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):

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

and its dual program:

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

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

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

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 $\text{SI}(X, \tau; I, J, u, v)$ 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.*

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

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 $\beta_u = \sqrt{d^2 \log[2(1+d^2K \max_h \min\{|I_h|,|J_h|\})/(\lambda\delta)]} + \sqrt{\lambda} d$. 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 $\zeta_{k,h}^1, \zeta_{k,h}^2$:

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

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

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

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. Note that our notion of meta algorithm for Algorithm 1 is different from that of meta learning (Finn et al., 2017; Xu et al., 2021).

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*

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

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

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

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 $\|\boldsymbol{\psi}\|_2 \le 1$ 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:

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

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

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

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

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

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

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*

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

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

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

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

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

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

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

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

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

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 $u \le \widehat{u}$, $v \le \widehat{v}$. Then the Subset Instability of $(\widehat{X}, \widehat{\tau})$ w.r.t. the utility (u, v) satisfies*

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

Proof of Lemma E.1. Define the function

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

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

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

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

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

Term **I** is nonpositive. To see this, note that by assumption $u \le \widehat{u}$ and $v \le \widehat{v}$. 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

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

where the inequality follows from the assumption that $u \le \widehat{u}$ and $v \le \widehat{v}$ 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*

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

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 $\xi_h^k := \mathbb{J}_h^\star\overline{Q}_h^k - \mathbb{J}_{k,h}\overline{Q}_h^k$. 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,

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

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

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

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

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

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 $\overline{V}_{h+1}^k$. The covering trick is a commonly used technique (Ling et al., 2019), and we will 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

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

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 $|\overline{V}_h^k| \le \sum_{l=h}^H W_l$. 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 $k \in [K]$,

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

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

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

□

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*

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

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

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

Suppose for now that there exists $L_{\mathbf{w}} > 0$ such that $\|\mathbf{w}_h^k\|_2 \le L_{\mathbf{w}}$ 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 $\widehat{\mathbb{P}}_h \overline{V}_{h+1}^k$:

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

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

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

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

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

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

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

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

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

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.,*

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

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

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

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 $\|\Delta_V\|_\infty \leq \epsilon$. □