Bandit Learning in Housing Markets

Shiyun Lin *

November, 2025

Abstract

The housing market, also known as one-sided matching market, is a classic exchange economy model where each agent on the demand side initially owns an indivisible good (a house) and has a personal preference over all goods. The goal is to find a core-stable allocation that exhausts all mutually beneficial exchanges among subgroups of agents. While this model has been extensively studied in economics and computer science due to its broad applications, little attention has been paid to settings where preferences are unknown and must be learned through repeated interactions. In this paper, we propose a statistical learning model within the multi-player multi-armed bandit framework, where players (agents) learn their preferences over arms (goods) from stochastic rewards. We introduce the notion of *core regret* for each player as the market objective. We study both centralized and decentralized approaches, proving $\mathcal{O}(N \log T/\Delta^2)$ upper bounds on regret, where N is the number of players, T is the time horizon and Δ is the minimum preference gap among players. For the decentralized setting, we also establish a matching lower bound, demonstrating that our algorithm is order-optimal.

1 Introduction

The housing market, also known as one-sided matching market, represents a foundational economic institution where agents exchange property rights to achieve more desirable arrangements. Classical research in this domain traces back to the seminal work of Shapley and Scarf [1974], and the model has been extensively studied in the literature [Roth, 2023] due to its wide range of applications such as student housing assignments [Sönmez and Ünver, 2010], school choice [Abdulkadiroğlu and Sönmez, 2003] and kidney exchange [Roth et al., 2004]. In a housing market, each agent starts with an indivisible good – typically representing a house – and has their own personal preferences over all available goods in the market. Unlike traditional markets with buyers and sellers, here, agents trade their initial endowments based on their preferences, seeking to obtain a house they value more, and the houses must be traded without monetary transactions. The key challenge is to find a stable and fair allocation where no agent can improve their outcome by trading with someone else. The Top Trading Cycle (TTC) algorithm introduced by Shapley and Scarf [1974] is a celebrated mechanism that ensures efficient, strategy-proof and Pareto optimal outcomes [Roth and Postlewaite, 1977]. Subsequent research has expanded this framework to accommodate different settings and provide various solutions to the problem [Hylland and Zeckhauser, 1979, Echenique et al., 2021, Garg et al., 2024]. Despite these advancements, the integration of adaptive learning mechanisms into housing market models remains an underexplored area, particularly in online environments where participants must navigate uncertainty in preferences.

*Center for Statistical Science, School of Mathematical Sciences, Peking University

The assumptions of a known, full preference profile in housing markets is often unrealistic. In practice, participants – such as tenants on LeaseSwap NYC, patients in kidney exchange, or traders in NFT markets – frequently lack well-defined preferences over items they have not experienced. Fortunately, these settings typically allow for repeated short-term interactions that yield immediate feedback, for example, through home visits, medical compatibility tests, or temporary NFT licensing. This enables agents to learn their uncertain preferences through iterative matchings, circumventing the limitations of classical models.

The multi-armed bandit (MAB) framework models how a player learns in an unknown environment with limited feedback [Auer et al., 2002, Lattimore and Szepesvári, 2020]. In the basic setup, a player faces K arms, each with an initially unknown reward distribution. Upon selecting an arm, the player observes a stochastic reward and updates their belief about the arm's preference. The goal is to maximize cumulative expected rewards, or equivalently, minimize cumulative expected regret – the difference between rewards from the optimal arm and the player's chosen arms over time. Classical strategies for balancing exploration (learning arm preferences) and exploitation (leveraging known rewards), such as explore-then-commit (ETC) [Garivier et al., 2016], upper confidence bound (UCB) [Auer et al., 2002] and Thompson sampling (TS) [Thompson, 1933], achieve sublinear regret, ensuring asymptotic optimality.

The online learning setting in housing market mirrors the exploration-exploitation trade-off central to MAB problems, where agents sequentially select actions to maximize cumulative rewards amid uncertainty. The dynamic and competitive nature of housing markets further complicates this learning process, as agents' decisions influence not only their own outcomes but also the opportunities available to others.

To formalize these challenges, we initiate *bandit learning in housing markets* in this paper, abstract the market as a multiplayer bandit problem. Here, players and arms correspond to agents and houses, each player has heterogeneous and unknown preferences over the arms, and each player is naturally associated with an arm, which represents their initially endowed house. When being matched to an arm, the player could learn the corresponding preference through the stochastic reward, but every arm could be matched to at most one player every time, when more than one player try to pull the same arm, all of them would get collision and receive no rewards. Taking the outcome from the TTC algorithm as a natural and desirable solution for the market, denoted as the core matching, our objective is to minimize the regret defined as the cumulative reward difference between the arm from the core matching and the player's selected arm. We develop matching and learning algorithms that can provably attain the core of the market in this setting. Our contributions are as follows:

- We introduce a novel model for housing markets in which agents initially lack knowledge of their preferences over houses but can repeatedly interact with the market to gradually learn these preferences. A key contribution is the definition of a natural notion of regret, grounded in the cooperative game-theoretic concept of the *core*, which quantifies the exploration-exploitation trade-off faced by individual players.

- When the horizon T of the bandit problem is known, we propose an ETC-type algorithm in the decentralized market setting, and prove an $\mathcal{O}(N \log T/\Delta_{\min}^2)$ problem-dependent upper bounds on the regret for every player, where N is the number of players, and $\Delta_{\min}$ is the players' minimum preference gap.

- When the horizon T is unknown, we provide a UCB-type algorithm in the centralized market setting, which is adaptive and anytime. We prove that centralized UCB achieves $\mathcal{O}(N \log T/\Delta_{\min}^2)$ problem-dependent upper bounds on the regret for every player.

- For the decentralized setting, we prove a matching lower bound of $\Omega(N \log T/\Delta_{\min}^2)$. To establish this, we construct an instance where a single player's exploration inevitably causes collisions for others. Specifically, any algorithm must spend $\Omega(\log T/\Delta_{\min}^2)$ rounds for a player to identify its optimal arm when the preference gap is $\Delta_{\min}$. In our construction, each such exploration step by one player forces a collision upon a specific, distinct player. Consequently, if $N-1$ players each have a minimum gap of $\Delta_{\min}$, their collective exploration imposes $\Omega(N \log T/\Delta_{\min}^2)$ total collisions – and thus regret – upon the remaining player.

2 Related Work

Multi-player Bandit Learning The multi-player bandit problem involves multiple decentralized players interacting with a shared multi-armed bandit environment. When players pull the same arm simultaneously, a collision occurs, resulting in a loss. In settings where arm means vary across players, the benchmark is typically the *maximum weight matching*, and the *system regret* – defined as the cumulative reward loss summed over all players – measures performances. Tibrewal et al. [2019] proposed an ETC type algorithm where players exploit the best-estimated matching, Mehrabian et al. [2020] combines forced collisions for implicit communication with matching eliminations, Shi et al. [2021] adapted the CUCB algorithm [Chen et al., 2013] to this setting. For a comprehensive survey, see Boursier and Perchet [2024].

While our model adopts the same reward and collision structure as multiplayer bandits with heterogeneous arm means, our benchmark diverges: we use the *core* of the housing market – a game-theoretic solution concept distinct from maximum weight matching – and define *individual regret* for each player rather than aggregate system regret. This shift ensures fairness but introduces new algorithmic challenges.

Competing Bandits in Two-sided Matching Markets Bandit problems in matching markets were first formalized by Das and Kamenica [2005], with subsequent works [Liu et al., 2020, 2021, Sankararaman et al., 2021, Kong and Li, 2023] exploring this model to achieve stable matching [Gale and Shapley, 1962]. In these two-sided markets, players (with unknown utilities) and arms (with known preferences) interact – when multiple players pull the same arm, only the top-ranked player by the arm's preference gets matched while others face collisions, with individual regrets defined accordingly. Since multiple stable matchings may exist, regrets are typically measured against either player-optimal or player-pessimal stable matchings.

In contrast, we study housing markets (one-sided matching) where only players have preferences over arms. This creates a distinct collision structure: when multiple players propose to the same arm, no one receives a reward. Moreover, the core matching in housing markets is unique, ensuring our regret definition is unambiguous. These differences from two-sided markets introduce novel algorithmic challenges.

3 Problem Setting

In this section, we formally introduce the problem setting of bandit learning in housing markets.

Denote N as the number of players in the market, and every player has an initial endowed arm. Let $\mathcal{N} = \{p_1, p_2, \cdots, p_N\}$ be the player set and $\mathcal{A} = \{a_1, a_2, \cdots, a_N\}$ be the arm set. The preferences of players on arms can be represented through a utility matrix $\boldsymbol{U}$, where $\boldsymbol{U}(i,j) \in [0,1]$ denotes the preference of player p_i on arm a_j. If $\boldsymbol{U}(i,j) > \boldsymbol{U}(i,j')$, player p_i prefers arm a_j over $a_{j'}$, and we denote as $a_j \succ_i a_{j'}$. In this paper, we assume all preferences are distinct, i.e.,

$\boldsymbol{U}(i,j) \neq \boldsymbol{U}(i,j')$ for any different arms $a_j \neq a_{j'}$. Additionally, the utility vectors can vary entirely between players, reflecting heterogeneous preferences over arms. In each round $t = 1, 2, \cdots$, every player p_i proposes to an arm $A_i(t) \in \mathcal{A}$. Each arm a_j then receives applications from the set of players $A_j^{-1}(t) := \{p_i : A_i(t) = a_j\}$. Since arm a_j is initially owned by player p_j, we assume only p_j observes the application profile $A_j^{-1}(t)$, while other players remain unaware of this information. Arms are indivisible goods with no preferences over players, if multiple players propose to a_j in the same round, i.e., $|A_j^{-1}(t)| > 1$, a collision occurs, and a_j is not matched to any player. In this case, the successfully matched player for a_j is $\bar{A}_j^{-1}(t) = \emptyset$, and each proposing player $p_i \in A_j^{-1}(t)$ receives $\bar{A}_i(t) = \emptyset$, along with a deterministic reward $X_i(t) = 0$ indicating they are blocked. Conversely, if only one player p_i proposes to a_j, i.e., $|A_j^{-1}(t)| = 1$, the match succeeds: $\bar{A}_j^{-1}(t) = p_i$, and p_i receives a random reward $X_i(t)$ characterizing its matching experience in this round, which we assume is 1-subgaussian with expectation $\boldsymbol{U}(i, \bar{A}_i(t))$. The set of all successfully matched player-arm pairs in round t forms a matching, denoted μ_t, where $\mu_t(p_i) = \bar{A}_i(t)$.

The core is a fundamental solution concept in housing markets [Shapley and Scarf, 1974], which is the set of feasible allocations where no coalition of agents can benefit by breaking away from the grand coalition. Formally,

Definition 1 (Core)**.** *A matching μ is in the core if no coalition $\mathcal{S} \subseteq \mathcal{N}$ can block μ, i.e., there does not exist an alternative matching μ' such that $\mu'(p_i) \succ_i \mu(p_i), \forall p_i \in \mathcal{S}$.*

In housing markets with strict preferences, the core is always nonempty, and consists of a unique matching [Roth and Postlewaite, 1977], which we refer to as the *core matching* and denote by μ^*. Our objective is to learn μ^* while minimizing the *core regret* for each player $p_i \in \mathcal{N}$. This regret is defined as the cumulative difference, over T rounds, between the expected rewards from being matched to $\mu^*(p_i)$ and the rewards actually obtained by p_i:

$$Reg_i(T) = T \cdot \boldsymbol{U}(i, \mu^*(p_i)) - \mathbb{E}\left[\sum_{t=1}^{T} X_i(t)\right]. \tag{1}$$

The expectation is taken over the randomness of the received reward and the players' strategy.

Offline Top-Trading-Cycle Algorithm In the *offline* setting where all players know their exact preferences, the Top-Trading-Cycle (TTC) algorithm [Shapley and Scarf, 1974] efficiently computes the unique core allocation of the housing market – a stable assignment where no coalition of players can improve their outcomes through mutual exchanges. The centralized TTC proceeds iteratively. First, each player identifies her most preferred available arm. Second, a directed graph is formed where players point to owners of their top choices. Third, at least one cycle (including possible self-cycles) is identified and implemented, with involved players exiting the market. This process terminates within N steps, guaranteed to produce a core allocation.

4 Decentralized Algorithm

In this section, we present an Explor-then-commit (ETC) algorithm for decentralized housing markets, operating in two phases. First, players explore arms in a round-robin fashion to estimate their preference rankings (Line 2-25), ensuring reliable ordinal estimates for matching. Next, players use these estimates to identify their assigned arms in the core matching via a decentralized adaptation of the *You Request My House, I Get Your Turn (YRMH-IGYT)* mechanism [Abdulkadiroğlu and

Algorithm 1 Decentralized Bandit Learning in Housing Markets (from view of player p_i)

Input: player set $\mathcal{P}$ and the corresponding arm set $\mathcal{A}$, horizon T.

Initialize: $\hat{\boldsymbol{U}}(i,j)=0$, $T_{i,j}=0$, $\forall j\in[N]$.
//Phase 1, learn the preferences
for $\ell=1,2,\cdots$ **do**
  $P_\ell=$ False // whether the preference has been estimated well
  for $t=\sum_{\ell'=1}^{\ell-1}(2^{\ell'}+N)+1,\cdots,\sum_{\ell'=1}^{\ell-1}(2^{\ell'}+N)+2^{\ell}$ **do**
    $A_i(t)=a_{(i+t-1)\%N+1}$.
    Observe $X_{i,A_i(t)}(t)$ and update $\hat{\boldsymbol{U}}(i,A_i(t)),T_{i,A_i(t)}$ if $\bar{A}_i(t)=A_i(t)$.
  end for
  Compute $\mathrm{UCB}_{i,j}$ and $\mathrm{LCB}_{i,j}$ for each $j\in[N]$.
  if $\exists\sigma$ such that $\mathrm{LCB}_{i,\sigma_k}>\mathrm{UCB}_{i,\sigma_{k+1}}$ for any $k\in[N]$ **then**
    $P_\ell=$ True and $\sigma_i=\sigma$.
  end if
  for $t=\sum_{\ell'=1}^{\ell-1}(2^{\ell'}+N)+2^{\ell}+1,\cdots,\sum_{\ell'=1}^{\ell}(2^{\ell'}+N)$ **do**
    $t'=t-\sum_{\ell'=1}^{\ell-1}(2^{\ell'}+N)-2^{\ell}$
    if $P_\ell==$ True **then**
      $A_i(t)=a_{t'}$.
    else
      $A_i(t)=\emptyset$.
    end if
    if $i==t'$ and $|A_i^{-1}(t)|==N$ **then**
      Enter in Phase 2 with $\sigma_i=(\sigma_{i,1},\cdots,\sigma_{i,N})$.
      $t_1=\sum_{\ell'=1}^{\ell}(2^{\ell'}+N)$ //The round when phase 1 ends.
    end if
  end for
end for
//Phase 2, find the core matching arm with σ_i
$t=t_1+1$
$F_i=$ True //Whether the initial endowed arm is still available in the market, this flag is known to all players
$k=1$, $\mathcal{A}_k=\{a_i:F_i==\text{True}\}$.
$Propose(i)=$ False. //A flag indicated whether player p_i proposed to some arm in the k-th epoch.
$i_{\min}=\min(\{i|a_i\in\mathcal{A}_k\})$.
while $|\mathcal{A}_k|>0$ and $F_i==$ True **do**
  $j_{\min}^{(i)}=\min\left\{j|a_{\sigma_{i,j}}\in\mathcal{A}_k\right\}$. //The best available arm for player p_i.
  if $t==t_1+1$ or $k(t)==k(t-1)+1$ **then**
    if $i==i_{\min}$ **then**
      $A_i(t)=a_{j_{\min}^{(i)}}$.
      $Propose(i)=$ True.
    end if
  else if $|A_i^{-1}(t-1)|>0$ **then**
    $A_i(t)=a_{j_{\min}^{(i)}}$, $Propose(i)=$ True.
    $j_r=\left\{j|a_j=A_i^{-1}(t-1)\right\}$.
  end if
  if ($Propose(i)==$ True and $|A_i^{-1}(t)|>0$) or ($F_{j_r}==$ False) **then** //In a cycle
    $F_i=$ False, $A_i(s)=a_{j_{\min}^{(i)}}$, $\forall s>t$.
  end if
  $\mathcal{B}=\{a_i:F_i==True\}$.
  if $\mathcal{B}\neq\mathcal{A}_k$ **then** //The available arm set is updated
    $k=k+1$, $\mathcal{A}_k=\mathcal{B}$, $Propose(i)=$ False.
    $i_{\min}=\min\{i|a_i\in\mathcal{A}_k\}$.
  end if
end while

Sönmez, 1999] (Line 26-51). Once matched, players commit exclusively to these arms in all future rounds.

The first phase of the Algorithm 1 consists of multiple sub-phases $\ell = 1, 2, \cdots$, each lasting $2^\ell + N$ rounds (Line 2-25). Each sub-phase ℓ begins with an exploration stage of 2^ℓ rounds (Line 5-8), followed by N communication rounds (Line 13-24). During the exploration stage, players aim to gather sufficient observations about the arms. The subsequent communication rounds allow them to verify whether all N players have accurately learned their preferences. Once this condition is met, players exit the exploration phase and proceed to the second phase, where they identify arms in the core matching (Line 21).

During the exploration stage of each sub-phase, players follow a round-robin strategy to propose to arms (Line 6). Since each player initially owns a distinct endowed arm with a unique index, this ensures that no two players select the same arm simultaneously. As a result, all proposals are successfully accepted without collisions. When player p_i receives an observation from the selected arm $A_i(t)$, it updates both the estimated preference value $\hat{\boldsymbol{U}}(i, A_i(t))$ and the observation count $T_{i,A_i(t)}$ for that arm (Line 7). The update rule is as follows,

$$\hat{\boldsymbol{U}}(i, A_i(t)) = \frac{\hat{\boldsymbol{U}}(i, A_i(t)) \cdot T_{i,A_i(t)} + X_{i,A_i(t)}(t)}{T_{i,A_i(t)} + 1}, \tag{2}$$

$$T_{i,A_i(t)} = T_{i,A_i(t)} + 1. \tag{3}$$

At the end of the exploration phase, players construct a confidence set for their estimated preference values using collected observations. Specifically, for player p_i, the confidence interval for the preference value of arm a_j is defined by an upper bound (UCB) and a lower bound (LCB), given as

$$\begin{aligned} \mathrm{UCB}_{i,j} &= \hat{\boldsymbol{U}}(i,j) + \sqrt{\frac{6 \log T}{\max\{T_{i,j}, 1\}}}, \\ \mathrm{LCB}_{i,j} &= \hat{\boldsymbol{U}}(i,j) - \sqrt{\frac{6 \log T}{\max\{T_{i,j}, 1\}}}. \end{aligned} \tag{4}$$

When the confidence intervals of two arms $a_j, a_{j'}$ become disjoint, that is, wehn $\mathrm{LCB}_{i,j} > \mathrm{UCB}_{i,j'}$ or vice versa, player p_i can establish its strict preference ordering between them. Once p_i successfully determines a complete preference ranking over all N arms (Line 10), it sets the flag P_ℓ to True for the current sub-phase and records this permutation as its estimated preference ranking σ_i.

Communication rounds enable players verify if all participants have accurately estimated their preference rankings. In the i-th communication round of every sub-phase, player p_i's endowed arm serves as a broadcast channel. A player p_j proposes to p_i's arm *if and only if* it has an accurate ranking, i.e., $P_\ell =$ True (Line 16); otherwise, it abstains (Line 18). While this may cause collisions, the goal is not to resolve them but to convey status information. If p_i observes proposals from all N players including itself (Line 20), it infers universal estimation success and triggers the transition to the second phase using its estimated ranking σ_i (Line 21).

The second phase (Line 26-51) adapts the YRMH-IGYT algorithm [Abdulkadiroğlu and Sönmez, 1999] to our decentralized setting for players to progressively identify their core allocations. The process unfolds in successive sub-phases. In each, the unassigned player with the smallest index initiates a proposal chain by proposing to their most preferred remaining arm (Lin 36-37). The recipient of a proposal then proposes to their own top choice among available arms, and this sequence continues. A top trading cycle is completed when a proposal is received by a player who is already part of the current chain (Line 43). The player who closes the cycle – the first to receive a repeat proposal

– then marks their endowed arm as allocated. This allocation status propagates backward through the chain. All players involved in the cycle permanently fix their allocations and remove their own endowed arms from the market (Line 44). The process repeats among the remaining players until all arms are allocated, thus achieving a stable core matching in a fully decentralized manner.

Remark 1. *Algorithm 1 relies on standard multiplayer bandit assumptions for implicit coordination: (1) Unique, globally-known arm IDs, which also serve as player identifiers for round-robin exploration and proposal-chain initiation. (2) A shared global clock to synchronize the transition to Phase 2. (3) Common knowledge of the available arms in each round, enabling players to reconstruct the flag F_i.*

4.1 Theoretical Analysis

Prior to presenting the formal regret analysis of Algorithm 1, we introduce the notion of preference gaps to quantify the intrinsic difficulty of the learning problem.

Definition 2 (Minimum Preference Gap)**.** *For each player p_i and arm $a_j \neq a_{j'}$, let $\Delta_{i,j,j'} = \boldsymbol{U}(i,j) - \boldsymbol{U}(i,j')$ be the preference gap of p_i between a_j and $a_{j'}$. Let r_i be the preference ranking of player p_i and $r_{i,k}$ be the k-th preferred arm in p_i's ranking for $k \in [N]$. Define $\Delta_{\min} = \min_{i\in[N];k\in[N]} \Delta_{i,r_{i,k},r_{i,k+1}}$ as the minimum preference gap among all workers and their preferences over the arms. $\Delta_{\min}$ is non-negative since all preferences are distinct.*

We now present the upper bound for the core regret for each player by following Algorithm 1.

Theorem 1. *Following Algorithm 1, the core regret of each player $p_i \in \mathcal{N}$ satisfies*

$$\begin{aligned} Reg_i(T) &\leq \left(\frac{192N\log T}{\Delta_{\min}^2} + N\log\left(\frac{192N\log T}{\Delta_{\min}^2}\right) + 3N^2\right)\cdot \Delta_{i,\max} \\ &= \mathcal{O}\left(\frac{N\log T}{\Delta_{\min}^2}\right) \end{aligned}$$

The regret bound consists of three key components. First, the cumulative regret from exploration rounds in phase 1. Second, the regret from communication rounds in phase 1. The final term combines two distinct elements: (1) the regret during phase 2, where the YRMH-IGYT mechanism guarantees convergence to the core matching within at most N epoches (with each epoch requirting $\leq N$ rounds to identify a top trading cycle), and (2) the regret contribution from rare concentration failure events.

For convenience, let $\hat{\boldsymbol{U}}^{(t)}(i,j), T_{i,j}^{(t)}, UCB_{i,j}^{(t)}, LCB_{i,j}^{(t)}$ be the value of $\hat{\boldsymbol{U}}(i,j), T_{i,j}, UCB_{i,j}, LCB_{i,j}$ at the end of round t. Define

$$\mathcal{F} = \left\{ \exists t \in [T], i \in [N], j \in [N] : |\hat{\boldsymbol{U}}^{(t)}(i,j) - \boldsymbol{U}(i,j)| > \sqrt{\frac{6\log T}{T_{i,j}^{(t)}}} \right\}$$

as the bad event that some preference is not estimated well during the horizon. Since all players would communicate whether they have a precise enough estimation after each sub-phase of phase 1, and determine to enter phase 2 once they find every player include themselves have good-enough estimations, we can conclude that all players would enter phase 2 at the same time. Denote $\ell_{\max}$ as the largest sub-phase number of phase 1. That is to say, players enter in phase 2 at the end of sub-phase $\ell_{\max}$. We then provide the proof of Theorem 1 as follows.

Proof of Theorem 1. Let $\Delta_{i,\max}$ be the maximum core regret that may be suffered by player p_i in all rounds, we have $\Delta_{i,\max} \leq 1$ by assumption on the utility matrix. The core regret of each player p_i by following Algorithm 1 satisfies

$$\begin{aligned}
Reg_i(T) =& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}(i,\mu^*(p_i))\right) - X_i(t)\right] \\
\leq& \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mu_t(i)\neq\mu^*(p_i)\}\cdot\Delta_{i,\max}\right] && (5)\\
\leq& \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mu_t(i)\neq\mu^*(p_i)\}\middle|\neg\mathcal{F}\right]\cdot\Delta_{i,\max} + \mathbb{P}(\mathcal{F})\cdot T\cdot\Delta_{i,\max} \\
\leq& \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\mu_t(i)\neq\mu^*(p_i)\}\middle|\neg\mathcal{F}\right]\cdot\Delta_{i,\max} + 2N^2\Delta_{i,\max} && (6)\\
\leq& \mathbb{E}\left[\sum_{\ell=1}^{\ell_{\max}}\left(2^\ell+N\right)+N^2\middle|\neg\mathcal{F}\right]\cdot\Delta_{i,\max} + 2N^2\Delta_{i,\max} && (7)\\
\leq& \left(\frac{192N\log T}{\Delta_{\min}^2} + N\log\left(\frac{192N\log T}{\Delta_{\min}^2}\right)\right)\cdot\Delta_{i,\max} + 3N^2\Delta_{i,\max}, && (8)
\end{aligned}$$

where Eq.(5) comes from the fact that in a housing market, there is a unique core matching and hence a unique core matching partner $\mu^*(p_i)$ for player p_i. Eq.(6) holds based on Lemma 1. Eq.(7) holds according to Algorithm 1 and the fact that we need at most N^2 rounds for the YRMH-IGYT procedure to reach the equilibrium (Lemma 2). Eq.(8) follows from Lemma 3. □

The following the technical lemmas underlying our analysis. Proofs are provided in Appendix A. We begin with Lemma 1, which bounds the probability of erroneous preference estimation.

Lemma 1 (Bad Concentration Event)**.**

$$\mathbb{P}\left(\mathcal{F}\right) \leq \frac{2N^2}{T}.$$

Lemma 2 shows that each player p_i finds their core-matched arm $\mu^*(p_i)$ within at most N^2 rounds in Phase 2.

Lemma 2. *Conditional on $\neg\mathcal{F}$, at most N^2 rounds are needed in phase 2 before $A_i(t) = \mu^*(p_i)$, and in all of the following rounds, $A_i(t)$ would not be updated and p_i would always be successfully accepted by $\mu^*(p_i)$.*

Lemma 3 bounds the duration of the exploration phase.

Lemma 3. *Condition on $\neg\mathcal{F}$, phase 1 will proceed in at most $\ell_{\max}$ sub-phases where*

$$\ell_{\max} = \min\left\{\ell : \sum_{\ell'=1}^{\ell} 2^{\ell'} \geq \frac{96N\log T}{\Delta_{\min}^2}\right\}, \tag{9}$$

which implies that $\sum_{\ell'=1}^{\ell_{\max}} 2^{\ell'} \leq \frac{192N\log T}{\Delta_{\min}^2}$ and $\ell_{\max} = \log\left(\frac{192N\log T}{\Delta^2}\right)$ since the sub-phase length grows exponentially. And all players will enter in phase 2 simultaneously at the end of sub-phase $\ell_{\max}$.

5 Regret Lower Bound

We establish a regret lower bound by adapting the method of Burnetas and Katehakis [1996] to our multi-player setting, demonstrating the tightness of the regret upper bound of Algorithm 1. Let $Reg(T;\boldsymbol{\nu},\pi)$ be the cumulative expected regret of policy π over all players and time T, for an instance with arm distributions $\boldsymbol{\nu} = \{\nu_{ij} : i, j \in [N]\}$. Denote by $\mathcal{P}$ the set of all probability distributions with support in $[0,1]$. For any $\boldsymbol{\nu} \in \mathcal{P}$, define $D_{\inf}(\boldsymbol{\nu}, x, \mathcal{P}) = \inf_{\boldsymbol{\nu}' \in \mathcal{P}} \{D(\boldsymbol{\nu}, \boldsymbol{\nu}') : \rho(\boldsymbol{\nu}') > x\}$, where $\rho : \mathcal{P} \to \mathbb{R}$ maps a distribution to its mean and $D(\cdot,\cdot)$ is the KL divergence.

Definition 3 (Uniformly Consistent Policies)**.** *A policy π is uniformly consistent* if and only if *for all $\boldsymbol{\nu} \in \mathcal{P}$, all $\alpha \in (0,1)$, the regret* $\limsup_{T\to\infty} \frac{Reg(T;\boldsymbol{\nu},\pi)}{T^\alpha} = 0$.

This ensures the policy is not overly tuned to the current instance at the expense of performance in others, ensuring robust performance across different problem instances [Burnetas and Katehakis, 1996, Lattimore and Szepesvári, 2020].

Single Top Trading Cycle Bandits Our regret lower bound applies to a subclass of bandits characterized by a single top trading cycle (STTCB). In these instances, each player has a unique top-ranked arm, and there exists a permutation σ of players such that $\mu^*(p_{\sigma_i}) = a_{\sigma_{i+1}}$ for $i = 1, \cdots, N-1$ and $\mu^*(p_{\sigma_N}) = a_{\sigma_1}$. This structure ensures all players form one top trading cycle in the core matching. For each player p_i, define the gap $\Delta_j^{(i)} = \boldsymbol{U}(i, \mu^*(p_i)) - \boldsymbol{U}(i,j)$ and the minimum gap $\Delta_{\min}^{(i)} = \min_{a_j \neq \mu^*(p_i)} \boldsymbol{U}(i, \mu^*(p_i)) - \boldsymbol{U}(i,j)$, which are non-negative in STTCB instances.

Lemma 4 (Regret Decomposition)**.** *For an STTCB instance $\boldsymbol{\nu} = \{\nu_{i,j} : i \in [N], j \in [N]\}$, and any uniformly consistent policy π, player p_i for $i \in [N]$ the following holds*

$$Reg_i(T;\boldsymbol{\nu},\pi) \geq \max\left\{ \sum_{i' \neq i} \Delta_{\min}^{(i)} \mathbb{E}_{\boldsymbol{\nu},\pi}\left[N_{\mu^*(p_i)}^{(i')}(T)\right], \sum_{a_j \neq \mu^*(p_i)} \Delta_j^{(i)} \mathbb{E}_{\boldsymbol{\nu},\pi}\left[N_j^{(i)}(T)\right] \right\}.$$

For an STTCB instance, when the p_i's core-matched partner $\mu^*(p_i)$ is allocated to another player, the optimal outcome is for p_i to receive its second-best arm. While this event occurs infrequently, it provides a sufficient bound for the first term above.

Theorem 2. *For any player p_i, $i \in [N]$, under any decentralized universally consistent algorithm π on an STTCB instance $\boldsymbol{\nu}$ satisfies*

$$\begin{aligned} &\liminf_{T\to\infty} \frac{Reg_i(T;\boldsymbol{\nu},\pi)}{\log T} \\ &\geq \max\left\{ \sum_{i' \neq i} \frac{\Delta_{\min}^{(i)}}{D_{\inf}(\boldsymbol{\nu}_{i',\mu^*(p_i)}, \boldsymbol{U}(i', \mu^*(p_{i'})), \mathcal{P})}, \sum_{a_j \neq \mu^*(p_i)} \frac{\Delta_j^{(i)}}{D_{\inf}(\boldsymbol{\nu}_{i,j}, \boldsymbol{U}(i, \mu^*(p_i)), \mathcal{P})} \right\}. \end{aligned} \tag{10}$$

Proof Idea of Theorem 2. For a given STTCB instance $\boldsymbol{\nu}$, we construct a confounding alternative $\boldsymbol{\nu}'$ that differs only in the utility of a single player-arm pair (p_i, a_j), where a_j is the core-matched arm of p_i in $\boldsymbol{\nu}'$ but not in $\boldsymbol{\nu}$. To identify the true instance and avoid linear regret, any algorithm must gather enough evidence to distinguish $\boldsymbol{\nu}$ from $\boldsymbol{\nu}'$, necessitating a number of samples inversely proportional to their KL-divergence. This directly implies a logarithmic regret lower bound. □

The detailed proof of Theorem 2 is deferred to Appendix B. The following corollary demonstrates that the $N \log T / \Delta_{\min}^2$ dependence in Theorem 1 cannot be improved, thus establishing $\Theta(N \log T / \Delta_{\min}^2)$ as the minimax regret rate for all players.

Corollary 1. *There exists an STTCB instance with Bernoulli rewards, where the regret of player p_i is lower bounded as $\Omega(\frac{N \log T}{\Delta_{\min}^2})$.*

Reward heterogeneity necessitates $\Omega(\log T/\Delta_{\min}^2)$ explorations for the player with minimum gap $\Delta_{\min}$, during which players with substantial gaps ($\Delta_{\min}^{(i')} = \Omega(1)$) incur significant regret. This core observation underpins our proof.

From Corollary 1, we can see that the regret upper bound (Theorem 1) for Algorithm 1 matches the regret lower bound, showing the significance of our proposed algorithm.

6 Centralized Anytime Algorithm

While Algorithm 1 achieves a regret upper bound that matches the lower bound in Theorem 2 for all players, it requires prior knowledge of the time horizon T to properly construct confidence intervals. Although one could apply the doubling trick [Auer et al., 1995] to make the algorithm anytime, this approach leads to non-monotonic expected instantaneous regret. Such behavior could raise concerns among players about the algorithm's consistency during execution.

In this section, we develop an adaptive anytime algorithm that operates through a centralized platform capable of aggregating player preferences and regulating allocations without observing individual reward realizations. The algorithm employs the principle of optimism in the face of uncertainty, where each player maintains upper confidence bounds (UCB) to rank arms and submits these rankings to the platform every round. The platform then computes a core matching by applying the TTC algorithm to the collected preferences, and players subsequently pull their assigned arms.

We begin by formally defining the UCB estimation method used by individual players and introducing several technical concepts necessary for the analysis. Building on this foundation, we derive a regret upper bound for the centralized approach. A key feature of this framework is that the platform's coordination eliminates potential conflicts between player proposals, allowing us to assume collision-free operation throughout the learning process.

Algorithm 2 Anytime Centralized Bandit Learning in Housing Markets

Input: player set $\mathcal{P}$ and the corresponding arm set $\mathcal{A}$.

for $t = 1, \cdots$ **do**
  The *platform* receives ranking $\hat{r}_{i,t}$ from all players p_i.
  The *platform* computes a core matching μ_t using the offline TTC algorithm with the ranking $\hat{r}_{i,t}$.
  for $i = 1, \cdots, N$ **do**//Players pull arms simultaneously
    $A_i(t) = \mu_t(p_i)$.
    Update $\hat{\boldsymbol{U}}^{(t)}(i, A_i(t))$ and $u^{(t)}(i, A_i(t))$ according to Eq.(2) and Eq.(11).
    Compute the current ranking $\hat{r}_{i,t+1}$ according to $u^{(t)}(i, \cdot)$.
  end for
end for

At iteration t, when a player p_i gets matched to arm $A_i(t)$, it would update the estimated preference value $\hat{\boldsymbol{U}}(i, A_i(t))$ and the observed time $T_{i,A_i(t)}$ for arm $A_i(t)$ according to Eq.(2) in Section 4.

Then the upper confidence bound, which is called the *index* for each (i, j)-th entry of the utility

matrix $\boldsymbol{U}$ is computed as

$$u^{(t)}(i,j) = \begin{cases} \infty & \text{if } T_{i,j}(t) = 0, \\ \hat{\boldsymbol{U}}^{(t)}(i,j) + \sqrt{\frac{3\log t}{2T_{i,j}^{(t-1)}}} & \text{otherwise.} \end{cases} \tag{11}$$

Each player p_i uses the index to compute a preference ranking $\hat{r}_{i,t+1}$, where arms are ordered by their UCBs in a decreasing order (e.g., $\text{argmax}_j\, u^{(t)}(i,j)$ is ranked first).

Let $T_\mu(t)$ denote the number of times a matching μ is played by time t. A matching is *achievable* at time t if it is core-stable according to the current estimated rankings $\{\hat{r}_{i,t}\}_{i\in[N]}$. A matching is *truly core-stable* if it is core-stable under the true utility matrix $\boldsymbol{U}$. For player p_i and arm a_j, define $M_{i,j}$ as the set of achievable (but not truly core-stable) matchings where p_i is matched to a_j. The *achievable preference gap* is defined as $\bar{\Delta}_{i,j} = \boldsymbol{U}(i,\mu^*(p_i)) - \boldsymbol{U}(i,j)$.

Since the truly core-stable matching μ^* incurs zero regret, we bound the regret of player p_i as follows:

$$Reg_i(T) \le \sum_{j:\bar{\Delta}_{i,j}>0} \bar{\Delta}_{i,j} \left(\sum_{\mu\in M_{i,j}} \mathbb{E}T_\mu(T) \right). \tag{12}$$

If a matching μ is not truly core-stable under the ground-truth utility matrix $\boldsymbol{U}$, then a *blocking coalition* must exist. This coalition – a subset of players and their endowed arms – can reassign arms among themselves to improve every member's utility. A *blocking triplet* (p_i, a_j, a_k) occurs when, under μ, player p_i is matched to a_j, yet the ground-truth utility satisfies $\boldsymbol{U}(i,k) > \boldsymbol{U}(i,j)$. Now, consider the set $M_{i,j}$ of all achievable matchings that are non-truly core-stable and where p_i receives a_j. For this set, there exists a corresponding set of arms $\mathcal{S}_{i,j} \subseteq \mathcal{A}$ such that: for every $a_k \in \mathcal{S}i,j$, the triplet (p_i, a_j, a_k) is blocking for any $\mu \in Mi,j$; while for any $a_{k'} \in \mathcal{A}\setminus\mathcal{S}_{i,j}$, we have $\mathbf{U}(i,j) \ge \mathbf{U}(i,k')$. Let $Q_{i,j}$ be the set of all such blocking triplets formed from $\mathcal{S}_{i,j}$. Crucially, for any implemented matching $\mu \in \mathbf{M}_{i,j}$, at least one triplet in $Q_{i,j}$ must be realized. This structure leads to the following regret upper bound.

Theorem 3. *When all players rank arms according to the index defined in Eq.(11) and submit their preferences to the centralized platform, and the platform uses the TTC algorithm with the received preferences to compute an allocation each round, the core regret of player p_i up to time T satisfies*

$$\begin{aligned} Reg_i(T) &\le \sum_{j:\bar{\Delta}_{i,j}>0} \bar{\Delta}_{i,j} \left[\sum_{(p_i,a_j,a_k)\in Q_{i,j}} \left(5 + \frac{6\log T}{\Delta_{i,k,j}^2} \right) \right] \\ &\le \max_j \bar{\Delta}_{i,j} \left(5N^2 + 12\frac{N\log T}{\Delta_{\min}^2} \right) \\ &= \mathcal{O}\left(\frac{N\log T}{\Delta_{\min}^2} \right) \end{aligned}$$

Theorem 3 offers a centralized problem-dependent $\mathcal{O}\left(\frac{N\log T}{\Delta_{\min}^2}\right)$ upper bound guarantees on the core regret of each player p_i, which also matches the lower bound derived in Section 5, and the proposed UCB-type algorithm is anytime and adaptive in the sense that the players do not need to know the time horizon T and the minimum reward gap $\Delta_{\min}$ in advance.

Proof. Let $T_{i,j,k}(T)$ be the number of times player p_i pulls arm a_j while the triplet (p_i, a_j, a_k) is blocking the matching selected by the platform. Since every time when a matching $\mu \in M_{i,j}$ is implemented, at least one of the blocking triplets in $Q_{i,j}$ happens, we have

$$\sum_{\mu \in M_{i,j}} T_\mu(T) \leq \sum_{(p_i,a_j,a_k) \in Q_{i,j}} T_{i,j,k}(T). \tag{13}$$

By the definition of a blocking triplet, we know that if p_i pulls arm a_j when (p_i, a_j, a_k) is blocking, p_i must have a higher upper confidence bound for a_j than for a_k, i.e., we have the index $u(i,j) > u(i,k)$ while according to the ground-truth utility matrix we have $\boldsymbol{U}(i,j) < \boldsymbol{U}(i,k)$. Standard analysis for the single player UCB [Bubeck et al., 2012] shows that

$$\mathbb{E}T_{i,j,k}(T) \leq 5 + \frac{6 \log T}{\Delta^2_{i,k,j}}. \tag{14}$$

Combining Eq.(12), (13) and (14) we get the first inequality on the regret upper bound.

Furthermore, as there are at most N arms such that $\bar{\Delta}_{i,j} > 0$, we have

$$\sum_{j:\bar{\Delta}_{i,j}>0} \bar{\Delta}_{i,j} \left[\sum_{(p_i,a_j,a_k) \in Q_{i,j}} \left(5 + \frac{6 \log T}{\Delta^2_{i,k,j}} \right) \right] \leq \max_j \bar{\Delta}_{i,j} \left[\sum_{(p_i,a_j,a_k) \in Q_{i,j}} \left(5N + \frac{6N \log T}{\Delta^2_{i,k,j}} \right) \right]$$

When we consider the triplet set composed of all possible $\boldsymbol{U}(i,j) < \boldsymbol{U}(i,k)$, we have $|Q_{i,j}| \leq N$ and

$$\sum_{k:\boldsymbol{U}(i,j)<\boldsymbol{U}(i,k)} \frac{1}{\Delta^2_{i,k,j}} \leq \sum_{k=1}^{N} \frac{1}{k^2 \Delta^2_{\min}} \leq \frac{2}{\Delta^2_{\min}},$$

which concludes the second inequality of the regret bound. □

7 Conclusion and Discussion

This paper introduces a novel online learning framework for housing markets with uncertain preferences, unifying two key objectives: core-stability and sample efficiency. We formalize this through core regret as our performance metric and propose two algorithms that bridge multi-armed bandit techniques with housing market mechanisms. These algorithms adapt to both centralized and decentralized settings, with guarantees for fixed-horizon and anytime environments. A matching regret lower bound proves the order-optimality of the algorithms, highlighting their theoretical and practical significance.

There are many additional questions that can be studied in this model.

Housing Markets with Existing Tenants This paper focuses on housing markets where each player initially possesses one endowed arm, maintaining an equal number of players and arms. However, real-world housing markets typically involve more complex scenarios: existing tenants with endowed arms coexist with new applicants lacking initial allocations, while vacant houses (free arms) may also be available [Abdulkadiroğlu and Sönmez, 1999]. In these generalized settings, the core losses its uniqueness, raising two fundamental challenges: first, identifying an appropriate tractable solution concept to serve as a benchmark, and second, determining whether efficient learning algorithms can achieve sublinear regret in this more realistic but complicated environment.

Housing Markets with Indifference While our current analysis assumes strict player preferences over arms, real-world housing markets often involve indifference between options. Our learning algorithm achieves a regret bound of $\Theta(N \log T/\Delta_{\min}^2)$, which becomes vacuous when $\Delta_{\min} = o(1/\sqrt{T})$ – precisely when indifference between arms creates vanishing preference gaps. This limitation highlights the need for new algorithmic approaches that can handle indifferences in housing market allocations, presenting an important direction for future research.

Incentive Compatibility in the Learning Setting While the top trading cycle algorithm is strategy-proof under known, deterministic preferences, its incentive compatibility remains unclear in learning settings where preferences must be discovered dynamically. In particular, the decentralized nature of these interactions may create opportunities for strategic manipulation through learning behavior. Understanding whether and how players can exploit the learning process to achieve better outcomes presents a significant open question for future research.

References

Atila Abdulkadiroğlu and Tayfun Sönmez. House allocation with existing tenants. *Journal of Economic Theory*, 88(2):233–260, 1999.

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

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

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Etienne Boursier and Vianney Perchet. A survey on multi-player bandits. *Journal of Machine Learning Research*, 25(137):1–45, 2024.

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

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013.

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

Federico Echenique, Antonio Miralles, and Jun Zhang. Constrained pseudo-market equilibrium. *American Economic Review*, 111(11):3699–3732, 2021.

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

Jugal Garg, Thorben Tröbst, and Vijay Vazirani. One-sided matching markets with endowments: equilibria and algorithms. *Autonomous Agents and Multi-Agent Systems*, 38(2):40, 2024.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 29, 2016.

Aanund Hylland and Richard Zeckhauser. The efficient allocation of individuals to positions. *Journal of Political economy*, 87(2):293–314, 1979.

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.

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

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.

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

Abbas Mehrabian, Etienne Boursier, Emilie Kaufmann, and Vianney Perchet. A practical algorithm for multiplayer bandits when arm means vary among players. In *International Conference on Artificial Intelligence and Statistics*, pages 1211–1221. PMLR, 2020.

Alvin E Roth. *Online and matching-based market design*. Cambridge University Press, 2023.

Alvin E Roth and Andrew Postlewaite. Weak versus strong domination in a market with indivisible goods. *Journal of Mathematical Economics*, 4(2):131–137, 1977.

Alvin E Roth, Tayfun Sönmez, and M Utku Ünver. Kidney exchange. *The Quarterly journal of economics*, 119(2):457–488, 2004.

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.

Lloyd Shapley and Herbert Scarf. On cores and indivisibility. *Journal of mathematical economics*, 1(1):23–37, 1974.

Chengshuai Shi, Wei Xiong, Cong Shen, and Jing Yang. Heterogeneous multi-player multi-armed bandits: Closing the gap and generalization. *Advances in neural information processing systems*, 34:22392–22404, 2021.

Tayfun Sönmez and M Utku Ünver. House allocation with existing tenants: A characterization. *Games and Economic Behavior*, 69(2):425–445, 2010.

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

Harshvardhan Tibrewal, Sravan Patchala, Manjesh K Hanawal, and Sumit J Darak. Distributed learning and optimal assignment in multiplayer heterogeneous networks. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 1693–1701. IEEE, 2019.

A Proof of Regret Upper Bound of the Decentralized Algorithm

A.1 Proof of Lemma 1

Proof.

$$\begin{aligned}
\mathbb{P}(\mathcal{F}) &= \mathbb{P}\left(\exists 1 \le t \le T, i \in [N], j \in [N] : |\hat{\boldsymbol{U}}^{(t)}(i,j) - \boldsymbol{U}(i,j)| > \sqrt{\frac{6\log T}{T_{i,j}^{(t)}}}\right) \\
&\le \sum_{t=1}^{T} \sum_{i\in[N]} \sum_{j\in[N]} \mathbb{P}\left(|\hat{\boldsymbol{U}}^{(t)}(i,j) - \boldsymbol{U}(i,j)| > \sqrt{\frac{6\log T}{T_{i,j}^{(t)}}}\right) \\
&\le \sum_{t=1}^{T} \sum_{i\in[N]} \sum_{j\in[N]} \sum_{s=1}^{t} \mathbb{P}\left(T_{i,j}^{(t)} = s, |\hat{\boldsymbol{U}}^{(t)}(i,j) - \boldsymbol{U}(i,j)| > \sqrt{\frac{6\log T}{s}}\right) \\
&\le \sum_{t=1}^{T} \sum_{i\in[N]} \sum_{j\in[N]} t \cdot 2\exp\left(-3\log T\right) \\
&\le \frac{2N^2}{T},
\end{aligned}$$

where the second last inequality results from Lemma 10. □

A.2 Proof of Lemma 2

Proof. According to Lemma 5 and Algorithm 1, when player p_i enters in phase 2 with ranking σ_i, we have

$$\boldsymbol{U}(i,\sigma_{i,j}) > \boldsymbol{U}(i,\sigma_{i,j+1}), \quad \forall j \in [N].$$

That is to say, player p_i successfully recovers the real preference ranking when she enters in phase 2. Further, according to Lemma 3, all players would enter in phase 2 simultaneously. Based on Lemma 6, we know that when given the true preference profile, YRMH-IGYT would stop in at most N^2 steps, and once the core matching is reached, players would focus on their core matching partner in all the remaining rounds. Since there is a unique core matching and no two players would share their core matching partners, p_i could always get matched to $\mu^*(p_i)$. □

Lemma 5. *Condition on $\neg\mathcal{F}$, $UCB_{i,j}^{(t)} < LCB_{i,j'}^{(t)}$ implies $\boldsymbol{U}(i,j) < \boldsymbol{U}(i,j')$.*

Proof. According to the definition of LCB and UCB, we have that conditional on $\neg\mathcal{F}$,

$$\begin{aligned}
\mathrm{LCB}_{i,j}^{(t)} &= \hat{mU}_{i,j}^{(t)} - \sqrt{\frac{6\log T}{T_{i,j}^{(t)}}} \\
&\le \boldsymbol{U}(i,j) \\
&\le \hat{\boldsymbol{U}}^{(t)}(i,j) + \sqrt{\frac{6\log T}{T_{i,j}^{(t)}}} \\
&= \mathrm{UCB}_{i,j}^{(t)}.
\end{aligned}$$

Therefore, if $\mathrm{UCB}_{i,j}^{(t)} < \mathrm{LCB}_{i,j'}^{(t)}$, we have that

$$\boldsymbol{U}(i,j) \leq \mathrm{UCB}_{i,j}^{(t)} \leq \mathrm{LCB}_{i,j}^{(t)} \leq \boldsymbol{U}(i,j').$$

□

Lemma 6 (YRMH-IGYT Procedure)**.** *Given the true preferences of all players, at most N epochs are needed to find the core matching, while in each epoch, at most N rounds are needed to reach a top trading cycle. Therefore, at most N^2 steps are needed for the YRMH-IGYT procedure to detect the core matching.*

Proof. In the YRMH-IGYT procedure, each epoch proceeds as follows:

1. Leader Selection: At the start of each epoch, the player with the smallest-index endowed arm among the remaining available arms is chosen as the leader to initiate the epoch.

2. Proposal Dynamics: The leader begins by proposing to their most preferred available arm. For subsequent rounds, any player who receives a proposal in round t will, in round $t+1$, propose to their own most favored remaining arm. If a player receives no proposal, they remain passive and do not make any offers. Critically, exactly one player proposes to one arm in every round, ensuring a structured progression.

3. Termination Condition: An epoch terminates when a player who previously made a proposal receives a new proposal from another player. This player must belong to a top trading cycle (TTC), as demonstrated in Lemma 7.

We could observe that each epoch identifies exactly one TTC, as the procedure ensures players always propose to their top remaining choices. Then in the worst case, in a single epoch, the request path may involve all N players, requiring up to N steps. On the other hand, since each TTC has length ≥ 1, at most N epochs are needed, leading to an overall worst-case runtime of N^2 steps. □

Lemma 7 (Cycle Detection)**.** *In any epoch during the YRMH-IGYT procedure, the first repeated player in the request path belongs to a top trading cycle.*

Proof. To formalize, consider the request path generated during the procedure:

- If a player p_i proposes to an arm owned by p_j, then p_j (upon receiving the proposal) will propose to their favorite remaining arm, say one owned by p_k. This creates a chain $p_i \to p_j \to p_k \to \cdots$.
- A cycle emerges when this chain loops back to a player already in the sequence (e.g. $p_k \to p_i$).

In a housing market, the cycle is the top trading cycle, and it is detected when a player who previously proposed receives a new proposal, closing the loop. □

A.3 Proof of Lemma 3

Proof. Since players propose to arms based on the indices of their own endowed arm in a round-robin way, no collision occurs and all players can be successfully accepted at each round in the exploration stage of phase 1. Therefore, at the end of the sub-phase $\ell_{\max}$ defined in Eq.(9), it holds that $T_{i,j} \geq 96 \log T/\Delta_{\min}^2$ for any $i, j \in [N]$.

By Lemma 8, once $T_{i,j} \geq 96 \log T/\Delta_{\min}^2$ for all arms a_j, player p_i can recover the true preference ranking under the good concentration event. That is, p_i obtains a permutation σ_i satisfying

$\mathrm{LCB}_{i,\sigma i,j} > \mathrm{UCB}_{i,\sigma i,j+1}$ for all $j \in [N]$. Consequently, during the communication stage of sub-phase $\ell_{\max}$, every player will propose to arm a_j in the j-th communication round. This ensures that in the i-th communication round, player p_i receives N proposals, confirming that all players have successfully learned the true preference profile. Thus, all players synchronously transition to Phase 2 at the end of sub-phase $\ell_{\max}$. □

Lemma 8. *In round t, let $T_i^{(t)} = \min_{j\in[K]} T_{i,j}^{(t)}$. Conditional on $\neg\mathcal{F}$, if $T_i^{(t)} > 96\log T/\Delta_{\min}^2$, we have $LCB_{i,r_{i,k}} > UCB_{i,r_{i,k+1}}^{(t)}$ for any $k \in [N]$.*

Proof. We prove it by contradiction, suppose that there exists $k \in [N]$ such that $LCB_{i,r_{i,k}} \leq UCB_{i,r_{i,k+1}}^{(t)}$. Without loss of generality, denote j as the arm on the LHS and j' as the arm on the RHS. Conditional on $\neg\mathcal{F}$ and by the definition of LCB and UCB, we have that

$$\begin{aligned}
&\boldsymbol{U}(i,j) - 2\sqrt{\frac{6\log T}{T_i^{(t)}}} \\
\leq &\mathrm{LCB}_{i,j}^{(t)} \\
\leq &\mathrm{UCB}_{i,j'}^{(t)} \\
\leq &\boldsymbol{U}(i,j') + 2\sqrt{\frac{6\log T}{T_i^{(t)}}}.
\end{aligned}$$

Therefore, $\Delta_{i,j,j'} = \boldsymbol{U}(i,j) - \boldsymbol{U}(i,j') \leq 4\sqrt{6\log T/T_i^{(t)}}$, which implies that $T_i^{(t)} \leq 96\log T/\Delta_{i,j,j'}^2 \leq 96\log T/\Delta_{\min}^2$, which is a contradiction. □

B Proof of Regret Lower Bound

We will use the following notations throughout the proof of the lower bound.

- For any time t, the number of times arm $j \in [N]$ is played by player p_i is $N_j^{(i)}(t)$.
- Distribution of player $i \in [N]$ and arm $j \in [N]$ is given by $\boldsymbol{\nu}_{i,j}$, which has mean $\boldsymbol{U}(i,j)$. And we assume that for all $i, j \in [N]$, $\boldsymbol{U}(i,j) \in [0,1]$.
- The core matching partner of any player $i \in [N]$ is given by $\mu^*(p_i)$.
- For any player $i \in [N]$, arm $j \in [N]$, $\Delta_j^{(i)} := \boldsymbol{U}(i,\mu^*(p_i)) - \boldsymbol{U}(i,j)$, the arm gap. This can be negative.

B.1 Divergence Decomposition

The proof of the divergence decomposition lemma builds upon the framework presented in Chapter 16 of Lattimore and Szepesvári [2020], extending it to accommodate multiple players. We first introduce some notations as follows.

Canonical multi-player with endowed arm bandit model: We now define the (N-player, N-arm, T-horizon) bandit models, where each player is associated with an arm. The canonical bandit model lies in a measurable space $\{\Omega, \mathcal{F}\}$. Let $A_i(t)$ denote the arm chosen by player p_i at time t. We denote the rejection by the symbol $\emptyset$. Therefore, $A_i(t) \in [N]$, and $X_i(t) \in \mathbb{R} \cup \{\emptyset\}$ for all $i \in [N]$ and $t \in [T]$. Also, $A_i(t)$ and $X_i(t)$ for all $i \in [N]$ and $t \in [T]$ are measurable with respect to $\mathcal{F}$. Let $H(t) = (A_i(t'), X_i(t') : \forall i \in [N], \forall t' \leq t)$ be the random variable representing the history of actions taken and rewards seen up to and including time t. We have $H(t) \in \mathcal{H}(t) \equiv \left([N]^N \times (\mathbb{R} \cup \{\emptyset\})^N\right)^t$. We may set $\Omega \equiv \mathcal{H}(T)$ and the sigma algebra generated by the history as $\mathcal{F} \equiv \sigma\left(H(T)\right)$.

Environment: The bandit environment is specified by $\boldsymbol{\nu} = (\boldsymbol{\nu}_{i,j} : \forall i \in [N], \forall j \in [N])$, where $\boldsymbol{\nu}_{i,j}$ is the distribution of rewards obtained when arm a_j is matched to player p_i in this environment.

Policy: A policy is a sequence of distribution of possible request to the arms from the players (which can assimilate any coordination among the players) conditioned on the past events. More formally, the policy $\boldsymbol{\pi} = \{\boldsymbol{\pi}_t(\cdot) : t \in [T]\}$ where $\boldsymbol{\pi} \equiv \{\pi_t(i, j|\cdot) : \forall i \in [N], \forall j \in [N]\}$ is the function that maps the history up to time $t-1$ to the action $A_i(t), \forall i \in [N]$. Further, $\pi_t(i, j|\cdot) : \mathcal{H}(t-1) \to [0, 1]$ denotes the probability, as a function of history $H(t-1)$ of player p_i playing arm a_j.

Probability Measure: Each environment $\boldsymbol{\nu}$ and policy $\boldsymbol{\pi}$ jointly induces a probability distribution over the measurable space $\{\Omega, \mathcal{F}\}$ denoted by $\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}$. Let $\mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\pi}}$ denote the expectation induced. The density of a particular history up to time T, under an environment $\boldsymbol{\nu}$ and a policy $\boldsymbol{\pi}$, can be defined as

$$\begin{aligned} & d\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left(\boldsymbol{A}(t), \boldsymbol{X}(t) : t \in [T]\right) \\ =& \prod_{t=1}^{T} \pi_t(\boldsymbol{A}(t)|h(t-1)) p_\nu\left(\boldsymbol{X}(t)|\boldsymbol{A}(t)\right) d\lambda\left(\boldsymbol{X}(t); \boldsymbol{\nu}\right) d\varrho\left(\boldsymbol{A}(t)\right). \end{aligned}$$

Here, $\lambda(\boldsymbol{X}; \boldsymbol{\nu}) = \prod_{i=1}^{N} \lambda_i\left(X_i\right)$ is the dominating measure over the rewards with $\lambda_i(X_i) = \delta_\emptyset + \sum_j \boldsymbol{\nu}_{i,j}$[1]. Also, $\varrho(\boldsymbol{A})$ is the counting measure on the collective action of the players.

Lemma 9 (Divergence Decomposition)**.** *For two bandit instances $\boldsymbol{\nu} = \{\boldsymbol{\nu}_{i,j} : i \in [N], j \in [N]\}$, and $\boldsymbol{\nu}' = \left\{\boldsymbol{\nu}'_{i,j} : i \in [N], j \in [N]\right\}$, and any admissible policy $\boldsymbol{\pi}$ the following divergence decomposition is true*

$$D(\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}, \mathbb{P}_{\boldsymbol{\nu}',\boldsymbol{\pi}}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left[N_j^{(i)}(T)\right] D\left(\boldsymbol{\nu}_{i,j}, \boldsymbol{\nu}'_{i,j}\right).$$

Proof. The divergence between two measure, which correspond to two different environments under

[1]Here $\delta_\emptyset$ is the dirac measure on $\emptyset$ denoting the rejection event.

a policy $\boldsymbol{\pi}$, $\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}$ and $\mathbb{P}_{\boldsymbol{\nu}',\boldsymbol{\pi}}$ can be expressed as

$$
\begin{aligned}
& D\left(\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}, \mathbb{P}_{\boldsymbol{\nu}',\boldsymbol{\pi}}\right) \\
=& \mathbb{E}_{\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}}\left[\sum_{t=1}^{T} \log\left(\frac{d\mathbb{P}_{\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}}}{d\mathbb{P}_{\mathbb{P}_{\boldsymbol{\nu}',\boldsymbol{\pi}}}}\right)\right] \\
=& \mathbb{E}_{\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}}\left[\log\left(\frac{p_{\boldsymbol{\nu}}(\boldsymbol{X}(t)|\boldsymbol{A}(t))}{p_{\boldsymbol{\nu}'}(\boldsymbol{X}(t)|\boldsymbol{A}(t))}\right)\right] \qquad (15) \\
=& \mathbb{E}_{\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}}\left[\sum_{t=1}^{T} \log\left(\frac{\prod_{i:X_i(t)\neq\emptyset} p_{\boldsymbol{\nu}}(X_i(t)|\boldsymbol{A}(t))}{\prod_{i:X_i(t)} p_{\boldsymbol{\nu}'}(X_i(t)|\boldsymbol{A}(t))}\right)\right] \qquad (16) \\
=& \mathbb{E}_{\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}}\left[\sum_{t=1}^{T} \sum_{i:X_i(t)\neq\emptyset} \mathbb{E}_{\boldsymbol{\nu}}\left[\log\left(\frac{p_{\boldsymbol{\nu}}(X_i(t)|\boldsymbol{A}(t))}{p_{\boldsymbol{\nu}'}(X_i(t)|\boldsymbol{A}(t))}\right)\middle|\boldsymbol{A}(t)\right]\right] \\
=& \mathbb{E}_{\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}}\left[\sum_{t=1}^{T} \sum_{i:X_i(t)\neq\emptyset} D\left(\boldsymbol{\nu}_{i,A_i(t)}, \boldsymbol{\nu}'_{i,A_i(t)}\right)\right] \qquad (17) \\
=& \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}}\left[\sum_{t=1}^{T} \mathbb{1}\{A_i(t)=j, X_i(t)\neq\emptyset\} D\left(\boldsymbol{\nu}_{i,j}, \boldsymbol{\nu}'_{i,j}\right)\right] \\
=& \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}}\left[N_j^{(i)}(T)\right] D\left(\boldsymbol{\nu}_{i,j}, \boldsymbol{\nu}'_{i,j}\right), \qquad (18)
\end{aligned}
$$

where Eq.(15) comes from the fact that the density of the policy cancels out for the two different environments. Eq.(16) holds because if for some set of actions $\boldsymbol{A}(t)$ player p_i observes $X_i(t)=\emptyset$, it indicates that player p_i is rejected on that round, which is independent of the environment. In particular, we have $p_{\boldsymbol{\nu}}\left(X_i(t)|\boldsymbol{A}(t)\right) = p_{\boldsymbol{\nu}'}\left(X_i(t)|\boldsymbol{A}(t)\right)$ if $X_i(t)=\emptyset$ for any $\boldsymbol{A}(t)$. Eq.(17) comes from the definition of divergence. Eq.(18) uses the definition of $N_j^{(i)}(t)$ the total number of times player p_i successfully plays arm a_j up to time T. □

B.2 Proof of Regret Decomposition (Lemma 4)

Proof. We fix any player p_i, $i \in [N]$ for the rest of the proof. Denote $C_i(T)$ as the number of times that player p_i experiences a collision up to time T. We have the expected regret for the player p_i, under a policy $\boldsymbol{\pi}$ and any bandit instance $\boldsymbol{\nu}$ as

$$
Reg_i(T;\boldsymbol{\nu},\boldsymbol{\pi}) = \sum_{j=1}^{N} \Delta_j^{(i)} \mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left[N_j^{(i)}(T)\right] + \sum_{j=1}^{N} \boldsymbol{U}(i,\mu^*(p_i)) \mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left[C_i(T)\right],
$$

where the first term comes from successfully pulling an arm that is not the core matching partner for player p_i and the second term reflects that for each collision the player p_i obtains 0 reward and hence $\boldsymbol{U}(i,\mu^*(p_i))$ regret in expectation. A trivial regret lower bound is

$$
\begin{aligned}
Reg_i(T;\boldsymbol{\nu},\boldsymbol{\pi}) &\geq \sum_{j=1}^{N} \Delta_j^{(i)} \mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left[N_j^{(i)}(T)\right] \\
&= \sum_{a_j \neq \mu^*(p_i)} \Delta_j^{(i)} \mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left[N_j^{(i)}(T)\right].
\end{aligned}
$$

For an STTCB instance, the core matching partner of a player is her optimal arm, and if two or more players propose to the same arm in one round, it would result in a collision and all of these players would get a reward of 0. Therefore, if any of the other $N-1$ players propose to pull $\mu^*(p_i)$, player p_i should either move to a sub-optimal arm or it is blocked. In the best possible scenario, the player p_i successfully plays its second best arm, in each of these instances. We have $\Delta_{\min}^{(i)} \leq \boldsymbol{U}(i, \mu^*(p_i))$ for non-negative rewards, denote $\bar{N}_{\mu^*(p_i)^{(i')}}$ as the number of times player $p_{i'}$ proposes to arm $\mu^*(p_i)$, the regret from the events when any of the other $N-1$ players propose to pull arm $\mu^*(p_i)$ is lower bounded by

$$\begin{aligned} Reg_i(T; \boldsymbol{\nu}, \boldsymbol{\pi}) &\geq \sum_{i' \neq i} \Delta_{\min}^{(i)} \mathbb{E}_{\boldsymbol{\nu}, \boldsymbol{\pi}} \left[\bar{N}_{\mu^*(p_i)}^{(i')}(T) \right] \\ &\geq \sum_{i' \neq i} \Delta_{\min}^{(i)} \mathbb{E}_{\boldsymbol{\nu}, \boldsymbol{\pi}} \left[N_{\mu^*(p_i)}^{(i')}(T) \right]. \end{aligned}$$

The last inequality comes from the fact that player $p_{i'}$ could successfully pull arm $\mu^*(p_i)$ when only one player proposes to this arm. Therefore, the combined regret bound is given as

$$Reg_i(T; \boldsymbol{\nu}, \pi) \geq \max \left\{ \sum_{i' \neq i} \Delta_{\min}^{(i)} \mathbb{E}_{\boldsymbol{\nu}, \pi} \left[N_{\mu^*(p_i)}^{(i')}(T) \right], \sum_{a_j \neq \mu^*(p_i)} \Delta_j^{(i)} \mathbb{E}_{\boldsymbol{\nu}, \pi} \left[N_j^{(i)}(T) \right] \right\}.$$

□

B.3 Proof of Regret Lower Bound (Theorem 2)

Proof. We consider any instance in the class of STTCB $\boldsymbol{\nu}$, universally consistent policy $\boldsymbol{\pi}$, player p_i for $i \in [N]$, and arm a_j for $j \in [N]$. Let us consider the instance $\boldsymbol{\nu'}$ (which is specific to the i and j pair) where $\boldsymbol{\nu}'_{i',j'} = \boldsymbol{\nu}_{i',j'}$ for all $i' \neq i, j' \neq j$, and $\boldsymbol{\nu}'_{i,j}$ such that $D\left(\boldsymbol{\nu}_{i,j}, \boldsymbol{\nu}'_{i,j}\right) \leq D_{\inf}\left(\boldsymbol{\nu}_{i,j}, \boldsymbol{U}(i, \mu^*(p_i)), \mathcal{P}\right) + \epsilon$ for some $\epsilon > 0$ and $\boldsymbol{U}'(i,j) \equiv \rho(\boldsymbol{\nu}'_{i,j}) > \boldsymbol{U}(i, \mu^*(p_i))$ where ρ is the operator mapping a distribution to its mean. Note, for $\boldsymbol{U}(i, \mu^*(p_i)) < 1$ and $\Delta_j^{(i)} > 0$, which holds by assumption, the distribution $\boldsymbol{\nu}'_{i,j}$ exists by definition of $D_{\inf}(\cdot)$. In short, for the i-th player we make the j-th arm optimal. The optimal arm for player i in the instance $\boldsymbol{\nu'}$ is the arm j.

For any event A (and its complement A^c), by Lemma 11 in Appendix C, we have

$$D\left(\mathbb{P}_{\boldsymbol{\nu}, \boldsymbol{\pi}}, \mathbb{P}_{\boldsymbol{\nu}', \boldsymbol{\pi}}\right) \geq \log \left(\frac{1}{2\left(\mathbb{P}_{\boldsymbol{\nu}, \boldsymbol{\pi}}(A) + \mathbb{P}_{\boldsymbol{\nu}', \boldsymbol{\pi}}(A^c)\right)} \right). \tag{19}$$

Let us now consider the event $A = \left\{ N_j^{(i)}(T) \geq \frac{T}{2} \right\}$. Then, due to the regret decomposition (Lemma 4), we have the regrets:

1. In instance $\boldsymbol{\nu}$ as $Reg(T; \boldsymbol{\nu}, \boldsymbol{\pi}) \geq Reg_i(T; \boldsymbol{\nu}, \boldsymbol{\pi}) \geq \Delta_j^{(i)} \frac{T}{2} \mathbb{P}_{\boldsymbol{\nu}, \boldsymbol{\pi}} \left(\left\{ N_j^{(i)}(T) \geq \frac{T}{2} \right\} \right)$.
2. In instance $\boldsymbol{\nu'}$ as $Reg(T; \boldsymbol{\nu}', \boldsymbol{\pi}) \geq Reg_i(T; \boldsymbol{\nu}', \boldsymbol{\pi}) \geq (\boldsymbol{U}'(i,j) - \boldsymbol{U}(i, \mu^*(p_i))) \frac{T}{2} \mathbb{P}_{\boldsymbol{\nu}, \boldsymbol{\pi}} \left(\left\{ N_j^{(i)}(T) < \frac{T}{2} \right\} \right)$.

As the only change in reward distribution happens in player p_i, arm a_j pair, by the divergence decomposition (Lemma 9), we have

$$\begin{aligned} & D(\mathbb{P}_{\boldsymbol{\nu}, \boldsymbol{\pi}}, \mathbb{P}_{\boldsymbol{\nu}', \boldsymbol{\pi}}) \\ =& D(\boldsymbol{\nu}_{i,j}, \boldsymbol{\nu}'_{i,j}) \mathbb{E}_{\boldsymbol{\nu}, \boldsymbol{\pi}} \left[N_j^{(i)}(T) \right] \\ \leq & \left(\epsilon + D_{\inf}\left(\boldsymbol{\nu}_{i,j}, \boldsymbol{U}(i, \mu^*(p_i)), \mathcal{P}\right)\right) \mathbb{E}_{\boldsymbol{\nu}, \boldsymbol{\pi}} \left[N_j^{(i)}(T) \right], \end{aligned}$$

where the last inequality comes from the construction of $\boldsymbol{\nu}'_{i,j}$.

Substituting the above three relations in Eq.(19) we obtain for any $\epsilon > 0$:

$$\begin{aligned}
&\left(\epsilon + D_{\inf}\left(\boldsymbol{\nu}_{i,j}, \boldsymbol{U}(i, \mu^*(p_i)), \mathcal{P}\right)\right) \mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left[N_j^{(i)}(T)\right] \\
&\geq \log\left(\frac{1}{2\left(\mathbb{P}_{\boldsymbol{\nu},\boldsymbol{\pi}}(A) + \mathbb{P}_{\boldsymbol{\nu}',\boldsymbol{\pi}}(A^c)\right)}\right) \\
&\geq \log\left(\frac{T \min\left\{\boldsymbol{U}'(i,j) - \boldsymbol{U}(i, \mu^*(p_i)), \Delta_j^{(i)}\right\}}{4\left(Reg_i(T; \boldsymbol{\nu}, \boldsymbol{\pi}) + Reg_i(T; \boldsymbol{\nu}', \boldsymbol{\pi})\right)}\right),
\end{aligned}$$

where the last inequality holds since the policy $\boldsymbol{\pi}$ is assumed to be universally consistent. Taking ϵ and T to the limit, we have

$$\begin{aligned}
&\lim_{\epsilon \to 0} \liminf_{T \to \infty} \frac{\mathbb{E}_{\boldsymbol{\nu},\boldsymbol{\pi}}\left[N_j^{(i)}(T)\right]}{\log T} \\
&\geq \lim_{\epsilon \to 0} \frac{1}{\epsilon + D_{\inf}\left(\boldsymbol{\nu}_{i,j}, \boldsymbol{U}(i, \mu^*(p_i)), \mathcal{P}\right)} \\
&= \frac{1}{D_{\inf}\left(\boldsymbol{\nu}_{i,j}, \boldsymbol{U}(i, \mu^*(p_i)), \mathcal{P}\right)}
\end{aligned}$$

The above bound is true for any uniformly consistent policy $\boldsymbol{\pi}$, and for player p_i, and arm a_j for $i, j \in [N]$. We use the regret decomposition (Lemma 4) to obtain the final asymptotic regret lower bound for any player p_i as

$$\begin{aligned}
&\liminf_{T \to \infty} \frac{Reg_i(T; \boldsymbol{\nu}, \pi)}{\log T} \\
&\geq \max\left\{\sum_{i' \neq i} \frac{\Delta_{\min}^{(i)}}{D_{\inf}(\boldsymbol{\nu}_{i',\mu^*(p_i)}, \boldsymbol{U}(i', \mu^*(p_{i'})), \mathcal{P})}, \sum_{a_j \neq \mu^*(p_i)} \frac{\Delta_j^{(i)}}{D_{\inf}(\boldsymbol{\nu}_{i,j}, \boldsymbol{U}(i, \mu^*(p_i)), \mathcal{P})}\right\}.
\end{aligned}$$

□

B.4 Proof of Corollary 1

The above corollary follows readily from Theorem 2. Consider the following STTCB instance, let the optimal arm for any player p_i be a_{i+1}, then all players together form a top trading cycle. For a given index i, let for any player $p_{i'}$, $i' \neq i$, set $\boldsymbol{U}(i', i'+1) = \frac{1}{2}$ while for any other arm a_j, $j \neq i'+1$, set $\boldsymbol{U}(i', j) = \frac{1}{2} - \Delta$, where $\Delta > 0$ is small enough. Also, let player p_i has the arm mean for a_{i+1} be $\boldsymbol{U}(i, i+1) = \frac{1}{2}$ and for any other arms a_j, $j \neq i+1$ set $\boldsymbol{U}(i, j) = \frac{1}{4}$. For $\mathcal{P}$ the class of Bernoulli rewards, we have

$$D_{\inf}\left(\boldsymbol{\nu}_{i',\mu^*(p_i)}, \boldsymbol{U}(i', \mu^*(p_i)), \mathcal{P}\right) \leq \frac{\Delta^2}{4}, \forall i' \neq i,$$

and $\Delta_{\min}^{(i)} = \frac{1}{4}$. Therefore, the regret for player p_i is lower bounded by $\frac{(N-1)\log T}{16\Delta^2}$.

C Technical Lemmas

Lemma 10 (Corollary 5.5 in Lattimore and Szepesvári [2020]). *Assume that $X_1, X_2, \cdots, X_n$ are independent σ-subgaussian random variables centered around μ. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu + \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$
$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq \mu - \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

Lemma 11 (Bretagnolle-Huber Inequality, Theorem 14.2 [Lattimore and Szepesvári, 2020]). *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp\left(-D(P,Q)\right),$$

where $A^c = \Omega \backslash A$ is the complement of A.