Learn to Match with No Regret: Reinforcement Learning in Markov Matching Markets

Yifei Min
Department of Statistics and Data Science
Yale University
New Haven, CT 06511
yifei.min@yale.edu

Tianhao Wang
Department of Statistics and Data Science
Yale University
New Haven, CT 06511
tianhao.wang@yale.edu

Ruitu Xu
Department of Statistics and Data Science
Yale University
New Haven, CT 06511
ruitu.xu@yale.edu

Zhaoran Wang
Departments of Industrial Engineering
and Management Sciences
Northwestern University
Evanston, IL 60208
zhaoranwang@gmail.com

Michael I. Jordan
Department of Electrical Engineering
and Computer Science
University of California, Berkeley
CA 94720
jordan@cs.berkeley.edu

Zhuoran Yang
Department of Statistics and Data Science
Yale University
New Haven, CT 06511
zhuoran.yang@yale.edu

Abstract

We study a Markov matching market involving a planner and a set of strategic agents on the two sides of the market. At each step, the agents are presented with a dynamical context, where the contexts determine the utilities. The planner controls the transition of the contexts to maximize the cumulative social welfare, while the agents aim to find a myopic stable matching at each step. Such a setting captures a range of applications including ridesharing platforms. We formalize the problem by proposing a reinforcement learning framework that integrates optimistic value iteration with maximum weight matching. The proposed algorithm addresses the coupled challenges of sequential exploration, matching stability, and function approximation. We prove that the algorithm achieves sublinear regret.

1 Introduction

Large-scale digital markets play a crucial role in modern economies. Increasingly, the understanding and design of such markets require tools from both economics and machine learning. Indeed, dynamically changing market environments require mechanisms to be adaptive, scalable, and incentive compatible in the face of significant nonstationarity. The data streams that arise from digital markets provide opportunities to cope with such challenges, via learning-based mechanism design. Recent work (Jagadeesan et al., 2021; Liu et al., 2021; Sankararaman et al., 2021; Basu et al., 2021) has begun to apply modern machine learning tools to problems in adaptive mechanism design. One particular area of focus in learning-aware market design has been matching, a class of problems

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

central to microeconomics (Mas-Colell et al., 1995). Existing work has focused on static matching markets, however, and the more challenging yet critically important setting of dynamic matching markets has been neglected. Such markets are our focus in the current paper, where we provide reinforcement learning (RL) techniques to address dynamic matching market problems.

Specifically, we propose a *Markov matching market* model involving a planner and a set of two-sided agents. Given the time horizon $H \in \mathbb{Z}_+$, a set of agents $I_h \cup J_h$ enter the market in context C_h at step $h \in [H]$, with unknown utility functions u_h and v_h for agents in I_h and J_h respectively. The agents' utilities u_h and v_h depend on the context C_h. Given the utility functions, the agents seek to achieve a myopic stable matching with transferable utilities (Shapley and Shubik, 1971). In particular, the contexts C_h are subject to a Markov transition kernel which is controlled by the planner's policy. The goal of the planner is to select an optimal policy, which together with the stable matching in each step maximizes the expected accumulated social welfare.

As an illustration, consider a simplified abstraction of ride-hailing platforms (Qin et al., 2020; Özkan and Ward, 2020; Hu and Zhou, 2021), where the horizon is set as the time span of a day. In this case, the platform is the planner, and the two-sided agents I_h and J_h are the drivers and the riders. The contexts may include information such as GPS location, car types, and pricing. The platform is required to ensure the stability of the matching at each step so that the agents do not prefer alternative outcomes, otherwise they may leave the platform. In addition, the platform needs to adjust the policy to maximize the accumulated social welfare as a measure of the satisfaction of both drivers and riders.

As illustrated in this example, there are several key challenges in Markov matching markets. First, the agents' utilities and the transition of contexts are unknown, so we need to perform efficient exploration to collect the required information. Second, we need to ensure the stability of the matching in each step, for which a consideration of only the difference in total utilities is insufficient, as discussed in Jagadeesan et al. (2021). To this end, we adopt a notion of *Subset Instability* (SI) from Jagadeesan et al. (2021) as a metric that quantifies the distance between a proposed matching and the optimal stable matching. This metric has the flavor of Shapley value, comparing the discrepancy between the total utilities over all subsets of the participating agents, while accounting for transfers. For efficient estimation, we need to take into account the function class of the utilities as well as the Markov transition kernel, thus demanding a systematic usage of function approximation.

To tackle these challenges, we develop a novel algorithm called *Sequential Optimistic Matching* (SOM), which features a combination of optimistic value iteration and max-weight matching. Note that the planner's problem becomes a standard Markov decision process (MDP) if we regard the value of the max-weight matching over the true utilities as the reward. Inspired by this observation, on the agents side, our algorithm applies the optimism principle to construct UCB estimates of the utilities. Based on these UCB estimates, the algorithm computes the corresponding max-weight matching, and the values of the resulting matching serve as the surrogate rewards. Although we do not have access to unbiased estimate of the rewards, the key property here is that the surrogate rewards upper bound the true social welfare, thus justifying optimistic planning by the planner. Interestingly, our framework can readily incorporate any online RL algorithm based on optimism.

On the theoretical side, we show that the suboptimality of the accumulated social welfare for our algorithm consists of two parts: (1) the planner's regret in terms of the suboptimality of the policy, and (2) the agents' regret in terms of SI. Jagadeesan et al. (2021) proved that SI can be bounded by the sum of the optimistic bonuses, and we further show that the planner's regret can also be bounded by the bonus sum. In this way, we reconcile the seemingly independent learning goals of the planner and the agents, and thereby provide a unified approach to controlling the suboptimality of the total social welfare. In particular, based on the above decomposition, we further show that in the case of linear function approximation, our algorithm enjoys a sublinear regret independent of the size of the context space.

Compared with existing methods on online RL such as `LSVI-UCB` (Jin et al., 2020), our framework incorporates a matching problem in each time step. Compared with the setup of matching bandits (Jagadeesan et al., 2021), our model provides an extension to a dynamic setup with transitions between contexts. More importantly, our work lies beyond a straightforward combination of online RL with matching bandits due to the unique technical challenges that we will discuss in Section 4 and further elaborate in Section C.

The main contributions are summarized as follows:

(i) We propose a novel Markov matching market model that captures a range of instances of centralized matching problems.

(ii) We develop a novel algorithm that combines optimistic value iteration with max-weight matching, such that any online RL algorithm based on optimism can be readily incorporated into the framework.

(iii) We provide a general analysis framework and show that our proposed algorithms achieve sublinear regret under proper structural assumptions on the underlying model.

2 Related Work

There is an emerging line of research on learning stable matchings with bandit feedback (Das and Kamenica, 2005; Liu et al., 2020, 2021; Sankararaman et al., 2021; Cen and Shah, 2021; Basu et al., 2021) using the mature tools from the bandit literature. Most of them focus on matchings with non-transferable utilities (Gale and Shapley, 1962), which fails to capture real-world markets with monetary transfers between agents, e.g., payments from passengers to drivers on ride-hailing platforms. The study of learning for matchings with transferable utilities is comparably limited, and our work extends Jagadeesan et al. (2021) to dynamic scenarios in this regime. Broadly speaking, our work is also related to the research on learning economic models via RL. In particular, Kandasamy et al. (2020); Rasouli and Jordan (2021) studied VCG mechanisms, and Guo et al. (2021) study exchange economies. Although similar in spirit, these models differ from our Markov matching markets in their mathematical structure, and thus have different solution concepts and planning methods.

Our model of Markov matching markets is related to the topic of dynamic matching in the economics literature (Taylor, 1995; Satterthwaite and Shneyerov, 2007; Niederle and Yariv, 2009; Ünver, 2010; Anderson et al., 2014; Lauermann and Nöldeke, 2014; Leshno, 2019; Akbarpour et al., 2020; Baccara et al., 2020; Loertscher et al., 2018; Doval and Szentes, 2019). Instead of studying the learning problem in matching markets, the goal therein is mainly focused on the problem of optimal mechanism design (Akbarpour et al., 2014) with known utilities or explicit modelling of agents' arrivals via queuing models (Zenios, 1999; Gurvich and Ward, 2015). In this work we focus on the notion of (static) stability of myopic matchings (Shapley and Shubik, 1971). There is also a line of literature on the notion of dynamic stability for matching markets (Damiano and Lam, 2005; Doval, 2014; Kadam and Kotowski, 2018; Doval, 2019; Kotowski, 2019; Kurino, 2020; Liu, 2020), and it is an interesting open problem to study learning for dynamically stable matchings.

Our methodology builds upon recent progress in online RL, where the "optimism in face of uncertainty" principle has engendered efficient algorithms that are either model-based (Jaksch et al., 2010; Osband et al., 2016; Azar et al., 2017; Dann et al., 2017) or model-free (Strehl et al., 2006; Jin et al., 2018; Fei et al., 2020, 2021a), and can be combined with function approximation techniques (Yang and Wang, 2019; Jin et al., 2020; Zanette et al., 2020; Ayoub et al., 2020; Wang et al., 2020; Fei et al., 2021b; Yang et al., 2020; Zhou et al., 2021; Min et al., 2021a,b; Du et al., 2021; Jin et al., 2021). We note that these approaches can be incorporated into our framework with proper modifications on structural assumptions and correspondingly the algorithm. We do not pursue these extensions here, however, as our regret analysis is already sufficiently challenging given the lack of an unbiased estimate of the reward, and the additional constraints imposed by the requirements of matching stability.

3 Preliminaries

3.1 Markov Matching Markets

We first review basic concepts for matching with transferable utilities (Shapley and Shubik, 1971). Consider a two-sided matching market where $\mathcal{I}$ denotes the set of all side-1 agents (e.g., buyers) and $\mathcal{J}$ denotes the set of all side-2 agents (e.g., sellers). Given any set of participating agents $I \times J \in 2^{\mathcal{I}} \times 2^{\mathcal{J}}$, a matching X is a set of pairs (i, j) indicating $i \in I$ is matched with $j \in J$, and each agent can be matched at most once. For any pair of agents $(i, j) \in I \times J$, we denote by $u(i, j)$ the utility of agent i and $v(i, j)$ the utility of agent j when they are matched. In addition to the matching X, we also allow transfers between agents, summarized by the transfer function

$\tau : I \cup J \to \mathbb{R}$. For each agent $i \in I \cup J$, $\tau(i)$ is the transfer that it receives. We assume that the transfers are within agents, so $\sum_{i \in I \cup J} \tau(i) = 0$.

The overall market outcome is denoted by a tuple (X, τ), where X represents the matching and τ represents transfers. For any $(i, j) \in X$, the total utilities are $u(i, j) + \tau(i)$ and $v(i, j) + \tau(j)$ for i and j respectively. Moreover, if no agents prefer any alternate outcome, then we say (X, τ) is stable (see Definition B.1 for details). The stable matching can be found by solving the corresponding max-weight matching, as will be explained in Section 4.1.

Based on these (classical) definitions, we formulate the notion of a *Markov matching market* involving a planner and a set of two-sided agents. Throughout the paper, we focus on matchings with transferable utilities between two-sided agents, and we may omit such descriptions for convenience.

Definition 3.1 (Markov matching markets)**.** A Markov matching market is denoted by a tuple $M = (\mathcal{C}, \Upsilon, \{I_h\}_{h=1}^H, \{J_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H, \{u_h\}_{h=1}^H, \{v_h\}_{h=1}^H)$. Here $\mathcal{C}$ is the set of contexts and Υ denotes the set of planner's actions. At each step $h \in [H]$, $I_h \cup J_h \subset \mathcal{I} \times \mathcal{J}$ is the set of participating agents, and $u_h : \mathcal{C} \times \Upsilon \times \mathcal{I} \times \mathcal{J} \to \mathbb{R}$ and $v_h : \mathcal{C} \times \Upsilon \times \mathcal{I} \times \mathcal{J} \to \mathbb{R}$ are utility functions for two sides of agents respectively. For each $h \in [H]$, $\mathbb{P}_h(C' \mid C, e)$ is the transition probability for context C to transit to C' given action e.

In such Markov matching markets, the learning goal is two-fold: (1) to learn the stable matching in each step and (2) to maximize the accumulated social welfare.

3.2 A Reinforcement Learning Approach to Markov Matching Markets

We now translate the problem into the language of RL. Given Definition 3.1, we consider an episodic setting with K episodes where each episode consists of H steps of sequential matchings. Each episode proceeds in the following way: at each step $h \in [H]$ under context C_h, a set of agents $I_h \cup J_h$ enter the market. The planner takes action e_h, implement the matching (X_h, τ_h), and observes the noisy feedback of utilities $u_h(C_h, e_h, i, j)$ and $v_h(C_h, e_h, i, j)$ for all $(i, j) \in X_h$. Then the context transitions according to $\mathbb{P}_h(\cdot \mid C_h, e_h)$, and the market proceeds to the next step.

Note that here the implemented matching in each step is myopic and we seek for stability for each of these matchings. The maximization of the accumulated social welfare is achieved through the planner's action $\{e_h\}_{h=1}^H$ that controls the transitions of contexts, which together with planner's actions determine the optimal values of the matchings. To apply RL to maximize the accumulated social welfare, let us specify the ingredients of the corresponding RL problem.

States and Actions. The state space is $\mathcal{S} = \mathcal{C} \times 2^{\mathcal{I}} \times 2^{\mathcal{J}}$. The action space is $\mathcal{A} = \Upsilon \times \mathcal{X} \times \mathcal{T}$ where $\mathcal{X}$ denotes the set of all matchings and $\mathcal{T}$ is the set of all possible transfers among the agents. For each step $h \in [H]$, the state $s_h = (C_h, I_h, J_h)$ contains the context and participating agents, and the action $a_h = (e_h, X_h, \tau_h)$ contains the planner's action e_h and the matching outcome (X_h, τ_h) for the agents.

Rewards. At each step $h \in [H]$, given the state $s_h = (C_h, I_h, J_h)$ and the action $a_h = (e_h, X_h, \tau_h)$, the immediate reward is the social welfare (i.e. sum of utilities):

$$r_h(s_h, a_h) \coloneqq \sum\nolimits_{(i,j) \in X_h} [u_h(C_h, e_h, i, j) + v_h(C_h, e_h, i, j)]. \tag{1}$$

Note that the transfer $\tau_h \in \mathcal{T}$ does not appear in the reward since the total transfer sums to zero.

Transition of States. The state consists of the context in $\mathcal{C}$ and the sets of agents in $2^{\mathcal{I}} \times 2^{\mathcal{J}}$. The transition of context at step h follows the heterogeneous transition function $\mathbb{P}_h(C_{h+1} \mid C_h, e_h)$, which only depends on the planner's action e_h and is independent of the matching (X_h, τ_h).

We assume that the sequence of agent sets $\{I_h, J_h\}_{h=1}^H$ is generated independently from other components in this matching market. We also assume the same sequence through all K episodes for the sake of clarity. Note that I_h and J_h can also be handled as part of the context C_h and covered by our current argument with more involved transition dynamics, which is often task-specific. Our minor simplification serves to build a generic framework and avoid detailed modeling of the agent sets.

Policies and Value Functions. A policy π is defined as $\pi = \{\pi_h\}_{h=1}^H$, where for each $s \in \mathcal{S}$, $\pi_h(\cdot|s)$ is a distribution on $\mathcal{A}$. The policy consists of two parts: the planner's part (i.e. choosing $e \in \Upsilon$ to influence the market context C), and the agents' part (i.e. determining the matching-transfer (X, τ)). We use Π to denote the set of such policies.

For any policy π, we define each value function $V_h^\pi(\cdot)$ as

$$V_h^\pi(s) \coloneqq \mathbb{E}_\pi\Big[\sum\nolimits_{l=h}^{H} r_l(s_l, a_l) \;\Big|\; s_h = s;\; a_l \sim \pi_l(\cdot|s_l),\; s_{l+1} \sim \mathbb{P}_l(\cdot|s_l, a_l), \forall\, h \le l \le H\Big]. \quad (2)$$

Maximizing the accumulated social welfare is equivalent to maximizing V_1^π over $\pi \in \Pi$, so the overall performance over K episodes is evaluated through the regret

$$R(K) \coloneqq \sum\nolimits_{k=1}^{K} \Big[\max_{\pi\in\Pi} V_1^\pi(s_1) - V_1^{\pi_k}(s_1)\Big], \quad (3)$$

where π_k denotes the policy in episode k. See Appendix A for detailed definitions of our notation.

4 Method: An Optimistic Meta Algorithm

The dynamic matching problem introduced in Sec. 3 suffers from several coupled challenges. Specifically: 1) we study the dynamic setting and therefore the algorithm proposed by (Jagadeesan et al., 2021) for the bandit setting (i.e., $H = 1$) does not apply; 2) the dynamic regime also requires us to handle the propagation of error through H steps in the analysis; 3) compared with standard MDPs in RL, here we aim to maximize the social welfare and thus the regret involves a new notion called Subset Instability, which is a nonlinear functional of the rewards of the agents. Regret decomposition involving such a complex quantity was never considered in previous works (Jaksch et al., 2010; Osband et al., 2016; Azar et al., 2017; Jin et al., 2020; Ayoub et al., 2020; Wang et al., 2020; Fei et al., 2021b; Yang et al., 2020; Zhou et al., 2021).

We now introduce an algorithm for the Markov matching market that addresses all these challenges. This algorithm involves solving a sequence of combinatorial optimization problems to obtain a UCB estimate (Sec. 4.1). Moreover, for the purpose of the theoretical analysis, as the regret is in terms of Subset Instability, we need to connect the suboptimality measured by this metric with the uncertainty of the value function estimation. We thus incorporate a novel decomposition of the planner and the agents in the algorithm design (Sec. 4.2).

In Sec. 4.3 we summarize our proposed Algorithm 1 which serves as a meta stereotype and can readily incorporate various existing RL methods.

4.1 Optimistic Estimation of Rewards

Note that we do not directly observe the rewards defined in (1) and have no unbiased estimates of them, so we cannot explicitly construct their optimistic estimates. However, thanks to the nature of the stable matching being a max-weight matching, we show that we can still obtain useful optimistic rewards estimates.

To see this, recall the definition of reward in (1). We know that there exists some stable matching (X_h, τ_h) that maximizes r_h and can be obtained by solving a linear program and its dual program (Shapley and Shubik, 1971). Define the following linear program $\mathcal{LP}(I, J, u, v)$:

$$\begin{aligned} &\max_{w\in\mathbb{R}^{|I|\times|J|}} \sum\nolimits_{(i,j)\in I\times J} w_{i,j}\,[u(i,j) + v(i,j)] \\ &\text{s.t.}\;\; \text{i)} \sum\nolimits_{j\in J} w_{i,j} \le 1, \forall\, i \in I;\; \text{ii)} \sum\nolimits_{i\in I} w_{i,j} \le 1, \forall\, j \in J;\; \text{iii)}\; w_{i,j} \ge 0, \forall\, (i,j) \in I\times J, \end{aligned} \quad (4)$$

and its dual program $\mathcal{DP}(I, J, u, v)$:

$$\min_{p: I\cup J \to \mathbb{R}^+} \sum\nolimits_{a\in I\cup J} p(a) \qquad \text{s.t.} \quad p(i) + p(j) \ge u(i,j) + v(i,j), \forall (i,j) \in I\times J. \quad (5)$$

Shapley and Shubik (1971) proved that the stable (X, τ) correspond to the solution to the linear program (4) (for X) and its dual program (5) (for τ). It is clear from (4) that the optimal value of the objective function is equal to the total social welfare of the stable matching. Now, suppose we have some optimistic estimates of the utilities, i.e., $\widehat{u}$ and $\widehat{v}$ such that $\widehat{u}(\cdot,\cdot) \ge u(\cdot,\cdot)$ and $\widehat{v}(\cdot,\cdot) \ge v(\cdot,\cdot)$. It is easy to see that when substituting $(\widehat{u}, \widehat{v})$ into the linear program (4), the resulting optimal value will be an upper bound of the original optimal value (see the proof of Lemma C.2).

Based on this observation, let us return to the reward in (1). The previous argument implies that as long as we have optimistic estimates of the utilities u_h and v_h, we can get optimistic estimates of the

reward by solving the max-weight matching based on the optimistic utilities. Moreover, it is further an upper bound of the following pseudo-reward:

$$\overline{r}_h(C_h, I_h, J_h, e_h) \coloneqq \max_{(X_h,\tau_h)\in\mathcal{M}_h} r_h(C_h, I_h, J_h, e_h, X_h, \tau_h), \tag{6}$$

where $\mathcal{M}_h \coloneqq \mathcal{M}(I_h, J_h, u_h, v_h, C_h, e_h)$ denotes the set of all myopic stable matching on (I_h, J_h) with utility functions $u_h(C_h, e_h, \cdot, \cdot)$ and $u_h(C_h, e_h, \cdot, \cdot)$. Finally, the optimistic estimates of the utilities can be constructed from noisy observations of agents' utilities via any standard approach in the online learning literature.

The definition of the pseudo-reward in (6) provides a way to decompose the total regret into the planner's regret and the agents' regret, as will be clear in the next subsection.

4.2 Decomposition of The Planner and The Agents

Recall that we require the matching in each step to be stable, which is an additional constraint apart from maximizing the social welfare. We need to separate these two entangled goals from each other. Indeed, we will show that the total regret consists of two parts: 1) the suboptimality of the planner's policy *over the entire episode*, and 2) the distance between the proposed matching and the optimal myopic stable matching *at each step*. We identify the former as the planner's problem and the latter the agents' problem.

The Planner's Problem. The planner's problem focuses on the transition of the contexts, so we need to partial out the effects from the actual matching. This has been done in the definition of the pseudo-reward in (6), and the corresponding pseudo-value function $\overline{V}_h^\pi$ for $h \in [H]$ is defined as

$$\overline{V}_h^\pi(s) \coloneqq \mathbb{E}_\pi\Big[\sum\nolimits_{l=h}^{H} \overline{r}_l(s_l, e_l) \;\Big|\; s_h = s,\ e_l \sim \pi_l(\cdot|s_l),\ s_{l+1} \sim \mathbb{P}_l(\cdot|s_l, e_l), \forall\, h \le l \le H\Big], \tag{7}$$

where we slightly abuse the notation $e_l \sim \pi_l(\cdot|s_l)$. Also note that we can write $s_{h+1} \sim \mathbb{P}(\cdot|s_h, e_h)$ instead of the more general $s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h)$ since we condition on (I_h, J_h) and the transition of C_h only depends on the planner's action e_h as $C_{h+1} \sim \mathbb{P}_h(\cdot|C_h, e_h)$.

Clearly, $\overline{V}_h^\pi$ is an upper bound of V_h^π and does not depend on the actual matching $\{X_h, \tau_h\}_{h\in[H]}$ since it has been maximized out. Now, we specify the planner's problem as trying to maximize the pseudo-value $\overline{V}_1^\pi$, and define the planner's regret given the initial state s_1 as

$$R^P(K) \coloneqq \sum\nolimits_{k=1}^{K}\Big[\max_\pi \overline{V}_1^\pi(s_1) - \overline{V}_1^{\pi_k}(s_1)\Big] = \sum\nolimits_{k=1}^{K}\Big[\overline{V}_1^\star(s_1) - \overline{V}_1^{\pi_k}(s_1)\Big]. \tag{8}$$

From a control-theoretic perspective, the planner's problem can be viewed as learning an MDP with the same state space $\mathcal{S}$, the action space reduced to Υ, and reward being the value of the myopic max-weight matching at each step. The reward cannot be observed, nor do we have an unbiased estimator. From an economic perspective, the planner's problem captures only the market context and not the specific market outcome (i.e. matching). Now, note that

$$R(K) = \underbrace{\sum\nolimits_{k=1}^{K}\Big[\max_{\pi\in\Pi} V_1^\pi(s) - \overline{V}_1^{\pi_k}(s_1)\Big]}_{\textbf{(Planner's regret)}} + \underbrace{\sum\nolimits_{k=1}^{K}\Big[\overline{V}_1^{\pi_k}(s_1) - V_1^{\pi_k}(s_1)\Big]}_{\textbf{(Utility difference)}},$$

where the planner's regret has been captured in (8), and it remains to control the utility difference on the agents' side.

Agents' Problem. The agents' problem amounts to controlling the suboptimality of each implemented matching, which boils down to SI proposed by Jagadeesan et al. (2021).

Definition 4.1 (Subset Instability, Jagadeesan et al. 2021)**.** Given any agent sets $I,\ J$ and utility functions $u,\ v : I \times J \to \mathbb{R}$, the Subset Instability SI$(X, \tau; I, J, u, v)$ of the matching and transfer (X, τ) is defined as

$$\max_{I'\times J'\subseteq I\times J}\Big[\Big(\max_{X'}\sum_{i\in I'} u(i, X'(i)) + \sum_{j\in J'} v(X'(j), j)\Big) - \sum_{i\in I'}\big(u(i, X(j)) + \tau(i)\big) - \sum_{j\in J'}\big(v(X(j), j) + \tau(j)\big)\Big],$$

where $X(\cdot)$ and $X'(\cdot)$ denotes the matched agent in matching X and X' respectively.

Algorithm 1 Sequential Optimistic Matching (SOM)

Require: $\lambda, \beta_u, \beta_V$
Initialize: $u_h^1 \equiv 1, v_h^1 \equiv 1$ and $\mathcal{D}_h^0 = \emptyset, \forall h \in [H]$
for episode $k = 1, 2, \ldots, K$ **do**
  Receive the initial state $s_1^k = (C_1^k, I_1, J_1)$
  Set $\overline{Q}_{H+1}^k \equiv 0$
  for stage $h = H, H-1, \ldots, 1$ **do**
    Esti. utilities $(u_h^k, v_h^k) \leftarrow \texttt{UE}(\mathcal{D}_h^{k-1}, \beta_u, \lambda)$
    Esti. pseudo-reward $\overline{r}_h^k \leftarrow \texttt{RE}(u_h^k, v_h^k, I_h, J_h)$
    Esti. Q-func. $\overline{Q}_h^k \leftarrow \texttt{QE}(\mathcal{D}_h^{k-1}, \overline{r}_h^k, \overline{Q}_{h+1}^k, \beta_V, \lambda)$
  end for
  for stage $h = 1 \ldots, H$ **do**
    Planner takes action $e_h^k \leftarrow \operatorname{argmax}_{e \in \Upsilon} \overline{Q}_h^k(s_h^k, e)$
    $(X_h^k, \tau_h^k) \leftarrow \texttt{OM}(u_h^k, v_h^k, I_h, J_h, C_h^k, e_h^k)$
    Implement matching (X_h^k, τ_h^k)
    Observe utilities $u_h^k(i,j), v_h^k(i,j)$ for $(i,j) \in X_h^k$
    Receive next state $s_{h+1}^k = (C_{h+1}^k, I_{h+1}, J_{h+1})$
    Update utility dataset $\mathcal{D}_h^k = \mathcal{D}_h^{k-1} \cup \{C_h^k, e_h^k\} \cup \{u_h^k(i,j), v_h^k(i,j)\}_{(i,j) \in X_h^k}$
  end for
end for

Subset Instability has several key properties for learning. Importantly, it can be shown that given (I, J, u, v), the utility difference between the optimal matching-transfer pair and (X, τ) is upper bounded by $\text{SI}(X, \tau; I, J, u, v)$.

With a slight abuse of notation, for $s_h = (C_h, I_h, J_h)$ and $a_h = (e_h, X_h, \tau_h)$, we denote by $\text{SI}(s_h, a_h, u_h, v_h)$ the SI of (X_h, τ_h) given I_h, J_h and $u_h(C_h, e_h, \cdot, \cdot)$ and $u_h(C_h, e_h, \cdot, \cdot)$. We define the regret of the agents as

$$R^M(K) := \sum_{k=1}^K \mathbb{E}_{\pi_k}\Big[\sum_{h=1}^H \text{SI}(s_h, a_h, u_h, v_h)\Big]. \tag{9}$$

Moreover, SI itself can be bounded by the sum of optimistic bonuses. Therefore, quite surprisingly, we can control the planner's regret and the agents' regret at the same time by bounding the bonus sums. In this way, the total regret can be controlled due to the following proposition.

Proposition 4.2 (Proof in Appendix D.1)**.** *For $R(K)$, $R^P(K)$ and $R^M(K)$ defined by* (3)*,* (8)*,* (9)*, it holds that $R(K) \leq R^P(K) + R^M(K)$.*

4.3 A Meta Algorithm

Now, we are ready to present our meta algorithm as displayed in Algorithm 1. As is clear from the previous derivations, it suffices to construct optimistic estimates of the utilities, which then induces 1) matchings of agents and 2) optimistic estimates of the value functions. The latter then enables the optimistic planning for the planner.

In particular, for the estimation part, Algorithm 1 first generates the Q-function estimates in a backward fashion. Then, using the estimates, Algorithm 1 computes optimistic estimates of utilities by calling the subroutine UE, which then leads to estimates of the pseudo-reward using the subroutine RE. Next, optimistic estimates of the Q functions are obtained via the subroutine QE. Next, for the planning part, Algorithm 1 chooses action in Υ greedily and the matching-transfer pair by calling OM. Finally, OM finds a matching-transfer pair which is stable w.r.t. the estimated utility functions.

As discussed in Section 4.1, the optimistic estimates of the rewards come from solving the linear program in (4) and its dual program in (5), which produce the optimal matching given the set of participating agents and utilities. Therefore, the RE oracle is defined as Algorithm 2, and the OM oracle is defined as Algorithm 3.

The remaining subroutines (UE and QE) are flexible and can be carefully calibrated for different model assumptions. The modular nature of Algorithm 1 facilitates incorporation of existing RL algorithms.

In particular, we study a special case of linear function approximation in the next section, where we provide explicit oracles for these subroutines, and show that the corresponding algorithm enjoys a sublinear regret.

5 Case Study: Markov Matching Markets with Linear Features

In this section, we illustrate the power of our framework under linear function approximation, which is the simplest case of function approximation, yet still a rich enough model.

5.1 Model Assumptions

Utility Model. We assume there are known feature mappings $\boldsymbol{\psi} : \mathcal{C} \times \Upsilon \to \mathbb{R}^d$ and $\boldsymbol{\phi} : \mathcal{I} \times \mathcal{J} \to \mathbb{R}^d$, such that for any $h \in [H]$ and (C_h, e_h, i, j), the utility functions are

$$u_h(C_h, e_h, i, j) = \langle \text{vec}(\boldsymbol{\psi}(C_h, e_h)\boldsymbol{\phi}(i, j)^\top), \boldsymbol{\theta}_h \rangle,$$
$$v_h(C_h, e_h, i, j) = \langle \text{vec}(\boldsymbol{\psi}(C_h, e_h)\boldsymbol{\phi}(i, j)^\top), \boldsymbol{\gamma}_h \rangle.$$

Here $\{\boldsymbol{\theta}_h, \boldsymbol{\gamma}_h\}_{h=1}^H$ are unknown parameters in $\mathbb{R}^{d^2}$. We further define the vectorized feature vector:

$$\boldsymbol{\Phi}(C_h, e_h, i, j) \coloneqq \text{vec}(\boldsymbol{\psi}(C_h, e_h)\boldsymbol{\phi}(i, j)^\top) \in \mathbb{R}^{d^2}.$$

Then the immediate reward defined in (1) can be written as

$$r_h(s_h, a_h) = \Big\langle \sum\nolimits_{(i,j) \in X_h} \boldsymbol{\Phi}(C_h, e_h, i, j), \boldsymbol{\theta}_h + \boldsymbol{\gamma}_h \Big\rangle.$$

Transition Model. Conditioning on the agents' sets $\{I_h, J_h\}_{h=1}^H$, the state transition reduces to that of the contexts. We assume a linear transition model (Jin et al., 2020):

$$\mathbb{P}_h(C_{h+1} | C_h, e_h) \coloneqq \langle \boldsymbol{\psi}(C_h, e_h), \boldsymbol{\mu}_h(C_{h+1}) \rangle, \tag{10}$$

for all $h \in [H]$, where $\boldsymbol{\mu}_h : \mathcal{C} \to \mathbb{R}^d$ is some unknown measure.

Next, we introduce some standard assumptions for matching and linear MDPs (Jin et al., 2020; Jagadeesan et al., 2021).

Algorithm 2 Reward Estimation (`RE`)

Input: u, v, I, J
Output: $\widehat{r}(C, I, J, e)$ as the solution to the $\mathcal{LP}(I, J, u, v)$, for any $(C, e) \in \mathcal{C} \times \Upsilon$

Algorithm 3 Optimal Matching (`OM`)

Input: u, v, I, J, C, e
Output: (X, τ) as the solution of the primal-dual program defined by
(4): $\mathcal{LP}(I, J, u(C, e, \cdot, \cdot), v(C, e, I, J))$, and
(5): $\mathcal{DP}(I, J, u(C, e, \cdot, \cdot), v(C, e, I, J))$.

Assumption 5.1. We assume WLOG that for any $h \in [H]$, $\|u_h(\cdot)\|, \|v_h(\cdot)\| \leq 1$. Assume that for any $(I_h, J_h) \in 2^{\mathcal{I}} \times 2^{\mathcal{J}}$, there exists $W_h > 0$ such that for any context $C \in \mathcal{C}$ and action $e \in \Upsilon$, the max-weight (possibly unstable) matching on (I_h, J_h) with utility functions $u_h(C, e, \cdot, \cdot)$ and $v_h(C, e, \cdot, \cdot)$ has total utility upper bounded by W_h.

The quantities $\{W_h\}_{h=1}^H$ can be viewed as a measure of complexity of the matching problem, and indeed $\sum_{h=1}^H W_h$ determines the magnitude of $V_1^{\pi^\star}$. Assumption 5.1 implies a trivial upper bound that $W_h \leq \min\{|I_h|, |J_h|\}$, but when the max-weight matching involves only a subset of the agents, W_h can be much smaller. Therefore, we regard $\{W_h\}_{h=1}^H$ as instance-dependent parameters.

Assumption 5.2. We assume WLOG that $\|\boldsymbol{\psi}(C, e)\|_2 \leq 1$ and $\|\boldsymbol{\phi}(i, j)\|_2 \leq 1$, implying $\|\boldsymbol{\Phi}(C, e, i, j)\|_2 \leq 1$, for any (C, e, i, j). We assume that for any $h \in [H]$, $\|\boldsymbol{\theta}_h\|_2 \leq d$, $\|\boldsymbol{\gamma}_h\|_2 \leq d$ and$\|\boldsymbol{\mu}_h(\cdot)\|_2 \leq \sqrt{d}$. Moreover, assume that $\max_{h \in [H]} \| \int_{\mathcal{C}} f(C) \mathrm{d}\mu_h(C)\|_2 \leq \sqrt{d}$ for any function $f : \mathcal{C} \to \mathbb{R}$ such that $\sup |f| \leq 1$.

Assumption 5.3. We assume that the observed utilities of matched pairs are the true utilities plus independent 1-subgaussian noise.

5.2 Algorithms

Based on previous model assumptions, we now present explicit computation oracles for Algorithm 1.

Utility Estimation. At the beginning of episode k, for any $h \in [H]$, denote the available data by $\mathcal{D}_h^{k-1}$ which consists of $\{u_h^t(i,j)\}_{t\in[k-1]}^{(i,j)\in X_h^t}$ and $\{v_h^t(i,j)\}_{t\in[k-1]}^{(i,j)\in X_h^t}$, where by default $\mathcal{D}_h^0 = \phi$. For the linear case, each $u_h^t(i,j) = \langle \Phi(C_h^t, e_h^t, i, j), \boldsymbol{\theta}_h \rangle + noise$, and similar for $v_h^t(i,j)$. So we can estimate $\boldsymbol{\theta}_h$ and $\boldsymbol{\gamma}_h$ by ridge regression:

$$\begin{aligned}
\boldsymbol{\theta}_h^k &= (\boldsymbol{\Sigma}_h^k)^{-1} \sum\nolimits_{t=1}^{k-1} \sum\nolimits_{(i,j)\in X_h^t} \Phi(C_h^t, e_h^t, i, j) u_h^t(i,j), \\
\boldsymbol{\gamma}_h^k &= (\boldsymbol{\Sigma}_h^k)^{-1} \sum\nolimits_{t=1}^{k-1} \sum\nolimits_{(i,j)\in X_h^t} \Phi(C_h^t, e_h^t, i, j) v_h^t(i,j), \\
\boldsymbol{\Sigma}_h^k &= \lambda \mathbf{I}_{d^2} + \sum\nolimits_{t=1}^{k-1} \sum\nolimits_{(i,j)\in X_h^t} \Phi(C_h^t, e_h^t, i, j) \Phi(C_h^t, e_h^t, i, j)^\top.
\end{aligned} \tag{11}$$

We then add a bonus to ensure optimism in the utility function estimates and apply the truncation:

$$\begin{aligned}
u_h^k(C, e, i, j) &= \left(\langle \Phi(C, e, i, j), \boldsymbol{\theta}_h^k \rangle + \beta_u \|\Phi(C, e, i, j)\|_{(\boldsymbol{\Sigma}_h^k)^{-1}} \right)_{[-1,1]}, \\
v_h^k(C, e, i, j) &= \left(\langle \Phi(C, e, i, j), \boldsymbol{\gamma}_h^k \rangle + \beta_u \|\Phi(C, e, i, j)\|_{(\boldsymbol{\Sigma}_h^k)^{-1}} \right)_{[-1,1]}.
\end{aligned} \tag{12}$$

The UE for the linear case is summarized in Algorithm 4.

Q-function Estimation. By the assumption on the context transition, for any function f, the function $\mathbb{P}_h f$ is linear in features $\boldsymbol{\psi}$, which induces the commonly used LSVI-type approach (Jin et al., 2020). Together with the reward estimates, we can estimate the Q-function using Bellman equation via backward ridge regression. For each (k, h), denote the estimate of the Q-function by $\overline{Q}_h^k$ and the value function by $\overline{V}_h^k$. Given $\overline{Q}_{h+1}^k$, maximizing over $e \in \Upsilon$ yields $\overline{V}_{h+1}^k$, then we solve the following ridge regression:

$$\mathbf{w}_h^k = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \sum\nolimits_{t=1}^{k-1} \left[\overline{V}_{h+1}^k(C_{h+1}^t) - \boldsymbol{\psi}(C_h^t, e_h^t)^\top \mathbf{w} \right]^2 + \lambda \|\mathbf{w}\|_2^2,$$

which further yields the estimated expectation of $\overline{V}_{h+1}^k$:

$$\widehat{\mathbb{P}}_h \overline{V}_{h+1}^k(\cdot,\cdot) = \boldsymbol{\psi}(\cdot,\cdot)^\top \mathbf{w}_h^k + \beta_V \|\boldsymbol{\psi}(\cdot,\cdot)\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}, \tag{13}$$

where $\boldsymbol{\Lambda}_h^k = \sum_{t=1}^{k-1} \boldsymbol{\psi}(C_h^t, e_h^t) \boldsymbol{\psi}(C_h^t, e_h^t)^\top + \lambda \mathbf{I}_d$. Then we estimate Q_h^k by the Bellman equation, as in Algorithm 5.

Algorithm 4 Utility Estimation (UE)

Input: $\mathcal{D}_h^{k-1}$, β_u, λ
if $\mathcal{D}_h^{k-1}$ is empty **then**
  Output: $u_h^k \equiv 1$ and $v_h^k \equiv 1, \forall h \in [H]$
end if
Compute $\boldsymbol{\theta}_h^k, \boldsymbol{\gamma}_h^k$ by (11)
Estimate utility functions with u_h^k and v_h^k by (12)
Output: the functions u_h^k and v_h^k

Algorithm 5 Q-function Estimation (QE)

Input: $\mathcal{D}_h^{k-1}$, $\overline{r}_h^k$, $\overline{Q}_{h+1}^k$, β_V, λ
$\overline{V}_{h+1}^k(C, I_h, J_h) = \max_e \overline{Q}_{h+1}^k(C, e, I_h, J_h)$
Compute $\widehat{\mathbb{P}}_h \overline{V}_{h+1}^k$ by (13)
$\overline{Q}_h^k(C, I_h, J_h, e) = (\overline{r}_h^k(C, I_h, J_h, e) + \widehat{\mathbb{P}}_h \overline{V}_{h+1}^k(C, e))_{[0, \sum_{l=h}^H W_l]}$
Output: function $\overline{Q}_h^k$

5.3 Theoretical Results

In this section, we present our theoretical results, with the proofs deferred to Appendix D. We start with the main theorems on the agents' regret and the planner's regret.

Theorem 5.4 (Agents' Regret)**.** *With probability at least $1 - 2\delta$, the agents' regret can be bounded as $R^M(K) \le \mathcal{O}(d^2(\sum_{h=1}^H \min\{|I_h|, |J_h|\})\kappa\sqrt{K})$ where $\kappa = \log(dK \min(|\mathcal{I}|, |\mathcal{J}|)/\delta)$.*

Comparing this result with Theorem 5.3 in Jagadeesan et al. (2021), which proves the regret of their MatchLinUCB algorithm designed for the linear utility class under $H = 1$, and ignoring the logarithmic term, we see that both have a linear dependence on the cardinality of the agents' set, while we have an extra summation over the H horizon due to the sequential setting. It might seem that their d-dependence is $\mathcal{O}(d)$ while ours is $\mathcal{O}(d^2)$. But this is because their feature is in $\mathbb{R}^d$ while our feature $\mathbf{\Phi} \in \mathbb{R}^{d^2}$. Therefore, the dominant term in Theorem 5.4 matches that of Jagadeesan et al. (2021), and our result can be viewed as an extension.

Theorem 5.5 (Planner's Regret)**.** *Under Assumption 5.1, 5.2, and $KH > 32$, there exists a problem-independent constant $\eta > 0$, such that for any $\delta > 0$, with the choice of parameters $\lambda = 1$, $\beta_V = \eta d^2\big(\sum_{h=1}^H W_h\big) \cdot \sqrt{\iota}$ where $\iota = \log(dKH \min\{|\mathcal{I}|, |\mathcal{J}|\}/\delta)$, and β_u as given in Lemma C.1, then with probability at least $1 - \delta$, the planner's regret is bounded by $R^P(K) \leq \mathcal{O}(\eta d^{5/2} H(\sum_{h=1}^H W_h)\iota\sqrt{K})$.*

Combining the above two theorems, we have shown that we find the optimal policy for both the planner and agents at sublinear rates, and accordingly for the total regret. Notably, the regret upper bound only depends on the size of the market through $\{W_h\}_{h=1}^H$, which is instance-dependent. Due to the sublinear regret, our algorithm can be further adapted to a PAC algorithm (cf. Jin et al., 2018). We also remark that our result with LSVI-type estimation can be naturally extended to Eluder dimension (Wang et al., 2020; Ayoub et al., 2020).

5.4 Analysis

The key step in our analysis is to show that the estimated pseudo-reward function $\overline{r}_h^k$ satisfies optimism, i.e., $\overline{r}_h^k \geq \overline{r}_h$, and we need to ensure that $\overline{r}_h^k$ is not too far away from $\overline{r}_h$ (Lemma C.2). Due to space limit, we refer interested readers to Appendix C for a proof sketch of the main theory and an introduction of all the key technical lemmas. The full proof is presented in Appendix D.

6 Conclusion

We propose a novel Markov matching market model and a general framework that incorporates max-weight matching and RL algorithms for efficient online learning. We prove that our proposed algorithms achieve sublinear regret under proper structural assumptions on the structural model.

Acknowledgments and Disclosure of Funding

Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports.

References

ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* **24** 2312–2320.

AKBARPOUR, M., LI, S. and GHARAN, S. O. (2014). Dynamic matching market design. *arXiv preprint arXiv:1402.3643* .

AKBARPOUR, M., LI, S. and GHARAN, S. O. (2020). Thickness and information in dynamic matching markets. *Journal of Political Economy* **128** 783–815.

ANDERSON, R., ASHLAGI, I., GAMARNIK, D. and KANORIA, Y. (2014). A dynamic model of barter exchange. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*. SIAM.

AYOUB, A., JIA, Z., SZEPESVARI, C., WANG, M. and YANG, L. (2020). Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*. PMLR.

AZAR, M. G., OSBAND, I. and MUNOS, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*. PMLR.

AZUMA, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series* **19** 357–367.

BACCARA, M., LEE, S. and YARIV, L. (2020). Optimal dynamic matching. *Theoretical Economics* **15** 1221–1278.

BAI, Y., XIE, T., JIANG, N. and WANG, Y.-X. (2019). Provably efficient q-learning with low switching cost. *arXiv preprint arXiv:1905.12849* .

BASU, S., SANKARARAMAN, K. A. and SANKARARAMAN, A. (2021). Beyond $\log^2(t)$ regret for decentralized bandits in matching markets. *arXiv preprint arXiv:2103.07501* .

CAI, Q., YANG, Z., JIN, C. and WANG, Z. (2020). Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*. PMLR.

CEN, S. H. and SHAH, D. (2021). Regret, stability, and fairness in matching markets with bandit learners. *arXiv preprint arXiv:2102.06246* .

CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction, learning, and games*. Cambridge university press.

DAMIANO, E. and LAM, R. (2005). Stability in dynamic matching markets. *Games and Economic Behavior* **52** 34–53.

DANN, C., LATTIMORE, T. and BRUNSKILL, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710* .

DAS, S. and KAMENICA, E. (2005). Two-sided bandits and the dating market. In *IJCAI*, vol. 5. Citeseer.

DOVAL, L. (2014). A theory of stability in dynamic matching markets. Tech. rep., Technical report, mimeo.

DOVAL, L. (2019). Dynamically stable matching. *Available at SSRN 3411717* .

DOVAL, L. and SZENTES, B. (2019). On the efficiency of queueing in dynamic matching markets. Tech. rep., Working paper, Columbia University.

DU, S. S., KAKADE, S. M., LEE, J. D., LOVETT, S., MAHAJAN, G., SUN, W. and WANG, R. (2021). Bilinear classes: A structural framework for provable generalization in RL. *arXiv preprint arXiv:2103.10897* .

FEI, Y., YANG, Z., CHEN, Y. and WANG, Z. (2021a). Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. In *Advances in Neural Information Processing Systems*.

FEI, Y., YANG, Z., CHEN, Y., WANG, Z. and XIE, Q. (2020). Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. In *Advances in Neural Information Processing Systems*.

FEI, Y., YANG, Z. and WANG, Z. (2021b). Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *International Conference on Machine Learning*. PMLR.

GALE, D. and SHAPLEY, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly* **69** 9–15.

GAO, M., XIE, T., DU, S. S. and YANG, L. F. (2021). A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494* .

GUO, W., KANDASAMY, K., GONZALEZ, J. E., JORDAN, M. I. and STOICA, I. (2021). Online learning of competitive equilibria in exchange economies. *arXiv preprint arXiv:2106.06616* .

GURVICH, I. and WARD, A. (2015). On the dynamic control of matching queues. *Stochastic Systems* **4** 479–523.

HU, M. and ZHOU, Y. (2021). Dynamic type matching. *Manufacturing & Service Operations Management* .

JAGADEESAN, M., WEI, A., WANG, Y., JORDAN, M. I. and STEINHARDT, J. (2021). Learning equilibria in matching markets from bandit feedback. *arXiv preprint arXiv:2108.08843* .

JAKSCH, T., ORTNER, R. and AUER, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* **11**.

JIN, C., ALLEN-ZHU, Z., BUBECK, S. and JORDAN, M. I. (2018). Is q-learning provably efficient? *arXiv preprint arXiv:1807.03765* .

JIN, C., LIU, Q. and MIRYOOSEFI, S. (2021). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815* .

JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*. PMLR.

KADAM, S. V. and KOTOWSKI, M. H. (2018). Multiperiod matching. *International Economic Review* **59** 1927–1947.

KANDASAMY, K., GONZALEZ, J. E., JORDAN, M. I. and STOICA, I. (2020). Mechanism design with bandit feedback. *arXiv preprint arXiv:2004.08924* .

KOTOWSKI, M. H. (2019). A perfectly robust approach to multiperiod matching problems .

KURINO, M. (2020). Credibility, efficiency, and stability: A theory of dynamic matching markets. *The Japanese Economic Review* **71** 135–165.

LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit Algorithms*. Cambridge University Press.

LAUERMANN, S. and NÖLDEKE, G. (2014). Stable marriages and search frictions. *Journal of Economic Theory* **151** 163–195.

LESHNO, J. (2019). Dynamic matching in overloaded waiting lists. *Available at SSRN 2967011* .

LIU, C. (2020). Stability in repeated matching markets. *arXiv preprint arXiv:2007.03794* .

LIU, L. T., MANIA, H. and JORDAN, M. (2020). Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

LIU, L. T., RUAN, F., MANIA, H. and JORDAN, M. I. (2021). Bandit learning in decentralized matching markets. *Journal of Machine Learning Research* **22** 1–34.

LOERTSCHER, S., MUIR, E. V. and TAYLOR, P. G. (2018). Optimal market thickness and clearing. *Unpublished paper, Department of Economics, University of Melbourne.[1224]* .

MAS-COLELL, A., WHINSTON, M. D., GREEN, J. R. ET AL. (1995). *Microeconomic theory*, vol. 1. Oxford university press New York.

MIN, Y., HE, J., WANG, T. and GU, Q. (2021a). Learning stochastic shortest path with linear function approximation. *arXiv preprint arXiv:2110.12727* .

MIN, Y., WANG, T., ZHOU, D. and GU, Q. (2021b). Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems* **34**.

NIEDERLE, M. and YARIV, L. (2009). Decentralized matching with aligned preferences. Tech. rep., National Bureau of Economic Research.

OSBAND, I., VAN ROY, B. and WEN, Z. (2016). Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*. PMLR.

ÖZKAN, E. and WARD, A. R. (2020). Dynamic matching for real-time ride sharing. *Stochastic Systems* **10** 29–70.

QIN, Z., TANG, X., JIAO, Y., ZHANG, F., XU, Z., ZHU, H. and YE, J. (2020). Ride-hailing order dispatching at didi via reinforcement learning. *INFORMS Journal on Applied Analytics* **50** 272–286.

RASOULI, M. and JORDAN, M. I. (2021). Data sharing markets. *arXiv preprint arXiv:2107.08630* .

SANKARARAMAN, A., BASU, S. and SANKARARAMAN, K. A. (2021). Dominate or delete: Decentralized competing bandits in serial dictatorship. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

SATTERTHWAITE, M. and SHNEYEROV, A. (2007). Dynamic matching, two-sided incomplete information, and participation costs: Existence and convergence to perfect competition. *Econometrica* **75** 155–200.

SHAPLEY, L. S. and SHUBIK, M. (1971). The assignment game i: The core. *International Journal of game theory* **1** 111–130.

STREHL, A. L., LI, L., WIEWIORA, E., LANGFORD, J. and LITTMAN, M. L. (2006). Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*.

TAYLOR, C. R. (1995). The long side of the market and the short end of the stick: Bargaining power and price formation in buyers', sellers', and balanced markets. *The Quarterly Journal of Economics* **110** 837–855.

ÜNVER, M. U. (2010). Dynamic kidney exchange. *The Review of Economic Studies* **77** 372–414.

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

WANG, R., SALAKHUTDINOV, R. and YANG, L. F. (2020). Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *arXiv preprint arXiv:2005.10804* .

WANG, T., ZHOU, D. and GU, Q. (2021). Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *arXiv preprint arXiv:2101.02195* .

YANG, L. and WANG, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*. PMLR.

YANG, Z., JIN, C., WANG, Z., WANG, M. and JORDAN, M. I. (2020). On function approximation in reinforcement learning: Optimism in the face of large state spaces. *arXiv preprint arXiv:2011.04622* .

ZANETTE, A., LAZARIC, A., KOCHENDERFER, M. and BRUNSKILL, E. (2020). Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*. PMLR.

ZENIOS, S. A. (1999). Modeling the transplant waiting list: A queueing model with reneging. *Queueing systems* **31** 239–251.

ZHOU, D., GU, Q. and SZEPESVARI, C. (2021). Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*. PMLR.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This work is seeking to develop a mathematical understanding of dynamic matching.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use existing theorems and cite them appropriately.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Clarification of Notation

In this section, we provide a comprehensive clarification on the use of notation in this paper.

Throughout the paper, we use $\mathcal{O}(\cdot)$ to hide problem-independent constants and use $(\cdot)_{[a,b]}$ to denote the truncation into the range $[a, b]$. $\mathcal{I}$ and $\mathcal{J}$ denote the sets of all side-1 and side-2 agents respectively. Further, we use $I \in 2^{\mathcal{I}}$ and $J \in 2^{\mathcal{J}}$ to denote any set of participating agents, which are the subsets of $\mathcal{I}$ and $\mathcal{J}$.

For any given stage $h \in [H]$, we use s_h and (C_h, I_h, J_h) interchangeably when describing any state $s_h \in \mathcal{S}$ where $\mathcal{S} = \mathcal{C} \times 2^{\mathcal{I}} \times 2^{\mathcal{J}}$. Analogously, we also use a_h and (e_h, X_h, τ_h) interchangeably for the action $a_h \in \mathcal{A}$ where $\mathcal{A} = \Upsilon \times \mathcal{X} \times \mathcal{T}$

We use $\pi = \{\pi_h\}_{h=1}^H$ to denote a policy, where each π_h is defined to be a mapping from $\mathcal{S}$ to a distribution $\Delta_{\mathcal{A}}$ on $\mathcal{A}$. Therefore, for any $h \in [H]$ and $s_h \in \mathcal{S}$, $\pi_h(\cdot|s_h)$ denotes a probability distribution on $\mathcal{A}$. Note that because $\mathcal{A} = \Upsilon \times \mathcal{X} \times \mathcal{T}$, the policy π is a joint policy. We may slightly abuse the notation in the paper and refer to π as the policy restricted on Υ only, whenever it is clear from the context. In such case, we refer to $\pi_h(\cdot|s_h)$ as a distribution on Υ only.

We also present the following table of notations. The π in the superscript can be replaced by π_k or $\pi_\star$, where the former refers to the policy in episode k, and the latter refers to the optimal policy.

Table 1: Notation

Notation	Meaning
$\mathcal{C}, \mathcal{I}, \mathcal{J}$	set of contexts, side-1 agents, side-2 agents
$\Upsilon, \mathcal{X}, \mathcal{T}$	set of planner's actions, all matchings and transfers over all possible subsets of $\mathcal{I} \times \mathcal{J}$
$r_h, \overline{r}_h$	reward, pseudo-reward functions
$V_h^\pi, Q_h^\pi, V_h^\star$	value, Q functions under π, optimal value functions w.r.t. the transition functions $\{\mathbb{P}_h\}_{h=1}^H$ and reward functions $\{r_h\}_{h=1}^H$
$\overline{V}_h^\pi, \overline{Q}_h^\pi, \overline{V}_h^\star$	pseudo-value, pseudo-Q functions under π, optimal pseudo-value functions w.r.t. the transition functions $\{\mathbb{P}_h\}_{h=1}^H$ and reward functions $\{\overline{r}_h\}_{h=1}^H$
$\overline{V}_h^k, \overline{Q}_h^k$	estimated value, Q functions for stage h in episode k in Algorithm 1
π_k	the policy followed by Algorithm 1 in episode k, where $\pi_k = \{\pi_{k,h}\}_{h=1}^H$
$\pi_{k,h}$	the policy followed by Algorithm 1 at stage h in episode k

B Supplementary Information on Matching and Stability

In this section, we review some basics on the matching problem. We first introduce the classic problem of (static) matching with transfers and the notion of stability. We then recap the primal-dual formulation that provides an efficient way to solve a stable matching (Shapley and Shubik, 1971). Finally, we give more details about Subset Instability and its properties.

B.1 Matching with Transferable Utilities

This section is a supplementary to Section 3.1 in the main text. We introduce the two-sided static matching with transferable utilities.

Denote the sets of participating agents by I and J for two sides respectively. A matching $X \subseteq I \times J$ is a set of pairs of agents, and $(i, j) \in X$ means $i \in I$ is matched to $j \in J$. Each agent is matched at most once. We denote by $X(i) = j$ and $X(j) = i$ for any matched pair $(i, j) \in X$, while for any unmatched agent $a \in I \cup J$, we write $X(a) = a$.

Matched agents receive utilities, denoted by $u : I \times J \to \mathbb{R}$ for agents in I and $v : I \times J \to \mathbb{R}$ for agents in J. Specifically, if $(i, j) \in X$, then agent i receives an utility $u(i, j)$ and agent j receives an

utility $v(i,j)$. Remaining unmatched agents receive zero utility. With a slight abuse of notation, we overwrite $u(i,i)=0$ and $v(j,j)=0$ for any $i \in I$ and $j \in J$.

In addition, there are utility transfers between (and only between) agents. We denote the transfer function by $\tau : I \times J \to \mathbb{R}$ such that for any agent $a \in I \cup J$, $\tau(a)$ is the transfer received by agent a. Since the transfers are within agents, we have

$$\sum_{a \in I \cup J} \tau(a) = 0.$$

We denote the market outcome by (X, τ), under which the net utility received by an agent $i \in I$ is $u(i,j) + \tau(i)$ if $(i,j) \in X$, and similarly for agents in J.

The notion of *stable matching* is as follows.

Definition B.1 (Stable matching)**.** A matching-transfer pair (X, τ) on I, J is stable if:

1. The net utility of of any agent is non-negative, i.e.

$$u(i, X(i)) + \tau(i) \geq 0,$$
$$v(X(j), j) + \tau(j) \geq 0,$$

   for all $i \in I$ and $j \in J$.

2. There are no blocking pairs, i.e.

$$[u(i, X(i)) + \tau(i)] + [v(X(j), j) + \tau(j)] \geq u(i,j) + v(i,j),$$

   for all pairs $(i,j) \in I \times J$.

Stable matching implies that no matched agents would rather be unmatched and no pair of agents can find a transfer between themselves so that both would rather match with each other than follow (X, τ). The following proposition provides a fundamental and important max-weight interpretation for stable matchings.

Proposition B.2 (Shapley and Shubik 1971)**.** *For the matching with transfer problem, if (X, τ) is a stable matching under Definition B.1, then X must be the max-weight matching, i.e.,*

$$X = \arg\max_{X'} \sum_{i \in I, j \in J} u(i, X'(i)) + v(X'(j), j)$$

where the maximum is over all matchings on $I \times J$.

Therefore, by Proposition B.2, to maximize the total social welfare (i.e. sum of utilities), it suffices to find a stable matching. But how? This is answered in the next subsection.

B.2 The Linear Program and Dual Program

In this subsection, we explain how to find a stable matching (X, τ) given input I, J, u, v, which gives rise to the algorithm `OM` (i.e. Algorithm 3) in the main text.

Shapley and Shubik (1971) showed that, assuming the utility functions are known, the stable (X, τ) can be found by solving the following linear program and its dual program (recapped from Section 4.1 in the main text):

$$\begin{aligned} \mathcal{LP}(I, J, u, v): \quad & \max_{w \in \mathbb{R}^{|I| \times |J|}} \sum_{(i,j) \in I \times J} w_{i,j} \left[u(i,j) + v(i,j) \right] \\ & \text{s.t.} \quad \sum_{j \in J_h} w_{i,j} \leq 1, \forall\, i \in I, \\ & \quad \sum_{i \in I_h} w_{i,j} \leq 1, \forall\, j \in J, \\ & \quad w_{i,j} \geq 0, \forall\, (i,j) \in I \times J, \end{aligned} \tag{14}$$

and its dual program:

$$\mathcal{DP}(I, J, u, v): \quad \min_{p: I \cup J \to \mathbb{R}^+} \sum_{a \in I \cup J} p(a) \quad \text{s.t.} \quad p(i) + p(j) \geq u(i,j) + v(i,j), \ \forall (i,j) \in I \times J. \tag{15}$$

Denote the solution pair to the primal-dual problems by (w, p). Shapley and Shubik (1971) proved that (w, p) leads to a max-weight stable matching-transfer pair (X, τ). Specifically, it is proved that the vector w must have integer entries, i.e., $w_{i,j} = 0$ or 1, which naturally induces a matching X such that $(i, j) \in X$ if and only if $w_{i,j} = 1$. Correspondingly, the transfers are $\tau(i) = p(i) - u(i, X(i))$ for $i \in I$, and similarly for $j \in J$.

The above procedure constitutes the subroutine oracle `OM` as displayed in Algorithm 3. It takes as input the sets of participating agents and estimated utility functions, then outputs the stable matching (X, τ) by solving the primal-dual linear program described above. Note that the matching is only stable with respect to the estimated utility functions.

B.3 Details about Subset Instability

Next, we review the notion of Subset Instability and its properties. We refer the interested reader to (Jagadeesan et al., 2021) for the full details.

Given a matching-pair (X, τ), define its utility difference as

$$\Big[\max_{X'} \sum_{i \in I, j \in J} u(i, X'(i)) + v(X'(j), j)\Big] - \Big[\sum_{i \in I, j \in J} u(i, X(i)) + v(X(j), j)\Big]. \tag{16}$$

Recall the definition of Subset Instability:

Definition 4.1 (Subset Instability, Jagadeesan et al. 2021)**.** *Given any agent sets $I,\ J$ and utility functions $u,\ v : I \times J \to \mathbb{R}$, the Subset Instability* $\mathrm{SI}(X, \tau; I, J, u, v)$ *of the matching and transfer (X, τ) is defined as*

$$\max_{I' \times J' \subseteq I \times J} \Bigg[\Big(\max_{X'} \sum_{i \in I'} u(i, X'(i)) + \sum_{j \in J'} v(X'(j), j)\Big) - \Big(\sum_{i \in I'} (u(i, X(j)) + \tau(i))\Big) - \Big(\sum_{j \in J'} (v(X(j), j) + \tau(j))\Big)\Bigg]$$

where $X(\cdot)$ and $X'(\cdot)$ denotes the matched agent in matching X and X' respectively.

By definition, Subset Instability indicates whether there exists any subset $I' \times J'$ of agents who can achieve a higher total utility by taking some alternative matching X' among themselves other than the current matching-transfer pair (X, τ). Thus, Subset Instability is an upper bound of the total utility difference, and quantifies the distance from a proposed matching to the optimal stable matching, as summarized in the following proposition.

Proposition B.3 (Proposition 4.4 in Jagadeesan et al. 2021)**.** *The following holds for Subset Instability*

1. *Subset Instability is always nonnegative and is zero if and only if (X, τ) is stable matching.*
2. *Subset Instability is Lipschitz continuous with respect to the ℓ_∞ norm of the utility functions.*

$$|\mathrm{SI}(X, \tau; I, J, u, v) - \mathrm{SI}(X, \tau; I, J, \widetilde{u}, \widetilde{v})| \leq 2\Big(\sum_{i \in I} \|u(i, \cdot) - \widetilde{u}(i, \cdot)\|_\infty - \sum_{j \in J} \|v(\cdot, j) - \widetilde{v}(\cdot, j)\|_\infty\Big).$$

3. *Subset Instability is always at least the utility difference* (16).

In our problem, this allows us to bound the total regret of the agents by the sum of Subset Instability of matchings across all episodes, as reflected in Proposition 4.2.

C Proof Sketch

C.1 Optimistic Utility and Reward Estimates

The key step in our analysis is to show that the estimated pseudo-reward function $\overline{r}_h^k$ satisfies optimism, i.e., $\overline{r}_h^k \geq \overline{r}_h$, and we need to ensure that $\overline{r}_h^k$ is not too far away from $\overline{r}_h$. The lemma below justifies the optimism of utility estimates.

Lemma C.1 (UCB for Utility Estimates; proof in E.1)**.** *For any $0 < \delta < 1$, set β_u as $\beta_u = \sqrt{d^2 \log[2(1 + d^2 K \max_h \min\{|I_h|, |J_h|\})/(\lambda\delta)]} + \sqrt{\lambda} d$. Then with probability at least $1 - \delta$, u_h^k and v_h^k in Algorithm 1 satisfy $u_h^k \geq u_h$ and $v_h^k \geq v_h$. Furthermore, $|u_h^k(\cdot) - u_h(\cdot)|$ and $|v_h^k(\cdot) - v_h(\cdot)|$ are bounded by $2\beta_u \left\|\mathbf{\Phi}(\cdot)\right\|_{(\mathbf{\Sigma}_h^k)^{-1}}$.*

Next, we explain why the optimism of utility estimates implies that of reward estimates. By Lemma C.1, we can write $u_h^k = u_h + b_{u,h}$ and $v_h^k = v_h + b_{v,h}$ where $b_{u,h}$ and $b_{v,h}$ are bonus functions satisfying

$$b_{u,h}(C, e, i, j),\ b_{v,h}(C, e, i, j) \in \left[0,\ 2\beta_u \left\|\mathbf{\Phi}(C, e, i, j)\right\|_{(\mathbf{\Sigma}_h^k)^{-1}}\right].$$

Lemma C.2 (Planner's Optimism; proof in E.2)**.** *Under the event of Lemma C.1, it holds that for any $(C, e) \in \mathcal{C} \times \Upsilon$,*

$$0 \leq \overline{r}_h^k(C, e, I_h, J_h) - \overline{r}_h(C, e, I_h, J_h) \leq \sum\nolimits_{(i,j) \in X_h^k} (b_{u,h}(C, e, i, j) + b_{v,h}(C, e, i, j)).$$

In the sequel, we denote by π_k the policy whose market making part is greedy w.r.t. $\overline{Q}_h^k$ and whose matching part chooses the max-weight stable matching given u_h^k and v_h^k.

C.2 Proof Sketch of Theorem 5.4

By definition (9), the agents' regret can be interpreted as the expected sum of total SI across all time steps, where the expectation is over the trajectory induced by π_k for $k \in [K]$.

To bound the regret, we relate the expected SI in (9) with the realized SI via a martingale difference sequence. Specifically, writing $\mathrm{SI}_h = \mathrm{SI}(s_h, a_h, u_h, v_h)$ and $\mathrm{SI}_h^k = \mathrm{SI}(s_h^k, a_h^k, u_h, v_h)$, we define the sum of differences as

$$\sum\nolimits_{k=1}^{K} \left\{ \mathbb{E}_{\pi_k} \left[\sum\nolimits_{h=1}^{H} \mathrm{SI}_h \right] - \sum\nolimits_{h=1}^{H} \mathrm{SI}_h^k \right\}. \tag{17}$$

We bound the difference (17) and the sum of realized SI $\sum_{k=1}^{K} \sum_{h=1}^{H} \mathrm{SI}_h^k$ separately, where the former is a sum of martingale difference sequences that concentrates and the latter can be bounded using the following lemma.

Lemma C.3 (Lemma 5.4 in Jagadeesan et al. 2021; proof in E.3)**.** *Under the event of Lemma C.1, we have*

$$\mathrm{SI}_h^k \leq \sum\nolimits_{(i,j) \in X_h^k} (b_{u,h}(C_h^k, e_h^k, i, j) + b_{v,h}(C_h^k, e_h^k, i, j)).$$

Remark C.4. Note that each implemented matching induces several utility observations at a time, so bounding the bonus sum for utilities has a resemblance to lazy policy updates in the online learning literature (Abbasi-Yadkori et al., 2011). It particularly is similar to techniques used in the low switching cost problem in RL (Bai et al., 2019; Wang et al., 2021; Gao et al., 2021). This will be clear in the proof of Theorem 5.4 in Appendix D.3.

C.3 Proof Sketch of Theorem 5.5

Define the following functions δ_h^k and terms $\zeta_{k,h}^1, \zeta_{k,h}^2$:

$$\begin{aligned}
\delta_h^k(C, e, I, J) &\coloneqq [\overline{r}_h + \mathbb{P}_h \overline{V}_{h+1}^k - \overline{Q}_h^k](C, e, I, J)\,, \\
\zeta_{k,h}^1 &\coloneqq (\overline{V}_h^k - \overline{V}_h^{\pi_k})(C_h^k, I_h, J_h) - (\overline{Q}_h^k - \overline{Q}_h^{\pi_k})(C_h^k, e_h^k, I_h, J_h)\,, \\
\zeta_{k,h}^2 &\coloneqq \mathbb{P}_h(\overline{V}_{h+1}^k - \overline{V}_{h+1}^{\pi_k})(C_h^k, e_h^k, I_h, J_h) - (\overline{V}_{h+1}^k - \overline{V}_{h+1}^{\pi_k})(C_{h+1}^k, I_{h+1}, J_{h+1}).
\end{aligned} \tag{18}$$

To simplify the notation, in the following, we omit I_h, J_h from the arguments of the functions since we are conditioning on $\{I_h, J_h\}_{h=1}^H$ being fixed.

Lemma C.5 (Regret Decomposition of Planner; proof in E.4)**.** *The planner's regret defined by* (8) *satisfies*

$$R^P(K) = \underbrace{\sum_{k=1}^K \sum_{h=1}^H \left(\zeta_{k,h}^1 + \zeta_{k,h}^2\right)}_{E_1} + \underbrace{\sum_{k=1}^K \sum_{h=1}^H \left[\mathbb{E}_{\pi^\star}\left[\delta_h^k(C_h, e_h)|C_1^k\right] - \delta_h^k(C_h^k, e_h^k)\right]}_{E_2} \tag{19}$$
$$+ \underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^\star}\left[\langle \overline{Q}_h^k(C_h, \cdot), \pi_h^\star(\cdot|C_h) - \pi_{k,h}(\cdot|C_h)\rangle_\Upsilon \middle| C_1^k\right]}_{E_3},$$

where the expectation $\mathbb{E}_{\pi^\star}[\cdot|C_1^k]$ is with respect to the trajectory $\{C_h, e_h\}_{h=1}^H$ induced by the policy $\pi^\star$ conditioning on $C_1 = C_1^k$ and $\langle \cdot, \cdot \rangle_\Upsilon$ means sum over all $e \in \Upsilon$.

In decomposition (19), term E_1 is controlled using a standard martingale concentration. Next, to bound E_2, we show that $\delta_h^k \leq 0$ with high probability, which implies that $E_1 \leq \sum_{k=1}^K \sum_{h=1}^H \left|\delta_h^k(C_h^k, e_h^k)\right|$. Bounding each $|\delta_h^k(C_h^k, e_h^k)|$ by the corresponding optimistic bonus $\|\boldsymbol{\psi}(C_h^k, e_h^k)\|_{(\boldsymbol{\Lambda}_h^{k+1})^{-1}}$, we then apply Elliptical Potential lemma to get

$$E_1 \leq 2\beta_V \sum\nolimits_{h=1}^H \sum\nolimits_{k=1}^K \sqrt{2}\left\|\boldsymbol{\psi}(C_h^k, e_h^k)\right\|_{(\boldsymbol{\Lambda}_h^{k+1})^{-1}} \leq 2\sqrt{2}\beta_V H \sqrt{Kd \log\left((K+d)/d\right)}.$$

Finally, for E_3, note that by Algorithm 1, the market-making part of the policy $\pi_{k,h}$ is the greedy policy with respect to $\overline{Q}_h^k$. Based on this observation, it follows that $\sum_{e\in\Upsilon} \overline{Q}_h^k(C_h, e)(\pi_h^\star(e|C_h) - \pi_{k,h}(e|C_h)) = \sum_{e\in\Upsilon} \overline{Q}_h^k(C_h, e)\pi_h^\star(e|C_h) - \max_{e\in\Upsilon} \overline{Q}_h^k(C_h, e) \leq 0$, so $E_3 \leq 0$. Combining yields the bound in Theorem 5.5. Full details are presented in Appendix C.4.

C.4 Proof of Theorem 5.5

To prove Theorem 5.5, we need the following two lemmas which is helpful for bounding E_1 and E_2.

Lemma C.6 (Proof in Section E.5)**.** *Under the setting of Theorem 5.5, with probability at least $1-\delta$, for all $(h,k) \in [H] \times [K]$ and $(C, e) \in \mathcal{C} \times \Upsilon$, it holds that*

$$-2\beta_V \cdot \|\boldsymbol{\psi}(C, e)\|_{(\boldsymbol{\Lambda}_h^k)^{-1}} \leq \delta_h^k(C, e) \leq 0.$$

Lemma C.7 (Proof in Section E.6)**.** *For any $\delta > 0$, with probability at least $1-\delta$, it holds that*

$$\sum\nolimits_{k=1}^K \sum\nolimits_{h=1}^H \left(\zeta_{k,h}^1 + \zeta_{k,h}^2\right) \leq 3\Big(\sum\nolimits_{h=1}^H W_h\Big)\sqrt{K \log 2/\delta}.$$

We are now ready to prove Theorem 5.5.

Proof of Theorem 5.5. By Lemma C.5, we bound the three terms separately.

Bound on E_3. According to Algorithm 1, the planner's policy $\pi_{k,h}$ is the greedy policy with respect to $\overline{Q}_h^k$. It follows that

$$\begin{aligned}&\langle \overline{Q}_h^k(C_h, I_h, J_h \cdot), \pi_h^\star(\cdot|C_h, I_h, J_h) - \pi_{k,h}(\cdot|C_h, I_h, J_h)\rangle_\Upsilon \\ &\quad = \langle \overline{Q}_h^k(C_h, I_h, J_h, \cdot), \pi_h^\star(\cdot|C_h, I_h, J_h)\rangle_\Upsilon - \max_{e\in\Upsilon} \overline{Q}_h^k(C_h, I_h, J_h, e) \leq 0.\end{aligned}$$

Therefore, we have $E_3 \leq 0$.

Bound on E_1. We apply Lemma C.6. The E_1 term can be bounded as

$$E_1 \leq \sum_{k=1}^{K}\sum_{h=1}^{H} -\delta_h^k(C_h^k, e_h^k) \leq 2\beta_V \cdot \sum_{k=1}^{K}\sum_{h=1}^{H} \left\|\boldsymbol{\psi}(C_h^k, e_h^k)\right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}.$$

Note that $\boldsymbol{\Lambda}_h^k = \lambda\mathbf{I} + \sum_{t=1}^{k-1}\boldsymbol{\psi}(C_h^t, e_h^t)\boldsymbol{\psi}(C_h^t, e_h^t)^\top$ by definition, where $\|\boldsymbol{\psi}\|_2 \leq 1$ by Assumption 5.2 and $\lambda = 1$ by our choice. Thus we have $\boldsymbol{\Lambda}_h^{k+1} = \boldsymbol{\Lambda}_h^k + \boldsymbol{\psi}(C_h^t, e_h^t)\boldsymbol{\psi}(C_h^t, e_h^t)^\top \preccurlyeq 2\boldsymbol{\Lambda}_h^k$, or equivalently, $(\boldsymbol{\Lambda}_h^k)^{-1} \preccurlyeq 2(\boldsymbol{\Lambda}_h^{k+1})^{-1}$, for all k. Applying this to the above inequality, we get the final bound for E_1:

$$E_1 \leq 2\beta_V \cdot \sum_{h=1}^{H}\sum_{k=1}^{K} \sqrt{2}\left\|\boldsymbol{\psi}(C_h^k, e_h^k)\right\|_{(\boldsymbol{\Lambda}_h^{k+1})^{-1}} \leq 2\sqrt{2}\beta_V H \cdot \sqrt{Kd \cdot \log\left(\frac{K+d}{d}\right)}, \tag{20}$$

where the second step is by the Elliptical Potential Lemma (Lemma G.2).

Bound on E_2. By Lemma C.7, we have

$$E_2 \leq 3\left(\sum_{h=1}^{H} W_h\right) \cdot \sqrt{K \cdot \log\frac{2}{\delta}}. \tag{21}$$

Combining Lemma C.5, (20), (21) and $E_3 \leq 0$, we get that, with probability at least $1 - 3\delta$,

$$R^P(K) \leq 6\eta d^{5/2} H\left(\sum_{h=1}^{H} W_h\right) \cdot \sqrt{K} \cdot \log\left(\frac{dKH\min\{|\mathcal{I}|, |\mathcal{J}|\}}{\delta}\right),$$

where the $1 - 3\delta$ probability is from the union bound on the events of Lemma C.1, Lemma C.6 and Lemma C.7. Since $dKH\min\{|\mathcal{I}|, |\mathcal{J}|\}/\delta > 3$, replacing δ with $\delta/3$ and absorbing the constant into the big-O notation, we finish the proof.

□

D Proof of the Main Theory

D.1 Proof of Proposition 4.2

Proof of Proposition 4.2. By (3), the total regret can be written as

$$R(K) = \sum_{k=1}^{K}\left[V_1^\star(s) - V_1^{\pi_k}(s)\right] = \sum_{k=1}^{K}\left[V_1^\star(s) - \overline{V}_1^{\pi_k}(s)\right] + \sum_{k=1}^{K}\left[\overline{V}_1^{\pi_k}(s) - V_1^{\pi_k}(s)\right],$$

where $\overline{V}_1^{\pi_k}$ is the pseudo-value function defined by (7) corresponding to the pseudo-reward $\overline{r}_h$ (defined by (6)) and induced by the policy π_k. Note that here we only care about the Υ part of π_k since matching-transfer has been maximized out by the definition of $\overline{r}_h$.

Now for any policy $\pi \in \Pi$ where $\pi = \{\pi_h\}_{h\in[H]}$, there exists a counterpart $\pi' = \{\pi'_h\}_{h\in[H]}$, such that $\pi_h(C|s) = \pi'_h(C|s)$ for all $C \in \Upsilon$ and $s \in \mathcal{S}$, whereas for the matching part π' always chooses the stable matching w.r.t. the true and unknown utility functions $u_h(C_h, e_h, \cdot, \cdot)$ and $v_h(C_h, e_h, \cdot, \cdot)$. Since π' can be viewed as a function of π, we write $\pi' = \pi'(\pi)$. Since the matching does not affect the context transition by assumption, and the stable matching maximizes total utility by Proposition B.2, we then have that

$$\begin{aligned}
V_1^\star &= \max_{\pi\in\Pi} V_1^\pi \\
&= \max_{\pi\in\Pi} \mathbb{E}_\pi\left[\sum_{h=1}^{H} r_h(s_h, a_h) \,\middle|\, s_1 = s;\ a_h \sim \pi_h(\cdot|s_h), s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h), \forall\, h \in [H]\right] \\
&= \max_{\pi'(\pi):\pi\in\Pi} \mathbb{E}_{\pi'}\left[\sum_{h=1}^{H} r_h(s_h, a_h) \,\middle|\, s_1 = s;\ a_h \sim \pi'_h(\cdot|s_h), s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h), \forall\, h \in [H]\right] \\
&= \max_{\pi'(\pi):\pi\in\Pi} \mathbb{E}_{\pi'}\left[\sum_{h=1}^{H} \overline{r}_h(s_h, e_h) \,\middle|\, s_1 = s;\ e_h \sim \pi'_h(\cdot|s_h), s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, e_h), \forall\, h \in [H]\right] = \overline{V}_1^\star,
\end{aligned}$$

where the fourth step is by definition of $\overline{r}_h$. It follows that

$$R(K) = \sum_{k=1}^{K} \left[\overline{V}_1^{\star}(s) - \overline{V}_1^{\pi_k}(s)\right] + \sum_{k=1}^{K} \left[\overline{V}_1^{\pi_k}(s) - V_1^{\pi_k}(s)\right] = R^P(K) + \sum_{k=1}^{K} \left[\overline{V}_1^{\pi_k}(s) - V_1^{\pi_k}(s)\right]. \tag{22}$$

The second term in the R.H.S. can be written as

$$\begin{aligned}
&\sum_{k=1}^{K} \left[\overline{V}_1^{\pi_k}(s) - V_1^{\pi_k}(s)\right] \\
&\quad = \mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \overline{r}_h(s_h, e_h) - r_h(s_h, a_h) \;\middle|\; s_1 = s;\ a_h \sim \pi_{k,h}(\cdot|s_h), s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h), \forall\, h \in [H]\right].
\end{aligned}$$

Note that $\overline{r}_h(s_h, e_h) - r_h(s_h, a_h)$ is exactly the utility difference defined by (16). Since Subset Instability is at least the utility difference by Proposition B.3, it follows that

$$\begin{aligned}
&\sum_{k=1}^{K} \left[\overline{V}_1^{\pi_k}(s) - V_1^{\pi_k}(s)\right] \\
&\quad \le \sum_{k=1}^{K} \mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \mathrm{SI}(s_h, a_h, u_h, v_h) \;\middle|\; s_1 = s;\ a_h \sim \pi_{k,h}(\cdot|s_h), s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h), \forall\, h \in [H]\right] \\
&\quad = R^M(K),
\end{aligned}$$

where the last step is by the definition of $R^M(K)$ in (9). Plugging into (22), we get

$$R(K) \le R^P(K) + R^M(K).$$

This completes the proof. □

D.2 Bounds for Regression Estimators

We first bound the norm of the regression estimators in the algorithm.

Lemma D.1. *The regression estimators $\mathbf{w}_h^k$ in Algorithm 1 satisfy*

$$\left\|\mathbf{w}_h^k\right\|_2 \le \left(\sum_{l=h}^{H} W_l\right) \cdot \sqrt{\frac{dk}{\lambda}}, \quad \left\|\boldsymbol{\theta}_h^k\right\|_2 \le \sqrt{\frac{d^2 k \cdot \min\{|\mathcal{I}|, |\mathcal{J}|\}}{\lambda}}, \quad \left\|\boldsymbol{\gamma}_h^k\right\|_2 \le \sqrt{\frac{d^2 k \cdot \min\{|\mathcal{I}|, |\mathcal{J}|\}}{\lambda}}.$$

Proof of Lemma D.1. Consider $\mathbf{w}_h^k$ for arbitrary h, k. For any vector $\mathbf{v} \in \mathbb{R}^d$,

$$\begin{aligned}
\left|\mathbf{v}^\top \mathbf{w}_h^k\right| &= \left|\left(\boldsymbol{\Lambda}_h^k\right)^{-1} \sum_{t=1}^{k-1} \boldsymbol{\psi}(C_h^t, e_h^t) \overline{V}_{h+1}^k(C_{h+1}^t)\right| \\
&\le \sum_{t=1}^{k-1} \left|\mathbf{v}^\top \left(\boldsymbol{\Lambda}_h^k\right)^{-1} \boldsymbol{\psi}(C_h^t, e_h^t)\right| \cdot \sum_{l=h}^{H} W_l \\
&\le \left(\sum_{l=h}^{H} W_l\right) \cdot \sqrt{\left[\sum_{t=1}^{k-1} \mathbf{v}^\top \left(\boldsymbol{\Lambda}_h^k\right)^{-1} \mathbf{v}\right] \cdot \left[\sum_{t=1}^{k-1} \boldsymbol{\psi}(C_h^t, e_h^t)^\top \left(\boldsymbol{\Lambda}_h^k\right)^{-1} \boldsymbol{\psi}(C_h^t, e_h^t)\right]} \\
&\le \left(\sum_{l=h}^{H} W_l\right) \cdot \sqrt{dk/\lambda} \cdot \|\mathbf{v}\|_2,
\end{aligned}$$

where the first inequality holds because of the truncation of $\overline{Q}_h^k$ and $\overline{V}_h^k(C, I, J) = \max_e \overline{Q}_h^k(C, e, I, J)$, the second inequality is by the Cauchy-Schwarz inequality, and the last inequality is by Lemma G.1. Since above holds for any $\mathbf{v}$, we conclude that

$$\left\|\mathbf{w}_h^k\right\|_2 = \max_{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 = 1} \left|\mathbf{v}^\top \mathbf{w}_h^k\right| \le \left(\sum_{l=h}^{H} W_l\right) \cdot \sqrt{dk/\lambda}.$$

For $\boldsymbol{\theta}_h^k$, let $\mathbf{v} \in \mathbb{R}^{d\times d}$. Then by the same analysis we have

$$\begin{aligned}
\left|\mathbf{v}^\top \boldsymbol{\theta}_h^k\right| &= \left|\mathbf{v}^\top \left(\boldsymbol{\Sigma}_h^k\right)^{-1} \sum_{t=1}^{k-1} \sum_{(i,j)\in X_h^t} \boldsymbol{\Phi}(C_h^t, e_h^t, i, j) u_h^t(i,j)\right| \\
&\leq \sum_{t=1}^{k-1} \sum_{(i,j)\in X_h^t} \left|\mathbf{v}^\top \left(\boldsymbol{\Sigma}_h^k\right)^{-1} \boldsymbol{\Phi}(C_h^t, e_h^t, i, j)\right| \cdot 1 \\
&\leq \sqrt{\left[\sum_{t=1}^{k-1} \sum_{(i,j)\in X_h^t} \mathbf{v}^\top \left(\boldsymbol{\Sigma}_h^k\right)^{-1} \mathbf{v}\right] \cdot \left[\sum_{t=1}^{k-1} \sum_{(i,j)\in X_h^t} \boldsymbol{\Phi}(C_h^t, e_h^t, i, j)^\top \left(\boldsymbol{\Sigma}_h^k\right)^{-1} \boldsymbol{\Phi}(C_h^t, e_h^t, i, j)\right]} \cdot \\
&\leq \|\mathbf{v}\|_2 \cdot \sqrt{k \min\{|\mathcal{I}|, |\mathcal{J}|\} \cdot d^2/\lambda},
\end{aligned}$$

which implies

$$\left\|\boldsymbol{\theta}_h^k\right\|_2 \leq \sqrt{\frac{d^2 k \cdot \min\{|\mathcal{I}|, |\mathcal{J}|\}}{\lambda}}.$$

The same holds for $\boldsymbol{\gamma}_h^k$.

□

D.3 Proof of Theorem 5.4

We present the complete proof of Theorem 5.4.

Proof of Theorem 5.4. Recall from Proposition B.3 that Subset Instability is at least the utility difference. We thus have

$$\begin{aligned}
R^M(K) &= \sum_{k=1}^{K} \left[\max_{\pi\in\Pi} V_1^\pi(s_1) - V_1^{\pi_k}(s_1)\right] \\
&\leq \sum_{k=1}^{K} \mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \mathrm{SI}(s_h, a_h, u_h, v_h)\right] \\
&= \sum_{k=1}^{K} \Bigg\{ \underbrace{\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H} \mathrm{SI}(s_h, a_h, u_h, v_h)\right]}_{R_k^M} - \underbrace{\sum_{h=1}^{H} \mathrm{SI}(s_h^k, a_h^k, u_h, v_h)}_{\widetilde{R}_k^M} \Bigg\} + \sum_{k=1}^{K}\sum_{h=1}^{H} \mathrm{SI}(s_h^k, a_h^k, u_h, v_h).
\end{aligned}$$

For each $k \in [K]$, let $\mathcal{F}_{k-1}$ denote all the history until the beginning of Episode k. Then $\pi_k \sim \mathcal{F}_{k-1}$ and $R_k^M \sim \mathcal{F}_{k-1}$. Furthermore, we have $R_k^M = \mathbb{E}[\widetilde{R}_k^M \mid \mathcal{F}_{k-1}]$, i.e., R_k^M is the conditional expectation of the realized quantity $\widetilde{R}_k^M$. We can view (17) as the sum of a martingale difference sequence. From Definition 4.1, it holds almost surely that

$$\left|\mathbb{E}_{\pi_k}\left[\mathrm{SI}(s_h, a_h, u_h, v_h)\right] - \mathrm{SI}(s_h^k, a_h^k, u_h, v_h)\right| \leq W_h.$$

By Lemma G.3, we have that, for any $0 < \delta < 1$, with probability at least $1-\delta$,

$$\sum_{l=1}^{K} \left(R_k^M - \widetilde{R}_k^M\right) \leq \left(\sum_{h=1}^{H} W_h\right)\sqrt{2K \log\left(\frac{2}{\delta}\right)}. \tag{23}$$

To bound the second term $\sum_{k=1}^{K} \widetilde{R}_k^M$, note that by Lemma C.1 and Lemma C.3, we have

$$\begin{aligned}
\sum_{k=1}^{K} \widetilde{R}_k^M &\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \sum_{(i,j)\in X_h^k} 4\beta_u \left\|\boldsymbol{\Phi}(C_h^k, e_h^k, i, j)\right\|_{(\boldsymbol{\Sigma}_h^k)^{-1}} \\
&= 4\beta_u \sum_{h=1}^{H} \Bigg(\sum_{k=1}^{K} \sum_{(i,j)\in X_h^k} \left\|\boldsymbol{\Phi}(C_h^k, e_h^k, i, j)\right\|_{(\boldsymbol{\Sigma}_h^k)^{-1}}\Bigg).
\end{aligned} \tag{24}$$

Note that by definition, we have

$$\mathbf{\Sigma}_h^{k+1} = \mathbf{\Sigma}_h^k + \sum_{(i,j)\in X_h^k} \mathbf{\Phi}(C_h^k, e_h^k, i, j)\mathbf{\Phi}(C_h^k, e_h^k, i, j)^\top.$$

By picking $\lambda = 1$ and the assumption that $\|\mathbf{\Phi}\| \le 1$, we have

$$\sum_{(i,j)\in X_h^k} \mathbf{\Phi}(C_h^k, e_h^k, i, j)\mathbf{\Phi}(C_h^k, e_h^k, i, j)^\top \preccurlyeq \min(|I_h|, |J_h|) \cdot \mathbf{I}_{d^2} \preccurlyeq \min(|I_h|, |J_h|) \cdot \mathbf{\Sigma}_h^k.$$

It follows that $\mathbf{\Sigma}_h^{k+1} \preccurlyeq (1 + \min(|I_h|, |J_h|)) \cdot \mathbf{\Sigma}_h^k$, and thus $(\mathbf{\Sigma}_h^k)^{-1} \preccurlyeq (1 + \min(|I_h|, |J_h|)) \cdot (\mathbf{\Sigma}_h^{k+1})^{-1}$. Combining with (24), we get that

$$\begin{aligned}\sum_{k=1}^K \widetilde{R}_k^M &\le 4\beta_u \sum_{h=1}^H \Big(\sqrt{1 + \min(|I_h|, |J_h|)} \cdot \sum_{k=1}^K \sum_{(i,j)\in X_h^k} \big\|\mathbf{\Phi}(C_h^k, e_h^k, i, j)\big\|_{(\mathbf{\Sigma}_h^{k+1})^{-1}}\Big) \\ &\le 4\sqrt{2}\beta_u \sum_{h=1}^H \sqrt{\min(|I_h|, |J_h|)} \cdot \Big(\sum_{k=1}^K \sum_{(i,j)\in X_h^k} \big\|\mathbf{\Phi}(C_h^k, e_h^k, i, j)\big\|_{(\mathbf{\Sigma}_h^{k+1})^{-1}}\Big), \qquad (25)\end{aligned}$$

where the second step is by $\min(|I_h|, |J_h|) \ge 1$. To bound the summation in the bracket, we seek to use the elliptical potential lemma. However, note tat the telescoping sum involves several utility observations (i.e. all $(i, j) \in X_h^k$) at one time, instead of a single observation. To address this issue, for any (k, h), let's index all the pairs $(i, j) \in X_h^k$ by an index $m = 1, \cdots, |X_h^k|$. Here the order of the indexing does not matter. With a slight abuse of notation, we denote $\mathbf{\Phi}_h^k(m) = \mathbf{\Phi}(C_h^k, e_h^k, i, j)$ if (i, j) has index m in our indexing. We also define, for $n = 1, \cdots, |X_h^k|$, that

$$\mathbf{\Sigma}_h^{k+1}(n) = \mathbf{\Sigma}_h^k + \sum_{m=1}^n \mathbf{\Phi}_h^k(m)\mathbf{\Phi}_h^k(m)^\top.$$

By the above definition, we have $\mathbf{\Sigma}_h^k(n_1) \preccurlyeq \mathbf{\Sigma}_h^k(n_2)$ for all $1 \le n_1 < n_2 \le |X_h^k|$, and $\mathbf{\Sigma}_h^{k+1} = \mathbf{\Sigma}_h^{k+1}(|X_h^k|)$. It follows that

$$\begin{aligned}\sum_{k=1}^K \sum_{(i,j)\in X_h^k} \big\|\mathbf{\Phi}(C_h^k, e_h^k, i, j)\big\|_{(\mathbf{\Sigma}_h^{k+1})^{-1}} &= \sum_{k=1}^K \sum_{(i,j)\in X_h^k} \big\|\mathbf{\Phi}(C_h^k, e_h^k, i, j)\big\|_{(\mathbf{\Sigma}_h^{k+1}(|X_h^k|))^{-1}} \\ &= \sum_{k=1}^K \sum_{m=1}^{|X_h^k|} \big\|\mathbf{\Phi}_h^k(m)\big\|_{(\mathbf{\Sigma}_h^{k+1}(|X_h^k|))^{-1}} \\ &\le \sum_{k=1}^K \sum_{m=1}^{|X_h^k|} \big\|\mathbf{\Phi}_h^k(m)\big\|_{(\mathbf{\Sigma}_h^{k+1}(m))^{-1}} \\ &\le \sqrt{K \cdot \min(|I_h|, |J_h|) \cdot d^2 \log\Big(\frac{K\min(|I_h|, |J_h|) + d^2}{d^2}\Big)},\end{aligned}$$

where the first inequality is by $\mathbf{\Sigma}_h^k(|X_h^k|)^{-1} \preccurlyeq \mathbf{\Sigma}_h^k(m)^{-1}$, and the second inequality is by Lemma G.2 with $\lambda = 1$, $\mathbf{\Phi} \in \mathbb{R}^{d^2}$, and $|X_h^k| \le \min(|I_h, J_h|)$. Combining with (25), we get that

$$\sum_{k=1}^K \widetilde{R}_k^M \le 4\sqrt{2}\beta_u d\Big(\sum_{h=1}^H \min(|I_h|, |J_h|)\Big) \cdot \sqrt{K \log\Big(\frac{K\min(|I_h|, |J_h|) + d^2}{d^2}\Big)}.$$

The final bound follows by plugging in the expression of β_u, and using the fact that $W_h \le \min\{|I_h|, |J_h|\}$. The $1 - 2\delta$ probability comes from the union bound of the event of Lemma C.1 and the event of (23).

□

E Proof of Lemmas

E.1 Proof of Lemma C.1

Proof of Lemma C.1. Fix arbitrary h. Since by Assumption 5.2, the noises in the observed utilities are independent and 1-sub-Gaussian, we can apply Theorem 2 in (Abbasi-Yadkori et al., 2011). We then have that, with probability at least $1-\delta/2$, for any $k \in [K]$,

$$\left\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\right\|_{\boldsymbol{\Sigma}_h^k} \leq \sqrt{d^2 \cdot \log\left(\frac{1 + k \cdot \min\{|I_h|, |J_h|\} \cdot d^2/\lambda}{\delta/2}\right)} + \sqrt{\lambda} d.$$

We then have that for any C, e, i, j,

$$\begin{aligned}
u_h^k(C,e,i,j) - u_h(C,e,i,j) &= \langle \boldsymbol{\Phi}(C,e,i,j), \boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h \rangle + \beta_u \|\Phi(C,e,i,j)\|_{(\boldsymbol{\Sigma}_h^k)^{-1}} \\
&= \langle (\boldsymbol{\Sigma}_h^k)^{-1/2} \boldsymbol{\Phi}(C,e,i,j), (\boldsymbol{\Sigma}_h^k)^{1/2} (\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h) \rangle + \beta_u \|\Phi(C,e,i,j)\|_{(\boldsymbol{\Sigma}_h^k)^{-1}} \\
&\geq \beta_u \|\Phi(C,e,i,j)\|_{(\boldsymbol{\Sigma}_h^k)^{-1}} - \|\boldsymbol{\Phi}(C,e,i,j)\|_{(\boldsymbol{\Sigma}_h^k)^{-1}} \cdot \left\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\right\|_{\boldsymbol{\Sigma}_h^k} \\
&\geq 0,
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality and the second in equality is my the choice of β_u. Similarly, we have

$$\begin{aligned}
u_h^k(C,e,i,j) - u_h(C,e,i,j) &\leq \beta_u \|\Phi(C,e,i,j)\|_{(\boldsymbol{\Sigma}_h^k)^{-1}} + \|\boldsymbol{\Phi}(C,e,i,j)\|_{(\boldsymbol{\Sigma}_h^k)^{-1}} \cdot \left\|\boldsymbol{\theta}_h^k - \boldsymbol{\theta}_h\right\|_{\boldsymbol{\Sigma}_h^k} \\
&\leq 2\beta_u \|\Phi(C,e,i,j)\|_{(\boldsymbol{\Sigma}_h^k)^{-1}}.
\end{aligned}$$

The same argument holds for $v_h^k - v_h$ with probability at least $1-\delta/2$.

Finally, we take a union bound and conclude that the event holds with probability at least $1-\delta$. □

E.2 Proof of Lemma C.2

Proof of Lemma C.2. Note that $\overline{r}_h$ is the maximum value of the linear program (4) with (I_h, J_h, u_h, v_h), and $\overline{r}_h^k$ is the maximum value of (4) with (I_h, J_h, u_h^k, v_h^k). Since $u_h \leq u_h^k$ and $v_h \leq v_h^k$ by Lemma C.1 and the weights in (4) are restricted to be, it immediately holds that $\overline{r}_h(C, e, I_h, J_h) \leq \overline{r}_h^k(C, e, I_h, J_h)$.

On the other hand, denote by X_h the matching corresponding to $\overline{r}_h$. It follows that

$$\begin{aligned}
\overline{r}_h(C,e,I_h,J_h) &= \sum_{(i,j)\in X_h} (u_h(C,e,i,j) + v_h(C,e,i,j)) \\
&\geq \sum_{(i,j)\in X_h^k} (u_h(C,e,i,j) + v_h(C,e,i,j)) \\
&= \sum_{(i,j)\in X_h^k} (u_h^k(C,e,i,j) + v_h^k(C,e,i,j)) \\
&\quad - \sum_{(i,j)\in X_h^k} (b_{u,h}(C,e,i,j) + b_{v,h}(C,e,i,j)),
\end{aligned}$$

where the inequality is due to the sub-optimality of X_h^k under (u_h, v_h). The result follows by using

$$\overline{r}_h^k(C,e,I_h,J_h) = \sum_{(i,j)\in X_h^k} (u_h^k(C,e,i,j) + v_h^k(C,e,i,j)),$$

and rearranging the terms. This completes the proof. □

E.3 Proof of Lemma C.3

Lemma C.3 is a restatement of Lemma 5.4 in Jagadeesan et al. (2021). For completeness, we present the proof here. Specifically, we prove a general version of Lemma C.3, which is Lemma E.1 below.

Lemma E.1. *Let (u, v) and $(\widehat{u}, \widehat{v})$ be two pairs of utility functions on the agents set I, J, such that each of u, v, $\widehat{u}$, $\widehat{v}$ maps from $I \times J$ to $\mathbb{R}$. Let $(\widehat{X}, \widehat{\tau})$ be a stable matching on (I, J) w.r.t. the utility functions $(\widehat{u}, \widehat{v})$. Suppose $u \leq \widehat{u}$, $v \leq \widehat{v}$. Then the Subset Instability of $(\widehat{X}, \widehat{\tau})$ w.r.t. the utility (u, v) satisfies*

$$\text{SI}(\widehat{X}, \widehat{\tau}; I, J, u, v) \leq \sum_{i \in I} \left| \widehat{u}(i, \widehat{X}(i)) - u(i, \widehat{X}(i)) \right| + \sum_{j \in J} \left| \widehat{v}(\widehat{X}(j), j) - v(\widehat{X}(j), j) \right|.$$

Proof of Lemma E.1. Define the function

$$\begin{aligned} f(I', J', X, \tau; u, v) = \max_{X'} & \left(\sum_{i \in I'} u(i, X'(i)) + \sum_{j \in J'} v(X'(j), j) \right) \\ & - \sum_{i \in I'} (u(i, X(i)) + \tau(i)) - \sum_{j \in J'} (v(X(j), j) + \tau(j)), \end{aligned}$$

where $I' \times J' \subset I \times J$. Then by Definition 4.1 of Subset Instability, $\text{SI}(X, \tau; I, J, u, v) = \max_{I' \times J' \subset I \times J} f(I', J', X, \tau; u, v)$, and thus

$$\text{SI}(\widehat{X}, \widehat{\tau}; I, J, u, v) - \text{SI}(\widehat{X}, \widehat{\tau}; I, J, \widehat{u}, \widehat{v}) \leq \max_{I' \times J' \subset I \times J} \left[f(I', J', \widehat{X}, \widehat{\tau}; u, v) - f(I', J', \widehat{X}, \widehat{\tau}; \widehat{u}, \widehat{v}) \right].$$

To bound $f(I', J', \widehat{X}, \widehat{\tau}; u, v) - f(I', J', \widehat{X}, \widehat{\tau}; \widehat{u}, \widehat{v})$, we decompose

$$\begin{aligned} & f(I', J', \widehat{X}, \widehat{\tau}; u, v) - f(I', J', \widehat{X}, \widehat{\tau}; \widehat{u}, \widehat{v}) \\ & = \max_{X'} \left(\sum_{i \in I'} u(i, X'(i)) + \sum_{j \in J'} v(X'(j), j) \right) - \max_{X'} \left(\sum_{i \in I'} \widehat{u}(i, X'(i)) + \sum_{j \in J'} \widehat{v}(X'(j), j) \right) \\ & \quad + \sum_{i \in I'} \left(\widehat{u}(i, \widehat{X}(i)) + \widehat{\tau}(i) \right) + \sum_{j \in J'} \left(\widehat{v}(\widehat{X}(j), j) + \widehat{\tau}(i) \right) - \sum_{i \in I'} \left(u(i, \widehat{X}(i)) + \widehat{\tau}(i) \right) \\ & \quad - \sum_{j \in J'} \left(v(\widehat{X}(j), j) + \widehat{\tau}(j) \right) \\ & = \underbrace{\max_{X'} \left(\sum_{i \in I'} u(i, X'(i)) + \sum_{j \in J'} v(X'(j), j) \right) - \max_{X'} \left(\sum_{i \in I'} \widehat{u}(i, X'(i)) + \sum_{j \in J'} \widehat{v}(X'(j), j) \right)}_{\mathbf{I}} \\ & \quad \underbrace{+ \sum_{i \in I'} \left(\widehat{u}(i, \widehat{X}(i)) - u(i, \widehat{X}(i)) \right) + \sum_{j \in J'} \left(\widehat{v}(\widehat{X}(j), j) - v(\widehat{X}(j), j) \right)}_{\mathbf{II}} \end{aligned}$$

Term **I** is nonpositive. To see this, note that by assumption $u \leq \widehat{u}$ and $v \leq \widehat{v}$. Thus the max-weight matching on $I' \times J'$ w.r.t. the utility functions (u, v) cannot exceed the max-weight matching on $I' \times J'$ w.r.t. $(\widehat{u}, \widehat{v})$.

To bound term **II**, note that all the transfers $\widehat{\tau}(i)$ and $\widehat{\tau}(j)$ in the expression cancel out, and it follows that

$$\mathbf{II} \leq \sum_{i \in I} \left(\widehat{u}(i, \widehat{X}(i)) - u(i, \widehat{X}(i)) \right) + \sum_{j \in J} \left(\widehat{v}(\widehat{X}(j), j) - v(\widehat{X}(j), j) \right)$$

where the inequality follows from the assumption that $u \leq \widehat{u}$ and $v \leq \widehat{v}$ and $I' \times J' \subseteq I \times J$. This finishes the proof. □

Proof of Lemma C.3. For any fixed $k \in [K]$ and $h \in [H]$, replace I, J in Lemma E.1 with I_h, J_h, and replace u, v, $\widehat{u}$, $\widehat{v}$ in with $u_h(C_h^k, e_h^k, \cdot, \cdot)$, $v_h(C_h^k, e_h^k, \cdot, \cdot)$, $u_h^k(C_h^k, e_h^k, \cdot, \cdot)$, $v_h^k(C_h^k, e_h^k, \cdot, \cdot)$. Since X_h^k is the stable matching w.r.t. $u_h^k(C_h^k, e_h^k, \cdot, \cdot)$, $v_h^k(C_h^k, e_h^k, \cdot, \cdot)$, we can replace $\widehat{X}$ with X_h^k. It then

follows from Lemma E.1 that

$$\begin{aligned}
\mathrm{SI}_h^k &\le \sum_{i\in I_h} \left|u_h^k(C_h^k, e_h^k, i, X_h^k(i)) - u_h(C_h^k, e_h^k, i, X_h^k(i))\right| \\
&\quad + \sum_{j\in J_h} \left|v_h^k(C_h^k, e_h^k, X_h^k(j), j) - v_h(C_h^k, e_h^k, , X_h^k(j), j)\right| \\
&= \sum_{(i,j)\in X_h^k} \left|u_h^k(C_h^k, e_h^k, i, j) - u_h(C_h^k, e_h^k, i, j)\right| + \left|v_h^k(C_h^k, e_h^k, i, j) - v_h(C_h^k, e_h^k, i, j)\right| \\
&= \sum_{(i,j)\in X_h^k} \left(b_{u,h}(C_h^k, e_h^k, i, j) + b_{v,h}(C_h^k, e_h^k, i, j)\right),
\end{aligned}$$

where the second step holds because the true utility and the estimated utility are zero for unmatched agents under X_h^k, and the last step is by the definition of the bonus function $b_{u,h}$ and $b_{v,h}$. □

E.4 Proof of Planner's Regret Decomposition

We first restate the lemma in its complete form.

Lemma C.5. *The planner's regret defined by* (8) *can be decomposed as*

$$\begin{aligned}
R^P(K) = &\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\left[\mathbb{E}_{\pi^\star}\left[\delta_h^k(C_h, I_h, J_h, e_h) \mid C_1 = C_1^k\right] - \delta_h^k(C_h^k, I_h, J_h, e_h^k)\right]}_{E_1} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\zeta_{k,h}^1 + \zeta_{k,h}^2\right)}_{E_2} \\
&+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[\langle \overline{Q}_h^k(C_h, I_h, J_h, \cdot), \pi_h^\star(\cdot|C_h, I_h, J_h) - \pi_{k,h}(\cdot|C_h, I_h, J_h)\rangle_\Upsilon \middle| C_1 = C_1^k\right]}_{E_3},
\end{aligned}$$

where the expectation is over the trajectory $\{C_h, e_h\}_{h\in[H]}$ induced by executing the policy $\pi^\star$ (on the choice of $e \in \Upsilon$ only), and conditioning on $\{I_h, J_h\}_{h\in[H]}$ being fixed.

Proof of Lemma C.5. Recall the definition of the planner's regret from (8). We write

$$\overline{V}_1^\star(s_1^k) - \overline{V}_1^{\pi_k}(s_1^k) = \underbrace{\overline{V}_1^\star(s_1^k) - \overline{V}_1^k(s_1^k)}_{\mathbf{I}} + \underbrace{\overline{V}_1^k(s_1^k) - \overline{V}_1^{\pi_k}(s_1^k)}_{\mathbf{II}},$$

where $s_h^k = (C_h^k, I_h, J_h)$ by our notation.

Term I. We define two operators $\mathbb{J}_h^\star$ and $\mathbb{J}_{k,h}$ as

$$\mathbb{J}_h^\star Q(s) = \langle Q(s,\cdot), \pi_h^\star(\cdot|s)\rangle_\Upsilon, \quad \mathbb{J}_{k,h}^\star Q(s) = \langle Q(s,\cdot), \pi_{k,h}(\cdot|s)\rangle_\Upsilon,$$

for all $(k,h) \in [H]\times[K]$, $s \in \mathcal{C}\times 2^{\mathcal{I}}\times 2^{\mathcal{J}}$, and function $Q : \mathcal{C}\times\mathcal{I}\times\mathcal{J}\times\Upsilon \to \mathbb{R}$. Then by definition, we have $\overline{V}_h^k = \mathbb{J}_{k,h}\overline{Q}_h^k$, and $\overline{V}_h^\star = \mathbb{J}_h^\star\overline{Q}_h^\star$. It follows that

$$\overline{V}_h^\star - \overline{V}_h^k = \mathbb{J}_h^\star\overline{Q}_h^\star - \mathbb{J}_{k,h}\overline{Q}_h^k = \left(\mathbb{J}_h^\star\overline{Q}_h^\star - \mathbb{J}_h^\star\overline{Q}_h^k\right) + \left(\mathbb{J}_h^\star\overline{Q}_h^k - \mathbb{J}_{k,h}\overline{Q}_h^k\right) = \left(\mathbb{J}_h^\star\overline{Q}_h^\star - \mathbb{J}_h^\star\overline{Q}_h^k\right) + \xi_h^k, \tag{26}$$

where $\xi_h^k \coloneqq \mathbb{J}_h^\star\overline{Q}_h^k - \mathbb{J}_{k,h}\overline{Q}_h^k$. Also, by the definition of δ_h^k in (18), we have

$$\overline{Q}_h^\star - \overline{Q}_h^k = \overline{r}_h + \mathbb{P}_h\overline{V}_{h+1}^\star - \left(\overline{r}_h + \mathbb{P}_h\overline{V}_{h+1}^k\right) + \delta_h^k = \mathbb{P}_h\left(\overline{V}_{h+1}^\star - \overline{V}_{h+1}^k\right) + \delta_h^k.$$

Combining with (26), we get

$$\overline{V}_h^\star - \overline{V}_h^k = \mathbb{J}_h^\star\mathbb{P}_h\left(\overline{V}_{h+1}^\star - \overline{V}_{h+1}^k\right) + \mathbb{J}_h^\star\delta_h^k + \xi_h^k.$$

Applying the above equation recursively, we have that for any C_1, I_1, J_1,

$$
\begin{aligned}
&\overline{V}_1^\star(C_1, I_1, J_1) - \overline{V}_1^k(C_1, I_1, J_1) \\
&\quad = \prod_{h=1}^{H} (\mathbb{J}_h^\star \mathbb{P}_h) \left(\overline{V}_{H+1}^\star - \overline{V}_{H+1}^k\right)(C_1, I_1, J_1) + \sum_{h=1}^{H} \left(\prod_{l=1}^{h-1} \mathbb{J}_l^\star \mathbb{P}_l\right) \mathbb{J}_h^\star \delta_h^k(C_1, I_1, J_1) \\
&\qquad + \sum_{h=1}^{H} \left(\prod_{l=1}^{h-1} \mathbb{J}_l^\star \mathbb{P}_l\right) \xi_h^k(C_1, I_1, J_1) \\
&\quad = \mathbb{E}_{\pi^\star}\left[\sum_{h=1}^{H} \delta_h^k(C_h, I_h, J_h, e_h) \middle| C_1, I_1, J_1\right] \\
&\qquad + \mathbb{E}_{\pi^\star}\left[\sum_{h=1}^{H} \langle \overline{Q}_h^k(C_h, I_h, J_h, \cdot), [\pi_h^\star - \pi_{k,h}](\cdot \mid C_h, I_h, J_h)\rangle_\Upsilon \middle| C_1, I_1, J_1\right],
\end{aligned} \tag{27}
$$

where the second step holds because $\overline{V}_{H+1}^\star = \overline{V}_{H+1}^k = 0$. By definition of the operators, it is clear that here the expectation $\mathbb{E}_{\pi^\star}$ is over the trajectory $\{(C_h, e_h)\}_{h\in[H]}$ induced by the planner executing the policy π^k to choose actions in Υ.

Term II. First note that by (18), the function $\delta_h^k : \mathcal{C} \times 2^{\mathcal{I}} \times 2^{\mathcal{J}} \times \Upsilon \to \mathbb{R}$ can be written as

$$
\delta_h^k = \overline{r}_h + \mathbb{P}_h \overline{V}_{h+1}^k - \overline{Q}_h^k = \overline{r}_h + \mathbb{P}_h \overline{V}_{h+1}^k - \overline{Q}_h^{\pi_k} + \overline{Q}_h^{\pi_k} - \overline{Q}_h^k = \mathbb{P}_h\left(\overline{V}_{h+1}^k - \overline{V}_h^{\pi_k}\right) + \left(\overline{Q}_h^{\pi_k} - \overline{Q}_h^k\right), \tag{28}
$$

where the last step is by $\overline{Q}_h^{\pi_k} = \overline{r}_h + \mathbb{P}_h V_{h+1}^{\pi_k}$. Then for any h, we can write

$$
\begin{aligned}
&\left[\overline{V}_h^k - \overline{V}_h^{\pi_k}\right](C_h^k, I_h, J_h) \\
&\quad = \left[\overline{V}_h^k - \overline{V}_h^{\pi_k} + \delta_h^k - \delta_h^k\right](C_h^k, I_h, J_h) \\
&\quad = \left[\overline{V}_h^k - \overline{V}_h^{\pi_k}\right](C_h^k, I_h, J_h) + \left[\overline{Q}_h^{\pi_k} - \overline{Q}_h^k\right](C_h^k, I_h, J_h, e_h^k) \\
&\qquad + \mathbb{P}_h\left[\overline{V}_{h+1}^k - \overline{V}_{h+1}^{\pi_k}\right](C_h^k, I_h, J_h, e_h^k) - \delta_h^k(C_h^k, I_h, J_h, e_h^k) \\
&\quad = \left[\overline{V}_h^k - \overline{V}_h^{\pi_k}\right](C_h^k, I_h, J_h) - \left[\overline{Q}_h^k - \overline{Q}_h^{\pi_k}\right](C_h^k, I_h, J_h, e_h^k) \\
&\qquad + \mathbb{P}_h\left[\overline{V}_{h+1}^k - \overline{V}_{h+1}^{\pi_k}\right](C_h^k, I_h, J_h, e_h^k) - [\overline{V}_{h+1}^k - \overline{V}_{h+1}^{\pi_k}](C_{h+1}^k, I_{h+1}, J_{h+1}) \\
&\qquad + [\overline{V}_{h+1}^k - \overline{V}_{h+1}^{\pi_k}](C_{h+1}^k, I_{h+1}, J_{h+1}) - \delta_h^k(C_h^k, I_h, J_h, e_h^k) \\
&\quad = [\overline{V}_{h+1}^k - \overline{V}_{h+1}^{\pi_k}](C_{h+1}^k, I_{h+1}, J_{h+1}) - \delta_h^k(C_h^k, I_h, J_h, e_h^k) \\
&\qquad + \underbrace{\left[\mathbb{P}_h[\overline{V}_h^k - \overline{V}_h^{\pi_k}](C_h^k, I_h, J_h, e_h^k) - [\overline{V}_{h+1}^k - \overline{V}_{h+1}^{\pi_k}](C_{h+1}^k, I_{h+1}, J_{h+1})\right]}_{\zeta_{k,h}^2} \\
&\qquad + \underbrace{[\overline{V}_h^k - \overline{V}_h^{\pi_k}](C_h^k, I_h, J_h) - [\overline{Q}_h^k - \overline{Q}_h^{\pi_k}](C_h^k, I_h, J_h, e_h^k)}_{\zeta_{k,h}^1},
\end{aligned}
$$

where the second step is by (28). Applying the above equation recursively, we get

$$
[\overline{V}_1^k - \overline{V}_1^{\pi_k}](C_1^k, I_1, J_1) = \sum_{h=1}^{H} \left(\zeta_{k,h}^1 + \zeta_{k,h}^2\right) - \sum_{h=1}^{H} \delta_h^k(C_h^k, I_h, J_h, e_h^k), \tag{29}
$$

where we use $\overline{V}_{H+1}^\star = \overline{V}_{H+1}^k = 0$ again.

Combining (27) and (29), we get

$$\begin{aligned}
&\sum_{k=1}^{K}\left[\overline{V}_1^{\star}(C_1^k, I_1, J_1) - \overline{V}_1^{\pi_k}(C_1^k, I_1, J_1)\right] \\
&\quad = \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[\delta_h^k(C_h, I_h, J_h, e_h)\middle| C_1, I_1, J_1\right] - \sum_{k=1}^{K}\sum_{h=1}^{H}\delta_h^k(C_h^k, I_h, J_h, e_h^k) \\
&\qquad + \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\zeta_{k,h}^1 + \zeta_{k,h}^2\right) \\
&\qquad + \mathbb{E}_{\pi^\star}\left[\sum_{h=1}^{H}\langle \overline{Q}_h^k(C_h, I_h, J_h, \cdot), [\pi_h^\star - \pi_{k,h}](\cdot \mid C_h, I_h, J_h)\rangle_{\Upsilon}\middle| C_1, I_1, J_1\right],
\end{aligned}$$

which finishes the proof.

□

E.5 Proof of Lemma C.6

Proof of Lemma C.6. By the definition of $\overline{Q}_h^k$ in Algorithm 1, the function δ_h^k satisfies

$$\begin{aligned}
\delta_h^k(C, I, J, e) &= \overline{r}_h(C, I, J, e) + \mathbb{P}_h\overline{V}_{h+1}^k(C, I, J, e) - \overline{Q}_h^k(C, I, J, e) \\
&= \overline{r}_h(C, I, J, e) + \mathbb{P}_h\overline{V}_{h+1}^k(C, I, J, e) - \overline{r}_h^k(C, I, J, e) - \widehat{\mathbb{P}}_h\overline{V}_{h+1}^k(C, I, J, e).
\end{aligned}$$

In the sequel, we will show that $\overline{r}_h^k$ and $\widehat{\mathbb{P}}_h\overline{V}_{h+1}^k$ upper bound $\overline{r}_h$ and $\mathbb{P}_h\overline{V}_{h+1}^k$ respectively.

We first consider the term $\overline{r}_h(C, e, I_h, J_h) - \overline{r}_h^k(C, e, I_h, J_h)$. It immediately follows from Lemma C.2 that, under the event of Lemma C.1,

$$\begin{aligned}
-4\sum_{(i,j)\in X_h^k}\beta_u\left\|\boldsymbol{\Phi}(C, e, i, j)\right\|_{(\boldsymbol{\Sigma}_h^k)^{-1}} &\le -\sum_{(i,j)\in X_h^k}\left(b_{u,h}(C, e, i, j) + b_{v,h}(C, e, i, j)\right) \\
&\le \overline{r}_h(C, e, I_h, J_h) - \overline{r}_h^k(C, e, I_h, J_h) \le 0. \qquad (30)
\end{aligned}$$

We now consider $\widehat{\mathbb{P}}_h\overline{V}_{h+1}^k$. Since we are conditioning on $\{I_h, J_h\}$ which is independent of anything else, we can essentially treat it as a deterministic sequence. By (10), for any $\overline{V}_{h+1}^k$, there exists $\overline{\mathbf{w}}_h^k \in \mathbb{R}^d$ such that for any C, e, $\mathbb{P}_h\overline{V}_{h+1}^k(C, e, I_h, J_h) = \boldsymbol{\psi}(C, e)^\top\overline{\mathbf{w}}_h^k$. This is because

$$\begin{aligned}
\mathbb{P}_h\overline{V}_{h+1}^k(C, e, I_h, J_h) &= \int \overline{V}_{h+1}^k(C', I_{h+1}, J_{h+1})\mathrm{d}\mathbb{P}(C'|C, e) \\
&= \int \overline{V}_{h+1}^k(C', I_{h+1}, J_{h+1})\langle\boldsymbol{\psi}(C, e), \mathrm{d}\boldsymbol{\mu}_h(C')\rangle \\
&= \langle\boldsymbol{\psi}(C, e), \int \overline{V}_{h+1}^k(C', I_{h+1}, J_{h+1})\mathrm{d}\boldsymbol{\mu}_h(C')\rangle.
\end{aligned}$$

Then we can write

$$\begin{aligned}
&\mathbb{P}_h\overline{V}_{h+1}^k(C, e, I_h, J_h) - \widehat{\mathbb{P}}_h\overline{V}_{h+1}^k(C, e, I_h, J_h) \\
&\quad = \boldsymbol{\psi}(C, e)^\top\overline{\mathbf{w}}_h^k - \boldsymbol{\psi}(C, e)^\top\left(\boldsymbol{\Lambda}_h^k\right)^{-1}\sum_{t=1}^{k-1}\boldsymbol{\psi}(C_h^t, e_h^t)\overline{V}_{h+1}^k(C_{h+1}^t, I_{h+1}, J_{h+1}) - \beta_V\cdot\left\|\boldsymbol{\psi}(C, e)\right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}} \\
&\quad = \boldsymbol{\psi}(C, e)^\top\left(\boldsymbol{\Lambda}_h^k\right)^{-1}\left[\boldsymbol{\Lambda}_h^k\overline{\mathbf{w}}_h^k - \sum_{t=1}^{k-1}\boldsymbol{\psi}(C_h^t, e_h^t)\overline{V}_{h+1}^k(C_{h+1}^t, I_{h+1}, J_{h+1})\right] - \beta_V\cdot\left\|\boldsymbol{\psi}(C, e)\right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}} \\
&\quad = \boldsymbol{\psi}(C, e)^\top\left(\boldsymbol{\Lambda}_h^k\right)^{-1}\left[\sum_{t=1}^{k-1}\boldsymbol{\psi}(C_h^t, e_h^t)\left(\mathbb{P}_h\overline{V}_{h+1}^k(C_h^t, e_h^t, I_h, J_h) - \overline{V}_{h+1}^k(C_{h+1}^t, I_{h+1}, J_{h+1})\right)\right] \\
&\qquad + \lambda\,\boldsymbol{\psi}(C, e)^\top\left(\boldsymbol{\Lambda}_h^k\right)^{-1}\overline{\mathbf{w}}_h^k - \beta_V\cdot\left\|\boldsymbol{\psi}(C, e)\right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}, \qquad (31)
\end{aligned}$$

where the first step uses the construction of $\mathbf{w}_h^k$ in Algorithm 1 and the last step uses the construction of $\mathbf{\Lambda}_h^k$. For convenience, we write $s_h^t = (C_h^t, I_h, J_h)$ and $(s_h^t, e_h^t) = (C_h^t, e_h^t, I_h, J_h)$. It follows from the Cauchy-Schwarz inequality that the first part on the R.H.S. of (31) satisfies

$$\begin{aligned}
&\left|\boldsymbol{\psi}(C,e)^\top \left(\mathbf{\Lambda}_h^k\right)^{-1} \left[\sum_{t=1}^{k-1} \boldsymbol{\psi}(C_h^t, e_h^t)\left(\mathbb{P}_h \overline{V}_{h+1}^k(s_h^t, e_h^t) - \overline{V}_{h+1}^k(s_{h+1}^t)\right)\right] + \lambda\, \boldsymbol{\psi}(C,e)^\top \left(\mathbf{\Lambda}_h^k\right)^{-1} \overline{\mathbf{w}}_h^k\right| \\
&\quad \le \|\boldsymbol{\psi}(C,e)\|_{(\mathbf{\Lambda}_h^k)^{-1}} \cdot \left\|\sum_{t=1}^{k-1} \boldsymbol{\psi}(C_h^t, e_h^t)\left(\mathbb{P}_h \overline{V}_{h+1}^k(s_h^t, e_h^t) - \overline{V}_{h+1}^k(s_{h+1}^t)\right)\right\|_{(\mathbf{\Lambda}_h^k)^{-1}} \\
&\qquad + \lambda\, \|\boldsymbol{\psi}(C,e)\|_{(\mathbf{\Lambda}_h^k)^{-1}} \cdot \left\|\overline{\mathbf{w}}_h^k\right\|_{(\mathbf{\Lambda}_h^k)^{-1}}. \qquad (32)
\end{aligned}$$

In the following, to bound (31), we first bound the self-normalized stochastic process using tools from self-normalized martingale. The issue is that, according to Algorithm 1, the function $\overline{V}_{h+1}^k$ depends on the first $(k-1)$ episodes and thus depends on the trajectory $\{(C_h^t, e_h^t, C_{h+1}^t)\}_{t\in[k-1]}$. We thus adopt a common approach to solve this issue by considering the function class containing each value function estimator $\overline{V}_{h+1}^k$. We discuss the detail of the construction of the function class and its covering in Section F. The covering trick allows us to get the following lemma.

Lemma E.2. *Under the setting of Theorem 5.5, with probability at least $1-\delta$, for any $(h,k) \in [H]\times[K]$,*

$$\begin{aligned}
&\left\|\sum_{t=1}^{k-1} \boldsymbol{\psi}(C_h^t, e_h^t)\left(\mathbb{P}_h \overline{V}_{h+1}^k(C_h^t, e_h^t, I_h, J_h) - \overline{V}_{h+1}^k(C_{h+1}^t, I_{h+1}, J_{h+1})\right)\right\|_{(\mathbf{\Lambda}_h^k)^{-1}} \\
&\quad \le 16d^2 \cdot \left(\sum_{h=1}^H W_h\right) \cdot \sqrt{\log\left(\frac{dKH\min\{|\mathcal{I}|, |\mathcal{J}|\}\cdot(\beta_V+\beta_u)}{\delta}\right)}.
\end{aligned}$$

Proof of Lemma E.2. See Appendix E.7 for the proof. □

For the term $\left\|\overline{\mathbf{w}}_h^k\right\|_{(\mathbf{\Lambda}_h^k)^{-1}}$, by Assumption 5.2 and $|\overline{V}_h^k| \le H\min\{|\mathcal{I}|, |\mathcal{J}|\}$, we have

$$\left\|\overline{\mathbf{w}}_h^k\right\|_2 = \left\|\int_{\mathcal{C}} \overline{V}_h^k(C, I_h, J_h)\, \mathrm{d}\mu_h(C)\right\|_2 \le \sqrt{d}\cdot\left(\sum_{h=1}^H W_h\right).$$

Combine the above inequality with Lemma E.2 and (32), and we get that

$$\begin{aligned}
&\left|\boldsymbol{\psi}(C,e)^\top \left(\mathbf{\Lambda}_h^k\right)^{-1} \left[\sum_{t=1}^{k-1} \boldsymbol{\psi}(C_h^t, e_h^t)\left(\mathbb{P}_h \overline{V}_{h+1}^k(s_h^t, e_h^t) - \overline{V}_{h+1}^k(s_{h+1}^t)\right)\right] + \lambda\, \boldsymbol{\psi}(C,e)^\top \left(\mathbf{\Lambda}_h^k\right)^{-1} \overline{\mathbf{w}}_h^k\right| \\
&\quad \le 17d^2 \cdot \left(\sum_{h=1}^H W_h\right)\sqrt{\chi}\cdot \|\boldsymbol{\psi}(C,e)\|_{(\mathbf{\Lambda}_h^k)^1},
\end{aligned}$$

where

$$\chi := \log\left(\frac{dKH\min\{|\mathcal{I}|, |\mathcal{J}|\}\cdot(\beta_V+\beta_u)}{\delta}\right).$$

It remains to show that there exists choice of β_V (or equivalently, the constant η in the description of Theorem 5.5) such that

$$17d^2 \cdot \left(\sum_{h=1}^H W_h\right)\sqrt{\chi} \le \beta_V.$$

Specifically, we show that we can pick some constant η and set

$$\beta_V = \eta d^2 \left(\sum_{h=1}^H W_h\right)\cdot\sqrt{\log\frac{dKH\min\{|\mathcal{I}|, |\mathcal{J}|\}}{\delta}}.$$

Indeed, plug in the expression of β_V and β_u and we get

$$\begin{aligned}\chi &\leq \log\left(\frac{3d^3KH^2\min\{|\mathcal{I}|,|\mathcal{J}|\}^2\cdot\eta\log\left(3dKH\min\{|\mathcal{I}|,|\mathcal{J}|\}/\delta\right)}{\delta}\right)\\ &\leq 3\log\left(\frac{3\eta dKH\min\{|\mathcal{I}|,|\mathcal{J}|\}\cdot\iota}{\delta}\right)\\ &= 3\iota+3\log(\eta)+3\log\left(\iota\right),\end{aligned}$$

where $\iota=\log(3dKH\min\{|\mathcal{I}|,|\mathcal{J}|\}/\delta)$. Since $\iota>\log 3$, it suffices to pick η such that

$$13\sqrt{3\log 3+3\log(\eta)+3\log\left(\log 3\right)}\leq\eta\cdot\log 3,$$

which finishes the proof. □

E.6 Proof of Lemma C.7

Proof of Lemma C.7. The lemma can be proven by standard martingale concentration, similar to the analysis in (Cai et al., 2020; Yang et al., 2020). Specifically, we define the σ-fields as

$$\begin{aligned}\mathcal{F}_{k,h,0}&=\sigma\left(\left\{(C_l^t,e_l^t)_{(l,t)\in[k-1]\times[H]}\right\}\cup\left\{(C_l^k,e_l^k)\right\}_{l\in[H]}\right),\\ \mathcal{F}_{k,h,1}&=\sigma\left(\left\{(C_l^t,e_l^t)_{(l,t)\in[k-1]\times[H]}\right\}\cup\left\{(C_l^k,e_l^k)\right\}_{l\in[H]}\cup\{C_{h+1}^k\}\right).\end{aligned}$$

By definition, it is clear that these σ-fields form a filtration under the dictionary order on the index tuple (k,h,o) where $o\in\{0,1\}$.

Note that for any $(k,h)\in[K]\times[H]$, since $\overline{V}_h^k$, $\overline{Q}_h^k$ and the policy π_k are all functions of the first $(k-1)$ episodes, they are all $\mathcal{F}_{k,1,1}$-measurable. As a result, $\zeta_{k,h}^1$ is $\mathcal{F}_{k,h,1}$-measurable and $\zeta_{k,h}^2$ is $\mathcal{F}_{k,h,2}$-measurable, for all (k,h).

According to Algorithm 1, the planner's action $e_h^k\sim\pi_k(\cdot|C_h^k)$. This indicates that condition on C_h^k, $\overline{V}_h^{\pi_k}(C_h^k,I_h,J_h)-\overline{Q}_h^{\pi_k}(C_h^k,e_h^k,I_h,J_h)=0$. Also, since $e_h^k\leftarrow\operatorname{argmax}_{e\in\Upsilon}\overline{Q}_h^k(s_h^k,e,I_h,J_h)$, and $\overline{V}_h^k(C,I,J)=\max_{e\in\Upsilon}\overline{Q}_h^k(C,e,I,J)$ for all (C,I,J) by the algorithm, we have $\overline{V}_h^k(C_h^k,I_h,J_h)-\overline{Q}_h^k(C_h^k,e_h^k,I_h,J_h)=0$. Altogether we have $\zeta_{k,h}^1=0$ for all $(k,h)\in[K]\times[H]$.

For $\zeta_{k,h}^2$, first note that there is no dependence issue between the value functions and the trajectory since $\overline{V}_{h+1}^k$ and $\overline{V}_{h+1}^{\pi_k}$ are functions of the first $(k-1)$ episodes. Then since $C_{h+1}^k\sim\mathbb{P}_h(\cdot|C_h^k,e_h^k)$, we have

$$\mathbb{E}\left[\zeta_{k,h}^2\mid\mathcal{F}_{k,h,1}\right]=0. \tag{33}$$

Thus we conclude that $\{(\zeta_{k,h}^1,\zeta_{k,h}^2)\}_{(k,h)\in[K]\times[H]}$ is a martingale difference sequence. Since $\overline{V}_h^k$, $\overline{Q}_h^k$, $\overline{V}_h^{\pi_k}$, $\overline{Q}_h^{\pi_k}$ are all bounded by W_h for all h,k, we apply the Azuma-Hoeffding inequality (Lemma G.3) and get that,

$$\mathbb{P}\left(|E_2|\geq\epsilon\right)\leq 2\exp\left(\frac{-\epsilon^2}{8K\sum_{h=1}^H W_h^2}\right).$$

Equivalently, with probability at least $1-\delta$, we have

$$|E_2|\leq\sqrt{8K\cdot\log\frac{2}{\delta}}\cdot\sqrt{\sum_{h=1}^H W_h^2}\leq\sqrt{8K\cdot\log\frac{2}{\delta}}\cdot\Big(\sum_{h=1}^H W_h\Big),$$

which finishes the proof. □

E.7 Proof of Lemma E.2

Proof of Lemma E.2. By the analysis in Section F, there exists a function class $\mathcal{V}$ containing all $\overline{V}_h^k$, and the ϵ-covering number of $\mathcal{V}$ is given by Lemma F.1. Also note that by the truncation, we have

$|\overline{V}_h^k| \leq \sum_{l=h}^H W_l$. Then we apply Lemma G.5 with $R = \sum_{l=h}^H W_l$ and combine with Lemma F.1, and get that, fix any $0 < \epsilon < 1$, with probability at least $1 - \delta/H$, for all $k \in [K]$,

$$\left\| \sum_{t=1}^{k-1} \boldsymbol{\psi}(C_h^t, e_h^t) \left(\mathbb{P}_h \overline{V}_{h+1}^k(C_h^t, e_h^t, I_h, J_h) - \overline{V}_{h+1}^k(C_{h+1}^t, I_{h+1}, J_{h+1}) \right) \right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}^2$$
$$\leq 4\Big(\sum_{l=h}^H W_l\Big)^2 \left[\frac{d}{2} \log\left(\frac{k+\lambda}{\lambda}\right) + 6d^2 \log\left(1 + \frac{dkH \min\{|\mathcal{I}|, |\mathcal{J}|\} \cdot \beta_V}{\min\{\lambda, 1\} \cdot \epsilon}\right) \right.$$
$$\left. + 4d^4 \log\left(1 + \frac{d \min\{|\mathcal{I}|, |\mathcal{J}|\} \cdot \beta_u}{\min\{\lambda, 1\} \cdot \epsilon}\right) + \log\frac{1}{\delta} \right] + \frac{8k^2\epsilon^2}{\lambda}.$$

Let $\lambda = 1$, pick $\epsilon = d^2 \left(\sum_{l=h}^H W_l\right)/K$ and then take a union bound over $h \in [H]$, we get

$$\left\| \sum_{t=1}^{k-1} \boldsymbol{\psi}(C_h^t, e_h^t) \left(\mathbb{P}_h \overline{V}_{h+1}^k(C_h^t, e_h^t, I_h, J_h) - \overline{V}_{h+1}^k(C_{h+1}^t, I_{h+1}, J_{h+1}) \right) \right\|_{(\boldsymbol{\Lambda}_h^k)^{-1}}$$
$$\leq 16d^2 \cdot \left(\sum_{l=h}^H W_l\right) \cdot \sqrt{\log\left(\frac{dKH \min\{|\mathcal{I}|, |\mathcal{J}|\} \cdot (\beta_V + \beta_u)}{\delta}\right)}.$$

□

F Covering Number of Function Classes

In this section, we will construct a function class $\mathcal{V}$ that provably contains $\overline{V}_h^k$ for all $(k, h) \in [K] \times [H]$. And we will compute the covering number of $\mathcal{V}$. The result is summarized by Lemma F.1 below.

Lemma F.1. *Assume $KH > 32$. For any $\epsilon < 1$, the ϵ-covering number of $\mathcal{V}$ is upper bounded by*

$$\log \mathcal{N}_\epsilon^{\mathcal{V}} \leq 6d^2 \log\left(1 + \frac{dKH \min\{|\mathcal{I}|, |\mathcal{J}|\} \cdot \beta_V}{\min\{\lambda, 1\} \cdot \epsilon}\right) + 4d^4 \log\left(1 + \frac{d \min\{|\mathcal{I}|, |\mathcal{J}|\} \cdot \beta_u}{\min\{\lambda, 1\} \cdot \epsilon}\right).$$

To prove Lemma F.1, we will first construct a function class $\mathcal{G}$ that contains all $\widehat{\mathbb{P}}_h \overline{V}_{h+1}^k$, $\mathcal{R}$ that contains all $\overline{r}_h^k$, and $\mathcal{Q}$ that contains all $\overline{Q}_h^k$. The formal definition of these classes will be given in the following.

We also introduce the following technical lemma.

Lemma F.2 (Covering Number of ℓ_2 Ball)**.** *For any $\epsilon > 0$, the ϵ-covering number of the ℓ_2 ball in $\mathbb{R}^d$ with radius L is upper bound by $(1 + 2L/\epsilon)^d$.*

The proof of this classical result can be found in, for example, Chapter 5 in (Vershynin, 2010). Now we prove Lemma F.1.

F.1 Proof of Lemma F.1

Covering of $\widehat{\mathbb{P}}_h \overline{V}_{h+1}^k$. The next lemma is helpful to bound the covering number of the function class containing the function $\widehat{\mathbb{P}}_h \overline{V}_{h+1}^k$. The proof is the same as that of Lemma D.6. in (Jin et al., 2020).

Lemma F.3. *Let $\mathcal{G} = \mathcal{G}(L, B)$ denote the function class with functions mapping from $\mathcal{C} \times \Upsilon$ to $\mathbb{R}$ and of the following parametric form*

$$g(\cdot, \cdot) = \boldsymbol{\psi}(\cdot, \cdot)^\top \mathbf{w} + \beta \cdot \|\boldsymbol{\psi}(\cdot, \cdot)\|_{\boldsymbol{\Lambda}^{-1}},$$

where $\|\mathbf{w}\| \leq L$, $\beta \in [0, B]$, and $\lambda_{\min}(\boldsymbol{\Lambda}) \geq \lambda > 0$. Assume $\|\boldsymbol{\psi}(\cdot, \cdot)\|_2 \leq 1$. Let $\mathcal{N}_\epsilon$ denote the ϵ-covering number of $\mathcal{G}$ with respect to the ℓ_∞ distance. Then we have

$$\log(\mathcal{N}_\epsilon^{\mathcal{G}}) \leq d \log(1 + 4L/\epsilon) + d^2 \log\left[1 + 8d^{1/2}B^2/(\lambda\epsilon^2)\right].$$

Suppose for now that there exists $L_{\mathbf{w}} > 0$ such that $\|\mathbf{w}_h^k\|_2 \leq L_{\mathbf{w}}$ for all (k, h). The value of $L_{\mathbf{w}}$ will be determined later. By applying Lemma F.3 with $L = L_{\mathbf{w}}$ and $B = \beta_V$, we get the following upper bound on the ϵ-covering number of the function class $\mathcal{G}(l_{\mathbf{w}}, \beta_V)$ which contains all $\widehat{\mathbb{P}}_h \overline{V}_{h+1}^k$:

$$\log(\mathcal{N}_\epsilon^{\mathcal{G}}) \leq d \log(1 + 4L_{\mathbf{w}}/\epsilon) + d^2 \log\left[1 + 8d^{1/2}\beta_V^2/(\lambda\epsilon^2)\right]. \tag{34}$$

Covering of $\overline{r}_h^k$. We now define a function class $\mathcal{R}$ which provably contains all the pseudo-reward estimates $\overline{r}_h^k$. Formally speaking, by Algorithm 2, the functions in $\mathcal{R}$ are parametrized by the utility function estimates u_h^k and v_h^k. Denote the functions class containing all these utility function estimates by $\mathcal{U}$. Then according to Algorithm 4, any function $u : \mathcal{C} \times \Upsilon \times \mathcal{I} \times \mathcal{J} \to \mathbb{R}$ in $\mathcal{U}$ can be written as

$$u(C, e, i, j) = \left(\langle \mathbf{\Phi}(C, e, i, j), \boldsymbol{\theta}\rangle + \beta_u \sqrt{\mathbf{\Phi}(C, e, i, j)^\top \mathbf{\Sigma}^{-1} \mathbf{\Phi}(C, e, i, j)}\right)_{[-1,1]}. \tag{35}$$

where $\mathbf{\Phi} \in \mathbb{R}^{d^2}$ satisfying $\|\mathbf{\Phi}\|_1 \leq 1$ by Assumption 5.2, $\|\boldsymbol{\theta}\| \leq L_u$ for some $L_u > 0$ to be determined, and $\mathbf{\Sigma} \in \mathbb{R}^d \times \mathbb{R}^d$ such that $\lambda_{\min}(\mathbf{\Sigma}) \geq \lambda$. Since the truncation is a contraction mapping, by Lemma F.3, the ϵ-covering number of $\mathcal{U}$ is upper bounded by

$$\log\left(\mathcal{N}_\epsilon^u\right) \leq d^2 \log\left(1 + 4L_u/\epsilon\right) + d^4 \log\left[1 + 8d\beta_u^2/(\lambda\epsilon^2)\right]. \tag{36}$$

We now consider the function class $\mathcal{R}$. We formally define $\mathcal{R}$ to be the function class such that any function $r \in \mathcal{R}$ can be represented by

$$r(C, e, I, J) = \texttt{RE}(u(C, e, \cdot, \cdot), v(C, e, \cdot, \cdot), I, J)$$

for some $u, v \in \mathcal{U}$. Let functions $r_1, r_2 \in \mathcal{R}$ be parametrized by u_1, v_1 and u_2, v_2 respectively, such that

$$\begin{aligned} r_1(C, e, I, J) &= \texttt{RE}(u_1(C, e, \cdot, \cdot), v_1(C, e, \cdot, \cdot), I, J), \\ r_2(C, e, I, J) &= \texttt{RE}(u_2(C, e, \cdot, \cdot), v_2(C, e, \cdot, \cdot), I, J), \end{aligned}$$

for all $C \in \mathcal{C}$, $e \in \upsilon$, $I \subset \mathcal{I}$ and $J \subset \mathcal{J}$. According to the linear program 4, there exist some weights $w_1 = \{w_{1,i,j}\}_{(i,j)\in I\times J}$ and $w_2 = \{w_{2,i,j}\}_{(i,j)\in I\times J}$, such that

$$\begin{aligned} r_1(C, e, I, J) &= \sum_{(i,j)\in I\times J} w_{1,i,j}\left[u_1(C, e, i, j) + v_1(C, e, i, j)\right], \\ r_2(C, e, I, J) &= \sum_{(i,j)\in I\times J} w_{2,i,j}\left[u_2(C, e, i, j) + v_2(C, e, i, j)\right]. \end{aligned}$$

It follows that

$$\begin{aligned} (r_1 - r_2)(C, e, I, J) &= \sum_{(i,j)\in I\times J} w_{1,i,j}\left[u_1(C, e, i, j) + v_1(C, e, i, j)\right] \\ &\quad - \sum_{(i,j)\in I\times J} w_{2,i,j}\left[u_2(C, e, i, j) + v_2(C, e, i, j)\right] \\ &\leq \sum_{(i,j)\in I\times J} w_{2,i,j}\left[u_1(C, e, i, j) + v_1(C, e, i, j)\right] \\ &\quad - \sum_{(i,j)\in I\times J} w_{2,i,j}\left[u_2(C, e, i, j) + v_2(C, e, i, j)\right] \\ &\leq \sum_{(i,j)\in I\times J} w_{2,i,j}\left[(u_1(C, e, i, j) - u_2(C, e, i, j)) + (v_1(C, e, i, j) - v_2(C, e, i, j))\right] \\ &\leq \min\{|I|, |J|\} \cdot \left(\|u_1 - u_2\|_\infty + \|v_1 - v_2\|_\infty\right), \end{aligned}$$

where the second step holds because w_1 is the optimal weight given u_1 and v_1 and w_2 satisfies the constraint of the linear program with u_1 and v_1. The same upper bound holds for the difference $(r_2 - r_1)$. Therefore, for any I, J, we have

$$\|r_1(\cdot, \cdot, I, J) - r_2(\cdot, \cdot, I, J)\|_\infty \leq \min\{|I|,\ |J|\} \cdot \left(\|u_1 - u_2\|_\infty + \|v_1 - v_2\|_\infty\right).$$

Since $I \subset \mathcal{I}$ and $J \subset \mathcal{J}$, in order for $\|r_1 - r_2\|_\infty \le \epsilon$ to hold, it suffices to have $\|u_1 - u_2\|_\infty \le \epsilon'$ and $\|v_1 - v_2\|_\infty \le \epsilon'$ where $\epsilon' = \epsilon/(2\min\{|\mathcal{I}|, |\mathcal{J}|\})$. Therefore, by (36), the ϵ-covering number of $\mathcal{R}$ satisfies

$$\begin{aligned}\log \mathcal{N}_\epsilon^{\mathcal{R}} &\le 2\log \mathcal{N}_{\epsilon'}^{u} \le 2d^2 \log\left(1 + 4L_u/\epsilon'\right) + 2d^4 \log\left[1 + 8d\beta_u^2/(\lambda {\epsilon'}^2)\right] \\ &\le 2d^2 \log\left(1 + \frac{8L_u \min\{|\mathcal{I}|, |\mathcal{J}|\}}{\epsilon}\right) + 2d^4 \log\left[1 + \frac{32d\beta_u^2 \left(\min\{|\mathcal{I}|, |\mathcal{J}|\}\right)^2}{\lambda\epsilon^2}\right].\end{aligned} \tag{37}$$

In the above we have shown that the function class $\mathcal{R}$ contains all $\overline{r}_h^k$ and $\mathcal{G}$ contains all $\widehat{\mathbb{P}}_h \overline{V}_{h+1}^k$. We now define the function class $\mathcal{Q} := \mathcal{R} + \mathcal{G}$ as

$$\mathcal{Q} := \left\{ (r+g)_{[0, \sum_{l=h}^H W_l]} \,\middle|\, r \in \mathcal{R},\ g \in \mathcal{G} \right\}.$$

Then it immediately follows from the algorithm that $\mathcal{Q}$ contains all $\overline{Q}_h^k$ functions. By (34) and (37), the ϵ-covering number of the function class $\mathcal{Q}$ can be upper bounded by

$$\begin{aligned}\log \mathcal{N}_\epsilon^{\mathcal{Q}} \le\ & d\log(1 + 4L_{\mathbf{w}}/\epsilon) + d^2 \log\left[1 + 8d^{1/2}\beta_V^2/(\lambda\epsilon^2)\right] \\ & + 2d^2 \log\left(1 + \frac{8L_u \min\{|\mathcal{I}|, |\mathcal{J}|\}}{\epsilon}\right) \\ & + 2d^4 \log\left[1 + \frac{32d\beta_u^2 \left(\min\{|\mathcal{I}|, |\mathcal{J}|\}\right)^2}{\lambda\epsilon^2}\right],\end{aligned} \tag{38}$$

Since by construction, $\overline{V}_h^k(C, I, J) = \max_e \overline{Q}_h^k(C, e, I, J)$ and taking the maximum is a contraction mapping, the upper bound in (38) also holds for $\log \mathcal{N}_\epsilon^{\mathcal{V}}$.

By Lemma D.1, we can pick

$$L_{\mathbf{w}} = \left(\sum_{h=1}^H W_h\right) \cdot \sqrt{dK/\lambda} \qquad \text{and} \qquad L_u = \sqrt{\frac{d^2 K \cdot \min\{|\mathcal{I}|, |\mathcal{J}|\}}{\lambda}}.$$

From the above analysis, we can simplify the R.H.S. of (38) and get the desired bound for the covering number of $\mathcal{V}$. This finishes the proof of Lemma F.1.

G Auxiliary Lemmas

Lemma G.1 (Lemma D.1 in Jin et al. 2020)**.** *For arbitrary d, let $\mathbf{\Lambda}_k = \lambda \mathbf{I}_d + \sum_{t=1}^{k-1} \mathbf{x}_t \mathbf{x}_t^\top$ where $\mathbf{x}_t \in \mathbb{R}^d$ and $\lambda > 0$. Then*

$$\sum_{t=1}^{k-1} \mathbf{x}_t^\top \left(\mathbf{\Lambda}_k\right)^{-1} \mathbf{x}_t \le d.$$

Proof of Lemma G.1. We can write

$$\sum_{t=1}^{k-1} \mathbf{x}_t^\top \left(\mathbf{\Lambda}_k\right)^{-1} \mathbf{x}_t = \sum_{t=1}^{k-1} \mathrm{tr}\left(\mathbf{x}_t^\top \left(\mathbf{\Lambda}_k\right)^{-1} \mathbf{x}_t\right) = \mathrm{tr}\Big(\left(\mathbf{\Lambda}_k\right)^{-1} \sum_{t=1}^{k-1} \mathbf{x}_t^\top \mathbf{x}_t\Big).$$

Denote the eigenvalue of $\sum_{t=1}^{k-1} \mathbf{x}_t^\top \mathbf{x}_t$ as $\{\lambda_1, \cdots, \lambda_d\}$, and decompose $\sum_{t=1}^{k-1} \mathbf{x}_t^\top \mathbf{x} = \mathbf{U}\mathrm{diag}(\lambda_1, \cdots, \lambda_d)\mathbf{U}^\top$. Then we have $\mathbf{\Lambda}_k = \mathbf{U}\mathrm{diag}(\lambda_1 + \lambda, \cdots, \lambda_d + \lambda)\mathbf{U}^\top$. It follows that $\mathrm{tr}((\mathbf{\Lambda}_k)^{-1} \sum_{t=1}^{k-1} \mathbf{x}_t^\top \mathbf{x}_t) = \sum_{j=1}^d \lambda_j/(\lambda_j + \lambda) \le d$. □

The next is the well-known Elliptical Potential Lemma (Cesa-Bianchi and Lugosi, 2006; Abbasi-Yadkori et al., 2011; Lattimore and Szepesvári, 2020).

Lemma G.2 (Elliptical Potential Lemma)**.** *For arbitrary d, let $\mathbf{\Lambda}_k = \lambda \mathbf{I}_d + \sum_{t=1}^{k-1} \mathbf{x}_t \mathbf{x}_t^\top$ where $\mathbf{x}_t \in \mathbb{R}^d$ and $\lambda > 0$. Then*

$$\sum_{t=1}^{k} \|\mathbf{x}_t\|_{(\mathbf{\Lambda}_{t+1})^{-1}} \leq \sqrt{kd \log\left(\frac{k + d\lambda}{d\lambda}\right)}.$$

Lemma G.3 (Azuma-Hoeffding inequality (Azuma, 1967))**.** *Let $\{X_t\}_{t=0}^{\infty}$ be a real-valued martingale such that for every $t \geq 1$, it holds that $|X_t - X_{t-1}| \leq B_t$ for some $B_t \geq 0$. Then*

$$\mathbb{P}\left(|X_t - X_0| \geq \epsilon\right) \leq 2 \exp\left(\frac{-\epsilon^2}{2\sum_{\tau=1}^{t} B_\tau^2}\right).$$

G.1 Concentration Inequalities for Self-normalized Martingales

Theorem G.4 (Hoeffding inequality for Self-normalized martingales (Abbasi-Yadkori et al., 2011))**.** *Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process. Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration, such that η_t is $\mathcal{F}_t$-measurable. Assume $\eta_t \mid \mathcal{F}_{t-1}$ is zero-mean and R-subgaussian for some $R > 0$, i.e.,*

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}\left[e^{\lambda \eta_t | \mathcal{F}_{t-1}}\right] \leq e^{\lambda^2 R^2/2}.$$

Let $\{\mathbf{x}_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where $\mathbf{x}_t$ is $\mathcal{F}_{t-1}$-measurable. Assume $\mathbf{\Lambda}_0$ is a $d \times d$ positive definite matrix, and let $\mathbf{\Lambda}_t = \mathbf{\Lambda}_0 + \sum_{s=1}^{t} \mathbf{x}_s \mathbf{x}_s^\top$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t > 0$,

$$\left\| \sum_{s=1}^{t} \mathbf{x}_s \eta_s \right\|_{\mathbf{\Lambda}_t^{-1}}^2 \leq 2R^2 \log\left(\frac{\det(\mathbf{\Lambda}_t)^{1/2} \det(\mathbf{\Lambda}_0)^{-1/2}}{\delta}\right).$$

Lemma G.5 (Lemma D.4 in Jin et al. 2020)**.** *Let $\mathcal{V}$ be a function class such that any $V \in \mathcal{V}$ maps from $\mathcal{S} \to \mathbb{R}$ and $\|V\|_\infty \leq R$. Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{s_t\}_{t=1}^{\infty}$ be a stochastic process in the space $\mathcal{S}$ such that s_t is $\mathcal{F}_t$-measurable. Let $\{\mathbf{x}\}_{t=0}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that $\mathbf{x}_t$ is $\mathcal{F}_{t-1}$-measurable and $\|\mathbf{x}\|_2 \leq 1$. Let $\mathbf{\Lambda}_k = \lambda \mathbf{I} + \sum_{t=1}^{k-1} \mathbf{x}_t \mathbf{x}_t^\top$. Then for any $\delta > 0$, with probability at least $1 - \delta$, for any k, and any $V \in \mathcal{V}$, we have*

$$\left\| \sum_{t=1}^{k-1} \mathbf{x}_t \left[V(s_t) - \mathbb{E}\left[V(s_t) \mid \mathcal{F}_{t-1}\right]\right] \right\|_{(\mathbf{\Lambda}_k)^{-1}}^2 \leq 4R^2 \left[\frac{d}{2} \log\left(\frac{k+\lambda}{\lambda}\right) + \log \frac{\mathcal{N}_\epsilon^{\mathcal{V}}}{\delta}\right] + \frac{8k^2\epsilon^2}{\lambda},$$

where $\mathcal{N}_\epsilon^{\mathcal{V}}$ is the ϵ-covering number of $\mathcal{V}$ with respect to the ℓ_∞ distance.

Proof of Lemma G.5. For any $V \in \mathcal{V}$, there exists V' in the ϵ-covering such that

$$V = V' + \Delta_V \quad \text{and} \quad \|\Delta_V\|_\infty \leq \epsilon.$$

Then we have

$$\begin{aligned}\left\| \sum_{t=1}^{k-1} \mathbf{x}_t \left[V(s_t) - \mathbb{E}\left[V(s_t) \mid \mathcal{F}_{t-1}\right]\right] \right\|_{(\mathbf{\Lambda}_k)^{-1}}^2 &\leq 2\left\| \sum_{t=1}^{k-1} \mathbf{x}_t \left[V'(s_t) - \mathbb{E}\left[V'(s_t) \mid \mathcal{F}_{t-1}\right]\right] \right\|_{(\mathbf{\Lambda}_k)^{-1}}^2 \\ &\quad + 2\left\| \sum_{t=1}^{k-1} \mathbf{x}_t \left[\Delta_V(s_t) - \mathbb{E}\left[\Delta_V(s_t) \mid \mathcal{F}_{t-1}\right]\right] \right\|_{(\mathbf{\Lambda}_k)^{-1}}^2.\end{aligned}$$

For the first term on the R.H.S., we apply Theorem G.4 and a union bound to the ϵ-covering. The second term can be bound by $8k^2\epsilon^2/\lambda$ by using $\|\mathbf{x}_t\|_2 \leq 1$, $\lambda_{\min}(\mathbf{\Lambda}_k) \geq \lambda$ and $\|\Delta_V\|_\infty \leq \epsilon$. □