{(\Delta_a(f))^2}\right)$

Now let's turn our attention to Term B which characterizes the number of times agent a has pruned the stable match. Using Lemma 20 we have

$$\mathbb{E}[\text{Term B}] \leq \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f_a^*}(t)\right)\right] + \mathcal{O}(\log(T))\right)$$

Thus the Term A is bounded by number of there can be potential collisions at the stable firm. This concludes the proof of this lemma. □

B.2.3 Proof of (L3) in Lemma 6

In this part, we prove a result which is more general than **(L3)** in Lemma 6.

Lemma 15. *Expected number of collisions faced by agent a on the set of firms $\mathcal{F}^\dagger \subseteq \mathcal{F}\backslash\{f_a^*\}$*

$$\sum_{f\in\mathcal{F}^\dagger} \mathbb{E}[C_{a,f}(T)] \leq \mathcal{O}\left(|\mathcal{F}^\dagger|\log(T) + \mathbb{E}[M_{a,\underline{\mathcal{F}}_a^\dagger}(T)] + \mathbb{E}[M_{a,\bar{\mathcal{F}}_a^\dagger}(T)] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f_a^*}(t)\right)\right]\right), \tag{B.3}$$

where $\underline{\mathcal{F}}_a^\dagger = \underline{\mathbb{F}}_a \cap \mathcal{F}^\dagger$ and $\bar{\mathcal{F}}_a^\dagger = \bar{\mathbb{F}}_a \cap \mathcal{F}^\dagger$. Additionally

$$\mathbb{E}\left[C_{a,f_a^*}(T)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f_a^*}(t)\right)\right] \tag{B.4}$$

Proof. To compute the number of collisions, we compute the following for $a \in \mathcal{A}$ and $f \in \mathcal{F}\backslash\{f_a^*\}$

$$\begin{aligned}
&\sum_{f\in\mathcal{F}^\dagger} C_{a,f}(T) = \sum_{f\in\mathcal{F}^\dagger}\sum_{t=1}^{T} \mathbb{1}\left(f_a(t) = f, H_{a,f}(t)\right) \\
&= \sum_{f\in\mathcal{F}^\dagger}\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}(t)\right) \\
&\quad + \sum_{f\in\mathcal{F}^\dagger}\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f'}^{(\mathsf{r})}(t) = 0 \;\forall\; f' \in \mathcal{F}, f_a(t) = f, H_{a,f}(t)\right) \\
&\leq \sum_{f\in\mathcal{F}^\dagger}\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}(t)\right) + \sum_{f\in\mathcal{F}^\dagger}\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f_a^*}^{(\mathsf{r})}(t) = 0, f_a(t) = f\right), \\
&\leq \sum_{f\in\mathcal{F}^\dagger}\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}(t)\right) + \sum_{t=1}^{T} \mathbb{1}\left(E_{a,f_a^*}^{(\mathsf{r})}(t) = 0\right),
\end{aligned}$$

where the first inequality holds because $\{E_{a,f'}^{(\mathsf{r})}(t) = 0 \;\forall\; f' \in \mathcal{F}\}$ implies that $\{E_{a,f_a^*}^{(\mathsf{r})}(t) = 0\}$.

Using (D.1) we have: for all $a \in \mathcal{A}$, $f \in \mathcal{F}$ and $\varpi \in (0, 32\eta) \subset (0,1)$

$$\begin{aligned}
&\sum_{f\in\mathcal{F}^\dagger} \mathbb{E}[C_{a,f}(T)] \\
&\leq \sum_{f\in\mathcal{F}^\dagger} \left((1+\varpi)\mathbb{E}[M_{a,f}(T)] + \mathcal{O}(\log(T)) + \varpi\mathbb{E}[C_{a,f}(T)] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(E^{(\mathsf{r})}_{a,f^*_a} = 0\right)\right]\right) \\
&\leq \mathcal{O}\left(|\mathcal{F}^\dagger|\log(T) + \sum_{f\in\mathcal{F}^\dagger}\mathbb{E}[M_{a,f}(T)]\right) + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] + \varpi\sum_{f\in\mathcal{F}^\dagger}\mathbb{E}[C_{a,f}(T)]
\end{aligned}$$

where the last inequality is due to Lemma 20. In summary,

$$\begin{aligned}
\sum_{f\in\mathcal{F}^\dagger}\mathbb{E}[C_{a,f}(T)] &\leq \mathcal{O}\left(|\mathcal{F}|\mathcal{O}(\log(T)) + \sum_{f\in\mathcal{F}^\dagger}(\mathbb{E}[M_{a,f}(T)])\right) + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] \\
&\leq \mathcal{O}\left(|\mathcal{F}^\dagger|\log(T) + \mathbb{E}[M_{a,\underline{\mathcal{F}}^\dagger_a}(T)] + \mathbb{E}[M_{a,\bar{\mathcal{F}}^\dagger_a}(T)] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)\right]\right)
\end{aligned}$$

This completes the proof of (B.3). We now prove (B.4). We note that

$$\mathbb{E}\left[C_{a,f^*_a}(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(f_a(t) = f, H_{a,f^*_a}(t)\right)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)\right].$$

This completes the proof. □

B.2.4 Proof of (L4) in Lemma 6

We restate **(L4)** from Lemma 6 below:

Lemma 16. *For any $i \in [K]$ we have*

$$\sum_{j=1}^{i}\sum_{a\in\mathcal{A}_j}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] = \mathcal{O}\left(C_i|\mathcal{F}|\left(\sum_{j=1}^{i}|\mathcal{A}_j|\right)\log(T)\left(1+\frac{1}{\Delta^2}\right)\right),$$

where C_i is a constant dependent on market $\mathcal{M}_i$ such that $C_1 < C_2 < ... < C_K$.

Proof. For any $k \in [K]$ define $S_k = \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)]$ and $Z(T,\Delta) = |\mathcal{F}|\log(T)\left(1+\frac{1}{\Delta^2}\right)$. Define $f(\theta;\ell) = \sum_{j=1}^{\ell}\theta^j$, $f(\theta;0) = 1$ and $g(\theta;\ell) = \sum_{j=0}^{\ell-1}\theta^j$. Moreover, let $\mathcal{H}_i = \sum_{a\in\mathcal{A}_i}\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)]$. Consequently $S_k = \sum_{i=1}^{k}\mathcal{H}_i$. We claim that

$$\begin{aligned}
S_K \leq S_{K-\ell} + f(\theta;\ell)\mathcal{H}_{K-\ell} + \sum_{p=1}^{\ell} g(\theta;p) \sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-1}\mathcal{A}_j}\mathbb{E}\left[M_{a',f^*_a}(T)\right] \\
+ Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}| \qquad \text{(B.5)}
\end{aligned}$$

We prove this via induction. We first show that this holds for $\ell = 1$. Indeed note that

$$
\begin{aligned}
S_K &= S_{K-1} + \mathcal{H}_K = S_{K-1} + \sum_{a\in\mathcal{A}_K} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\big(H_{a,f_a^*}(t)\big)\right] \\
&\underset{(a)}{\leq} S_{K-1} + \sum_{a\in\mathcal{A}_K} \sum_{a'\in\cup_{j=1}^{K-2}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] + \sum_{a\in\mathcal{A}_K} \sum_{a'\in\mathcal{A}_{K-1}} \mathbb{E}\left[M_{a',f_a^*}(T)\right] \\
&\underset{(b)}{=} S_{K-1} + \sum_{a\in\mathcal{A}_K} \sum_{a'\in\cup_{j=1}^{K-2}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] + \sum_{a'\in\mathcal{A}_{K-1}} \sum_{f\in\mathcal{F}_K} \mathbb{E}\left[M_{a',f}(T)\right] \\
&\underset{(c)}{\leq} S_{K-1} + \sum_{a\in\mathcal{A}_K} \sum_{a'\in\cup_{j=1}^{K-2}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] + \sum_{a'\in\mathcal{A}_{K-1}} \mathbb{E}\left[M_{a',\underline{\mathbb{F}}_{a'}}(T)\right] \\
&\underset{(d)}{\leq} S_{K-1} + \theta \sum_{a'\in\mathcal{A}_{K-1}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\Big(H_{a',f_{a'}^*}(t)\Big)\right] + \sum_{a\in\mathcal{A}_K} \sum_{a'\in\cup_{j=1}^{K-2}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] + \theta|\mathcal{A}_{K-1}|Z(T,\Delta) \\
&= S_{K-1} + \theta\mathcal{H}_{K-1} + \sum_{a\in\mathcal{A}_K} \sum_{a'\in\cup_{j=1}^{K-2}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] + \theta|\mathcal{A}_{K-1}|Z(T,\Delta)
\end{aligned}
$$

where the (a) holds due to $\alpha-$reducible structure which says that any agent in $\mathcal{A}_K$ will only get collided at stable arm if some agent from $\cup_{j=1}^{k-1}\mathcal{A}_j$ has also requested the stable firm. Next, (b) holds due to the fact that for any agent $a \in \mathcal{A}_k$, the corresponding stable match $f_a^* \in \mathcal{F}_k$(see Remark 3). Next, (c) follows because for agents in $\mathcal{A}_{K-1}$, the set of suboptimal firms is super set of $\mathcal{F}_K$. This is again a property of $\alpha-$reducible structure. Finally (d) follows from **(L2)** in Lemma 6 where θ is the corresponding constant from big-oh notation.

Suppose the bound in (B.5) holds for $\ell = L$ for some integer $\ell \in \{2, 3, ..., K\}$. Then we show it also holds for $\ell + 1$. That is,

$$
\begin{aligned}
S_K &\underset{(a)}{\leq} S_{K-\ell} + f(\theta;\ell)\mathcal{H}_{K-\ell} + \sum_{p=1}^{\ell} g(\theta;p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{a'\in\cup_{j=1}^{K-\ell-1}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] \\
&\qquad + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}| \\
&\underset{(b)}{=} S_{K-\ell-1} + g(\theta;\ell+1)\mathcal{H}_{K-\ell} + \sum_{p=1}^{\ell} g(\theta;p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{a'\in\cup_{j=1}^{K-\ell-1}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] \\
&\qquad + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}| \\
&\underset{(c)}{\leq} S_{K-\ell-1} + g(\theta;\ell+1)\left(\mathcal{H}_{K-\ell} + \sum_{p=1}^{\ell} \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{a'\in\mathcal{A}_{K-\ell-1}} \mathbb{E}\left[M_{a',f_a^*}(T)\right]\right) \\
&\qquad + \sum_{p=1}^{\ell} g(\theta;p) \sum_{a\in\mathcal{A}_{K-p+1}} \sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j} \mathbb{E}\left[M_{a',f_a^*}(T)\right] + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|
\end{aligned}
$$

$$
\begin{aligned}
&\underset{(d)}{\leq} S_{K-\ell-1} + g(\theta;\ell+1)\left(\sum_{p=1}^{K-\ell-1}\sum_{a'\in\mathcal{A}_p}\sum_{a\in\mathcal{A}_{K-\ell}}\mathbb{E}[M_{a',f_a^*}] + \sum_{p=1}^{\ell}\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\mathcal{A}_{K-\ell-1}}\mathbb{E}\left[M_{a',f_a^*}(T)\right]\right)\\
&\qquad + \sum_{p=1}^{\ell} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right] + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|\\
&\underset{(e)}{=} S_{K-\ell-1} + g(\theta;\ell+1)\left(\sum_{p=1}^{K-\ell-2}\sum_{a'\in\mathcal{A}_p}\sum_{a\in\mathcal{A}_{K-\ell}}\mathbb{E}[M_{a',f_a^*}] + \sum_{p=1}^{\ell+1}\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\mathcal{A}_{K-\ell-1}}\mathbb{E}\left[M_{a',f_a^*}(T)\right]\right)\\
&\qquad + \sum_{p=1}^{\ell} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right] + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|\\
&\underset{(f)}{\leq} S_{K-\ell-1} + g(\theta;\ell+1)\left(\sum_{p=1}^{K-\ell-2}\sum_{a'\in\mathcal{A}_p}\sum_{a\in\mathcal{A}_{K-\ell}}\mathbb{E}[M_{a',f_a^*}] + \sum_{a'\in\mathcal{A}_{K-\ell-1}}\mathbb{E}\left[M_{a',\underline{\mathbb{F}}_{a'}}(T)\right]\right)\\
&\qquad + \sum_{p=1}^{\ell} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right] + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|\\
&\underset{(g)}{=} S_{K-\ell-1} + g(\theta;\ell+1)\left(\sum_{a'\in\mathcal{A}_{K-\ell-1}}\mathbb{E}\left[M_{a',\underline{\mathbb{F}}_{a'}}(T)\right]\right)\\
&\qquad + \sum_{p=1}^{\ell+1} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right] + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|\\
&\underset{(h)}{\leq} S_{K-\ell-1} + g(\theta;\ell+1)\left(\theta|\mathcal{F}|Z(T,\Delta)|\mathcal{A}_{K-\ell-1}| + \theta\mathcal{H}_{K-\ell-1}\right)\\
&\qquad + \sum_{p=1}^{\ell+1} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right] + Z(T,\Delta)\sum_{r=1}^{\ell} f(\theta;r)|\mathcal{A}_{K-r}|\\
&\underset{(i)}{=} S_{K-\ell-1} + f(\theta;\ell+1)\mathcal{H}_{K-\ell-1} + \sum_{p=1}^{\ell+1} g(\theta;p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{a'\in\cup_{j=1}^{K-\ell-2}\mathcal{A}_j}\mathbb{E}\left[M_{a',f_a^*}(T)\right]\\
&\qquad + Z(T,\Delta)\sum_{r=1}^{\ell+1} f(\theta;r)|\mathcal{A}_{K-r}|
\end{aligned}
$$

where (a) holds by induction hypothesis, (b) holds by definition of S_k and $f(\theta;\ell), g(\theta;\ell)$, (c) holds by moving some terms around and noting that $g(\theta;\cdot)$ is increasing. Next, (d) holds by $\alpha-$reducbility and definition of $\mathcal{H}_k$ (same analysis as in base case of induction). Next, (e) holds by splitting the terms. Next, (f) holds by $\alpha-$reducilibility definition. Next (g) holds by combining similar terms. Next (h) holds by **(L2)** in Lemma 6. Next, (i) holds due to combining similar terms.

Thus we conclude that induction claim (B.5) holds true. We know that $S_1 = 0$ therefore from (B.5) we obtain

$$S_k \leq Z(T,\Delta)\sum_{r=1}^{K-1} f(\theta;r)|\mathcal{A}_{K-r}| \leq \left(\sum_{j=1}^{K-1}|\mathcal{A}_j|\right) K\theta^{K-1} Z(T,\Delta). \tag{B.6}$$

The term $C_k = k\theta^{k-1}$ in the statement. This completes the proof.

□

B.2.5 Proof of (L5) in Lemma 6

So only thing to bound is matching with superoptimal firms.

Lemma 17. *For any $k \in [K]$ we have*

$$\sum_{j=1}^{k}\sum_{a\in\mathcal{A}_j}\sum_{f\in\bar{\mathbb{F}}_a}\mathbb{E}[M_{a,f}(T)] \leq \mathcal{O}\left(C_i\left(\sum_{j=1}^{k-1}|\mathcal{A}_j|\right)|\mathcal{F}|\log(T)\left(1+\frac{1}{\Delta^2}\right)\right),$$

where C_i is a constant dependent on market $\mathcal{M}_i$ such that $C_1 < C_2 < ... < C_K$.

Proof. For any $k \in [K]$, define $\tilde{S}_k = \sum_{i=1}^{k}\sum_{a\in\mathcal{A}_i}\mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)]$ and $Z(T,\Delta) = |F|\log(T)\left(1+1/\Delta^2\right)$. Define $f(\theta;\ell) = \sum_{j=1}^{\ell}\theta^j$, $f(\theta;0) = 1$ and $g(\theta;\ell) = \sum_{j=0}^{\ell-1}\theta^j$. Let $\mathcal{H}_i = \sum_{a\in\mathcal{A}_i}\mathbb{E}[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)]$ and $\mathbb{M}_i = \sum_{a\in\mathcal{A}_i}\mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)]$ then $\tilde{S}_k = \sum_{i=1}^{k}\mathbb{M}_i$. We claim that

$$\tilde{S}_k \leq \mathcal{O}\left(\tilde{\theta}^{k-1}\left(\sum_{j=1}^{k-1}|\mathcal{A}_j|\right)|\mathcal{F}|Z(T,\Delta)\right) \tag{B.7}$$

where $\tilde{\theta}$ is a constant greater than 1. Note that the bound holds for $k = 1$ as there is not super-optimal firms for those agents. Let (B.7) holds till some integer $K-1$ then we show that it holds for K as well. Indeed,

We claim that

$$\tilde{S}_K \leq \tilde{S}_{K-\ell} + f(\tilde{\theta};\ell)\mathbb{M}_{K-\ell} + \sum_{p=1}^{\ell} g(\tilde{\theta};p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{f\in\cup_{j\leq K-\ell-1}\mathcal{F}_j}\mathbb{E}\left[M_{a,f}\right] + \sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p}$$
$$+ Z(T,\Delta)\sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}| \tag{B.8}$$

We prove (B.7) by induction. First, consider the case $\ell = 1$

$$\begin{aligned}
\tilde{S}_K &= \sum_{i=1}^{K} \sum_{a \in \mathcal{A}_i} \mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)] \\
&\underset{(a)}{=} \tilde{S}_{K-1} + \sum_{a \in \mathcal{A}_K} \mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)] \\
&\underset{(b)}{\leq} \tilde{S}_{K-1} + \sum_{a \in \mathcal{A}_K} \sum_{f \in \cup_{j \leq K-2} \mathcal{F}_j} \mathbb{E}[M_{a,f}(T)] + \sum_{a \in \mathcal{A}_K} \sum_{f \in \mathcal{F}_{K-1}} \mathbb{E}[M_{a,f}(T)] \\
&\underset{(c)}{=} \tilde{S}_{K-1} + \sum_{a \in \mathcal{A}_K} \sum_{f \in \cup_{j \leq K-2} \mathcal{F}_j} \mathbb{E}[M_{a,f}(T)] + \sum_{a' \in \mathcal{A}_{K-1}} \sum_{a \in \mathcal{A}_K} \mathbb{E}[M_{a,f^*_{a'}}(T)] \\
&\underset{(d)}{\leq} \tilde{S}_{K-1} + \tilde{\theta} \sum_{a' \in \mathcal{A}_{K-1}} \mathbb{E}[M_{a',\bar{\mathbb{F}}_{a'}}(T)] + \sum_{a \in \mathcal{A}_K} \sum_{a' \in \cup_{j \leq K-2} \mathcal{A}_j} \mathbb{E}[M_{a,f^*_{a'}}(T)] \\
&\qquad + \sum_{a' \in \mathcal{A}_{K-1}} \tilde{\theta}\Big(H_{a',f^*_{a'}} + Z(T,\Delta)\Big) \\
&\underset{(e)}{=} \tilde{S}_{K-1} + \tilde{\theta}\mathbb{M}_{K-1} + \sum_{a \in \mathcal{A}_K} \sum_{f \in \cup_{j \leq K-2} \mathcal{F}_j} \mathbb{E}[M_{a,f}(T)] + \tilde{\theta}\mathcal{H}_{K-1} + Z(T,\Delta)\tilde{\theta}|\mathcal{A}_{K-1}|
\end{aligned}$$

where (a) holds by definition, (b) holds by using $\alpha-$reducilbe structure which ensures that set of superoptimal firms of any agent will lie in markets before it. Next, (c) holds by property of alpha-reducible markets which ensures that for firm $f \in \mathcal{F}_{K-1}$ there exists agent $a' \in \mathcal{A}_{K-1}$ such that $f = f^*_{a'}$. Next, (d) holds by Lemma 21. Next (e) holds by rearrangement of terms. Next, we show that if (B.7) holds for some ℓ then it holds for $\ell + 1$ as well. That is,

$$\begin{aligned}
\tilde{S}_K &\underset{(a)}{\leq} \tilde{S}_{K-\ell} + f(\tilde{\theta};\ell)\mathbb{M}_{K-\ell} + \sum_{p=1}^{\ell} g(\tilde{\theta};p) \sum_{a \in \mathcal{A}_{K-p+1}} \sum_{f \in \cup_{j \leq K-\ell-1} \mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p} \\
&\qquad + Z(T,\Delta) \sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}| \\
&\underset{(b)}{=} \tilde{S}_{K-\ell-1} + g(\tilde{\theta};\ell+1)\mathbb{M}_{K-\ell} + \sum_{p=1}^{\ell} g(\tilde{\theta};p) \sum_{a \in \mathcal{A}_{K-p+1}} \sum_{f \in \cup_{j \leq K-\ell-1} \mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p} \\
&\qquad + Z(T,\Delta) \sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}| \\
&\underset{(c)}{=} \tilde{S}_{K-\ell-1} + g(\tilde{\theta};\ell+1) \left(\sum_{a \in \mathcal{A}_{K-\ell}} \sum_{f \in \cup_{j \leq K-\ell-2} \mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{a \in \mathcal{A}_{K-\ell}} \sum_{f \in \mathcal{F}_{K-\ell-1}} \mathbb{E}\left[M_{a,f}(T)\right] \right) \\
&\qquad + \sum_{p=1}^{\ell} g(\tilde{\theta};p) \sum_{a \in \mathcal{A}_{K-p+1}} \sum_{f \in \cup_{j \leq K-\ell-1} \mathcal{F}_j} \mathbb{E}\left[M_{a,f}\right] + \sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p} \\
&\qquad + Z(T,\Delta) \sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}|
\end{aligned}$$

$$
\begin{aligned}
&\underset{(d)}{\leq} \tilde{S}_{K-\ell-1} + g(\tilde{\theta};\ell+1)\left(\sum_{p=1}^{\ell+1}\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{f\in\mathcal{F}_{K-\ell-1}}\mathbb{E}\left[M_{a,f}(T)\right]\right)\\
&\quad+\sum_{p=1}^{\ell+1} g(\tilde{\theta};p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{f\in\cup_{j\leq K-\ell-2}\mathcal{F}_j}\mathbb{E}\left[M_{a,f}\right]+\sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p}\\
&\quad+Z(T,\Delta)\sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}|\\
&\underset{(e)}{=} \tilde{S}_{K-\ell-1} + g(\tilde{\theta};\ell+1)\left(\sum_{a'\in\mathcal{A}_{K-\ell-1}}\sum_{p=1}^{\ell+1}\sum_{a\in\mathcal{A}_{K-p+1}}\mathbb{E}\left[M_{a,f^*_{a'}}(T)\right]\right)\\
&\quad+\sum_{p=1}^{\ell+1} g(\tilde{\theta};p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{f\in\cup_{j\leq K-\ell-2}\mathcal{F}_j}\mathbb{E}\left[M_{a,f}\right]+\sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p}\\
&\quad+Z(T,\Delta)\sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}|\\
&\underset{(f)}{\leq} \tilde{S}_{K-\ell-1} + g(\tilde{\theta};\ell+1)\left(\tilde{\theta}\mathcal{H}_{K-\ell-1}+\tilde{\theta}\mathbb{M}_{K-\ell-1}+\tilde{\theta}Z(T,\Delta)|\mathcal{A}_{K-\ell-1}|\right)\\
&\quad+\sum_{p=1}^{\ell+1} g(\tilde{\theta};p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{f\in\cup_{j\leq K-\ell-2}\mathcal{F}_j}\mathbb{E}\left[M_{a,f}\right]+\sum_{p=1}^{\ell} f(\tilde{\theta},p)\mathcal{H}_{K-p}\\
&\quad+Z(T,\Delta)\sum_{p=1}^{\ell} f(\tilde{\theta},p)|\mathcal{A}_{K-p}|\\
&\underset{(g)}{=} \tilde{S}_{K-\ell-1} + f(\tilde{\theta};\ell+1)\mathbb{M}_{K-\ell-1}+\\
&\quad+\sum_{p=1}^{\ell+1} g(\tilde{\theta};p)\sum_{a\in\mathcal{A}_{K-p+1}}\sum_{f\in\cup_{j\leq K-\ell-2}\mathcal{F}_j}\mathbb{E}\left[M_{a,f}\right]+\sum_{p=1}^{\ell+1} f(\tilde{\theta},p)\mathcal{H}_{K-p}\\
&\quad+Z(T,\Delta)\sum_{p=1}^{\ell+1} f(\tilde{\theta},p)|\mathcal{A}_{K-p}|
\end{aligned}
$$

where (a) is by induction hypothesis, (b) is by decomposing $\tilde{S}_{K-\ell}$, (c) is by using definition of $\mathbb{M}_{K-\ell}$, (d) is by rearrangement of terms and using the fact that $g(\tilde{\theta},\cdot)$ is increasing, (e) is by rearrangement of terms and using the fact that for any $f\in\mathcal{F}_k$ for some k there exists $a'\in\mathcal{A}_k$ such that $f=f^*_{a'}$. Next, (f) is by Lemma 21. Next, (g) is by combining similar terms. This concludes the induction proof.

We know that $\tilde{S}_1=\mathbb{M}_1=0$ because of $\alpha-$reducible structure which ensures that these firms do not have superoptimal firms. Thus in (B.7) if take $\ell=K-1$ then we get

$$
\begin{aligned}
\tilde{S}_K &\leq \sum_{p=1}^{K-1} f(\tilde{\theta},p)\mathcal{H}_{K-p}+Z(T,\Delta)\sum_{p=1}^{K-1} f(\tilde{\theta},p)|\mathcal{A}_{K-p}|\\
&\leq \sum_{p=1}^{K-1}\sum_{j=1}^{p}\tilde{\theta}^j\mathcal{H}_{K-p}+Z(T,\Delta)\sum_{p=1}^{K-1} f(\tilde{\theta},p)|\mathcal{A}_{K-p}|
\end{aligned}
$$

$$\begin{aligned}
\tilde{S}_K &\leq \sum_{j=1}^{K-1} \tilde{\theta}^j \sum_{p=j}^{K-1} \mathcal{H}_{K-p} + Z(T,\Delta) \sum_{p=1}^{K-1} f(\tilde{\theta},p)|\mathcal{A}_{K-p}| \\
&\underset{(a)}{=} \sum_{j=1}^{K-1} \tilde{\theta}^j S_{K-j} + Z(T,\Delta) \left(\sum_{j=1}^{K-1} |\mathcal{A}_j| \right) K\tilde{\theta}^{K-1} \\
&\underset{(b)}{\leq} Z(T,\Delta) \left(\sum_{j=1}^{K-1} |\mathcal{A}_j| \right) \sum_{j=1}^{K-1} \tilde{\theta}^j (K-j)\theta^{K-j-1} + Z(T,\Delta) \left(\sum_{j=1}^{K-1} |\mathcal{A}_j| \right) K\tilde{\theta}^{K-1}
\end{aligned}$$

where S_{K-j} in (a) is from proof of **(L4)** in Lemma 6 and (b) is by (B.6). Define $\tilde{C}_k = k\tilde{\theta}^{k-1} + \sum_{j=1}^{k-1} \tilde{\theta}^j (k-j)\theta^{k-j-1}$. Thus we see that

$$\tilde{S}_K \leq |\mathcal{F}| \log(T) \left(1 + \frac{1}{\Delta^2}\right) \left(\sum_{j=1}^{K-1} |\mathcal{A}_j| \right) \tilde{C}_K$$

□

C Proof of Theorem 5

We now look at the joint regret for any $k \in [K]$. Define $Z(T,\Delta) = |F| \log(T) \left(1 + \frac{1}{\Delta^2}\right)$

$$\begin{aligned}
\sum_{i=1}^{k} \sum_{a \in \mathcal{A}_i} R_a &\underset{(a)}{=} \mathcal{O}\Bigg(\sum_{i=1}^{k} \sum_{a \in \mathcal{A}_i} \mathbb{E}[M_{a,\underline{\mathbb{F}}_a}(T)] + \sum_{i=1}^{k} \sum_{a \in \mathcal{A}_i} \sum_{f \in F \setminus \{f_a^*\}} \mathbb{E}[C_{a,f}(T)] \\
&\quad + \sum_{i=1}^{k} \sum_{a \in \mathcal{A}_i} \mathbb{E}[\sum_{t=1}^{T} H_{a,f_a^*}(t)] \Bigg) \\
&\underset{(b)}{=} \mathcal{O}\left(\sum_{i=1}^{k} \sum_{a \in \mathcal{A}_i} \mathbb{E}[M_{a,\underline{\mathbb{F}}_a}(T)] + \sum_{i=1}^{k} \sum_{a \in \mathcal{A}_i} \mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)] + \sum_{i=1}^{k} \sum_{a \in \mathcal{A}_i} \mathbb{E}[\sum_{t=1}^{T} H_{a,f_a^*}(t)] \right) \\
&\quad + \mathcal{O}\left(|\mathcal{F}| \sum_{i=1}^{k} |\mathcal{A}_i| \log(T) \right) \\
&\underset{(c)}{=} \mathcal{O}\left(\sum_{i=1}^{k} \sum_{a \in \mathcal{A}_i} \mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)] + \sum_{i=1}^{k} \sum_{a \in \mathcal{A}_i} \mathbb{E}[\sum_{t=1}^{T} H_{a,f_a^*}(t)] \right) + \mathcal{O}(\sum_{i=1}^{k} \sum_{a \in \mathcal{A}_i} |\underline{\mathbb{F}}_a| Z(T,\Delta)) \\
&\quad + \mathcal{O}\left(|F| \sum_{i=1}^{k} |\mathcal{A}_i| \log(T) \right) \\
&\underset{(d)}{=} \mathcal{O}(\tilde{C}_k \left(\sum_{p=1}^{k} |\mathcal{A}_p| \right) Z(T,\Delta)) + \mathcal{O}(\left(\sum_{p=1}^{k} |\mathcal{A}_p| \right) C_k Z(T,\Delta)) + \mathcal{O}(\sum_{p=1}^{k} \sum_{a \in \mathcal{A}_p} |\underline{\mathbb{F}}_a| Z(T,\Delta)) \\
&\quad + \mathcal{O}\left(|F| \sum_{p=1}^{k} |\mathcal{A}_p| \log(T) \right) \\
&\underset{(e)}{=} \mathcal{O}\left((C_k + \tilde{C}_k)|\mathcal{F}| \left(\sum_{p=1}^{k} |\mathcal{A}_p| \right) \right) \log(T) \left(1 + \frac{1}{\Delta^2}\right)
\end{aligned}$$

where (a) holds due to **(L1)** in Lemma 6, (b) holds due to **(L3)** in Lemma 6, (c) is due to **(L2)** in Lemma 6. Next, (d) is due to **(L4)-(L5)** in Lemma 6. Finally, (e) follows by combining terms.

D Technical lemmas

In this section we present some technical lemmas which are helpful in the proofs in next section.

Lemma 18. *(Lemma 8.2,[LS20]) Let $X_1, X_2, \ldots, X_T$ be a sequence of independent 1-subgaussian random variable, and $\hat{\mu}^{(t)} \coloneqq \frac{1}{t}\sum_{s=1}^{t} X_s, \epsilon > 0, a > 0$ and*

$$\kappa \coloneqq \sum_{t=1}^{n} \mathbb{1}\left(\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \epsilon\right), \quad \kappa' \coloneqq u + \sum_{t=\lceil u \rceil}^{T} \mathbb{1}\left(\hat{\mu}_t + \sqrt{\frac{2a}{t}} \geq \epsilon\right)$$

where $u = \frac{2a}{\epsilon^2}$. Then

$$\mathbb{E}[\kappa] \leq \mathbb{E}[\kappa'] \leq 1 + \frac{2}{\epsilon^2}(a + \sqrt{\pi a} + 1)$$

Lemma 19. *Suppose we use the AB subroutine Algorithm 3 with $\eta \leq 1/50$ then the following two inequalities hold:*

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f}^{(r)}(t) = 1, E_{a,f}^{(c)}(t) = 1, H_{a,f}(t)\right)\right] \\ &\qquad \leq (1+\varpi)\mathbb{E}[M_{a,f}(T)] + \mathcal{O}(\log(T)) + \varpi\mathbb{E}[C_{a,f}(T)], \end{aligned} \tag{D.1}$$

where $0 < \varpi \leq 32\eta < 1$ and

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f}^{(r)}(t) = 0, E_{a,f}^{(c)}(t) = 1, H_{a,f}^{\mathsf{c}}(t)\right)\right] \\ &\qquad \leq \mathcal{O}\left(\log(T) + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f}(t)\right)\right] + \mathbb{E}[C_{a,f}^{\star}(T)]\right). \end{aligned} \tag{D.2}$$

Proof. To simplify the presentation of proof, let's define

$$L_{a,f}^{(\mathsf{adv})}(T) \coloneqq \sum_{t=1}^{T}\left(\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}(t)\right) - \mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}^{\mathsf{c}}(t)\right)\right)$$

The regret bound for adversarial bandit algorithm from Lemma 10 under $\eta \leq 1/50$ implies

$$\begin{aligned} \mathbb{E}\left[L_{a,f}^{(\mathsf{adv})}(T)\right] &\leq \mathcal{O}(\log(T)) + \varpi\mathbb{E}\left[\min\left\{M_{a,f}^{\star}(T), C_{a,f}^{\star}(T), M_{a,f}(T) + C_{a,f}(T)\right\}\right] \\ \mathbb{E}\left[L_{a,f}^{(\mathsf{adv})}(T) - \ell_{a,f}(T)\right] &\leq \mathcal{O}(\log(T)) + \varpi\mathbb{E}\left[\min\left\{M_{a,f}^{\star}(T), C_{a,f}^{\star}(T), M_{a,f}(T) + C_{a,f}(T)\right\}\right] \end{aligned} \tag{D.3}$$

where $\varpi \leq 32\eta$ and

$$\ell_{a,f}(T) = \sum_{t=1}^{T}\left(\mathbb{1}\left(E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}(t)\right) - \mathbb{1}\left(E_{a,f}^{(\mathsf{c})}(t) = 1, H_{a,f}^{\mathsf{c}}(t)\right)\right)$$

which denotes the total loss received by the adversarial bandit subroutine associated with (a, f) in time T *if* it never take pruning action. Therefore, in (D.3) LHS in first inequality is the regret associated with always pruning. While LHS in second inequality is the regret associated with never pruning.

In the following proof we shall analyze each of the equations in (D.3) separately.

1. The first inequality in (D.3) implies

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=1, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}(t)\right)-\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=1, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}^{\mathsf{c}}(t)\right)\right)\right]$$
$$\leq \mathcal{O}(\log(T))+\varpi\left(\mathbb{E}[M_{a,f}(T)+C_{a,f}(T)]\right).$$

This in turn leads to

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\left(\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=1, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}(t)\right)\right)\right]\\&\leq \mathbb{E}\left[\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=1, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}^{\mathsf{c}}(t)\right)\right]+\mathcal{O}(\log(T))+\frac{1}{2}\left(\mathbb{E}[M_{a,f}(T)+C_{a,f}(T)]\right)\\&\leq (1+\varpi)\,\mathbb{E}[M_{a,f}(T)]+\mathcal{O}(\log(T))+\varpi\mathbb{E}[C_{a,f}(T)]\end{aligned}$$

2. Using the definition of $\ell_{a,f}(T)$ in the second inequality in (D.3) we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(-\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=0, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}(t)\right)+\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=0, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}^{\mathsf{c}}(t)\right)\right)\right]$$
$$\leq \mathcal{O}(\log(T)+\mathbb{E}[\min\{M_{a,f}^{\star}(T), C_{a,f}^{\star}(T)\}])$$

which implies

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=0, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}^{\mathsf{c}}(t)\right)\right]\\&\leq \mathcal{O}\Bigg(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t)=0, E_{a,f}^{(\mathsf{c})}(t)=1, H_{a,f}(t)\right)\right]+\mathcal{O}(\log(T))\\&\qquad+\mathbb{E}[\min\{M_{a,f}^{\star}(T), C_{a,f}^{\star}(T)\}]\Bigg)\\&\leq \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f}(t)\right)\right]+\log(T)+\mathbb{E}[\min\{M_{a,f}^{\star}(T), C_{a,f}^{\star}(T)\}]\right)\end{aligned}$$

This concludes the proof. □

Lemma 20 (Pruning stable match)**.** *For any $a \in \mathcal{A}$,*

$$\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(E_{a,f_a^*}^{(r)}(t)=0, E_{a,f_a^*}^{(c)}(t)=1\right)\right]}_{\mathbb{E}[\mathit{Term\ I}]}\leq \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f_a^*}(t)\right)\right]+\log(T)\right)$$

Proof. We note that

$$\begin{aligned}\mathbb{E}[\text{Term I}] &\leq \mathbb{E}\Bigg[\sum_{t=1}^{T}\mathbb{1}\left(E^{(\mathsf{r})}_{a,f^*_a}(t)=0, E^{(\mathsf{c})}_{a,f^*_a}(t)=1, H_{a,f^*_a}(t)\right)\\ &\qquad + \sum_{t=1}^{T}\mathbb{1}\left(E^{(\mathsf{r})}_{a,f^*_a}(t)=0, E^{(\mathsf{c})}_{a,f^*_a}(t)=1, H^{\mathsf{c}}_{a,f^*_a}(t)\right)\Bigg]\\ &\leq \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] + \mathcal{O}(\log(T)) + \mathbb{E}[C^{\star}_{a,f^*_a}(T)]\right)\\ &\leq \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] + \mathcal{O}(\log(T))\right)\end{aligned}$$

where the first inequality is due to (D.2) and the last inequality holds due to Lemma 15. □

Lemma 21. *For any $a \in \mathcal{A}$ and $a' \in \mathcal{A}\backslash\{a\}$ we have*

$$\sum_{a'\in\mathcal{A}}\mathbb{E}[M_{a',f^*_a}(T)] \leq \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] + |\mathcal{F}|Z(T,\Delta) + \mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)]\right)$$

Proof. For any agent $a \in \mathcal{A}$ we know that at every time step it either gets matched with some firm or gets collided. This implies

$$\sum_{f'\in\mathcal{F}}\mathbb{E}[C_{a,f'}(T)] + \sum_{f'\in\mathcal{F}\backslash\{f^*_a\}}\mathbb{E}[M_{a,f'}(T)] + \mathbb{E}[M_{a,f^*_a}(T)] = T. \tag{D.4}$$

Furthermore, in T steps the firm f^*_a can get matched with some agents or remain unmatched. This implies

$$\sum_{a'\in\mathcal{A}\backslash\{a\}}\mathbb{E}[M_{a',f^*_a}(T)] + \mathbb{E}[M_{a,f^*_a}(T)] \leq T. \tag{D.5}$$

Combining (D.4), (D.5) and Lemma 15 we see that

$$\begin{aligned}&\sum_{a'\in\mathcal{A}}\mathbb{E}[M_{a',f^*_a}(T)] \leq \sum_{f'\in\mathcal{F}}\mathbb{E}[C_{a,f'}(T)] + \sum_{f'\in\mathcal{F}\backslash\{f^*_a\}}\mathbb{E}[M_{a,f'}(T)]\\ &\leq \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] + |\mathcal{F}|\log(T)\right) + \mathcal{O}\left(\mathbb{E}[M_{a,\underline{\mathbb{F}}_a}(T)] + \mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)]\right).\end{aligned}$$

Note that from Lemma 14 we have

$$\begin{aligned}\sum_{a'\in\mathcal{A}}\mathbb{E}[M_{a',f^*_a}(T)] &\leq \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] + |\mathcal{F}|\log(T) + |\underline{\mathbb{F}}_a|Z(T,\Delta) + \mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)]\right)\\ &\leq \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\left(H_{a,f^*_a}(t)\right)\right] + |\mathcal{F}|Z(T,\Delta) + \mathbb{E}[M_{a,\bar{\mathbb{F}}_a}(T)]\right)\end{aligned}$$

This completes the proof.

□

E Thompson Sampling based Decentralized Matching Algorithm

E.1 Algorithmic Description

In this section we present a variant of Algorithm 2 but with Thompson sampling based stochastic bandit subroutine. For simplicity, we consider the scenario where the noise in (3.1) is sampled from a normal distribution. To compute the Thompson sampling index each agent a maintains an empirical average of utility generated from any firm f till time t which is $\hat{\mu}_{a,f}(t-1)$. At time step t any agent $a \in \mathcal{A}$ will maintain an index of every firm $f \in \mathcal{F}$ by sampling it from a normal distribution with mean $\hat{\mu}_{a,f}(t-1)$ and variance $\frac{1}{\sum_{f\in\mathcal{F}} M_{a,f}}$ (refer line 3 in Algorithm 5).

Algorithm 5: Thompson Sampling based Decentralized Matching Algorithm (TS-DMA)

Initialize : $\hat{\mu}_{a,f} = 0, M_{a,f} = 0, p_{a,f} = 0.5, x_{a,f} = 0.5, L_{a,f} = 0, \forall a \in \mathcal{A}, f \in \mathcal{F}$

1 **for** $t = 1, \ldots, T$ **do**
2 **for** $f \in \mathcal{F}$ **do**
3 Sample $\mathcal{T}_{a,f} \sim \mathcal{N}\left(\hat{\mu}_{a,f}, \frac{1}{\bar{M}_a}\right)$, where $\bar{M}_a = \sum_{f\in\mathcal{F}} M_{a,f}$
4 **end**
5 Set $\mathcal{T}_a = \mathsf{ArgDescendingSort}(\{\mathcal{T}_{a,f}\}_{f\in\mathcal{F}})$, $i = 1$
6 **while** $i \leq n$ **do**
7 Set $f = \mathcal{T}_a^{[i]}$
8 Sample $P_{a,f} \sim \mathsf{Bernoulli}(p_{a,f})$
9 **if** $P_{a,f} = 0$ **then**
10 Update $(x_{a,f}, p_{a,f}, L_{a,f}) \longrightarrow \mathsf{AB\_Subroutine}(P_{a,f}, x_{a,f}, p_{a,f}, L_{a,f}, Y_a)$
11 **end**
12 **if** $P_{a,f} = 1$ **then**
13 Query firm f and receive (U_a, Y_a)
14 Update $\hat{\mu}_{a,f} \longrightarrow Y_a \frac{\hat{\mu}_{a,f} M_{a,f} + U_a}{M_{a,f}+1} + (1 - Y_a)\hat{\mu}_{a,f}$ and $M_{a,f} \longrightarrow M_{a,f} + Y_a$,
15 Update $(x_{a,f}, p_{a,f}, L_{a,f}) \longrightarrow \mathsf{AB\_Subroutine}(P_{a,f}, x_{a,f}, p_{a,f}, L_{a,f}, Y_a)$
16 break while;
17 **end**
18 $i \longrightarrow i + 1$
19 **end**
20 **if** $i = |\mathcal{F}| + 1$ **then**
21 Query a firm $\mathcal{T}_a^{[1]}$ and receive (U_a, Y_a)
22 Update $\hat{\mu}_{a,f} \longrightarrow Y_a \frac{\hat{\mu}_{a,f} M_{a,f} + U_a}{M_{a,f}+1} + (1 - Y_a)\hat{\mu}_{a,f}$, $M_{a,f} \longrightarrow M_{a,f} + Y_a$
23 **end**
24 **end**

E.2 Bounds for Algorithm 5

We first present the regret bound for Algorithm 5.

Theorem 22. *Suppose every agent $a \in \mathcal{A}$ uses Algorithm 5. Then for any $i \in [K]$:*

$$\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \mathbb{E}[\mathcal{R}_a(T)] = \mathcal{O}\left(C_i |\mathcal{F}||\mathcal{A}| \left(\frac{1}{\Delta^2} \log\left(\frac{1}{\Delta}\right) + \frac{\log(T)}{\Delta^2} + \log(T)\right)\right)$$

where $\Delta = \min_{a,f} \Delta_{a,f}$ and C_i is a constant dependent on market $\mathcal{M}_i$ and $C_1 < C_2 < ... < C_K$.

The only difference between proof of Theorem 5 and Theorem 22 is the bound on expected number of matchings with suboptimal firms (refer **(L2)** in Lemma 6). We now present the analogue of **(L2)** of Lemma 6 below.

Lemma 23. *For any $i \in [K]$, the expected matches with suboptimal firm satisfies*

$$\begin{aligned}
&\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \mathbb{E}[M_{a,\underline{\mathbb{F}}_a}(T)] \\
&= \mathcal{O}\left(\sum_{j=1}^{i} \sum_{a \in \mathcal{A}_j} \left(|\underline{\mathbb{F}}_a| \left(\frac{1}{\Delta^2} \log\left(\frac{1}{\Delta}\right) + \frac{\log(T)}{\Delta^2} + \log(T)\right) + \mathbb{E}\left[\sum_{t=1}^{T} H_{a,f_a^*}(t)\right]\right)\right)
\end{aligned}$$

where $\Delta = \min_{a,f} \Delta_a(f)$

Proof. Note that we call an agent a matches with firm f at time t if $Y_a(t) = 1$ and $f_a(t) = f$. Therefore the total number of matchings between a and f till time T is $M_{a,f}(T) = \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f\right)$. Therefore from Lemma 12 and Remark 13 the following holds for every $f \in \underline{\mathbb{F}}_a$:

$$\begin{aligned}
M_{a,\underline{\mathbb{F}}_a}(T) &= \sum_{f \in \underline{\mathbb{F}}_a} \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f\right) \\
&\leq \sum_{f \in \underline{\mathbb{F}}_a} \sum_{t=1}^{T} \left(\mathbb{1}\left(Y_a(t) = 1, f_a(t) = f, \mathcal{T}_{a,f}(t) \geq \mathcal{T}_{a,f_a^*}(t)\right) + \mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f_a^*}^{(\mathsf{r})} = 0\right)\right) \\
&\leq \sum_{f \in \underline{\mathbb{F}}_a} \sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f, \mathcal{T}_{a,f}(t) \geq \mathcal{T}_{a,f_a^*}(t)\right) \\
&\qquad + \sum_{t=1}^{T} \sum_{f \in \underline{\mathbb{F}}_a} \mathbb{1}\left(E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f_a^*}^{(\mathsf{r})} = 0\right) \\
&\leq \sum_{f \in \underline{\mathbb{F}}_a} \underbrace{\sum_{t=1}^{T} \mathbb{1}\left(Y_a(t) = 1, f_a(t) = f, \mathcal{T}_{a,f}(t) \geq \mathcal{T}_{a,f_a^*}(t)\right)}_{\text{Term A}} + \underbrace{\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f_a^*}^{(\mathsf{r})} = 0\right)}_{\text{Term B}}
\end{aligned}$$

Let's first analyze Term A. Define $\mathcal{F}_{t-1} = \{\{f_a(\tau), Y_a(\tau), U_a(\tau)\}_{\tau=1}^{t-1}\}_{a \in \mathcal{A}}$. We first observe that

$$\begin{aligned}
&\mathbb{1}\left(Y_a(t) = 1, E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, \mathcal{T}_{a,f_a^*} \leq \mathcal{T}_{a,f}(t)\right) \\
&= \underbrace{\mathbb{1}\left(Y_a(t) = 1, E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, \mathcal{T}_{a,f_a^*} \leq \mathcal{T}_{a,f}(t), \mathcal{T}_{a,f}(t) < \hat{\mu}_{a,f_a^*} - \epsilon\right)}_{\text{Term C}} \\
&+ \underbrace{\mathbb{1}\left(Y_a(t) = 1, E_{a,f}^{(\mathsf{r})}(t) = 1, E_{a,f}^{(\mathsf{c})}(t) = 1, \mathcal{T}_{a,f_a^*} \leq \mathcal{T}_{a,f}(t), \mathcal{T}_{a,f}(t) \geq \hat{\mu}_{a,f_a^*} - \epsilon\right)}_{\text{Term D}}
\end{aligned} \tag{E.1}$$

We first provide a bound on Term C. Prior to that let's define some notations. Let's define $G^{(s)}_{a,f}(\epsilon) = 1 - F^{(s)}_{a,f}(\hat{\mu}_{a,f^*_a} - \epsilon)$. Furthermore, conditioned on the event that atleast one arm is pulled, for any agent a let's define $\mathcal{P}_a(t)$ to be the set of arms that are pruned before one is chosen to be played at time t. Moreover let $\tilde{A}^{\mathsf{select}}_{a,f}(t)$ be a random variable such that $\tilde{A}^{\mathsf{select}}_{a,f}(t) = 1$ iff f is the firm with maximum index value in all of the non-pruned arms at time t. That is, $\tilde{A}^{\mathsf{select}}_{a,f}(t) = \mathbb{1}\left(f \in \arg\max_{f' \in \mathcal{F}\setminus\{\mathcal{P}(t)\cup\{f^*_a\}\}} \mathcal{T}_{a,f'}(t)\right)$. Using this the following holds:

$$\begin{aligned}
&\mathbb{E}[\text{Term C}] = \mathbb{E}[\mathbb{E}[\text{Term C}|\mathcal{F}_{t-1}]] \\
&= \mathbb{E}[\mathsf{Pr}\left(Y_a(t)=1, E^{(\mathsf{r})}_{a,f}(t)=1, E^{(\mathsf{c})}_{a,f}(t)=1, \mathcal{T}_{a,f^*_a} \le \mathcal{T}_{a,f}(t), \mathcal{T}_{a,f}(t) < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)] \\
&\le \mathbb{E}\left[\mathsf{Pr}\left(\mathcal{T}_{a,f^*_a} < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)\mathsf{Pr}\left(Y_a(t)=1, \tilde{A}^{\mathsf{select}}_{a,f}(t)=1, \mathcal{T}_{a,f}(t) < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)\right]
\end{aligned} \tag{E.2}$$

Moreover note that

$$\begin{aligned}
&\mathsf{Pr}\left(Y_a(t)=1, E^{(\mathsf{c})}_{a,f^*_a}(t)=1, \mathcal{T}_{a,f}(t)(t) < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right) \\
&\ge \mathsf{Pr}\left(Y_a(t)=1, \tilde{A}^{\mathsf{select}}_{a,f}(t)=1, \mathcal{T}_{a,f}(t)(t) < \hat{\mu}_{a,f^*_a} - \epsilon, \mathcal{T}_{a,f^*_a}(t) > \hat{\mu}_{a,f^*} - \epsilon|\mathcal{F}_{t-1}\right) \\
&= \mathsf{Pr}\left(\mathcal{T}_{a,f^*_a}(t) > \hat{\mu}_{a,f^*_a}(t-1) - \epsilon|\mathcal{F}_{t-1}\right)\mathsf{Pr}\left(Y_a(t)=1, \tilde{A}^{\mathsf{select}}_{a,f}(t)=1, \mathcal{T}_{a,f}(t)(t) < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)
\end{aligned} \tag{E.3}$$

Using (E.3) in (E.2) we obtain the following

$$\begin{aligned}
\mathbb{E}[\text{Term C}] &= \mathbb{E}\Bigg[\frac{\mathsf{Pr}\left(\mathcal{T}_{a,f^*_a} < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)}{\mathsf{Pr}\left(\mathcal{T}_{a,f^*_a}(t) > \hat{\mu}_{a,f^*_a}(t-1) - \epsilon|\mathcal{F}_{t-1}\right)} \cdot \\
&\qquad \mathsf{Pr}\left(Y_a(t)=1, E^{(\mathsf{c})}_{a,f^*_a}(t)=1, \mathcal{T}_{a,f}(t)(t) < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)\Bigg] \\
&= \mathbb{E}\left[\frac{1 - G^{(M_{a,f^*_a}(t-1))}_{a,f^*_a}(\epsilon)}{G^{(M_{a,f^*_a}(t-1))}_{a,f^*_a}(\epsilon)}\mathsf{Pr}\left(Y_a(t)=1, E^{(\mathsf{c})}_{a,f^*_a}(t)=1, \mathcal{T}_{a,f}(t)(t) < \hat{\mu}_{a,f^*_a} - \epsilon|\mathcal{F}_{t-1}\right)\right] \\
&\le \mathbb{E}\left[\frac{1 - G^{(M_{a,f^*_a}(t-1))}_{a,f^*_a}(\epsilon)}{G^{(M_{a,f^*_a}(t-1))}_{a,f^*_a}(\epsilon)}\mathsf{Pr}\left(Y_a(t)=1, E^{(\mathsf{c})}_{a,f^*_a}(t)=1|\mathcal{F}_{t-1}\right)\right]
\end{aligned}$$

Further evaluating the expectation of Term C we have:

$$\begin{aligned}
\mathbb{E}[\text{Term C}] &= \sum_{t=1}^{T}\mathbb{E}\left[\frac{1 - G^{(M_{a,f^*_a}(t-1))}_{a,f^*_a}(\epsilon)}{G^{(M_{a,f^*_a}(t-1))}_{a,f^*_a}(\epsilon)}\mathbb{1}\left(E^{(\mathsf{c})}_{a,f^*_a}(t)=1, E^{(\mathsf{r})}_{a,f^*_a}(t)=1, Y_a(t)=1\right)\right] \\
&= \sum_{t=1}^{T}\sum_{s=1}^{t}\mathbb{E}\left[\frac{1 - G^{(s)}_{a,f^*_a}(\epsilon)}{G^{(s)}_{a,f^*_a}(\epsilon)}\mathbb{1}\left(E^{(\mathsf{c})}_{a,f^*_a}(t)=1, E^{(\mathsf{r})}_{a,f^*_a}(t)=1, Y_a(t)=1, M_{a,f^*_a}(t-1)=s\right)\right] \\
&\le \mathbb{E}\left[\sum_{s=1}^{T}\frac{1 - G^{(s)}_{a,f^*_a}(\epsilon)}{G^{(s)}_{a,f^*_a}(\epsilon)}\sum_{t=s+1}^{T}\mathbb{1}\left(M_{a,f}(t-1)=s, M_{a,f}(t)=s+1\right)\right] \\
&\le \sum_{s=0}^{\infty}\frac{1 - G^{(s)}_{a,f^*_a}(\epsilon)}{G^{(s)}_{a,f^*_a}(\epsilon)} \le \frac{1}{\epsilon^2}\log(\frac{1}{\epsilon})
\end{aligned}$$

where the last inequality is due to [LS20]. Now let's look at Term D. Let's set of time indices when $\mathcal{J}_{a,f} = \{t : G_{a,f}^{(M_{a,f}(t-1))}(\epsilon) > 1/T\}$.

$$\begin{aligned}\mathbb{E}[\text{Term D}] &= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left(Y_a(t)=1, E_{a,f}^{(\mathsf{r})}(t)=1, E_{a,f}^{(\mathsf{c})}(t)=1, \mathcal{T}_{a,f_a^*} \le \mathcal{T}_{a,f}(t), \mathcal{T}_{a,f}(t) \ge \hat{\mu}_{a,f_a^*} - \epsilon\right)\right] \\ &\le \underbrace{\sum_{t\in\mathcal{J}_{a,f}} \mathbb{E}\left[\mathbb{1}\left(Y_a(t)=1, E_{a,f}^{(\mathsf{r})}(t)=1\right)\right]}_{\text{Term E}} + \underbrace{\sum_{t\notin\mathcal{J}_{a,f}} \mathbb{E}\left[\mathbb{1}\left(\mathcal{T}_{a,f}(t) \ge \hat{\mu}_{a,f_a^*} - \epsilon\right)\right]}_{\text{Term F}}\end{aligned}$$

Let's first analyze the Term E above. Note that

$$\begin{aligned}&\sum_{t\in\mathcal{J}_{a,f}} \mathbb{1}\left(Y_a(t)=1, E_{a,f}^{(\mathsf{r})}(t)=1\right) \\ &\le \sum_{t=1}^{T}\sum_{s=1}^{t-1} \mathbb{1}\left(Y_a(t)=1, E_{a,f}^{(\mathsf{r})}(t)=1, G_{a,f}^{s}(\epsilon) > \frac{1}{T}, M_{a,f}(t-1)=s, M_{a,f}(t)=s+1\right) \\ &= \sum_{s=0}^{T-1} \mathbb{1}\left(G_{a,f}^{(s)}(\epsilon) > \frac{1}{T}\right) \sum_{t=s+1}^{T} \mathbb{1}\left(M_{a,f}(t-1)=s, M_{a,f}(t)=s+1\right) \\ &= \sum_{s=0}^{T-1} \mathbb{1}\left(G_{a,f}^{(s)}(\epsilon) > \frac{1}{T}\right) \le \mathcal{O}\left(\frac{\log(T)}{(\Delta_{a,f}-\epsilon)^2} + \log(T)\right)\end{aligned}$$

where the last property is a property of concentration of normal distribution and is standard in frequentist Thompson sampling analysis. For reader's reference we point to the book [LS20]. Next, we bound Term F below:

$$\begin{aligned}&\sum_{t\notin\mathcal{J}_{a,f}} \mathbb{E}\left[\mathbb{1}\left(\mathcal{T}_{a,f}(t) \ge \hat{\mu}_{a,f_a^*} - \epsilon\right)\right] = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}\left(\mathcal{T}_{a,f}(t) \ge \hat{\mu}_{a,f_a^*} - \epsilon, G_{a,f}^{(M_{a,f}(t-1))}(\epsilon) \le \frac{1}{T}\right)\right] \\ &= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\mathcal{T}_{a,f}(t) \ge \hat{\mu}_{a,f_a^*} - \epsilon, G_{a,f}^{(M_{a,f}(t-1))}(\epsilon) \le \frac{1}{T}\right)\right] |\mathcal{F}_{t-1}\right] \\ &= \sum_{t=1}^{T} \mathbb{E}\left[G_{a,f}^{(M_{a,f}(t-1))}(\epsilon)\mathbb{1}\left(G_{a,f}^{(M_{a,f}(t-1))}(\epsilon) < \frac{1}{T}\right)\right] \\ &\le 1\end{aligned}$$

Combining the bounds on Term C, Term E and Term F and choosing $\epsilon = \frac{\Delta}{2}$ we have

$$\begin{aligned}\sum_{f\in\underline{\mathbb{F}}_a} \mathbb{E}[M_{a,f}(T)] &\le |\underline{\mathbb{F}}_a|\mathcal{O}\left(\frac{1}{\Delta^2}\log\left(\frac{1}{\Delta}\right) + \frac{\log(T)}{\Delta^2} + \log(T)\right) \\ &\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(E_{a,f_a^*}^{(\mathsf{c})}(t)=1, E_{a,f_a^*}^{(\mathsf{r})}(t)=0\right)\right] \\ &\le |\underline{\mathbb{F}}_a|\mathcal{O}\left(\frac{1}{\Delta^2}\log\left(\frac{1}{\Delta}\right) + \frac{\log(T)}{\Delta^2} + \log(T)\right) + \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left(H_{a,f_a^*}(t)\right)\right]\right)\end{aligned}$$

where the second inequality is due to Lemma 20. This concludes the proof. □

F Table of Notations

We have accumulated all the main notations used in the paper in form of table below

Notation	Description
$\mathcal{A}$	Set of agents
$\mathcal{F}$	Set of firms/arms
$\mathcal{M}$	Union of agents and firms
$u_a(f)$	Utility for agent a when matched with firm f
$u_f(a)$	Utility for firm f when matched with agent a
$f_a(t)$	Firm chosen by agent a at time t
f_a^*	Stable match of agent a
$\overline{\mathbb{F}}_a$	Set of super-optimal firms for agent a
$\underline{\mathbb{F}}_a$	Set of sub-optimal firms for agent a
K	Number of markets formed by decomposition as stated in Remark 3
$\mathcal{A}_i$	Agents forming fixed pairs after $i-1$ rounds of elimination (Remark 3)
$\mathcal{F}_i$	Firms forming fixed pairs after $i-1$ rounds of elimination (Remark 3)
$U_{a,f}$	Noisy reward that agent a receives on getting matched with firm f
$\mathbb{A}_f$	Set of agents that pull firm f
$M_{a,f}(T)$	Number of times agent a has successfully matched with firm f till time T
$C_{a,f}(T)$	Number of times agent a has collided on firm f till time T
$p_{a,f}(t)$	Probability that agent a will pull firm f at time t
$P_{a,f}(t)$	An indicator if agent a has pulled arm f at time t
$Y_a(t)$	An indicator if agent a got successfully matched at time t
$\hat{\mu}_{a,f}(t)$	Empirical mean of utility derived by agent a on matching with f
$\mathsf{UCB}_{a,f}(t)$	UCB estimate of reward from firm f to agent a at time t
$\mathcal{T}_{a,f}(t)$	Thompson Sampling index of reward from firm f to agent a at time t
$E_{a,f}^{(\mathsf{r})}(t)$	An indicator if agent a pulled firm f at time t
$E_{a,f}^{(\mathsf{c})}(t)$	An indicator if all the firms with higher index than f got pruned at time t
$\tau_{a,f}(T)$	Time steps during which $E_{a,f}^{(\mathsf{c})}(t)=1$
$\Delta_{a,f}$	$u_a(f_a^*)-u_a(f)$

Table 1: Table of notations