Two-Sided Time-Independent Regret for Matching Markets with Limited Interviews

Amirmahdi Mirfakhar
University of Massachusetts Amherst
smirfakhar@umass.edu

Xuchuang Wang
Hong Kong Baptist University
xuchuangw@hkbu.edu.hk

Mengfan Xu
University of Massachusetts Amherst
mengfanxu@umass.edu

Hedyeh Beyhaghi
University of Massachusetts Amherst
hbeyhaghi@umass.edu

Mohammad Hajiesmaili
University of Massachusetts Amherst
mhajiesmaili@umass.edu

Abstract

Two-sided matching platforms depend on the preferences of both sides, yet each participant can evaluate only a small fraction of potential partners. In practice, participants rely on low-cost pre-match screening, such as interviews, profile views, or trial tasks, to form noisy impressions before committing to applications and offers. We study bandit learning in matching markets with n agents and m firms, modeling this screening as interviews that act as queried *hints*: they give both sides noisy observations of match quality, while an agent may apply only to firms it has interviewed. Firms, like agents, are uncertain about their preferences and may make early hiring mistakes. To handle this, we introduce strategic deferral, a firm-side action that leaves a position vacant for a round; it lets a firm undo a premature hire and, because vacancies are publicly observed, supports decentralized learning under coarse anonymous feedback. We design centralized and decentralized algorithms and show that two interviews per round suffice for horizon-independent regret (three in general markets without coordination), improving on the $O(\log T)$ guarantees known without interviews. The centralized bound is within a factor m of an information-theoretic lower bound; decentralized algorithms pay an additional $O(n^2)$ factor for distributing Gale–Shapley in structured markets and remain horizon-independent in general markets.

1 Introduction

Two-sided matching markets [26] provide a foundational model for settings in which outcomes depend on the preferences and decisions of both sides, from classical applications such as labor markets [25] and school choice [1] to modern digital platforms such as creator–brand sponsorship markets, reciprocal recommenders, and research internship or collaboration markets [11, 4, 12, 3, 24]. Across these domains, participants face uncertainty about match quality and often rely on limited low-cost screening before committing to a full match. Formally, a market consists of n agents and m firms, each with preferences over the other side, and a matching is *stable* if no agent–firm pair would both prefer to deviate from their current assignments [9]. When these preferences are not known a priori, participants must learn them through interaction while evaluating only a small fraction of potential partners, motivating *bandit learning* in matching markets [18, 21, 19, 13, 10]. Much of the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

bandit-matching literature studies this problem under the simplifying assumptions that only agents are uncertain, while firms have known preferences. These assumptions miss two central features of the motivating settings above.

Two missing ingredients: pre-match interviews and uncertain firms. (i) First, participants often engage in low-cost pre-match interactions before submitting a final application or contract. We call these interactions *interviews*, but use the term broadly to include screening calls, profile views, and trial tasks. Such interactions produce *pre-application* noisy side-observations of match quality for both sides and can accelerate learning. This resembles *bandits with hints* [6, 22] and, more broadly, *algorithms with predictions* [20]. However, unlike prior hinted models focusing on single-agent settings, interviews in matching markets simultaneously *reveal information* and *restrict feasible actions*, since agents may only apply to interviewed firms [16, 5]. Thus, informative interviews can still intensify the matching competition.

(ii) Second, firms may also be uncertain about their preferences. A brand, employer, or user may initially misrank applicants based on noisy screening signals; if forced to always hire, early mistakes can reject strong candidates, slow convergence to stability, and increase regret. To model how firms hedge against such uncertainty, we allow a firm to *defer* hiring in a round, i.e., to remain vacant rather than commit to an applicant it currently deems suboptimal. We introduce *deferral* as a firm-side action that can be added to existing algorithms. It expands the firm's action space and helps decentralized learning reach a stable matching under limited feedback. A related phenomenon appears in static (non-learning) markets, where strategic rejections by a single firm can shift the stable outcome [14]; here, deferral serves a different purpose, letting a firm correct its own estimation errors. The closest prior work [23] also allows uncertainty on both sides, but without interviews or deferral; using explore-then-commit, it obtains $O(\log T)$ regret when agents share a "communication blackboard" and $O(T^{\gamma})$ regret for a tunable $\gamma \in (0, 1)$ otherwise.

Model at a glance. We study two-sided bandit learning with interviews and uncertainty on both sides. Our model follows the platform workflow above: *limited screening/interview first, then a higher-commitment application, then acceptance or deferral, followed by coarse firm-side status feedback*. In round t, every agent a selects k firms to interview. Each interviewed pair (a, f) yields stochastic noisy feedback to both sides with unknown means $u_{a,f}$ and $u_{f,a}$, representing their expected utilities from matching. After interviewing, each agent applies to one of its interviewed firms; a firm may accept its most-preferred applicant or defer and remain vacant, with both sides receiving stochastic rewards upon a match and zero otherwise. We evaluate performance via agent regret: the gap between the expected reward of the agent's stable match and its accrued reward over T rounds. Since stability is bilateral, agent-side regret bounds induce analogous firm-side guarantees under the corresponding stable benchmark. We study both centralized and decentralized settings. In the centralized setting, an interview allocator coordinates interviews so that no two agents apply to the same firm. In the decentralized settings, agents choose interview sets independently and observe only coarse anonymous signals from firms: vacancy-only feedback, indicating whether a firm filled its position, or hiring-change feedback, indicating whether a firm's hire changed. Without such a signal, an agent rejected by a firm cannot tell when that firm's hire later changes, and the market can remain at an unstable matching indefinitely (Example 2.4).

A guiding question and technical challenges. Existing bandit-matching work without interviews typically yields regret scaling at least logarithmically in T [18, 21, 19, 8, 17], with problem-dependent constants hidden in $O(\log T)$ term. Interviews provide side observations before applications are made, so learning about a firm no longer requires matching with it. This removes the per-sample exploration cost behind logarithmic regret in standard bandits and suggests that stable outcomes may be learned much faster. This leads to our central question: *Can a two-sided market learn a stable matching with near-optimal horizon-independent regret using only a constant number of low-cost pre-match interactions per participant in each round, in a decentralized manner with only anonymous firm-side status signals, even when firms are uncertain and may defer hiring?* Answering this requires addressing two challenges: (1) *Uncertain firms and controlled deferral:* when firms learn their rankings, early misrankings can induce rejections of strong applicants, creating feedback loops that redirect agents and slow convergence. Deferral can mitigate premature hires, but must be controlled so that it corrects instability without becoming a new source of regret. (2) *Interviews as hints under competition and stability:* in single-agent hinted bandits [6, 29], hints provide side-information alongside a reward each round. In our multi-agent two-sided setup, interviews provide side-information but do not guarantee rewards: when interview sets overlap, competition intensifies

Table 1: Summary of regret bounds for n agents and m firms; Δ is the minimum reward gap and ϵ is defined in Theorem 4.2; the exact gap dependence is detailed in the corresponding theorem.

Setting	Algorithm	Market	Feedback	Regret
Centralized	Algorithm 6	General	N/A	$O(nm^2\Delta^{-2})$
Coordinated decentralized	Algorithm 7	α-reducible General	Vacancy-only	$O(n^3m^2\Delta^{-2})$ $O(n^4m^2\Delta^{-2})$
Coordination-free decentralized	Algorithm 3 Algorithm 9	α-reducible General	Hiring-change	$O(n^3m^2\Delta^{-2})$ $O\big(\epsilon^{-1}n^5m^4\Delta^{-2}\big)$
Lower bound (information-theoretic)	—	General	N/A	$\Omega(nm\Delta^{-2})$

and many agents may remain unmatched since each firm hires at most one. Combined with two-sided uncertainty, deferral, and anonymous status signals, this makes horizon-independent regret substantially harder than in either classical bandit matching or single-agent hinted bandits.

Contributions. We introduce a two-sided bandit model with interviews and uncertain firms, showing for the first time that a constant number of interviews enable near-optimal *horizon-independent* regret for stable matchings in both centralized and decentralized markets, improving upon the $O(\log T)$ bounds of prior work [18, 19, 21] under significantly weaker assumptions.

► *Modeling uncertain firms with a deferral option.* We introduce *deferral* as a novel modular primitive for two-sided learning (Section 2), the first to extend firms' action space beyond always hiring (Example 2.4), pluggable into prior works to relax the strong assumption of certain firms. We design a deferral policy (Algorithm 1) provably avoiding uncontrolled deferrals while enabling stable two-sided learning: a firm defers only when an agent it displaced since its last vacancy is now estimated above its current top applicant, so every deferral is charged to an earlier estimation error, with broader implications for strategic firm manipulations in matching markets.

► *Two interviews per round suffice for horizon-independent regret.* We formalize interviews as bandit hints in a two-sided market and show that *two interviews per round* using empirical mean estimators suffice for horizon-independent regret (Theorem 3.1), provably converging to an *actual stable matching*, not merely matchings with zero regret that need not be stable. One interview goes to the firm the agent applies to and the other cycles over all firms in round-robin order, decoupling exploration from applications: every pair is sampled once every m rounds however the market evolves, so the expected number of rounds with invalid preference estimates is $O(m\Delta^{-2})$ per pair, independent of T. Even in the simpler single-agent case with guaranteed rewards, the result resolves an open conjecture of [6], who proved three hints sufficient via UCB-V but conjectured three necessary under empirical means; we show *two suffice*. We extend this guarantee to the richer two-sided setting with uncertain firms and strategic deferral under anonymous signals.

► *Centralized allocation: learning the agent-optimal stable matching.* In the centralized setting (Section 3), we design a central interview allocator (`CIA`) that coordinates interviews and applications. Building on deferred acceptance [9], we give an algorithm that learns the agent-optimal stable matching in $O(nm^2)$ regret under both certain and uncertain firms (Theorem 3.1). We show this bound is within a factor m of the information-theoretic lower bound $\Omega(nm)$, and conjecture $O(nm^2)$ is tight, leaving the matching $\Omega(nm^2)$ lower bound as an open problem.

► *Decentralized learning.* Under vacancy-only feedback $\mathcal{V}$ (Section 4.2), a coordinated algorithm (Algorithm 7) achieves $O(n^3m^2)$ regret in structured and $O(n^4m^2)$ in general markets (Theorem 4.1), reflecting the natural $O(n^2)$ cost of distributing `GS`; beyond this cost, the bounds are within factors m and nm of the lower bound. Under anonymous hiring-change feedback $\mathcal{V}^+$ (Section 4.3), a coordination-free algorithm (Algorithm 3) matches the coordinated bound in structured markets and achieves horizon-independent regret in general markets with three interviews (Theorem 4.2). There, cyclic blocking-pair dynamics introduce large constants, as in [19]; in contrast to [21], our structured-market bounds avoid such constants. Tight lower bounds here remain a compelling open problem. All results hold under significantly weaker assumptions than [19], which requires firm preferences known to both sides a priori and match identities revealed each round.

2 Model and Preliminaries

Consider a two-sided market $\mathcal{M}(\mathcal{A}, \mathcal{F})$ with n agents $\mathcal{A}$ and m firms $\mathcal{F}$, where $n \leq m$. The terminology is generic: in the platform applications motivating this work, agents may be creators, job seekers, users, or mentees, while firms may be brands, employers, campaigns, mentors, or reciprocal-recommendation opportunities; Appendix A gives a detailed mapping.

Utilities and Preferences. Each agent–firm pair (a, f) is associated with two reward distributions: $\mathcal{D}_{a,f}$ for agent $a \in \mathcal{A}$, with mean $u_{a,f}$, and $\mathcal{D}_{f,a}$ for firm $f \in \mathcal{F}$, with mean $u_{f,a}$. Rewards are independent across agent–firm pairs and time steps, and are bounded in $[0, 1]$. We assume a heterogeneous market with no ties: for every agent a and distinct firms $f, f' \in \mathcal{F}$, $u_{a,f} \neq u_{a,f'}$, and for every firm f and distinct agents $a, a' \in \mathcal{A}$, $u_{f,a} \neq u_{f,a'}$. Thus preferences are strict and induce well-defined rankings on both sides.

- *Ground-truth preferences.* For each agent $a \in \mathcal{A}$, the *ground-truth preference list* O_a ranks firms by decreasing $u_{a,f}$, and we write $f \underset{O_a}{>} f'$ if $u_{a,f} > u_{a,f'}$. Analogously, each firm $f \in \mathcal{F}$ has a ground-truth preference list O_f ranking agents by decreasing $u_{f,a}$, with $a \underset{O_f}{>} a'$ defined similarly.
- *Estimated preferences.* Agents do not know their preferences *a priori* and must learn them from observations. On the firm side, we consider two cases: *certain firms*, which know their preference lists in advance, and *uncertain firms*, which must learn them over time. Each agent maintains an estimate $\hat{u}_{a,f}(t)$ of $u_{a,f}$, and each firm maintains an estimate $\hat{u}_{f,a}(t)$ of $u_{f,a}$. The induced estimated preference lists are denoted by $\hat{O}_a(t)$ and $\hat{O}_f(t)$. For a certain firm f, $\hat{u}_{f,\cdot}(t) = u_{f,\cdot}$ and hence $\hat{O}_f(t) = O_f$ for all $t \in \mathcal{T}$.

While agents and firms act on their estimated preferences, stability is defined with respect to the ground-truth preferences, as follows.

Definition 2.1 (Stable matching)**.** A *stable matching* is a one-to-one matching μ between agents and firms such that no *blocking pair* exists, i.e., no agent a and firm f that prefer each other to their current matches under the ground-truth preferences: $f \underset{O_a}{>} \mu(a)$ and $a \underset{O_f}{>} \mu(f)$.

Decision Process. Agents and firms interact over T rounds, indexed by $\mathcal{T}$ with $|\mathcal{T}| = T$; each round $1 \leq t \leq T$ consists of four stages, *interview*, *application*, *hiring*, and *firm-side feedback revelation*, as follows:

- *Interview stage.* Each agent $a \in \mathcal{A}$ selects a subset of firms $\mathcal{I}_a(t) \subseteq \mathcal{F}$ to interview[1], with size $2 \leq |\mathcal{I}_a(t)| \leq k$, where $k \in \mathbb{N}^+$ is the *interview budget*. For each interviewed firm $f \in \mathcal{I}_a(t)$, the agent observes a stochastic signal $X_{a,f}(t) \sim \mathcal{D}_{a,f}$ drawn from distribution $\mathcal{D}_{a,f}$ with mean $u_{a,f}$; notably, these signals are not counted as rewards. Similarly, each firm f interviews agents in $\mathcal{I}_f(t)$ and observes $X_{f,a}(t) \sim \mathcal{D}_{f,a}$ for each $a \in \mathcal{I}_f(t)$.
- *Application stage.* After obtaining the interview results, each agent a applies to a set of interviewed firms. In all but one of our algorithms this set is a single firm, which we denote by $f_a^{\texttt{apply}}(t) \in \mathcal{I}_a(t)$; the exception is Algorithm 9, where, with some probability, the set consists of two interviewed firms, and the agent accepts the better offer according to its current estimates.
- *Hiring stage.* Each firm f collects the applications it receives from agents in round t, denoted by $\mathcal{P}_f(t) = \{a \in \mathcal{A} : f_a^{\texttt{apply}}(t) = f\}$. If firm f decides to hire, it admits its most preferred applicant within $\mathcal{P}_f(t)$ according to its current preference as

$$a_f^{\texttt{match}}(t) = \underset{a \in \mathcal{P}_f(t)}{\arg\max}\, \hat{u}_{f,a}(t). \tag{1}$$

  If $a_{f_a^{\texttt{apply}}(t)}^{\texttt{match}}(t) = a$, then agent a is hired by firm $f_a^{\texttt{apply}}(t)$ and receives a stochastic reward $X_{a,f_a^{\texttt{apply}}(t)}(t) \sim \mathcal{D}_{a,f_a^{\texttt{apply}}(t)}$; otherwise, agent a receives no reward. We denote the match of agent a at round t by $f_a^{\texttt{match}}(t)$ where $f_a^{\texttt{match}}(t) = f_a^{\texttt{apply}}(t)$ if a is admitted, and $f_a^{\texttt{match}}(t) = \emptyset$ otherwise. The match of firm f is similarly denoted $a_f^{\texttt{match}}(t)$, and agent a's reward is

$$X_{a,f_a^{\texttt{match}}(t)} = X_{a,f_a^{\texttt{apply}}(t)} \mathbb{1}\left\{f_a^{\texttt{match}}(t) \neq \emptyset\right\}. \tag{2}$$

[1] Although, we refer to this stage as an *interview*, it is more akin to an open house or participation in a career fair.

- *Firm-side feedback revelation stage.* At the end of round t, agents observe one of two types of firm-side feedback:

$$\mathcal{V}(t) \doteq \{f \in \mathcal{F} : a_f^{\texttt{match}}(t) = \emptyset\}, \quad (3) \qquad \mathcal{V}^+(t) \doteq \mathcal{V}(t) \cup \{f \in \mathcal{F} : a_f^{\texttt{match}}(t-1) \neq a_f^{\texttt{match}}(t)\}. \quad (4)$$

  Here, the *vacancy-only* feedback $\mathcal{V}(t)$ is the set of firms that are vacant at time t, whereas *anonymous hiring changes* $\mathcal{V}^+(t)$ includes firms whose hiring status has changed since the previous round, without revealing their current matches, on top of the vacancies; i.e., $\mathcal{V} \subseteq \mathcal{V}^+$. We view $\mathcal{V}(t)$ as the weaker signal and $\mathcal{V}^+(t)$ as the stronger one; this feedback determines what information agents observe and plays a central role in our algorithm design.

Each round thus yields a matching $(a, f_a^{\texttt{match}}(t))_{a\in\mathcal{A}}$ between agents and firms, and our goal is for it to become stable (Definition 2.1).

Optimal and Pessimal Regrets. For policy π, we use regret to quantify performance from the agents' perspective. Since multiple stable matchings may exist, we focus on agent-optimal and agent-pessimal stable matches with corresponding analysis depending on whether we are in the centralized or decentralized model. More precisely, following the prior work [19], we define the *optimal* and *pessimal* regrets under policy π for agent a over T rounds as follows

$$\underbrace{\mathbb{E}[\overline{R}_a(\pi,T)]}_{a\text{-Optimal}} \doteq \sum_{t=1}^{T} u_{a,\overline{f_a^*}} - \mathbb{E}[X_{a,f_a^{\texttt{match}}(t)}(t)], \qquad \underbrace{\mathbb{E}[\underline{R}_a(\pi,T)]}_{a\text{-Pessimal}} \doteq \sum_{t=1}^{T} u_{a,\underline{f_a^*}} - \mathbb{E}[X_{a,f_a^{\texttt{match}}(t)}(t)], \quad (5)$$

where $\overline{f_a^*}$ and $\underline{f_a^*}$ are the best and worst stable matches of agent a. For brevity, we omit the policy π from the regret and use $\mathbb{E}[\overline{R}_a(T)]$ and $\mathbb{E}[\underline{R}_a(T)]$, hereinafter. For each agent $a \in \mathcal{A}$ and firm $f \in \mathcal{F}$, we define the *optimal reward gap* as $\overline{\Delta}_{a,f} \doteq |u_{a,\overline{f_a^*}} - u_{a,f}|$ and the *pessimal reward gap* as $\underline{\Delta}_{a,f} \doteq |u_{a,\underline{f_a^*}} - u_{a,f}|$. Similarly, for firms, we define the optimal reward gap as $\overline{\Delta}_{f,a} \doteq |u_{f,\underline{a_f^*}} - u_{f,a}|$ and the pessimal reward gap as $\underline{\Delta}_{f,a} \doteq |u_{f,\overline{a_f^*}} - u_{f,a}|$, where $\overline{a_f^*}$ and $\underline{a_f^*}$ denote the firm-optimal and pessimal stable matches for firm f.

Remark 2.2. The agent-optimal stable matching is firm-pessimal, and vice versa [9]; hence the firm-side gaps $\overline{\Delta}_{f,a}$ and $\underline{\Delta}_{f,a}$ are measured against $\underline{a_f^*}$ and $\overline{a_f^*}$, respectively.

When the market admits a unique stable matching, the optimal and pessimal regrets coincide, and we write $\mathbb{E}[R_a(T)]$, $\Delta_{a,f}$, and $\Delta_{f,a}$; α-reducible markets, introduced next, are a key example.

α-Reducible Markets. A key class of markets with a unique stable matching is α-reducible markets, introduced for uniqueness by [7] and later used by [21] for coordination-free decentralized bandit learning. This structure is less restrictive than serial dictatorship [27, 28], while still covering a broad family of unique-stable-matching markets. Formally, a pair $(a, f) \in \mathcal{A} \times \mathcal{F}$ is a *fixed pair* if a and f are mutual top choices, i.e., $f \underset{O_a}{>} f'$ for all $f' \in \mathcal{F} \setminus \{f\}$, and $a \underset{O_f}{>} a'$ for all $a' \in \mathcal{A} \setminus \{a\}$.

Definition 2.3. A market $\mathcal{M}(\mathcal{A}, \mathcal{F})$ is *α-reducible* if every $\mathcal{M}(\mathcal{A}', \mathcal{F}')$ with $\mathcal{A}' \subseteq \mathcal{A}$ and $\mathcal{F}' \subseteq \mathcal{F}$ contains a fixed pair.

In α-reducible markets, we can iteratively remove fixed pairs: there exists a fixed pair (a_1, f_1); removing it yields an α-reducible sub-market; repeating yields pairs $(a_i, f_i)_{i\in[n]}$, which match all agents and induce the unique stable matching with firms in $\mathcal{F} \setminus \{f_i\}_{i\in[n]}$ left unmatched. This layered structure enables a cleaner analysis, which highlights our techniques and yields sharper bounds than in general markets.

2.1 Extended Action Space for Firms' Uncertainty

A key departure from prior work is that firms may be *uncertain* about their own preferences. If an uncertain firm is forced to hire whenever it has applicants, early estimation errors can "lock in" an unstable outcome: the firm keeps hiring the agent it currently, but incorrectly, believes is best, while agents may receive no local evidence that the firm has learned otherwise. We therefore extend the firm action space to allow a firm to *strategically abstain* from hiring in selected rounds. In platform terms, this corresponds to leaving a campaign slot open or postponing a hire or mentorship. Crucially, abstention makes the firm publicly vacant through $\mathcal{V}(t)$ and $\mathcal{V}^+(t)$, giving previously rejected agents a coarse signal that can unwind decentralized deadlocks.

Example 2.4 (Why abstention is necessary)**.** *Let $\mathcal{M}(\mathcal{A}, \mathcal{F})$ have agents $\mathcal{A} = \{a_1, a_2\}$ and (strategic) uncertain firms $\mathcal{F} = \{f_1, f_2\}$ with ground-truth preferences*

$$O_{a_1} : \langle f_1^\dagger \succ f_2 \rangle \quad O_{f_1} : \langle a_1^\dagger \succ a_2 \rangle$$
$$O_{a_2} : \langle f_1 \succ f_2^\dagger \rangle \quad O_{f_2} : \langle a_2^\dagger \succ a_1 \rangle.$$

The unique stable matching is $(a_i, f_i)_{i\in[2]}$ (marked by †s). Suppose that at time t the estimated preferences are

$$\hat{O}_{a_1}(t) : \langle f_1 \succ f_2^* \rangle \quad \hat{O}_{f_1}(t) : \langle a_2^* \succ a_1 \rangle$$
$$\hat{O}_{a_2}(t) : \langle f_1^* \succ f_2 \rangle \quad \hat{O}_{f_2}(t) : \langle a_2 \succ a_1^* \rangle,$$

so only f_1 is wrong (it swaps a_1 and a_2). The stable matching under $\hat{O}(t)$ is (a_1, f_2) and (a_2, f_1), marked by $(*)$s, which is unstable under the ground truth.

Now assume that at some later time $t' > t$, firm f_1 learns the correct order, i.e., $\hat{O}_{f_1}(t') = O_{f_1}$. In a decentralized setting, this correction may be unobservable *to agents, so if firms always hire, the market can remain stuck at $(a_1, f_2), (a_2, f_1)$. If instead f_1 can abstain at t', it can* reject *its current hire: by not hiring, f_1 becomes vacant and appears in the firm-side feedback, which all agents observe. This vacancy triggers a_2 to move to f_2, which displaces a_1 back to f_1, restoring the stable matching $(a_i, f_i)_{i\in[2]}$.*

Formally, we allow each firm $f \in \mathcal{F}$ to choose a private binary hiring decision $\gamma_f(t) \in \{0, 1\}$ each round, where $\gamma_f(t) = 1$ means hire and $\gamma_f(t) = 0$ means remain vacant. The decision is taken after interviews, once f observes its applicant set $\mathcal{P}_f(t)$. With this modification, the realized reward of agent a becomes

$$X_{a,f_a^{\texttt{match}}(t)} = \gamma_{f_a^{\texttt{apply}}(t)}(t)\, X_{a,f_a^{\texttt{apply}}(t)}\, \mathbb{1}\{f_a^{\texttt{match}}(t) \neq \emptyset\}. \tag{6}$$

Accordingly, beyond agent-side learning, we introduce a *strategic rejection policy* (Algorithm 1) that selects $\gamma_f(t)$ each round. We call any firm running Algorithm 1 *strategic*, even if it never abstains; for a strategic *certain* firm, Algorithm 1 never triggers abstention, so $\gamma_f(t) = 1$ for all t. Firms that hire every round are *non-strategic* (even if uncertain). We defer the formal policy to Section 4.1.

2.2 Algorithmic Paradigms and Preliminaries

Here, we first formally present two different algorithm design paradigms *centralized* and *decentralized*, and then introduce the preliminaries common to both.

Centralized vs. Decentralized Algorithms. We study two settings. In the *centralized* setting, a coordinator allocates interviews to avoid collisions. In the *decentralized* setting, agents choose interview sets independently, so multiple agents may apply to the same firm; such collisions must be resolved using firm-side feedback, either $\mathcal{V}(t)$ or $\mathcal{V}^+(t)$.

- *Centralized algorithms.* We assume an omniscient **C**entral **I**nterview **A**llocator (`CIA`) that observes the estimated preference lists of all agents and firms and selects the interview sets $\mathcal{I}_a(t)$ at each round.
- *Decentralized algorithms.* Without `CIA`, each agent a selects $\mathcal{I}_a(t)$ from its *local observations* (Definition D.1), which include: (i) its estimated preference lists, (ii) whether its application was rejected, and (iii) the firm-side feedback from the previous round.

Algorithmic Preliminaries. Here, we introduce the preliminaries of our algorithmic designs.

Empirical means as estimators. We use empirical averages to estimate expected utilities for each agent–firm pair (a, f). For agent a and firm f at time t, the estimator is $\hat{u}_{a,f}(t) = \frac{\sum_{t'=1}^{t} X_{a,f}(t')\, \mathbb{1}\{f \in \mathcal{I}_a(t')\}}{N_{a,f}(t)}$, where $N_{a,f}(t) = \sum_{t'=1}^{t} \mathbb{1}\{f \in \mathcal{I}_a(t')\}$ is the number of interviews of a at f up to time t. For uncertain firms, the estimator $\hat{u}_{f,a}(t)$ is defined analogously.

The structure of the interview set $\mathcal{I}_a(t)$. Across all our algorithms, each agent interviews between two and k firms per round, i.e., $2 \leq |\mathcal{I}_a(t)| \leq k$. In the main body, we focus on the case $k = 2$, i.e., showing only two interviews is sufficient for time-independent regret, and defer a special extension to unstructured markets with $k = 3$ to Section G.5. The interviewing set is $\mathcal{I}_a(t) = \{f_a^{\texttt{RR}}(t), f_a^{\texttt{apply}}(t)\}$, where $f_a^{\texttt{apply}}(t)$ is the firm to which agent a applies and $f_a^{\texttt{RR}}(t)$ is selected in a round-robin manner for exploration, as follows.

- *Interviewing round-robin firm* $f_a^{\texttt{RR}}(t)$. Each agent interviews a designated firm $f_a^{\texttt{RR}}(t)$ such that over each block of m rounds, every firm appears exactly once as the round-robin firm for every agent, ensuring uniform coverage. Formally, for agents indexed as a_i and firms as f_j, we set $f_{a_i}^{\texttt{RR}}(t) = f_{(t+i \bmod m)+1}$. This round-robin exploration resembles the exploration phase of Explore-then-Commit (ETC) schemes [15], but here it is specifically designed to achieve *time-independent regret* rather than the $O(\log T)$ behavior typical of ETC.
- *Interviewing to-be-applied firm* $f_a^{\texttt{apply}}(t)$. The other firm $f_a^{\texttt{apply}}(t)$ that agent a interviews with and applies to is chosen via the Gale–Shapley algorithm (GS) (or its distributed variant), which computes a stable matching with respect to the given preference lists of both sides.

3 Centralized Learning

We study the centralized setting where CIA controls interview assignments, and present *Centralized Interview Allocation with Round-Robin Interviews* (Algorithm 6), which also creates a backbone for our decentralized design. At each round t, CIA runs GS [9] on the estimated lists $\{\hat{O}_a(t)\}_{a\in\mathcal{A}}$ and $\{\hat{O}_f(t)\}_{f\in\mathcal{F}}$, producing a stable matching that assigns each agent a a target firm $f_a^{\texttt{apply}}(t)$. The allocator then sets $\mathcal{I}_a(t) = \{f_a^{\texttt{apply}}(t), f_a^{\texttt{RR}}(t)\}$; agents interview both firms, update their estimates, and apply only to $f_a^{\texttt{apply}}(t)$. The following theorem gives *time-independent* regret for all agents under Algorithm 6 with uncertain firms with the full proof presented in Section E.4. We show a factor m tightness of the regret in Theorem 3.1 in Section I.1.

Theorem 3.1. *In market $\mathcal{M}(\mathcal{A},\mathcal{F})$ with non-strategic firms, agent a's optimal regret under Algorithm 6 is $\mathbb{E}[\overline{R}_a(\mathcal{T})] \in O(nm^2 \min\{\overline{\Delta}_{\mathcal{A}}, \underline{\Delta}_{\mathcal{F}}\}^{-2})$, with $\overline{\Delta}_{\mathcal{A}} \doteq \min_{a\in\mathcal{A}} \overline{\Delta}_a$ and $\underline{\Delta}_{\mathcal{F}} \doteq \min_{f\in\mathcal{F}} \underline{\Delta}_f$.*

Intuitively, an agent incurs regret only until it ranks its stable match above each of the (at most m) firms it prefers less, each comparison settled within $O(m)$ rounds, giving $O(nm^2)$ overall; with the symmetric firm-side $O(n^2m)$, this is within a factor m of the $\Omega(nm)$ lower bound (Proposition I.1).

Proof Sketch. Following [18], for each agent $a \in \mathcal{A}$ and firm $f \in \mathcal{F}$, let $\mathcal{H}_{a,f}$ (resp. $\mathcal{L}_{a,f}$) denote the firms ranked above (resp. below) f in O_a, with estimated analogues $\hat{\mathcal{H}}_{a,f}(t)$ and $\hat{\mathcal{L}}_{a,f}(t)$ defined w.r.t. $\hat{O}_a(t)$; firm-side sets are defined analogously.

Definition 3.2 (Valid Preference Lists)**.** In a matching market $\mathcal{M}(\mathcal{A},\mathcal{F})$, $\hat{O}_a(t)$ is *valid* w.r.t. $f \in \mathcal{F}$ if $\hat{\mathcal{H}}_{a,f}(t) \subseteq \mathcal{H}_{a,f}$. Analogously, $\hat{O}_f(t)$ is *valid* w.r.t. $a \in \mathcal{A}$ if $\hat{\mathcal{H}}_{f,a}(t) \subseteq \mathcal{H}_{f,a}$.

Let $\bar{\mathcal{E}}_{a,f}$ denote the set of invalid rounds for pair (a,f) (with $\bar{\mathcal{E}}_{f,a}$ defined similarly). Regret accrues only in invalid rounds, and the following lemma (proved in Section E) bounds their number:

Lemma 3.3. *In a matching market $\mathcal{M}(\mathcal{A},\mathcal{F})$, for any agent $a \in \mathcal{A}$ with optimal stable match $\overline{f_a^*} \in \mathcal{F}$, Algorithm 6 guarantees $\mathbb{E}[|\bar{\mathcal{E}}_{a,\overline{f_a^*}}|] \in O(|\mathcal{L}_{a,\overline{f_a^*}}| \cdot m \cdot \Delta_a^{-2})$, where $\Delta_a \doteq \min_{f\neq\overline{f_a^*}} \overline{\Delta}_{a,f}$.*

An analogous bound $O(n|\mathcal{L}_{f,\underline{a_f^*}}| \cdot \underline{\Delta}_f^{-2})$ holds for firms. Summing over all agents and firms and bounding $|\mathcal{L}_{a,\overline{f_a^*}}| \leq m$ and $|\mathcal{L}_{f,\underline{a_f^*}}| \leq n$ yields Theorem 3.1. □

4 Decentralized Learning

We study decentralized learning without CIA, where interview decisions rely solely on local observations. Agents select interviewed firms $f_a^{\texttt{RR}}(t)$ via round-robin and apply to $f_a^{\texttt{apply}}(t)$ computed through distributed Gale-Shapley (GS) executions. Since GS is inherently global, each execution spans multiple rounds and incurs regret, so agents should trigger it only when local signals justify it. We use firm-side feedback as a shared coordination signal and design algorithms under both feedback models: *anonymous hiring changes* ($\mathcal{V}^+$) and *vacancy-only* ($\mathcal{V}$).

We introduce a firm-side *strategic rejection policy* (Algorithm 1, Section 4.1) applicable under both feedback models. Under *vacancy-only* feedback $\mathcal{V}(t)$ (Section 4.2), agents jointly trigger distributed GS, yielding Algorithm 7 with $O(n^3m^2)$ and $O(n^4m^2)$ regret for structured and general markets. Under the richer $\mathcal{V}^+(t)$ (Section 4.3), agents run independent GS threads without explicit coordination, yielding Algorithm 3 with $O(n^3m^2)$ and Algorithm 9 with $O(\epsilon^{-1}n^5m^4)$ regret, respectively.

Algorithm 1 Decentralized Strategic Rejection

Input: f, $\mathcal{A}$, $\mathcal{T}$, switching conditions $\mathcal{S}_f$ (8)
Initialize: $c_f(1) \leftarrow 0$, $r_{f,a}(1) \leftarrow 0$ for all $a \in \mathcal{A}$
for $t \in \mathcal{T}$ **do**
  $\gamma_f(t) \leftarrow 1$
  Interview with agents in $\mathcal{I}_f(t)$ and update $\hat{u}_{f,a}(t+1)$ for all $a \in \mathcal{I}_f(t)$
  Receive the set of applicants $\mathcal{P}_f(t)$
  $a_f^*(t) \leftarrow \arg\max_{a \in \mathcal{P}_f(t)} \hat{u}_{f,a}(t)$
  if $t \notin \mathcal{S}_f$ **then** $a_f^{\texttt{match}}(t) \leftarrow a_f^*(t)$ ▷ hiring top applicant
  else $\gamma_f(t) \leftarrow 0$, $a_f^{\texttt{match}}(t) \leftarrow 0$ ▷ rejecting all applicants
  UPDATEFIRMREJVARS(f, t) ▷ Subroutine 4

4.1 Strategic Firm's Rejection Policy

We introduce a firm-side policy, given in Algorithm 1, that specifies whether a firm hires in each round and, if so, which applicant it hires. The policy is the same under both firm-side feedback models, $\mathcal{V}(t)$ and $\mathcal{V}^+(t)$. At a high level, a strategic uncertain firm $f \in \mathcal{F}$ abstains from hiring when its updated estimates suggest it may have mis-ranked a previously rejected agent relative to its current top applicant. After interviewing and observing the applicant set, the firm compares its current top applicant to previously rejected agents: if the top applicant is not estimated to dominate every such *active* rejected agent, the firm sets $\gamma_f(t) = 0$, rejects all applicants, and becomes vacant (appearing in $\mathcal{V}$ and $\mathcal{V}^+$); otherwise, it hires its estimated top applicant and updates its rejection records.

We now formalize the conditions under which a strategic uncertain firm performs a strategic rejection. Each firm $f \in \mathcal{F}$ maintains two private state variables, $r_{f,a}(t)$ and $c_f(t)$. The variable $r_{f,a}(t)$ records the most recent round prior to t in which f rejected agent a in favor of another hire, while $c_f(t)$ records the most recent round prior to t in which f remained vacant. We interpret $r_{f,a}(t) \geq c_f(t)$ as: agent a was rejected after the last vacancy, so f may need to reconsider a as estimates evolve. Let $a_f^*(t) \doteq \arg\max_{a \in \mathcal{P}_f(t)} \hat{u}_{f,a}(t)$ denote the firm's current top applicant in $\mathcal{P}_f(t)$. Formally,

$$r_{f,a}(t) \doteq \arg\max_{t'<t} \left\{ \gamma_f(t') = 1 \ \wedge \ \begin{matrix} f_a^{\texttt{apply}}(t') = f \\ a_f^{\texttt{match}}(t') \neq a \end{matrix} \right\}, \qquad c_f(t) \doteq \arg\max_{t''<t} \left\{ a_f^{\texttt{match}}(t'') = \emptyset \right\}. \tag{7}$$

The set of rounds where firm f abstains from hiring is

$$\mathcal{S}_f \doteq \{ t \in \mathcal{T} : \exists a \in \hat{\mathcal{H}}_{f,a_f^*(t)}(t) \ \wedge \ r_{f,a}(t) \geq c_f(t) \}. \tag{8}$$

This condition captures the case where a previously rejected agent (displaced by another hire) is now estimated to be preferred to the current top applicant, triggering a *strategic rejection* (deferred hiring) rather than a permanent one. One can observe that under Algorithm 1 a certain firm always hires.

4.2 Coordinated Decentralized Algorithm with Vacancy-Only Feedback $\mathcal{V}(t)$

In this section, we design a novel *coordinated* decentralized learning algorithm under vacancy-only feedback $\mathcal{V}(t)$. We first highlight the high-level idea and the main technical challenges of designing such algorithms under limited feedback while maintaining robustness to strategic rejections by firms. We then present the key algorithmic components and the resulting regret guarantees. An extensive explanation of the details and analysis is deffered to the Section F.2.

Necessity of Coordination. A key challenge under vacancy-only feedback $\mathcal{V}(t)$ is that hiring changes remain hidden. This can obscure critical market transitions — for instance, two agents may simultaneously swap matches without changing the vacancy set, leaving others to act on stale information (see Example F.1). An agent rejected by a firm, for example, may never learn that the firm's match has since changed in their favor.

Achieving $O(1)$ Regret under Indistinguishable Vacancy Signals. Under $\mathcal{V}(t)$, the vacancy signal is the only public information shared among agents, so coordination must be embedded within it. An agent can signal by abstaining from applying, thereby creating a vacancy. However, since this feedback is anonymous, agents cannot distinguish coordination signals from ordinary interviewing

Algorithm 2 Modular Coordinated Learning Process

Initialize: $t^{\mathtt{GS}} \leftarrow 1$, $\rho_a \leftarrow 0$, and $r_{a,f} \leftarrow 0$ for all $f \in \mathcal{F}$
for $t \in \mathcal{T}$ **do**
  if $\rho_a = 0$ **then**
    Run CoordinatedUpdate $3n^2$ rounds from $t^{\mathtt{GS}}$ using $\hat{u}_{a,f}(t^{\mathtt{GS}})$; then set $\rho_a \leftarrow 1$
  if $\rho_a = 1$ **then**
    Commit to the output of the last update
  if $t \in \mathcal{S}_a^{\mathrm{inc}} \cup \mathcal{S}_a^{\mathrm{rej}} \cup \mathcal{S}_a^{\mathrm{vac}}$ **then**
    $f_a^{\mathtt{apply}}(t) \leftarrow \emptyset$, $r_{a,f} \leftarrow 0$, $t^{\mathtt{GS}} \leftarrow t+1$, and $\rho_a \leftarrow 0$

set change or strategic firm rejections. The challenge is to design a scheme that triggers and encodes coordination correctly while maintaining time-independent regret and bounded coordination cost.

Algorithmic Design. To address the challenges above, we design a coordinated framework that partitions time into alternating *updating* and *committing* phases, encoded by $\rho_a(t) \in \{0,1\}$, where $\rho_a(t) = 0$ denotes updating and $\rho_a(t) = 1$ denotes committing. The process starts with an updating phase: agents coordinate on the interview sets for the next committing phase using estimated preference lists fixed at $t^{\mathtt{GS}}$, the synchronously maintained start of the most recent updating phase, which is initially set to be 0. After updating, they commit to this outcome while acting independently. During committing, agents monitor local events indicating that the lists fixed at $t^{\mathtt{GS}}$ may have become invalid, and hence that the coordinated matching may no longer be stable. Upon detecting such an event, an agent signals coordination by refusing to apply, thereby altering the public vacancy feedback $\mathcal{V}(t)$, causing all agents to synchronously set a new $t^{\mathtt{GS}}$ and enter another updating phase.

It remains to specify how agents coordinate, when an updating phase ends, and which local events during committing should trigger coordination. These rules must be robust to strategic rejections. We address this via fixed-length updating phases and a small set of local inconsistency and vacancy-based triggers, all defined through a private rejection-time variable $r_{a,f}(t)$, which records the most recent time $t' < t$ at which agent a applied to firm f and was non-strategically rejected.

- *Updating (Coordination).* Agents coordinate for a fixed $3n^2$ rounds starting at $t^{\mathtt{GS}}$, using a common snapshot of their estimated preferences at time $t^{\mathtt{GS}}$. At the start of this phase, all agents reset $r_{a,f}$ to zero. Each agent a then selects actions from firms that have not rejected it since $t^{\mathtt{GS}}$: $\mathcal{B}_a(t) \doteq \{f \in \mathcal{F} : r_{a,f}(t) < t^{\mathtt{GS}}\}$, $f_a^{\mathtt{apply}}(t) = \arg\max_{f \in \mathcal{B}_a(t)} \hat{u}_{a,f}(t^{\mathtt{GS}})$. This phase is shown, by Lemma F.2, to end with applications $(a, f_a^{\mathtt{apply}}(t^{\mathtt{GS}} + 3n^2))_{a \in \mathcal{A}}$ forming a perfect matching, which agents then commit to in the subsequent committing phase.
- *Switching to Updating.* During committing, agents continue applying to the firm selected at the end of the last updating phase. Agents monitor a small set of local triggers; upon detecting any such event, an agent signals coordination by altering $\mathcal{V}(t)$, e.g., by not applying, which creates an observable vacancy because the committed applications form a matching. This causes all agents to synchronously set $t^{\mathtt{GS}} \leftarrow t+1$ and enter a new updating phase. The set of switching events are (i) preference-list inconsistencies $\mathcal{S}_a^{\mathrm{inc}}$, (ii) rejections due to strategic firm behavior $\mathcal{S}_a^{\mathrm{rej}}$, and (iii) unexpected vacancy as coordination signals $\mathcal{S}_a^{\mathrm{vac}}$.

Regret Guarantee. Under the above policies, summarized in Algorithm 2 and given in full as Algorithm 7, we obtain time-independent regret for both α-reducible and general markets under $\mathcal{V}(t)$, with a proof sketch (full proofs in Section F.6 and Section F.7.4).

Theorem 4.1. *In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, with firms and agents following Algorithm 1 and Algorithm 7, the expected regret is time-independent. In particular,*

$$\mathbb{E}[\underline{R}_a(\mathcal{T})] \in \begin{cases} O(n^3 m^2 \min\{\overline{\Delta}_{\mathcal{A}}, \underline{\Delta}_{\mathcal{F}}\}^{-2} + m\sum_{j=1}^{i}(|\mathcal{L}_{a_j,f_j}|\,\overline{\Delta}_{a_j}^{-2} + |\mathcal{L}_{f_j,a_j}|\,\underline{\Delta}_{f_j}^{-2})), & \alpha\text{-reducible,} \\ O(n^4 m^2 \Delta^{-2}), & \text{General Markets.} \end{cases}$$

The extra n^2 factor over Theorem 3.1 is the decentralization cost of spreading each GS run over $O(n^2)$ rounds, and general markets lose one more factor n, as regret can no longer be charged to a unique stable match; hence the bounds are within factors m and nm of optimal for α-reducible and general markets, respectively (Section I.2).

Algorithm 3 Coordination-Free Decentralized Learning

Input: a, $\mathcal{F}$, anonymous hiring changes $\mathcal{V}^+$, **Initialize:** $r_{a,f}(1) \leftarrow 0$ for all $f \in \mathcal{F}$
for $t \in \mathcal{T}$ **do**
  Construct candidate set $\mathcal{B}'_a(t)$ according to (9)
  Set $f_a^{\texttt{apply}}(t) \leftarrow \arg\max_{f \in \mathcal{B}'_a(t)} \hat{u}_{a,f}(t)$
  Interview with $\mathcal{I}_a(t) = \{f_a^{\texttt{apply}}(t), f_a^{\texttt{RR}}(t)\}$ and apply to $f_a^{\texttt{apply}}(t)$
  UPDATEAGENTREJVARS(a, t)

4.3 Coordination-Free Decentralization with Anonymous Hiring Changes $\mathcal{V}^+(t)$ Feedback

In this section, we design a *coordination-free* decentralized learning algorithm under the richer firm-side feedback $\mathcal{V}^+(t)$. Utilizing globally revealed hiring changes, the algorithm avoids coordination while remaining robust to uncertain firms' strategic rejections, and achieves time-independent regret. We first outline the main idea and technical challenges, then present the key algorithmic components and regret analysis. Detailed descriptions and analysis are deferred to the Section G.1.

Heterogeneous reactions and instability under anonymity. In the absence of coordination, agents react independently to observed hiring changes based on their local estimates, leading to inconsistent updates that can slow or block convergence to stable matching. This is further complicated by anonymous firm-side feedback and strategic rejections, which may induce several misleading signals.
Cyclic blocking-pair dynamics. Independent reactions may also lead to simultaneous attempts to resolve blocking pairs. In general markets with multiple stable matchings, this can induce persistent cycles and linear regret, requiring additional randomization to break such dynamics.

Algorithmic Design. To address the challenges above, we use a simple coordination-free rule for agents, whose analysis is substantially more delicate. Each agent maintains the same rejection-time variable $r_{a,f}(t)$ and forms a candidate set of firms that either have never rejected it or have exhibited a hiring change since their last rejection:

$$\mathcal{B}'_a(t) \doteq \left\{f : \exists\, t' \in [r_{a,f}(t), t),\ f \in \mathcal{V}^+(t')\right\}, \qquad f_a^{\texttt{apply}}(t) = \arg\max_{f \in \mathcal{B}'_a(t)} \hat{u}_{a,f}(t). \tag{9}$$

As in Algorithm 3, the agent interviews $\mathcal{I}_a(t) = \{f_a^{\texttt{apply}}(t), f_a^{\texttt{RR}}(t)\}$ and applies to $f_a^{\texttt{apply}}(t)$ that suffices for α-reducible markets to achieve time independent regret. For general markets, where cyclic blocking-pair resolutions may arise, we use a randomized variant deferred to the appendix (Algorithm 9) with $k = 3$ interviews: $\mathcal{I}_a(t) = \{f_a^{\texttt{apply}}(t), f_a^{\texttt{match}}(t-1), f_a^{\texttt{RR}}(t)\}$. With probability $\lambda \in (0, 1)$, the agent applies to both its current choice and its previous match, accepting the better offer; otherwise it keeps applying to its previous match.

Regret Guarantee. Under the above policies, given in Algorithm 3 and its extension Algorithm 9, we obtain time-independent regret for both α-reducible and general markets under $\mathcal{V}^+(t)$ (full proofs in Section G.4 and Section G.5.6 respectively).

Theorem 4.2. *In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, with firms following Algorithm 1 and agents following coordination-free Algorithm 3 (for α-reducible markets with $k = 2$) and Algorithm 9 (for general markets with $k = 3$, where an agent may apply to two interviewed firms per round), the expected regret is time-independent. In particular,*

$$\mathbb{E}[\underline{R}_a(\mathcal{T})] \in \begin{cases} O\big((\sum_{j\in[i]} |\mathcal{H}_{a_j,f_j}| + 1) \cdot (\sum_{j\in[i]} (m^2 \overline{\Delta}_{a_j}^{-2} + nm\, \underline{\Delta}_{f_j}^{-2}) + 1)\big), & \alpha\text{-reducible } \mathcal{M}(\mathcal{A}, \mathcal{F}), \\ O\left(\epsilon^{-1} n^5 m^4 \Delta^{-2}\right), \text{ for } \epsilon = \left(\lambda(1-\lambda)^{n-1}\right)^{n^4 m + nm} \text{ with } \lambda \in (0,1), & \text{General Markets}. \end{cases}$$

Without coordination, the α-reducible bound adapts to the rounds GS needs, at most $O(n^3 m^2)$, or $O(n^2 m^2)$ if GS ends in $O(n)$ rounds; in general markets, ϵ^{-1} is the price of breaking blocking-pair cycles by randomization (Section I.3).

Conclusion. We study bandit learning in matching markets with interviews and two-sided uncertainty. We present centralized and decentralized algorithms that achieve time-independent regret with an anonymous firm-side signal. The key ingredient is handling firm uncertainty via an extended action space enabling strategic deferral, which stabilizes decentralized learning under limited feedback and improves prior work along multiple dimensions. Incentive compatibility, optimality, comparisons, and future directions are in Appendixes H, I, J, and K.

Acknowledgments

This research is supported by National Science Foundation grants SaTC-2512128, CNS-2533814, and Expeditions-2325956.

References

[1] Atila Abdulkadiroğlu and Tayfun Sönmez. School choice: A mechanism design approach. *American economic review*, 93(3):729–747, 2003.

[2] Hernan Abeledo and Uriel G Rothblum. Paths to marriage stability. *Discrete applied mathematics*, 63(1):1–12, 1995.

[3] Nick Arnosti, Ramesh Johari, and Yash Kanoria. Managing congestion in matching markets. *Manufacturing & Service Operations Management*, 23(3):620–636, 2021.

[4] Itai Ashlagi, Mark Braverman, Yash Kanoria, and Peng Shi. Clearing matching markets efficiently: Informative signals and match recommendations. *Management Science*, 66(5):2163–2193, 2020.

[5] Hedyeh Beyhaghi and Éva Tardos. Randomness and fairness in two-sided matching with limited interviews. In *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

[6] Aditya Bhaskara, Sreenivas Gollapudi, Sungjin Im, Kostas Kollias, and Kamesh Munagala. Online learning and bandits with queried hints. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference, ITCS 2023, MIT, Cambridge, Massachusetts, USA, January 10-13, 2023*, LIPIcs, pages 16:1–16:24. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023.

[7] Simon Clark. The uniqueness of stable matchings. *Contributions in Theoretical Economics*, 6(1):1–28, 2006.

[8] Xiaowu Dai and Michael I. Jordan. Learning strategies in decentralized matching markets under uncertain preferences. *Journal of Machine Learning Research*, 22:260:1–260:50, 2021.

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

[10] Avishek Ghosh, Abishek Sankararaman, Kannan Ramchandran, Tara Javidi, and Arya Mazumdar. Competing bandits in non-stationary matching markets. *IEEE Transactions on Information Theory*, 70(4):2831–2850, 2024.

[11] Günter J Hitsch, Ali Hortaçsu, and Dan Ariely. Matching and sorting in online dating. *American Economic Review*, 100(1):130–163, 2010.

[12] Yash Kanoria and Daniela Saban. Facilitating the search for partners on matching platforms. *Management Science*, 67(10):5990–6029, 2021.

[13] Fang Kong and Shuai Li. Player-optimal stable regret for bandit learning in matching markets. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1512–1522. SIAM, 2023.

[14] Ron Kupfer. The influence of one strategic agent on the core of stable matchings. *arXiv preprint arXiv:1806.04034*, 2018.

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

[16] Robin S Lee and Michael Schwarz. Interviewing in two-sided matching markets. *The RAND Journal of Economics*, 48(3):835–855, 2017.

[17] Yuantong Li, Guang Cheng, and Xiaowu Dai. Two-sided competing matching recommendation markets with quota and complementary preferences constraints. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 27589–27614. PMLR / OpenReview.net, 2024.

[18] 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.

[19] Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *The Journal of Machine Learning Research*, 22(1):9612–9645, 2021.

[20] Thodoris Lykouris and Sergei Vassilvitskii. Competitive caching with machine learned advice. *Journal of the ACM (JACM)*, 68(4):1–25, 2021.

[21] Chinmay Maheshwari, Shankar Sastry, and Eric Mazumdar. Decentralized, communication-and coordination-free learning in structured matching markets. *Advances in Neural Information Processing Systems*, 35:15081–15092, 2022.

[22] Amirmahdi Mirfakhar, Xuchuang Wang, Jinhang Zuo, Yair Zick, and Mohammad Hajiesmaili. Heterogeneous multi-agent bandits with parsimonious hints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19467–19474, 2025.

[23] Tejas Pagare and Avishek Ghosh. Explore-then-commit algorithms for decentralized two-sided matching markets. In *IEEE International Symposium on Information Theory, ISIT 2024, Athens, Greece, July 7-12, 2024*, pages 2092–2097. IEEE, 2024.

[24] Ignacio Rios, Daniela Saban, and Fanyin Zheng. Improving match rates in dating markets through assortment optimization. *Manufacturing & Service Operations Management*, 25(4):1304–1323, 2023.

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

[26] Alvin E Roth and Marilda Sotomayor. Two-sided matching. *Handbook of game theory with economic applications*, 1:485–541, 1992.

[27] 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.

[28] Zilong Wang and Shuai Li. Optimal analysis for bandit learning in matching markets with serial dictatorship. *Theoretical Computer Science*, 1010:114703, 2024.

[29] Chen-Yu Wei, Haipeng Luo, and Alekh Agarwal. Taking a hint: How to leverage loss predictors in contextual bandits? In *Conference on Learning Theory*, pages 3583–3634. PMLR, 2020.

[30] Yirui Zhang and Zhixuan Fang. Decentralized two-sided bandit learning in matching market. pages 4173–4191, 2024.

A Motivating Applications

Our model is motivated by several real-world matching applications. Below, we highlight a few representative examples that capture the main primitives of our model: low-cost pre-match screening, higher-commitment applications, firm-side acceptance or deferral, stochastic rewards, and limited feedback about other participants' outcomes. While our framework is designed to abstract these features, a detailed domain-specific mapping and performance analysis for any particular application would require substantial additional modeling effort and is beyond the scope of this work.

Creator–Brand Sponsorship Markets. Consider a platform that matches content creators to brands with scarce campaign slots. Each round t corresponds to a sponsorship cycle: every creator selects a small number of brands to screen or interview—through product seeding, trial affiliate links, or limited test posts—before committing to one pitch. These interactions are cheaper than a full sponsorship contract, but they provide noisy signals of match quality to both sides. The creator learns whether a brand fits their audience and long-run reputation, while the brand learns whether the creator is likely to deliver strong campaign performance with acceptable content quality and brand-safety risk. After screening, the creator pitches to one screened brand; the brand then accepts its currently most preferred applicant or defers by leaving the campaign slot open when the evidence is weak. A successful match yields stochastic value to both sides, such as compensation and audience fit for the creator and predicted campaign performance or product–audience alignment for the brand. The learning goal is to approach a stable benchmark under the true two-sided preferences while using only a constant number of low-cost pre-match observations per round.

Reciprocal Recommender and Recruiting Platforms. Consider a platform that recommends bilateral connections, such as online dating, professional networking, or recruiting [11, 3, 24]. Agents may be users, job seekers, mentees, or candidates, while firms may be potential partners or employers. Each round t corresponds to a search or recruiting cycle: every agent screens a small number of potential counterparties—through profile views, likes, resume screens, introductory calls, or technical screens—before committing to one serious request or application. These screening interactions are lower cost than a final match, but they provide noisy signals of match quality to both sides. The agent learns which counterparty best fits their preferences, career goals, or compatibility, while the firm learns which applicant is most promising given its own preferences, capacity constraints, and expected match quality. After screening (interview), the agent applies to one screened counterparty; the firm then accepts its currently most preferred applicant or defers, for example, by keeping a slot open, or continuing to search. A successful match yields a stochastic value to both sides. The learning goal is again to approach a stable benchmark under the true two-sided preferences using only a constant number of pre-match observations per round, with coarse feedback such as whether a slot remains open or whether a counterparty's match status changed.

Academic Research Internships. Consider a market in which research labs (agents) place their students and postdocs into industry internship positions (firms) each cycle. Each round t corresponds to a hiring cycle: every research lab sends one of its members to interview at a small number of industry positions—through campus visits, informational calls, or trial projects—before the member commits to a single application. These interviews serve as low-cost explorations that yield noisy signals of match quality for both sides, without constituting a formal commitment. After interviewing, the member applies to one position; the industry host then decides whether to extend an offer to its most preferred applicant or defer if uncertain about the fit. A successful match yields a reward reflecting the quality of the collaboration for both sides. Importantly, it is the lab as an institution—not just any individual member—that bears the objective: across cycles, different members may be placed, but the lab accumulates experience and refines its understanding of which industry partners best complement its research agenda. Performance is measured via regret: the gap between the cumulative reward the lab would have obtained by always matching with its best stable industry partner (in hindsight) and its actual accumulated reward over T cycles. The goal is to minimize this regret while using only a constant number of interviews per cycle, operating in a decentralized manner with limited feedback about other labs' outcomes.

B General Notation, Lemmas, and Observations for the Regret Analysis

In this section, we introduce general notation and collect auxiliary lemmas and observations used throughout our proofs, in particular in the proofs of Theorem 4.2 and Theorem 4.1 for unstructured general markets.

B.1 Top-k Ground-Truth and Estimated Preferences

We define, for each agent (resp., firm), the set of its top-k firms (resp., agents) under the ground-truth and estimated preference lists.

Definition B.1 (Top-k agents and firms)**.** In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, for each agent $a \in \mathcal{A}$ and integer $1 \leq k \leq m$, let $\mathcal{F}_a^{(k)}$ denote the top-k firms in a's ground-truth list O_a, and let $\hat{\mathcal{F}}_a^{(k)}(t)$ denote the top-k firms under a's estimated list $\hat{O}_a(t)$ at time t. Formally,

$$\mathcal{F}_a^{(k)} \doteq \{f \in \mathcal{F} : |\mathcal{H}_{a,f}| \leq k-1\}, \tag{10}$$

$$\hat{\mathcal{F}}_a^{(k)}(t) \doteq \left\{f \in \mathcal{F} : \left|\hat{\mathcal{H}}_{a,f}(t)\right| \leq k-1\right\}. \tag{11}$$

Similarly, for each firm $f \in \mathcal{F}$ and integer $1 \leq k \leq n$, let $\mathcal{A}_f^{(k)}$ denote the top-k agents in f's ground-truth list O_f, and let $\hat{\mathcal{A}}_f^{(k)}(t)$ denote the top-k agents under f's estimated list $\hat{O}_f(t)$ at time t. Formally,

$$\mathcal{A}_f^{(k)} \doteq \{a \in \mathcal{A} : |\mathcal{H}_{f,a}| \leq k-1\}, \tag{12}$$

$$\hat{\mathcal{A}}_f^{(k)}(t) \doteq \left\{a \in \mathcal{A} : \left|\hat{\mathcal{H}}_{f,a}(t)\right| \leq k-1\right\}. \tag{13}$$

B.2 Rounds of Top-k Alignment

We next define the rounds in which estimated preferences agree with the ground truth up to the top-k items. We first introduce an agreement operator and then relate it to validity events.

Definition B.2 (Top-k set agreement)**.** Fix an integer k. For an agent $a \in \mathcal{A}$ and time t, we write

$$\hat{\mathcal{F}}_a^{(k)}(t) \;\equiv\; \mathcal{F}_a^{(k)}$$

to mean that the top-k firms under $\hat{O}_a(t)$ and O_a agree both element-wise and order-wise, i.e., $\hat{\mathcal{F}}_a^{(k)}(t) = \mathcal{F}_a^{(k)}$ and $\hat{O}_a(t)$ and O_a induce the same order on $\mathcal{F}_a^{(k)}$. Similarly, for a firm $f \in \mathcal{F}$ and time t, we write

$$\hat{\mathcal{A}}_f^{(k)}(t) \;\equiv\; \mathcal{A}_f^{(k)}$$

to mean that the top-k agents under $\hat{O}_f(t)$ and O_f agree both element-wise and order-wise, i.e., $\hat{\mathcal{A}}_f^{(k)}(t) = \mathcal{A}_f^{(k)}$ and $\hat{O}_f(t)$ and O_f induce the same order on $\mathcal{A}_f^{(k)}$.

Lemma B.3. *Fix an integer k. For any agent $a \in \mathcal{A}$ and time t, we have $\hat{\mathcal{F}}_a^{(k)}(t) \equiv \mathcal{F}_a^{(k)}$ iff*

$$t \in \bigcap_{f \in \mathcal{F}_a^{(k)}} \mathcal{E}_{a,f}.$$

Likewise, for any firm $f \in \mathcal{F}$ and time t, we have $\hat{\mathcal{A}}_f^{(k)}(t) \equiv \mathcal{A}_f^{(k)}$ iff

$$t \in \bigcap_{a \in \mathcal{A}_f^{(k)}} \mathcal{E}_{f,a}.$$

Proof. We prove the claim for agents; the proof for firms is analogous.

First, assume $\hat{\mathcal{F}}_a^{(k)}(t) \equiv \mathcal{F}_a^{(k)}$. By definition of $\equiv$, for every $f \in \mathcal{F}_a^{(k)}$ the relative order of the firms above f is correct under $\hat{O}_a(t)$. Equivalently, for each such f we have $\hat{\mathcal{H}}_{a,f}(t) = \mathcal{H}_{a,f}$, so $\hat{O}_a(t)$ is valid with respect to f at time t, i.e., $t \in \mathcal{E}_{a,f}$. Hence $t \in \bigcap_{f \in \mathcal{F}_a^{(k)}} \mathcal{E}_{a,f}$.

Subroutine 4 UPDATEFIRMREJVARS(f, t)

if $t \notin \mathcal{S}_f$ **then**
  $r_{f,a}(t+1) \leftarrow t$ for all $a \in \mathcal{P}_f(t) \setminus \{a_f^{\texttt{match}}(t)\}$
  $c_f(t+1) \leftarrow c_f(t)$
else
  $c_f(t+1) \leftarrow t$ ▷ update latest vacancy

Subroutine 5 UPDATEAGENTREJVARS(a, t)

for all firms $f \in \mathcal{F}$ **do**
  $r_{a,f}(t+1) \leftarrow r_{a,f}(t)$
if $f_a^{\texttt{match}}(t) = \emptyset$ **and** $f_a^{\texttt{apply}}(t) \notin \mathcal{V}(t)$ **then**
  $r_{a,f_a^{\texttt{apply}}(t)}(t+1) \leftarrow t$ ▷ rejected in favor of another hire

Conversely, assume $t \in \bigcap_{f \in \mathcal{F}_a^{(k)}} \mathcal{E}_{a,f}$ but $\hat{\mathcal{F}}_a^{(k)}(t) \not\equiv \mathcal{F}_a^{(k)}$. Then there exists some $f' \in \mathcal{F}_a^{(k)}$ such that $\hat{O}_a(t)$ and O_a induce different orders on the relevant set $\mathcal{L}_{a,f'}$. Equivalently, there exists $f'' \in \mathcal{L}_{a,f'}$ with $f'' \in \hat{\mathcal{H}}_{a,f'}(t)$, which implies $t \in \bar{\mathcal{E}}_{a,f'}$, contradicting $t \in \mathcal{E}_{a,f'}$. Therefore $\hat{\mathcal{F}}_a^{(k)}(t) \equiv \mathcal{F}_a^{(k)}$. □

By Lemma B.3, top-k alignment can be characterized in terms of validity events. We therefore define the rounds in which *all* agents (resp., *all* firms) have top-k alignment.

Definition B.4 (Rounds of global top-k alignment)**.** Fix an integer k. We define $\Gamma_{\mathcal{A}}^{(k)}$ (resp., $\Gamma_{\mathcal{F}}^{(k)}$) as the set of time steps at which *all* agents (resp., *all* firms) have their top-k preferences aligned with the ground truth (element-wise and order-wise). Formally,

$$\Gamma_{\mathcal{A}}^{(k)} \doteq \left\{ t \in \mathcal{T} : \forall a \in \mathcal{A},\ t \in \bigcap_{f \in \mathcal{F}_a^{(k)}} \mathcal{E}_{a,f} \right\}, \tag{14}$$

$$\Gamma_{\mathcal{F}}^{(k)} \doteq \left\{ t \in \mathcal{T} : \forall f \in \mathcal{F},\ t \in \bigcap_{a \in \mathcal{A}_f^{(k)}} \mathcal{E}_{f,a} \right\}. \tag{15}$$

In particular, when k equals the full list size, top-k alignment reduces to exact recovery of the entire preference lists.

Observation 1. *We have $t \in \Gamma_{\mathcal{A}}^{(m)} \cap \Gamma_{\mathcal{F}}^{(n)}$ if and only if every agent and every firm has fully learned its preference list at time t, i.e., $\hat{O}_a(t) = O_a$ for all $a \in \mathcal{A}$ and $\hat{O}_f(t) = O_f$ for all $f \in \mathcal{F}$. Equivalently, $\Gamma_{\mathcal{A}}^{(m)} \cap \Gamma_{\mathcal{F}}^{(n)}$ is exactly the set of rounds at which the estimated market induced by $\hat{O}(t)$ coincides with the ground-truth matching market induced by O.*

The last one is an observation for the strategic rejection policy Algorithm 1.

Observation 2. *For any certain firm f, we have $t \notin \mathcal{S}_f$ for all t, so under Algorithm 1, $\gamma_f(t) = 1$ in every round.*

C Rejection Variables Update Pseudo-Codes (Algorithms 1, 7,3, and 9)

We collect two short subroutines for updating the *rejection variables* used throughout our decentralized algorithms. Algorithm 4 updates the firm-side private state: $r_{f,a}(t)$ is the last round before t in which firm f rejected agent a while hiring someone else, and $c_f(t)$ is the last round before t in which f remained vacant. Algorithm 5 updates the agent-side rejection time $r_{a,f}(t)$, the last round before t in which agent a applied to firm f, remained unmatched, and f did *not* appear vacant (i.e., f hired another agent).

D Deferred Concepts from the Model (Section 2) and Preliminaries (Section 2.2)

This section collects definitions and technical details deferred from Sections 2 and 2.2 to streamline the main presentation.

D.1 Definition of Local Observations

We define the local observations available to each agent which determine the information used to update estimates and make decisions in our decentralized algorithms.

Definition D.1 (Local Observations)**.** In the *decentralized* setup, each agent $a \in \mathcal{A}$ updates its interviewing set $\mathcal{I}_a(t)$ solely based on its own local observations at time t, which include:

1. Changes in its estimated preference list $\hat{O}_a(t-1)$ to $\hat{O}_a(t)$.
2. Whether it was rejected by the firm $f_a^{\texttt{apply}}(t-1)$.
3. The firm-side feedback $\mathcal{V}(t-1)$.

D.2 Definition of Convergence

We formalize the notion of convergence used throughout the paper, specifying when the sequence of matchings stabilizes under our learning dynamics.

Definition D.2 (Convergence of a Learning Algorithm)**.** A learning algorithm π is said to *converge* to a perfect matching $(a, f_a^{\texttt{match}}(t))_{a\in\mathcal{A}}$ if there exists a time step $t \in [T]$ such that for all $t' \geq t$, we have $f_a^{\texttt{match}}(t') = f_a^{\texttt{match}}(t)$ and $f_a^{\texttt{match}}(t') \neq \emptyset$ for every agent $a \in \mathcal{A}$.

E Deferred Concepts and Proofs from Section 3

This section collects additional definitions and deferred proofs from Section 3 to streamline the main presentation.

E.1 Pseudocode for the Centralized Algorithm

Here we provide the pseudocode of the centralized algorithm referenced in Section 3.

Algorithm 6 Centralized Interview Allocation

Input: $\mathcal{A}, \mathcal{F}, \mathcal{T}$
for $t \in \mathcal{T}$ **do**
  CIA collects the current estimated preference lists $\{\hat{O}_a(t)\}_{a\in\mathcal{A}}$ and $\{\hat{O}_f(t)\}_{f\in\mathcal{F}}$
  CIA runs $\texttt{GS}\Big(\{\hat{O}_a(t)\}_{a\in\mathcal{A}}, \{\hat{O}_f(t)\}_{f\in\mathcal{F}}\Big)$ and sets $f_a^{\texttt{apply}}(t)$ to be agent a's match in the resulting estimated stable matching
  CIA assigns the interview set $\mathcal{I}_a(t) = \{f_a^{\texttt{apply}}(t), f_a^{\texttt{RR}}(t)\}$ to each agent $a \in \mathcal{A}$
  Each agent interviews the firms in $\mathcal{I}_a(t)$ and observes the corresponding feedback
  Each agent $a \in \mathcal{A}$ observes $X_{a,f_a^{\texttt{apply}}(t)}(t)$
  Agents update $\hat{u}_{a,f_a^{\texttt{apply}}(t)}(t+1)$ and $\hat{u}_{a,f_a^{\texttt{RR}}(t)}(t+1)$
  Each firm $f \in \mathcal{F}$ observes $X_{f,a}(t)$ and $X_{a,f_a^{\texttt{RR}}(t)}(t)$ from interviewed agents $a \in \mathcal{I}_f(t)$
  Agents apply to $f_a^{\texttt{apply}}(t)$ and receive reward $X_{a,f_a^{\texttt{match}}(t)}(t)$
  Each firm admits its most preferred applicant $a_f^{\texttt{match}}(t)$

E.2 Proof of an Auxiliary Lemma

We first state and prove a simple auxiliary bound used in the regret analysis.

Lemma E.1. *Let $a > 0$, $c > 0$, and $\Delta > 0$ be constants, and define $R(m, \Delta) = \frac{a\, e^{-\frac{\Delta^2}{cm}}}{1-e^{-\frac{\Delta^2}{cm}}}$. Then*

$$R(m, \Delta) \in O\Big(\frac{m}{\Delta^2}\Big) \quad \textit{for } m > 0.$$

Proof. Let $x = \frac{\Delta^2}{cm}$. Then $x > 0$ and we can rewrite

$$R(m, \Delta) = \frac{a\, e^{-x}}{1 - e^{-x}} = \frac{a}{e^x - 1}.$$

Using the standard inequality valid for all $x > 0$, $e^x - 1 \geq x$, we obtain

$$R(m, \Delta) \leq \frac{a}{x} = \frac{acm}{\Delta^2}.$$

Thus, there exists a constant $K = ac$ such that

$$R(m, \Delta) \leq K \cdot \frac{m}{\Delta^2},$$

and hence $R(m, \Delta) \in O\big(\frac{m}{\Delta^2}\big)$. □

E.3 Proof of Lemma 3.3

We next present the proof of Lemma 3.3, which is a key ingredient in our time-independent regret analysis.

Lemma 3.3. *In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, for any agent $a \in \mathcal{A}$ with optimal stable match $\overline{f_a^*} \in \mathcal{F}$, Algorithm 6 guarantees $\mathbb{E}[|\bar{\mathcal{E}}_{a,\overline{f_a^*}}|] \in O(|\mathcal{L}_{a,\overline{f_a^*}}| \cdot m \cdot \Delta_a^{-2})$, where $\Delta_a \doteq \min_{f \neq \overline{f_a^*}} \overline{\Delta}_{a,f}$.*

Proof. By union bounding over firms and applying Hoeffding's inequality to estimation errors, we obtain

$$\begin{aligned}
\mathbb{E}[|\bar{\mathcal{E}}_{a,\overline{f_a^*}}|] &= \sum_{t=1}^{T} \Pr\Big(\exists f' \in \mathcal{L}_{a,\overline{f_a^*}} :\ f' \in \hat{\mathcal{H}}_{a,\overline{f_a^*}}(t)\Big), \\
&= \sum_{t=1}^{T} \Pr\Big(\exists f' \in \mathcal{L}_{a,\overline{f_a^*}} :\ \hat{u}_{a,f'}(t) > \hat{u}_{a,\overline{f_a^*}}(t)\Big), \\
&\leq \sum_{t=1}^{T} \sum_{f' \in \mathcal{L}_{a,\overline{f_a^*}}} \Pr\Big(\hat{u}_{a,f'}(t) - u_{a,f'} + u_{a,\overline{f_a^*}} - \hat{u}_{a,\overline{f_a^*}}(t) > \overline{\Delta}_{a,f'}\Big), \\
&\leq \sum_{t=1}^{T} \sum_{f' \in \mathcal{L}_{a,\overline{f_a^*}}} \Pr\left(\big|\hat{u}_{a,f'}(t) - u_{a,f'}\big| > \tfrac{\overline{\Delta}_{a,f'}}{2}\right) + \Pr\left(\big|u_{a,\overline{f_a^*}} - \hat{u}_{a,\overline{f_a^*}}(t)\big| > \tfrac{\overline{\Delta}_{a,f'}}{2}\right), \\
&\overset{(a)}{\leq} \sum_{t=1}^{T} \sum_{f' \in \mathcal{L}_{a,\overline{f_a^*}}} 4\exp\left(-\frac{t\overline{\Delta}_{a,f'}^2}{2m}\right), \\
&\leq \frac{4|\mathcal{L}_{a,\overline{f_a^*}}| \exp\Big(-\frac{(\overline{\Delta}_a)^2}{2m}\Big)}{1 - \exp\Big(-\frac{(\overline{\Delta}_a)^2}{2m}\Big)}, \\
&\leq 8m\,|\mathcal{L}_{a,\overline{f_a^*}}| \cdot \big(\overline{\Delta}_a\big)^{-2} \in O\Big(m\,|\mathcal{L}_{a,\overline{f_a^*}}| \cdot \big(\overline{\Delta}_a\big)^{-2}\Big),
\end{aligned} \tag{16}$$

where $\overline{\Delta}_a \doteq \min_{f \neq \overline{f_a^*}} \overline{\Delta}_{a,f}$ denotes the minimum optimal utility gap. Step (a) follows from Hoeffding's inequality. Step (b), implied by Lemma E.1, shows that the expected number of invalid time steps for agent a is bounded by a constant independent of T, completing the proof. □

E.4 Proof of Theorem 3.1

Theorem 3.1. *In market $\mathcal{M}(\mathcal{A},\mathcal{F})$ with non-strategic firms, agent a's optimal regret under Algorithm 6 is $\mathbb{E}[\overline{R}_a(\mathcal{T})] \in O(nm^2 \min\{\overline{\Delta}_{\mathcal{A}}, \underline{\Delta}_{\mathcal{F}}\}^{-2})$, with $\overline{\Delta}_{\mathcal{A}} \doteq \min_{a\in\mathcal{A}} \overline{\Delta}_a$ and $\underline{\Delta}_{\mathcal{F}} \doteq \min_{f\in\mathcal{F}} \underline{\Delta}_f$.*

Proof. We track invalidity on both sides. For each agent $a \in \mathcal{A}$, let $\bar{\mathcal{E}}_{a,\overline{f_a^*}}$ be the rounds in which $\hat{O}_a(t)$ is invalid w.r.t. $\overline{f_a^*}$, and for each firm $f \in \mathcal{F}$, let $\bar{\mathcal{E}}_{f,\underline{a_f^*}}$ be the rounds in which $\hat{O}_f(t)$ is invalid w.r.t. $\underline{a_f^*}$ (defined analogously). Define the globally invalid rounds by

$$\bar{\mathcal{E}} \doteq \Big(\bigcup_{a\in\mathcal{A}} \bar{\mathcal{E}}_{a,\overline{f_a^*}}\Big) \cup \Big(\bigcup_{f\in\mathcal{F}} \bar{\mathcal{E}}_{f,\underline{a_f^*}}\Big), \qquad \mathcal{E} \doteq [T] \setminus \bar{\mathcal{E}}.$$

Decompose the expected regret as

$$\mathbb{E}[\overline{R}_a(T)] = \sum_{t\in[T]} u_{a,\overline{f_a^*}} - \Big(\sum_{t\in\mathcal{E}} \mathbb{E}[X_{a,f_a^{\mathtt{match}}(t)}(t)] + \sum_{t\in\bar{\mathcal{E}}} \mathbb{E}[X_{a,f_a^{\mathtt{match}}(t)}(t)]\Big). \tag{17}$$

Valid rounds ($t \in \mathcal{E}$). Fix any $t \in \mathcal{E}$ and consider the agent-optimal stable matching under the ground-truth preferences, $(a,\overline{f_a^*})_{a\in\mathcal{A}}$. If this matching had a blocking pair under the estimated preferences at time t, then the same pair would also block under the ground-truth preferences: on $\mathcal{E}$, any agent (resp., firm) that is ranked above the relevant stable partner in the estimated list is also ranked above it in the ground truth. This contradicts stability of $(a,\overline{f_a^*})_{a\in\mathcal{A}}$ under O. Hence $(a,\overline{f_a^*})_{a\in\mathcal{A}}$ is also stable under $(\hat{O}_a(t),\hat{O}_f(t))$.

Since `CIA` runs `GS` on the estimated market at time t, it returns the agent-optimal stable matching w.r.t. $(\hat{O}_a(t),\hat{O}_f(t))$. Therefore, for every agent a, the resulting match $f_a^{\mathtt{match}}(t)$ is either $\overline{f_a^*}$ or a firm that a ranks strictly above $\overline{f_a^*}$ in $\hat{O}_a(t)$; by validity, this firm is also strictly preferred in the ground truth. Thus agents incur zero regret on all rounds $t \in \mathcal{E}$.

Invalid rounds ($t \in \bar{\mathcal{E}}$). By a union bound,

$$\mathbb{E}[|\bar{\mathcal{E}}|] \le \sum_{a\in\mathcal{A}} \mathbb{E}\Big[\Big|\bar{\mathcal{E}}_{a,\overline{f_a^*}}\Big|\Big] + \sum_{f\in\mathcal{F}} \mathbb{E}\Big[\Big|\bar{\mathcal{E}}_{f,\underline{a_f^*}}\Big|\Big]. \tag{18}$$

Lemma 3.3 gives, for every $a \in \mathcal{A}$, $\mathbb{E}\Big[\Big|\bar{\mathcal{E}}_{a,\overline{f_a^*}}\Big|\Big] \in O\Big(m|\mathcal{L}_{a,\overline{f_a^*}}| \cdot \overline{\Delta}_a^{-2}\Big)$, where $\overline{\Delta}_a \doteq \min_{f\neq\overline{f_a^*}} \overline{\Delta}_{a,f}$. Moreover, since each (strategic uncertain) firm participates in round-robin interviews, the same argument yields the firm-side analogue

$$\mathbb{E}\Big[\Big|\bar{\mathcal{E}}_{f,\underline{a_f^*}}\Big|\Big] \in O\Big(m|\mathcal{L}_{f,\underline{a_f^*}}| \cdot \underline{\Delta}_f^{-2}\Big), \tag{19}$$

where $\underline{\Delta}_f \doteq \min_{a\neq\underline{a_f^*}} \underline{\Delta}_{f,a}$. Using $|\mathcal{L}_{a,\overline{f_a^*}}| \le m$ and $|\mathcal{L}_{f,\underline{a_f^*}}| \le n$, we obtain $\mathbb{E}[|\bar{\mathcal{E}}|] \in O\left(nm^2 \min\left\{\overline{\Delta}_{\mathcal{A}}, \underline{\Delta}_{\mathcal{F}}\right\}^{-2}\right)$, where $\overline{\Delta}_{\mathcal{A}} \doteq \min_{a\in\mathcal{A}} \overline{\Delta}_a$ and $\underline{\Delta}_{\mathcal{F}} \doteq \min_{f\in\mathcal{F}} \underline{\Delta}_f$. Since regret is zero on $\mathcal{E}$ and per-round regret is at most 1, we have $\mathbb{E}[\overline{R}_a(T)] \le \mathbb{E}[|\bar{\mathcal{E}}|] \in O\left(nm^2 \min\left\{\overline{\Delta}_{\mathcal{A}}, \underline{\Delta}_{\mathcal{F}}\right\}^{-2}\right)$, completing the proof. □

F Deferred Concepts and Proofs from Section 4.2

This section collects extensive explanation of the algorithm followed by additional definitions and deferred proofs from Section 4.2.

F.1 Necessary Coordination under $\mathcal{V}(t)$ compared to $\mathcal{V}^+(t)$

Under the only-vacancy firm-side feedback $\mathcal{V}(t)$, agents do not observe firms' hiring updates—information that is available under $\mathcal{V}^+(t)$. Since $\mathcal{V}(t)$ only announces vacancies, agents must coordinate when adjusting their strategies. Thus, while $\mathcal{V}^+(t)$ allows independent adaptation, the limited feedback in $\mathcal{V}(t)$ requires coordinated distributed executions of `GS`. Example F.1 demonstrates why such coordination becomes essential under $\mathcal{V}(t)$ and clarifies the form of interaction our algorithm adopts.

Example F.1. *Consider a market with agents $\mathcal{A} = \{a_1, a_2, a_3\}$ and certain firms $\mathcal{F} = \{f_1, f_2, f_3\}$, whose true preferences are:*

$$O_{a_1} : \langle f_1^\dagger \succ f_2 \succ f_3 \rangle \quad O_{f_1} : \langle a_1^\dagger \succ a_2 \succ a_3 \rangle$$
$$O_{a_2} : \langle f_2^\dagger \succ f_1 \succ f_3 \rangle \quad O_{f_2} : \langle a_1 \succ a_2^\dagger \succ a_3 \rangle$$
$$O_{a_3} : \langle f_1 \succ f_3^\dagger \succ f_2 \rangle \quad O_{f_3} : \langle a_1 \succ a_2 \succ a_3^\dagger \rangle$$

The unique stable matching is $(a_1, f_1), (a_2, f_2), (a_3, f_3)$, as indicated by daggers. Assume agent a_2 has already learned its true preferences. At some time step t, the estimated preference lists are:

$$\hat{O}_{a_1}(t) : \langle f_2 \succ f_1 \succ f_3 \rangle \quad I_{a_1}(t) = \{f_2, f_{a_1}^{\texttt{RR}}(t)\}$$
$$\hat{O}_{a_2}(t) : \langle f_2 \succ f_1 \succ f_3 \rangle \quad I_{a_2}(t) = \{f_1, f_{a_2}^{\texttt{RR}}(t)\}$$
$$\hat{O}_{a_3}(t) : \langle f_3 \succ f_1 \succ f_2 \rangle \quad I_{a_3}(t) = \{f_3, f_{a_3}^{\texttt{RR}}(t)\}$$

Since a_1 is the most preferred by all firms, it applies to $f_{a_1}^{\texttt{apply}}(t) = f_2$, which causes a_2 to be rejected by f_2 when it considers a_2. Consequently, a_2 proceeds to consider f_1, and since a_2 is the second most preferred agent by all firms, the only remaining option for a_3 is to consider f_3.

Now suppose that at time step $t' > t$, the updated estimates become:

$$\hat{O}_{a_1}(t') : \langle f_3 \succ f_1 \succ f_2 \rangle \quad I_{a_1}(t') = \{f_3, f_{a_1}^{\texttt{RR}}(t')\}$$
$$\hat{O}_{a_2}(t') : \langle f_2 \succ f_1 \succ f_3 \rangle \quad I_{a_2}(t') = \{f_1, f_{a_2}^{\texttt{RR}}(t')\}$$
$$\hat{O}_{a_3}(t') : \langle f_2 \succ f_3 \succ f_1 \rangle \quad I_{a_3}(t') = \{f_1, f_{a_3}^{\texttt{RR}}(t')\}$$

Here, a_1 switches to applying to f_3, again causing other applicants to be rejected due to being the most preferred agent. Meanwhile, a_3, which was never rejected by f_2, now reconsiders f_2 as its most preferred estimated firm at that round. Since a_1 and a_3 simultaneously change their targets, and a_2 observes no hiring updates under $\mathcal{V}(t')$ (only vacancies), it may assume that f_2 remains unavailable and commit to the suboptimal firm f_1.

Under $\mathcal{V}^+(t')$, the hiring update at f_2 would be visible to a_2, prompting it to reconsider f_2 and avoid a suboptimal match. Hence, the limited feedback in $\mathcal{V}(t')$ prevents agents from detecting such preference updates, necessitating coordinated distributed runs of `GS`. If, for instance, f_1—upon detecting a change in its local estimates $\hat{O}_{f_1}(t')$—could signal others to initiate a joint distributed `GS`, then a_2 would reassess f_2, ensuring convergence to a stable matching.

Otherwise, a_2 must continually reconsider firms it was previously rejected from (with some positive probability), and in some scenarios this prevents achieving $O(1)$ regret.

F.2 Extended Coordinated Decentralized Algorithm with Vacancy-Only $\mathcal{V}(t)$ Feedback

We now design decentralized learning algorithms under the only-vacancy feedback $\mathcal{V}(t)$, which agents use as the coordination signal. Unlike $\mathcal{V}^+(t)$, which reveals hiring changes, $\mathcal{V}(t)$ only announces vacancies, so important market transitions can remain invisible to agents not directly involved (see Example F.1); this weaker feedback makes learning more challenging. Accordingly, under $\mathcal{V}(t)$ we rely on explicit coordination to ensure stability and avoid linear regret; a formal justification is deferred to Section F.1.

We next describe a learning framework in which agents alternate between independent behavior and deliberate signaling through their application decisions; these signals shape the vacancy feedback and trigger coordinated updates.

Learning Framework. The time horizon is divided into two alternating stages: *committing* and *updating*. In a committing stage, agents adhere to their current interviewing sets $\mathcal{I}_a(t)$ and act independently. In an updating stage, agents coordinate to run a distributed execution of `GS` on their current estimated preference lists in order to revise their interviewing decisions.

Each agent $a \in \mathcal{A}$ maintains a binary state $\rho_a(t)$, where $\rho_a(t) = 1$ indicates a committing stage and $\rho_a(t) = 0$ indicates an updating stage. This state is updated at the end of each round after observing $\mathcal{V}(t)$. Since $\mathcal{V}(t)$ is globally revealed, all agents update their states identically, synchronizing transitions between the two stages. Formally, we define the sets of updating and committing rounds as $\mathcal{U} \doteq \{\, t \in \mathcal{T} : \forall a \in \mathcal{A},\ \rho_a(t) = 0 \,\}$ and $\mathcal{C} \doteq \{\, t \in \mathcal{T} : \forall a \in \mathcal{A},\ \rho_a(t) = 1 \,\}$.

The central challenge is deciding when to switch stages. Transitions from $\rho_a(t) = 1$ (committing) to $\rho_a(t) = 0$ (updating) occur in two ways. First, an agent may detect a market change and intentionally refrain from applying, creating a vacancy that serves as a coordination signal via the firm-side feedback. Second, a strategic firm may abstain from hiring under Algorithm 1 and appear in $\mathcal{V}(t)$. Since agents do not observe the source of new vacancies, both events are treated as coordination signals. Conversely, switching back is deterministic: agents wait $O(n^2)$ rounds for the distributed GS run and return to committing. The procedure is detailed in Algorithm 7.

Algorithm 7 Coordinated Decentralized Learning (agent a)

Input: a, $\mathcal{F}$, $\mathcal{V}$, switching conditions $\mathcal{S}_a^1, \mathcal{S}_a^2, \mathcal{S}_a^3$
Initialize: $t \leftarrow 1$, $t^{\texttt{GS}} \leftarrow 1$, $r_{a,f}(1) \leftarrow 0$ for all $f \in \mathcal{F}$
while $t \leq T$ **do**
  Construct the candidate set $\mathcal{B}_a(t)$ according to (21)
  if $t \in [t^{\texttt{GS}}, t^{\texttt{GS}} + 3n^2)$ **then**
    $\rho_a(t) \leftarrow 0$ ▷ updating stage
    $f_a^{\texttt{apply}}(t) \leftarrow \arg\max_{f \in \mathcal{B}_a(t)} \hat{u}_{a,f}(t^{\texttt{GS}})$
    $\mathcal{I}_a(t) \leftarrow \{f_a^{\texttt{apply}}(t), f_a^{\texttt{RR}}(t)\}$ and interview
  else if $t \notin \mathcal{S}_a^{\text{inc}} \cup \mathcal{S}_a^{\text{rej}} \cup \mathcal{S}_a^{\text{vac}}$ **then**
    $\rho_a(t) \leftarrow 1$ ▷ committing stage
    $f_a^{\texttt{apply}}(t) \leftarrow \arg\max_{f \in \mathcal{B}_a(t)} \hat{u}_{a,f}(t)$
    $\mathcal{I}_a(t) \leftarrow \{f_a^{\texttt{apply}}(t), f_a^{\texttt{RR}}(t)\}$ and interview
  else
    $\rho_a(t+1) \leftarrow 0$ ▷ switching to updating stage
    $t^{\texttt{GS}} \leftarrow t + 1$
    $r_{a,f}(t+1) \leftarrow 0$ for all $f \in \mathcal{F}$ ▷ resetting
    $f_a^{\texttt{apply}}(t) \leftarrow \emptyset$ ▷ not applying to signal others
  Apply to $f_a^{\texttt{apply}}(t)$
  UPDATEAGENTREJVARS(a, t) ▷ Subroutine 5
  $t \leftarrow t + 1$

At each round t, agent a interviews $\mathcal{I}_a(t) = \{f_a^{\texttt{apply}}(t), f_a^{\texttt{RR}}(t)\}$, where $f_a^{\texttt{RR}}(t)$ is chosen via round-robin and $f_a^{\texttt{apply}}(t)$ is the firm to which a applies. We next describe how agents choose their applied firms and interview sets within each stage, and how transitions between committing and updating are triggered. Each agent a maintains a private rejection-time variable $r_{a,f}(t)$ for each firm f, recording the most recent round $t' < t$ in which a was *non-strategically rejected* by f:

$$r_{a,f}(t) \doteq \arg\max_{t'<t} \left\{ \gamma_f(t') = 1 \wedge \begin{matrix} f_a^{\texttt{apply}}(t') = f \\ f_a^{\texttt{match}}(t') = \emptyset \end{matrix} \right\}. \tag{20}$$

Thus, if a applies to $f_a^{\texttt{apply}}(t) = f$ and remains unmatched, it updates $r_{a,f}(t+1) = t$ whenever $f \notin \mathcal{V}(t)$. We also maintain a global variable $t^{\texttt{GS}}$ denoting the first round of the most recent updating stage, i.e., the start of the current GS execution. This value is common to all agents and is updated whenever they switch from committing to updating.

Interview Set During Updating. The goal of the updating phase is to execute a fresh coordinated run of GS using a fixed snapshot of the agents' current estimated preference lists. At the beginning of an updating stage, all agents reset their rejection variables by setting $r_{a,f}(t^{\texttt{GS}}) = 0$ for every firm f so that past rejections do not affect the new run of GS. Throughout the updating phase, agents coordinate to execute a distributed version of GS and make all interviewing decisions based on their estimated preference lists fixed at time $t^{\texttt{GS}}$, namely $\hat{u}_{a,f}(t^{\texttt{GS}})$. In contrast, firms continue to follow Algorithm 1 and make hiring decisions using their current estimates at each round.

To coordinate a new run of GS, agents enter an updating phase of fixed length $3n^2$ rounds starting at $t^{\texttt{GS}}$. During this phase, all agents follow a synchronized distributed execution of GS. Specifically, for all $t \in \{t^{\texttt{GS}}, t^{\texttt{GS}} + 1, \ldots, t^{\texttt{GS}} + 3n^2 - 1\}$, each agent a selects its applied firm from candidate set

$$\mathcal{B}_a(t) \doteq \{f \in \mathcal{F} : r_{a,f}(t) < t^{\texttt{GS}}\}, \tag{21}$$

which consists of all firms that have not rejected agent a since the beginning of the current updating phase. The agent then chooses the most preferred firm in $B'_a(t)$ according to its estimated preferences at time $t^{\texttt{GS}}$, formally as $f_a^{\texttt{apply}}(t) = \arg\max_{f \in \mathcal{B}_a(t)} \hat{u}_{a,f}(t^{\texttt{GS}})$.

If agent a is rejected because firm $f_a^{\texttt{apply}}(t)$ hires another agent, that is, if $a^{\texttt{match}}_{f_a^{\texttt{apply}}(t)}(t) \neq \emptyset$, the rejection variable is updated as $r_{a,f_a^{\texttt{apply}}(t)}(t) = t$. This procedure continues until round $t^{\texttt{GS}} + 3n^2 - 1$, after which all agents set $\rho_a(t^{\texttt{GS}} + 3n^2) = 1$ and transition back to the committing stage. By Lemma F.2, each updating phase ends with a matching $(a, f_a^{\texttt{apply}}(t^{\texttt{GS}} + 3n^2))_{\forall a \in \mathcal{A}}$ that matches all agents; hence, during the subsequent committing phase no two agents apply to the same firm.

Interview Set During Committing. After an updating phase, agents enter a committing stage, acting independently under their current estimates until a coordination trigger is detected. During committing, at any time step $t \geq t^{\texttt{GS}} + 3n^2 + 1$, each agent sets $f_a^{\texttt{apply}}(t) = \arg\max_{f \in \mathcal{B}_a(t)} \hat{u}_{a,f}(t)$, where the candidate set $\mathcal{B}_a(t)$ is fixed throughout the committing stage. As long as the maximizer does not change, agent a keeps applying to the same firm, even as estimates continue to update.

When a switching condition is detected at some $t \in \mathcal{C}$, agent a sets $f_a^{\texttt{apply}}(t) = \emptyset$; all agents then synchronously switch to an updating stage in the next round. For each agent a, define the committing rounds that trigger this transition as follows (the first two are detected locally, while the third is a global signal common to all agents):

(i) **Preference-list inconsistency.** Learning may change agent a's estimated top choice relative to the firm selected at the end of the most recent updating phase, suggesting that the last distributed GS run used an invalid list. Formally,

$$\mathcal{S}_a^{\text{inc}} \doteq \left\{ t \in \mathcal{C} : f_a^{\texttt{apply}}(t) \neq f_a^{\texttt{apply}}(t^{\texttt{GS}} + 3n^2) \right\}.$$

(ii) **Strategic rejection.** Agent a may be rejected due to a firm's strategic abstention after the beginning of the last updating phase; this is detected when the relevant firm appears in the vacancy feedback, invalidating the matching obtained at the end of the updating phase.

$$\mathcal{S}_a^{\text{rej}} \doteq \left\{ t \in \mathcal{C} \ : \ \exists\, t' \in [t^{\texttt{GS}}, t) \text{ s.t. } f_a^{\texttt{apply}}(t') \in \mathcal{V}(t') \right\}.$$

(iii) **New vacancy signal.** A matched firm becomes newly vacant, so vacancies exceed $m - n$. Formally,

$$\mathcal{S}_a^{\text{vac}} \doteq \{ t \in \mathcal{C} : |\mathcal{V}(t)| > m - n \}.$$

If $t \in \mathcal{S}_a^{\text{inc}} \cup \mathcal{S}_a^{\text{rej}} \cup \mathcal{S}_a^{\text{vac}}$, agent a sets $\rho_a(t+1) = 0$ and updates $t^{\texttt{GS}} = t + 1$. All agents then synchronously enter a new updating stage in the following round.

F.3 Committing to a Perfect Matching Lemma F.2

This lemma shows that after each updating phase, the agents' applications induce a perfect matching, which serves as the baseline structure for the subsequent regret analysis.

Lemma F.2. *In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$ where firms follow Algorithm 1 and agents follow Algorithm 7, at the end of any updating phase that starts at time $t^{\texttt{GS}}$, the applications $(a, f_a^{\texttt{apply}}(t^{\texttt{GS}} + 3n^2))_{a \in \mathcal{A}}$ form a matching that matches all agents.*

Proof. Since the beginning of the updating phase at time $t^{\texttt{GS}}$, each agent a selects its applied firm according to

$$f_a^{\texttt{apply}}(t) = \arg\max_{f \in \mathcal{B}_a(t)} \hat{u}_{a,f}(t^{\texttt{GS}}),$$

that is, with respect to its estimated preference list $\hat{O}_a(t^{\texttt{GS}})$. Consequently, during the updating phase, each agent effectively traverses $\hat{O}_a(t^{\texttt{GS}})$ in descending order, applying to increasingly less preferred firms until reaching one that does not reject it in favor of another agent. Once an agent a becomes matched to a firm $f = f_a^{\texttt{apply}}(t)$, it continues to apply to f in subsequent rounds until either the updating phase ends or f rejects a by hiring another agent. Therefore, whenever the applications $(a, f_a^{\texttt{apply}}(t))_{a \in \mathcal{A}}$ do not form a perfect matching that matches all agents, there exists at least one agent who has not been hired in that round.

In addition to rejections caused by competing applicants, agents may experience strategic rejections. However, by construction, the rejection variable $r_{a,f}$ is updated only when firm f hires another agent,

i.e., when $\gamma_f(t) = 1$, and not when f strategically abstains from hiring. As a result, following a strategic rejection, agents reapply to the same firm in the subsequent round. Moreover, under the rejection policy Algorithm 1, no firm performs strategic rejections in two consecutive rounds during an updating phase. Further, by the definition of the candidate set (21) under Algorithm 7, an agent does not apply to a firm after being rejected in favor of another hire. Hence, each firm can reject a given agent at most twice during the updating phase starting from $t^{\texttt{GS}}$: at most once strategically, and at most once in favor of hiring another agent.

It follows that as long as the applications do not form a perfect matching, within at most three rounds there exists at least one agent who moves to a less preferred firm. Indeed, in the worst case an agent is first strategically rejected at time t (and thus re-applies to the same firm at time $t+1$), and then rejected at time $t+1$ in favor of another hire (which updates $r_{a,f}$), in which case it can consider a new firm only at time $t+2$.

Since, by Observation 6, each agent's match during the updating phase is restricted to the top n firms in $\hat{O}_a(t^{\texttt{GS}})$, the applications form a perfect matching within at most $3n^2$ rounds. In particular, after at most $3n^2$ rounds from the start of an updating phase, the applications $(a, f_a^{\texttt{apply}}(t))_{a\in\mathcal{A}}$ form a perfect matching. The $O(n^2)$ bound follows since each agent applies to at most the top n firms (with respect to $\hat{O}_a(t^{\texttt{GS}})$) during a single updating phase. □

F.4 Counting Updating Stages via Triggered Events

This observation expresses the number of updating stages in terms of the total number of triggered inconsistency/rejection events across all agents, which is later used to bound the overall regret.

Observation 3. *The total number of updating stages is given by* $\mathbb{E}\big[\big|\bigcup_{a\in\mathcal{A}}\left(\mathcal{S}_a^{\mathrm{inc}} \cup \mathcal{S}_a^{\mathrm{rej}}\right)\big|\big]$.

Proof. This holds because whenever an agent $a \in \mathcal{A}$ detects an event in $\mathcal{S}_a^{\mathrm{inc}}$, it abstains from applying in that round, which is subsequently revealed to all other agents through their detection of events in $\mathcal{S}_{a'}^{\mathrm{vac}}$. Similarly, if a detects an event in $\mathcal{S}_a^{\mathrm{rej}}$, it indicates that at that time step the agent was strategically not hired, an event that is also reflected to others via $\mathcal{S}_{a'}^{\mathrm{vac}}$ in the firm-side feedback. Therefore, every initialization of a new distributed GS run is ultimately triggered by at least one agent a detecting the current time step t as belonging to either $\mathcal{S}_a^{\mathrm{inc}}$ or $\mathcal{S}_a^{\mathrm{rej}}$. □

F.5 Bounding the Number of Updating Stages in α-Reducible Markets

Lemma F.3. *In an α-reducible market, with strategic firms following Algorithm 1 and agents following Algorithm 7, the total number of updating stages is $O\Big(nm^2 \min\{\overline{\Delta}_{\mathcal{A}}, \underline{\Delta}_{\mathcal{F}}\}^{-2}\Big)$, where $\overline{\Delta}_{\mathcal{A}} \doteq \min_{a\in\mathcal{A}} \overline{\Delta}_a$ and $\underline{\Delta}_{\mathcal{F}} \doteq \min_{f\in\mathcal{F}} \underline{\Delta}_f$.*

Proof. Consider the unique stable matching $(a_i, f_i)_{i\in[n]}$ in an α-reducible market. For each $i \in [n]$, define the set of time steps

$$\mathcal{G}_i \doteq \{t \in \mathcal{C} : \forall j \in [i],\ f_{a_j}^{\texttt{apply}}(t) = f_j\},$$

namely, the rounds in which agents $a_1, \ldots, a_i$ all apply to their respective stable partners. Its complement $\bar{\mathcal{G}}_i$ consists of rounds in which at least one of these agents is not matched to its stable partner.

We now bound the expected number of switching events across all agents. In particular, with respect to Observation 3, we bound $\mathbb{E}\big[\big|\bigcup_{a_i\in\mathcal{A}} \mathcal{S}^{\mathrm{inc}}_{a_i} \cup \mathcal{S}^{\mathrm{rej}}_{a_i}\big|\big]$ as follows:

$$
\begin{aligned}
\mathbb{E}\left[\left|\bigcup_{a_i\in\mathcal{A}} \mathcal{S}^{\mathrm{inc}}_{a_i} \cup \mathcal{S}^{\mathrm{rej}}_{a_i}\right|\right] &= \mathbb{E}\left[\left|\bigcup_{a_i\in\mathcal{A}} \Big(\{t\in\mathcal{S}^{\mathrm{inc}}_{a_i}\cup\mathcal{S}^{\mathrm{rej}}_{a_i} : t\in\bar{\mathcal{G}}_i\} \cup \{t\in\mathcal{S}^{\mathrm{inc}}_{a_i}\cup\mathcal{S}^{\mathrm{rej}}_{a_i} : t\in\mathcal{G}_i\}\Big)\right|\right] \\
&\le \mathbb{E}\left[\left|\bigcup_{a_i\in\mathcal{A}} \Big(\bar{\mathcal{G}}_i \cup \{t\in\mathcal{S}^{\mathrm{inc}}_{a_i}\cup\mathcal{S}^{\mathrm{rej}}_{a_i} : t\in\mathcal{G}_i\}\Big)\right|\right] \\
&\overset{(a)}{\le} \mathbb{E}\left[\left|\bigcup_{a_i\in\mathcal{A}} (\bar{\mathcal{E}}_{a_i,f_i}\cup\bar{\mathcal{E}}_{f_i,a_i}) \cup \bigcup_{a_i\in\mathcal{A}} \Big(\{t\in\mathcal{S}^{\mathrm{inc}}_{a_i}\cup\mathcal{S}^{\mathrm{rej}}_{a_i} : t\in\mathcal{G}_i\}\Big)\right|\right] \\
&\overset{(b)}{\le} \mathbb{E}\left[\left|\bigcup_{a_i\in\mathcal{A}} (\bar{\mathcal{E}}_{a_i,f_i}\cup\bar{\mathcal{E}}_{f_i,a_i}) \cup \bigcup_{a_i\in\mathcal{A}} \Big(\{t\in\mathcal{S}^{\mathrm{inc}}_{a_i}\cup\mathcal{S}^{\mathrm{rej}}_{a_i} : (t\in\mathcal{G}_i)\wedge(t,t-1\in\mathcal{E}_{a_i,f_i}\cap\mathcal{E}_{f_i,a_i})\}\Big)\right|\right] \\
&\overset{(c)}{\le} \sum_{i\in[n]} \left(\mathbb{E}\big[\!\big[\bar{\mathcal{E}}_{a_i,f_i}\big]\!\big] + \mathbb{E}\big[\!\big[\bar{\mathcal{E}}_{f_i,a_i}\big]\!\big] + \sum_{f\in\mathcal{H}_{a_i,f_i}} \mathbb{E}\Big[\!\Big[\{t : f\in\hat{\mathcal{L}}_{a_i,f_i}(t-1)\cap\hat{\mathcal{H}}_{a_i,f_i}(t)\}\Big]\!\Big]\right) \\
&\overset{(d)}{\le} \sum_{a_i\in\mathcal{A}} \frac{4m\exp\Big(-\frac{\overline{\Delta}^2_{a_i}}{2m}\Big)}{1-\exp\Big(-\frac{\overline{\Delta}^2_{a_i}}{2m}\Big)} + \sum_{f_i\in\mathcal{F}} \frac{4n\exp\Big(-\frac{\underline{\Delta}^2_{f_i}}{2m}\Big)}{1-\exp\Big(-\frac{\underline{\Delta}^2_{f_i}}{2m}\Big)} \\
&\overset{(e)}{\le} \sum_{a_i\in\mathcal{A}} O\Big(m^2\overline{\Delta}^{-2}_{a_i}\Big) + \sum_{f_i\in\mathcal{F}} O\big(nm\underline{\Delta}^{-2}_{f_i}\big) \in O\Big(nm^2\min\big\{\overline{\Delta}_{\mathcal{A}},\underline{\Delta}_{\mathcal{F}}\big\}^{-2}\Big), \qquad (22)
\end{aligned}
$$

Here, inequality (a) follows from the fact that if $t \in \bar{\mathcal{G}}_i$, then either some agent a_j has an invalid estimated preference list at time $t^{\mathtt{GS}}$, or there exists a time step $t' \in [t^{\mathtt{GS}}, t^{\mathtt{GS}} + 3n^2]$ at which the firm f_j, for some $j \in [i]$, has an invalid estimated preference list. Otherwise, if the estimated preference lists of agents $a_1, \ldots, a_i$ are valid at time $t^{\mathtt{GS}}$ and the estimated preference lists of firms $f_1, \ldots, f_i$ remain valid throughout $[t^{\mathtt{GS}}, t^{\mathtt{GS}} + 3n^2]$, then by the definition of α-reducibility (and its inductive structure), these agents must be matched to their unique stable partners, which implies $t \in \mathcal{G}_i$.

The justification of inequality (b) follows by considering the possible cases when $t - 1 \in \mathcal{G}_i$.

(i) Either $\hat{O}_{a_i}(t-1)$ or $\hat{O}_{f_i}(t-1)$ is invalid. In this case, $t - 1 \in \bar{\mathcal{E}}_{a_i,f_i} \cup \bar{\mathcal{E}}_{f_i,a_i}$, and the subsequent execution of $\mathtt{GS}$ can be charged to round $t - 1$ as an instance of an invalid preference list.

(ii) Both $\hat{O}_{a_i}(t-1)$ and $\hat{O}_{f_i}(t-1)$ are valid, but at least one becomes invalid at time t. If $\hat{O}_{a_i}(t)$ becomes invalid, then $t \in \mathcal{S}^{\mathrm{inc}}_{a_i}$, corresponding to a switching event by the agent. If instead $\hat{O}_{f_i}(t)$ becomes invalid, firm f_i may perform a strategic rejection, yielding $t \in \mathcal{S}^{\mathrm{rej}}_{a_i}$. In either case, the ensuing updating phase, and hence the next execution of $\mathtt{GS}$, is associated with $t \in \bar{\mathcal{E}}_{a_i,f_i} \cup \bar{\mathcal{E}}_{f_i,a_i}$.

(iii) Both $\hat{O}_{a_i}(t-1)$ and $\hat{O}_{a_i}(t)$, as well as $\hat{O}_{f_i}(t-1)$ and $\hat{O}_{f_i}(t)$, are valid. Suppose first that firm f_i performs a strategic rejection at time t, i.e., $t \in \mathcal{S}_{f_i}$. This action is observed by agent a_i through $\mathcal{V}^+(t)$, and therefore implies $t \in \mathcal{S}^{\mathrm{rej}}_{a_i}$.
Since all agents a_j for $j \in [i-1]$ are matched to their stable partners f_j, and since $t \in \mathcal{E}_{f_i,a_i}$, we claim that there must exist some agent a_j whose estimated preference list was invalid at some time during the previous updating phase $[t^{\mathtt{GS}}, t^{\mathtt{GS}} + 3n^2]$. Indeed, the strategic rejection by f_i can only be triggered if there exists an agent a_j that f_i truly prefers (i.e., $a_j \in \mathcal{H}_{f_i,a_i}$) but that was previously rejected by f_i during the updating phase. Concretely, during the updating phase, agent a_j applied to f_i and was rejected at some time t', which means that at the time of rejection $a_j \in \hat{\mathcal{L}}_{f_i,a_i}$. After this rejection, a_j subsequently applied to (and matched with) some firm f_j that it estimated to be less preferred. Given that f_i later performs a strategic rejection, this sequence of events implies that a_j's estimated preference list must have been invalid at time t'.
Otherwise, if no strategic rejection occurs at time t and $t \in \mathcal{S}^{\mathrm{inc}}_{a_i}$, then since $t-1, t \in \mathcal{E}_{a_i,f_i} \cap \mathcal{E}_{f_i,a_i}$, a switching event of type $\mathcal{S}^{\mathrm{inc}}_{a_i}$ implies that there exists a firm $f \in \mathcal{H}_{a_i,f_i}$ such that $f \in \hat{\mathcal{L}}_{a_i,f_i}(t^{\mathtt{GS}})$ but $f \in \hat{\mathcal{H}}_{a_i,f_i}(t)$.

These cases together establish inequality (c). Inequality (d) then follows with the same approach as used in the proof of Lemma 3.3. □

F.6 Proof of Theorem 4.1 for α-reducible Markets

Theorem 4.1. *In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, with firms and agents following Algorithm 1 and Algorithm 7, the expected regret is time-independent. In particular,*

$$\mathbb{E}[\underline{R}_a(\mathcal{T})] \in \begin{cases} O(n^3 m^2 \min\{\overline{\Delta}_{\mathcal{A}}, \underline{\Delta}_{\mathcal{F}}\}^{-2} + m\sum_{j=1}^{i}(|\mathcal{L}_{a_j,f_j}|\,\overline{\Delta}_{a_j}^{-2} + |\mathcal{L}_{f_j,a_j}|\,\underline{\Delta}_{f_j}^{-2})), & \alpha\text{-reducible}, \\ O(n^4 m^2 \Delta^{-2}), & \text{General Markets}. \end{cases}$$

Proof. We decompose the regret of agent a_i into two components: (i) regret incurred during updating phases, denoted by $\mathcal{U}$, and (ii) regret incurred during committing phases, denoted by $\mathcal{C}$. Formally, the expected regret satisfies

$$\begin{aligned} \mathbb{E}[R_{a_i}(T)] &= \mathbb{E}[R_{a_i}(\mathcal{U})] + \mathbb{E}[R_{a_i}(\mathcal{C})] \\ &= \underbrace{\sum_{t\in\mathcal{U}} \left(u_{a_i,f_i} - \mathbb{E}[X_{a_i,f_{a_i}^{\mathtt{match}}(t)}]\right)}_{\text{updating regret}} + \underbrace{\sum_{t\in\mathcal{C}} \left(u_{a_i,f_i} - \mathbb{E}[X_{a_i,f_{a_i}^{\mathtt{match}}(t)}]\right)}_{\text{committing regret}}. \end{aligned} \tag{23}$$

We bound these two terms separately.

We begin with the cumulative regret incurred during the updating phases. By construction, each updating phase lasts at most $3n^2$ rounds, and in each such round an agent incurs regret at most one. Moreover, by Observation 3 and Lemma F.3, the expected number of updating phases is bounded by the number of switching events across all agents. Consequently,

$$\sum_{t\in\mathcal{U}} \left(u_{a_i,f_i} - \mathbb{E}[X_{a_i,f_{a_i}^{\mathtt{match}}(t)}]\right) \le 3n^2 \cdot \mathbb{E}\left[\left|\bigcup_{a\in\mathcal{A}} \left(\mathcal{S}_a^{\mathrm{inc}} \cup \mathcal{S}_a^{\mathrm{rej}}\right)\right|\right]. \tag{24}$$

We now focus on the second term, corresponding to regret incurred during committing phases. Let

$$\bigcup_{a\in\mathcal{A}} \left(\mathcal{S}_a^{\mathrm{inc}} \cup \mathcal{S}_a^{\mathrm{rej}}\right) = \{t_1, t_2, \ldots\}$$

denote the set of switching times, ordered increasingly. The j-th committing phase is defined as the interval

$$[t'_j : t_{j+1}], \qquad \text{where } t'_j \doteq t_j + 3n^2 + 1,$$

corresponding to the period between the j-th and $(j+1)$-th distributed executions of `GS`. The collection of committing intervals is therefore

$$\mathcal{K} \doteq \left\{ [t'_j,\, t_{j+1}] : j \in \left[\left|\bigcup_{a\in\mathcal{A}} (\mathcal{S}_a^{\mathrm{inc}} \cup \mathcal{S}_a^{\mathrm{rej}})\right|\right] \right\}.$$

The expected regret incurred during committing phases can then be written as

$$
\begin{aligned}
\mathbb{E}[R_{a_i}(\mathcal{C})] &= \sum_{t\in\mathcal{C}} \Big(u_{a_i,f_i} - \mathbb{E}[X_{a_i,f_{a_i}^{\mathtt{match}}(t)}]\Big) \\
&= \mathbb{E}\left[\sum_{j=1}^{|\mathcal{K}|} \sum_{t\in[t'_j,\, t_{j+1}]} \Big(u_{a_i,f_i} - X_{a_i,f_{a_i}^{\mathtt{match}}(t'_j)}(t)\Big)\right], \\
&\overset{(a)}{\leq} \mathbb{E}\left[\left|\bigcup_{a\in\mathcal{A}} \left(\mathcal{S}_a^{\mathrm{inc}} \cup \mathcal{S}_a^{\mathrm{rej}}\right)\right|\right] + \mathbb{E}\left[\sum_{j\in[|\mathcal{K}|]} \left(t_{j+1} - 1 - t'_j\right) \,\middle|\, f_{a_i}^{\mathtt{match}}(t'_j) \in \mathcal{L}_{a_i,f_i}\right], \\
&\leq \mathbb{E}\left[\left|\bigcup_{a\in\mathcal{A}} \left(\mathcal{S}_a^{\mathrm{inc}} \cup \mathcal{S}_a^{\mathrm{rej}}\right)\right|\right] + \mathbb{E}\left[\sum_{j\in[|\mathcal{K}|]} \left(t_{j+1} - 1 - t'_j\right) \,\middle|\, f_{a_i}^{\mathtt{match}}(t'_j) \in \mathcal{L}_{a_i,f_i} \wedge t'_j \in \mathcal{G}_{i-1}\right] \\
&\quad + \mathbb{E}\left[\sum_{j\in[|\mathcal{K}|]} \left(t_{j+1} - 1 - t'_j\right) \,\middle|\, t'_j \in \bar{\mathcal{G}}_{i-1}\right], \\
&\overset{(b)}{\leq} \mathbb{E}\left[\left|\bigcup_{a\in\mathcal{A}} \left(\mathcal{S}_a^{\mathrm{inc}} \cup \mathcal{S}_a^{\mathrm{rej}}\right)\right|\right] + \mathbb{E}\left[\left|\bar{\mathcal{E}}_{a_i,f_i}\right|\right] + \mathbb{E}\left[\left|\bar{\mathcal{E}}_{f_i,a_i}\right|\right] + \mathbb{E}\left[\sum_{j\in[|\mathcal{K}|]} \left(t_{j+1} - 1 - t'_j\right) \,\middle|\, t'_j \in \bar{\mathcal{G}}_{i-1}\right], \\
&\overset{(c)}{\leq} \mathbb{E}\left[\left|\bigcup_{a\in\mathcal{A}} \left(\mathcal{S}_a^{\mathrm{inc}} \cup \mathcal{S}_a^{\mathrm{rej}}\right)\right|\right] + \sum_{j=1}^{i} \mathbb{E}\left[\left|\bar{\mathcal{E}}_{a_j,f_j}\right|\right] + \mathbb{E}\left[\left|\bar{\mathcal{E}}_{f_j,a_j}\right|\right], \qquad (25)
\end{aligned}
$$

Inequality (a) follows by accounting for the single round at the end of each committing phase, t_{j+1}, in which an agent intentionally refrains from applying in order to signal coordination to other agents. Such a signaling round occurs at most once per committing phase, immediately before the subsequent updating phase. By Observation 3, the expected number of these signaling rounds is bounded by $\mathbb{E}\left[\left|\bigcup_{a\in\mathcal{A}} \left(\mathcal{S}_a^{\mathrm{inc}} \cup \mathcal{S}_a^{\mathrm{rej}}\right)\right|\right]$. This term also includes the cases when a_i was strategically rejected during the last updating stage.

Inequality (b) follows from the structure of α-reducible markets. Condition on $t'_j \in \mathcal{G}_{i-1}$, so that the first $i-1$ agents are matched to their optimal firms, while agent a_i is matched suboptimally to some firm in $\mathcal{L}_{a_i,f_i}$. By α-reducibility, one of the following must hold.

(i) Either agent a_i has applied to its optimal firm f_i and was rejected during the last updating stage. In this case, firm f_i must be matched to a suboptimal agent due to an invalid estimated preference list during the preceding updating phase. This firm-side invalidity can persist during the committing phase, but at most until f_i's preference list becomes valid, at which point it strategically rejects its current match since a_i, which was previously rejected, now becomes preferred to its current applicant.

(ii) Alternatively, agent a_i has not applied to f_i while updating and is matched to a less preferred firm. This implies that a_i's estimated preference list was already invalid at the start of the updating phase, and this agent-side invalidity may persist until it is corrected.

In both cases, the committing regret incurred while $t'_j \in \mathcal{G}_{i-1}$ can be uniquely charged to an invalidity event in either the agent's or the firm's estimated preference list. Consequently, the total committing regret of agent a_i over all such disjoint intervals is bounded by $\mathbb{E}\left[\left|\bar{\mathcal{E}}_{a_i,f_i}\right|\right] + \mathbb{E}\left[\left|\bar{\mathcal{E}}_{f_i,a_i}\right|\right]$.

Moreover, to establish inequality (c), we bound

$$\begin{aligned}
\mathbb{E}\left[\sum_{j\in[|\mathcal{K}|]}\left(t_{j+1}-1-t'_j\right)\,\middle|\, t'_j\in\bar{\mathcal{G}}_{i-1}\right] &\le \mathbb{E}\left[\sum_{j\in[|\mathcal{K}|]}\left(t_{j+1}-1-t'_j\right)\,\middle|\, f^{\texttt{match}}_{a_{i-1}}(t'_j)\neq f_{i-1}\ \wedge\ t'_j\in\mathcal{G}_{i-2}\right]\\
&\quad+\mathbb{E}\left[\sum_{j\in[|\mathcal{K}|]}\left(t_{j+1}-1-t'_j\right)\,\middle|\, t'_j\in\bar{\mathcal{G}}_{i-2}\right]\\
&\le \mathbb{E}\left[\left|\bar{\mathcal{E}}_{a_{i-1},f_{i-1}}\right|\right]+\mathbb{E}\left[\left|\bar{\mathcal{E}}_{f_{i-1},a_{i-1}}\right|\right]+\mathbb{E}\left[\sum_{j\in[|\mathcal{K}|]}\left(t_{j+1}-1-t'_j\right)\,\middle|\, t'_j\in\bar{\mathcal{G}}_{i-2}\right]\\
&\ \ \vdots\\
&\le \sum_{j=1}^{i-1}\mathbb{E}\left[\left|\bar{\mathcal{E}}_{a_j,f_j}\right|\right]+\mathbb{E}\left[\left|\bar{\mathcal{E}}_{f_j,a_j}\right|\right],
\end{aligned}$$

where each step follows with the same rationale used to obtain inequality (b).

Finally, applying Lemma 3.3 to (25) and using (24) and Lemma F.3, the total regret is bounded as

$$\begin{aligned}
\mathbb{E}[R_{a_i}(T)] &\le \left(3n^2+1\right)\cdot O\left(n^3m^2\min\left\{\overline{\Delta}_{\mathcal{A}},\underline{\Delta}_{\mathcal{F}}\right\}^{-2}\right)+\left(\sum_{j=1}^{i}\frac{4\left|\mathcal{L}_{a_j,f_j}\right|\exp\left(-\frac{\overline{\Delta}^2_{a_j}}{2m}\right)}{1-\exp\left(-\frac{\overline{\Delta}^2_{a_j}}{2m}\right)}+\frac{4\left|\mathcal{L}_{f_j,a_j}\right|\exp\left(-\frac{\underline{\Delta}^2_{f_j}}{2m}\right)}{1-\exp\left(-\frac{\underline{\Delta}^2_{f_j}}{2m}\right)}\right)\\
&\in O\left(n^3m^2\min\left\{\overline{\Delta}_{\mathcal{A}},\underline{\Delta}_{\mathcal{F}}\right\}^{-2}+m\cdot\left(\sum_{j=1}^{i}\left|\mathcal{L}_{a_j,f_j}\right|\overline{\Delta}^{-2}_{a_j}+\left|\mathcal{L}_{f_j,a_j}\right|\underline{\Delta}^{-2}_{f_j}\right)\right)
\end{aligned}$$

□

F.7 Extension of Algorithm 7's Analysis to General Markets

We extend the results established for α-reducible markets in the main body to general, unstructured matching markets, where no structural assumptions are imposed and multiple stable matchings may exist.

F.7.1 An Upper Bound on the Expected Number of Updating Stages

We first extend Lemma F.3 to general matching markets in order to bound the expected number of updating stages. In the absence of structural assumptions, this bound becomes market-independent and necessarily larger than in the α-reducible case.

Lemma F.4. *In a matching market $\mathcal{M}(\mathcal{A},\mathcal{F})$, with strategic firms following Algorithm 1 and agents following Algorithm 7, the total number of distributed runs of `GS` (i.e., the number of updating phases) is time-independent and of $O(n^2m^2)$.*

Proof. In a general matching market $\mathcal{M}(\mathcal{A},\mathcal{F})$, as introduced earlier, we denote the agent-optimal and agent-pessimal stable matches of agent $a\in\mathcal{A}$ by $\overline{f^*_a}$ and $\underline{f^*_a}$, respectively. By the lattice structure of stable matchings, the agent-optimal stable matching coincides with the firm-pessimal one: in particular, if $\overline{f^*_a}=f$, then $\underline{a^*_f}=a$, and if $\underline{f^*_a}=f$, then $\overline{a^*_f}=a$.

Before proceeding, we introduce the event $\psi_{\mathcal{F}}\doteq\psi^{(1)}_{\mathcal{F}}\cap\psi^{(2)}_{\mathcal{F}}$. We first define

$$\psi^{(1)}_{\mathcal{F}}\doteq\left\{t\in[T]:\forall f\in\mathcal{F},\ t\in\bigcap_{a\in\mathcal{L}_{f,\overline{a^*_f}}}\mathcal{E}_{f,a}\right\},$$

$$\psi^{(2)}_{\mathcal{F}}\doteq\left\{t\in[T]:\forall f\in\mathcal{F},\ \left(t\in\mathcal{E}_{f,\overline{a^*_f}}\ \wedge\ \forall a\in\mathcal{H}_{f,\overline{a^*_f}},\ a\in\hat{\mathcal{H}}_{f,\overline{a^*_f}}(t)\right)\right\}.$$

Accordingly, $\psi_{\mathcal{F}}$ is the set of time steps at which, for every firm $f \in \mathcal{F}$, the estimated preference list of f correctly orders all agents that are strictly less preferred than its agent-optimal match $\overline{a_f^*}$, and every agent strictly preferred to $\overline{a_f^*}$ is ranked above it (possibly without being ordered among themselves).

We also define the event $\zeta_{\mathcal{A}}$ as

$$\zeta_{\mathcal{A}} \doteq \left\{ t \in [T] : \forall a \in \mathcal{A},\ t \in \bigcap_{f \in \{\underline{f_a^*}\} \cup \mathcal{H}_{a,\underline{f_a^*}}} \mathcal{E}_{a,f} \right\}.$$

That is, $\zeta_{\mathcal{A}}$ is the set of time steps at which, for every agent a, the estimated preference list correctly orders its agent-pessimal stable match $\underline{f_a^*}$ and all firms strictly preferred to it, according to the ground-truth preferences.

We also define $t^{\mathtt{GS}}(t)$ as the initial time step of the most recent updating stage by time t.

We now follow the proof idea in Observation 3 that the expected number of updating phases is bounded by $\mathbb{E}\left[\left|\bigcup_{a_i \in \mathcal{A}} \mathcal{S}_{a_i}^{\mathrm{inc}} \cup \mathcal{S}_{a_i}^{\mathrm{rej}}\right|\right]$. Thus we write

$$\begin{aligned}
\mathbb{E}\left[\left|\bigcup_{a_i \in \mathcal{A}} \mathcal{S}_{a_i}^{\mathrm{inc}} \cup \mathcal{S}_{a_i}^{\mathrm{rej}}\right|\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{ t \in \bigcup_{a_i \in \mathcal{A}} \mathcal{S}_{a_i}^{\mathrm{inc}} \cup \mathcal{S}_{a_i}^{\mathrm{rej}} \right\}\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{ t \in \bigcup_{a_i \in \mathcal{A}} \mathcal{S}_{a_i}^{\mathrm{inc}} \cup \mathcal{S}_{a_i}^{\mathrm{rej}} \wedge \exists t' \in [t^{\mathtt{GS}}(t), t-1], t' \in \bar{\psi}_{\mathcal{F}} \cup \bar{\zeta}_{\mathcal{A}} \right\}\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{ t \in \bigcup_{a_i \in \mathcal{A}} \mathcal{S}_{a_i}^{\mathrm{inc}} \cup \mathcal{S}_{a_i}^{\mathrm{rej}} \wedge \forall t' \in [t^{\mathtt{GS}}(t), t-1], t' \in \psi_{\mathcal{F}} \cap \zeta_{\mathcal{A}} \right\}\right] \\
&\overset{(a)}{\le} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{ \exists t' \in [t^{\mathtt{GS}}(t), t-1], t' \in \bar{\psi}_{\mathcal{F}} \cup \bar{\zeta}_{\mathcal{A}} \right\}\right] \\
&\qquad + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{ t \in \bigcup_{a_i \in \mathcal{A}} \mathcal{S}_{a_i}^{\mathrm{inc}} \cup \mathcal{S}_{a_i}^{\mathrm{rej}} \wedge t \in \bar{\psi}_{\mathcal{F}} \cup \bar{\zeta}_{\mathcal{A}} \right\}\right] \\
&\overset{(b)}{\le} \mathbb{E}\left[\left|\bar{\psi}_{\mathcal{F}} \cup \bar{\zeta}_{\mathcal{A}}\right|\right].
\end{aligned}$$

Inequality (a) follows by a case analysis.

(i) If there exists a time step $t' \in [t^{\mathtt{GS}}(t), t-1]$ such that $t' \in \bar{\psi}_{\mathcal{F}} \cup \bar{\zeta}_{\mathcal{A}}$, then the updating phase initiated at $t^{\mathtt{GS}}(t)$ is charged to this violation.

(ii) Otherwise, for all $t' \in [t^{\mathtt{GS}}(t), t-1]$ we have $t' \in \psi_{\mathcal{F}} \cap \zeta_{\mathcal{A}}$. In this case, throughout the updating phase starting at $t^{\mathtt{GS}}(t)$, all firms have estimated preference lists that are valid with respect to their agent-optimal matches $\overline{a_f^*}$ in the sense of $\psi_{\mathcal{F}}$, and all agents have estimated preference lists that are valid with respect to their agent-pessimal matches $\underline{f_a^*}$ in the sense of $\zeta_{\mathcal{A}}$. Consequently, the matching $(a, \underline{f_a^*})_{\forall a \in \mathcal{A}}$ remains stable with respect to the estimated preferences at all time steps in $[t^{\mathtt{GS}}(t), t-1]$, and no switching condition can be triggered.

Moreover, since the distributed GS run by the agents returns the agent-optimal (and firm-pessimal) stable matching under the current estimated preferences, it follows that for every agent $a \in \mathcal{A}$, $f_a^{\mathtt{apply}}(t)$ is ranked no worse than $\underline{f_a^*}$ under $\hat{O}_a(t)$, and for every firm $f \in \mathcal{F}$, the agent matched to f is ranked no better than $\overline{a_f^*}$ under $\hat{O}_f(t)$.

Therefore, any switching event at time t must correspond to a violation of either $\zeta_{\mathcal{A}}$ or $\psi_{\mathcal{F}}$:

- If $t \in \mathcal{S}_a^{\mathrm{inc}}$ for some agent a, then a detects a preference invalidity with respect to some $f \in \mathcal{H}_{a,\underline{f_a^*}}$, implying that $t \in \bar{\zeta}_{\mathcal{A}}$.
- If $t \in \mathcal{S}_a^{\mathrm{rej}}$ for some agent a, then the firm $f = f_a^{\mathtt{apply}}(t)$ performs a strategic rejection because its estimated preference list becomes invalid with respect to some $a' \in \mathcal{L}_{f,\overline{a_f^*}}$, implying that $t \in \bar{\psi}_{\mathcal{F}}$.

Inequality (b) then follows since each updating phase can be uniquely associated with a time step in $\bar{\psi}_{\mathcal{F}} \cup \bar{\zeta}_{\mathcal{A}}$, and such violations are counted at most once.

We complete the proof by bounding the two terms in $\mathbb{E}\big[\big|\bar{\psi}_{\mathcal{F}} \cup \bar{\zeta}_{\mathcal{A}}\big|\big] \;\; \leq \;\; \mathbb{E}\big[\big|\bar{\zeta}_{\mathcal{A}}\big|\big] + \mathbb{E}\big[\big|\bar{\psi}_{\mathcal{F}}\big|\big]$ separately.

To bound $\mathbb{E}\big[\big|\bar{\zeta}_{\mathcal{A}}\big|\big]$, we write

$$\begin{aligned}
\mathbb{E}\big[\big|\bar{\zeta}_{\mathcal{A}}\big|\big] &= \mathbb{E}\Big[\Big|\Big\{t \in [T] : \exists a \in \mathcal{A},\ \exists f \in \{\underline{f_a^*}\} \cup \mathcal{H}_{a,\underline{f_a^*}},\ t \in \bar{\mathcal{E}}_{a,f}\Big\}\Big|\Big], \\
&\leq \sum_{a \in \mathcal{A}} \sum_{f \in \{\underline{f_a^*}\} \cup \mathcal{H}_{a,\underline{f_a^*}}} \mathbb{E}\big[\big|\bar{\mathcal{E}}_{a,f}\big|\big], \\
&\leq \sum_{a \in \mathcal{A}} \sum_{f \in \{\underline{f_a^*}\} \cup \mathcal{H}_{a,\underline{f_a^*}}} \frac{4|\mathcal{L}_{a,f}| \exp\Big(-\frac{(\Delta^{\min}_{a,f})^2}{2m}\Big)}{1 - \exp\Big(-\frac{(\Delta^{\min}_{a,f})^2}{2m}\Big)}, \\
&\in \sum_{a \in \mathcal{A}} \sum_{f \in \{\underline{f_a^*}\} \cup \mathcal{H}_{a,\underline{f_a^*}}} O(m \cdot |\mathcal{L}_{a,f}|) \;\in\; O(n^2 m^2),
\end{aligned} \tag{26}$$

where the final bound follows from Lemma 3.3.

For $\mathbb{E}\big[\big|\bar{\psi}_{\mathcal{F}}\big|\big]$, we then write

$$\begin{aligned}
\mathbb{E}\big[\big|\bar{\psi}_{\mathcal{F}}\big|\big] &= \mathbb{E}\Big[\Big|\bar{\psi}_{\mathcal{F}}^{(1)} \cup \bar{\psi}_{\mathcal{F}}^{(2)}\Big|\Big], \\
&\leq \mathbb{E}\Big[\Big|\bar{\psi}_{\mathcal{F}}^{(1)}\Big|\Big] + \mathbb{E}\Big[\Big|\bar{\psi}_{\mathcal{F}}^{(2)}\Big|\Big], \\
&\leq \mathbb{E}\Big[\Big|\bar{\psi}_{\mathcal{F}}^{(1)}\Big|\Big] + \mathbb{E}\Big[\Big|\Big\{t \in [T] : \exists f \in \mathcal{F},\ t \in \bar{\mathcal{E}}_{f,\overline{a_f^*}} \;\vee\; \Big(\exists a \in \mathcal{H}_{f,\overline{a_f^*}},\ a \in \hat{\mathcal{L}}_{f,\overline{a_f^*}}(t)\Big)\Big\}\Big|\Big], \\
&\leq \mathbb{E}\Big[\Big|\bar{\psi}_{\mathcal{F}}^{(1)}\Big|\Big] + \sum_{f \in \mathcal{F}} \mathbb{E}\Big[\Big|\bar{\mathcal{E}}_{f,\overline{a_f^*}}\Big|\Big] + \mathbb{E}\Big[\Big|\Big\{t \in [T] : \exists a \in \mathcal{H}_{f,\overline{a_f^*}},\ a \in \hat{\mathcal{L}}_{f,\overline{a_f^*}}(t)\Big\}\Big|\Big], \\
&\leq \sum_{f \in \mathcal{F}} \sum_{a \in \mathcal{L}_{f,\overline{a_f^*}}} \frac{4|\mathcal{L}_{f,a}| \exp\Big(-\frac{(\Delta^{\min}_{f,a})^2}{2m}\Big)}{1 - \exp\Big(-\frac{(\Delta^{\min}_{f,a})^2}{2m}\Big)} + \frac{4n \exp\Big(-\frac{(\overline{\Delta}^{\min}_{f})^2}{2m}\Big)}{1 - \exp\Big(-\frac{(\overline{\Delta}^{\min}_{f})^2}{2m}\Big)}, \\
&\in \sum_{f \in \mathcal{F}} \sum_{a \in \mathcal{L}_{f,\overline{a_f^*}}} O(m \cdot |\mathcal{L}_{f,a}| + nm) \;\in\; O(n^2 m^2),
\end{aligned} \tag{27}$$

where the final bound follows by applying the same argument used to bound $\mathbb{E}\big[\big|\bar{\zeta}_{\mathcal{A}}\big|\big]$ to $\mathbb{E}\Big[\Big|\bar{\psi}_{\mathcal{F}}^{(1)}\Big|\Big]$, and then invoking Lemma 3.3 for the remaining term. Finally, we got

$$\begin{aligned}
\mathbb{E}\left[\left|\bigcup_{a_i \in \mathcal{A}} \mathcal{S}_{a_i}^{\text{inc}} \cup \mathcal{S}_{a_i}^{\text{rej}}\right|\right] \leq \sum_{a \in \mathcal{A}} \sum_{f \in \{\underline{f_a^*}\} \cup \mathcal{H}_{a,\underline{f_a^*}}} & \frac{4|\mathcal{L}_{a,f}| \exp\Big(-\frac{(\Delta^{\min}_{a,f})^2}{2m}\Big)}{1 - \exp\Big(-\frac{(\Delta^{\min}_{a,f})^2}{2m}\Big)} \\
+ \sum_{f \in \mathcal{F}} \sum_{a \in \mathcal{L}_{f,\overline{a_f^*}}} & \frac{4|\mathcal{L}_{f,a}| \exp\Big(-\frac{(\Delta^{\min}_{f,a})^2}{2m}\Big)}{1 - \exp\Big(-\frac{(\Delta^{\min}_{f,a})^2}{2m}\Big)} + \frac{4n \exp\Big(-\frac{(\overline{\Delta}^{\min}_{f})^2}{2m}\Big)}{1 - \exp\Big(-\frac{(\overline{\Delta}^{\min}_{f})^2}{2m}\Big)},
\end{aligned}$$

which is of $O(n^2 m^2)$. □

F.7.2 Proof of Convergence to a Stable Matching

We outline here the key convergence argument and discuss the impossibility of guaranteeing convergence to the agent-optimal stable matching.

Proposition F.5. *In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, where agents follow Algorithm 7 and strategic firms follow Algorithm 1, the algorithms converge to a stable matching, which is not necessarily agent-optimal.*

Proof. By Definition D.2, convergence implies the existence of a time t such that $f_a^{\texttt{match}}(t') = f_a^{\texttt{match}}(t'-1)$ for all $t' > t$ and all agents a. Let t_1 denote the start of the final committing phase. By Observation F.2, agents commit to a perfect matching $(a, f_a^{\texttt{apply}}(t_1))_{\forall a \in \mathcal{A}}$ throughout $[t_1, T]$.

Since no switching condition is triggered after t_1, no agent detects a preference-list inconsistency and no firm performs a strategic rejection. Consequently, each agent a has already been rejected by every firm in $\hat{\mathcal{H}}_{a, f_a^{\texttt{apply}}(t_1)}(t_1)$, and each firm f is matched with its most preferred agent among those who applied during the final updating phase.

Therefore, no blocking pair exists, and the matching $(a, f_a^{\texttt{apply}}(t_1))_{\forall a \in \mathcal{A}}$ is stable. Finally, as illustrated in Example F.1, the limiting stable matching need not be agent-optimal. □

F.7.3 Crucial Observations

We now present two crucial observations that will be used repeatedly in the proof of Theorem 4.1.

The following observations clarify how the set of top-n firms constrains stable outcomes in general markets.

Observation 4. *In any matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, for every agent $a \in \mathcal{A}$, the agent-optimal stable match $\overline{f_a^*}$ belongs to the set of top-n firms $\mathcal{F}_a^{(n)}$. Formally,*

$$\overline{f_a^*} \in \mathcal{F}_a^{(n)}.$$

Proof. Running GS on the ground-truth preference lists yields the agent-optimal stable matching. Since there are n agents, each agent can be rejected at most $n-1$ times during the execution of GS. Therefore, the agent-optimal stable match of any agent must lie among its top n ranked firms, that is, in $\mathcal{F}_a^{(n)}$. □

Observation 5. *In any matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, if for some agent $a \in \mathcal{A}$ the agent-pessimal stable match $\underline{f_a^*}$ does not belong to the set $\mathcal{F}_a^{(n)}$, that is, $\underline{f_a^*} \notin \mathcal{F}_a^{(n)}$, then the market admits more than one stable matching. Equivalently, $\underline{f_a^*} \neq \overline{f_a^*}$.*

Proof. If $\underline{f_a^*} \notin \mathcal{F}_a^{(n)}$ while $\overline{f_a^*} \in \mathcal{F}_a^{(n)}$ by Observation 4, then the agent-optimal and agent-pessimal stable matches of a must be distinct. Hence, the market admits more than one stable matching. □

Observation 6. *In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$ with strategic firms following Algorithm 1 and agents following Algorithm 7, consider any updating stage starting at time $t^{\texttt{GS}}$, and let $t' \doteq t^{\texttt{GS}} + 3n^2 + 1$ denote the start of the subsequent committing phase. Then, in the perfect matching $(a, f_a^{\texttt{match}}(t'))_{\forall a \in \mathcal{A}}$ to which agents commit, each agent a is matched to a firm $f_a^{\texttt{match}}(t')$ that belongs to its top-n estimated firms at time $t^{\texttt{GS}}$, i.e.,*

$$f_a^{\texttt{match}}(t') \in \hat{\mathcal{F}}_a^{(n)}(t^{\texttt{GS}}).$$

Proof. The key observation is that, throughout the updating stage, agents select their actions based on the estimated preference lists fixed at time $t^{\texttt{GS}}$. As a result, during the ensuing distributed execution of GS, each agent applies only to firms among its top-n estimated choices at time $t^{\texttt{GS}}$, which implies the claim. □

Now we conclude this section with the following lemma, which plays a crucial role in the proof of Theorem 4.1.

Lemma F.6. *In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$ with strategic firms following Algorithm 1 and agents following Algorithm 7, consider any updating stage starting at time $t^{\texttt{GS}}$, and let $t' \doteq t^{\texttt{GS}} + 3n^2 + 1$ denote the start of the subsequent committing phase. Suppose that the perfect matching $(a, f_a^{\texttt{match}}(t'))_{\forall a \in \mathcal{A}}$ to which agents commit is not stable with respect to the ground-truth preferences. If there exists a time step $t'' \geq t'$ during the committing phase such that*

$$t'' \in \Gamma_{\mathcal{A}}^{(n)} \cap \Gamma_{\mathcal{F}}^{(n)},$$

then agents initiate the next updating phase at time $t'' + 1$.

Proof. The claim follows from Observation 6 together with the definition of $\Gamma_{\mathcal{A}}^{(n)}$ and $\Gamma_{\mathcal{F}}^{(n)}$ in Definition B.4.

Since the perfect matching $(a, f_a^{\texttt{match}}(t'))_{\forall a \in \mathcal{A}}$ is not stable with respect to the ground-truth preferences, there must exist at least one agent or firm whose top-n estimated preference list was not aligned with the ground truth (either element-wise or order-wise) during the updating phase starting at $t^{\texttt{GS}}$. Equivalently, there exists a time step $t_1 \in [t^{\texttt{GS}}, t^{\texttt{GS}} + 3n^2]$ such that $t_1 \notin \Gamma_{\mathcal{A}}^{(n)} \cap \Gamma_{\mathcal{F}}^{(n)}$.

Now consider the committing phase, and let $t'' \geq t'$ be the first time step such that $t'' \in \Gamma_{\mathcal{A}}^{(n)} \cap \Gamma_{\mathcal{F}}^{(n)}$. At time t'', all agents' top-n estimated firms and all firms' top-n estimated agents coincide with the ground truth. Since the matching $(a, f_a^{\texttt{match}}(t'))_{\forall a \in \mathcal{A}}$ is unchanged throughout the committing phase and is unstable with respect to the ground-truth preferences, it is also unstable with respect to the estimated preferences at time t''. Hence a blocking pair exists.

Let (a_i, f) be such a blocking pair at time t'', where $f = f_{a_j}^{\texttt{match}}(t')$ for some a_j. By definition of a blocking pair, $f \underset{O_{a_i}}{>} f_{a_i}^{\texttt{match}}(t')$ and $a_i \underset{O_f}{>} a_j$. We consider the following cases.

(i) If $f_{a_i}^{\texttt{match}}(t') \notin \mathcal{F}_{a_i}^{(n)}$, then since $\mathcal{F}_{a_i}^{(n)} = \hat{\mathcal{F}}_{a_i}^{(n)}(t'')$ element-wise and order-wise, it follows that $f_{a_i}^{\texttt{match}}(t') \notin \hat{\mathcal{F}}_{a_i}^{(n)}(t'')$, which induces a preference-list inconsistency for agent a_i, and therefore $t'' \in \mathcal{S}_{a_i}^{\text{inc}}$.

(ii) If $f_{a_i}^{\texttt{match}}(t') \in \mathcal{F}_{a_i}^{(n)}$ and agent a_i did not apply to f during the previous updating phase, then necessarily $f \underset{O_{a_i}}{<} f_{a_i}^{\texttt{match}}(t')$ at time $t^{\texttt{GS}}$. At time t'', top-n alignment implies the ordering reverses to $f \underset{O_{a_i}}{>} f_{a_i}^{\texttt{match}}(t')$, which again yields a preference-list inconsistency for agent a_i, and hence $t'' \in \mathcal{S}_{a_i}^{\text{inc}}$.

(iii) If $f_{a_i}^{\texttt{match}}(t') \in \mathcal{F}_{a_i}^{(n)}$ and agent a_i did apply to f during the previous updating phase but was rejected, then f must have rejected a_i in favor of its current match a_j due to a preference-list inconsistency. At time t'', the estimated list of f is fully aligned with the ground truth and satisfies $a_i \underset{O_f}{>} a_j$, which triggers a firm-side strategic rejection, and therefore $t'' \in \mathcal{S}_{a_j}^{\text{rej}}$.

In all cases, a switching condition is triggered at time t'', and therefore $t'' + 1$ is the first round of the next updating phase. □

F.7.4 Proof of Theorem 4.1 for General Markets

Theorem 4.1. *In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, with firms and agents following Algorithm 1 and Algorithm 7, the expected regret is time-independent. In particular,*

$$\mathbb{E}[\underline{R}_a(\mathcal{T})] \in \begin{cases} O(n^3 m^2 \min\{\overline{\Delta}_{\mathcal{A}}, \underline{\Delta}_{\mathcal{F}}\}^{-2} + m \sum_{j=1}^{i} (|\mathcal{L}_{a_j, f_j}| \, \overline{\Delta}_{a_j}^{-2} + |\mathcal{L}_{f_j, a_j}| \, \underline{\Delta}_{f_j}^{-2})), & \alpha\text{-reducible}, \\ O(n^4 m^2 \Delta^{-2}), & \text{General Markets}. \end{cases}$$

Proof. We bound the pessimal regret $\mathbb{E}[\underline{R}_a(T)]$ by decomposing it into regret incurred during the updating and committing phases and bounding each term separately:

$$\mathbb{E}[\underline{R}_a(T)] = \mathbb{E}[\underline{R}_a(\mathcal{U})] + \mathbb{E}[\underline{R}_a(\mathcal{C})] \,.$$

For $\mathbb{E}[\underline{R}_a(\mathcal{U})]$, we write:

$$\mathbb{E}[\underline{R}_a(\mathcal{U})] = \sum_{t \in \mathcal{U}} \left(u_{a, \underline{f_a^*}} - \mathbb{E}[X_{a, f_a^{\texttt{match}}(t)}] \right) \leq 3n^2 \cdot \mathbb{E}\left[\left| \bigcup_{a' \in \mathcal{A}} \left(\mathcal{S}_{a'}^{\text{inc}} \cup \mathcal{S}_{a'}^{\text{rej}} \right) \right| \right]. \tag{28}$$

Before bounding the committing regret, we introduce the following notation. Let

$$\bigcup_{a \in \mathcal{A}} \left(\mathcal{S}_a^{\text{inc}} \cup \mathcal{S}_a^{\text{rej}} \right) = \{t_1, t_2, \ldots\}$$

denote the set of switching times, ordered increasingly. The j-th committing phase is defined as the interval

$$[t'_j : t_{j+1}], \qquad \text{where } t'_j \doteq t_j + 3n^2 + 1,$$

which corresponds to the period between the j-th and $(j+1)$-th distributed executions of GS. Accordingly, the collection of committing intervals is

$$\mathcal{K} \doteq \left\{ [t'_j,\, t_{j+1}] : j \in \left[\left| \bigcup_{a \in \mathcal{A}} (\mathcal{S}_a^{\text{inc}} \cup \mathcal{S}_a^{\text{rej}}) \right| - 1 \right] \right\}.$$

Now for the committing regret we write

$$\begin{aligned}
\mathbb{E}[\underline{R_a}(\mathcal{C})] &= \sum_{t \in \mathcal{C}} \left(u_{a,\underline{f_a^*}} - \mathbb{E}[X_{a,f_a^{\texttt{match}}(t)}] \right) \\
&= \mathbb{E}\left[\sum_{j=1}^{|\mathcal{K}|} \sum_{t \in [t'_j,\, t_{j+1}]} \left(u_{a,f} - X_{a,f_a^{\texttt{match}}(t'_j)}(t) \right) \right], \\
&\overset{(a)}{\leq} \mathbb{E}\left[\left| \bigcup_{a' \in \mathcal{A}} \left(\mathcal{S}_{a'}^{\text{inc}} \cup \mathcal{S}_{a'}^{\text{rej}} \right) \right| \right] + \mathbb{E}\left[\sum_{j \in [|\mathcal{K}|]} \left(t_{j+1} - 1 - t'_j \right) \,\middle|\, f_a^{\texttt{match}}(t'_j) \in \mathcal{L}_{a,\underline{f_a^*}} \wedge (a'', f_{a''}^{\texttt{match}}(t'_j))_{\forall a'' \in \mathcal{A}} \text{ is not stable} \right], \\
&\overset{(b)}{\leq} \mathbb{E}\left[\left| \bigcup_{a' \in \mathcal{A}} \left(\mathcal{S}_{a'}^{\text{inc}} \cup \mathcal{S}_{a'}^{\text{rej}} \right) \right| \right] + \mathbb{E}\left[\left| \bar{\Gamma}_{\mathcal{A}}^{(n)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)} \right| \right]. \qquad (29)
\end{aligned}$$

Inequality (a) is obtained by separating, within each committing phase, the single signaling round during which an agent intentionally abstains from applying in order to signal a switching condition. Each committing phase contributes at most one such round, and therefore the total number of signaling rounds is bounded by $\mathbb{E}\left[\left| \bigcup_{a' \in \mathcal{A}} \left(\mathcal{S}_{a'}^{\text{inc}} \cup \mathcal{S}_{a'}^{\text{rej}} \right) \right| \right]$. The second term captures the expected regret accumulated during the remaining rounds of each committing phase. Since we analyze the pessimal regret of agent a, regret is incurred only when a is matched to a firm that it strictly prefers less than its pessimal stable match $\underline{f_a^*}$, i.e., only when $f_a^{\texttt{match}}(t'_j) \in \mathcal{L}_{a,\underline{f_a^*}}$. In this case, the perfect matching reached at the end of the most recent updating stage, $(a'', f_{a''}^{\texttt{match}}(t'_j))_{\forall a'' \in \mathcal{A}}$, is necessarily unstable, which justifies the conditioning in the second term of (a).

Inequality (b) follows from Lemma F.6, since the matching $(a'', f_{a''}^{\texttt{match}}(t'_j))_{\forall a'' \in \mathcal{A}}$ is unstable. Specifically, consider the committing interval $[t'_j,\, t_{j+1}]$. If every time step in this interval belongs to $\bar{\Gamma}_{\mathcal{A}}^{(n)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}$, then the entire interval is charged to the second term of (29), while the signaling round is accounted for separately in the first term. Otherwise, there exists a time step $t'' \in [t'_j,\, t_{j+1}]$ such that $t'' \in \Gamma_{\mathcal{A}}^{(n)} \cap \Gamma_{\mathcal{F}}^{(n)}$. By Lemma F.6, this time step must be the signaling round t_{j+1} that triggers the next updating phase, and hence it is counted in the first term of (29).

The first term of (29) is bounded by Lemma F.4. We therefore bound the remaining term $\mathbb{E}\left[\left| \bar{\Gamma}_{\mathcal{A}}^{(n)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)} \right| \right]$ as follows:

$$\mathbb{E}\left[\left| \bar{\Gamma}_{\mathcal{A}}^{(n)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)} \right| \right] \leq \mathbb{E}\left[\left| \bar{\Gamma}_{\mathcal{A}}^{(n)} \right| \right] + \mathbb{E}\left[\left| \bar{\Gamma}_{\mathcal{F}}^{(n)} \right| \right] \leq \sum_{a' \in \mathcal{A}} \sum_{f' \in \mathcal{F}_{a'}^{(n)}} \mathbb{E}\left[\left| \bar{\mathcal{E}}_{a',f'} \right| \right] + \sum_{f'' \in \mathcal{F}} \sum_{a'' \in \mathcal{A}} \mathbb{E}\left[\left| \bar{\mathcal{E}}_{f'',a''} \right| \right]. \qquad (30)$$

The final inequality follows by expanding the definitions of $\bar{\Gamma}_{\mathcal{A}}^{(n)}$ and $\bar{\Gamma}_{\mathcal{F}}^{(n)}$ and applying a union bound.

Thus, with respect to (28) and (30), we obtain

$$\begin{aligned}
\mathbb{E}[\underline{R}_a(T)] &= \mathbb{E}[\underline{R}_a(\mathcal{U})] + \mathbb{E}[\underline{R}_a(\mathcal{C})] \\
&\overset{(a)}{\leq} (3n^2+1) \cdot \mathbb{E}\left[\left| \bigcup_{a' \in \mathcal{A}} \left(\mathcal{S}_{a'}^{\text{inc}} \cup \mathcal{S}_{a'}^{\text{rej}} \right) \right| \right] + \sum_{a' \in \mathcal{A}} \sum_{f' \in \mathcal{F}_{a'}^{(n)}} \mathbb{E}\left[\left| \bar{\mathcal{E}}_{a',f'} \right| \right] + \sum_{f'' \in \mathcal{F}} \sum_{a'' \in \mathcal{A}} \mathbb{E}\left[\left| \bar{\mathcal{E}}_{f'',a''} \right| \right] \\
&\in O\big(n^4 m^2 \Delta^{-2}\big),
\end{aligned}$$

where $\Delta \doteq \min_{a \in \mathcal{A}, f \in \mathcal{F}} \min(\Delta_{a,f}, \Delta_{f,a})$. □

G Deferred Concepts and Proofs from Section 4.3

This section collects additional definitions and deferred proofs from Section 4.3 to streamline the main presentation.

G.1 Extended Coordination-Free Decentralized Algorithm with Anonymous Hiring Changes $\mathcal{V}^+(t)$ as Feedback

We now design Algorithm 8 for the anonymous hiring-changes feedback $\mathcal{V}^+(t)$. Under this feedback, an agent treats any firm that appears in $\mathcal{V}^+(t)$ (either because it hired a new agent or became vacant) as a local signal that it may be worth reconsidering. Accordingly, an agent may revisit firms ranked above its previous match whenever those firms appear in $\mathcal{V}^+(t)$, viewing such events as opportunities to compete with the new hire and improve its match.

Algorithm 8 Coordination-Free Decentralized Learning

Input: a, $\mathcal{F}$, anonymous hiring changes $\mathcal{V}^+$
Initialize: $r_{a,f}(1) \leftarrow 0$ for all $f \in \mathcal{F}$
for $t \in \mathcal{T}$ **do**
  Construct $\mathcal{B}'_a(t)$ according to (31)
  $f_a^{\texttt{apply}}(t) \leftarrow \arg\max_{f \in \mathcal{B}'_a(t)} \hat{u}_{a,f}(t)$
  Interview with $\mathcal{I}_a(t) = \{f_a^{\texttt{apply}}(t), f_a^{\texttt{RR}}(t)\}$ and apply to $f_a^{\texttt{apply}}(t)$
  UPDATEAGENTREJVARS(a, t)

To choose $f_a^{\texttt{apply}}(t)$, the agent first constructs a *candidate set* $\mathcal{B}'_a(t)$ consisting of firms that either have never rejected a, or have exhibited a hiring change since the last time they rejected a to hire another agent:

$$\mathcal{B}'_a(t) \doteq \left\{ f : \exists\, t' \in [r_{a,f}(t), t), f \in \mathcal{V}^+(t') \right\} \tag{31}$$

The agent then applies to the firm in $\mathcal{B}'_a(t)$ with the highest estimated utility, $f_a^{\texttt{apply}}(t) = \arg\max_{f \in \mathcal{B}'_a(t)} \hat{u}_{a,f}(t)$.

For the agent-dependent bound in Theorem 4.2, In the worst case, $\sum_{j' \in [i]} |\mathcal{H}_{a_{j'}, f_{j'}}| = O(i^2)$, so agent a_i has regret $O(i^3 m^2)$; hence the worst-case over agents is $O(n^3 m^2)$.

To handle general (not necessarily structured) markets in a coordination-free manner, we extend Algorithm 3 to the randomized Algorithm 9, which uses $k = 3$ interviews per round and a fixed randomization parameter $\lambda \in (0, 1)$.

The exponentially large factor ϵ^{-1} reflects the need to rule out blocking-pair resolution cycles that may arise in general markets under coordination-free dynamics; such cycles do not occur in α-reducible markets. Similar large constants also appear in the $O(\log T)$ bounds of [19], albeit under more informative feedback than our $\mathcal{V}^+$.

G.2 Crucial Structural Observation

Observation 7. *By any time t, each agent $a \in \mathcal{A}$ must have been rejected by every firm $f \in \hat{\mathcal{H}}_{a, f_a^{\texttt{apply}}(t)}(t)$ since the most recent hiring change at time t', i.e., $f \in \mathcal{V}^+(t')$ and $r_{a,f}(t) \geq t'$.*

Proof. Algorithm 3 starts with $\mathcal{B}'_a(1) = \mathcal{F}$. By definition of $\mathcal{B}'_a(t)$, whenever the winner of a firm f changes at time t', we have $f \in \mathcal{B}'_a(t'+1)$. Thus, at $t'+1$, $f_a^{\texttt{apply}}(t'+1)$ is chosen so that $\hat{u}_{a, f_a^{\texttt{apply}}(t'+1)} \geq \hat{u}_{a,f}$, and a applies to a more preferred firm than f. Since $f_a^{\texttt{apply}}$ moves downward over time toward less preferred firms, once $f \in \hat{\mathcal{H}}_{a, f_a^{\texttt{apply}}(t)}$, there must exist some t'' with $t' < t'' < t$ where $f_a^{\texttt{apply}}(t'') = f$. At that round, a applies to f, is rejected, and does not reconsider f until its winner changes again. □

G.3 Proof Sketch of Theorem 4.2 for the agent a_1 in α-reducible Markets

Proof Sketch. Our proof has three steps: (i) exclude the (time-independent) *invalid* rounds for the pair (a_1, f_1) and reduce regret to counting them; (ii) show that on any two consecutive *valid* rounds, a_1 is matched to f_1 in the second; and (iii) conclude that within each maximal valid interval a_1 can incur regret only in its first round, and the number of such intervals is at most the number of invalid rounds plus one. Concretely, recalling $|\mathcal{H}_{a_1,f_1}| = 0$, we obtain the time-independent bound $\mathbb{E}[R_{a_1}(\mathcal{T})] \le 2\,\mathbb{E}\big[|\bar{\mathcal{E}}_{a_1,f_1} \cup \bar{\mathcal{E}}_{f_1,a_1}|\big] + 1$, where $\bar{\mathcal{E}}_{a_1,f_1} \cup \bar{\mathcal{E}}_{f_1,a_1}$ is the set of rounds in which either a_1 or f_1 has an invalid estimated list w.r.t. their optimal match, and its expected size is time-independent by Lemma 3.3.

Consider two consecutive rounds t and $t+1$ with both lists valid w.r.t. the optimal match, i.e., $t, t+1 \in \mathcal{E}_{a_1,f_1} \cap \mathcal{E}_{f_1,a_1}$. We show $f_{a_1}^{\texttt{match}}(t+1) = f_1$ (while $f_{a_1}^{\texttt{match}}(t)$ may or may not equal f_1). This implies that within any maximal *valid* interval (a consecutive block of rounds outside $\bar{\mathcal{E}}_{a_1,f_1} \cup \bar{\mathcal{E}}_{f_1,a_1}$), a_1 can incur regret in at most its first round; moreover, since valid and invalid intervals alternate, the number of valid intervals is at most $|\bar{\mathcal{E}}_{a_1,f_1} \cup \bar{\mathcal{E}}_{f_1,a_1}| + 1$.

To prove $f_{a_1}^{\texttt{match}}(t+1) = f_1$, first note that if $f_{a_1}^{\texttt{apply}}(t) = f_1$, then $f_{a_1}^{\texttt{match}}(t) = f_1$ immediately. Otherwise, by the definition of the candidate set $\mathcal{B}'_{a_1}(t)$ and Observation 7, at the most recent time a_1 applied to f_1 and was rejected, firm f_1 had already hired some agent a_i with $i \neq 1$, and this agent remained matched to f_1 up to round t. By the strategic rejection condition $\mathcal{S}_{f_1}$ in (8), agent a_1 is now estimated preferable to a_i, and since $r_{f_1,a_1}(t) \ge c_{f_1}(t)$, firm f_1 sets $\gamma_{f_1}(t) = 0$. This abstention is revealed to a_1 as a hiring change, prompting a_1 to reconsider f_1 next round, hence $f_{a_1}^{\texttt{apply}}(t+1) = f_{a_1}^{\texttt{match}}(t+1) = f_1$. □

G.4 Proof of Theorem 4.2 in α-reducible Markets

Theorem 4.2. *In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, with firms following Algorithm 1 and agents following coordination-free Algorithm 3 (for α-reducible markets with $k = 2$) and Algorithm 9 (for general markets with $k = 3$, where an agent may apply to two interviewed firms per round), the expected regret is time-independent. In particular,*

$$\mathbb{E}[\underline{R}_a(\mathcal{T})] \in \begin{cases} O\big((\sum_{j\in[i]} |\mathcal{H}_{a_j,f_j}| + 1) \cdot (\sum_{j\in[i]} (m^2 \overline{\Delta}_{a_j}^{-2} + nm\, \underline{\Delta}_{f_j}^{-2}) + 1)\big), & \alpha\text{-reducible } \mathcal{M}(\mathcal{A},\mathcal{F}), \\ O\big(\epsilon^{-1} n^5 m^4 \Delta^{-2}\big)\text{, for } \epsilon = \big(\lambda(1-\lambda)^{n-1}\big)^{n^4 m + nm} \text{ with } \lambda \in (0,1), & \text{General Markets}. \end{cases}$$

Proof. We present the proof for the case of uncertain firms; the case of certain firms then follows directly by setting the firm-uncertainty parameters to zero. Now int the α-reducible market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, for agent a_i, we write

$$\begin{aligned}
\mathbb{E}[R_{a_i}(T)] &\le \mathbb{E}\big[\big|\big\{\, t \in \mathcal{T} : f_{a_i}^{\texttt{match}}(t) \in \emptyset \cup \mathcal{L}_{a_i,f_i} \,\big\}\big|\big]\,, \\
&\le \mathbb{E}\big[\big|\bar{\mathcal{E}}_{a_i,f_i} \cup \big\{\, t \in \mathcal{E}_{a_i,f_i} : f_{a_i}^{\texttt{match}}(t) \in \emptyset \cup \mathcal{L}_{a_i,f_i} \,\big\}\big|\big]\,, \\
&\le \mathbb{E}\left[\left|\bar{\mathcal{E}}_{a_i,f_i} \cup \left\{\, t \in \mathcal{E}_{a_i,f_i} : f_{a_i}^{\texttt{apply}}(t) \underset{\hat{O}_{a_i}(t)}{<} f_i \,\right\} \cup \left\{\, t \in \mathcal{E}_{a_i,f_i} : f_{a_i}^{\texttt{apply}}(t) \underset{\hat{O}_{a_i}(t)}{\ge} f_i \wedge f_{a_i}^{\texttt{match}}(t) = \emptyset \,\right\}\right|\right], \\
&\overset{(a)}{\le} \mathbb{E}\left[\left|\bar{\mathcal{E}}_{a_i,f_i} \cup \left(\bar{\mathcal{E}}_{f_i,a_i} \cup \bigcup_{j\in[i-1]} \big(\bar{\mathcal{E}}_{a_j,f_j} \cup \bar{\mathcal{E}}_{f_j,a_j}\big)\right) \cup \left\{\, t \in \mathcal{E}_{a_i,f_i} : f_{a_i}^{\texttt{apply}}(t) \underset{\hat{O}_{a_i}}{\ge} f_i \wedge f_{a_i}^{\texttt{match}}(t) = \emptyset \,\right\}\right|\right], \\
&\le \mathbb{E}\left[\left|\bigcup_{j\in[i]} \big(\bar{\mathcal{E}}_{a_j,f_j} \cup \bar{\mathcal{E}}_{f_j,a_j}\big) \cup \left\{\, t \in \bigcap_{j'\in[i]} \Big(\mathcal{E}_{a_{j'},f_{j'}} \cap \mathcal{E}_{f_{j'},a_{j'}}\Big) : f_{a_i}^{\texttt{match}}(t) = \emptyset \,\right\}\right|\right], \\
&\le \mathbb{E}\left[\left|\bigcup_{j\in[i]} \big(\bar{\mathcal{E}}_{a_j,f_j} \cup \bar{\mathcal{E}}_{f_j,a_j} \cup \bar{\Phi}_j\big)\right|\right] + \mathbb{E}\left[\left|\left\{\, t \in \bigcap_{j'\in[i]} \Phi_{j'} : f_{a_i}^{\texttt{match}}(t) = \emptyset \,\right\}\right|\right], \qquad (32)
\end{aligned}$$

where Φ_i is defined as

$$\Phi_i \doteq \Big\{ t \in (\mathcal{E}_{a_i,f_i} \cap \mathcal{E}_{f_i,a_i}) : \Big(\nexists f \in \mathcal{F} \text{ s.t. } f \in \mathcal{H}_{a_i,f_i} \wedge f \in \hat{\mathcal{L}}_{a_i,f_i}(t)\Big) \wedge \Big(\nexists a \in \mathcal{A} \text{ s.t. } a \in \mathcal{H}_{f_i,a_i} \wedge a \in \hat{\mathcal{L}}_{f_i,a_i}(t)\Big) \Big\}. \qquad (33)$$

That is, Φ_i denotes the set of time steps during which both f_i and a_i possess valid preference lists with respect to their optimal stable match, and all peers preferred to f_i or a_i under the ground truth are also ranked higher in their estimated lists (though not necessarily in the correct internal order).

Then the inequality (a) holds because at each time step within the set $\left\{ t \in \mathcal{E}_{a_i,f_i} : f_{a_i}^{\texttt{apply}}(t) \underset{\hat{O}_{a_i}(t)}{<} f_i \right\}$, the chosen firm $f_{a_i}^{\texttt{apply}}(t)$ is estimated to be less preferred than f_i. By Observation 7, this implies that agent a_i must have applied to f_i at some earlier round $t' < t$ and been rejected and by definition of $\mathcal{B}'_a(t)$, the hired agent $a_{f_i}^{\texttt{match}}(t)$ has not changed since then. We now examine why the most recent time t' this rejection occurred which then hold that then hired agent $a_{f_i}^{\texttt{match}}(t')$ remains the same at t.

i) If $\hat{O}_{f_i}(t')$ was invalid at that round, then $\hat{O}_{f_i}(t)$ must also remain invalid at all subsequent rounds, since the agent hired by f_i at time t', namely $a_{f_i}^{\texttt{match}}(t')$, has not changed until t. Hence, $t \in \bar{\mathcal{E}}_{f_i,a_i}$.

ii) If $\hat{O}_{f_i}(t')$ was valid, then a_i must have been blocked by some other agent $a_j \in \mathcal{H}_{f_i,a_i}$ with $j \in [i-1]$. In this case, the hired agent $a_{f_i}^{\texttt{match}}(t')$ has also not changed throughout $[t', t]$, so the blocking of a_i persists. This persistence implies that either some higher-layer agent a_j has an invalid preference list $\hat{O}_{a_j}(t'')$, or some firm with higher index f_j has an invalid preference list $\hat{O}_{f_j}(t'')$ for all $t'' \in [t', t]$. Otherwise, by α-reducibility, $a_{f_i}^{\texttt{match}}(t')$ would eventually have been forced to leave f_i, since f_i would be suboptimal for that agent, contradicting the fact that $a_{f_i}^{\texttt{match}}(t')$ remained unchanged until t. Putting i) and ii) together, we conclude that every such t must fall in one of the following categories: (i) a_i or f_i has an invalid list, or (ii) some higher-ranked agent a_j or firm f_j has an invalid list. Formally,

$$\left\{ t \in \mathcal{E}_{a_i,f_i} : f_{a_i}^{\texttt{apply}}(t) \underset{\hat{O}_{a_i}(t)}{<} f_i \right\} \subseteq \left(\bar{\mathcal{E}}_{f_i,a_i} \cup \bigcup_{j \in [i-1]} \left(\bar{\mathcal{E}}_{a_j,f_j} \cup \bar{\mathcal{E}}_{f_j,a_j} \right) \right).$$

To establish the regret bound for the case of uncertain firms, we bound inequality (32) differently. For the first term, we have

$$\begin{aligned} \mathbb{E}\left[\left| \bigcup_{j \in [i]} \left(\bar{\mathcal{E}}_{a_j,f_j} \cup \bar{\mathcal{E}}_{f_j,a_j} \cup \bar{\Phi}_j \right) \right| \right] &\leq \sum_{j \in [i]} \frac{4m \exp\left(-\frac{\overline{\Delta}_{a_j}^2}{2m} \right)}{1 - \exp\left(-\frac{\overline{\Delta}_{a_j}^2}{2m} \right)} + \frac{4n \exp\left(-\frac{\underline{\Delta}_{f_j}^2}{2m} \right)}{1 - \exp\left(-\frac{\underline{\Delta}_{f_j}^2}{2m} \right)} \\ &\in \sum_{j \in [i]} O\left(m^2 \overline{\Delta}_{a_j}^{-2} + nm\, \underline{\Delta}_{f_j}^{-2} \right). \end{aligned} \tag{34}$$

To bound the second term of inequality (32), recall that, by Observation 2, strategic certain firms always hire. Hence, for such firms, we can bound the second term as

$$\mathbb{E}\left[\left| \left\{ t \in \bigcap_{j' \in [i]} \Phi_{j'} : f_{a_i}^{\texttt{match}}(t) = \emptyset \right\} \right| \right] \leq \left(1 + \sum_{j' \in [i]} \left| \mathcal{H}_{a_{j'},f_{j'}} \right| \right) \cdot \left(\mathbb{E}\left[\left| \bigcup_{j \in [i]} \left(\bar{\mathcal{E}}_{a_j,f_j} \cup \bar{\mathcal{E}}_{f_j,a_j} \cup \bar{\Phi}_j \right) \right| \right] + 1 \right). \tag{35}$$

This inequality holds as follows. Consider a time interval $[t_1, t_2]$ such that, for all $t' \in [t_1 : t_2]$, we have $t' \in \bigcap_{j' \in [i]} \Phi_{j'}$. We now argue that there are at most $1 + \sum_{j' \in [i]} \left| \mathcal{H}_{a_{j'},f_{j'}} \right|$ initial consecutive rounds within this interval during which agent a_i is not matched with f_i, i.e., $f_{a_i}^{\texttt{match}}(t') \neq f_i$ for all $t' \in \left[t_1 : t_1 + 1 + \sum_{j' \in [i]} \left| \mathcal{H}_{a_{j'},f_{j'}} \right| \right]$.

For agent a_1, as outlined in the proof sketch of Theorem 4.2 in Section G.1, the α-reducibility structure at time step t_1 implies that either $f_{a_1}^{\texttt{apply}}(t_1) = f_1$ or $\gamma_{f_1}(t_1) = 1$. Consequently, $f_{a_1}^{\texttt{match}}(t') = f_1$ for all $t' \in [t_1 + 1, t_2]$, according to the decision-making rules of agents under Algorithm 3 and firms under Algorithm 1. Thus, agent a_1 incurs regret only at time step t_1.

We now proceed by induction to establish the argument for agent a_i. Before doing so, we highlight a crucial observation. Let t_i^* denote the last time step within the interval $[t_1 : t_2]$ at which there

exists some agent a_j for $j \in [i-1]$ such that $f_{a_j}^{\texttt{match}}(t_i) \neq f_j$. Then, for all subsequent rounds $t' \in [t_i^* + 1 : t_2]$, the strategic uncertain firm f_i will hire a_i whenever it receives an application from it. This follows because, due to the validity condition of the preference lists of both sides, a_i remains the most preferred agent among those who still consider f_i as an active option, while all agents in $\mathcal{H}_{f_i,a_i}$ have already secured their stable matches and no longer apply to f_i.

By the inductive hypothesis, the latest time at which any agent $a_{j'}$ for $j' \in [i-1]$ is not matched with its optimal firm happens within at most $1 + \sum_{k' \in [i-1]} \left|\mathcal{H}_{a_{k'},f_{k'}}\right|$ rounds, thus $t_i^* \leq 1 + \sum_{k' \in [i-1]} \left|\mathcal{H}_{a_{k'},f_{k'}}\right|$. Hence, during the interval $t'' \in \left[t_1 + 1 + \sum_{k' \in [i-1]} \left|\mathcal{H}_{a_{k'},f_{k'}}\right| : t_2\right]$, we have $f_{a_k}^{\texttt{match}}(t'') = f_k$ for all $k \in [i-1]$.

Then, from the definition of $\bigcap_{j' \in [i]} \Phi_{j'}$, it follows that throughout $\left[t_1 + 1 + \sum_{k' \in [i-1]} \left|\mathcal{H}_{a_{k'},f_{k'}}\right| : t_2\right]$, there are at most $|\mathcal{H}_{a_i,f_i}|$ consecutive rounds in which a_i applies to firms and gets rejected. This is because, starting from time step $t_1 + 1 + \sum_{k' \in [i-1]} \left|\mathcal{H}_{a_{k'},f_{k'}}\right|$, all firms in $\mathcal{H}_{a_{k'},f_{k'}}$ are matched with their optimal agents and no longer change their hires during $[t_1 + \left|\mathcal{H}_{a_{k'},f_{k'}}\right| + 1 : t_2]$. Therefore, under Algorithm 3, agent a_i applies to each such firm at most once in sequence until eventually reaching f_i. Hence, $t_1 + 1 + \sum_{j' \in [i]} \left|\mathcal{H}_{a_{j'},f_{j'}}\right|$ is the last time step at which a_i incurs regret within $[t_1 : t_2]$, resulting in a total regret of $1 + \sum_{j' \in [i]} \left|\mathcal{H}_{a_{j'},f_{j'}}\right|$ during the interval. Finally, as the total number of such intervals $[t_1 : t_2]$ is bounded by $\mathbb{E}\left[\left|\bigcup_{j \in [i]} \left(\bar{\mathcal{E}}_{a_j,f_j} \cup \bar{\mathcal{E}}_{f_j,a_j} \cup \bar{\Phi}_j\right)\right|\right] + 1$, we can prove the inequality (35).

Thus, with respect to inequalit (32) for an strategic certain agent a_i, the regret is bounded by merging inequalitues (34) and (35) and get

$$\begin{aligned}
\mathbb{E}[R_{a_i}(T)] &\leq 4 \cdot \left(\sum_{j' \in [i]} \left|\mathcal{H}_{a_{j'},f_{j'}}\right| + 2\right) \cdot \left(\sum_{j \in [i]} \frac{4m \exp\left(-\frac{\overline{\Delta}_{a_j}^2}{2m}\right)}{1 - \exp\left(-\frac{\overline{\Delta}_{a_j}^2}{2m}\right)} + \frac{4n \exp\left(-\frac{\underline{\Delta}_{f_j}^2}{2m}\right)}{1 - \exp\left(-\frac{\underline{\Delta}_{f_j}^2}{2m}\right)}\right) + \sum_{j' \in [i]} \left|\mathcal{H}_{a_{j'},f_{j'}}\right| + 1 \\
&\leq 4 \cdot \left(\sum_{j' \in [i]} \left|\mathcal{H}_{a_{j'},f_{j'}}\right| + 2\right) \cdot \sum_{j \in [i]} \left(O\left(m^2 \overline{\Delta}_{a_j}^{-2} + nm\, \underline{\Delta}_{f_j}^{-2}\right)\right) + \sum_{j' \in [i]} \left|\mathcal{H}_{a_{j'},f_{j'}}\right| + 1 \\
&\in O\left(\left(\sum_{j \in [i]} |\mathcal{H}_{a_j,f_j}| + 1\right) \cdot \left(\sum_{j \in [i]} \left(m^2 \overline{\Delta}_{a_j}^{-2} + nm\, \underline{\Delta}_{f_j}^{-2}\right) + 1\right)\right).
\end{aligned}$$

□

G.5 Extension of Coordination-Free Algorithm 3 to General Markets

We now extend our results to general (unstructured) markets and focus on bounding the pessimal regret. We first present an example showing that, consistent with standard impossibility phenomena for decentralized learning in matching markets, when the underlying market admits multiple stable matchings, interviews alone may not suffice to ensure convergence to the agent-optimal stable matching. Consequently, in markets with potentially multiple stable outcomes, our goal is to bound the pessimal regret, namely the regret incurred when the dynamics converge to a stable matching that is suboptimal from the agents' perspective. We then address two algorithmic necessities that arise in this general setting. First, we show that moving beyond $k = 2$ interviews per round (in particular, allowing $k = 3$) is necessary to avoid unresolved application cycles caused by concurrent attempts to eliminate blocking pairs. Second, we show that even with $k = 3$, convergence can fail under the anonymous firm-side feedback $\mathcal{V}^+$ unless agents are allowed to apply to more than one firm in a round and then select among the firms that admit them, so as to remain matched without inducing spurious vacancy signals.

G.5.1 Impossibility of Guaranteeing Convergence to the Agent-Optimal Stable Matching

Example G.1. *Let the market consist of agents $\mathcal{A} = \{a_1, a_2, a_3\}$ and firms $\mathcal{F} = \{f_1, f_2, f_3\}$, with true preference lists*

$$O_{a_1} : (f_1^* \succ f_2 \succ f_3^\dagger) \quad O_{f_1} : (a_2^\dagger \succ a_3 \succ a_1^*)$$
$$O_{a_2} : (f_2^* \succ f_3 \succ f_1^\dagger) \quad O_{f_2} : (a_3^\dagger \succ a_1 \succ a_2^*)$$
$$O_{a_3} : (f_3^* \succ f_2^\dagger \succ f_1) \quad O_{f_3} : (a_1^\dagger \succ a_2 \succ a_3^*).$$

Under these preferences, multiple stable matchings exist. The agent-optimal (firm-pessimal) stable matching is $(a_1, f_1), (a_2, f_2), (a_3, f_3)$ (marked by *), whereas the agent-pessimal (firm-optimal) stable matching is $(a_1, f_3), (a_2, f_1), (a_3, f_2)$ (marked by †).

Now consider a time step t in which Algorithm 3 is executed using the estimated preferences $\hat{O}(t)$. Suppose agent a_1 has not yet learned its true preference list, while agents a_2 and a_3 have already converged. The estimated lists and current interview sets are

$$\hat{O}_{a_1}(t) : (f_2 \succ f_1 \succ f_3) \quad I_{a_1}(t) = \{f_3, f_{a_1}^{RR}\}$$
$$\hat{O}_{a_2}(t) : (f_2 \succ f_3 \succ f_1) \quad I_{a_2}(t) = \{f_1, f_{a_2}^{RR}\}$$
$$\hat{O}_{a_3}(t) : (f_3 \succ f_2 \succ f_1) \quad I_{a_3}(t) = \{f_2, f_{a_3}^{RR}\}.$$

Thus, a_1 is the only agent with an incorrect estimate at time t. At a later time $t' > t$, round-robin interviewing corrects a_1's estimate to $\hat{O}_{a_1}(t') : (f_1 \succ f_2 \succ f_3)$. However, since the hired agents at f_1 and f_2 remain unchanged from time t, agent a_1 continues to set $f_{a_1}^{\texttt{apply}}(t') = f_3$. Consequently, the dynamics converge to the agent-pessimal stable matching $(a_1, f_3), (a_2, f_1), (a_3, f_2)$. This shows that when multiple stable matchings exist, decentralized learning may fail to reach the agent-optimal outcome, and one can only guarantee regret with respect to the agent-pessimal stable matching.

G.5.2 Moving from $k = 2$ to $k = 3$

This section explains why allowing each agent to interview $k = 3$ firms per round is necessary to obtain time-independent regret bounds in general (unstructured) markets. The core issue is concurrency: multiple agents may attempt to resolve blocking pairs simultaneously, and their actions can interfere. The following Example G.2 shows that when $k = 2$, Algorithm 3 can enter an *application cycle* that never resolves, even after all agents' estimated preferences fully aligned with the ground truth.

Example G.2. *Consider a market with agents $\mathcal{A} = \{a_1, a_2\}$ and firms $\mathcal{F} = \{f_1, f_2\}$. Suppose that from some time t' onward the estimated preferences are exactly aligned with the ground truth and remain fixed, i.e., for all $t \geq t'$ we have $\hat{O}_a(t) = O_a$ for all $a \in \mathcal{A}$ and $\hat{O}_f(t) = O_f$ for all $f \in \mathcal{F}$. In particular, for every $t \geq t'$,*

$$\hat{O}_{a_1}(t) : (f_1^* \succ f_2) \quad \hat{O}_{f_1}(t) : (a_2 \succ a_1)$$
$$\hat{O}_{a_2}(t) : (f_2 \succ f_1^*) \quad \hat{O}_{f_2}(t) : (a_1 \succ a_2).$$

Here, the $*$ superscripts indicate the firms $f_{a_i}^{\texttt{apply}}(t)$ chosen by agents under Algorithm 3.

Fix any round $t \geq t'$, and assume that at time $t-1$ agent a_2 applied to f_2 and was rejected because f_2 admitted a_1; hence $r_{a_2,f_2}(t) \geq t-1$, and since $f_2 \notin \mathcal{V}^+(t-1)$ under this anonymous feedback, we have $f_2 \notin \mathcal{B}'_{a_2}(t)$. Starting from round t, the algorithm can enter the following cycle:

- *At time t, both agents apply to f_1, and f_1 admits a_2 and rejects a_1; hence $f_2 \in \mathcal{V}^+(t)$.*
- *At time $t+1$, agent a_1 applies to f_2 since $f_1 \notin \mathcal{B}'_{a_1}(t+1)$ after the rejection, and a_2 also applies to f_2 because it is unmatched and $f_2 \in \mathcal{V}^+(t)$. Firm f_2 admits a_1 and rejects a_2, so $f_1 \in \mathcal{V}^+(t+1)$.*
- *At time $t+2$, by the same reasoning both agents apply to f_1 again, and the process repeats indefinitely.*

Thus, Example G.2 shows that even with perfectly learned preferences, Algorithm 3 with $k = 2$ can cycle forever. Such a cycle could be broken, for instance, by allowing agents to deviate from Algorithm 3 with some probability and stick with $f_a^{\texttt{match}}(t-1)$. Instead, we adopt a deterministic

remedy by enlarging each agent's choice set: we allow agents to apply among multiple interviewed firms, including the firm they were matched to in the previous round. Concretely, we add a third option by requiring each agent to always include $f_a^{\texttt{match}}(t-1)$ in its interview set. Accordingly, in our analysis we treat agents as also reconsidering $f_a^{\texttt{match}}(t-1)$ at every round.

We then extend Algorithm 3 to $k=3$ by setting the interview set to include the current candidate, the previous match, and a round-robin firm, namely

$$\mathcal{I}_a(t) = \{f_a^{\texttt{apply}}(t), f_a^{\texttt{match}}(t-1), f_a^{\texttt{RR}}(t)\}.$$

This captures both remedies discussed above: a randomized deviation that lets the agent occasionally stick with $f_a^{\texttt{match}}(t-1)$, and a deterministic enlargement of the agent's option set. In contrast to the $k=2$ version—where the agent effectively has a single action, applying to $f_a^{\texttt{apply}}(t)$—under $k=3$ the agent interviews three firms and may apply to either $f_a(t)$ or $f_a^{\texttt{match}}(t-1)$, choosing between them based on the updated information revealed by $\mathcal{I}_a(t)$ to choose $f_a^{\texttt{apply}}(t)$.

G.5.3 Necessity of applying to more than one firm per round for coordination-free decentralized learning

We explain why agents must be allowed to apply to *more than one* firm per round (i.e., apply to a set of interviewed firms), and why merely increasing the interview budget from $k=2$ to $k=3$ is still insufficient for convergence in general. The key obstacle is the *anonymous* nature of the firm-side feedback $\mathcal{V}^+$: observing that a firm becomes vacant does not reveal *which* agent left or whether the vacancy was caused by a strategic rejection. As a result, agents may repeatedly "chase" the same vacancy and create persistent application cycles, even when all estimated preference lists have already converged to the ground truth. The following example illustrates this phenomenon and shows how allowing an agent to apply to a *set* of firms in a round can eliminate such cycles by preventing spurious vacancies from appearing in $\mathcal{V}^+$.

Example G.3. *Consider a market with agents $\mathcal{A}=\{a_1,a_2\}$ and firms $\mathcal{F}=\{f_1,f_2\}$. Suppose that from some time t' onward the estimated preferences are exactly aligned with the ground truth and remain fixed, i.e., for all $t \geq t'$ we have $\hat{O}_a(t) = O_a$ for all $a \in \mathcal{A}$ and $\hat{O}_f(t) = O_f$ for all $f \in \mathcal{F}$. In particular, for every $t \geq t'$,*

$$\hat{O}_{a_1}(t) : (f_1 \succ f_2^*) \quad \hat{O}_{f_1}(t) : (a_2 \succ a_1)$$
$$\hat{O}_{a_2}(t) : (f_2 \succ f_1^*) \quad \hat{O}_{f_2}(t) : (a_1 \succ a_2).$$

Here, the * superscripts indicate the firms $f_{a_i}^{\texttt{apply}}(t)$ chosen by agents under Algorithm 3. We also assume that f_2 rejected a_2 at some earlier time, so f_2 remains in a_2's candidate set until a_2 applies to it again.

Fix any round $t \geq t'$. Consider the following two cases.

1. ***Applying to a single firm can create a cycle.*** *Suppose that at time $t+1$ agent a_1 interviews $\mathcal{I}_{a_1}(t+1) = \{f_1, f_2, f_{a_1}^{\texttt{RR}}(t+1)\}$ and applies only to f_1, i.e., $f_{a_1}^{\texttt{apply}}(t+1) = f_1$, while agent a_2 keeps applying to f_2. Since f_1 prefers a_2 to a_1, firm f_1 rejects a_1, and hence f_2 is reported in $\mathcal{V}^+(t+1)$ as a vacant firm. Because $\mathcal{V}^+(t+1)$ is anonymous, agent a_2 cannot distinguish whether f_2 appeared due to a strategic rejection or because its current match applied elsewhere; consequently, a_2 must keep f_2 in its candidate set and eventually applies to f_2 again. This induces a back-and-forth dynamic as in Example G.2, yielding an application cycle that can persist indefinitely despite perfectly learned preferences.*

2. ***Applying to a set breaks the cycle.*** *Suppose instead that agents may apply to a* set *of firms in a round. In the same situation at time $t+1$, let a_1 apply to both firms, i.e., $f_{a_1}^{\texttt{apply}}(t+1) = \{f_1, f_2\}$. Even if f_1 rejects a_1, the simultaneous application to f_2 allows a_1 to remain (or become) matched to f_2 in the same round. Consequently, f_2 does* not *appear in $\mathcal{V}^+(t+1)$, so a_2 receives no spurious vacancy signal and is not pulled into chasing f_2 again.*

In summary, under anonymous firm-side feedback, allowing an agent to apply to multiple firms in a round can suppress spurious vacancy signals and thereby eliminate persistent application cycles that would otherwise lead to linear regret. .

Algorithm 9 Extended Decentralized Learning with $\mathcal{V}^+(t)$ (Agent a)

Input: a, $\mathcal{F}$, anonymous hiring changes $\mathcal{V}^+$
Initialize: $r_{a,f}(1) \leftarrow 0$ for all $f \in \mathcal{F}$
for $t \in \mathcal{T}$ **do**
    Construct $\mathcal{B}'_a(t)$ according to (31)
    $f_a(t) \leftarrow \arg\max_{f \in \mathcal{B}'_a(t)} \hat{u}_{a,f}(t)$
    $\mathcal{I}_a(t) \leftarrow \{f_a(t), f_a^{\texttt{match}}(t-1), f_a^{\texttt{RR}}(t)\}$ and interview

$$f_a^{\texttt{apply}}(t) = \begin{cases} \{f_a(t), f_a^{\texttt{match}}(t-1)\}, & \text{with probability } \lambda, \\ \{f_a^{\texttt{match}}(t-1)\}, & \text{with probability } 1-\lambda \end{cases}$$

    Apply to all firms in $f_a^{\texttt{apply}}(t)$; accept an offer from $f_a(t)$ if admitted, else accept an offer from $f_a^{\texttt{match}}(t-1)$ if admitted; otherwise remain unmatched
    UPDATEAGENTREJVARS(a, t) (with the two-firm convention of Section G.5.4)

G.5.4 Extended Algorithm 3

With $k = 3$, each agent a forms the interview set as

$$\mathcal{I}_a(t) = \{f_a(t), f_a^{\texttt{match}}(t-1), f_a^{\texttt{RR}}(t)\},$$

where $f_a(t)$ is chosen exactly as in the $k = 2$ version of Algorithm 3, namely

$$f_a(t) = \arg\max_{f \in \mathcal{B}'_a(t)} \hat{u}_{a,f}(t).$$

The candidate set $\mathcal{B}'_a(t)$ is defined in (31) as

$$\mathcal{B}'_a(t) \doteq \{\, f : r_{a,f}(t) = 0 \ \vee \ \exists\, t' \in [r_{a,f}(t), t),\ f \in \mathcal{V}^+(t') \,\}.$$

Under this extended algorithm, each agent a maintains a fixed probability parameter $\lambda \in (0, 1)$ and, at the end of each round, updates its application as follows:

$$f_a^{\texttt{apply}}(t) = \begin{cases} \{f_a(t), f_a^{\texttt{match}}(t-1)\}, & \text{with probability } \lambda, \\ \{f_a^{\texttt{match}}(t-1)\}, & \text{with probability } 1-\lambda. \end{cases}$$

When $f_a^{\texttt{apply}}(t)$ is a singleton, the application dynamics coincide with the original Algorithm 3: agent a applies only to that firm and is matched according to the firm's accept/reject decision. When $f_a^{\texttt{apply}}(t)$ is a set, the agent applies to *all* firms in $f_a^{\texttt{apply}}(t)$ simultaneously. The agent then prioritizes $f_a(t)$: if $f_a(t)$ admits a, then a accepts that offer and is matched to $f_a(t)$ for that round; otherwise, a attempts to remain matched with $f_a^{\texttt{match}}(t-1)$. If $f_a^{\texttt{match}}(t-1)$ also rejects a, then a remains unmatched in that round. If a is unmatched at round $t-1$, i.e., $f_a^{\texttt{match}}(t-1) = \emptyset$, then $\{f_a^{\texttt{match}}(t-1)\}$ is empty and a does not apply in that round; in particular, $f_a^{\texttt{match}}(0) = \emptyset$ at $t = 1$.

Finally, the rejection variables are updated exactly as before: whenever agent a is rejected by a firm f in a *non-strategic* manner, the rejection indicator $r_{a,f}(\cdot)$ is updated accordingly (whereas strategic rejections are treated separately and do not trigger the same update). In particular, when a applies to $\{f_a(t), f_a^{\texttt{match}}(t-1)\}$ and is rejected by $f_a(t)$ while remaining matched with $f_a^{\texttt{match}}(t-1)$, this counts as a non-strategic rejection by $f_a(t)$, and $r_{a,f_a(t)}$ is updated. On the firm side, a firm whose admitted applicant accepts another offer has no hire in that round, and c_f is updated as in (7). The following Algorithm 9 presents the pseudocode for this extended procedure.

G.5.5 Crucial Theorems, Lemmas, and Observations

We begin by defining the notion of an *agent-consistent* (a.k.a. player-consistent) blocking pair, adapted from [2, 19].

Definition G.4 (Agent-consistent blocking pair)**.** A blocking pair (a_i, f_j) in a matching μ is *player-consistent* if $f_j \underset{O_{a_i}}{>} f$ for every firm f such that (a_i, f) is a blocking pair in μ.

We next recall a crucial result of [2] (stated here for completeness) showing that, under fixed preferences, repeatedly resolving blocking pairs in a player-consistent order reaches a stable matching.

Theorem G.5 ([2])**.** *Given any unstable matching $\mathcal{G}$, there exists a sequence of blocking pairs of length at most n^4 such that resolving this sequence reaches a stable matching. Moreover, this sequence can be chosen to resolve blocking pairs in a player-consistent order, i.e., every blocking pair (a_i, f_j) resolved in the current matching μ is player-consistent with respect to μ.*

We now present Lemma G.6, which bounds the expected maximum number of rounds for which the matching $(a, f_a^{\texttt{match}}(t))_{a\in\mathcal{A}}$ remains unstable within any interval $[t_1 : t_2]$ such that $t \in \Gamma_{\mathcal{A}}^{(m)} \cap \Gamma_{\mathcal{F}}^{(n)}$ for all $t \in [t_1 : t_2]$ under Algorithm 9. By Definition B.4, these are precisely the time steps at which the estimated preference lists of all agents and firms coincide with the ground truth. This lemma plays a central role in the proof of Theorem 4.2.

Lemma G.6. *In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$ with strategic firms following Algorithm 1 and agents following Algorithm 9, consider any interval $[t_1 : t_2]$ such that $t \in \Gamma_{\mathcal{A}}^{(m)} \cap \Gamma_{\mathcal{F}}^{(n)}$ for all $t \in [t_1 : t_2]$. Then, for any $\lambda \in (0, 1)$,*

$$\mathbb{E}\left[\sum_{t\in[t_1:t_2]} \mathbb{1}\left\{(a, f_a^{\texttt{match}}(t))_{\forall a\in\mathcal{A}} \textit{ is not stable}\right\}\right] \le 1 + \frac{n^4m + nm}{\left(\lambda\cdot(1-\lambda)^{n-1}\right)^{n^4m+nm}}.$$

Proof. We first account for rounds in $[t_1 : t_2]$ in which a firm-side strategic rejection can make the realized matching $(a, f_a^{\texttt{match}}(t))_{a\in\mathcal{A}}$ unstable, regardless of the stability of the intended matching $(a, f_a^{\texttt{apply}}(t))_{a\in\mathcal{A}}$. We claim that for each firm $f \in \mathcal{F}$, at most one such rejection can occur over $[t_1 : t_2]$, and it can only occur at the first round t_1. Indeed, by the definition of $\Gamma_{\mathcal{F}}^{(n)}$, for every $t \in [t_1 : t_2]$ each firm's estimated list coincides with its ground-truth list, so a firm never rejects in order to hire an agent that is worse under the ground truth. Hence, whenever the hiring state of f changes during $[t_1 : t_2]$, it either (i) hires an agent it truly prefers to its current match, or (ii) becomes vacant. Therefore, under Algorithm 1 the only round at which a firm may strategically reject is t_1, which explains the additive 1 term in the lemma statement.

It remains to bound

$$\mathbb{E}\left[\sum_{t\in[t_1+1:t_2]} \mathbb{1}\left\{(a, f_a^{\texttt{match}}(t))_{\forall a\in\mathcal{A}} \text{ is not stable}\right\}\right].$$

Fix any $t \in [t_1 + 1 : t_2]$ and let $\mathcal{G}(t) = (a, f_a^{\texttt{match}}(t))_{a\in\mathcal{A}}$ be the realized matching at time t. By Theorem G.5, if $\mathcal{G}(t)$ is unstable then there exists a sequence $S = \{s_1, \ldots, s_{n^4}\}$ of agent-consistent blocking pairs such that resolving these pairs in order reaches a stable matching.

We lower bound the probability of resolving this sequence by exhibiting a sufficient event for resolving each blocking pair. Assume inductively that $s_1, \ldots, s_{i-1}$ have already been resolved, and let $s_i = (a', f'')$. Define $\mathcal{X}_i$ as the event that, for at most m consecutive rounds after $s_1, \ldots, s_{i-1}$ are resolved, agent a' chooses the "move" option (hence applies according to $f_{a'}(\cdot)$) in every round, while every other agent $a \neq a'$ chooses the "stay" option (hence applies to $f_a^{\texttt{match}}(\cdot - 1)$) in every round. Conditioned on $\mathcal{X}_i$, the agent a' keeps advancing through its candidate set and, in the worst case, reaches and applies to f'' within m rounds; when a' applies to f'', the agent-consistency of s_i implies that f'' admits a' and rejects its current match, thereby resolving the blocking pair.

By independence of agents' randomized choices across rounds,

$$\Pr(\mathcal{X}_i) \;\ge\; \left(\lambda\cdot(1-\lambda)^{n-1}\right)^m.$$

Let

$$\mathcal{X} \;\doteq\; \bigcap_{i=1}^{n^4} \mathcal{X}_i.$$

By construction, $\mathcal{X}$ implies that all blocking pairs in S are resolved in order, and hence the matching becomes stable after at most n^4m rounds. Moreover,

$$\Pr(\mathcal{X}) \;\ge\; \left(\lambda\cdot(1-\lambda)^{n-1}\right)^{n^4m}.$$

After $\mathcal{X}$ occurs, there exists a time t' at which the matching has no blocking pair. However, due to the anonymous revelation of hiring changes in $\mathcal{V}^+$, an agent may still have firms in its candidate set that it prefers to its current match $f_a^{\mathtt{match}}(t')$. Since there is no blocking pair, any such application is rejected provided the other agents remain with their stable matches. Under Algorithm 9, when an agent applies to a preferred firm while also applying to its current match $f_a^{\mathtt{match}}(t')$, it remains matched with $f_a^{\mathtt{match}}(t')$ upon rejection, and neither firm necessarily appears in $\mathcal{V}^+$. Thus, it may take up to an additional nm rounds for all agents to exhaust such futile applications and reach a state in which they keep applying to (and remaining matched with) the stable matching.

Formally, define $\mathcal{J}$ as the event that, after $\mathcal{X}$ occurs, within the next nm rounds the process reaches a time τ such that $\mathcal{G}(\tau)$ is stable and $\mathcal{G}(t) = \mathcal{G}(\tau)$ for all subsequent rounds in the interval. By the same independence argument,

$$\Pr(\mathcal{J} \mid \mathcal{X}) \;\geq\; \big(\lambda \cdot (1-\lambda)^{n-1}\big)^{nm}.$$

Therefore,

$$\Pr(\mathcal{X} \cap \mathcal{J}) \;=\; \Pr(\mathcal{J} \mid \mathcal{X})\,\Pr(\mathcal{X}) \;\geq\; \big(\lambda \cdot (1-\lambda)^{n-1}\big)^{n^4 m + nm}.$$

Hence, with probability at least $p \doteq \big(\lambda \cdot (1-\lambda)^{n-1}\big)^{n^4 m + nm}$, within the next $n^4 m + nm$ rounds the process reaches a stable matching that remains unchanged thereafter. Viewing this as a Bernoulli trial with success probability p, the expected number of rounds (starting from $t_1 + 1$) until such a stable-and-absorbing state is reached is at most $1/p$. Consequently,

$$\mathbb{E}\left[\sum_{t \in [t_1 : t_2]} \mathbb{1}\left\{(a, f_a^{\mathtt{match}}(t))_{\forall a \in \mathcal{A}} \text{ is not stable}\right\}\right] \leq 1 + \frac{n^4 m + nm}{\big(\lambda \cdot (1-\lambda)^{n-1}\big)^{n^4 m + nm}}.$$

□

G.5.6 Proof of Theorem 4.2 for General Markets

Theorem 4.2. *In a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, with firms following Algorithm 1 and agents following coordination-free Algorithm 3 (for α-reducible markets with $k = 2$) and Algorithm 9 (for general markets with $k = 3$, where an agent may apply to two interviewed firms per round), the expected regret is time-independent. In particular,*

$$\mathbb{E}[\underline{R}_a(\mathcal{T})] \in \begin{cases} O\big((\sum_{j \in [i]} |\mathcal{H}_{a_j, f_j}| + 1) \cdot (\sum_{j \in [i]} (m^2 \overline{\Delta}_{a_j}^{-2} + nm\, \underline{\Delta}_{f_j}^{-2}) + 1)\big), & \alpha\text{-reducible } \mathcal{M}(\mathcal{A}, \mathcal{F}), \\ O\big(\epsilon^{-1} n^5 m^4 \Delta^{-2}\big)\text{, for } \epsilon = \big(\lambda(1-\lambda)^{n-1}\big)^{n^4 m + nm} \text{ with } \lambda \in (0,1), & \text{General Markets}. \end{cases}$$

Proof. Since we aim to bound the *pessimal regret* $\mathbb{E}[\underline{\mathrm{R}}_a(T)]$, it suffices to measure regret with respect to agent a's pessimal stable match, denoted by $\underline{f_a^*}$. Accordingly, throughout the proof we measure regret relative to $\underline{f_a^*}$. Once the algorithm converges to a stable matching, agent a is matched either to $\underline{f_a^*}$ or to a strictly better stable match; therefore, all subsequent rounds incur zero (or negative) regret relative to $\underline{f_a^*}$. Hence, it suffices to bound the regret incurred prior to convergence in order to bound $\mathbb{E}[\underline{\mathrm{R}}_a(T)]$. We also denote by $\overline{a_f^*}$ the firm-optimal stable match of firm f.

Next, with respect to Definition B.4, consider the sets of time steps $\Gamma_{\mathcal{A}}^{(m)}$ and $\Gamma_{\mathcal{F}}^{(n)}$ in which all agents' and all firms' estimated preference lists, respectively, coincide with the ground truth. We first account for regret incurred on the complement rounds $\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}$, and then bound the maximum regret within each maximal interval $[t_i : t_i']$ such that every $t \in [t_i : t_i']$ satisfies $t \in \Gamma_{\mathcal{A}}^{(m)} \cap \Gamma_{\mathcal{F}}^{(n)}$.

Let $\mathcal{Y}$ denote the set of starting time steps of these maximal aligned intervals. Formally,

$$\mathcal{Y} \doteq \left\{ t \in \mathcal{T} : t \in \Gamma_{\mathcal{A}}^{(m)} \cap \Gamma_{\mathcal{F}}^{(n)} \text{ and } \big(t = 1 \text{ or } t - 1 \notin \Gamma_{\mathcal{A}}^{(m)} \cap \Gamma_{\mathcal{F}}^{(n)}\big) \right\}.$$

For each $t_i \in \mathcal{Y}$, let

$$t_i' \doteq \max\left\{ t' \in [T] : [t_i : t'] \subseteq \Gamma_{\mathcal{A}}^{(m)} \cap \Gamma_{\mathcal{F}}^{(n)} \right\}$$

denote the corresponding end time of the maximal interval starting at t_i. We also denote by

$$B \doteq \max_{[t_1, t_2] \subseteq \Gamma_{\mathcal{A}}^{(m)} \cap \Gamma_{\mathcal{F}}^{(n)}} \mathbb{E}\left[\sum_{t \in [t_1, t_2]} \mathbb{1}\left\{(a', f_{a'}^{\mathtt{match}}(t))_{\forall a' \in \mathcal{A}} \text{ is not stable}\right\}\right]$$

the largest expected number of unstable rounds over deterministic intervals $[t_1, t_2]$ satisfying $t'' \in \Gamma_{\mathcal{A}}^{(m)} \cap \Gamma_{\mathcal{F}}^{(n)}$ for all $t'' \in [t_1, t_2]$. By Lemma G.6, we obtain

$$B \le 1 + \frac{n^4 m + nm}{\left(\lambda \cdot (1-\lambda)^{n-1}\right)^{n^4 m + nm}}. \tag{36}$$

$$\begin{aligned}
\mathbb{E}[\underline{\mathbf{R}}_a(T)] &= \mathbb{E}\left[\underline{\mathbf{R}}_a\left(\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}\right)\right] + \mathbb{E}\left[\underline{\mathbf{R}}_a\left(\Gamma_{\mathcal{A}}^{(m)} \cap \Gamma_{\mathcal{F}}^{(n)}\right)\right] \\
&= \mathbb{E}\left[\underline{\mathbf{R}}_a\left(\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}\right)\right] + \sum_{t_i \in \mathcal{Y}} \mathbb{E}[\underline{\mathbf{R}}_a([t_i, t_i'])] \\
&\le \mathbb{E}\left[\left|\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}\right|\right] + \mathbb{E}\left[\sum_{t_i \in \mathcal{Y}} \underline{\mathbf{R}}_a([t_i, t_i'])\right] \\
&\le \mathbb{E}\left[\left|\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}\right|\right] + \mathbb{E}\left[\sum_{t_i \in \mathcal{Y}} \sum_{t \in [t_i, t_i']} \mathbb{1}\left\{f_a^{\texttt{match}}(t) \in \emptyset \cup \mathcal{L}_{a, \underline{f_a^*}}\right\}\right] \\
&\le \mathbb{E}\left[\left|\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}\right|\right] + \mathbb{E}\left[\sum_{t_i \in \mathcal{Y}} \sum_{t \in [t_i, t_i']} \mathbb{1}\left\{(a', f_{a'}^{\texttt{match}}(t))_{\forall a' \in \mathcal{A}} \text{ is not stable}\right\}\right] \\
&= \mathbb{E}\left[\left|\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}\right|\right] + \mathbb{E}\left[\sum_{t_i \in \mathcal{Y}} \mathbb{E}\left[\sum_{t \in [t_i, t_i']} \mathbb{1}\left\{(a', f_{a'}^{\texttt{match}}(t))_{\forall a' \in \mathcal{A}} \text{ is not stable}\right\} \,\middle|\, \mathcal{Y}, \{[t_i, t_i']\}\right]\right] \\
&\overset{(a)}{\le} \mathbb{E}\left[\left|\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}\right|\right] + \mathbb{E}[|\mathcal{Y}|] \cdot B \\
&\overset{(b)}{\le} \mathbb{E}\left[\left|\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}\right|\right] + \left(\mathbb{E}\left[\left|\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}\right|\right] + 1\right) \cdot B.
\end{aligned} \tag{37}$$

Here, the equality couples each maximal aligned interval with its unique starting time step $t_i \in \mathcal{Y}$, so the sum captures all aligned-round regret exactly, and applies the tower rule conditioning on $\mathcal{Y}$ and the interval endpoints. Inequality (a) bounds each conditional expectation by B, which is uniform over the time step and the initial matching by Lemma G.6; since the maximum is taken outside the expectation, no independence between $|\mathcal{Y}|$ and the interval lengths is required. Inequality (b) follows since each maximal interval $[t_i : t_i'] \subseteq \Gamma_{\mathcal{A}}^{(m)} \cap \Gamma_{\mathcal{F}}^{(n)}$ has, whenever defined, its boundary time steps $t_i - 1$ and $t_i' + 1$ belonging to $\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}$ (and only one boundary exists when $t_i = 1$ or $t_i' = T$). Hence, the number of such maximal intervals is at most $\left|\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}\right| + 1$.

It remains to bound $\mathbb{E}\left[\left|\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}\right|\right]$. We write

$$\begin{aligned}
\mathbb{E}\left[\left|\bar{\Gamma}_{\mathcal{A}}^{(m)} \cup \bar{\Gamma}_{\mathcal{F}}^{(n)}\right|\right] &\le \mathbb{E}\left[\left|\bar{\Gamma}_{\mathcal{A}}^{(m)}\right|\right] + \mathbb{E}\left[\left|\bar{\Gamma}_{\mathcal{F}}^{(n)}\right|\right] \\
&\le \mathbb{E}\left[\left|\left\{t \in \mathcal{T} : \exists a \in \mathcal{A}\ \exists f \in \mathcal{F},\ t \in \bar{\mathcal{E}}_{a,f}\right\}\right|\right] + \mathbb{E}\left[\left|\left\{t \in \mathcal{T} : \exists a' \in \mathcal{A}\ \exists f' \in \mathcal{F},\ t \in \bar{\mathcal{E}}_{f',a'}\right\}\right|\right] \\
&\le \sum_{a \in \mathcal{A}} \sum_{f \in \mathcal{F}} \frac{4|\mathcal{L}_{a,f}| \exp\left(-\frac{\Delta^2}{2m}\right)}{1 - \exp\left(-\frac{\Delta^2}{2m}\right)} + \frac{4|\mathcal{L}_{f,a}| \exp\left(-\frac{\Delta^2}{2m}\right)}{1 - \exp\left(-\frac{\Delta^2}{2m}\right)} \\
&\le \sum_{a \in \mathcal{A}} \sum_{f \in \mathcal{F}} O\left(m(|\mathcal{L}_{a,f}| + |\mathcal{L}_{f,a}|)\, \Delta^{-2}\right) \\
&\in O\left(nm^3 \Delta^{-2}\right),
\end{aligned} \tag{38}$$

where the last inequality follows by Lemma 3.3.

Finally, combining (37), (36), and (38), we obtain

$$\mathbb{E}[\underline{\mathbf{R}}_a(T)] \in O\left(\frac{n^5 m^4 \Delta^{-2}}{\left(\lambda \cdot (1-\lambda)^{n-1}\right)^{n^4 m + nm}}\right).$$

□

H Discussion on Incentive Compatibility in Decentralized Learning

We study double-sided learning in a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$, where both agents and firms update their behavior over time based on their own observations. Since stability is a shared objective on both sides, a natural question is whether the decentralized dynamics we propose are *incentive compatible* in the following sense: if all other players follow a fixed prescribed algorithm, does any single agent or firm have an incentive to deviate in order to obtain a more favorable stable outcome?

Concretely, we consider agents following the decentralized Algorithms 7, 3, and 9, together with firms implementing the strategic rejection policy in Algorithm 1. In markets that may admit multiple stable matchings, where agent-optimal/agent-pessimal and firm-optimal/firm-pessimal stable matches are well-defined, we ask whether following these algorithms—given that others do the same—leads each side to converge to the stable matching that is optimal (or pessimal) from its own perspective.

We next give a general notion of *incentive compatibility* for a player on either side of the market, defined relative to fixed algorithms adopted by all other players, and under the informational constraints of our learning model.

Definition H.1 (Incentive compatibility under partial information)**.** Consider a matching market $\mathcal{M}(\mathcal{A}, \mathcal{F})$ in which the underlying (ground-truth) preference profile is unknown to all players. At each time step, each agent and firm only maintains its own estimated preference list (formed from its observed samples), and players observe outcomes only through the firm-side feedback stream (e.g., $\mathcal{V}(t)$ or $\mathcal{V}^+(t)$, depending on the model). Fix a pair of policies $(\pi_{\mathcal{A}}, \pi_{\mathcal{F}})$, where $\pi_{\mathcal{A}}$ specifies the agents' learning policy and $\pi_{\mathcal{F}}$ specifies the firms' rejection (and, if applicable, learning) policy. For a side $S \in \{\mathcal{A}, \mathcal{F}\}$, we say that policy π_S is *incentive compatible* for side S (relative to π_{-S}) if, under these information constraints, when all players on side S follow π_S and all players on the opposite side follow π_{-S}, the induced dynamics converge to the S-optimal stable matching (equivalently, the $(-S)$-pessimal stable matching).

We now discuss the incentive compatibility of agents and firms under our decentralized learning dynamics, distinguishing between markets with a unique stable matching and markets with multiple stable matchings.

H.1 Incentive compatibility in matching markets with a unique stable matching

When $\mathcal{M}(\mathcal{A}, \mathcal{F})$ admits a *unique* stable matching, incentive considerations are comparatively clean: convergence to a stable outcome uniquely determines the optimal stable outcome for both sides. We formalize this under our decentralized information structure and show that, conditional on convergence, the prescribed agent- and firm-side policies are incentive compatible in the sense of Definition H.1.

Proposition H.2 (Incentive compatibility under a unique stable matching)**.** *Suppose $\mathcal{M}(\mathcal{A}, \mathcal{F})$ admits a unique stable matching $\mu^\star$ under the ground-truth preferences (this includes α-reducible markets). Assume the decentralized information structure where each player knows only its own estimated preference list and observes only the firm-side feedback. Fix any agent learning policy $\pi_1 \in \{$Algorithms 7, 3, 9$\}$ and the firm rejection policy $\pi_2 =$ Algorithm 1. If, when all agents follow π_1 and all firms follow π_2, the induced dynamics converge to a stable matching, then they converge to $\mu^\star$, and π_1 is incentive compatible for agents and π_2 is incentive compatible for firms with respect to Definition H.1.*

Proof. Since $\mathcal{M}(\mathcal{A}, \mathcal{F})$ has a unique stable matching, every stable outcome equals $\mu^\star$. Hence, under (π_1, π_2), convergence to any stable matching implies convergence to $\mu^\star$. Because no alternative stable matching exists, no agent or firm can improve its stable outcome by deviating in order to steer the dynamics toward a different stable limit; deviations can only affect transient behavior (e.g., delaying coordination or prolonging suboptimal commitments). Therefore, following π_1 (resp., π_2) is incentive compatible for agents (resp., firms) in the sense of Definition H.1. □

H.2 Incentive compatibility in matching markets with multiple stable matchings

When $\mathcal{M}(\mathcal{A}, \mathcal{F})$ admits *multiple* stable matchings, incentive considerations become inherently subtle: a guarantee of convergence to *a* stable matching does not identify *which* stable outcome is selected,

and different stable matchings can yield different payoffs for both sides (the number of stable matchings may be exponential in n and m). Thus, a policy profile may be stability-preserving while still failing to be incentive compatible in the sense of Definition H.1, which singles out a particular stable outcome (agent-pessimal / firm-optimal in our definition).

Agents. In our decentralized learning model, agents are the proposing side and each player has only partial information (it knows only its own estimated preference list and observes only the firm-side feedback). Under this information structure, even if all agents follow the prescribed learning policy and firms follow the prescribed rejection policy, convergence need not be to the *agent-optimal* stable matching; see Example G.1. Therefore, in markets with multiple stable matchings, the prescribed agent learning policy is *not* incentive compatible for agents under Definition H.1: an agent whose objective is to secure its agent-optimal stable partner cannot generally view "following the algorithm" as guaranteeing convergence to that target stable outcome.

Firms. A similar selection issue arises on the firm side. Our firm-side rejection policy Algorithm 1 is designed to preserve stability under limited feedback by triggering corrective updating phases when a firm detects local inconsistencies; it is *not* designed to perform outcome-steering strategic rejections. Thus, in general markets with multiple stable matchings, convergence to *some* stable matching under Algorithm 1 does not imply convergence to the agent-pessimal (equivalently, firm-optimal) stable matching.

Importantly, even when firms are *certain*, the literature shows that strategic rejections can change which stable matching is selected and may improve firms' final matches; see, e.g., [14]. In particular, there exist instances where allowing firms to use strategic rejections leads to a stable outcome that is strictly better for all firms than the outcome obtained without such rejections. Since under certainty our policy does not employ such outcome-improving strategic rejections, a firm may have an incentive to deviate in markets with multiple stable matchings, in an attempt to steer the dynamics toward a more firm-preferred stable outcome. Consequently, without an additional guarantee that the induced dynamics converge to the firm-optimal stable matching, following Algorithm 1 is *not* incentive compatible for firms in the sense of Definition H.1.

Proposition H.3 (Lack of incentive compatibility under multiple stable matchings)**.** *Suppose $\mathcal{M}(\mathcal{A}, \mathcal{F})$ admits more than one stable matching under the ground-truth preferences, and consider the decentralized information structure where each player knows only its own estimated preference list and observes only the firm-side feedback. Fix any agent learning policy $\pi_1 \in \{$Algorithms 7, 3, 9$\}$ and the firm rejection policy $\pi_2 =$* Algorithm 1.

- *There exists an instance (e.g., Example G.1) in which the induced dynamics under (π_1, π_2) converge to a stable matching that is* not *agent-optimal, then π_1 is not incentive compatible for agents in the sense of Definition H.1.*
- *There exists an instance in which the induced dynamics under (π_1, π_2) converge to a stable matching that is* not *agent-pessimal (equivalently, not firm-optimal), then π_2 is not incentive compatible for firms in the sense of Definition H.1.*

Taken together with Proposition H.2, this highlights a sharp contrast: under a unique stable matching, convergence to stability pins down a unique long-run outcome and yields incentive compatibility for both sides, whereas under multiple stable matchings, stable-outcome selection becomes endogenous and the same decentralized policies need not be incentive compatible for *general* underlying markets.

I Optimality of the Results

In this section, we discuss the optimality of our regret guarantees for both the centralized and decentralized algorithms. Our goal is to position the obtained bounds relative to natural lower-bound benchmarks, and to clarify which gaps are inherent to the learning problem versus artifacts of coordination and information constraints. In particular, we compare our upper bounds to immediate information-theoretic baselines and outline the remaining gaps as concrete lower-bound questions.

I.1 Optimality of the Centralized Bounds Theorem 3.1

We prove that the regret guarantee of Algorithm 6 is near-optimal by establishing a matching lower bound (up to a factor m) that applies to any policy achieving stable outcomes.

Proposition I.1. *Algorithm 6's regret $O(nm^2 \cdot \min\{\overline{\Delta}_{\mathcal{A}}, \underline{\Delta}_{\mathcal{F}}\}^{-2})$ is within a factor m of the information-theoretic optimal regret.*

Proof. Consider any policy π under which the market converges to a stable matching with sublinear regret. To certify stability, each agent a must learn, for every firm f and competing agent a', whether f prefers a to a', since these comparisons determine blocking pairs. This information requires at least one round in which a applies to f, so any policy must incur $\Omega(nm)$ regret per agent over the $O(nm)$ such triplets (a, a', f). Since Algorithm 6 achieves $O(nm^2)$ regret, it is within a factor m of optimal. The $\min\{\overline{\Delta}_{\mathcal{A}}, \underline{\Delta}_{\mathcal{F}}\}^{-2}$ factor reflects the minimum number of samples needed to distinguish any two partners on the opposite side. □

By Theorem 3.1, our centralized algorithm Algorithm 6 achieves a time-independent regret bound of $O(nm^2)$. This guarantee is within a factor m of the immediate benchmark $O(nm)$, and Proposition I.1 shows that Algorithm 6 is nearly optimal under that benchmark.

We further conjecture that the $O(nm^2)$ rate is in fact asymptotically optimal for learning in this setting. Intuitively, learning ordinal preferences requires, for each of the n agents, distinguishing between $\Theta(m^2)$ firm pairs. In the worst case, this can demand a constant number of observations per pair (depending on the gaps between the firms' expected rewards), which suggests an inherent $\Omega(nm^2)$ difficulty.

Conjecture I.2. *The $O(nm^2)$ regret bound of the centralized Algorithm 6 is asymptotically optimal.*

I.2 Optimality of the Coordinated Decentralized Bounds Theorem 4.1

Compared to the centralized algorithm Algorithm 6, coordinated decentralized learning necessarily incurs an additional overhead due to distributed executions of `GS`. In the centralized setting, `GS` is effectively run in a single round at each time step, which underlies the time-independent $O(nm^2)$ regret guarantee of Theorem 3.1. By contrast, under the coordinated decentralized algorithm Algorithm 7, agents must coordinate to jointly run `GS`, and each such distributed execution can take up to $O(n^2)$ rounds, during which agents may incur regret.

Consequently, a direct comparison suggests an $O(n^2)$ multiplicative overhead relative to the centralized benchmark: combining the $O(n^2)$ duration of a distributed `GS` run with the $O(nm^2)$ learning cost yields an $O(n^3m^2)$ regret bound, matching the guarantee in α-reducible markets in Theorem 4.1. Under this interpretation, the coordinated decentralized guarantee remains nearly optimal up to the inherent coordination cost.

For general (unstructured) markets, Theorem 4.1 yields an $O(n^4m^2)$ bound, which is an additional factor n larger than the α-reducible guarantee. Equivalently, relative to the immediate $O(nm)$ benchmark (corresponding to the idealized centralized baseline that resolves matching each round), the general-market bound is larger by a factor $O(n^3m)$.

I.3 Optimality of the Coordination-Free Decentralized Bounds Theorem 4.2

As noted earlier, the richer firm-side feedback $\mathcal{V}^+$ enables the design of coordination-free decentralized learning. In structured markets (in particular, α-reducible markets), the market-dependent regret bound of Theorem 4.2 for Algorithm 3 is at most $O(n^3m^2)$ in the worst case, and therefore remains within an $O(n^2m)$ factor of the immediate benchmark $O(nm)$.

A key difference from the coordinated algorithm is that coordination-free executions do not incur a fixed $O(n^2)$ overhead from explicitly coordinating a distributed run of `GS`. Instead, the regret bound in Theorem 4.2 adapts to the intrinsic time complexity of `GS` under the underlying market structure: in markets where `GS` completes in $O(n)$ rounds, the same analysis yields an $O(n^2m^2)$ bound. Hence, for α-reducible markets, the coordination-free guarantees can be viewed as adaptively nearly optimal with respect to the number of rounds required by `GS` under the ground-truth preferences. This leads to a substantial improvement over the coordination-free approach of [21] in such structured markets.

For general (unstructured) markets, the extension Algorithm 9 admits a time-independent bound in Theorem 4.2 whose leading term is dominated by a λ-dependent constant that is exponential in n and m, namely $\frac{1}{\left(\lambda\cdot(1-\lambda)^{n-1}\right)^{n^4m+nm}}$. While this dependence is consistent with the phenomena highlighted in [19], it is far larger than the immediate lower bound $O(n^3m)$ for coordination-free decentralized learning. As discussed by [19, 2], such large constants are unavoidable in fully general markets under coordination-free information constraints as simultaneous blocking pair resolution may cause infinite cycles. Establishing tight lower bounds for coordination-free algorithms in unstructured matching markets therefore remains an interesting and challenging open problem.

J Related Work

We present here the main ways our results strengthen prior work: we obtain *time-independent* guarantees under *limited firm-side feedback*, while explicitly modeling *firm uncertainty* via a strategic action space and a rejection policy. This yields a more general framework for decentralized double-sided learning, and our firm-side rejection mechanism can be incorporated into existing learning-to-stability approaches to preserve convergence even when firms update their estimates over time.

J.1 Applicability of Strategic Firm's Rejection Policy Algorithm 1

A key conceptual contribution of our framework is the firm-side rejection policy Algorithm 1, which is designed to handle *firm uncertainty* while preserving convergence to stability under decentralized learning. The policy uses controlled (strategic) rejections as a mechanism for re-initiating correction phases exactly when a firm detects that its current local estimates may be inconsistent with the true preference ordering relevant to stability.

Beyond its role in our algorithms, Algorithm 1 can be viewed as a modular primitive that can be adapted to prior decentralized matching-bandit frameworks [19, 21, 30]. Many existing works establish convergence to a stable matching under the assumption that firms' preferences are fixed and known (or, more generally, that firms act as passive acceptors without learning). Under such assumptions, stability is guaranteed only because the firm side introduces no additional uncertainty into the dynamics. Our rejection policy provides a principled way to relax this restriction: it allows firms to be uncertain and to learn their own preferences, while still enabling the overall process to converge to a stable outcome by triggering corrective updates whenever firm-side estimation errors would otherwise derail stability.

To our knowledge, this use of strategic firm-side rejections specifically to incorporate *learning and uncertainty on the firm side* while maintaining convergence-to-stability guarantees is first introduced in this work. Consequently, incorporating Algorithm 1 into prior approaches offers a direct path to strengthening their results by extending them from settings with fully known firm preferences to more realistic environments in which firms are uncertain and must learn.

J.2 Improvements with Limited Firm-Side Feedback

The firm-side feedback model studied here, together with our treatment of firm uncertainty via an extended firm action space, yields substantial improvements over the guarantees in [19]. In particular, their model assumes that firms have fixed preference lists that are known to both firms and agents throughout the learning process, which is a strong and often unrealistic assumption. They also assume a much richer feedback structure in which agents observe not only vacancies and hiring changes, but also the identity of each firm's current match at the end of every round.

In contrast, our bounds are obtained under anonymous firm-side feedback and while allowing firms to be both strategic and uncertain. Taken together, these features lead to nearly optimal regret guarantees under significantly weaker informational assumptions, improving upon [19] along multiple dimensions.

J.3 Model and Result Differences with Recent Work with Two-Sided Uncertainty

An ETC-then-GS algorithm for two-sided online learning was proposed in [23]. Although their ETC algorithmic structure is similar to ours, our problem setup differs from theirs in several key respects:

(a) firms in our setting can strategically defer hiring, which complicates the matching process; (b) agents can coordinate through mutual signaling; and (c) agents undergo an interview phase prior to matching. These distinctions necessitate fundamentally different algorithmic techniques, and our algorithms further achieve horizon-independent regret, substantially improving upon their logarithmic bounds.

K Future Directions and Open Questions

Having detailed our improvements over prior work, we now outline several future directions that remain as interesting open problems.

• **Firm-side incentives beyond the unique-stable regime.** Our analysis focuses on agent regret and introduces the firm-side rejection policy Algorithm 1, which guarantees convergence to stability under limited firm-side feedback. However, as discussed in Sections H.1 and H.2, this policy is incentive compatible for firms only in markets with a unique stable matching (cf. Proposition H.2), while incentive alignment becomes subtle when multiple stable matchings exist. A natural open question is to design *alternative* firm-side rejection policies that (i) preserve convergence to stability under the same decentralized information structure (each firm knows only its own estimated preference list and observes the firm-side feedback) and (ii) are more closely aligned with firm objectives in multi-stable markets, i.e., can bias equilibrium selection toward firm-preferred stable outcomes. This direction connects to the influence of firm-side strategic rejections in markets with certain firms (e.g., [14]) and raises the challenge of extending such outcome-improvement phenomena to an *online* learning-to-stability setting, potentially also under our extended firm action space that models uncertainty.

• **Reward-adaptive application decisions and stability notions.** In our model, agents choose whom to apply/interview based on their current estimates (e.g., empirical means) *before* observing the realized interview reward in that round. An interesting extension is a *reward-sensitive* variant in which an agent can condition its final application decision on the realized interview outcomes within the same round (e.g., after interviewing a subset of firms). This change can reduce regret by enabling within-round adaptation, but it also suggests stronger, reward-contingent notions of stability and benchmarking: the relevant comparison may no longer be regret to a fixed stable matching (agent-optimal/pessimal), but to an outcome concept that accounts for the additional adaptivity enabled by realized rewards. Analyzing this setting would likely require tracking not only mean gaps but also distributional properties (e.g., variance or tail behavior), and may call for estimators and concentration tools beyond the empirical-mean-based arguments used throughout our paper.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main claims of the paper, which are supported by the theoretical results and analyses presented in the subsequent sections. The claims are consistent with the contributions outlined in the paper, and any assumptions or limitations are also acknowledged.

   Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a discussion of the limitations in the conclusion and Appendix K.

   Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper provides all necessary assumptions and complete proofs (in appendix) for the theoretical results.

   Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification:

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification:

   Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

• The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
• The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
• The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
• At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
• Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [N/A]

Justification:

Guidelines:

• The answer [N/A] means that the paper does not include experiments.
• The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
• The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification:

Guidelines:

• The answer [N/A] means that the paper does not include experiments.
• The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
• The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
• The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
• The assumptions made should be given (e.g., Normally distributed errors).
• It should be clear whether the error bar is the standard deviation or the standard error of the mean.
• It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
• For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
• If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [N/A]

Justification: The paper is mainly theoretical and does not include large-scale experiments.

Guidelines:

• The answer [N/A] means that the paper does not include experiments.
• The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper adheres to the NeurIPS Code of Ethics.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is mainly theoretical and does not have direct societal impacts.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper is mainly theoretical and does not pose such risks.

    Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

• We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper is mainly theoretical and does not use existing assets.

Guidelines:

• The answer [N/A] means that the paper does not use existing assets.
• The authors should cite the original paper that produced the code package or dataset.
• The authors should state which version of the asset is used and, if possible, include a URL.
• The name of the license (e.g., CC-BY 4.0) should be included for each asset.
• For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
• If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
• For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
• If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper is mainly theoretical and does not release new assets.

Guidelines:

• The answer [N/A] means that the paper does not release new assets.
• Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
• The paper should discuss whether and how consent was obtained from people whose asset is used.
• At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper is mainly theoretical and does not involve crowdsourcing nor research with human subjects.

Guidelines:

• The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
• Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
• According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper is mainly theoretical and does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The paper is mainly theoretical and does not involve LLMs usage beyond writing, editing, or formatting purposes.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.