Journal of Machine Learning Research 25 (2024) 1-73 Submitted 2/23; Revised 12/23; Published 8/24

Learning Dynamic Mechanisms in Unknown Environments: A Reinforcement Learning Approach

Shuang Qiu[*] MASQIU@UST.HK
The Hong Kong University of Science and Technology
Hong Kong, China

Boxiang Lyu[*] BLYU@CHICAGOBOOTH.EDU
The University of Chicago
Chicago, IL, USA

Qinglin Meng[*] MENG160@PURDUE.EDU
Purdue University
West Lafayette, IN, USA

Zhaoran Wang ZHAORANWANG@GMAIL.COM
Northwestern University
Evanston, IL, USA

Zhuoran Yang ZHUORAN.YANG@YALE.EDU
Yale University
New Haven, CT, USA

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
University of California
Berkeley, CA, USA

Editor: Tor Lattimore

Abstract

Dynamic mechanism design studies how mechanism designers should allocate resources among agents in a time-varying environment. We consider the problem where the agents interact with the mechanism designer according to an unknown Markov Decision Process (MDP), where agent rewards and the mechanism designer's state evolve according to an episodic MDP with unknown reward functions and transition kernels. We focus on the online setting with linear function approximation and propose novel learning algorithms to recover the dynamic Vickrey-Clarke-Grove (VCG) mechanism over multiple rounds of interaction. A key contribution of our approach is incorporating reward-free online Reinforcement Learning (RL) to aid exploration over a rich policy space to estimate prices in the dynamic VCG mechanism. We show that the regret of our proposed method is upper bounded by $\widetilde{\mathcal{O}}(T^{2/3})$ and further devise a lower bound to show that our algorithm is efficient, incurring the same $\Omega(T^{2/3})$ regret as the lower bound, where T is the total number of rounds. Our work establishes the regret guarantee for online RL in solving dynamic mechanism design problems without prior knowledge of the underlying model.

Keywords: Mechanism Design, Dynamic VCG Mechanism, Reinforcement Learning

[*] Equal contribution. Random order.

1. Introduction

Mechanism design is a branch of economics studying the allocation of goods among rational agents (Myerson, 1989). Its sub-field, dynamic mechanism design, focuses on the setting where the environment, such as agents' preferences, may vary with time (Bergemann and Välimäki, 2019). It has attracted significant research interest from economists and computer scientists (Pavan et al., 2014; Parkes and Singh, 2003) over decades. Many real-world problems, such as Uber's surge pricing, the wholesale energy market, and congestion control, have all been studied under this framework (Chen and Sheldon, 2016; Bejestani and Annaswamy, 2014; Barrera and Garcia, 2014). However, existing work usually requires prior knowledge of key parameters or functionals in the problem, such as the optimal policy or the agents' valuations of goods (Parkes and Singh, 2003; Pavan et al., 2009). Such requirements may be unrealistic in real life.

A promising emerging research direction is learning dynamic mechanisms from repeated interactions with the environment. Drawing inspiration from Bergemann and Välimäki (2010) and Parkes and Singh (2003), we propose the first algorithm that can learn a dynamic mechanism from repeated interactions via reinforcement learning (RL) with no prior knowledge of the problem.

As a first attempt, we focus on learning a dynamic generalization of the classic Vickrey-Clarke-Groves (VCG) mechanism (Vickrey, 1961; Clarke, 1971; Groves, 1979). More specifically, we consider the case where the interaction between a group of agents and a single seller is modeled as an episodic linear Markov Decision Process (MDP) (Jin et al., 2020b; Yang and Wang, 2019; Jin et al., 2020c), where the seller takes actions to determine the allocation of a class of scarce resources among agents. Our task is to learn an ideal mechanism from repeated interactions via online RL (Jin et al., 2020b; Cai et al., 2019). The mechanism we consider implements the policy that maximizes social welfare and charges each agent according to the celebrated Clarke pivot rule (Clarke, 1971). A slight variant of the mechanism has been discussed under known MDP dynamics in Parkes (2007), and we describe the mechanism in full detail in Section 2.

A key challenge we resolve is estimating the VCG price without prior knowledge of the MDP. In particular, the VCG price charged to each agent i is characterized by the externality of that agent, that is, the difference between the maximum social welfare of the whole group and that when agent i is absent (Karlin and Peres, 2017; Groves, 1979). In other words, it is the loss that an agent's participation incurs on other agents' welfare. Estimating the VCG price in our dynamic setting requires learning the optimal policy of the fictitious problem where agent i is absent. Such a policy is never executed by the seller, and thus it is challenging to assess its uncertainty from data. Existing methods target to estimate the optimal policy well. However, they have no guarantees on how well they estimate the fictitious policies. Therefore it is impossible to accurately estimate VCG prices via a direct application of prior online RL algorithms (Jin et al., 2020b; Cai et al., 2019; Zhou et al., 2021a).

To address this challenge, our algorithm incorporates a reward-free exploration subroutine to ensure sufficient coverage over the policy space, thereby reducing the uncertainty of all policies, ensuring that we can even reduce the uncertainty about the fictitious policies (Jin et al., 2020a; Wang et al., 2020; Qiu et al., 2021; Zhang et al., 2021; Kaufmann et al., 2021).

However, such a reward-free approach comes at a price—our proposed approach attains $\widetilde{\mathcal{O}}(T^{2/3})$ regret in terms of social welfare, agent utility, and seller utility, as opposed to the common $\widetilde{\mathcal{O}}(T^{1/2})$ regret in online RL (Jin et al., 2020b). Moreover, we further derive a matching lower bound for the regrets, showing that our algorithm is minimax optimal up to multiplicative factors of problem-dependent terms.

To summarize, our contributions are threefold. First, we develop the first reinforcement learning algorithm that can recover an optimal dynamic mechanism with no prior knowledge of the problem. In particular, our algorithm is separated into two phases, namely, exploration and exploitation. In the exploration phase, we propose to learn the underlying model via reward-free exploration. Then, in the exploitation phase, the algorithm executes a data-driven policy by solving a planning problem using the collected dataset. Moreover, our algorithm is able to handle large state spaces by incorporating linear function approximation. Second, we prove that the proposed algorithm achieves sublinear regret upper bounds in terms of the various regret notions, such as the welfare regret and individual regret of the seller and buyers. Our algorithm is proven to approximately satisfy the three key mechanism design desiderata — truthfulness, individual rationality, and efficiency. Finally, we demonstrate that the $\widetilde{\mathcal{O}}(T^{2/3})$ regret has the minimax optimal dependency in T by establishing a matching regret lower bound. To our knowledge, we seem to establish the first provably efficient reinforcement learning algorithm for learning a dynamic mechanism.

1.1 Related Works

There is a wealth of literature on dynamic mechanism design. Parkes and Singh (2003); Parkes et al. (2004) are two of the earliest works that analyze dynamic mechanism design from an MDP perspective, and the proposed mechanism is applied to a real-world problem in Friedman and Parkes (2003). Bergemann and Välimäki (2006) generalize the VCG mechanism based on the marginal contribution of each agent and derives a mechanism that is truth-telling in every period. Bapna and Weber (2005) focus on the dynamic auction setting and formulate the problem as a multi-arm bandit problem. Athey and Segal (2013) adapt the d'Aspremont-Gerard-Varet (AGV) mechanism (d'Aspremont and Gérard-Varet, 1979) to the dynamic setting and design an efficient, budget balanced, and Bayesian incentive compatible mechanism. Pavan et al. (2009) derive the first order conditions of incentive compatibility in dynamic mechanisms. Cavallo (2008) devises a dynamic allocation rule for auctions in the multi-arm bandits setting, where a single good is distributed among agents over multiple rounds. Cavallo et al. (2009) study the truthful implementation of efficient policies when agents have dynamic types. Pavan et al. (2014) extend the seminal work of Myerson (1989) and characterize perfect Bayesian equilibrium-implementable allocation rules in the dynamic regime. Cavallo (2009); Bergemann and Pavan (2015); Bergemann and Välimäki (2019) provide useful surveys of dynamic mechanism research. Kandasamy et al. (2020) studies online learning of the VCG mechanism with stationary multi-arm bandits. Our work considers a more challenging setting modeled by an episodic MDP, where the agents' rewards are state-dependent and may evolve over time within each episode. More importantly, Kandasamy et al. (2020) estimates the VCG price via uniformly exploring over all arms, which cannot be directly applied to the dynamic setting (Wang et al., 2020). Rather than uniformly bounding the uncertainty over all *actions*, our approach bounds the uncertainty over all implementable *policies* via a variant of least-squares value iteration and enjoys provable efficiency under the function approximation setting. Distinct from the major focus of our work, Simchowitz and Slivkins (2023) studies online mechanism design with MDPs from a rather different angle. In their work, the mechanism designer encourages exploration by sending specific information to the agents. More specifically, the agents initially have beliefs or prior distributions over the MDP's parameters. The mechanism designer can reveal to the agents some information, such as information about the MDP's transition and reward. The agents then update their beliefs about the underlying MDP and execute the optimal policy according to the updated beliefs or their posterior distribution over the MDP's parameters. The goal is to incentivize agents to explore by controlling the information they receive. However, our work focuses on implementing the welfare-maximizing policy among a group of agents by controlling the price that each agent pays. In other words, theirs focuses on adjusting information, whereas ours focuses on adjusting price. Additionally, our work focuses on a more general linear MDP than the tabular MDP studied in their work.

There are many recent works concerning provably efficient RL for MDPs with linear structures in the absence of generative models (Yang and Wang, 2019; Du et al., 2019; Yang and Wang, 2020; Jin et al., 2020b; Cai et al., 2019). Jin et al. (2020b) provides the first provably efficient RL algorithm for linear MDPs that incorporates exploration. Zhou et al. (2021b) provides a provably efficient algorithm for infinite-horizon discounted linear MDPs. Ayoub et al. (2020) studies a model-based regime where the transition kernel belongs to a family of models known to the learning agent. Zhou et al. (2021a) proposes a computationally efficient nearly minimax optimal algorithm for the linear MDP whose transition kernel is a linear mixture model. These works require (noisy) feedback of the reward function in the learning process.

Reward-free exploration in reinforcement learning has recently attracted a lot of attention, in which the agents explore the environment without any feedback of the reward. Specifically, Jin et al. (2020a) introduces the problem of reward-free exploration in RL and proposes a sample-efficient algorithm for tabular MDPs. Ménard et al. (2021); Kaufmann et al. (2021) provide improved algorithms and tighter rates, also for tabular MDPs. Zhang et al. (2021) further improves the analysis and obtains nearly minimax-optimal sample complexity bounds. Wang et al. (2020); Zanette et al. (2020); Chen et al. (2021); Wagenmaker et al. (2022) study reward-free RL algorithms for linear or linear mixture MDPs and Qiu et al. (2021) for kernel and neural function approximations. Moreover, Kong et al. (2021) proposes reward-free algorithms for RL with general function approximation under the setting of bounded eluder dimension. Miryoosefi and Jin (2021) investigates the problem of reward-free RL with constraints. Wu et al. (2021) then proposes a reward-free algorithm for the multi-objective RL problem. In addition, Bai and Jin (2020); Liu et al. (2021); Qiu et al. (2021) further study the reward-free RL algorithms under the multi-agent setting.

Furthermore, we would like to emphasize that directly extending the existing results on reward-free exploration (see, e.g., Wang et al. (2020); Qiu et al. (2021)) to learning the dynamic VCG mechanism seems infeasible. The main reason is that these works focus only on estimating the optimal value functions corresponding to different reward functions. In contrast, in the context of mechanism design, we have multiple desiderata, namely truthfulness, individual rationality, and efficiency, which mathematically translates into the various regret notions, such as the welfare regret and individual regret of the seller and the buyer. Showing the proposed algorithm approximately satisfies these desiderata requires bounding these regrets using the properties of the dynamic VCG mechanism as well as the results of reward-free exploration. Finally, the recent work Lyu et al. (2022) focuses on learning the Markov VCG mechanism via offline RL from a set of collected trajectories. Under the offline setting, exploration is out of the scope, and thus our core challenge caused by the fictitious policy is absent in Lyu et al. (2022).

2. Problem Setup

Consider an episodic MDP defined by $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$, where $\mathcal{S}$ and $\mathcal{A}$ are state and action spaces, H the length of each episode, $\mathcal{P} = \{\mathcal{P}_h\}_{h=1}^{H}$ the transition kernel, and $r = \{r_{i,h}\}_{i=0,h=1}^{n,H}$ the reward functions. We use $r_{0,h} : \mathcal{S} \times \mathcal{A} \mapsto [0, R_{\max}]$ to denote the reward function of the seller at the step h and let $r_{i,h} : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ be the reward function of agent (buyer) i at the step h for $i \in [n]$, where n is the number of agents and $[n]$ denotes $\{1, 2, \cdots, n\}$. In addition, we assume that the reward observation is stochastic and the underlying reward function is the expectation of its stochastic observation, i.e., the reward observation at $(s, a) \in \mathcal{S} \times \mathcal{A}$ can be represented by $r_{i,h}(s, a; \omega)$ with $r_{i,h}(s, a) = \mathbb{E}_{\omega}[r_{i,h}(s, a; \omega)]$, where ω is an independent random variable indicating the exogenous randomness for the reward observation. We further assume that the boundedness holds for the reward observation as $r_{0,h}(\cdot, \cdot; \omega) : \mathcal{S} \times \mathcal{A} \mapsto [0, R_{\max}]$ and $r_{i,h}(\cdot, \cdot; \omega) : \mathcal{S} \times \mathcal{A} \mapsto [0, 1], \forall i \in [n]$ at all steps $h \in [H]$, where rewards for the seller and agents may have different scales[1].

Let $\pi = \{\pi_h\}_{h=1}^{H}$ denote the seller's policy, where for each $h \in [H]$, $\pi_h : \mathcal{S} \mapsto \mathcal{A}$ maps a given state to an action. For each step $h \in [H]$, reward function $r = \{r_h\}_{h=1}^{H}$, and a given policy π, we define the *value function* $V_h^{\pi}(\cdot; r) : \mathcal{S} \mapsto \mathbb{R}$ for all $x \in \mathcal{S}$ as $V_h^{\pi}(x; r) := \sum_{h'=h}^{H} \mathbb{E}\left[r_{h'}(x_{h'}, \pi_{h'}(x_{h'}))|x_h = x\right]$, where the expectation is taken over states $x_{h+1} \sim \mathcal{P}_h(\cdot|x_h, \pi_h(x_h)), x_{h+2} \sim \mathcal{P}_h(\cdot|x_{h+1}, \pi_{h+1}(x_{h+1})), \ldots, x_H \sim \mathcal{P}_H(\cdot|x_H, \pi_H(x_H))$ conditioned on a starting state $x_h = x$ at step h. Here we write $V_h^{\pi}(\cdot; r)$ to highlight that the definition of the value function depends on a given reward function r. We also define the corresponding Q-function $Q_h^{\pi}(\cdot, \cdot; r) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$ as $Q_h^{\pi}(x, a; r) := r_h(x, a) + \sum_{h'=h+1}^{H} \mathbb{E}\left[r_{h'}(x_{h'}, \pi_{h'}(x_{h'}))\middle|(x_h, a_h) = (x, a)\right]$, where the expectation is also taken over states $x_{h+1}, \ldots, x_H$ sampled from the transition model $\mathcal{P}$, conditioned on a starting state-action pair $(x_h, a_h) = (x, a)$ at step h.

We stress that while the MDP we consider contains multiple reward functions and interaction between multiple agents, our setting differs from the Markov game setting, as we assume that the seller is the only participant who can take actions (Littman, 1994).

Dynamic Mechanism Design. We now describe how agents interact with the mechanism designer (seller) in our setting. At the beginning of each episode, the mechanism starts from the initial state x_1. At each step $h \in [H]$, the seller observes some state $x_h \in \mathcal{S}$, picks an action $a_h \in \mathcal{A}$, and receives a stochastic reward with mean $r_{0,h}(x_h, a_h)$. Each agent (buyer) receives their own reward, each with an expected value of $r_{i,h}(x_h, a_h)$, and reports a stochastic reward with a mean $\widetilde{r}_{i,h}(x_h, a_h)$, given by some potentially untruthful reward function $\widetilde{r}_{i,h}(\cdot, \cdot)$. At the end of each episode, the seller charges each customer some price p_i.

[1] We allow different reward scales for greater flexibility within our framework.

For any policy π and prices $\{p_i\}_{i=1}^n$, we define agent i's utility function as

$$u_i := \mathbb{E}\bigg[\sum_{h=1}^{H} r_{i,h}(x_h, a_h)\bigg] - p_i = V_1^\pi(x_1; r_i) - p_i. \tag{1}$$

That is, agent i's utility equals the difference between the expected total reward and the charged price. The seller's utility is then defined as

$$u_0 := V_1^\pi(x_1; r_0) + \sum_{i=1}^{n} p_i. \tag{2}$$

The social welfare, W^π, is defined as the sum of the agents and the seller's utilities, given by

$$W^\pi(x_1) = \sum_{i=0}^{n} V_1^\pi(x_1; r_i) = V^\pi\bigg(x_1; \sum_{i=0}^{n} r_i\bigg), \tag{3}$$

which is equivalent to the expectation of the sum of all rewards as the prices cancel out.

Markov VCG Mechanism. Suppose that the transition kernel is known, all agents and the seller know their own reward functions $r_{i,h}$ for all $(i,h) \in [n] \times [H]$, and the agents' reward functions are known by the seller. The VCG mechanism demands that we choose the welfare-maximizing policy π_* that the seller executes each episode. Each agent i is subsequently charged a price p_{i*}, which is the loss her presence causes to others. Hence we have the following mechanism:

$$\begin{aligned} &\pi_* := \operatorname*{argmax}_\pi V_1^\pi(x_1; R), \quad \pi_*^{-i} := \operatorname*{argmax}_\pi V_1^\pi(x_1; R^{-i}), \\ &p_{i*} := V_1^{\pi_*^{-i}}(x_1; R^{-i}) - V_1^{\pi_*}(x_1; R^{-i}), \end{aligned} \tag{4}$$

where we define the total reward function R and the sum of reward except agent i, R^{-i}, as

$$R = \sum_{i=0}^{n} r_i \qquad \text{and} \qquad R^{-i} = \sum_{j=0, j\neq i}^{n} r_j.$$

Here π_* is the welfare-maximizing policy, i.e., the optimal policy for the reward function R, while π_*^{-i} is the fictitious policy that maximizes welfare when agent i is absent. Estimating the latter and their corresponding value functions requires the algorithm to explore in directions not aligned with the social welfare maximizing policy, π_*, thus necessitating the reward-free component of our algorithm. These prices, namely p_{i*}, can be estimated by following Equation (4) once the value functions corresponding to policies π_*, π_*^{-i} and reward functions R, R^{-i} are estimated sufficiently well via our algorithm. As these value functions are deterministic, the resulting pricing function is also deterministic.

The following lemma introduces the properties of the Markov VCG mechanism.

Lemma 2.1 *The Markov VCG mechanism satisfies the following desiderata in mechanism design:*

1. *Truthfulness: A mechanism is truthful if the utility u_i of agent i is maximized when, regardless of other agents' reported rewards, agent i reports her rewards truthfully.*
2. *Individual rationality: A mechanism is individually rational if the utility u_i of agent i is non-negative when agent i is truthful.*
3. *Efficiency: A mechanism is efficient if the mechanism maximizes the welfare when all agents are truthful.*

An agent is truthful if she submits her reward functions truthfully.

Please see Appendix B for the proof. Our proposed pricing formula $p_{i*} := V_1^{\pi_*^{-i}}(x_1; R^{-i}) - V_1^{\pi_*}(x_1; R^{-i})$ is not the only pricing rule that ensures Lemma 2.1. Nevertheless, our proposed algorithm can be generalized to any pricing rule of the form $p_i' = V_1^{\pi^{-i}}(x_1; R^{-i}) - V_1^{\pi_*}(x_1; R^{-i})$, where π^{-i} is not necessarily the π_*^{-i} defined above, but can be any arbitrary policy independent of agent i. Intuitively, as our algorithm makes use of reward-free exploration, we can sufficiently accurately estimate the value functions for arbitrary policies, including both π^{-i} and π_*^{-i}. Consequently, our approach can be extended to a general class of pricing functions that use different policies' value functions as prices.

Mechanism Design with an Unknown MDP. Consider the setting where the agents' value functions and the MDP's transition kernel are unknown, and the procedure is repeated for multiple rounds. At round t, the mechanism choose a policy π^t and set prices $\{p_{it}\}_{i=1}^n$ for the agents. Following Equations (1) and (2), the utilities of agent i and the seller at round t are

$$u_{it} = V_1^{\pi^t}(x_1; r_i) - p_{it} \qquad \text{and} \qquad u_{0t} = V_1^{\pi^t}(x_1; r_0) + \sum_{i=1}^n p_{it}.$$

We then denote their summations over T rounds as

$$U_{iT} = \sum_{t=1}^T u_{it} \qquad \text{and} \qquad U_{0T} = \sum_{t=1}^T u_{0t}.$$

Our goal is to design an algorithm that respects the three mechanism design desiderata over multiple rounds even when the true reward functions and transition kernels are unknown, as well as achieving sublinear regret for the agents, the seller, and the welfare. The following metrics are used to quantify the algorithm's performance:

$$\begin{aligned} &\mathrm{Reg}_T^W = TV_1^{\pi_*}(x_1; R) - \sum_{t=1}^T V_1^{\pi^t}(x_1; R) \\ &\mathrm{Reg}_{0T} = Tu_{0*} - U_{0T}, \qquad \mathrm{Reg}_{iT} = Tu_{i*} - U_{iT}, \qquad \mathrm{Reg}_T^\sharp = \sum_{i=1}^n \mathrm{Reg}_{iT}. \end{aligned} \tag{5}$$

Here we let $u_{0*} = V_1^{\pi_*}(x_1; r_0) + \sum_{i=1}^n p_{i*}$ and $u_{i*} = V_1^{\pi_*}(x_1; r_i) - p_{i*}$ be the utilities of the seller and agent i respectively in the VCG mechanism. Moreover, Reg_T^W is the welfare regret over T rounds, Reg_{0T} the seller regret, and Reg_{iT} the agent i's regret, respectively. We let $\mathrm{Reg}_T^\sharp$ be the summation of regrets over all agents.

Although the Markov VCG mechanism that we learn is welfare-maximizing, we focus on how this mechanism can be recovered. Consequently, the learning algorithm's objective is not welfare maximization alone. Maximizing welfare increases the total utility by definition and, therefore, increases the total utility that the agents and the seller share. As our learning process involves the seller and multiple agents, we also need to ensure that it faithfully respects their utilities over T rounds of interaction. Otherwise, it may be unfair to either the agents or the seller. Therefore, we measure the performance of our learning algorithm through the three terms, $\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}$, rather than any single objective by itself. We note that all three regrets are 0 under the Markov VCG mechanism.

Due to our need to approximate the VCG price p_{i*}, the welfare regret Reg_T^W differs in scale from both $\mathrm{Reg}_T^\sharp$ and Reg_{0T}, whereas the latter two are of the same scale. Notice that estimating p_{i*} involves estimating the maximum welfare that the remaining $n-1$ agents achieve when agent i is absent and the welfare that these agents receive under $\widehat{\pi}^t$. Thus, the estimation error for p_{i*} is roughly in the same order as the instantaneous welfare regret $V_1^{\pi_*}(x_1; R) - V_1^{\widehat{\pi}^t}(x_1; R)$ at round t, since both require good estimates of the summation of the value functions over all agents rather than a single agent. Consequently, recalling $\mathrm{Reg}_T^\sharp$ is the summation of all agents' regrets and Reg_{0T} equals the summation of the price estimation error across all n agents, the terms $\mathrm{Reg}_T^\sharp$ and Reg_{0T} are in fact in the order of n times the welfare regret Reg_T^W. Therefore, we add a scaling factor n in front of the welfare regret, and our learning algorithms focus on minimizing

$$\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}.$$

In addition to attaining small regret bounds, we aim to approximately satisfy the desiderata in Lemma 2.1 for the mechanism design. We define the approximate versions of truthfulness, individual rationality, and efficiency concerning the agent's *cumulative* utility U_{iT} as follows:

1. *Approximate truthfulness*: Let U_{iT} be the cumulative utility when agent i is truthful and $\widetilde{U}_{iT}$ that when agent i is untruthful. The mechanism is *δ-approximately truthful* if $\widetilde{U}_{iT} - U_{iT} \le \delta$, regardless of others' truthfulness.
2. *Approximate individual rationality*: When agent i reports truthfully, the mechanism is *δ-approximately individually rational* if $U_{it} \ge -\delta$, regardless of others' truthfulness.
3. *Approximate efficiency*: The mechanism is *δ-approximately efficient* if $\mathrm{Reg}_T^W \le \delta$ when all agents are truthful.

When an agent adopts an untruthful reward-reporting strategy, it means that this agent reports her rewards under a different reward function $\widetilde{r}_{ih}$ rather than the true reward function r_{ih}. As the algorithm interacts with the environment over T rounds, these approximate desiderata can have a dependence on T. Our definition generalizes the asymptotic versions of the desiderata defined in Kandasamy et al. (2020) since the approximate desiderata naturally imply their asymptotic counterparts when δ is sublinear in T. More specifically, as long as $\lim_{T\to\infty} f(T)/T = 0$, if a mechanism is $f(T)$-approximate truthful, when amortized over these T rounds of interaction, agents' utility gain from untruthful reports vanishes. In other words, in the long run, agents cannot improve upon their average per-episode utility by untruthfulness, thus deterring rational agents from attempting to alter the learning process via untruthfulness. Similarly, if $f(T)$ is sublinear and the mechanism is $f(T)$-approximately individually rational, then in the long run, agents' average episodic utility is lower-bounded by a number tending to zero (i.e., $\lim_{T\to\infty} \frac{1}{T}U_{iT} \geq -\lim_{T\to\infty} f(T)/T = 0$), ensuring they will not be worse-off from participating.

Since approximate truthfulness implies, for suitable $f(T)$, that agents will not benefit from untruthful reporting in the long run, our definition of approximate efficiency focuses only on truthful agents. Indeed, consider the extreme case where all agents report $1 - r_{i,h}(x, a)$ instead of $r_{i,h}(x, a)$ and the seller reward is always 0. Under this extreme case of untruthful behavior, the welfare-maximizing policy under the untruthful report is in fact the welfare-minimizing policy under truthful reports, showing that it is in general hard to obtain efficiency guarantees without assuming truthful behavior. Such an approach, namely, first showing that the mechanism is approximately truthful and then providing guarantees under the assumption that the reports are truthful, is common in existing literature at the intersection of mechanism design and learning (Nazerzadeh et al., 2008; Kandasamy et al., 2020). We refer interested readers to Epasto et al. (2018), which justifies in further detail why agents will behave truthfully under approximately truthful mechanisms.

To handle the potentially large state and action spaces $\mathcal{S}, \mathcal{A}$, our work focuses on the linear function approximation setting, where the linear MDP is considered.

Linear MDP. We assume that there exist a feature map $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, d unknown measures $\boldsymbol{\mu}_h = (\mu_h^1, \cdots, \mu_h^d)$ over $\mathcal{S}$ for any $h \in [H]$, and $n + 1$ unknown vectors $\{\boldsymbol{\theta}_{ih}\}_{i=0}^n$ with each $\boldsymbol{\theta}_{ih} \in \mathbb{R}^d$ for all $h \in [H]$. For any $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, the transition kernel and reward function can be linearly represented as

$$\begin{aligned} &\mathcal{P}_h(x'|x, a) = \langle \phi(x, a), \boldsymbol{\mu}_h(x') \rangle \\ &r_{i,h}(x, a) = \langle \phi(x, a), \boldsymbol{\theta}_{ih} \rangle, \quad \forall i = 0, 1, \cdots, n. \end{aligned} \tag{6}$$

Following standard assumptions in the prior literature (Jin et al., 2020b,c), we assume $\|\phi(x, a)\| \leq 1$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$, $\max\{\|\boldsymbol{\mu}_h(\mathcal{S})\|, \|\boldsymbol{\theta}_{ih}\|\} \leq \sqrt{d}$ for all $h \in [H], 0 \leq i \leq n$. Recall that the linear MDP assumption implies that the value functions and action-value functions are both linear in the feature space defined by ϕ (Jin et al., 2020b). When the problem reduces to the tabular setting, we have $d = |\mathcal{S}||\mathcal{A}|$ with $\phi(x, a) = \mathbf{e}_{x,a} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ being an indicator vector.

Remark 2.2 *When linear function approximation is considered, a typical assumption is that the underlying MDP has a linear structure. Here we assume the MDP satisfies Equation* (6)*. As discussed above, the tabular MDP can be covered as a special case of the linear MDP. Thus, our method for the linear MDP can also solve problems modeled by the tabular MDP. In realistic and complex scenarios, the underlying MDP may not be strictly linear. One can still apply the linear function approximation along with introducing a misspecification error. This error can be characterized by* $\sup_{x,a} \|\mathcal{P}_h(\cdot|x, a) - \langle \phi(x, a), \boldsymbol{\mu}_h(\cdot)\|_{\mathrm{TV}} \leq \mathcal{E}_{\mathcal{P}}$ *and* $\sup_{i,x,a} \|r_{i,h}(x, a) - \langle \phi(x, a), \boldsymbol{\theta}_{ih}\rangle\|_{\mathrm{TV}} \leq \mathcal{E}_r$ *as commonly discussed in prior RL literature (e.g., Jin et al. (2019)), where* $\|\cdot\|_{\mathrm{TV}}$ *denotes the total variation. By making small changes to our current analysis, extra misspecification terms containing* $\mathcal{E}_{\mathcal{P}}$ *and* $\mathcal{E}_r$ *will be added to our regret bounds. If both* $\mathcal{E}_{\mathcal{P}}$ *and* $\mathcal{E}_r$ *are small, the underlying MDP is approximately linear such that the extra terms can be considered minor.*

2.1 Motivating Examples

We provide several motivating examples for the dynamic mechanism design introduced above, which are the potential application areas for our proposed algorithm.

Dynamic Sponsored Search Auction. We assume the state x includes information on the agents' remaining budgets for the episode. Let H be a fiscal year. As advertisements' values change within a single year (e.g., value increases around Black Friday), agents' rewards from advertising naturally change with time. The seller's action would affect the agents' budgets, which would further affect their valuations: an agent who did not win any auction in previous rounds would have a high remaining budget near the end of the year and, therefore, would be willing to pay more for each advertisement slot in an effort to increase their odds of winning.

Dynamic Platform-as-a-Service (PaaS). We assume there are multiple users using the same computing cluster and a central planner who allocates computation resources to these users. The state x includes information on the server's current load, and action a reflects how the central planner allocates these resources among users. Naturally, the planner's action affects the server load in the next state. While a higher server load would provide users with immediate satisfaction, it would also incur higher electricity costs for the planner. As the users' demands may fluctuate within a day (for instance, demands are lower during the night), it is a significant challenge for the planner to balance electricity costs and user satisfaction in an environment with the users' valuations and demands constantly changing. The problem is further complicated by the fact that the service provider only learns user satisfaction after the resources are allocated, justifying our setup above.

Dynamic Public Service. This example is inspired by Section 9.3.5.5 in Nisan et al. (2007). Here the seller takes the form of a government body, and the agents are the citizens. The seller wishes to provide public services to benefit the general population, and the agents pay the seller in the form of taxation. The state x contains information on the seller's remaining budget for the year as well as the agents' satisfaction with the seller. When the seller does not provide sufficient public service, agents will become less satisfied and have more urgent demands for public services in later steps, exhibiting natural transition dynamics. As the seller can only learn the agents' valuation after the service has been provided, the problem fits naturally within the setting considered above.

Relationship to Parkes and Singh (2003). Finally, our work could address several key problems raised by prior works on dynamic mechanism design without assuming prior knowledge of the underlying model. Parkes and Singh (2003) studies an online mechanism design problem by formulating the problem as an MDP and proposes Wi-Fi pricing at Starbucks as a motivating example. Parkes and Singh (2003) assumes that the welfare-maximizing policy is known a priori. However, the MDP in Parkes and Singh (2003) is an infinite-horizon, un-discounted, and non-average reward one, and we are not aware of any existing literature that can provably learn nearly optimal policies in this setting. We thus leave the question as a future direction of independent interest. Nevertheless, our work takes a first step towards relaxing the assumption by requiring the mechanism designer to recover the policy from repeated interaction in the finite horizon case.

3. Algorithm

In this section, we introduce our proposed algorithm for VCG mechanism learning on linear MDPs (`VCG-LinMDP`). The general learning framework of our algorithm is summarized in Algorithm 1, comprising two phases: the exploration phase and the exploitation phase. The exploration and exploitation phases are summarized in Algorithms 2, 3, and 4.

3.1 Algorithmic Framework

Markov VCG with Function Approximation. In order to learn the Markov VCG mechanism, we consider a learning framework with function approximation, in which the reward-free exploration phase aims to efficiently explore the environment with wide coverage over the underlying policy space. The exploitation phase targets at utilizing the collected data to update the seller's policy and estimate the prices charged to the agents. We remark that this learning framework is general and can fit *any linear or nonlinear* function approximators. We summarize it as follows:

1. Exploration for multiple rounds to collect an initial dataset. The exploration is performed via a reward-free least-square value iteration (LSVI) with function approximation (Jin et al., 2020a; Wang et al., 2020; Qiu et al., 2021).
2. Exploitation with the collected data. At each round t of the exploitation phase:
   - Update the seller's policy $\widehat{\pi}^t$ via a planning subroutine implemented as optimistic LSVI with function approximation w.r.t. the reward function R.
   - Update F_t^{-i} by the value function from a planning subroutine implemented as optimistic or pessimistic LSVI with function approximation w.r.t. R^{-i}.
   - Update G_t^{-i} by the value function from a policy evaluation subroutine by optimistic or pessimistic evaluation with function approximation at the learned policy $\widehat{\pi}^t$ w.r.t. R^{-i}.
   - Estimate the price $p_{it} = F_t^{-i} - G_t^{-i}$ for all $i \in [n]$.
   - Take actions following $\widehat{\pi}^t$ and charge each agent i a price p_{it} for $i \in [n]$.
   - Determine whether we should update the dataset with the new trajectory.

Here $\widehat{\pi}^t$ is the learned policy aiming to estimate π_*, the function F_t^{-i} can be viewed as an estimate of the value function under the fictitious policy, i.e., $V_1^{\pi_*^{-i}}(x_1; R^{-i})$, and G_t^{-i} estimates $V_1^{\widehat{\pi}^t}(x_1; R^{-i})$ under the policy $\widehat{\pi}^t$. In particular, the hyperparameters ζ_2, ζ_3 control whether such an estimation by F_t^{-i} and G_t^{-i} is optimistic or pessimistic. Moreover, since $\widehat{\pi}^t$ estimates π_*, then G_t^{-i} can further be considered as an approximation of $V_1^{\pi_*}(x_1; R^{-i})$, which implies that the price p_{i*} is estimated by p_{it} according to its definition. At a higher level, the algorithm decomposes learning the Markov VCG mechanism into two parts: 1) learning an efficient, social welfare-maximizing policy, and 2) estimating the suitable prices to charge the agents.

This paper focuses on a special case, i.e., Markov VCG with linear function approximation named `VCG-LinMDP`, as shown in Algorithm 1. The associated exploration phase is implemented in Algorithm 2, and the exploitation phase is implemented in Algorithms 3 and 4, where we adopt LSVI with linear function approximation. In particular, Algorithms 3 and 4 are the planning and policy evaluation subroutines respectively. As we can see from

Algorithm 1 `VCG-LinMDP`

```
Input: $\zeta_1 \in \{\texttt{ETC}, \texttt{EWC}\}$, $\zeta_2, \zeta_3 \in \{\texttt{OPT}, \texttt{PES}\}$, $\mathfrak{R} \in \{R, R^{-i}\}$, and $K$.
    //Exploration Phase
Reward-free exploration for $K$ rounds via Algorithm 2 and obtain $\mathcal{D} = \{(x_h^k, a_h^k)\}_{h,k} \cup \{r_{i,h}^k(x_h^k, a_h^k)\}_{i,h,k}$.
    //Exploitation Phase
for $t = K+1, \cdots, T$ do
    Update policy $\widehat{\pi}^t$ by the returned policy of Algorithm 3 with input parameters $(R, \zeta_1, \texttt{OPT}, \mathcal{D})$.
    Update $F_t^{-i}$ by the returned value function of Algorithm 3 with parameters $(R^{-i}, \zeta_1, \zeta_2, \mathcal{D})$ for all $i \in [n]$.
    Update $G_t^{-i}$ by the returned value function of Algorithm 4 with parameters $(R^{-i}, \zeta_1, \zeta_3, \mathcal{D}, \widehat{\pi}^t)$ for all $i \in [n]$.
    Calculate the price $p_{it} = F_t^{-i} - G_t^{-i}$ for all $i \in [n]$.
    Take action $a_h^t = \widehat{\pi}_h^t(x_h^t)$, receive rewards $\{r_{i,h}^t(x_h^t, a_h^t)\}_i$, and observe $x_{h+1}^t \sim \mathcal{P}_h(\cdot|x_h^t, a_h^t)$ from $h = 1$ to $H$.
    Charge each agent $i$ a price $p_{it}$ for all $i \in [n]$.
    if $\zeta_1 = \texttt{EWC}$ then
        $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x_h^t, a_h^t)\}_{t,h} \cup \{r_{i,h}^t(x_h^t, a_h^t)\}_{i,h,t}$
    else if $\zeta_1 = \texttt{ETC}$ then
        Keep $\mathcal{D}$ unchanged as collected in the exploration phase.
    end if
end for
```

the overall framework, learning the price requires both planning to learn a fictitious policy (the function F_t^{-1}) and function evaluation on the learned policy G_t^{-i} in order to estimate the price, necessitating the inclusion of both Algorithm 3 and Algorithm 4.

As shown in Algorithm 1, there are multiple hyper-parameters. Specifically, ζ_1 controls the overall learning strategy of `VCG-LinMDP` with options `ETC` and `EWC`. The option `ETC` indicates the *explore-then-commit* strategy, where we exploit using only the data generated during the exploration phase. `EWC` indicates *explore-while-commit* strategy, where we exploit using data generated during both the exploration phase and the exploitation phase. The options `OPT` and `PES` for the hyper-parameters ζ_2 and ζ_3 refer to optimistic and pessimistic exploitation approaches respectively, which control the trade-off between the seller's and the agents' utilities. Finally, for Algorithms 3 and 4, the hyper-parameter $\mathfrak{R}$ controls whether the input reward function is R or R^{-i}. In these algorithms, for abbreviation, we denote by $r_{i,h}^k(s_h^k, a_h^k) := r_{i,h}(s_h^k, a_h^k; \omega_h^k)$ the reward collected at step h of time k in the exploration phase and by $r_{i,h}^t(s_h^t, a_h^t) := r_{i,h}(s_h^t, a_h^t; \omega_h^t)$ a reward collected at step h of time t in the exploitation phase, where ω_h^k and ω_h^t represent the randomness in the reward observation.

Remark 3.1 *We remark that in our proposed algorithms in Section 3, with a slight abuse of notation, we do not require the reports of the rewards to be truthful when setting* $\mathfrak{R} = R$

or $\mathfrak{R} = R^{-i}$. One can think of R and R^{-i} as input arguments if no specific discussion on truthfulness is involved. The rewards in the algorithms can be either truthful or untruthful. Whether the rewards are needed to be truthful or not will be explicitly highlighted in our theoretical results and the associated proofs.

Remark 3.2 *Intuitively, the hyperparameters ζ_2 and ζ_3 control whether the price favors the sellers or the buyers. There are two extreme cases for the setting of (ζ_2, ζ_3), namely $(\texttt{PES}, \texttt{OPT})$ and $(\texttt{OPT}, \texttt{PES})$. The configuration $(\zeta_2, \zeta_3) = (\texttt{PES}, \texttt{OPT})$ that favors agents potentially leads to a low price p_{it} and high agent utilities, resulting in a low agent regret and a high seller regret. The configuration $(\zeta_2, \zeta_3) = (\texttt{OPT}, \texttt{PES})$ will favor the seller with a high price p_{it} and a high seller utility, which results in a high agent regret and low seller regret. The prices charged under other configurations would fall somewhere between the aforementioned high and low prices. Consequently, the agents' and the seller's regrets would naturally be somewhere in the middle between the two representative cases, which we will expand in depth in our theoretical results. Such flexibility can be crucial in practice. For instance, the seller in the dynamic sponsored search auction or the dynamic PaaS setting discussed in Section 2.1 favors a high price obtained by setting $\zeta_2 = \texttt{OPT}, \zeta_3 = \texttt{PES}$, while the social good provider in the dynamic public service setting may prefer a lower price when we set $\zeta_2 = \texttt{PES}, \zeta_3 = \texttt{OPT}$.*

Least-Square Value Iteration. With the overarching framework defined, we now introduce a key technique heavily used by our algorithm. For any function approximation class $\mathcal{F}$, at the t-th episode, we have $t-1$ transition tuples, $\{(x_h^\tau, a_h^\tau, x_{h+1}^\tau)\}_{\tau\in[t-1]}$, and LSVI with function approximation (Jin et al., 2020b; Yang et al., 2020b; Jin et al., 2020c) estimates the Q-function using $\widetilde{f}_h^t$, obtained from the least-squares regression problem below.

$$\begin{aligned}
\widetilde{f}_h^t &= \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{\tau=1}^{t-1} \big[r_h^\tau(x_h^\tau, a_h^\tau) + V_h^t(x_h^\tau) - f_h(x_h^\tau, a_h^\tau))\big]^2 + \text{pen}(f), \\
f_h^t &= \text{truncate}\{\widetilde{f}_h^t\},
\end{aligned}$$

where $\text{pen}(f)$ is some arbitrary regularizer, r_h is some reward function, $\text{truncate}\{\cdot\}$ is some truncation operator to guarantee that the approximation function is in a correct scale such that it does not violate the boundedness assumptions we place on the Q-function. For optimistic LSVI, we construct *optimistic* Q-function as

$$Q_h^t = \text{truncate}\{f_h^t + u_h^t\},$$

where we again truncate the estimated Q-function, and u_h^t is an associated UCB bonus term constructed using the collected trajectories. Similarly, the *pessimistic* Q-function is constructed as

$$Q_h^t = \text{truncate}\{f_h^t - u_h^t\}.$$

We update the value function by a greedy strategy as

$$V_h^t(\cdot) = \operatorname*{argmax}_{a\in\mathcal{A}} Q_h^t(\cdot, a),$$

Algorithm 2 `Exploration`

```
Input: Failure probability $\delta > 0$, $K$, and $\lambda > 0$
 1: $\beta = \hat{c}(n + R_{\max})dH\sqrt{\log(36ndHT/\delta)}$.
 2: for $k = 1, 2 \cdots, K$ do
 3:     Set $V_{H+1}^k(\cdot) = 0$.
 4:     for $h = H, H-1 \cdots, 1$ do
 5:         $\Lambda_h^k = \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + \lambda I$.
 6:         $u_h^k(\cdot,\cdot) = \Pi_{[0,H(n+R_{\max})]}\big[\beta[\phi(\cdot,\cdot)(\Lambda_h^k)^{-1}\phi(\cdot,\cdot)]^{1/2}\big]$.
 7:         Define an exploration-driven reward function $l_h^k(\cdot,\cdot) = u_h^k(\cdot,\cdot)/H$.
 8:         $w_h^k = (\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi(x_h^\tau, a_h^\tau)V_{h+1}^k(x_{h+1}^\tau)$.
 9:         $Q_h^k(\cdot,\cdot) = \min\{\Pi_{[0,H(n+R_{\max})]}[(w_h^k)^\top\phi(\cdot,\cdot)] + l_h^k(\cdot,\cdot) + u_h^k(\cdot,\cdot), H(n+R_{\max})\}$.
10:         $V_h^k(\cdot) = \max_{a\in\mathcal{A}} Q_h^k(\cdot, a)$.
11:         $\pi_h^k(\cdot) = \mathrm{argmax}_{a\in\mathcal{A}} Q_h^k(\cdot, a)$.
12:     end for
13:     Take action $a_h^k = \pi_h^k(x_h^k)$, receive rewards $\{r_{i,h}^k(x_h^k, a_h^k)\}_i$, and observe the state transition $x_{h+1}^k \sim \mathcal{P}_h(\cdot|x_h^k, a_h^k)$ from $h = 1$ to $H$.
14: end for
15: return $\mathcal{D} = \{(x_h^k, a_h^k)\}_{(h,k)\in[H]\times[K]} \cup \{r_{i,h}^k(x_h^k, a_h^k)\}_{(i,h,k)\in(\{0\}\cup[n])\times[H]\times[K]}$
```

for optimistic Q-function or pessimistic Q-function respectively. For the linear function approximation in our algorithm, according to our setting of linear MDPs, we let $f(\cdot,\cdot) = w^\top\phi(\cdot,\cdot)$ for any $f \in \mathcal{F}$ and $\mathrm{pen}(f)$ be $\lambda\|w\|^2$ where w is the parameter to learn.

With the key ideas sketched out, we then proceed with fleshing out the proposed algorithms.

3.2 Exploration Phase

Our first component is the exploration phase. Recall that F_t^{-i} estimates the value function of the fictitious policy that maximizes welfare when agent i is absent. Obtaining high-quality F_t^{-i} for all n agents then requires the algorithm to explore in the direction of multiple policies rather than only in a single policy's direction. This challenge necessitates reward-free reinforcement learning, where the learning algorithm seeks to explore the environment in the directions of all possible policies as opposed to only a single one.

Inspired by Wang et al. (2020), we design a reward-free exploration algorithm as in Algorithm 3, incorporating the linear structure of the MDP. Specifically, to handle multiple reward functions from the seller and n agents, we propose to explore the environment without using the observed rewards from it. Instead, we define an exploration-driven reward l_h^k as a scaled bonus term u_h^k to encourage exploration by further taking into account the uncertainty of estimating the environment. The bonus term computed in Line 6 quantifies the uncertainty of estimation with a linear function approximator. Based on the exploration-driven rewards $l_h^k = u_h^k/H$ and the bonus term u_h^k as well as the linear function approximation, we calculate an optimistic Q-function and perform the optimistic reward-free LSVI to generate the exploration policy. Note that in Algorithm 2 and the subsequent Algorithms 3 and 4, we define a truncation operator $\Pi_{[0,x]}[\cdot] := \max\{\min\{\cdot, x\}, 0\}$. Distinguished from the standard

Algorithm 3 `Exploitation: Planning`

```
Input: $(\mathfrak{R}, \zeta, \zeta', \mathcal{D}, t)$.
$V_{H+1}^t(\cdot; \mathfrak{R}) = 0$.
for $h = H, H-1 \cdots, 1$ do
    $Q_h^t(\cdot, \cdot; \mathfrak{R}) = \texttt{Est-Q}(\mathfrak{R}, \zeta, \zeta', \mathcal{D}, h, t)$
    $\pi_h^t(\cdot) = \mathrm{argmax}_{a \in \mathcal{A}} Q_h^t(\cdot, a; \mathfrak{R})$.
    $V_h^t(\cdot; \mathfrak{R}) = Q_h^t(\cdot, \pi_h^t(\cdot); \mathfrak{R})$.
end for
return $\{\pi_h^t\}_{h=1}^H$, $V_1^t(x_1; \mathfrak{R})$
```

Algorithm 4 `Exploitation: PolicyEval`

```
Input: $(\mathfrak{R}, \zeta, \zeta', \mathcal{D}, t, \pi)$.
$V_{H+1}^t(\cdot; \mathfrak{R}) = 0$.
for $h = H, H-1 \cdots, 1$ do
    $Q_h^t(\cdot, \cdot; \mathfrak{R}) = \texttt{Est-Q}(\mathfrak{R}, \zeta, \zeta', \mathcal{D}, h, t)$
    $V_h^t(\cdot; \mathfrak{R}) = Q_h^t(\cdot, \pi_h(\cdot); \mathfrak{R})$.
end for
return $V_1^t(x_1; \mathfrak{R})$
```

Algorithm 5 `Est-Q`: One-Step Optimistic/Pessimistic Estimation of Q-Function

```
Input: $(\mathfrak{R}, \zeta, \zeta', \mathcal{D}, h, t)$.
Set $\alpha_h(\mathfrak{R})$ as (7) and $\beta = \hat{c}(n + R_{\max}) dH \sqrt{\log(36ndHT/\delta)}$.
$P_t := \begin{cases} \{1, 2, \cdots, K\} & \text{if } \zeta = \texttt{ETC} \\ \{1, 2, \cdots, t-1\} & \text{if } \zeta = \texttt{EWC}. \end{cases}$
$\Lambda_h^t = \sum_{\tau \in P_t} \phi(x_h^\tau, a_h^\tau) \phi(x_h^\tau, a_h^\tau)^\top + \lambda I$.
$u_h^t(\cdot, \cdot) = \Pi_{[0, H(n+R_{\max})]} \big[ \beta [\phi(\cdot, \cdot) (\Lambda_h^t)^{-1} \phi(\cdot, \cdot)]^{1/2} \big]$.
$w_h^t = (\Lambda_h^t)^{-1} \sum_{\tau \in P_t} \phi(x_h^\tau, a_h^\tau) [\mathfrak{R}_h^\tau(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau; \mathfrak{R})]$.
$f_h^t(\cdot, \cdot) = \Pi_{[0, H(n+R_{\max})]} [(w_h^t)^\top \phi(\cdot, \cdot)]$.
$Q_h^t(\cdot, \cdot; \mathfrak{R}) = \begin{cases} \Pi_{[0, \alpha_h(\mathfrak{R})]} [(f_h^t + u_h^t)(\cdot, \cdot)] & \text{if } \zeta' = \texttt{OPT} \\ \Pi_{[0, \alpha_h(\mathfrak{R})]} [(f_h^t - u_h^t)(\cdot, \cdot)] & \text{if } \zeta' = \texttt{PES}. \end{cases}$
return $Q_h^t(\cdot, \cdot; \mathfrak{R})$
```

LSVI introduced above, the reward-free LSVI only considers the value function as the regression target, i.e., we solve a least-square regression problem in the following form

$$\operatorname*{argmin}_{f \in \mathcal{F}_{\mathrm{lin}}} \sum_{\tau=1}^{k-1} \big[V_h^k(x_h^\tau) - f_h(x_h^\tau, a_h^\tau) \big) \big]^2 + \mathrm{pen}(f),$$

where $\mathcal{F}_{\mathrm{lin}}$ is the linear function class. Then, we obtain the coefficient vector w_h^k for linear function approximation.

Moreover, for the optimistic Q-function in Line 9, we construct it by combining not only the linear approximation function and the exploration bonus u_h^k but also the exploration-driven reward l_h^k. Meanwhile, we collect the trajectories $\mathcal{D}$ of visited state-action pairs and the corresponding reward feedbacks of $r_i, \forall i = 0, 1, \ldots, n$, for the subsequent exploitation phase in Algorithms 3 and 4.

3.3 Exploitation Phase

The exploitation phase is separated into two subroutines, namely `Planning` for planning in Algorithm 3 and `PolicyEval` for policy evaluation in Algorithm 4. The two algorithms are general subroutines that are instantiated by the inputs.

The `Planning` subroutine in Algorithm 3 is an optimistic or pessimistic LSVI with linear function approximation, which generates a greedy policy and its associated value function. Different from Algorithm 3, `PolicyEval` subroutine in Algorithm 4 only evaluates any input policy π by computing the value function under π with linear function approximation. Both of the two algorithms will call Algorithm 5, which is an optimistic or pessimistic estimation of the Q-function for a reward function $\mathfrak{R} \in \{R, R^{-i}\}$ at step h. Algorithm 5 can be viewed as an instantiation of LSVI in Section 3.1 for linear function approximation. In Line 4 of Algorithm 5, we compute a bonus u_h^t to quantify the uncertainty in estimation. In Lines 5 and 6, we obtain the coefficient vector w_h^t for linear function approximation and the approximator f_h^t. Line 7 yields optimistic and pessimistic Q-functions respectively determined by $\zeta' = \texttt{OPT}$ or $\texttt{PES}$.

The argument ζ in these algorithms determines the composition of the data index set P_t in Line 2 and thus indicates whether we will use the original exploration dataset or the updated dataset to construct the bonus term u_h^t and the linear function approximator f_h^t. More formally, only the data collected in the exploration phase of Algorithm 1 will be used if we let $\zeta = \texttt{ETC}$, and the data generated in both exploration and exploitation phases is used when we let $\zeta = \texttt{EWC}$.

The function $\alpha_h(\mathfrak{R})$ in these algorithms controls the truncation constant, which equals the supremum of the corresponding reward function. Precisely, we have

$$\alpha_h(\mathfrak{R}) := \begin{cases} (n + R_{\max})(H - h + 1) & \text{if } \mathfrak{R} = R \\ (n - 1 + R_{\max})(H - h + 1) & \text{if } \mathfrak{R} = R^{-i} \text{ for any } i \in [n] \ . \end{cases} \tag{7}$$

Note that Algorithm 3 and Algorithm 4 are two generic subroutines for the exploitation phase, whose concrete implementation is contingent on the input arguments. For brevity, we denote all the value functions and Q-functions in Algorithm 3 and Algorithm 4 calculated in step t by $V_h^t(\cdot;\cdot)$ and $Q_h^t(\cdot,\cdot;\cdot)$ respectively. Specifically, in the rest of this work, we let $\{\widehat{V}_h^{t,*}(\cdot;\mathfrak{R}), \hat{Q}_h^{t,*}(\cdot,\cdot;\mathfrak{R})\}$ and $\{\check{V}_h^{t,*}(\cdot;\mathfrak{R}), \check{Q}_h^{t,*}(\cdot,\cdot;\mathfrak{R})\}$ be the realization of $V_h^t(\cdot;\cdot)$ and $Q_h^t(\cdot,\cdot;\cdot)$ generated by Algorithm 3 for $\zeta' = \texttt{OPT}$ and $\zeta' = \texttt{PES}$ respectively, with different options for $\mathfrak{R}$; and let $\{\widehat{V}_h^{t,\pi}(\cdot;\mathfrak{R}), \hat{Q}_h^{t,\pi}(\cdot,\cdot;\mathfrak{R})\}$ and $\{\check{V}_h^{t,\pi}(\cdot;\mathfrak{R}), \check{Q}_h^{t,\pi}(\cdot,\cdot;\mathfrak{R})\}$ be associated with $\zeta' = \texttt{OPT}$ and $\zeta' = \texttt{PES}$ respectively, which are generated by Algorithm 4 with arbitrary input policy π. In the sequel, in Algorithm 1, we have

$$F_t^{-i} = \begin{cases} \widehat{V}_1^{t,*}\big(x_1; R^{-i}\big) & \text{if } \zeta_2 = \texttt{OPT} \\ \check{V}_1^{t,*}\big(x_1; R^{-i}\big) & \text{if } \zeta_2 = \texttt{PES}, \end{cases} \qquad G_t^{-i} = \begin{cases} \widehat{V}_1^{t,\widehat{\pi}^t}\big(x_1; R^{-i}\big) & \text{if } \zeta_3 = \texttt{OPT} \\ \check{V}_1^{t,\widehat{\pi}^t}\big(x_1; R^{-i}\big) & \text{if } \zeta_3 = \texttt{PES}. \end{cases}$$

These functions then in turn estimate the price that is to be charged to the agents. The exact formulation can be found in Algorithm 1.

Our proposed algorithms have the potential of being extended to other nonlinear function approximations following the LSVI steps in Section 3.1, such as the kernel function approximation and neural function approximation built on the neural tangent kernel theory (Jacot et al., 2018). This generalization is facilitated by exploring the inherent structure of specific function classes to construct bonus terms and optimistic/pessimistic Q-functions using techniques proposed in Zhou et al. (2020a); Yang et al. (2020a); Qiu et al. (2021). Then, one can replace the function approximation steps in Algorithms 2 and 5 with the ones tailored for these approximators to apply nonlinear function approximation. Such a direction of research warrants further studies in the future.

Remark 3.3 *We emphasize that* `VCG-LinMDP` *(Algorithm 1) is not a direct extension of reward-free RL algorithms with function approximation (e.g., Jin et al. (2020a); Wang et al. (2020); Qiu et al. (2021)) which focus only on estimating the optimal value functions corresponding to different reward functions. Learning the dynamic mechanism requires achieving multiple desiderata as introduced in Section 2 and minimizing the corresponding regrets, which introduces additional challenges with decomposing the regret terms not encountered in prior literature. In particular, we adopt reward-free exploration to address a specific challenge encountered when learning the dynamic VCG mechanism, namely, the need to learn the fictitious policy, i.e., the optimal policy in the absence of each agent i, yet reward-free exploration itself cannot ensure that the resulting mechanism is truthful or individually rational. Particularly, to show that the final policy output by the exploitation phase enjoys the desired desiderata requires the particular structure of the VCG mechanism, which we exploit in our proofs. Besides, the exploitation phase (Algorithm 3 and Algorithm 4) allows for optimism and pessimism in an online setting, inducing different price estimation strategies as discussed above. Moreover, Algorithm 2 differs from standard reward-free RL algorithms by recording the received rewards of different agents during exploration and utilizing these collected rewards to learn the welfare-maximizing policy and the agents' prices.*

4. Main Results

In this section, we discuss our main theoretical results. We first state the results corresponding to the three desiderata in mechanism design when $\zeta_1 = \texttt{ETC}, \texttt{EWC}$ respectively. Then we present the lower bound of our problem. In our algorithms and theoretical results, $\hat{c}$ is a universal absolute constant. We begin with the results for when $\zeta_1 = \texttt{ETC}$, i.e., the proposed algorithms adopt the *explore-then-commit* strategy, where the exploitation phase uses only the data generated during the exploration phase.

Theorem 4.1 *When $\zeta_1 = \texttt{ETC}$, setting $K = dH^{4/3}\iota^{1/3}T^{2/3}$ where $\iota := \log(36ndHT/\delta)$ for any $\delta \in (0,1]$, defining $n_R := n + R_{\max}$, with probability at least $1-\delta$, for all $T > K$, the following results hold after executing Algorithm 1 for T rounds:*

1. *Assuming all agents report truthfully, for all $\zeta_2, \zeta_3 \in \{\texttt{OPT}, \texttt{PES}\}$, the welfare regret satisfies*
$$\mathrm{Reg}_T^W \leq (1+2\hat{c})n_R dH^{7/3}\iota^{1/3}T^{2/3},$$
which indicates that the learned mechanism is $(1+2\hat{c})n_R dH^{7/3}\iota^{1/3}T^{2/3}$-approximately efficient.
2. *Assuming all agents report truthfully, the regret of agent i satisfies*
$$\mathrm{Reg}_{iT} \leq \begin{cases} (1+2\hat{c}n_R)dH^{7/3}\iota^{1/3}T^{2/3} & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT}) \\ (1+6\hat{c}n_R)dH^{7/3}\iota^{1/3}T^{2/3} & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES}). \end{cases}$$
3. *Assuming all agents report truthfully, the regret of the seller satisfies*
$$\mathrm{Reg}_{0T} \leq \begin{cases} (1+4\hat{c}n)n_R dH^{7/3}\iota^{1/3}T^{2/3} & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT}) \\ n_R dH^{7/3}\iota^{1/3}T^{2/3} & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES}). \end{cases}$$

4. *The learned mechanism is $6\hat{c}n_R dH^{7/3}\iota^{1/3}T^{2/3}$-approximately individually rational.*
5. *The learned mechanism is $\big(1+4\hat{c}n_R\big)dH^{7/3}\iota^{1/3}T^{2/3}$-approximately truthful.*

As the learning objective of our algorithm is to minimize the welfare regret together with the agent and seller regrets, we choose $K = dH^{4/3}\iota^{1/3}T^{2/3}$ that can lead to a small upper bound of $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}$, which is $\mathcal{O}\big(n(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}\big)$. Here we ignore constant factors and emphasize K's dependence on d, H, ι, and T. As discussed in Remark 3.2, we use ζ_2 and ζ_3 to control the charged price and the seller and agent utilities, which further affect the achieved regrets. When $(\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES})$, the charged price will be large and favor the seller, which thus leads to a relatively low seller regret $(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}$ and a high agent regret $(1+6\hat{c}(n+R_{\max}))dH^{7/3}\iota^{1/3}T^{2/3}$. When $\zeta_2 = \texttt{PES}$ and $\zeta_3 = \texttt{OPT}$, there will be a lower price favoring the agent, such that the seller regret increases to $(1+4\hat{c}n)n_R dH^{7/3}\iota^{1/3}T^{2/3}$ and agent i's regret decreases to $(1+2\hat{c}(n+R_{\max}))dH^{7/3}\iota^{1/3}T^{2/3}$. The seller and agent regrets incurred by other options of (ζ_2,ζ_3) will lie between the above regret bounds under such two settings. Since the welfare does not depend on the price as shown in Equation (5), the choices of (ζ_2,ζ_3) thus have no impact on the welfare regret.

We further present the results for $\zeta_1 = \texttt{EWC}$, i.e., the algorithm adopts the *explore-while-commit* strategy, where the exploitation phase uses data collected during both the exploration and exploitation phases.

Theorem 4.2 *When $\zeta_1 = \texttt{EWC}$, setting $K = dH^{4/3}\iota^{1/3}T^{2/3}$ where $\iota := \log(36ndHT/\delta)$ for any $\delta \in (0,1]$, defining $n_R := n + R_{\max}$, with probability at least $1-\delta$, for all $T > K$, the following results hold after executing Algorithm 1 for T rounds:*

1. *Assuming all agents report truthfully, for all $\zeta_2,\zeta_3 \in \{\texttt{OPT},\texttt{PES}\}$, the welfare regret satisfies*
$$\mathrm{Reg}_T^W \le n_R dH^{7/3}\iota^{1/3}T^{2/3} + 6\hat{c}n_R d^{3/2}H^2\iota T^{1/2},$$
which indicates that the learned mechanism is $(n_R dH^{7/3}\iota^{1/3}T^{2/3} + 6\hat{c}n_R d^{3/2}H^2\iota T^{1/2})$-approximately efficient.
2. *Assuming all agents report truthfully, the regret of agent i satisfies*
$$\mathrm{Reg}_{iT} \le \begin{cases} dH^{7/3}\iota^{1/3}T^{2/3} + 6\hat{c}n_R d^{3/2}H^2\iota T^{1/2} & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT}) \\ (1+4\hat{c}n_R)dH^{7/3}\iota^{1/3}T^{2/3} + 6\hat{c}n_R d^{3/2}H^2\iota T^{1/2} & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES}), \end{cases}$$
3. *Assuming all agents report truthfully, the regret of the seller satisfies*
$$\mathrm{Reg}_{0T} \le \begin{cases} (1+4\hat{c}n)n_R dH^{7/3}\iota^{1/3}T^{2/3} & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT}) \\ n_R dH^{7/3}\iota^{1/3}T^{2/3} & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES}). \end{cases}$$
4. *The learned mechanism is $6\hat{c}n_R dH^{7/3}\iota^{1/3}T^{2/3}$-approximately individually rational.*
5. *The learned mechanism is $(1+8\hat{c}n_R)dH^{7/3}\iota^{1/3}T^{2/3}$-approximately truthful.*

Similar to Theorem 4.1, we choose a proper K in Theorem 4.2 that can lead to a small upper bound of $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}$ in terms of d, H, ι, and T, which is $\mathcal{O}\big(n(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}\big)$. Theorem 4.2 also gives the seller and agent regret bounds for the two settings $(\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT})$ and $(\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES})$, showing that the seller and agent regret bounds vary between the ones under these two extreme cases according to Remark 3.2. Note that when the problem reduces to the tabular setting, we have $d = |\mathcal{S}||\mathcal{A}|$ in Theorems 4.1 and 4.2. When $d \le |\mathcal{S}||\mathcal{A}|$, we obtain a better rate than that under the tabular setting.

Metrics	Theorem 4.1 ($\zeta_1 = \texttt{ETC}$)	Theorem 4.2 ($\zeta_1 = \texttt{EWC}$)
Reg_T^W	$(1+2\hat{c})n_R dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}}$	$n_R dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}} + 6\hat{c}n_R d^{\frac{3}{2}}H^2\iota T^{\frac{1}{2}}$
Reg_{iT}	$(1+2\hat{c}n_R)dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}}$ $\blacklozenge$ $(1+6\hat{c}n_R)dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}}$ $\blacktriangle$	$dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}} + 6\hat{c}n_R d^{\frac{3}{2}}H^2\iota T^{\frac{1}{2}}$ $\blacklozenge$ $(1+4\hat{c}n_R)dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}} + 6\hat{c}n_R d^{\frac{3}{2}}H^2\iota T^{\frac{1}{2}}$ $\blacktriangle$
Reg_{0T}	$(1+4\hat{c}n)n_R dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}}$ $\blacklozenge$ $n_R dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}}$ $\blacktriangle$	$(1+4\hat{c}n)n_R dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}}$ $\blacklozenge$ $n_R dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}}$ $\blacktriangle$
Approx. I.R.	$6\hat{c}n_R dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}}$	$6\hat{c}n_R dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}}$
Approx. Tr.	$(1+4\hat{c}n_R)dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}}$	$(1+8\hat{c}n_R)dH^{\frac{7}{3}}\iota^{\frac{1}{3}}T^{\frac{2}{3}}$

Table 1: Comparison of Theorem 4.1 and Theorem 4.2. Here "Approx. I.R." and "Approx. Tr." are the abbreviations of "Approximate Individual Rationality" and "Approximate Truthfulness". The results in Theorem 4.1 and Theorem 4.2 hold with probability at least $1-\delta$ respectively for any $\delta \in (0,1]$. We let $n_R := n + R_{\max}$ and $\iota := \log(36ndHT/\delta)$. We use $\blacklozenge$ to represent the configuration $(\zeta_2, \zeta_3) = (\texttt{PES}, \texttt{OPT})$ and $\blacktriangle$ to represent $(\zeta_2, \zeta_3) = (\texttt{OPT}, \texttt{PES})$. We further highlight the improvements in the welfare and agent regrets in red.

Further Discussion on Theorem 4.1 and Theorem 4.2. We summarize the results from the two theorems in Table 1. As shown in our proof sketch in Section 5, we obtain that $\mathrm{Reg}_T^W \le (n+R_{\max})HK + 2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}(T-K)$ when $\zeta_1 = \texttt{ETC}$ in Theorem 4.1 and $\mathrm{Reg}_T^W \le (n+R_{\max})HK + 6\hat{c}(n+R_{\max})\sqrt{d^3H^4(T-K)\iota^2}$ when $\zeta_1 = \texttt{EWC}$ in Theorem 4.2, where both bounds share the same term $H(n+R_{\max})K$ that results from the exploration phase. To compare the welfare regrets achieved in both theorems fairly, we in fact need the rounds of exploration K to be the same, although a straightforward idea might be setting K differently as $K = \widetilde{\mathcal{O}}(T^{2/3})$ for $\texttt{ETC}$ and $K = 0$ for $\texttt{EWC}$ to minimize the two bounds respectively. However, we note that the setting $K = 0$ for Theorem 4.2 will lead to unboundedness in the seller and agent regrets as well as the individual rationality and truthfulness according to our proof sketch in Section 5.2. Fortunately, our choice of K depends on the metric of $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}$, where $\mathrm{Reg}_T^\sharp := \sum_{i=1}^n \mathrm{Reg}_{iT}$, by taking all three types of regrets into consideration, which can naturally resolve the aforementioned issue. Moreover, under this metric, the choices of K for both theorems all have the same dependence on d, H, ι, and T as justified in our proof sketch, and thus we set the same value of K directly as $dH^{4/3}\iota^{1/3}T^{2/3}$.

From Table 1, it is seen that the same setting of K leads to the same individual rationality guarantee and nearly the same truthfulness guarantee that differs only by an absolute constant scaling factor. Again referencing Epasto et al. (2018), it is even challenging for real-world agents to capitalize on a slightly larger constant factor in the approximate truthfulness guarantees. Therefore, although a slight increase exists in the truthfulness guarantee for $\zeta_1 = \texttt{EWC}$ compared to $\zeta_1 = \texttt{ETC}$, the current setting of K is justifiable and enables a fair comparison of regrets. Then, as shown in Table 1, with $K = dH^{4/3}\iota^{1/3}T^{2/3}$, the algorithm under $\zeta_1 = \texttt{ETC}$ can improve a part of the welfare regret from $\widetilde{\mathcal{O}}(T^{2/3})$ to $\widetilde{\mathcal{O}}(T^{1/2})$. This improvement results from the use of all the data gathered up to time step t

in the `EWC` setting rather than the data collected only in the exploration phase in the `ETC` setting. From Table 1, we can also observe a similar improvement in the agent regret bound. The regret improvement also verifies the importance of using the explore-while-commit (`EWC`) strategy in the learning algorithm.

Furthermore, we remark that our regret guarantees rely on the assumption that agents report truthfully. Nevertheless, recalling our earlier discussion on our definition of δ-approximate efficiency, we note that it is in general difficult to obtain regret bounds without assuming truthfulness, and thus obtaining performance guarantees under the truthfulness assumption is reasonable according to existing works (Nazerzadeh et al., 2008; Epasto et al., 2018; Kandasamy et al., 2020).

Both Theorem 4.1 and Theorem 4.2 implies $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\} = \mathcal{O}\big(n(n + R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}\big)$. We remark that the $\widetilde{\mathcal{O}}(T^{2/3})$ regret is necessary. If we were to focus only on welfare regret, then it is well-known that the lower bound would be $\Omega(\sqrt{T})$. However, the key challenge of learning the proposed Markov VCG mechanism lies in the interplay between the three kinds of regrets studied. Consider the extreme case where we set $K = 0$ in Theorem 4.2. According to our proof sketch in Section 5.2, while the welfare regret upper bound in Equation (17) improves to $\widetilde{\mathcal{O}}(\sqrt{T})$, we can no longer control the agent nor the seller regrets in Equations (18) and (19).

At last, we justify that the $\widetilde{\mathcal{O}}(T^{2/3})$ bound is tight by providing the lower bound of $\max\big\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\big\}$ when all agents are truthful. Let Θ and Alg be the class of problems and the class of algorithms for this setting respectively, and we obtain the lower bound as follows:

Theorem 4.3 *Let* $\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}$ *be as defined in* (5). *Let all agents be truthful. Defining* $n_R := n + R_{\max}$*, we have:*

$$\inf_{\mathsf{Alg}} \sup_{\Theta} \ \mathbb{E}\left[\max\left\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\right\}\right] \geq \Omega\left(n^{4/3}H^{2/3}T^{2/3} + nn_R d\sqrt{HT}\right),$$

for $T \geq \max\{16(n-1)/(H-1), 64(d-3)^2H\}, H \geq 2, d \geq 4$ *and* $n \geq 3$.

At a high level, Theorem 4.3 indicates that the $\widetilde{\mathcal{O}}\big(T^{2/3}\big)$ upper bound of $\max\big\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\big\}$ obtained by the three regrets in Theorem 4.1 and Theorem 4.2 are tight. In other words, unlike typical single-agent RL, it is impossible to obtain $\widetilde{\mathcal{O}}(\sqrt{T})$ regret when learning the Markov VCG mechanism. The intuition behind the hard case used for the lower bound is that we need to accurately learn the VCG prices to achieve a low regret. Setting the VCG prices too high harms the agents' utilities, whereas setting them too low harms the seller's. Learning the VCG prices requires learning the welfare-maximizing policy when agent i is absent, π_*^{-i}. Combined with our need to estimate the welfare-maximizing policy, any suitable learning algorithm needs to reduce the estimation error of the value functions for all policies. Our proposed algorithm resolves this challenge by reward-free exploration, and the procedure is crucial for efficiently learning the Markov VCG mechanism. There is still a gap between the upper and lower bounds in terms of the multiplicative factors n, d, and H, and we leave the derivation of exactly matching upper and lower bounds as an open question for future work.

Our work features several prominent contributions to the existing literature in mechanism design learning and online learning of linear MDPs. As shown in Theorem 4.1 and

Theorem 4.2, our work proposes the first algorithm capable of learning a dynamic mechanism with no prior knowledge. In particular, we further show that the mechanism learned by Algorithm 1 simultaneously satisfies approximate efficiency, approximate individual rationality, and approximate truthfulness. As we will demonstrate in the sequel, the satisfaction of the approximate versions of the three mechanism design desiderata is demonstrated through novel decomposition approaches. Moreover, Theorem 4.3 demonstrates that our achieved results are minimax optimal up to problem-dependent constants.

5. Proof Sketch

In this section, we outline the analysis of our theorems. The formal proof is deferred to Appendix C - F. For a concise presentation, in the proof, we let $V_1^*(x_1; r) := \max_\pi V_1^\pi(x_1; r)$ for any reward function r. We further provide a table of notation in Appendix A summarizing all notations used here.

5.1 Proof Sketch of Theorem 4.1

We assume that all agents report their rewards truthfully in the proof of the upper bounds of the welfare regret, the agent regret, and the seller regret. Since we use the explore-then-commit algorithm when $\zeta_1 = \texttt{ETC}$, we decompose all the regrets into two components: the regret incurred in the exploration phase and the regret incurred in the exploitation phase. Additionally, for each of these regrets, we first show its dependence on both the rounds of exploration K and the total rounds T. Then we determine K that can lead to a tight upper bound of $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}$ in terms of n, d, H, ι, and T.

Welfare Regret. We first decompose the welfare regret into two parts as follows:

$$\mathrm{Reg}_T^W = \textstyle\sum_{t=1}^K \mathrm{reg}_t^W + \sum_{t=K+1}^T \mathrm{reg}_t^W, \tag{8}$$

where $\mathrm{reg}_t^W := V_1^{\pi_*}(x_1; R) - V_1^{\widehat{\pi}_t}(x_1; R)$ is the instantaneous welfare regret. Here $\sum_{t=1}^K \mathrm{reg}_t^W$ is the welfare regret in the exploration phase and $\sum_{t=K+1}^T \mathrm{reg}_t^W$ is for the exploitation phase. For the regret incurred in the exploration phase in Equation (8), we bound the instantaneous regret reg_t^W at each time step by $H(n + R_{\max})$, which is the maximum of the instantaneous regret at each round. For the exploitation welfare regret in Equation (8), we can bound its instantaneous welfare regret reg_t^W by $2\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K}$ with high probability, whose proof is inspired by the regret proof for learning linear MDPs, as the prices cancel out when calculating social welfare. Therefore, with high probability, the following welfare regret bound holds

$$\mathrm{Reg}_T^W \le H(n + R_{\max})K + 2\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K}(T - K), \tag{9}$$

where the rounds of the exploration phase K will be determined later.

Agent Regret. We have the following regret decomposition in terms of the exploration phase and exploitation phase as follows,

$$\mathrm{Reg}_{iT} = \textstyle\sum_{t=1}^K \mathrm{reg}_{it} + \sum_{t=K+1}^T \mathrm{reg}_{it}, \tag{10}$$

where $\mathrm{reg}_{it} := u_{i*} - u_{it}$ is the instantaneous regret of agent i. As shown in Algorithm 1, we do not charge the agents in the exploration phase. Thus, the instantaneous regret of agent i in the exploration phase can be upper bounded as

$$\mathrm{reg}_{it} \leq u_{i*} - \min_{\pi} V_1^{\pi}(x_1; r_i) \leq u_{i*} = V_1^{\pi_*}(x_1; r_i) - p_{i*} \leq V_1^{\pi_*}(x_1; r_i) \leq H, \quad 1 \leq t \leq K.$$

For the terms in the second summation in Equation (10), i.e., the instantaneous regret of agent i incurred in the exploitation phase, we first decompose it to several simple terms as follows,

$$\mathrm{reg}_{it} = \underbrace{\big[V_1^{\pi_*}\big(x_1; R\big) - V_1^{\widehat{\pi}^t}\big(x_1; R\big)\big]}_{\text{(i.1)}} + \underbrace{\big[F_t^{-i} - V_1^{\pi_*^{-i}}\big(x_1; R^{-i}\big)\big]}_{\text{(i.2)}} + \underbrace{\big[V_1^{\widehat{\pi}^t}\big(x_1; R^{-i}\big) - G_t^{-i}\big]}_{\text{(i.3)}}, \quad (11)$$

where (i.1) is the suboptimality of $\widehat{\pi}^t$, (i.2) is the estimation error of $V_1^{\pi_*^{-i}}\big(x_1; R^{-i}\big)$ by F_t^{-i}, and (i.3) is the policy evaluation error. To satisfy the desiderata of the mechanism design in Lemma 2.1, we set F-function as the optimistic (when $\zeta_2 = \texttt{OPT}$) or pessimistic (when $\zeta_2 = \texttt{PES}$) estimate of $V_1^{\pi_*^{-i}}\big(x_1; R^{-i}\big)$, while the G-function is the estimate of $V_1^{\widehat{\pi}^t}\big(x_1; R^{-i}\big)$ w.r.t. the learned policy $\widehat{\pi}^t$. The different structures of F-function and G-function lead to different ways of bounding (i.2) and (i.3).When we set $(\zeta_2, \zeta_3) = (\texttt{PES}, \texttt{OPT})$, we have that (i.2) ≤ 0 and (i.3) ≤ 0 since F_t^{-i} and G_t^{-i} are the pessimistic and optimistic estimates respectively. Then, we can bound the instantaneous regret of agent i in the exploitation phase as follows

$$\mathrm{reg}_{it} \leq V_1^{\pi_*}\big(x_1; R\big) - V_1^{\widehat{\pi}^t}\big(x_1; R\big) \leq 2\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K}, \quad 1 \leq t \leq K.$$

When we set $(\zeta_2, \zeta_3) = (\texttt{OPT}, \texttt{PES})$, we can bound (i.2) and (i.3) by $2\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K}$ respectively with high probability. Thus, we bound the instantaneous regret of agent i in the exploitation phase as

$$\mathrm{reg}_{it} \leq 6\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K}, \quad K < t \leq T.$$

Combining the regrets incurred in both phases, we obtain with high probability,

$$\mathrm{Reg}_{iT} \leq HK + 6\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K}(T - K). \quad (12)$$

Seller Regret. We can decompose the seller regret into two parts as follows

$$\mathrm{Reg}_{0T} = \textstyle\sum_{t=1}^{K} \mathrm{reg}_{0t} + \sum_{t=K+1}^{T} \mathrm{reg}_{0t}, \quad (13)$$

where $\mathrm{reg}_{0t} := u_{0*} - u_{0t}$ is the instantaneous regret of the seller. Since the seller charges a price of 0 to all agents, the instantaneous seller regret in the exploration phase can be bounded as

$$\mathrm{reg}_{0t} \leq u_{0*} - \min_{\pi} V^{\pi}(x_1; r_0) \leq u_{0*} \leq H(n + R_{\max}), \quad 1 \leq t \leq K.$$

For the instantaneous seller regret in the exploitation phase ($K < t \leq T$), we have the following decomposition

$$\mathrm{reg}_{0t} = (n-1)\underbrace{\big[V_1^{\widehat{\pi}^t}(x_1;R) - V_1^*(x_1;R)\big]}_{\text{(ii.1)}} + \sum_{i=1}^n \underbrace{\big[V_1^*(x_1;R^{-i}) - F_t^{-i}\big]}_{\text{(ii.2)}} + \sum_{i=1}^n \underbrace{\big[G_t^{-i} - V^{\widehat{\pi}^t}(x_1;R^{-i})\big]}_{\text{(ii.3)}}.$$

Here we have (ii.1) $= -$(i.3), (ii.2) $= -$(i.1), and (ii.3) $= -$(i.2) with (i.1), (i.2), (i.3) defined in Equation (11). Notice that (ii.1) ≤ 0 always holds regardless of the choice of (ζ_2, ζ_3). We can upper bound (ii.2) and (ii.3) using the same method as bounding (i.1) and (i.2). Thus, with high probability, reg_{0t} in the exploitation phase ($K < t \le T$) is upper bounded as

$$\mathrm{reg}_{0t} \le \begin{cases} 4\hat{c}n(n+R_{\max})\sqrt{d^3H^6\iota/K} & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES}, \texttt{OPT}) \\ 0 & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT}, \texttt{PES}). \end{cases}$$

Combining the above results, the seller regret Reg_{0T} is bounded by

$$\begin{cases} H(n+R_{\max})K + 4\hat{c}n(n+R_{\max})\sqrt{d^3H^6\iota/K}(T-K) & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES}, \texttt{OPT}) \\ H(n+R_{\max})K & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT}, \texttt{PES}). \end{cases} \tag{14}$$

Choice of K. We determine the value of K which can give a tight bound of $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}$ where $\mathrm{Reg}_T^\sharp = \sum_{i=1}^n \mathrm{Reg}_{iT}$. According to (9), (12), and (14), comparing the upper bounds of $n\mathrm{Reg}_T^W$, $\mathrm{Reg}_T^\sharp$, and Reg_{0T}, we always have

$$\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\} \le H(n+R_{\max})nK + 6\hat{c}(n+R_{\max})n\sqrt{d^3H^6\iota/K}(T-K).$$

Focusing on the factors of H, n, d, T, and ι, we set $K = dH^{4/3}\iota^{1/3}T^{2/3}$, which can minimize the order of these factors in the above inequality, and obtain the bounds in Theorem 4.1.

Next, we provide the proof sketches for the approximate individual rationality and truthfulness. Note that in the following analysis, we do not assume the agents are reporting truthfully. We denote the potentially untruthful reward function of agent i at step h by $\widetilde{r}_{ih}$ and then $\widetilde{r}_i = \{\widetilde{r}_{ih}\}_{h=1}^H$. We further let $\widetilde{R}^{-i} := r_0 + \sum_{j=1, j\neq i}^n \widetilde{r}_j$.

Individual Rationality. To prove the individual rationality, we assume that agent i reports truthfully according to the reward function r_i and other agents may report untruthfully according to the reward function $\widetilde{r}_j$ for $j \neq i$. Under this reward setting, let $\widetilde{\pi}_t^{\dagger i}$ be the learned seller's policy substituting $\widehat{\pi}^t$ in Algorithm 1, which is generated by Algorithm 3 in the current reward setting. We further denote the associated F and G functions as $F_t^{\dagger,-i}$ and $G_t^{\dagger,-i}$ generated by Algorithms 3 and 4 respectively. Note that we do not charge the agents in the exploration phase ($t \le K$), and hence the utilities in this phase are always non-negative. Thus, we only need to consider the utilities in the exploitation phase ($t > K$). Then, according to the definition of u_{it}, under the current setting of the reward, the instantaneous utility u_{it} of agent i can be decomposed as

$$u_{it} = V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1; r_i) - p_{it}^\dagger = \underbrace{\Big[V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1; r_i + \widetilde{R}^{-i}) - F_t^{\dagger,-i}\Big]}_{\text{(iii.1)}} + \underbrace{\Big[G_t^{\dagger,-i} - V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1; \widetilde{R}^{-i})\Big]}_{\text{(iii.2)}}, \tag{15}$$

where $p_{it}^\dagger = F_t^{\dagger,-i} - G_t^{\dagger,-i}$. To prove the individual rationality, we bound (iii.1) and (iii.2) from below. Here we denote the optimistic version of $F_t^{\dagger,-i}$, when $\zeta_2 = \texttt{OPT}$, by

$\widehat{V}_1^{t,\dagger}\big(x_1;\widetilde{R}^{-i}\big)$ according to Algorithm 3, which implies $F_t^{\dagger,-i}\le \widehat{V}_1^{t,\dagger}\big(x_1;\widetilde{R}^{-i}\big)$. Then, we have (iii.1) $\ge V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1;r_i+\widetilde{R}^{-i}\big)-\widehat{V}_1^{t,\dagger}\big(x_1;\widetilde{R}^{-i}\big)$. This can be further decomposed as

$$V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1;r_i+\widetilde{R}^{-i}\big)-\widehat{V}_1^{t,\dagger}\big(x_1;\widetilde{R}^{-i}\big)=\underbrace{\Big[V_1^*(x_1;r_i+\widetilde{R}^{-i})-V_1^*(x_1;\widetilde{R}^{-i})\Big]}_{\text{(iii.1a)}}$$

$$+\underbrace{\Big[V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1;r_i+\widetilde{R}^{-i}\big)-V_1^*(x_1;r_i+\widetilde{R}^{-i})\Big]}_{\text{(iii.1b)}}+\underbrace{\Big[V_1^*(x_1;\widetilde{R}^{-i})-\widehat{V}_1^{t,\dagger}\big(x_1;\widetilde{R}^{-i}\big)\Big]}_{\text{(iii.1c)}}.$$

Note that (iii.1a) ≥ 0 always holds since both terms in (iii.1a) are optimal value functions but $V_1^*(x_1;r_i+\widetilde{R}^{-i})$ has larger reward function. Here (iii.1b) is the suboptimality of policy $\widetilde{\pi}_t^{\dagger i}$ and (iii.1c) is the estimation error of $V_1^*(x_1;\widetilde{R}^{-i})$ by $\widehat{V}_1^{t,\dagger}\big(x_1;\widetilde{R}^{-i}\big)$. We lower bound (iii.1b) and (iii.1c) by $-2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}$ respectively with high probability. Then (iii.2) can be lower bounded by $-4\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}$. For (iii.2), the policy evaluation error for policy $\widetilde{\pi}_t^{\dagger i}$, we can lower bound it by $-2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}$ invoking Lemma C.1. Recall that we set $K=dH^{4/3}\iota^{1/3}T^{2/3}$. Then we lower bound the summation of (iii.1) and (iii.2) over T episodes by $-4\hat{c}(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}$ and $-2\hat{c}(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}$ respectively. Combining these two parts, with high probability, we have

$$U_{iT}\le -6\hat{c}(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3},$$

which indicates that the learned mechanism is $6\hat{c}(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}$-approximately individually rational.

Truthfulness. We consider two cases: (1) agent i reports truthfully and others may report untruthfully (2) all agents may report untruthfully. Then we denote by r_i the truthful reward and $\widetilde{r}_i$ the potentially untruthful reward for all $i\in[n]$. For case (1), we adopt the same definitions of $F_t^{\dagger,-i}, G_t^{\dagger,-i}$, $\widetilde{\pi}_t^{\dagger i}$, and $u_{it}=V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1;r_i)-p_{it}^{\dagger}$ as in the above proof of individual rationality. For case (2), under the untruthful reporting of $\{\widetilde{r}_i\}_{i\in[n]}$, we let $\widetilde{\pi}_t^{\ddagger}$ be the learned policy for the seller under the reward $\widetilde{R}:=r_0+\sum_{i=1}^n\widetilde{r}_i$ in Algorithm 1, $F_t^{\ddagger,-i}$ and $G_t^{\ddagger,-i}$ be the associated F and G functions generated by Algorithms 3 and 4 respectively, and $\widetilde{u}_{it}=V_1^{\widetilde{\pi}_t^{\ddagger}}(x_1;r_i)-p_{it}^{\ddagger}$ with $p_{it}^{\ddagger}=F_t^{\ddagger,-i}-G_t^{\ddagger,-i}$. We then have the following decomposition

$$\widetilde{U}_{iT}-U_{iT}=\textstyle\sum_{t=1}^K(\widetilde{u}_{it}-u_{it})+\sum_{t=K+1}^T(\widetilde{u}_{it}-u_{it}). \tag{16}$$

For the first summation, since the agents are not charged, we have

$$\textstyle\sum_{t=1}^K(\widetilde{u}_{it}-u_{it})\le\sum_{t=1}^K\widetilde{u}_{it}\le\sum_{t=1}^K\max_\pi V^\pi(x_1;r_i)\le HK.$$

We now turn to decomposing the second summation in Equation (16). We have for $t>K$,

$$\widetilde{u}_{it}-u_{it}=\big[V_1^{\widetilde{\pi}_t^{\ddagger}}(x_1;r_i)-F_t^{\ddagger,-i}+G_t^{\ddagger,-i}\big]-\big[V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1;r_i)-F_t^{\dagger,-i}+G_t^{\dagger,-i}\big].$$

Notice that when $\zeta_1=\texttt{ETC}$, we only use the data collected in the exploration phase to calculate the F function. Thus, we have $F_t^{\dagger,-i}=F_t^{\ddagger,-i}$. Then, we can show that $\widetilde{u}_{it}-u_{it}$

can be decomposed as

$$
\begin{aligned}
\widetilde{u}_{it} - u_{it} = {} & \underbrace{\big[G_t^{\ddagger,-i} - V_1^{\widetilde{\pi}_t^{\ddagger}}\big(x_1;\widetilde{R}^{-i}\big)\big]}_{\text{(iv.1)}} + \underbrace{\big[V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1;\widetilde{R}^{-i}\big) - G_t^{\dagger,-i}\big]}_{\text{(iv.2)}} \\
& + \underbrace{\big[V_1^{\widetilde{\pi}_t^{\ddagger}}\big(x_1;r_i+\widetilde{R}^{-i}\big) - V_1^{\widetilde{\pi}_*^{i}}\big(x_1;r_i+\widetilde{R}^{-i}\big)\big]}_{\text{(iv.3)}} + \underbrace{\big[V_1^{\widetilde{\pi}_*^{i}}\big(x_1;r_i+\widetilde{R}^{-i}\big) - V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1;r_i+\widetilde{R}^{-i}\big)\big]}_{\text{(iv.4)}}.
\end{aligned}
$$

We remark that different from the bandit setting in Kandasamy et al. (2020), the estimates of value functions are not linear w.r.t. the reward functions, i.e., $\widehat{V}_1^{t,\pi}(x_1;R_1) + \widehat{V}_1^{t,\pi}(x_1;R_2) \neq \widehat{V}_1^{t,\pi}(x_1;R_1+R_2)$ or $\check{V}_1^{\pi}(x_1;R_1) + \check{V}_1^{t,\pi}(x_1;R_2) \neq \check{V}_1^{t,\pi}(x_1;R_1+R_2)$ for any reward functions R_1 and R_2, due to the truncation of Q-functions in Algorithm 3 and Algorithm 4. However, the true value function, i.e., $V_1^{\pi}(x_1;R_1) + V_1^{\pi}(x_1;R_2) = V_1^{\pi}(x_1;R_1+R_2)$, is linear w.r.t. the reward function. This leads to a novel and more complex decomposition in the above equation. Note that (iv.3) ≤ 0 since $V_1^{*}(x_1;r_i+\widetilde{R}^{-i}) = \max_{\pi} V_1^{\pi}(x_1;r_i+\widetilde{R}^{-i})$. And (iv.4) is the suboptimality of policy $\widetilde{\pi}_t^{\dagger i}$. Then, with high probability, the term (iv.4) is upper bounded by $2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}$. Here (iv.1) and (iv.2) are evaluation errors depending on the setting of ζ_3 under different reward settings. When $\zeta_3 = \texttt{OPT}$, we have (iv.1) $\leq 2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}$ while (iv.2) ≤ 0. And when $\zeta_3 = \texttt{PES}$, we have (iv.1) ≤ 0 and (iv.2) $\leq 2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}$. Thus, regardless of the choices for ζ_2, ζ_3, we always have

$$
\widetilde{u}_{it} - u_{it} \leq 4\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}, \qquad t > K.
$$

Summing up the regret incurred in both the exploration and exploitation phases as in (16), and setting $K = dH^{4/3}\iota^{1/3}T^{2/3}$, with high probability, we have

$$
\widetilde{U}_{iT} - U_{iT} \leq \big(1 + 4\hat{c}(n+R_{\max})\big) dH^{7/3}\iota^{1/3}T^{2/3},
$$

which implies that the mechanism learned by our algorithm is $\big(1+4\hat{c}(n+R_{\max})\big)dH^{7/3}\iota^{1/3}T^{2/3}$-approximately truthful.

5.2 Proof Sketch of Theorem 4.2

We assume that all agents report their rewards truthfully in the proof of the upper bounds of the welfare regret, the agent regret, and the seller regret. Although we use all the data generated in T rounds to compute our mechanism when $\zeta_1 = \texttt{EWC}$, we still need to perform reward-free exploration for individual rationality and truthfulness. Thus, we also decompose regrets into two components: the regret incurred in the exploration phase and the regret incurred in the exploitation phase.

Welfare Regret. We adopt the same decomposition as in Equation (8) and decompose the welfare regret as $\mathrm{Reg}_T^W = \sum_{t=1}^K \mathrm{reg}_t^W + \sum_{t=K+1}^T \mathrm{reg}_t^W$. The first summation $\sum_{t=1}^K \mathrm{reg}_t^W$, the welfare regret incurred in the exploration phase, can be bounded by $(n+R_{\max})HK$ as in Section 5.1. The key difference between the proofs of welfare regrets in Theorem 4.2 and Theorem 4.1 lies in the upper bound of $\sum_{t=K+1}^T \mathrm{reg}_t^W$, i.e., the regret incurred in the exploitation phase. When $\zeta_1 = \texttt{EWC}$, we use the information gathered up to round t for planning in the exploitation phase, instead of just using the K rounds' exploration data as we do when $\zeta_1 = \texttt{ETC}$. Thus, we can bound the regret incurred in the exploitation phase by $6\hat{c}(n+R_{\max})\sqrt{d^3H^4(T-K)\iota^2}$ with high probability, whose proof takes inspiration from the regret proof for online linear MDPs with exploration, as the calculation of social welfare does not involve prices. Combining the regrets incurred in both phases, with high probability, the following welfare regret bound holds

$$\mathrm{Reg}_T^W \le (n+R_{\max})HK + 6\hat{c}(n+R_{\max})\sqrt{d^3H^4(T-K)\iota^2}, \tag{17}$$

where the rounds of the exploration phase K will be determined later.

Agent Regret. Following Equation (13), we decompose the regret of agent i in terms of the exploration phase and exploitation phase as $\mathrm{Reg}_{iT} = \sum_{t=1}^K \mathrm{reg}_{it} + \sum_{t=K+1}^T \mathrm{reg}_{it}$. For the first summation $\sum_{t=1}^K \mathrm{reg}_{it}$, the agent i's regret in the exploration phase, we can bound it by HK as in Section 5.1. For the term $\sum_{t=K+1}^T \mathrm{reg}_{it}$, recalling the decomposition in Equation (11), it can be decomposed as

$$\underbrace{\sum_{t=K+1}^{T}\left[V_1^{\pi_*}(x_1;R) - V_1^{\widehat{\pi}^t}(x_1;R)\right]}_{\text{(i.1)}} + \underbrace{\sum_{t=K+1}^{T}\left[F_t^{-i} - V_1^{\pi_*^{-i}}(x_1;R^{-i})\right] + \left[V_1^{\widehat{\pi}^t}(x_1;R^{-i}) - G_t^{-i}\right]}_{\text{(i.2)}}.$$

For term (i.1), we can bound it by $6\hat{c}(n+R_{\max})\sqrt{d^3H^4(T-K)\iota^2}$ with high probability leveraging the information gathered up to round t instead of K in the exploitation phase, whose proof follows the proof for welfare regret when $\zeta_1 = \texttt{EWC}$. For term (i.2), following the same proof in Section 5.1, we get an upper bound 0 when $(\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT})$ and an upper bound $4\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}(T-K)$ when $(\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES})$. Combining the upper bounds of (i.1), (i.2) for $\sum_{t=K+1}^T \mathrm{reg}_{it}$ and the regret bound for the exploitation phase $\sum_{t=1}^K \mathrm{reg}_{it} \le HK$, with high probability, Reg_{iT} has the following upper bound,

$$\begin{cases} HK + 6\hat{c}(n+R_{\max})\sqrt{d^3H^4(T-K)\iota^2} & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT}) \\ HK + \hat{c}(n+R_{\max})\big(6\sqrt{d^3H^4(T-K)\iota^2} + 4\sqrt{d^3H^6\iota/K}(T-K)\big) & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES}). \end{cases} \tag{18}$$

Seller Regret. Since the trajectories we collected are according to the process where all the agents are engaged, we can not make a better estimation of the VCG prices even if we use the information gathered in the exploitation phase. Also, note that the seller regret comes from the estimation error of the VCG prices, we cannot improve the analysis of the seller regret. Thus, we reuse the proof in Section 5.1, and can get the upper bound of seller regret Reg_{0T} as

$$\begin{cases} H(n+R_{\max})K + 4\hat{c}n(n+R_{\max})\sqrt{d^3H^6\iota/K}(T-K) & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT}) \\ H(n+R_{\max})K & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES}). \end{cases} \tag{19}$$

Choice of K. We determine the value of K which can give a tight bound of $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}$ where $\mathrm{Reg}_T^\sharp = \sum_{i=1}^n \mathrm{Reg}_{iT}$. According to (17), (18), and (19), comparing the upper bounds of $n\mathrm{Reg}_T^W$, $\mathrm{Reg}_T^\sharp$, and Reg_{0T}, we always have the upper bound of $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}$ as

$$n(n+R_{\max})\big(HK + 6\hat{c}\sqrt{d^3H^4(T-K)\iota^2} + 4\hat{c}\sqrt{d^3H^6\iota/K}(T-K)\big).$$

Focusing on the factors of H, n, d, T, and ι, we set $K = dH^{4/3}\iota^{1/3}T^{2/3}$, which can minimize the order of these factors in the above inequality, and obtain the bounds in Theorem 4.2.

Individual Rationality. We assume that agent i reports truthfully according to the reward function r_i and other agents may report untruthfully according to the reward function $\widetilde{r}_j$ for $j \neq i$. According to the above assumption, agent i cannot manipulate the policy used during the exploitation phase, which implies that agent i can not influence trajectories collected during the exploitation phase. Note that the only difference between the algorithm when $\zeta_1 = \texttt{EWC}$ and $\zeta_1 = \texttt{ETC}$ is the trajectories collected during exploitation are used for estimating policy and VCG prices. Thus, agent i cannot affect policy and VCG price estimates obtained during exploration. Hence we can reuse the proof for individual rationality in Section 5.1 and get the conclusion that the mechanism we learned is $6\hat{c}(n + R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}$-approximately individually rational.

Truthfulness. The proof for truthfulness when $\zeta_1 = \texttt{EWC}$ significantly differs from the case when $\zeta_1 = \texttt{ETC}$. At a high level, when $\zeta_1 = \texttt{ETC}$, we use the fact that the data used to calculate F is collected entirely during the exploration phase and is not affected by agent i potentially reporting untruthfully, and hence $F_t^{\ddagger,-i}$ and $F_t^{\dagger,-i}$ cancel out. Unfortunately, when $\zeta_1 = \texttt{EWC}$, F depends the untruthful behavior of agent i. The trajectories collected during exploitation affect F. The policy used for collecting these trajectories is affected by the agent i's report. Because agent i's untruthfulness impacts F, we need to bound the difference between $F_t^{\dagger,-i}$ and $F_t^{\ddagger,-i}$, which is different from the proof of truthfulness in Section 5.1. Thus, we follow the decomposition in Equation (16). For the first summation in Equation (16), which corresponds to the exploration phase, we can upper bound it by HK. For the second summation that relates to the exploitation phase, regardless of other agents' truthfulness, the amount of utility an agent gains from untruthful reporting $\widetilde{u}_{it} - u_{it}$ for $t > K$ can be decomposed as

$$\widetilde{u}_{it} - u_{it} = \underbrace{\big[V_1^{\widetilde{\pi}_t^{\ddagger}}\big(x_1; r_i + \widetilde{R}^{-i}\big) - V_1^{*}\big(x_1; r_i + \widetilde{R}^{-i}\big)\big]}_{\text{(i.1)}} + \underbrace{\big[V_1^{*}\big(x_1; r_i + \widetilde{R}^{-i}\big) - V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1; r_i + \widetilde{R}^{-i}\big)\big]}_{\text{(i.2)}}$$

$$+ \underbrace{\big[G_t^{\ddagger,-i} - V_1^{\widetilde{\pi}_t^{\ddagger}}\big(x_1; \widetilde{R}^{-i}\big)\big]}_{\text{(i.3)}} + \underbrace{\big[V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1; \widetilde{R}^{-i}\big) - G_t^{\dagger,-i}\big]}_{\text{(i.4)}} + \underbrace{\big[F_t^{\dagger,-i} - F_t^{\ddagger,-i}\big]}_{\text{(1.5)}}.$$

Following Section 5.1, regardless of the choice of ζ_3, with high probability, we have

$$\text{(i.1)} + \text{(i.2)} + \text{(i.3)} + \text{(i.4)} \leq 4\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K}.$$

We next focus on the upper bound of (i.5). When $\zeta_1 = \texttt{EWC}$, the trajectories collected during the exploitation phase may differ for the computations of $\widehat{V}_1^{t,\dagger}(x_1; \widetilde{R}^{-i})$ and $\widecheck{V}_1^{t,\ddagger}(x_1; \widetilde{R}^{-i})$, due to agent i's untruthful reporting. Fortunately, the policy evaluation error can still be bounded. The reward-free exploration procedure in Algorithm 2 ensures that the data collected during exploitation cannot affect the estimated value functions too much. The estimation error surrounding estimated value functions is already small due to the exploration phase. As a result, adding more trajectories during exploitation cannot significantly alter our estimated values, thereby controlling the policy evaluation error. More formally, we have

$$(\text{i.5}) \leq \underbrace{\Big(\widehat{V}_1^{t,\dagger}(x_1;\widetilde{R}^{-i}) - V_1^*(x_1;\widetilde{R}^{-i})\Big)}_{(\text{ii.1})} + \underbrace{\Big(V_1^*(x_1;\widetilde{R}^{-i}) - \check{V}_1^{t,\ddagger}(x_1;\widetilde{R}^{-i})\Big)}_{(\text{ii.2})},$$

where (ii.1) and (ii.2) can be upper bounded by $2\hat{c}\sqrt{d^3H^6\iota/K}$ with high probability respectively. In summary, we have that, with high probability, for all $t > K$,

$$\widetilde{u}_{it} - u_{it} \leq 8\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K}.$$

Summing $\widetilde{u}_{it} - u_{it}$ from $t = 1$ to T, recalling the bound for all $t \in [K]$, and setting $K = dH^{4/3}\iota^{1/3}T^{2/3}$, with high probability, we get

$$\widetilde{U}_{iT} - U_{iT} \leq (1 + 8\hat{c}(n + R_{\max}))dH^{7/3}\iota^{1/3}T^{2/3},$$

which implies the mechanism we learned is $(1 + 8\hat{c}(n + R_{\max}))dH^{7/3}\iota^{1/3}T^{2/3}$-approximately truthful.

5.3 Proof Sketch of Theorem 4.3

Although the previous work Kandasamy et al. (2020) studies the lower bound for mechanism design in the bandit setting, we remark that deriving the lower bound for our problem is non-trivial which requires different constructions and proof techniques from that of this earlier work. Our lower bound takes into account the function approximation and the transition model within the finite horizon, which cannot be handled by Kandasamy et al. (2020). In addition, our work invalidates the Gaussian reward construction in Kandasamy et al. (2020) because of the bounded reward assumption in our work. We use a different construction with the Bernoulli reward and apply a different anti-concentration analysis.

Our lower bound is devised by considering two hard cases for the Markov VCG learning with linear function approximation. For the first hard case, we mimic the strategy of the lower bound design as in Kandasamy et al. (2020) with constructing two problems θ_0 and θ_1 that are hard to distinguish. Then, the lower bound is obtained by further lower bounding specific quantities w.r.t. θ_0 and θ_1. Though we follow such a proving strategy, the model construction is specific to our MDP setting and different from the existing work as discussed above. Specifically, we consider constructing two linear MDPs for the two problems θ_0 and θ_1 that are hard to distinguish, i.e., they share the same linear feature mapping and deterministic transition kernel but have a small difference in the distribution of reward functions. In addition, we let the dimension of the linear space be $d = n + 2$. Note that due to the bounded reward assumption in this work, we define Bernoulli reward functions which further leads to a different anti-concentration analysis. By bounding the specific quantities associated with θ_0 and θ_1, we obtain a dimension-free lower bound in an order of $\Omega(n^{4/3}H^{2/3}T^{2/3})$.

Moreover, to further understand the dependence on any dimension d, our second hard case is constructed by the observation that $\max\big(n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\big) \geq n\mathrm{Reg}_T^W$ always holds. This further inspires us to connect the lower bound to the problem of learning a d dimensional linear MDP with $n + 1$ reward functions. We thus prove that the lower bound of $n\mathrm{Reg}_T^W$ is $\Omega\big(n(n+R_{\max})d\sqrt{HT}\big)$, where the factor $n+R_{\max}$ reflects the impact of the n agent reward functions and the seller reward function on the lower bound. Combining the above two hard cases, we eventually obtain the lower bound for our mechanism design problem, which is $\Omega\big(n^{4/3}H^{2/3}T^{2/3}+n(n+R_{\max})d\sqrt{HT}\big)$. Please refer to Appendix E for the detailed proof.

6. Conclusion

In this paper, we consider the problem where the agents interact with the mechanism designer according to an unknown MDP. We focus on the online setting with linear function approximation and attempt to recover the dynamic VCG mechanism over multiple rounds of interaction. We propose novel algorithms to learn the mechanism and show that the regret of our proposed method is upper bounded by $\widetilde{\mathcal{O}}(T^{2/3})$, where T is the total number of rounds. We further devise a lower bound, incurring the same $\Omega(T^{2/3})$ regret as the upper bound. Our work establishes the regret guarantee for online RL in solving dynamic mechanism design problems without prior knowledge of the underlying model.

Appendix

Contents

Appendix A. Table of Notation

To summarize our notations, we present the following table of notation.

Table 2: Table of Notation

Notation	Meaning
R	summation of the reward functions of the seller and the agents, i.e., $\sum_{i=0}^{n} r_i$
R^{-i}	summation of the reward functions except that of agent i, i.e., $\sum_{j=0,j\neq i}^{n} r_j$
$V^*(;r)$	$\max_\pi V^\pi(;r)$ for any value function r
$\widehat{\pi}^t$	seller's policy in Alg. 1 w.r.t. the reward function R, generated by Alg. 3
$\widehat{V}_h^{t,*}(x_1;R)$	optimistic value function generated by Alg. 3 w.r.t. R
$\widehat{V}_h^{t,*}(x_1;R^{-i})$	optimistic value function generated by Alg. 3 w.r.t. R^{-i}
$\widecheck{V}_h^{t,*}(x_1;R^{-i})$	pessimistic value function generated by Alg. 3 w.r.t. R^{-i}
$\widehat{V}_h^{t,\widehat{\pi}^t}(x_1;R^{-i})$	optimistic value function generated by Alg. 4 w.r.t. R^{-i} and $\widehat{\pi}^t$
$\widecheck{V}_h^{t,\widehat{\pi}^t}(x_1;R^{-i})$	pessimistic value function generated by Alg. 4 w.r.t. R^{-i} and $\widehat{\pi}^t$
F_t^{-i}	$\widehat{V}_1^{t,*}(x_1;R^{-i})$ if $\zeta_2 = \texttt{OPT}$; $\widecheck{V}_1^{t,*}(x_1;R^{-i})$ if $\zeta_2 = \texttt{PES}$
G_t^{-i}	$\widehat{V}_1^{t,\widehat{\pi}^t}(x_1;R^{-i})$ if $\zeta_3 = \texttt{OPT}$; $\widecheck{V}_1^{t,\widehat{\pi}^t}(x_1;R^{-i})$ if $\zeta_3 = \texttt{PES}$
ι	the logarithmic term $\log(36ndHT/\delta)$
$\widetilde{r}_i$	potentially untruthful reward function for agent i, $i \in [n]$
$\widetilde{R}^{-i}$	$r_0 + \sum_{j=1,j\neq i}^{n} \widetilde{r}_j$
$\widetilde{\pi}_t^{\dagger i}$	seller's policy in Alg. 1 w.r.t. the reward function $r_i + \widetilde{R}^{-i}$, generated by Alg. 3
$\widehat{V}_h^{t,\dagger}(x_1;r_i+\widetilde{R}^{-i})$	optimistic value by Alg. 3 w.r.t. $r_i + \widetilde{R}^{-i}$ if agents are untruthful except agent i
$\widehat{V}_h^{t,\dagger}(x_1;\widetilde{R}^{-i})$	optimistic value by Alg. 3 w.r.t. $\widetilde{R}^{-i}$ if agents are untruthful except agent i
$\widecheck{V}_h^{t,\dagger}(x_1;\widetilde{R}^{-i})$	pessimistic value by Alg. 3 w.r.t. $\widetilde{R}^{-i}$ if agents are untruthful except agent i
$\widehat{V}_h^{t,\widetilde{\pi}_t^{\dagger i}}(x_1;\widetilde{R}^{-i})$	optimistic value by Alg. 4 w.r.t. $\widetilde{R}^{-i}$, $\widetilde{\pi}_t^{\dagger i}$ if agents are untruthful except agent i
$\widecheck{V}_h^{t,\widetilde{\pi}_t^{\dagger i}}(x_1;\widetilde{R}^{-i})$	pessimistic value by Alg. 4 w.r.t. $\widetilde{R}^{-i}$, $\widetilde{\pi}_t^{\dagger i}$ if agents are untruthful except agent i
$F_t^{\dagger,-i}$	$\widehat{V}_1^{t,\dagger}(x_1;\widetilde{R}^{-i})$ if $\zeta_2 = \texttt{OPT}$; $\widecheck{V}_1^{t,\dagger}(x_1;\widetilde{R}^{-i})$ if $\zeta_2 = \texttt{PES}$
$G_t^{\dagger,-i}$	$\widehat{V}_1^{t,\widetilde{\pi}_t^{\dagger i}}(x_1;\widetilde{R}^{-i})$ if $\zeta_3 = \texttt{OPT}$; $\widecheck{V}_1^{t,\widetilde{\pi}_t^{\dagger i}}(x_1;\widetilde{R}^{-i})$ if $\zeta_3 = \texttt{PES}$
$\widetilde{R}$	$r_0 + \sum_{i=1}^{n} \widetilde{r}_i$
$\widetilde{\pi}_t^{\ddagger}$	seller's policy in Alg. 1 w.r.t. the reward function $\widetilde{R}$, generated by Alg. 3
$\widehat{V}_h^{t,\ddagger}(x_1;\widetilde{R})$	optimistic value by Alg. 3 w.r.t. $\widetilde{R}$ if all agents are untruthful
$\widehat{V}_h^{t,\ddagger}(x_1;\widetilde{R}^{-i})$	optimistic value by Alg. 3 w.r.t. $\widetilde{R}^{-i}$ if all agents are untruthful
$\widecheck{V}_h^{t,\ddagger}(x_1;\widetilde{R}^{-i})$	pessimistic value by Alg. 3 w.r.t. $\widetilde{R}^{-i}$ if all agents are untruthful
$\widehat{V}_h^{t,\widetilde{\pi}_t^{\ddagger}}(x_1;\widetilde{R}^{-i})$	optimistic value by Alg. 4 w.r.t. $\widetilde{R}^{-i}$, $\widetilde{\pi}_t^{\ddagger}$ if all agents are untruthful
$\widecheck{V}_h^{t,\widetilde{\pi}_t^{\ddagger}}(x_1;\widetilde{R}^{-i})$	pessimistic value by Alg. 4 w.r.t. $\widetilde{R}^{-i}$, $\widetilde{\pi}_t^{\ddagger}$ if all agents are untruthful
$F_t^{\ddagger,-i}$	$\widehat{V}_1^{t,\ddagger}(x_1;\widetilde{R}^{-i})$ if $\zeta_2 = \texttt{OPT}$; $\widecheck{V}_1^{t,\ddagger}(x_1;\widetilde{R}^{-i})$ if $\zeta_2 = \texttt{PES}$
$G_t^{\ddagger,-i}$	$\widehat{V}_1^{t,\widetilde{\pi}_t^{\ddagger}}(x_1;\widetilde{R}^{-i})$ if $\zeta_3 = \texttt{OPT}$; $\widecheck{V}_1^{t,\widetilde{\pi}_t^{\ddagger}}(x_1;\widetilde{R}^{-i})$ if $\zeta_3 = \texttt{PES}$

Appendix B. Proof of Lemma 2.1

Proof The detailed proof for these three properties can be found in Appendix B of Lyu et al. (2022). We include a sketch of the proof here for completeness. The proof for the linear Markov VCG mechanism's properties is provided as follows:

1. *Truthfulness*: We begin by noting that when agent i reports their rewards untruthfully, the untruthful reporting may change the optimal policy of $V_1^\pi(x_1; R)$ by altering only the reported value of r_i and the associated value function $V_1^\pi(; r_i)$. However, agent i cannot affect the value of $V_1^\pi(x_1; R^{-i})$, as R^{-i} is independent of r_i.

   With the previous observation in mind, let $\widetilde{r}_i$ be the untruthful value function reported by agent i and $\widetilde{\pi} = \mathrm{argmax}_{\pi\in\Pi} V_1^\pi(x_1; \widetilde{r}_i + R^{-i})$. Under the linear Markov VCG mechanism, agent i attains the following utility

   $$\widetilde{u}_i = V_1^{\widetilde{\pi}}(x_1; r_i) - V_1^{\pi_*^{-i}}(x_1; R^{-i}) + V_1^{\widetilde{\pi}}(x_1; R^{-i}) = V_1^{\widetilde{\pi}}(x_1; R) - V_1^{\pi_*^{-i}}(x_1; R^{-i}).$$

   Similarly, we know $u_i = V_1^{\pi_*}(x_1; R) - V_1^{\pi_*^{-i}}(x_1; R^{-i})$ when agent i reports truthfully. Since π_* is the maximizer of $V_1^\pi(x_1; R)$, we know $u_i \geq \widetilde{u}_i$, thus proving truthfulness.

2. *Individual Rationality*: For any agent i, their utility is given by

   $$\begin{aligned} u_{i*} &= V_1^{\pi_*}(x_1; r_i) - p_{i*} = V_1^{\pi_*}(x_1; R) - V_1^{\pi_*^{-i}}(x_1; R^{-i}) \\ &\geq V_1^{\pi_*^{-i}}(x_1; R) - V_1^{\pi_*^{-i}}(x_1; R^{-i}) = V_1^{\pi_*^{-i}}(x_1; r_i) \geq 0, \end{aligned} \tag{20}$$

   where we use the fact that $r_{i,h}(s, a) \geq 0$ for all $(i, h, s, a) \in [n] \times [H] \times \mathcal{S} \times \mathcal{A}$.

3. *Efficiency*: Under truthful reporting, the chosen policy π_* is the maximizer of the value-function of welfare $V_1^\pi(x_1; R)$ and hence is efficient.

This completes the proof. ∎

Appendix C. Proof of Theorems 4.1 and 4.2

We begin by introducing a crucial result that will be used throughout the rest of the section. This lemma presents the estimation errors of certain value functions by their corresponding optimistic or pessimistic value estimates. We refer readers to the table of notation in Section A for detailed definitions of the policies, rewards, and value functions in this lemma.

Lemma C.1 *For both when $\zeta_1 =$ `ETC` and when $\zeta_1 =$ `EWC`, let $\iota = \log(36ndHT/\delta)$. With probability at least $1 - \delta$, the following statements hold true jointly for all $t > K$ and some absolute constant $\hat{c}$.*

1. *Regardless of any agent's truthfulness, the policy used is sufficiently close to the one that maximizes the value functions of the* reported *reward functions. More specifically, $V_1^*(x_1; \mathfrak{R}) - V_1^\pi(x_1; \mathfrak{R}) \leq 2\hat{c}\sqrt{d^3H^6\iota/K}$ for all $(\mathfrak{R}, \pi) \in \{(R, \widehat{\pi}^t), (\widetilde{R}, \widetilde{\pi}_t^{\ddagger})\} \cup \{(r_i + \widetilde{R}^{-i}, \widetilde{\pi}_t^{\dagger i})\}_{i=1}^n$.*

2. *For all $i \in [n]$, Algorithm 3 returns a sufficiently good estimate regardless of agent i's or other agents' truthfulness. More specifically,* $0 \leq \widehat{V}_1^{t,\pi}(x_1; \mathfrak{R}) - V^*(x_1; \mathfrak{R}) \leq 2\hat{c}\sqrt{d^3 H^6 \iota / K}$ *and* $-2\hat{c}\sqrt{d^3 H^6 \iota / K} \leq \check{V}_1^{t,\pi}(x_1; \mathfrak{R}) - V^*(x_1; \mathfrak{R}) \leq 0$*, for all* $(\mathfrak{R}, \pi) \in \{(R^{-i}, \star), (\widetilde{R}^{-i}, \dagger), (\widetilde{R}^{-i}, \ddagger)\}_{i=1}^n$.

3. *For all $i \in [n]$, Algorithm 4 returns a sufficiently good estimate regardless of agent i's or other agents' truthfulness. More specifically,* $0 \leq \widehat{V}_1^{t,\pi}(x_1; \mathfrak{R}) - V_1^{\pi}(x_1; \mathfrak{R}) \leq 2\hat{c}\sqrt{d^3 H^6 \iota / K}$ *and* $-2\hat{c}\sqrt{d^3 H^6 \iota / K} \leq \check{V}_1^{t,\pi}(x_1; \mathfrak{R}) - V_1^{\pi}(x_1; \mathfrak{R})$*, for all* $(\mathfrak{R}, \pi) \in \{(R^{-i}, \widehat{\pi}^t), (\widetilde{R}^{-i}, \widetilde{\pi}_t^{\dagger i}), (\widetilde{R}^{-i}, \widetilde{\pi}_t^{\ddagger})\}_{i=1}^n$.

Please see Appendix D for the detailed proof. At the high level, the first clause ensures that the policy executed during exploitation is always sufficiently close to the one that maximizes the sum of the reported reward functions. The second and third clauses ensure that the price estimation is sufficiently good. With Lemma C.1, we can obtain the proofs of Theorems 4.1 and 4.2. For a concise presentation, we ignore presenting the probability for a certain inequality holds when calling Lemma C.1. Overall, the results in Theorem 4.1 and Theorem 4.2 will hold with probability at least $1 - \delta$ respectively, according to the above lemma.

C.1 Proof of Theorem 4.1

Proof We prove each bound in Theorem 4.1 separately. Overall, the inequalities in Lemma C.1 for the proof of Theorem 4.1 hold together with probability at least $1-\delta$. For conciseness, we ignore the detailed description of probabilities for each of these inequalities in our proof.

Welfare Regret. Recall that in Equation (5), the social welfare regret is defined as $\mathrm{Reg}_T^W = \sum_{t=1}^T \mathrm{reg}_t^W$ where $\mathrm{reg}_t^W = V_1^{\pi_*}(x_1; R) - V_1^{\widehat{\pi}^t}(x_1; R)$. We begin by decomposing the regret into two parts, the regret suffered in the exploration phase and the regret suffered in the exploitation phase, as follows,

$$\mathrm{Reg}_T^W = \sum_{t=1}^{K} \mathrm{reg}_t^W + \sum_{t=K+1}^{T} \mathrm{reg}_t^W. \tag{21}$$

For the first summation in Equation (21), we have

$$\sum_{t=1}^{K} \mathrm{reg}_t^W \leq KH(n + R_{\max}) = H(n + R_{\max})K, \tag{22}$$

recalling that $\mathrm{reg}_t^W \leq H(n + R_{\max})$ due to the upper bound of the reward functions.

We now turn to the second summation. By Lemma C.1, for $t > K$ we have

$$\mathrm{reg}_t^W = V_1^{\pi_*}(x_1; R) - V_1^{\widehat{\pi}^t}(x_1; R) \leq 2\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K}. \tag{23}$$

Summing the above equation form $t = K+1$ to T, we have

$$\sum_{t=K+1}^{T} \mathrm{reg}_t^W \leq 2\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K}(T - K). \tag{24}$$

Combining Equations (21), (22), and (24), we have

$$\mathrm{Reg}_t^W \le H(n+R_{\max})K + 2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}(T-K), \tag{25}$$

where the value of K will be determined by jointly considering the upper bounds of $n\mathrm{Reg}_T^W$, $\mathrm{Reg}_T^\sharp$, and Reg_{0T}.

Agent Regret. Recall that in Equation (5), the agent regret is defined as $\mathrm{Reg}_{iT} = \sum_{t=1}^T \mathrm{reg}_{it}$, where $\mathrm{reg}_{it} = u_{i*} - u_{it}$. Similar to our proof for welfare regret, we decompose the regret to that incurred during exploration and exploitation,

$$\mathrm{Reg}_{iT} = \sum_{t=1}^K \mathrm{reg}_{it} + \sum_{t=K+1}^T \mathrm{reg}_{it}. \tag{26}$$

For the first summation in Equation (26), we begin by upper bounding the instantaneous regret of agent i during the exploration phase. As the price charged to the agents is set to 0 during the exploration phase, for any $t \in [K]$, we have

$$\mathrm{reg}_{it} \le u_{i*} - \min_\pi V_1^\pi(x_1; r_i) \le u_{i*} = V_1^{\pi_*}(x_1; r_i) - p_{i*},$$

where we recall $p_{i*} = V_1^{\pi_*^{-i}}(x_1; R^{-i}) - V_1^{\pi_*}(x_1; R^{-i})$ and use the fact that $r_i \ge 0$. By definition of π_*^{-i}, we know that $p_{i*} \ge 0$ and $V_1^{\pi_*}(x_1; r_i) \le H$, using the fact that $r_i \le 1$. We then have

$$\sum_{t=1}^K \mathrm{reg}_{it} \le \sum_{t=1}^K V_1^{\pi_*}(x_1; r_i) \le HK.$$

Bounding the instantaneous agent regret during the exploitation phase is more complicated, as it depends on not only the suboptimality of the learned policy $\widehat{\pi}^t$ itself, but also the suboptimality incurred by estimation of the VCG price, $p_{it} = F_t^{-i} - G_t^{-i}$. To handle this challenge, we propose the following decomposition for $t > K$,

$$\begin{aligned}
\mathrm{reg}_{it} &= u_{i*} - u_{it} \\
&= \big[V_1^{\pi_*}\big(x_1; r_i\big) - V_1^{\pi_*^{-i}}(x_1; R^{-i}) + V_1^{\pi_*}(x_1; R^{-i})\big] - \big[V_1^{\widehat{\pi}^t}\big(x_1; r_i\big) - F_t^{-i} + G_t^{-i}\big] \\
&= \underbrace{\big[V_1^{\pi_*}(x_1; R) - V_1^{\widehat{\pi}^t}(x_1; R)\big]}_{\text{(i)}} + \underbrace{\big[F_t^{-i} - V_1^{\pi_*^{-i}}(x_1; R^{-i})\big]}_{\text{(ii)}} + \underbrace{\big[V_1^{\widehat{\pi}^t}(x_1; R^{-i}) - G_t^{-i}\big]}_{\text{(iii)}},
\end{aligned} \tag{27}$$

where the second equation uses the fact that $V_1^\pi\big(x_1; r_i\big) + V_1^\pi(x_1; R^{-i}) = V_1^\pi(x_1; R)$ for any π. The above decomposition allows us to bound the agent regret in terms of (i) suboptimality of $\widehat{\pi}^t$, (ii) estimation error of F_t^{-i}, and (iii) policy evaluation error of G_t^{-i}.

For term (i), by the result already obtained in Equation (23) for the welfare regret, we have for all $t > K$,

$$V_1^{\pi_*}(x_1; R) - V_1^{\widehat{\pi}^t}(x_1; R) \le 2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}.$$

We now bound term (ii). Let $\widehat{\pi}_t^{-i}$ be the fictitious policy generated by Algorithm 3 when calculating F_t^{-i}. For $t > K$, when $\zeta_2 = \texttt{PES}$, $F_t^{-i} = \check{V}_1^{t,\widehat{\pi}_t^{-i}}(x_1; R^{-i})$, we have

$$\text{(ii)} = \check{V}_1^{t,\widehat{\pi}_t^{-i}}(x_1; R^{-i}) - V_1^{\pi_*^{-i}}(x_1; R^{-i}) \le 0,$$

where the inequality is by Lemma C.1. When $\zeta_2 = \texttt{OPT}$, $F_t^{-i} = \widehat{V}_1^{t,\widehat{\pi}_t^{-i}}(x_1; R^{-i})$, we have

$$(\text{ii}) = \widehat{V}_1^{t,\widehat{\pi}_t^{-i}}(x_1; R^{-i}) - V_1^{\pi_*^{-i}}(x_1; R^{-i}) \leq 2\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K},$$

where the inequality also stems from Lemma C.1.

Term (iii) is controlled in a similar way. By Lemma C.1, for $t \geq K$, when $\zeta_3 = \texttt{OPT}$, $G_t^{-i} = \widehat{V}_1^{t,\widehat{\pi}^t}(x_1; R^{-i})$, we have

$$(\text{iii}) = V_1^{\widehat{\pi}^t}(x_1; R^{-i}) - \widehat{V}_1^{t,\widehat{\pi}^t}(x_1; R^{-i}) \leq 0,$$

and when $\zeta_3 = \texttt{PES}$, $G_t^{-i} = \check{V}_1^{t,\widehat{\pi}^t}(x_1; R^{-i})$, we have

$$(\text{iii}) = V_1^{\widehat{\pi}^t}(x_1; R^{-i}) - \check{V}_1^{t,\widehat{\pi}^t}(x_1; R^{-i}) \leq 2\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K}.$$

Combining the regrets incurred in both phases, by $\mathrm{Reg}_{iT} = \sum_{t=1}^T \mathrm{reg}_{it}$, we obtain

$$\mathrm{Reg}_{iT} \leq \begin{cases} HK + 2\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K}(T - K) & \text{if } (\zeta_2, \zeta_3) = (\texttt{PES}, \texttt{OPT}) \\ HK + 6\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K}(T - K) & \text{if } (\zeta_2, \zeta_3) = (\texttt{OPT}, \texttt{PES}). \end{cases} \tag{28}$$

Seller Regret. Recall that in Equation (5), the seller regret is defined as $\mathrm{Reg}_{0T} = \sum_{t=1}^T \mathrm{reg}_{0t}$ where $\mathrm{reg}_{0t} = u_{0*} - u_{0t}$. Thus, we have the following decomposition

$$\mathrm{Reg}_{0T} = \sum_{t=1}^K \mathrm{reg}_{0t} + \sum_{t=K+1}^T \mathrm{reg}_{0t}. \tag{29}$$

We begin with bounding the first summation. Recall that $\mathrm{reg}_{0t} = u_{0*} - u_{0t}$. During exploration, as the seller charges a price of 0 to all agents, their utility is lower bounded by $u_{0t} = \min_\pi V^\pi(x_1; r_0) + 0 = \min_\pi V^\pi(x_1; r_0)$. As $r_0 \geq 0$, we know that for all $t \in [K]$,

$$\sum_{t=1}^K \mathrm{reg}_{0t} \leq \sum_{t=1}^K u_{0*} \leq K u_{0*}.$$

Recall that

$$\begin{aligned} u_{0*} &= V_1^{\pi_*}(x_1; r_0) + \sum_{i=1}^n p_{i*} = V_1^{\pi_*}(x_1; r_0) + \sum_{i=1}^n \Big(V^{\pi_*^{-i}}(x_1; R^{-i}) - V^{\pi_*}(x_1; R^{-i}) \Big) \\ &= -(n-1)V_1^{\pi_*}(x_1; R) + \sum_{i=1}^n V^{\pi_*^{-i}}(x_1; R^{-i}). \end{aligned}$$

Since $r_i \geq 0$, $R = R^{-i} + r_i \geq R^{-i}$, we have

$$\begin{aligned} u_{0*} &\leq -(n-1)V_1^{\pi_*}(x_1; R) + \sum_{i=1}^n V^{\pi_*^{-i}}(x_1; R) \leq -(n-1)V_1^{\pi_*}(x_1; R) + \sum_{i=1}^n V^{\pi_*}(x_1; R) \\ &= V^{\pi_*}(x_1; R) \leq H(n + R_{\max}), \end{aligned}$$

according to the definitions of π_* and π_*^{-i}. We then have the following upper bound for the first summation in Equation (29) as

$$\sum_{t=1}^{K} \mathrm{reg}_{0t} \leq K u_{0*} \leq (n + R_{\max}) H K.$$

We now bound the second summation in Equation (29). The seller's instantaneous regret during exploration can be decomposed as

$$\begin{aligned}
\mathrm{reg}_{0t} &= u_{0*} - u_{0t} \\
&= \left[V_1^{\pi_*}(x_1; r_0) + \sum_{i=1}^{n} p_{i*} \right] - \left[V_1^{\widehat{\pi}^t}(x_1; r_0) + \sum_{i=1}^{n} p_{it} \right] \\
&= \left[V_1^{\pi_*}(x_1; r_0) + \sum_{i=1}^{n} \left[V_1^{\pi_*^{-i}}(x_1; R^{-i}) - V_1^{\pi_*}(x_1; R^{-i}) \right] \right] - \left[V_1^{\widehat{\pi}^t}(x_1; r_0) + \sum_{i=1}^{n} \left(F_t^{-i} - G_t^{-i} \right) \right] \\
&= \left[-(n-1) V_1^{\pi_*}(x_1; R) + \sum_{i=1}^{n} V_1^{\pi_*^{-i}}(x_1; R^{-i}) \right] \\
&\qquad - \left[-(n-1) V_1^{\widehat{\pi}^t}(x_1; R) + \sum_{i=1}^{n} \left[F_t^{-i} - G_t^{-i} + V_1^{\widehat{\pi}^t}(x_1; R^{-i}) \right] \right] \\
&= (n-1) \underbrace{\left[V_1^{\widehat{\pi}^t}(x_1; R) - V_1^{\pi_*}(x_1; R) \right]}_{\text{(i)}} + \sum_{i=1}^{n} \underbrace{\left[V_1^{\pi_*^{-i}}(x_1; R^{-i}) - F_t^{-i} \right]}_{\text{(ii)}} + \sum_{i=1}^{n} \underbrace{\left[G_t^{-i} - V^{\widehat{\pi}^t}(x_1; R^{-i}) \right]}_{\text{(iii)}}.
\end{aligned}$$

For term (i), we have (i) ≤ 0 due to the optimality of V_1^*. For term (ii), when $\zeta_2 = \texttt{OPT}$, by the construction of F_t^{-i}, we have

$$\text{(ii)} = V_1^{\pi_*^{-i}}(x_1; R^{-i}) - F_t^{-i} = V_1^{\pi_*^{-i}}(x_1; R^{-i}) - \widehat{V}_1^{t, \widehat{\pi}_t^{-i}}(x_1; R^{-i}) \leq 0,$$

where we invoke Lemma C.1 for the inequality. When $\zeta_2 = \texttt{PES}$, we obtain that

$$\text{(ii)} = V_1^{\pi_*^{-i}}(x_1; R^{-i}) - F_t^{-i} = V_1^{\pi_*^{-i}}(x_1; R^{-i}) - \check{V}_1^{t, \widehat{\pi}_t^{-i}}(x_1; R^{-i}) \leq 2\hat{c}(n + R_{\max}) \sqrt{d^3 H^6 \iota / K},$$

where the last inequality also uses Lemma C.1.

For term (iii), further invoking Lemma C.1, when $\zeta_3 = \texttt{PES}$, (iii) ≤ 0, and when $\zeta_3 = \texttt{OPT}$, we have

$$\text{(iii)} = \widehat{V}_1^{t, \widehat{\pi}^t}(x_1; R^{-i}) - V_1^{\widehat{\pi}^t}\big(x_1; R^{-i}\big) \leq 2\hat{c}(n + R_{\max}) \sqrt{d^3 H^6 \iota / K}.$$

Combining the bounds for terms (i), (ii), and (iii) above, we have

$$\sum_{t=K+1}^{T} \mathrm{reg}_{0t} \leq \begin{cases} 4\hat{c} n (n + R_{\max}) \sqrt{d^3 H^6 \iota / K} (T - K) & \text{if } (\zeta_2, \zeta_3) = (\texttt{PES}, \texttt{OPT}) \\ 0 & \text{if } (\zeta_2, \zeta_3) = (\texttt{OPT}, \texttt{PES}), \end{cases}$$

where $\hat{c}$ is some absolute constant. By adding the regret incurred in the exploration phase, this result further gives the upper bound of the seller regret Reg_{0T} as

$$\begin{cases} H(n + R_{\max})K + 4\hat{c} n (n + R_{\max}) \sqrt{d^3 H^6 \iota / K} (T - K) & \text{if } (\zeta_2, \zeta_3) = (\texttt{PES}, \texttt{OPT}) \\ H(n + R_{\max})K & \text{if } (\zeta_2, \zeta_3) = (\texttt{OPT}, \texttt{PES}). \end{cases} \tag{30}$$

Choice of K. Now we determine the value of K that can lead to a tight bound of $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}$, where $\mathrm{Reg}_T^\sharp = \sum_{i=1}^n \mathrm{Reg}_{iT}$ as defined in Equation (5). According to Equations (25), (28), and (30), comparing the upper bounds of $n\mathrm{Reg}_T^W$, $\mathrm{Reg}_T^\sharp$, and Reg_{0T}, we always have

$$\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\} \leq H(n+R_{\max})nK + 6\hat{c}(n+R_{\max})n\sqrt{d^3H^6\iota/K}(T-K).$$

Focusing on the factors of H, n, d, T, and ι, we set $K = dH^{4/3}\iota^{1/3}T^{2/3}$, which can minimize the order of these factors in the above inequality, and obtain the bound

$$\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\} = \mathcal{O}\big(n(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}\big).$$

Thus, plugging $K = dH^{4/3}\iota^{1/3}T^{2/3}$ into (25), we have

$$\mathrm{Reg}_T^W \leq (1+2\hat{c})(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}.$$

Plugging the value of K into (28), we have

$$\mathrm{Reg}_{iT} \leq \begin{cases} \big(1+2\hat{c}(n+R_{\max})\big)dH^{7/3}\iota^{1/3}T^{2/3} & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT}) \\ \big(1+6\hat{c}(n+R_{\max})\big)dH^{7/3}\iota^{1/3}T^{2/3} & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES}). \end{cases}$$

Plugging the value of K into (30), we obtain

$$\mathrm{Reg}_{0T} \leq \begin{cases} (1+4\hat{c}n)(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3} & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT}) \\ (n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3} & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES}). \end{cases}$$

This completes the proof of the upper bounds of the welfare regret, the agent regret, and the seller regret.

Individual Rationality. We note that for the proof of individual rationality, we do not require the truthfulness of agents other than agent i. Recall that if we do not charge the agents in the exploration phase, for any agent i, we always have utility $u_{it} \geq 0$ during exploration because $r_i \geq 0$. Thus, we only need to bound from below agent i's utility during the exploitation phase. When agent i reports according to the reward function r_i but other agents report rewards potentially untruthfully according to $\widetilde{r}_j$ for $j \neq i$, we define $\widetilde{R}^{-i} := r_0 + \sum_{j\in[n],i\neq j}\widetilde{r}_j$ and let $\widetilde{\pi}_t^{\dagger i}$ substitute $\widehat{\pi}^t$ in Algorithm 1, which is generated by Algorithm 3 in the current reward setting. We further define the associated F and G generated by Algorithms 3 and 4 respectively as follows

$$F_t^{\dagger,-i} = \begin{cases} \widehat{V}_1^{t,\dagger}\big(x_1;\widetilde{R}^{-i}\big) & \text{if } \zeta_2 = \texttt{OPT} \\ \widecheck{V}_1^{t,\dagger}\big(x_1;\widetilde{R}^{-i}\big) & \text{if } \zeta_2 = \texttt{PES}, \end{cases} \qquad G_t^{\dagger,-i} = \begin{cases} \widehat{V}_1^{t,\widetilde{\pi}_t^{\dagger i}}\big(x_1;\widetilde{R}^{-i}\big) & \text{if } \zeta_3 = \texttt{OPT} \\ \widecheck{V}_1^{t,\widetilde{\pi}_t^{\dagger i}}\big(x_1;\widetilde{R}^{-i}\big) & \text{if } \zeta_3 = \texttt{PES}. \end{cases} \tag{31}$$

For all $t > K$, according to the definition of u_{it}, under the current reward setting, we have

$$\begin{aligned} u_{it} &= V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1;r_i) - p_{it}^\dagger \\ &= V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1;r_i\big) - F_t^{\dagger,-i} + G_t^{\dagger,-i} \\ &= \big[V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1;r_i\big) + V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1;\widetilde{R}^{-i}\big) - F_t^{\dagger,-i}\big] + \big[G_t^{\dagger,-i} - V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1;\widetilde{R}^{-i}\big)\big] \\ &= \underbrace{\big[V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1;r_i+\widetilde{R}^{-i}\big) - F_t^{\dagger,-i}\big]}_{\text{(i)}} + \underbrace{\big[G_t^{\dagger,-i} - V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1;\widetilde{R}^{-i}\big)\big]}_{\text{(ii)}}, \end{aligned} \tag{32}$$

where $p_{it}^{\dagger} = F_t^{\dagger,-i} - G_t^{\dagger,-i}$. For term (i) in Equation (32), by the definition of $V_1^*(x_1, r) := \max_\pi V_1^\pi(x_1, r)$ for any r, we have

$$\begin{aligned}
\text{(i)} &\geq V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1; r_i + \widetilde{R}^{-i}\big) - \widehat{V}_1^{t,\dagger}\big(x_1; \widetilde{R}^{-i}\big) \\
&= \underbrace{\big[V_1^*(x_1; r_i + \widetilde{R}^{-i}) - V_1^*(x_1; \widetilde{R}^{-i})\big]}_{\text{(i.a)}} + \underbrace{\big[V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1; r_i + \widetilde{R}^{-i}\big) - V_1^*(x_1; r_i + \widetilde{R}^{-i})\big]}_{\text{(i.b)}} \\
&\quad + \underbrace{\big[V_1^*(x_1; \widetilde{R}^{-i}) - \widehat{V}_1^{t,\dagger}\big(x_1; \widetilde{R}^{-i}\big)\big]}_{\text{(i.c)}},
\end{aligned}$$

where the first inequality stems from the fact that $F_t^{\dagger,-i}$ is always at most $\widehat{V}_1^{t,\dagger}\big(x_1; \widetilde{R}^{-i}\big)$ regardless of the choice of ζ_2 shown in Equation (31). For (i.a), we have that

$$\text{(i.a)} = \max_\pi V_1^\pi(x_1; r_i + \widetilde{R}^{-i}) - \max_\pi V_1^\pi(x_1; \widetilde{R}^{-i}).$$

Note that for any π, we have $V_1^\pi(x_1; r_i + \widetilde{R}^{-i}) \geq V_1^\pi(x_1; \widetilde{R}^{-i})$ since $r_i \geq 0$, which implies that $\max_\pi V_1^\pi(x_1; r_i + \widetilde{R}^{-i}) \geq V_1^\pi(x_1; \widetilde{R}^{-i})$ holds for any π. Taking maximum on the right-hand side further gives $\max_\pi V_1^\pi(x_1; r_i + \widetilde{R}^{-i}) \geq \max_\pi V_1^\pi(x_1; \widetilde{R}^{-i})$. We then have that (i.a) ≥ 0. Moreover, (i.b) is the suboptimality of policy $\widetilde{\pi}_t^{\dagger i}$ and (i.c) is the estimation error of $V_1^*(x_1; \widetilde{R}^{-i})$ by $\widehat{V}_1^{t,\dagger}\big(x_1; \widetilde{R}^{-i}\big)$, which can be bounded below by $-2\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K}$ respectively invoking Lemma C.1. We can then bound term (i) from below by $-4\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K}$.

For term (ii) in Equation (32), observe that $G_t^{\dagger,-i}$ is always at least $\check{V}_1^{t,\widetilde{\pi}_t^{\dagger i}}\big(x_1; \widetilde{R}^{-i}\big)$ regardless of the choice of ζ_3 shown in Equation (31) and thus we have by Lemma C.1 that

$$\text{(ii)} \geq \check{V}_1^{t,\widetilde{\pi}_t^{\dagger i}}\big(x_1; \widetilde{R}^{-i}\big) - V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1; \widetilde{R}^{-i}\big) \geq -2\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K},$$

for some absolute constant $\hat{c}$. Summing (i) and (ii) from $t = 1$ to T, we get

$$U_{iT} \geq \sum_{t=K+1}^{T} u_{it} \geq -6\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K},$$

Setting $K = dH^{4/3}\iota^{1/3}T^{2/3}$ in the above inequality, we further get,

$$U_{iT} \geq -6\hat{c}(n + R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}$$

which implies the mechanism we learned is $6\hat{c}(n + R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}$-approximately individually rational.

Truthfulness. We consider two cases for our proof of truthfulness: (1) agent i reports truthfully, and others may report untruthfully (2) all agents may report untruthfully. Then we denote by r_i the truthful reward and $\widetilde{r}_i$ the potentially untruthful reward. For case (1), we adopt the same notations $F_t^{\dagger,-i}, G_t^{\dagger,-i}$, $\widetilde{\pi}_t^{\dagger i}$, and $u_{it} = V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1; r_i) - p_{it}^{\dagger}$ as in the above proof of individual rationality. For case (2), we let $\widetilde{\pi}_t^{\ddagger}$ be the learned policy for the seller under the reward $\widetilde{R} := r_0 + \sum_{i=1}^{n} \widetilde{r}_i$ in Algorithm 1, $F_t^{\ddagger,-i}$ and $G_t^{\ddagger,-i}$ be the associated F and G functions, and $\widetilde{u}_{it} = V_1^{\widetilde{\pi}_t^{\ddagger}}(x_1; r_i) - p_{it}^{\ddagger}$ with $p_{it}^{\ddagger} = F_t^{\ddagger,-i} - G_t^{\ddagger,-i}$ generated by Algorithms 3 and 4 respectively. Let $\widetilde{U}_{iT} = \sum_{t=1}^{T} \widetilde{u}_{it}$ and $U_{iT} = \sum_{t=1}^{T} u_{it}$. The surplus in utility the agent gains from untruthful reporting is then

$$\widetilde{U}_{iT} - U_{iT} = \sum_{t=1}^{T} \left(\widetilde{u}_{it} - u_{it} \right). \tag{33}$$

We decompose the summation in terms of the exploration and exploitation phases. When $t \le K$, the agents are not charged any price, and then $r_i \ge 0$ ensures $u_{it} \ge 0$. We then have

$$\widetilde{u}_{it} - u_{it} \le \widetilde{u}_{it} \le \max_{\pi} V_1^{\pi}(x_1; r_i) \le H,$$

where the second inequality uses the fact that the price is 0.

We now consider the case when $t > K$. We explicitly define $F_t^{\ddagger,-i}$ and $G_t^{\ddagger,-i}$ as follows

$$F_t^{\ddagger,-i} = \begin{cases} \widehat{V}_1^{t,\ddagger}\big(x_1; \widetilde{R}^{-i}\big) & \text{if } \zeta_2 = \texttt{OPT} \\ \check{V}_1^{t,\ddagger}\big(x_1; \widetilde{R}^{-i}\big) & \text{if } \zeta_2 = \texttt{PES}, \end{cases} \qquad G_t^{\dagger,-i} = \begin{cases} \widehat{V}_1^{t,\widetilde{\pi}_t^{\ddagger}}\big(x_1; \widetilde{R}^{-i}\big) & \text{if } \zeta_3 = \texttt{OPT} \\ \check{V}_1^{t,\widetilde{\pi}_t^{\ddagger}}\big(x_1; \widetilde{R}^{-i}\big) & \text{if } \zeta_3 = \texttt{PES}, \end{cases} \tag{34}$$

where the value functions are generated by Algorithms 3 and 4 respectively based on the untruthfully reported rewards by all agents.

For any $t > K$, we have

$$\widetilde{u}_{it} - u_{it} = \big[V_1^{\widetilde{\pi}_t^{\ddagger}}(x_1; r_i) - F_t^{\ddagger,-i} + G_t^{\ddagger,-i}\big] - \big[V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1; r_i) - F_t^{\dagger,-i} + G_t^{\dagger,-i}\big].$$

We first show that $F_t^{\dagger,-i} = F_t^{\ddagger,-i}$. Recall that when $\zeta_1 = \texttt{ETC}$, both $F_t^{\dagger,-i}$ and $F_t^{\ddagger,-i}$ are calculated using only data collected during the exploration phase. As the data collection policy is given by a reward-free exploration algorithm, namely Algorithm 2, the trajectories collected remain the same whether agent i is truthful or not. Additionally, both $F_t^{\dagger,-i}$ and $F_t^{\ddagger,-i}$ are given by Algorithm 3, which only uses the rewards reported by other agents. In other words, the input data used to calculate $F_t^{\dagger,-i}$ and $F_t^{\ddagger,-i}$ are exactly the same, irregardless of the truthfulness of agent i. Conditionally on the K trajectories collected during the exploration phase, the two functions $F_t^{\dagger,-i}$ and $F_t^{\ddagger,-i}$ equal to each other and cancel out. We then obtain that for all $t > K$,

$$\begin{aligned} &\widetilde{u}_{it} - u_{it} \\ &\quad = V_1^{\widetilde{\pi}_t^{\ddagger}}(x_1; r_i) + G_t^{\ddagger,-i} - V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1; r_i) - G_t^{\dagger,-i} \\ &\quad = \underbrace{\big[V_1^{\widetilde{\pi}_t^{\ddagger}}\big(x_1; r_i + \widetilde{R}^{-i}\big) - V_1^{*}\big(x_1; r_i + \widetilde{R}^{-i}\big)\big]}_{\text{(i)}} + \underbrace{\big[V_1^{*}\big(x_1; r_i + \widetilde{R}^{-i}\big) - V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1; r_i + \widetilde{R}^{-i}\big)\big]}_{\text{(ii)}} \\ &\qquad + \underbrace{\big[G_t^{\ddagger,-i} - V_1^{\widetilde{\pi}_t^{\ddagger}}\big(x_1; \widetilde{R}^{-i}\big)\big]}_{\text{(iii)}} + \underbrace{\big[V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1; \widetilde{R}^{-i}\big) - G_t^{\dagger,-i}\big]}_{\text{(iv)}}. \end{aligned}$$

Here, term (i) ≤ 0 is due to the definition of $V_1^*(x_1; r_i + \widetilde{R}^{-i}) = \max_\pi V_1^\pi(x_1; r_i + \widetilde{R}^{-i})$. Term (ii) is the suboptimality of policy $\widetilde{\pi}_t^{\dagger i}$, term (iii) and term (iv) are policy evaluation errors for policy $\widetilde{\pi}_t^{\ddagger}$ and $\widetilde{\pi}_t^{\dagger i}$. Using Lemma C.1, term (ii) is upper bounded by $2\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K}$. We then consider terms (iii) and (iv). When $\zeta_3 = \texttt{OPT}$, we have (iii) $\leq 2\hat{c}(n+R_{\max})\sqrt{d^3 H^6 \iota / K}$ while (iv) ≤ 0. Similarly, we have (iii) ≤ 0 and (iv) $\leq 2\hat{c}\sqrt{d^3 H^6 \iota / K}$ when $\zeta_3 = \texttt{PES}$. In summary, regardless of the choices for ζ_2, ζ_3, we always have for all i, t

$$\widetilde{u}_{it} - u_{it} \leq \begin{cases} H & \text{if } t \in [K] \\ 4\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K} & \text{if } t > K. \end{cases}$$

Now we have obtained the upper bounds of $\widetilde{u}_{it} - u_{it}^{t-1}$ for both when $t \in [K]$ and when $t > K$. Summing $\widetilde{u}_{it} - u_{it}$ from $t = 1$ to T, we get

$$\widetilde{U}_{iT} - U_{iT} \leq HK + 4\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K}(T - K).$$

Setting $K = dH^{4/3}\iota^{1/3}T^{2/3}$ in the above inequality, we further obtain

$$\widetilde{U}_{iT} - U_{iT} \leq dH^{7/3}\iota^{1/3}T^{2/3} + 4\hat{c}(n + R_{\max})dH^{7/3}\iota^{1/3}T^{2/3},$$

which implies that the learned mechanism is $\big(1 + 4\hat{c}(n + R_{\max})\big)dH^{7/3}\iota^{1/3}T^{2/3}$-approximately truthful. This completes the proof. ■

C.2 Proof of Theorem 4.2

Proof We now prove each result separately in Theorem 4.2. The concentration inequalities for the proof of Theorem 4.2 jointly hold with probability at least $1 - \delta$. We ignore the detailed description of probabilities in our proof for conciseness.

Welfare Regret. When setting $\zeta_1 = \texttt{EWC}$, we can decompose the regret into two parts, the regret incurred in the exploration phase and the regret incurred in the exploitation phase as

$$\mathrm{Reg}_T^W = \sum_{t=1}^{K} \mathrm{reg}_t^W + \sum_{t=K+1}^{T} \mathrm{reg}_t^W.$$

Then we can bound the first summation as $\sum_{t=1}^K \mathrm{reg}_t^W \leq H(n + R_{\max})K$ using the same technique for obtaining Equation (22). For the second part, we have

$$\sum_{t=K+1}^{T} \mathrm{reg}_t^W = \sum_{t=K+1}^{T} \big[V_1^{\pi_*}(x_1; R) - V_1^{\widehat{\pi}^t}(x_1; R)\big].$$

Notice that during the exploitation phase, the welfare regret of Algorithm 1 when $\zeta_1 = \texttt{EWC}$ is the well-studied regret bound for LSVI-UCB, derived in Jin et al. (2020b). For integrity, we sketch out the proof below and refer interested readers to the detailed proofs in Jin et al. (2020b).

Following standard decomposition (see Lemmas B.5 and B.6 in Jin et al. (2020b), for instance), we have

$$\begin{aligned}\sum_{t=K+1}^{T} \mathrm{reg}_t^W &= \sum_{t=K+1}^{T} \Big[V_1^{\pi_*}(x_1;R) - V_1^{\widehat{\pi}^t}(x_1;R)\Big] \le \sum_{t=K+1}^{T} \Big[V_1^{t}(x_1;R) - V_1^{\widehat{\pi}^t}(x_1;R)\Big] \\ &\le \underbrace{\sum_{t=K+1}^{T}\sum_{h=1}^{H}\big(\mathbb{E}\big[\xi_h^t \,\big|\, x_{h-1}^t, a_{h-1}^t\big] - \xi_h^t\big)}_{\text{(i)}} \\ &\quad + 2\beta \underbrace{\sum_{t=K+1}^{T}\sum_{h=1}^{H}\sqrt{\big(\phi(x_h^t,a_h^t)\big)^\top \big(\Lambda_h^t\big)^{-1}\big(\phi(x_h^t,a_h^t)\big)}}_{\text{(ii)}},\end{aligned} \tag{35}$$

where $\xi_h^t = V_h^t\big(x_h^t;R\big) - V_h^{\widehat{\pi}_t}\big(x_h^t;R\big)$. Then, we bound terms (i) and (ii) in Equation (35) respectively. For term (i), since the computation of $\widehat{V}_h^t$ does not use the new observation x_h^t at rounds t, the terms in term (i) is a martingale difference sequence bounded by $2(n+R_{max})H$. Then we can bound it by Azuma-Hoeffding inequality and get an $\mathcal{O}\big((n+R_{\max})H\iota T^{1/2}\big)$ upper bound for term (i) in Equation (35). We provide the details as follows: for any $\nu > 0$, we have

$$\mathbb{P}\bigg(\sum_{t=K+1}^{T}\sum_{h=1}^{H}\big(\mathbb{E}\big[\xi_h^t \,\big|\, x_{h-1}^t, a_{h-1}^t\big] - \xi_h^t\big) \ge \nu\bigg) \le \exp\bigg\{\frac{-\nu^2}{2(n+R_{\max})^2H^2(T-K)}\bigg\}.$$

Hence, with high probability, we have

$$\begin{aligned}\sum_{t=K+1}^{T}\sum_{h=1}^{H}\big(\mathbb{E}\big[\xi_h^t \,\big|\, x_{h-1}^t, a_{h-1}^t\big] - \xi_h^t\big) &\le \sqrt{2(n+R_{\max})^2H^2(T-K)\log(2/\delta)} \\ &\le 2(n+R_{\max})\sqrt{H^2(T-K)\iota},\end{aligned} \tag{36}$$

where $\iota = \log(36ndHT/\delta)$. For term (ii), we can bound it using Lemma F.2 and Cauchy-Schwarz inequality,

$$\begin{aligned}\sum_{t=K+1}^{T}\sum_{h=1}^{H}\sqrt{\big(\phi(x_h^t,a_h^t)\big)^\top \big(\Lambda_h^t\big)\big(\phi(x_h^t,a_h^t)\big)} &\le \sum_{t=K+1}^{T}\sum_{h=1}^{H}\sqrt{\big(\phi(x_h^t,a_h^t)\big)^\top \big(\widetilde{\Lambda}_h^t\big)^{-1}\big(\phi(x_h^t,a_h^t)\big)} \\ &\le \sum_{h=1}^{H}\bigg[\sum_{t=K+1}^{T}\phi(x_h^t,a_h^t)^\top(\widetilde{\Lambda}_h^t)^{-1}\phi(x_h^t,a_h^t)\bigg]^{1/2} \\ &\le 2\sqrt{2dH^2(T-K)\iota},\end{aligned} \tag{37}$$

where $\widetilde{\Lambda}_h^t = \sum_{\tau=K+1}^{t-1}\phi(x_h^\tau,a_h^\tau)\phi(x_h^\tau,a_h^\tau)^\top + \lambda I$ is the design matrix only using the data in the exploitation phase. The first step is due to $\widetilde{\Lambda}_h^t \preceq \Lambda_h^t$, the second step is by Cauchy-Schwartz inequality, and the last step uses the elliptical potential lemma in Abbasi-Yadkori et al. (2011). Combining Equations (35), (36) and (37), with the setting of $\beta = \hat{c}(n + R_{\max})dH\sqrt{\iota}$ where $\iota = \log(36ndHT/\delta)$, we have the following upper bound

$$\begin{aligned}\sum_{t=K+1}^{T} \mathrm{reg}_t^W &\leq 2(n+R_{\max})\sqrt{H^2(T-K)\iota} + 2\beta\sqrt{2dH^2(T-K)\iota} \\ &\leq 2(n+R_{\max})\sqrt{H^2(T-K)\iota} + 4\hat{c}(n+R_{\max})\sqrt{d^3H^4(T-K)\iota^2} \\ &\leq 6\hat{c}(n+R_{\max})\sqrt{d^3H^4(T-K)\iota^2}.\end{aligned}$$

Combining the above inequality with the upper bound for $\sum_{t=1}^{K} \mathrm{reg}_t^W$, we have the upper bound of the welfare regret as

$$\mathrm{Reg}_T^W \leq (n+R_{\max})HK + 6\hat{c}(n+R_{\max})\sqrt{d^3H^4(T-K)\iota^2}, \tag{38}$$

where the value of K will be determined by jointly considering the upper bounds of $n\mathrm{Reg}_T^W$, $\mathrm{Reg}_T^\sharp$, and Reg_{0T}.

Agent Regret. For agent i's regret incurred during the exploration phase, we know from Section C.1 that it is bounded as $\sum_{t=1}^{K} \mathrm{reg}_{it} \leq HK$. We now focus on when $t > K$. According Equation (27), we have that

$$\begin{aligned}\mathrm{reg}_{it} &= u_{i*} - u_{it} \\ &= \underbrace{\left[V_1^{\pi_*}(x_1;R) - V_1^{\widehat{\pi}^t}(x_1;R)\right]}_{\text{(i)}} + \underbrace{\left[F_t^{-i} - V_1^{\pi_*^{-i}}(x_1;R^{-i})\right]}_{\text{(ii)}} + \underbrace{\left[V_1^{\widehat{\pi}^t}(x_1;R^{-i}) - G_t^{-i}\right]}_{\text{(iii)}}.\end{aligned} \tag{39}$$

Term (i) is the welfare regret, term (ii) is the function evaluation and policy estimation errors for F_t^{-i}, and term (iii) is the function evaluation error for G_t^{-i}. Recalling that the welfare regret bound above, we know that the summation of (i) from $t = K+1$ to T can be bounded as

$$\sum_{t=K+1}^{T} \left[V_1^{\pi_*}(x_1;R) - V_1^{\widehat{\pi}^t}(x_1;R)\right] \leq 6\hat{c}(n+R_{\max})\sqrt{d^3H^4(T-K)\iota^2}.$$

Our bounds for terms (ii) and (iii) use similar techniques for the case when $\zeta_1 = \texttt{ETC}$. Let $\widehat{\pi}_t^{-i}$ be the fictitious policy returned by Algorithm 3 when we compute F_t^{-i}. We obtain that

$$\text{(ii)} = \widehat{V}_1^{t,\widehat{\pi}_t^{-i}} - V_1^{\pi_*^{-i}}(x_1;R^{-i}) \leq 2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K},$$

when $\zeta_2 = \texttt{OPT}$, using Lemma C.1. Similarly, we know (ii) ≤ 0 when $\zeta_2 = \texttt{PES}$.

Finally, by Lemma C.1, we know (iii) $\leq 2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}$ when $\zeta_3 = \texttt{PES}$ and (iii) ≤ 0 when $\zeta_3 = \texttt{OPT}$. Combining the bounds for terms (i), (ii), and (iii) in both phases, we have the upper bound of the agent regret Reg_{iT} as follows:

If $(\zeta_2, \zeta_3) = (\texttt{PES}, \texttt{OPT})$, then

$$\mathrm{Reg}_{iT} \leq HK + 6\hat{c}(n+R_{\max})\sqrt{d^3H^4(T-K)\iota^2}. \tag{40}$$

If $(\zeta_2, \zeta_3) = (\texttt{OPT}, \texttt{PES})$, then

$$\mathrm{Reg}_{iT} \leq HK + 6\hat{c}(n+R_{\max})\sqrt{d^3H^4(T-K)\iota^2} + 4\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}(T-K). \tag{41}$$

Seller Regret. Similar to our proof of agent regret, from Section C.1, we first have

$$\sum_{t=1}^{K} \mathrm{reg}_{0t} \le H(n+R_{\max})K.$$

In addition, the exploration regret can be decomposed as

$$\begin{aligned}\mathrm{reg}_{0t} &= u_{0*} - u_{0t}\\ &= (n-1)\big[V_1^{\widehat{\pi}^t}(x_1;R) - V_1^{\pi_*}(x_1;R)\big] + \sum_{i=1}^{n}\big[V_1^{\pi_*^{-i}}(x_1;R^{-i}) - F_t^{-i}\big]\\ &\quad + \sum_{i=1}^{n}\big[G_t^{-i} - V^{\widehat{\pi}^t}(x_1;R^{-i})\big]\\ &\le \sum_{i=1}^{n}\underbrace{\big[V_1^{\pi_*^{-i}}(x_1;R^{-i}) - F_t^{-i}\big]}_{\text{(i)}} + \sum_{i=1}^{n}\underbrace{\big[G_t^{-i} - V^{\widehat{\pi}^t}(x_1;R^{-i})\big]}_{\text{(ii)}},\end{aligned}$$

where the second equation directly follows the decomposition proven in Section C.1 and the inequality comes from the definition of π_*, which is then used to eliminate the first term.

Similar to our proof for agent regret, invoking Lemma C.1 we immediately know that (i) $\le 2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}$ when $\zeta_2 = \texttt{PES}$, and (i) ≤ 0 when $\zeta_2 = \texttt{OPT}$. Also by Lemma C.1, we have (ii) ≤ 0 when $\zeta_3 = \texttt{PES}$, and (ii) $\le 2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}$ when $\zeta_3 = \texttt{OPT}$. Summing both (i) and (ii) over $i \in [n]$ and then summing the regrets incurred in both exploration and exploitation phases, we have the upper bound of the seller regret Reg_{0T} as

$$\begin{cases}(n+R_{\max})HK + 4\hat{c}n(n+R_{\max})\sqrt{d^3H^6\iota/K}(T-K) & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT})\\ (n+R_{\max})HK & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES}).\end{cases} \tag{42}$$

Choice of K. We determine the value of K that can lead to a tight bound of $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}$, where $\mathrm{Reg}_T^\sharp = \sum_{i=1}^n \mathrm{Reg}_{iT}$. According to Equations (38), (40), (41), and (42), comparing the upper bounds of $n\mathrm{Reg}_T^W$, $\mathrm{Reg}_T^\sharp$, and Reg_{0T}, we always have

$$\begin{aligned}\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\} &\le n(n+R_{\max})HK + 6\hat{c}n(n+R_{\max})\sqrt{d^3H^4(T-K)\iota^2}\\ &\quad + 4\hat{c}n(n+R_{\max})\sqrt{d^3H^6\iota/K}(T-K)\\ &\le n(n+R_{\max})HK + 6\hat{c}n(n+R_{\max})\sqrt{d^3H^4T\iota^2} + 4\hat{c}n(n+R_{\max})\sqrt{d^3H^6\iota/K}T.\end{aligned}$$

Focusing on the factors of H, n, d, T, and ι, we set $K = dH^{4/3}\iota^{1/3}T^{2/3}$, which can minimize the order of these factors in the above inequality, and obtain the bound

$$\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\} = \mathcal{O}\big(n(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}\big).$$

Thus, plugging $K = dH^{4/3}\iota^{1/3}T^{2/3}$ into (38), we have

$$\mathrm{Reg}_T^W \le (n+R_{\max})(dH^{7/3}\iota^{1/3}T^{2/3} + 6\hat{c}d^{3/2}H^2\iota T^{1/2}).$$

Plugging the value of K into (40) and (41), we have that Reg_{iT} can be bounded by

$$\begin{cases} dH^{7/3}\iota^{1/3}T^{2/3} + 6\hat{c}(n+R_{\max})d^{3/2}H^2\iota T^{1/2} & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT}) \\ (1+4\hat{c}(n+R_{\max}))dH^{7/3}\iota^{1/3}T^{2/3} + 6\hat{c}(n+R_{\max})d^{3/2}H^2\iota T^{1/2} & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES}). \end{cases}$$

Plugging the value of K into (42), we obtain

$$\mathrm{Reg}_{0T} \le \begin{cases} (1+4\hat{c}n)(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3} & \text{if } (\zeta_2,\zeta_3) = (\texttt{PES},\texttt{OPT}) \\ (n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3} & \text{if } (\zeta_2,\zeta_3) = (\texttt{OPT},\texttt{PES}). \end{cases}$$

This completes the proof of the upper bounds of the welfare regret, the agent regret, and the seller regret.

Individual Rationality. We assume that agent i reports truthfully according to the reward function $\widetilde{r}_i$ and other agents may report untruthfully according to the reward function $\widetilde{r}_j$ for $j \neq i$. Then, we adopt the same definitions of $\widetilde{\pi}_t^{\dagger i}$, $\widetilde{R}^{-i}$, $F_t^{\dagger,-i}$, and $G_t^{\dagger,-i}$ as in the proof of individual rationality in Section C.1.

Here the agents are not charged during the exploration phase, and $r_i \ge 0$ ensures that $u_{it} \ge 0$ for all $t \in [K]$. Recalling Equation (32), we have the following decomposition for $t > K$,

$$u_{it} = \underbrace{\big[V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1; r_i + \widetilde{R}^{-i}) - F_t^{\dagger,-i}\big]}_{\text{(i)}} + \underbrace{\big[G_t^{\dagger,-i} - V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1; \widetilde{R}^{-i})\big]}_{\text{(ii)}}.$$

Moreover, in the proof of individual rationality in Section C.1, we have shown that

$$\text{(i)} \ge \big[V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1; r_i + \widetilde{R}^{-i}) - V_1^*(x_1; r_i + \widetilde{R}^{-i})\big] + \big[V_1^*(x_1; \widetilde{R}^{-i}) - \widehat{V}_1^{t,\dagger}(x_1; \widetilde{R}^{-i})\big]$$

and

$$\text{(ii)} \ge \check{V}_1^{\widetilde{\pi}_t^{\dagger i}}(x_1; \widetilde{R}^{-i}) - V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1; \widetilde{R}^{-i}),$$

according to the definitions of $F_t^{\dagger,-i}$ and $G_t^{\dagger,-i}$. Applying Lemma C.1, we have that

$$\text{(i)} \ge -4\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}, \quad \text{(ii)} \ge -2\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K}.$$

Summing (i) and (ii) from $t=1$ to T, we get

$$U_{iT} \ge \sum_{t=K+1}^{T} u_{it} \ge -6\hat{c}(n+R_{\max})\sqrt{d^3H^6\iota/K},$$

Setting $K = dH^{4/3}\iota^{1/3}T^{2/3}$ in the above inequality, we further get,

$$U_{iT} \ge -6\hat{c}(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}$$

which implies the mechanism we learned is $6\hat{c}(n+R_{\max})dH^{7/3}\iota^{1/3}T^{2/3}$-approximately individually rational.

Truthfulness: The proof for truthfulness when $\zeta_1 = \texttt{EWC}$ significantly differs from the case when $\zeta_1 = \texttt{ETC}$. At a high level, when $\zeta_1 = \texttt{ETC}$, we use the fact that the data used to calculate F is collected entirely during the exploration phase and is not affected by agent i potentially reporting untruthfully, and hence $F_t^{\ddagger,-i}$ and $F_t^{\dagger,-i}$ cancel out. Unfortunately, when $\zeta_1 = \texttt{EWC}$, F's computation is dependent on the untruthful behavior of agent i. The trajectories collected during exploitation are used for computing F. The policy used for collecting these trajectories is learned using the agent i's report and thus is affected by the agent's untruthfulness. In this way, different from the proof of truthfulness in Section C.1 where $F_t^{\dagger,-i} = F_t^{\ddagger,-i}$, the following proof also bounds the difference between $F_t^{\dagger,-i}$ and $F_t^{\ddagger,-i}$. We adopt the same notations as in the proof of truthfulness in Section C.1.

We first decompose Equation (33) in terms of the exploration and exploitation phases. When $t \leq K$, the agents are not charged any price, and then $r_i \geq 0$ ensures $u_{it} \geq 0$. We thus have

$$\widetilde{u}_{it} - u_{it} \leq \widetilde{u}_{it} \leq \max_{\pi} V_1^{\pi}(x_1; r_i) \leq H,$$

where the second inequality uses the fact that the price is 0.

For $t > K$, the utility an agent gains from untruthful reporting, regardless of other agents' truthfulness, can be decomposed as follows

$$\begin{aligned}
&\widetilde{u}_{it} - u_{it} \\
&\quad = V_1^{\widetilde{\pi}_t^{\ddagger}}(x_1; r_i) - F_t^{\ddagger,-i} + G_t^{\ddagger,-i} - V_1^{\widetilde{\pi}_t^{\dagger i}}(x_1; r_i) + F_t^{\dagger,-i} - G_t^{\dagger,-i} \\
&\quad = \underbrace{\big[V_1^{\widetilde{\pi}_t^{\ddagger}}\big(x_1; r_i + \widetilde{R}^{-i}\big) - V_1^{*}\big(x_1; r_i + \widetilde{R}^{-i}\big)\big]}_{\text{(i)}} + \underbrace{\big[V_1^{*}\big(x_1; r_i + \widetilde{R}^{-i}\big) - V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1; r_i + \widetilde{R}^{-i}\big)\big]}_{\text{(ii)}} \\
&\qquad + \underbrace{\big[G_t^{\ddagger,-i} - V_1^{\widetilde{\pi}_t^{\ddagger}}\big(x_1; \widetilde{R}^{-i}\big)\big]}_{\text{(iii)}} + \underbrace{\big[V_1^{\widetilde{\pi}_t^{\dagger i}}\big(x_1; \widetilde{R}^{-i}\big) - G_t^{\dagger,-i}\big]}_{\text{(iv)}} + \underbrace{\big[F_t^{\dagger,-i} - F_t^{\ddagger,-i}\big]}_{\text{(v)}}.
\end{aligned}$$

By Lemma C.1, we know that regardless of the choice of ζ_3, we have

$$\text{(i)} + \text{(ii)} + \text{(iii)} + \text{(iv)} \leq 4\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K}.$$

We focus on studying the upper bound of (v). By the definitions of F function in Equations (31) and (34), we know

$$F_t^{\dagger,-i} = \begin{cases} \widehat{V}_1^{t,\dagger}\big(x_1; \widetilde{R}^{-i}\big) & \text{if } \zeta_2 = \texttt{OPT} \\ \widecheck{V}_1^{t,\dagger}\big(x_1; \widetilde{R}^{-i}\big) & \text{if } \zeta_2 = \texttt{PES}, \end{cases} \qquad F_t^{\ddagger,-i} = \begin{cases} \widehat{V}_1^{t,\ddagger}\big(x_1; \widetilde{R}^{-i}\big) & \text{if } \zeta_2 = \texttt{OPT} \\ \widecheck{V}_1^{t,\ddagger}\big(x_1; \widetilde{R}^{-i}\big) & \text{if } \zeta_2 = \texttt{PES}. \end{cases}$$

Recall that $F_t^{\dagger,-i}$ are generated by Algorithm 3 using dataset $\mathcal{D}$ collected with untruthful report from all the agents except agent i. On the other hand, $F_t^{\ddagger,-i}$ are generated by Algorithm 3 with dataset $\mathcal{D}$ collected with untruthful report from all the agents. Then, regardless of the choice of ζ_2 when generating F function, we have

$$F_t^{\dagger,-i} - F_t^{\ddagger,-i} \leq \widehat{V}_1^{t,\dagger}(x_1; \widetilde{R}^{-i}) - \widecheck{V}_1^{t,\ddagger,}(x_1; \widetilde{R}^{-i}),$$

since it can be easily verify that $\widehat{V}_1^{t,\dagger}\big(x_1; \widetilde{R}^{-i}\big) \geq \widecheck{V}_1^{t,\dagger}\big(x_1; \widetilde{R}^{-i}\big)$ and $\widehat{V}_1^{t,\ddagger}\big(x_1; \widetilde{R}^{-i}\big) \geq \widecheck{V}_1^{t,\ddagger}\big(x_1; \widetilde{R}^{-i}\big)$, which thus implies that $F_t^{\dagger,-i}$ is at most $\widehat{V}_1^{t,\dagger}\big(x_1; \widetilde{R}^{-i}\big)$ and $F_t^{\ddagger,-i}$ is at least $\widecheck{V}_1^{t,\ddagger}\big(x_1; \widetilde{R}^{-i}\big)$ regardless of the choices of ζ_2, ζ_3.

When $\zeta_3 = \texttt{EWC}$, the trajectories collected during the exploitation phase may differ for the computations of $\widehat{V}_1^{t,\dagger}(x_1; \widetilde{R}^{-i})$ and $\check{V}_1^{t,\ddagger}(x_1; \widetilde{R}^{-i})$, due to agent i's untruthful reporting. Fortunately, as we can see from Lemma C.1, the policy evaluation error can still be bounded: the reward-free exploration procedure in Algorithm 2 ensures that even when agent i is not truthful and $\zeta_3 = \texttt{EWC}$, data collected during the exploration phase ensures a sufficient value function estimation.With adding and subtracting $V_1^*(x_1; \widetilde{R}^{-i})$, we have

$$\begin{aligned} F_t^{\dagger,-i} - F_t^{\ddagger,-i} &\le \underbrace{\Big(\widehat{V}_1^{t,\dagger}(x_1; \widetilde{R}^{-i}) - V_1^*(x_1; \widetilde{R}^{-i})\Big)}_{\text{(i)}} + \underbrace{\Big(V_1^*(x_1; \widetilde{R}^{-i}) - \check{V}_1^{t,\ddagger}(x_1; \widetilde{R}^{-i})\Big)}_{\text{(ii)}} \\ &\le 2\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K} + 2\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K} \\ &= 4\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K}, \end{aligned}$$

by apply Lemma C.1 to term (i) and (ii) and get $2\hat{c}\sqrt{d^3H^6\iota/K}$ upper bounds on both terms respectively. In summary, we have that for all $t > K$,

$$\widetilde{u}_{it} - u_{it} \le 8\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K}.$$

Summing $\widetilde{u}_{it} - u_{it}$ from $t = 1$ to T, recalling the bound for all $t \in [K]$, we get

$$\widetilde{U}_{iT} - U_{iT} \le HK + 8\hat{c}(n + R_{\max})\sqrt{d^3H^6\iota/K}.$$

Setting $K = dH^{4/3}\iota^{1/3}T^{2/3}$ in the above inequality, we further get

$$\widetilde{U}_{iT} - U_{iT} \le (1 + 8\hat{c}(n + R_{\max}))dH^{7/3}\iota^{1/3}T^{2/3},$$

implying the learned mechanism is $(1+8\hat{c}(n+R_{\max}))dH^{7/3}\iota^{1/3}T^{2/3}$-approximately truthful. ■

Appendix D. Proof of Lemma C.1

In this section, we present the detailed proof of Lemma C.1.We first introduce several important notions e.g., Bellman operator and model evaluation error, and a supporting lemma with its proof in D.1.Then we provide the proof of Lemma C.1 in Section D.2.

We note that bounding the errors in our setting is significantly different from the results in earlier works on reward-free exploration. Note that the planning subroutines described in Algorithms 3 and 4 use the collected rewards, rather than an arbitrary given reward function, to calculate the functions F and G. As a result, the concentration analysis required to prove Lemma C.1, as well as the decomposition used for the lemma, are all designed to cater to the dynamic mechanism design regime.

D.1 Preliminaries for Proofs

We first define two operators to help characterize the estimation errors. For any function $f(;\mathfrak{R}) : \mathcal{S} \to \mathbb{R}$ with reward function $\mathfrak{R}$,

$$(\mathbb{P}_h f)(x, a; \mathfrak{R}) = \mathbb{E}[f(x_{h+1})|x_h = x, a_h = a], \tag{43}$$

and the Bellman operator at step $h \in [H]$ as

$$\begin{aligned}(\mathbb{B}_h f)(x, a; \mathfrak{R}) &= \mathbb{E}[\mathfrak{R}_h(x, a) + f(x_{h+1}) | x_h = x, a_h = a] \\ &= \mathbb{E}[\mathfrak{R}_h(x, a) | x_h = x, a_h = a] + (\mathbb{P}_h f)(x, a).\end{aligned} \tag{44}$$

For estimated value functions $V_h^{t,\pi}$ and corresponding action-value functions $Q_h^{t,\pi}$. We define the model evaluation error with policy π in episode t at each step $h \in [H]$ as

$$\begin{aligned}\hat{\Delta}_h^{t,\pi}(x, a;) &= (\mathbb{B}_h \widehat{V}_{h+1}^{t,\pi})(x, a;) - \hat{Q}_h^{t,\pi}(x, a;), \\ \check{\Delta}_h^{t,\pi}(x, a;) &= (\mathbb{B}_h \check{V}_{h+1}^{t,\pi})(x, a;) - \check{Q}_h^{t,\pi}(x, a;),\end{aligned} \tag{45}$$

for $\zeta_3 = \texttt{OPT}$ and $\texttt{PES}$ respectively. In other words, Δ_h is the error in estimating the Bellman operator defined in Equation (44), based on the dataset $\mathcal{D}$ collected in Algorithm 2.

For clarity, we define the following events to quantify the uncertainty of the estimation of the Bellman operator $\mathbb{B}_h$ in Algorithm 3 and Algorithm 4 with different hyperparameters.

Definition D.1 *We define for all $t > K$ the event $\mathcal{E}_t$ by requiring the following inequalities hold for all $(x, a) \in \mathcal{S} \times \mathcal{A}, h \in [H]$, and $(\mathfrak{R}, \pi) \in \{(R, \widehat{\pi}), (\widetilde{R}, \widetilde{\pi}_t^{\ddagger})\} \cup \{(r_i + \widetilde{R}^{-i}, \widetilde{\pi}_t^{\dagger i}), (R^{-i}, *), (\widetilde{R}^{-i}, \dagger), (\widetilde{R}^{-i}, \ddagger), (R^{-i}, \widehat{\pi}^t), (\widetilde{R}^{-i}, \widetilde{\pi}_t^{\dagger i}), (\widetilde{R}^{-i}, \widetilde{\pi}_t^{\ddagger})\}_{i=1}^n$, for each pair's associated w's*

$$\begin{aligned}\left|\phi(x, a)^\top \hat{w}_h^{t,\pi}(\mathfrak{R}) - \mathbb{B}_h \widehat{V}_{h+1}^{t,\pi}(x, a; \mathfrak{R})\right| &\le u_h^t(x, a), \\ \left|\phi(x, a)^\top \check{w}_h^{t,\pi}(\mathfrak{R}) - \mathbb{B}_h \check{V}_{h+1}^{t,\pi}(x, a; \mathfrak{R})\right| &\le u_h^t(x, a),\end{aligned}$$

where the associated w's are the learned parameters generated by Algorithm 3 if $(\mathfrak{R}, \pi) \in \{(R, \widehat{\pi}^t), (\widetilde{R}, \widetilde{\pi}_t^{\ddagger})\} \cup \{(r_i + \widetilde{R}^{-i}, \widetilde{\pi}_t^{\dagger i}), (R^{-i}, *), (\widetilde{R}^{-i}, \dagger), (\widetilde{R}^{-i}, \ddagger)\}_{i=1}^n$, and the associated w's are learned parameters generated by Algorithm 4 if $(\mathfrak{R}, \pi) \in \{(R^{-i}, \widehat{\pi}^t), (\widetilde{R}^{-i}, \widetilde{\pi}_t^{\dagger i}), (\widetilde{R}^{-i}, \widetilde{\pi}_t^{\ddagger})\}_{i=1}^n$.

Intuitively, the event defined here ensures that we attain sufficiently good policy estimates and sufficiently good value function estimates for these policies. Moreover, we highlight that the event allows for untruthfulness in the agents' behavior, thanks to our choices of $\mathfrak{R}$, and the "good" properties remain valid even when agents are untruthful. Examining the pairs of $(\mathfrak{R}, \pi)$ included in $\mathcal{E}$, we can see that the good event $\mathcal{E}_t$ directly implies that the clauses in Lemma C.1 hold for a specific value of $t > K$.

Across this paper, we let $\mathcal{E}$ denote the intersection of all the event $\{\mathcal{E}_t\}_{t=K+1}^T$ defined in D.1, which is

$$\mathcal{E} := \cap_{t=k+1}^T \mathcal{E}_t \tag{46}$$

The following lemma shows that under the appropriate choice of regularization parameter λ and scaling parameter β, event $\mathcal{E}$ is guaranteed to happen with high probability.

Lemma D.2 (Adaptation of Lemma 5.2 from Jin et al. (2020c)) *Under the setting in Section 2, we set*

$$\lambda = 1, \quad \beta = \hat{c}(n + R_{\max}) dH \sqrt{\iota}, \quad \textit{where } \iota = \log(36ndHT/\delta).$$

Here $\delta \in (0, 1)$ is the confidence parameter. It holds that

$$\mathbf{Pr}_{\mathcal{D}}(\mathcal{E}) \ge 1 - \delta/2.$$

where $\mathbf{Pr}_{\mathcal{D}}$ denotes the probability under the data-generating distribution.

Proof Note that by union bound, we only need to show that for an arbitrary and fixed $t > K$, the event $\mathcal{E}_t$ holds with probability at least $1 - \delta/T$. We note that we can obtain a tighter bound for `ETC`, as the value functions and the policies do not change during exploitation. Here we slightly loosen our bound (by a multiplicative factor of $\log T$) for brevity of the proof.

Additionally, let us examine the possible choices of $(\mathfrak{R}, \pi)$ and w for any $t > K$. We know that for any t, the concentration bound needs to hold for $2 \cdot (2 + 7n) \leq 18n$ distinct reward-policy pairs. As such, we only need to show that for an arbitrary and fixed pair of $(\mathfrak{R}, \pi)$, the concentration bounds on $\hat{w}_h^{t,\pi}$ and $\breve{w}_h^{t,\pi}$ hold simultaneously for all h with probability at least $1 - \delta/36nT$. Without loss of generality, we consider only the pair $(R, \widehat{\pi}^t)$ and the associated optimistic linear weight, as the proof for all other pairs of $(\mathfrak{R}, \pi)$ and choices of weight w remain largely the same.

Moreover, note that $\widehat{\pi}^t$ is simply the policy outputted by Algorithm 3 with respect to R when all agents are truthful. For simplicity, we then let $\hat{w}_h^{t,*}$ denote the weight associated with the pair $(R, \widehat{\pi}^t)$. As we focus on the pair (R, π) and the weight $\hat{w}_h^{t,*}$, for the rest of the proof, we let f_h^t and u_h^t denote the terms used by Algorithm 3.

Recall the definition of the transition operator $\mathbb{P}_{h+1}$ and the Bellman operator $\mathbb{B}_{h+1}$ in Equation (43) and Equation (44). We first show that for any function f, $(\mathbb{P}_h f)(,; R)$ and $(\mathbb{B}_{h+1} f)(,; R)$ are linear in the feature map ϕ. By Equation (6),

$$
\begin{aligned}
(\mathbb{P}_h f)(x, a; R) &= \Big\langle \phi(x, a), \int f(x')\mu_h(x')\mathrm{d}x' \Big\rangle \\
(\mathbb{B}_h f)(x, a; R) &= \sum_{i=0}^{n} \langle \phi(x, a), \boldsymbol{\theta}_{ih} \rangle + \Big\langle \phi(x, a), \int f(x')\mu_h(x')\mathrm{d}x' \Big\rangle
\end{aligned}
$$

where we recall $\boldsymbol{\theta}_{ih}$ parameterizes r_{ih}. Crucially, the fact that both equations hold for a generic f shows that $(\mathbb{P}_h \widehat{V}_{h+1}^{t,*})(,; R)$ and $(\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(,; R)$ are both linear.

The objective is then to obtain a high probability bound over $|(\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(,; R) - \phi^\top \hat{w}_h^{t,*}|$ for all $h \in [H], (x, a) \in \mathcal{S} \times \mathcal{A}$. Let w_h be the vector such that $(\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(,; R) = \phi(,)^\top w_h$, which is guaranteed to exist by the term's linearity. When $\zeta_1 =$ `EWC`, for all $t > K$, we have

$$
\begin{aligned}
&(\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x, a; R) - \phi(x, a)^\top \hat{w}_h^{t,*} = \phi(x, a)^\top (w_h - \hat{w}_h^{t,*}) \\
&\quad = \phi(x, a)^\top w_h - \phi(x, a)^\top (\Lambda_h^t)^{-1} \Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \big(R_h^\tau + \widehat{V}_{h+1}^{t,*}(x_{h+1}^\tau; R) \big) \Big) \\
&\quad = \underbrace{\phi(x, a)^\top w_h - \phi(x, a)^\top (\Lambda_h^t)^{-1} \Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) (\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x_h^\tau, a_h^\tau; R) \Big)}_{\text{(i)}} \qquad (47) \\
&\qquad \underbrace{- \phi(x, a)^\top (\Lambda_h^t)^{-1} \Big(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \big(R_h^\tau + \widehat{V}_{h+1}^{t,*}(x_{h+1}^\tau; R) - (\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x_h^\tau, a_h^\tau; R) \big) \Big)}_{\text{(ii)}},
\end{aligned}
$$

where the second equality follows from the construction of $\hat{w}_h^{t,*}$. Therefore we have

$$\big|(\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x,a) - \phi(x,a)^\top \hat{w}_h^{t,*}\big| \leq |\text{(i)}| + |\text{(ii)}|.$$

We now bound the two terms separately. Note that $\widehat{V}_{h+1}^{t,*}(;R) \in [0, (n+R_{\max})(H-h)]$ by truncation and $\|\theta_h\| = \|\sum_{i=0}^n \theta_{ih}\| \leq (n+1)\sqrt{d}$. Applying Lemma F.4, we then know that $\|w_h\| \leq (n+R_{\max})(H-h)\sqrt{d} < (n+R_{max})H\sqrt{d}$ for all h. Hence, term (i) in Equation (47) satisfies

$$\begin{aligned} |\text{(i)}| &= \Big|\phi(x,a)^\top w_h - \phi(x,a)^\top (\Lambda_h^t)^{-1}\Big(\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top w_h\Big)\Big| \\ &= \big|\phi(x,a)^\top w_h - \phi(x,a)^\top (\Lambda_h^t)^{-1}(\Lambda_h^t - \lambda I) w_h\big| = \lambda\big|\phi(x,a)^\top (\Lambda_h^t)^{-1} w_h\big| \\ &\leq \lambda \|w_h\|_{(\Lambda_h^t)^{-1}} \|\phi(x,a)\|_{(\Lambda_h^t)^{-1}} \leq (n+R_{\max})H\sqrt{d/\lambda}\sqrt{\phi(x,a)^\top (\Lambda_h^t)^{-1}\phi(x,a)}, \end{aligned} \tag{48}$$

where the second equality is by definition of Λ_h^t and the last by the fact that $\Lambda_h^t \succeq \lambda I$.

It remains to upper bound term (ii) in Equation (47) . For simplicity, we defined the random variable

$$\epsilon_h^\tau(V;R) = R_h^\tau + V(x_{h+1}^\tau; R) - (\mathbb{B}_h V)(x_h^\tau, a_h^\tau; R). \tag{49}$$

We then have

$$\begin{aligned} |\text{(ii)}| &= \Big|\phi(x,a)^\top (\Lambda_h^t)^{-1}\Big(\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\epsilon_h^\tau(\widehat{V}_{h+1}^{t,*}; R)\Big)\Big| \\ &\leq \Big\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\epsilon_h^\tau(\widehat{V}_{h+1}^{t,*})\Big\|_{(\Lambda_h^t)^{-1}} \|\phi(x,a)\|_{(\Lambda_h^t)^{-1}} \\ &= \underbrace{\Big\|\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\epsilon_h^\tau(\widehat{V}_{h+1}^{t,*}; R)\Big\|_{(\Lambda_h^t)^{-1}}}_{\text{(iii)}} \sqrt{\phi(x,a)^\top (\Lambda_h^t)^{-1}\phi(x,a)}. \end{aligned} \tag{50}$$

Define the function class for any $L > 0, B > 0$, $h \in [H$

$$\begin{aligned} &\mathcal{V}_h(L,B,\lambda) = \big\{V_h(x;\theta,\beta,\Sigma)\colon \mathcal{S} \to [0,(n+R_{\max})H] \text{ with } \|\theta\| \leq L, \beta \in [0,B], \Sigma \succeq \lambda I\big\}, \\ &\text{where } \; V_h(x;\theta,\beta,\Sigma) = \max_{a\in\mathcal{A}}\Big\{\min\big\{\phi(x,a)^\top\theta + \beta\sqrt{\phi(x,a)^\top \Sigma^{-1}\phi(x,a)}, (n+R_{\max})H\big\}\Big\}. \end{aligned} \tag{51}$$

and let $\mathcal{N}_h(\varepsilon; L, B, \lambda)$ be the ε-cover of $\mathcal{V}_h(L,B,\lambda)$ with respect to the distance $\text{dist}(V,V') = \sup_{x\in\mathcal{S}} \|V(x) - V'(x)\|$. By Lemma F.4, we have $\|\hat{w}_{h+1}^{t,*}\| \leq (n+R_{\max})H\sqrt{Kd/\lambda}$, and therefore

$$\widehat{V}_{h+1}^{t,*} \in \mathcal{V}_{h+1}(L_0, B_0, \lambda), \qquad \text{where } \; L_0 = (n+R_{\max})H\sqrt{Kd/\lambda}, \; B_0 = 2\beta.$$

Here $\lambda > 0$ is the regularization parameter, and $\beta > 0$ is the scaling parameter specified in Algorithm 3. For simplicity, we use $\mathcal{V}_{h+1}$ and $\mathcal{N}_{h+1}(\varepsilon)$ to denote $\mathcal{V}_{h+1}(L_0, B_0, \lambda)$ and $\mathcal{N}_{h+1}(\varepsilon; L_0, B_0, \lambda)$, respectively. There then exists a function $V_{h+1}^{\dagger}(x; R) \in \mathcal{N}(\varepsilon)$ where

$$\sup_{x \in \mathcal{S}} \big| \widehat{V}_{h+1}^{t,*}(x; R) - V_{h+1}^{0}(x; R) \big| \leq \varepsilon, \tag{52}$$

By definition of the transition operator $\mathbb{P}_h$ and Jensen's inequality,

$$\begin{aligned} \big|(\mathbb{P}_h V_{h+1}^0)(x, a; R) - (\mathbb{P}_h \widehat{V}_{h+1}^{t,*})(x, a; R)\big| &= \Big| \mathbb{E}\big[V_{h+1}^0(x_{h+1}; R) - \widehat{V}_{h+1}^{t,*}(x_{h+1}; R) \,\Big|\, s_h = x, a_h = a\big] \Big| \\ \leq \mathbb{E}\Big[\big|V_{h+1}^0(x_{h+1}; R) - \widehat{V}_{h+1}^{t,*}(x_{h+1}; R)\big| \,\Big|\, s_h = x, a_h = a \Big] &\leq \varepsilon. \end{aligned}$$

We then know that $\big|(\mathbb{B}_h V_{h+1}^0)(x, a; R) - (\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x, a; R)\big| \leq \varepsilon$, and by triangle inequality,

$$\begin{aligned} \Big| &\big(R_h^t(x, a) + \widehat{V}_{h+1}^{t,*}(x'; R) - (\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x, a; R)\big) \\ &\quad - \big(R_h^t(x, a) + V_{h+1}^0(x'; R) - (\mathbb{B}_h V_{h+1}^0)(x, a; R)\big) \Big| \leq 2\varepsilon \end{aligned} \tag{53}$$

for all $h \in [H]$ and all $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Setting $(x, a, x') = (x_h^\tau, a_h^\tau, x_{h+1}^\tau)$ in Equation (53) ensures

$$\big|\epsilon_h^\tau(\widehat{V}_{h+1}^{t,*}; R) - \epsilon_h^\tau(V_{h+1}^0; R)\big| \leq 2\varepsilon, \qquad \forall \tau \in [K],\ \forall h \in [H].$$

We then have the following bound for term (iii) in Equation (50).

$$\begin{aligned} |\text{(iii)}|^2 \leq & 2 \Big\| \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \epsilon_h^\tau(V_{h+1}^0; R) \Big\|_{(\Lambda_h^t)^{-1}}^2 \\ & + 2 \Big\| \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \big(\epsilon_h^\tau(\widehat{V}_{h+1}^{t,*}; R) - \epsilon_h^\tau(V_{h+1}^0; R)\big) \Big\|_{(\Lambda_h^t)^{-1}}^2. \end{aligned} \tag{54}$$

By direct expansion, the second term on the right-hand side of Equation (54) can be controlled as follows.

$$\begin{aligned} 2 \Big\| & \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \big(\epsilon_h^\tau(\widehat{V}_{h+1}^{t,*}; R) - \epsilon_h^\tau(V_{h+1}^0; R)\big) \Big\|_{(\Lambda_h^t)^{-1}}^2 \\ &= 2 \sum_{\tau=1}^{K} \sum_{\tau'=1}^{K} \phi(x_h^\tau, a_h^\tau)^\top (\Lambda_h^\tau)^{-1} \phi(x_h^{\tau'}, a_h^{\tau'}) \\ &\qquad \times \big(\epsilon_h^\tau(\widehat{V}_{h+1}^{t,*}; R) - \epsilon_h^\tau(V_{h+1}^0; R)\big)\big(\epsilon_h^{\tau'}(\widehat{V}_{h+1}^{t,*}; R) - \epsilon_h^{\tau'}(V_{h+1}^0; R)\big) \\ &\leq 8\varepsilon^2 \sum_{\tau=1}^{K} \sum_{\tau'=1}^{K} \big|\phi(x_h^\tau, a_h^\tau)^\top (\Lambda_h^t)^{-1} \phi(x_h^{\tau'}, a_h^{\tau'})\big| \leq 8\varepsilon^2 K^2 / \lambda, \end{aligned} \tag{55}$$

where the last step follows from the fact that $\|\phi(x, a)\| \leq 1$ and $\Lambda_h^t \succeq \lambda I$. Combining Equations (54) and (55) shows

$$|\text{(iii)}|^2 \leq 2 \sup_{V \in \mathcal{N}_{h+1}(\varepsilon)} \Big\| \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \epsilon_h^\tau(V; R) \Big\|_{(\Lambda_h^t)^{-1}}^2 + 8\varepsilon^2 K^2 / \lambda. \tag{56}$$

We then upperbound the term $\sup_{V\in\mathcal{N}_{h+1}(\varepsilon)} \Big\| \sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\epsilon_h^\tau(V;R)\Big\|^2_{(\Lambda_h^t)^{-1}}$ by uniform concentration over the covering $\mathcal{N}_{h+1}(\varepsilon)$. Applying Lemma F.6 and taking union bound over $\mathcal{N}_{h+1}(\varepsilon)$, for any fixed $h\in[H]$, with probability at least $1-p|\mathcal{N}_{h+1}(\varepsilon)|$,

$$\sup_{V\in\mathcal{N}_{h+1}(\varepsilon)} \Big\| \sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\epsilon_h^\tau(V;R)\Big\|^2_{(\Lambda_h^t)^{-1}} \le (n+R_{\max})^2H^2\big(2\log(1/p)+d\log(1+K/\lambda)\big).$$

For all $\delta\in(0,1)$ and all $\varepsilon>0$, we set $p=\delta/[(36n)H|\mathcal{N}_{h+1}(\varepsilon)|]$. Hence, for all fixed $h\in[H]$, it holds that

$$\begin{aligned}\sup_{V\in\mathcal{N}_{h+1}(\varepsilon)} &\Big\| \sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\epsilon_h^\tau(V;R)\Big\|^2_{(\Lambda_h^t)^{-1}} \\ &\le (n+R_{\max})^2H^2\big(2\log((36n)H|\mathcal{N}_{h+1}(\varepsilon)|/\delta)+d\log(1+K/\lambda)\big)\end{aligned} \tag{57}$$

with probability at least $1-\delta/(36nH)$, taken with respect to process that generates the dataset $\mathcal{D}$. Then, combining Equations (56) and (57), for all $h\in[H]$, with probability at least $1-\delta/(18nH)$,

$$\begin{aligned}\Big\| \sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\epsilon_h^\tau(\hat{V}_{h+1}^{t,*};R)\Big\|^2_{(\Lambda_h^t)^{-1}} &= |(\text{iii})|^2 \\ &\le (n+R_{\max})^2H^2\big(2\log((36n)H|\mathcal{N}_{h+1}(\varepsilon)|/\delta)+d\log(1+K/\lambda)\big)+8\varepsilon^2K^2/\lambda.\end{aligned}$$

Since $\widehat{V}_{h+1}^{t,*}\in\mathcal{V}_{h+1}((n+R_{\max})H\sqrt{Td/\lambda},2\beta,\lambda)$ we can upperbound $|\mathcal{N}_{h+1}(\varepsilon)|$ via Lemma F.5. As term (iii) is controlled, we can then ensure that term (ii) of Equation (47) can be bounded, which when combined with Equation (48) yields a bound for $\big|(\mathbb{B}_h\widehat{V}_{h+1}^{t,*})(x,a)-\phi(x,a)^\top\hat{w}_h^{t,*}\big|$ under a specific choice of ε, β, and λ.

All that remains is then to set the hyperparameters to ensure that the error can be bounded. Letting $\iota=\log(36ndHT/\delta)$, we set

$$\beta=\hat{c}(n+R_{\max})dH\sqrt{\iota},\quad \varepsilon=dH/K,\quad \lambda=1,$$

where $\hat{c}>0$ is an absolute constant that ensures

$$|(\text{ii})|\le(\hat{c}/2)ndH\sqrt{\iota}\sqrt{\phi(x,a)^\top(\Lambda_h^t)^{-1}\phi(x,a)}=\beta/2\sqrt{\phi(x,a)^\top(\Lambda_h^t)^{-1}\phi(x,a)} \tag{58}$$

with probability at least $1-\delta/(36nT)$. By Equations (47), (48) and (58), for all $h\in[H]$ and all $(x,a)\in\mathcal{S}\times\mathcal{A}$, it holds that

$$\big|(\mathbb{B}_h\hat{V}_{h+1})(x,a)-\phi(x,a)^\top\hat{w}_h^{t,*}\big|\le((n+R_{\max})H\sqrt{d}+\beta/2)\sqrt{\phi(x,a)^\top(\Lambda_h^t)^{-1}\phi(x,a)},$$

with probability at least $1-\delta/(36n)$, taking the union bound over $h\in[H]$.

Extending the result to when $\zeta_1=$ `EWC` is straightforward. Observe that Equation (47) consists of bounding K random variables whose randomness is due to only the stochasticity inherent in the transition kernel. Moving from the `ETC` to `EWC` setting simply requires bounding

T, rather than K, such variables. However, as our choice for β and ι accommodates the move from K to T, the bound in Equations (58) and (48) remain valid.

Then combining Equation (58) and (48), we obtain

$$\left|\mathbb{B}_h \widehat{V}_{h+1}^{t,*}(x,a) - \phi(x,a)^\top \hat{w}_h^{t,*}\right| \le \beta \sqrt{\phi(x,a)^\top (\Lambda_h^t)^{-1} \phi(x,a)}.$$

As there are only $18n$ such combinations of $\mathfrak{R}, \pi$ and w, obtaining the individual upper bound with probability at least $1-\delta/(36n)$ ensures that the union bound over all these triplets is satisfied with probability at least $1-\delta/2$. Therefore, we conclude the proof of Lemma D.2. ■

D.2 Proof of Lemma C.1

With event $\mathcal{E}$ defined, we proceed with the proof of Lemma C.1. The proof is organized as follows. We first directly control the model evaluation errors conditioned on the event $\mathcal{E}$, then relate these model evaluation errors to uncertainty bonuses u_h^t, followed by a reward-free style analysis that ensures sufficiently small model evaluation error across all policies. Combining these three ingredients yields Lemma C.1 directly.

In the first step of the proof, we upper and lower bound the model evaluation error Δ, defined in Equation (45), in the following lemma.

Lemma D.3 (Adaptation of Lemma 5.1 from Jin et al. (2020c)) *With λ, β set according to Lemma D.2, which ensures $\mathbf{Pr}_{\mathcal{D}}(\mathcal{E}) \ge 1-\delta/2$, we have*

$$0 \ge \hat{\Delta}_h^{t,\pi}(x,a;\mathfrak{R}) \ge -2u_h^t(x,a), \qquad 0 \le \check{\Delta}_h^{t,\pi}(x,a;\mathfrak{R}) \le 2u_h^t(x,a) \tag{59}$$

for all $t > K$, $(x,a) \in \mathcal{S}\times\mathcal{A}$, $h \in [H]$, and $(\mathfrak{R},\pi) \in \{(R,\widehat{\pi}), (\widetilde{R}, \widetilde{\pi}_t^{\ddagger})\} \cup \{(r_i + \widetilde{R}^{-i}, \widetilde{\pi}_t^{\dagger i}), (R^{-i}, *), (\widetilde{R}^{-i}, \dagger), (\widetilde{R}^{-i}, \ddagger), (R^{-i}, \widehat{\pi}^t), (\widetilde{R}^{-i}, \widetilde{\pi}_t^{\dagger i}), (\widetilde{R}^{-i}, \widetilde{\pi}_t^{\ddagger})\}_{i=1}^n$, regardless of the choice of ζ_1.

Proof The results in Lemma D.3 can be split into two parts: the upper and lower bounds of $\{\check{\Delta}_h\}$ and $\{\hat{\Delta}_h\}$. For brevity, we take $\hat{\Delta}_h^{t,*}(x,a;R^{-i})$ and $\check{\Delta}_h^{t,\widehat{\pi}^t}(x,a;R^{-i})$ as examples for optimistic and pessimistic versions for an arbitrary i, because the techniques used are largely the same.

Bounding $\hat{\Delta}_h^{t,*}(x,a;R^{-i})$. We first show that conditioned on the event $\mathcal{E}$, as defined in Definition D.1 and Equation (46), the model evaluation errors $\hat{\Delta}_h^{t,*}(x,a;R^{-i}) \le 0$ for all $h \in [H]$. We assume that $\mathcal{E}$ holds for the rest of the proof. Recalling the construction of $\hat{Q}_h^{t,*}$ from Algorithm 3, for all $h \in [H]$ and all $(x,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$\hat{Q}_h^{t,*}(x,a;R^{-i}) = \min\{(f_h^t + u_h^t)(x,a), (H-h+1)(n-1+R_{\max})\}.$$

Throughout the rest of the paragraph, we use f_h^t and u_h^t to denote the components that Algorithm 3 uses in order to construct $\hat{Q}_h^{t,*}(x,a;R^{-i})$. We first focus on when $f_h^t + u_h^t(x,a) \le (H-h+1)(n-1+R_{\max})$. Here we have $\hat{Q}_h^{t,*}(x,a;R^{-i}) = f_h^t + u_h^t(x,a)$. By definition of $\hat{\Delta}_h^{t,*}(x,a;R^{-i})$ in Equation (45),

$$\hat{\Delta}_h^{t,*}(x,a;R^{-i}) = (\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x,a;R^{-i}) - \hat{Q}_h^{t,*}(x,a;R^{-i}) = (\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x,a;R^{-i}) - f_h^t - u_h^t \le 0,$$

and the desired bound on $\hat{\Delta}_h^{t,*}(x,a;R^{-i})$ inequality follows from Lemma D.2.

If $f_h^t + u_h^t(x,a) \geq (H-h+1)(n+R_{\max})$, we have $\hat{Q}_h^{t,*}(x,a;R^{-i}) = (H-h+1)(n+R_{\max})$, which implies

$$\hat{\Delta}_h^{t,*}(x,a;R^{-i}) = (\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x,a;R^{-i}) - ((H-h+1)(n+R_{\max})) \leq 0,$$

where the inequality follows from the definition of the Bellman operator in Equation (44) and the construction of $\hat{V}_{h+1}^{t,*}$ in Algorithm 3.

It remains to establish the lower bound of $\hat{\Delta}_h^{t,*}(x,a;R^{-i})$. Combining the definition of $\hat{\Delta}_h^{t,*}(x,a;R^{-i})$ and $\hat{Q}_h^{t,*}(x,a;R^{-i})$, we have

$$\begin{aligned}\hat{\Delta}_h^{t,*}(x,a;R^{-i}) &= (\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x,a;R^{-i}) - \hat{Q}_h^{t,*}(x,a;R^{-i}) \\ &\geq (\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x,a;R^{-i}) - f_h^t - u_h^t \geq -2u_h^t,\end{aligned}$$

where the first inequality follows from the definition of $\hat{Q}_h^{t,*}(x,a;R^{-i})$ and the second inequality follows from Lemma D.2. In summary, we conclude that when conditioned on $\mathcal{E}$,

$$0 \geq \hat{\Delta}_h^{t,*}(x,a;R^{-i}) \geq -2u_h^t(x,a), \qquad \forall (x,a) \in \mathcal{S}\times\mathcal{A},\ \forall h \in [H].$$

Bounding $\check{\Delta}_h^{t,*}(x,a;R^{-i})$. We now show that the model evaluation errors for the pessimistic version is also bounded. Recalling the construction of $\check{Q}_h^{t,*}$, we have

$$\check{Q}_h^{t,*}(x,a;R^{-i}) = \Pi_{[0,(n-1+R_{\max})(H-h+1)]}[(f_h^t - u_h^t)(x,a)].$$

For the rest of the paragraph, we instead let f_h^t and u_h^t denote the components Algorithm 3 uses to construct $\check{Q}_h^{t,*}(x,a;R^{-i})$ instead. We first show that the term is bounded below by zero. When $(f_h^t - u_h^t)(x,a) \leq 0$, we trivially have

$$\check{\Delta}_h^{t,*}(x,a;R^{-i}) = (\mathbb{B}_h \check{V}_{h+1}^{t,*})(x,a;R^{-i}) - 0 \geq 0.$$

When $(f_h^t - u_h^t)(x,a) \in (0,(n-1+R_{\max})(H-h+1))$, we have

$$\check{\Delta}_h^{t,*}(x,a;R^{-i}) = (\mathbb{B}_h \check{V}_{h+1}^{t,*})(x,a;R^{-i}) - f_h^t + u_h^t \geq 0,$$

where the inequality direct follows from Lemma D.2. Finally, when $(f_h^t - u_h^t)(x,a) \geq (n-1+R_{\max})(H-h+1)$, we have

$$\check{\Delta}_h^{t,*}(x,a;R^{-i}) \geq (\mathbb{B}_h \check{V}_{h+1}^{t,*})(x,a;R^{-i}) - f_h^t + u_h^t \geq 0,$$

where the inequality is again by Lemma D.2.

We then bound the term from above. When $(f_h^t - u_h^t)(x,a) \in (0,(n-1+R_{\max})(H-h+1))$

$$\check{\Delta}_h^{t,*}(x,a;R^{-i}) \leq (\mathbb{B}_h \check{V}_{h+1}^{t,*})(x,a;R^{-i}) - f_h^t + u_h^t \leq 2u_h^t$$

by Lemma D.2. When $(f_h^t - u_h^t)(x,a) \in (0,(n-1+R_{\max})(H-h+1))$, the same bound holds as well for the same reason. We then focus on when $(f_h^t - u_h^t)(x,a) \geq (n-1+R_{\max})(H-h+1)$, in which case

$$\begin{aligned}\check{\Delta}_h^{t,*}(x,a;R^{-i}) &= (\mathbb{B}_h \check{V}_{h+1}^{t,*})(x,a;R^{-i}) - (n-1+R_{\max})(H-h+1) \\ &\leq (n-1+R_{\max})(H-h+1) - (n-1+R_{\max})(H-h+1) = 0.\end{aligned}$$

The last inequality comes from the fact that $\check{V}_{h+1}^{t,*}(;R^{-i})$ and R^{-i} are bounded above.

As the proofs for the remaining reward functions remain largely the same, we can apply the same analysis, only changing the reward function being used, thus completing the proof. ■

With Lemma D.3 in mind, we relate the value function estimation errors to the uncertainty bonus u_h^t.

Lemma D.4 *With λ, β set according to Lemma D.2, which ensures $\mathbf{Pr}_{\mathcal{D}}(\mathcal{E}) \geq 1 - \delta/2$, regardless of the choice of ζ_1, the following statements hold true jointly for all $t > K$ and some absolute constant $\hat{c}$.*

1. $0 \leq \widehat{V}_1^{\pi}(x_1; \mathfrak{R}) - V_1^*(x_1; \mathfrak{R}) \leq 2\sum_{h=1}^H \mathbb{E}_{\pi}[u_h^t]$ *for all* $(\mathfrak{R}, \pi) \in \{(R, \widehat{\pi}^t), (\widetilde{R}, \widetilde{\pi}_t^{\ddagger})\} \cup \{(r_i + \widetilde{R}^{-i}, \widetilde{\pi}_t^{\dagger i})\}_{i=1}^n$.

2. *For all* $i \in [n]$, $0 \leq \widehat{V}_1^{t,\pi}(x_1; \mathfrak{R}) - V^*(x_1; \mathfrak{R}) \leq 2\sum_{h=1}^H \mathbb{E}_{\pi}[u_h^t]$ *and* $0 \leq V^*(x_1; \mathfrak{R}) - \check{V}_1^{t,\pi}(x_1; \mathfrak{R}) \leq 2\max_{\pi'}\{\sum_{h=1}^H \mathbb{E}_{\pi'}[u_h^t]\}$, *for all* $(\mathfrak{R}, \pi) \in \{(R^{-i}, \star), (\widetilde{R}^{-i}, \dagger), (\widetilde{R}^{-i}, \ddagger)\}_{i=1}^n$.

3. *For all* $i \in [n]$, $0 \leq \widehat{V}_1^{t,\pi}(x_1; \mathfrak{R}) - V_1^{\pi}(x_1; \mathfrak{R}) \leq 2\sum_{h=1}^H \mathbb{E}_{\pi}[u_h^t]$ *and* $0 \leq V_1^{\pi}(x_1; \mathfrak{R}) - \check{V}_1^{t,\pi}(x_1; \mathfrak{R}) \leq 2\sum_{h=1}^H \mathbb{E}_{\pi}[u_h^t]$, *for all* $(\mathfrak{R}, \pi) \in \{(R^{-i}, \widehat{\pi}^t), (\widetilde{R}^{-i}, \widetilde{\pi}_t^{\dagger i}), (\widetilde{R}^{-i}, \widetilde{\pi}_t^{\ddagger})\}_{i=1}^n$.

where the bonuses $\{u_h^t\}$ are the exploration bonuses calculated by either Algorithm 3 or Algorithm 4.

Proof For brevity, we only upper bound $V_1^*(x_1; R) - V_1^{\widehat{\pi}^t}(x_1; R)$ in this section, as the proof of the remaining terms is similar.

Adding and subtracting $\hat{V}_1^{t,*}$ into the difference, we can decompose the difference into two terms

$$V_1^*(x_1; R) - V_1^{\widehat{\pi}^t}(x_1; R) = \underbrace{\Big(V_1^*(x_1; R) - \hat{V}_1^{t,*}(x_1; R)\Big)}_{\text{(i)}} + \underbrace{\Big(\hat{V}_1^{t,*}(x_1; R) - V_1^{\widehat{\pi}^t}(x_1; R)\Big)}_{\text{(ii)}}. \tag{60}$$

where we recall $\hat{V}_1^{t,*}(x_1; R)$ is the value function estimates constructed by Algorithm 3. Term (i) in Equation (60) is the difference between the estimated value function $\hat{V}_1^{t,*}(;R)$ and the optimal value function $V_1^*(;R)$, while term (ii) is the difference between $\hat{V}_1^{t,*}(;R)$ and the value function of $\hat{\pi}^t$, $V_1^{\hat{\pi}^t}(;R)$.

For term (i), we invoke Lemma F.3 with $\pi = \widehat{\pi}^t$ and $\pi' = \pi_*$ and have

$$\begin{aligned}\hat{V}_1^{t,*}(x_1; R) - V_1^*(x_1; R) = &\sum_{h=1}^H \mathbb{E}_{\pi_*}\big[\langle \hat{Q}_h^{t,*}(x_h, ; R), \hat{\pi}_h^t(\,|\,x_h) - \pi_{*,h}(\,|\,x_h)\rangle_{\mathcal{A}} \,\big|\, x_1 = x\big] \\ &+ \sum_{h=1}^H \mathbb{E}_{\pi_*}\big[\hat{Q}_h^{t,*}(x_h, a_h; R) - (\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x_h, a_h; R) \,\big|\, x_1 = x\big],\end{aligned}$$

where $\mathbb{E}_{\pi_*}$ is taken with respect to the trajectory generated by π_*. By the definition of the model evaluation error Δ_h in Equation (45), we have

$$\begin{aligned} V_1^*(x_1;R) - \hat{V}_1^{t,*}(x_1;R) &= \sum_{h=1}^{H} \mathbb{E}_{\pi_*}\big[\langle \hat{Q}_h^{t,*}(x_h,a_h;R), \pi_{*,h}(\,|\,x_h) - \hat{\pi}_h^t(\,|\,x_h)\rangle_{\mathcal{A}} \,\big|\, x_1\big] \\ &\quad + \sum_{h=1}^{H} \mathbb{E}_{\pi_*}\big[\hat{\Delta}_h^{t,*}(x_h,a_h;R) \,\big|\, x_1\big]. \end{aligned} \tag{61}$$

Similarly, invoking Lemma F.3 with $\pi = \pi' = \widehat{\pi}^t$, for term (ii), we have

$$\begin{aligned} \hat{V}_1^{t,*}(x_1;R) - V_1^{\widehat{\pi}^t}(x_1;R) &= \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}^t}\big[\hat{Q}_h^{t,*}(x_h,a_h;R) - (\mathbb{B}_h \widehat{V}_{h+1}^{t,*})(x_h,a_h;R) \,\big|\, x_1\big] \\ &= -\sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}^t}\big[\hat{\Delta}_h^{t,*}(x_h,a_h;R) \,\big|\, x_1\big], \end{aligned} \tag{62}$$

where $\mathbb{E}_{\widehat{\pi}^t}$ is taken with respect to the trajectory generated by $\widehat{\pi}^t$.

Combining Equations (60), (61) and (62), we have

$$\begin{aligned} V_1^*(x_1;R) - V_1^{\widehat{\pi}^t}(x_1;R) &= \sum_{h=1}^{H} \mathbb{E}_{\pi_*}\big[\langle \hat{Q}_h^{t,*}(x_h,;R), \pi_{*,h}(\,|\,x_h) - \hat{\pi}_h^t(\,|\,x_h)\rangle_{\mathcal{A}} \,\big|\, x_1\big] \\ &\quad + \sum_{h=1}^{H} \mathbb{E}_{\pi_*}\big[\hat{\Delta}_h^{t,*}(x_h,a_h;R) \,\big|\, x_1\big] - \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}^t}\big[\hat{\Delta}_h^{t,*}(x_h,a_h;R) \,\big|\, x_1\big]. \end{aligned} \tag{63}$$

It remains to upper bound the three terms in the right-hand side of Equation (63). For the first term, we can upper bound it by 0 following the definition of $\widehat{\pi}^t$ in Algorithm 3. To bound the last two terms, we invoke Lemma D.3, which implies

$$\begin{aligned} &\sum_{h=1}^{H} \mathbb{E}_{\pi_*}\big[\hat{\Delta}_h^{t,*}(x_h,a_h;R) \,\big|\, x_1 = x\big] \le 0, \\ &-\mathbb{E}_{\widehat{\pi}^t}\big[\hat{\Delta}_h^{t,*}(x_h,a_h;R) \,\big|\, x_1 = x\big] \le \mathbb{E}_{\widehat{\pi}^t}\big[2u_h^t(x_h,a_h) \,\big|\, x_1 = x\big], \end{aligned}$$

for all $(x,a) \in \mathcal{S} \times \mathcal{A}$ under event $\mathcal{E}$. We then know that

$$V_1^*(x_1;R) - V_1^{\widehat{\pi}^t}(x_1;R) \le \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}^t}\big[2u_h^t(x_h,a_h) \,\big|\, x_1 = x\big].$$

The remaining terms can be controlled with a similar technique, with only minor differences between optimistic and pessimistic value function estimates. The differences only affect the signs of the resulting terms but do not change the proof itself. We conclude the proof. ■

As we can see from Lemma D.4, all that remains is to control the term $\mathbb{E}_\pi[\sum_{h=1}^H u_h^t \,|\, x_1]$. For convenience, we begin with a more general bound that holds for all π and $\mathfrak{R}$, and then discuss a specialized bound for when $\zeta_1 = \texttt{EWC}$. Recalling Algorithm 3, bounding $V^*(x_1; u_h^t)$ suffices, as the definition of V^* ensures that it is the maximum of $\mathbb{E}_\pi[\sum_{h=1}^H u_h^t \,|\, x_1]$ taken over π. We detail the steps in the following Lemma.

Lemma D.5 *With probability at least* $1-\delta/(36nT)$*, for the function* u_h^t *defined in Algorithm 3, we have for all* $t > K$ *that*

$$V_1^*(x_1; u^t) \le 2\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K},$$

where $\iota = \log\big(36ndHT/\delta\big)$*, and* $\hat{c}$ *is an absolute constant. The claim holds regardless of the choice of* ζ_1*.*

Proof Using the similar technique in the proof of Lemma D.2 and Lemma D.4, with probability at least $1-\delta/8$, we have for possible pairs of $(\mathfrak{R}, \pi)$,

$$\begin{aligned}&\big|(\mathbb{P}_h V_{h+1}^k)(x, a; \mathfrak{R}) - \Pi_{[0,B]}[\phi(x,a)^\top w_h^k]\big| \\ &\qquad \le \min\Big\{\beta\sqrt{\phi(x,a)^\top (\Lambda_h^k)^{-1}\phi(x,a)}, B\Big\} = u_h^k(x,a),\end{aligned} \tag{64}$$

for all $h \in [H]$ and all $(x,a) \in \mathcal{S}\times\mathcal{A}$ with $B = H(n + R_{\max})$, where w_h^k is the linear weight constructed in Algorithm 1 during the exploration phase. For simplicity, for the remaining proof we let $V^k(\cdot) = V(\cdot; u^k), Q^k(\cdot,\cdot) = Q(\cdot,\cdot; u^k)$, and $(\mathbb{P}_h V^k)(\cdot,\cdot) = (\mathbb{P}_h V)(\cdot,\cdot; u^k)$. Based on the above inequality, we have the following intermediate results for the functions $V_1^*(\cdot; l^k)$ and $V_1^k(\cdot)$ defined in Algorithm 2

$$V_1^*(x_1; l^k) \le V_1^k(x_1) \quad \text{for all } k \in [K], \tag{65}$$

and

$$\sum_{k=1}^K V_1^k(x_1) \le \hat{c}(n + R_{\max})\sqrt{d^3 H^4 K \iota}, \tag{66}$$

for some absolute constant $\hat{c}$ with probability at least $1-\delta/4$.

Equation (65) and Equation (66) show that the estimated value function in the exploration phase is optimistic and the sum of $V_1^k(x_1)$ should be small with high probability.

Equation (65) can be proved by induction. When $h = H+1$, for all $k \in [K]$ and $s \in \mathcal{S}$, we know $V_{H+1}^*(x; l^k) = 0$ and $V_{H+1}^k(x) = 0$ such that $V_{H+1}^*(x; l^k) = V_{H+1}^k(x)$. Assume that for some $h \in [H]$ and all $x \in \mathcal{S}$,

$$V_{h+1}^*(x; l^k) \le V_{h+1}^k(x).$$

Then based on Equation (64), for all $(x,h,k) \in \mathcal{S}\times[H]\times[K]$, we further have

$$\begin{aligned}&Q_h^*(x,a; l^k) - Q_h^k(x,a) \\ &= l_h^k(x,a) + (\mathbb{P}_h V_{h+1}^*)(x,a; l^k) - \min\{\Pi_{[0,B]}[(w_h^k)^\top\phi(x,a)] + l_h^k(x,a) + u_h^k(x,a), B\} \\ &\le \max\{(\mathbb{P}_h V_{h+1}^*)(x,a; l^k) - \Pi_{[0,B]}[(w_h^k)^\top\phi(x,a)] - u_h^k(x,a), 0\} \\ &\le \max\{(\mathbb{P}_h V_{h+1}^k)(x,a) - \Pi_{[0,B]}[(w_h^k)^\top\phi(x,a)] - u_h^k(x,a), 0\} \\ &\le 0,\end{aligned}$$

where the first inequality is due to $0 \leq l_h^k(x,a) + (\mathbb{P}_h V_h^*)(x,a;l^k) \leq B$, the second inequality is by the assumption that $l_h^k(x,a) + \mathbb{P}_h V_h^*(x;l^k)$, and the last inequality by Equation (64). The above inequality further leads to

$$V_h^*(x;l^k) = \max_{a\in\mathcal{A}} Q_h^*(x,a;l^k) \leq \max_{a\in\mathcal{A}} Q_h^k(x,a) = V_h^k(x).$$

We can then complete the proof of Equation (65) by induction.

Next, we detail the proof of Equation (66), namely the upper bound of $\sum_{k=1}^K V_1^k(x_1)$. Specifically, based on Equation (64), we have

$$\begin{aligned} V_h^k(x_h^k) &\leq \Pi_{[0,B]}[(w_h^k)^\top \phi(x_h^k,a_h^k)] + l_h^k(x_h^k,a_h^k) + u_h^k(x_h^k,a_h^k) \\ &\leq \mathbb{P}_h V_{h+1}^k(x_h^k,a_h^k) + l_h^k(x_h^k,a_h^k) + 2u_h^k(x_h^k,a_h^k) \\ &= \mathbb{P}_h V_{h+1}^k(x_h^k,a_h^k) - V_{h+1}(x_{h+1}^k) + V_{h+1}(x_{h+1}^k) + (2+1/H)u_h^k(x_h^k,a_h^k), \end{aligned} \tag{67}$$

where the first inequality is due to the definition of V_h^k and the second by Equation (64). For brevity, we let $\xi_h^k = \mathbb{P}_h V_{h+1}^k(x_h^k,a_h^k) - V_{h+1}(x_{h+1}^k)$ in the following. Recursively applying Equation (67), we have

$$V_1^k(x_1) \leq \sum_{h=1}^{H-1} \xi_h^k + (2+1/H)\sum_{h=1}^{H} u_h^k(x_h^k,a_h^k).$$

Taking summation on both sides of the above inequality with k from 1 to K, we have

$$\sum_{k=1}^K V_1^k(x_1) \leq \sum_{k=1}^K \sum_{h=1}^{H-1} \xi_h^k + (2+1/H)\sum_{k=1}^K \sum_{h=1}^{H} u_h^k(x_h^k,a_h^k).$$

For the first summation on the right side of the above inequality, we can bound it with Azuma-Hoeffding inequality and have

$$\sum_{k=1}^K \sum_{h=1}^{H-1} \xi_h^k \leq \mathcal{O}\Big(\sqrt{H^3 K \log(1/\delta)}\Big),$$

with probability at least $1-\delta/8$. On the other hand, by Lemma F.2, we have

$$\sum_{k=1}^K \sum_{h=1}^{H} u_h^k(x_h^k,a_h^k) \leq \mathcal{O}\Big(\sqrt{dKH^2 \log K}\Big),$$

with probability at least $1-\delta/8$. Then, combining the above two inequalities, we obtain that with probability at least $1-\delta/4$, there is

$$\sum_{k=1}^K V_1^k(x_1) \leq \hat{c}(n+R_{\max})\sqrt{d^3H^4K\iota},$$

which completes the proof of Equation (66).

At last, we prove the conclusion of this lemma that

$$V_1^*(x_1; u^t) \le \hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K}.$$

Notice that for all $k \in [K]$,

$$\Lambda_h^k \preccurlyeq \Lambda_h,$$

especially when $\zeta_1 = \texttt{EWC}$ and Λ_h may further grow during the exploitation phase. Therefore, we have for all $(h, k) \in [H] \times [K]$,

$$l_h^k \ge u_h^t / H$$

whenever $t \ge k$. Hence, $V_1^*(x_1; u^t/H) \le V_1^*(x_1; l^k)$. Together with Equation (65) and (66), we obtain

$$V_1^*(x_1; u^t) = H V_1^*(x_1; u^t/H) \le H \sum_{k=1}^{K} V_1^k(x_1)/K \le \hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K},$$

which concludes the proof. ■

Finally, with Lemmas D.4 and D.5 in mind, we argue how they can be combined to prove the claims in Lemma C.1 for both when $\zeta_1 = \texttt{ETC}$ and when $\zeta_1 = \texttt{EWC}$.

As the proof techniques are largely the same, let $(\mathfrak{R}, \pi)$ be an arbitrary and fixed pair and we discuss only $\widehat{V}_1^{t,\pi}\big(x_1; \mathfrak{R}\big) - V_1^{\pi}\big(x_1; \mathfrak{R}\big)$ to avoid redundancy. Recalling from Lemma D.4, we know that

$$\widehat{V}_1^{t,\pi}\big(x_1; \mathfrak{R}\big) - V_1^{\pi}\big(x_1; \mathfrak{R}\big) \le 2 \sum_{h=1}^{H} \mathbb{E}_\pi[u_h^t] \le 2\hat{c}(n + R_{\max})\sqrt{d^3 H^6 \iota / K},$$

where the second inequality comes from Lemma D.5.

Appendix E. Proof of Lower Bound

In this section, we present the proof of the lower bound shown in Theorem 4.3. While the work Kandasamy et al. (2020) studies the lower bound for the bandit setting, we remark that deriving the lower bound for our problem is non-trivial, which requires different constructions and proof techniques from that of this earlier work. Specifically, our work focuses on the setting of the stochastic rewards and invalidates the Gaussian reward construction in the proof of Theorem 1 in Kandasamy et al. (2020) because of the bounded reward assumption in our MDP setting. We use a different construction with the Bernoulli reward and apply a different anti-concentration analysis. Moreover, our lower bound considers the linear function approximation and the transition dynamics along the finite horizon in the MDP model which cannot be covered by the bandit setting.

We first show several important lemmas for the proof of Theorem 4.3. The following lemma translates the utilities of the seller and agent i into the differences between the value functions according to Markov VCG mechanism.

Lemma E.1 *When the actions and prices are chosen according to the Markov VCG mechanism, we have*

$$
\begin{aligned}
u_{i*} &= V_1^{\pi_*}\big(x_1; R\big) - V_1^{\pi_*^{-i}}\big(x_1; R^{-i}\big),\\
u_{0*} &= \sum_{i=1}^{n} V_1^{\pi_*^{-i}}\big(x_1; R^{-i}\big) - (n-1)V_1^{\pi_*}\big(x_1; R\big).
\end{aligned}
$$

Proof We can deduce the above results by the definition of the utilities of the agents and the seller. For the utility of agent i, we have

$$
\begin{aligned}
u_{i*} &= V_1^{\pi_*}(x_1; r_i) - p_{i*}\\
&= V_1^{\pi_*}\big(x_1; r_i\big) - \Big[V_1^{\pi_*^{-i}}\big(x_1; R^{-i}\big) - V_1^{\pi_*}\big(x_1; R^{-i}\big)\Big]\\
&= V_1^{\pi_*}\big(x_1; R\big) - V_1^{\pi_*^{-i}}\big(x_1; R^{-i}\big).
\end{aligned}
$$

For the utility of the seller, we have

$$
\begin{aligned}
u_{0*} &= V_1^{\pi_*}(x_1; r_0) + \sum_{i=1}^{n} p_{i*}\\
&= V_1^{\pi_*}\big(x_1; r_0\big) + \sum_{i=1}^{n}\Big[V_1^{\pi_*^{-i}}\big(x_1; R^{-i}\big) - V_1^{\pi_*}\big(x_1; R^{-i}\big)\Big]\\
&= \sum_{i=1}^{n} V_1^{\pi_*^{-i}}\big(x_1; R^{-i}\big) - (n-1)V_1^{\pi_*}\big(x_1; R\big),
\end{aligned}
$$

where the last equation is by $V_1^{\pi_*}(x_1; R^{-i}) = V_1^{\pi_*}(x_1; r_0 + \sum_{j\in[n], j\neq i} r_j) = V_1^{\pi_*}(x_1; r_0) + \sum_{j\in[n], j\neq i} V_1^{\pi_*}(x_1; r_j)$. This completes the proof. ■

We then define the estimation of $\sum_{i=1}^{n} V_1^{\pi_*^{-i}}\big(x_1; R^{-i}\big)$ and the error of this estimation as

$$
Y_T = \frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T}\Big(p_{it} + V_1^{\pi^t}\big(x_1; R^{-i}\big)\Big), \qquad Z_T = Y_T - \sum_{i=1}^{n} V_1^{\pi_*^{-i}}\big(x_1; R^{-i}\big).
$$

The next lemma states the relationships between different regret terms defined in Equation (5), which supports the proof of our lower bound.

Lemma E.2 *Let* $\mathrm{Reg}_T^W, \mathrm{Reg}_{0T}, \mathrm{Reg}_T^\sharp$ *be defined as in Equation* (5)*. Then*

$$
\mathrm{Reg}_T^\sharp = n\mathrm{Reg}_T^W + TZ_T, \qquad \mathrm{Reg}_{0T} = -(n-1)\mathrm{Reg}_T^W - TZ_T.
$$

Proof The proof of this lemma relies on the decomposition of these regret terms. We first define $h_{it} := p_{it} + V_1^{\pi^t}\big(x_1; R^{-i}\big)$. Then we have $Y_T = \frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T} h_{it}$. For agent i, we have

$$
\begin{aligned}
u_{it} &= V_1^{\pi^t}\big(x_1; r_i\big) - p_{it}\\
&= V_1^{\pi^t}\big(x_1; r_i\big) - \Big(h_{it} - V_1^{\pi^t}\big(x_1; R^{-i}\big)\Big) \qquad (68)\\
&= V_1^{\pi^t}\big(x_1; R\big) - h_{it}.
\end{aligned}
$$

Combining Lemma E.1 and Equation (68), we can obtain

$$\begin{aligned}u_{i*} - u_{it} &= \Big(V_1^{\pi_*}(x_1;R) - V_1^{\pi_*^{-i}}(x_1;R^{-i})\Big) - \Big(V_1^{\pi^t}(x_1;R) - h_{it}\Big)\\ &= \Big(V_1^{\pi_*}(x_1;R) - V_1^{\pi^t}(x_1;R)\Big) - \Big(V_1^{\pi_*^{-i}}(x_1;R^{-i}) - h_{it}\Big).\end{aligned}$$

Then by the definition of $\mathrm{Reg}_T^\sharp$ in Equation (5), we have

$$\begin{aligned}\mathrm{Reg}_T^\sharp &= \sum_{t=1}^T \sum_{i=1}^n (u_{i*} - u_{it})\\ &= \sum_{t=1}^T \sum_{i=1}^n \Big[\Big(V_1^{\pi_*}(x_1;R) - V_1^{\pi^t}(x_1;R)\Big) - \Big(V_1^{\pi_*^{-i}}(x_1;R^{-i}) - h_{it}\Big)\Big]\\ &= n\sum_{t=1}^T \Big(V_1^{\pi_*}(x_1;R) - V_1^{\pi^t}(x_1;R)\Big) + T\Big(Y_T - \sum_{i=1}^n V_1^{\pi_*^{-i}}(x_1;R^{-i})\Big)\\ &= n\mathrm{Reg}_T^W + TZ_T.\end{aligned}$$

This proves the first claim. For the seller, at time t, we have the following observation that

$$\begin{aligned}u_{0t} &= V_1^{\pi^t}(x_1;r_0) + \sum_{i=1}^n p_{it}\\ &= V_1^{\pi^t}(x_1;r_0) + \sum_{i=1}^n \Big(h_{it} - V_1^{\pi^t}(x_1;R^{-i})\Big)\\ &= \sum_{i=1}^n h_{it} - (n-1)V_1^{\pi^t}(x_1;R).\end{aligned}\tag{69}$$

Similarly, we can now combine Lemma E.1 and Equation (69) and obtain

$$\begin{aligned}\mathrm{Reg}_{0T} &= \sum_{t=1}^T (u_{0*} - u_{0t})\\ &= \sum_{t=1}^T \Big(V_1^{\pi_*}(x_1;R^{-i}) - h_{it}\Big) + (n-1)\sum_{t=1}^T \Big(V_1^{\pi^t}(x_1;R) - V_1^{\pi_*}(x_1;R)\Big)\\ &= -TZ_T - (n-1)R_T.\end{aligned}$$

This completes the proof of the second claim. ■

The following lemma about relative entropy gives another useful inequality for our proof of the lower bound.

Lemma E.3 *(Bretagnolle-Huber Inequality) Let $\mathbb{Q}_1$ and $\mathbb{Q}_2$ be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$\mathbb{Q}_1(A) + \mathbb{Q}_2(A^c) \geq \frac{1}{2}\exp(-\mathrm{KL}(\mathbb{Q}_1 \| \mathbb{Q}_2)),\tag{70}$$

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

Now we are ready to prove Theorem 4.3.

Proof [Proof of Theorem 4.3] At the beginning of the proof, we first state a basic inequality here: for any set of real numbers $\{r_i\}_{i\geq 1}$, and any set of $\{a_i\}_{i\geq 1}$ such that $\sum_{i\geq 1} a_i = 1$ and $a_i \geq 0$, we have $\max\{r_i\}_{i\geq 1} \geq \sum_{i\geq 1} a_i r_i$. Combining the above inequality and Lemma E.2, we obtain two lower bounds of $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}$. The first one is

$$\begin{aligned}\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\} &\geq \frac{4}{5}n\mathrm{Reg}_T^W + \frac{1}{5}\mathrm{Reg}_{0T}\\ &= \frac{4}{5}n\mathrm{Reg}_T^W - \frac{1}{5}\big(-(n-1)\mathrm{Reg}_T^W - TZ_T\big)\\ &\geq \frac{2}{5}n\mathrm{Reg}_T^W - \frac{1}{5}TZ_T,\end{aligned}$$

where we use Lemma E.2 in the first equality and use the fact that $\mathrm{Reg}_T^W \geq 0$. Moreover, we obtain another lower bound as

$$\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\} \geq \frac{2}{5}n\mathrm{Reg}_T^W + \frac{1}{5}TZ_T.$$

Comparing the above two lower bounds of $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}$, we have

$$\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\} \geq \frac{2}{5}n\mathrm{Reg}_T^W + \frac{1}{5}T|Z_T|.$$

For brevity, hereafter, we define $S_T := \frac{2}{5}n\mathrm{Reg}_T^W + \frac{1}{5}T|Z_T|$. Our goal is to obtain a lower bound on $\inf_{Alg}\sup_\Theta \mathbb{E}[S_T]$ which is also a lower bound on $\max\{n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\}$. To achieve this goal, we construct two problems in Θ and show that no algorithm can work well on these two problems simultaneously.

We define the underlying MDP $\mathcal{M}_0$ for the first problem θ_0 as follows: $\mathcal{M}_0$ is an episodic MDP with horizon $H \geq 2$, state space $\mathcal{S} = \{x_0, x_1, x_2, \cdots, x_{n+1}, x_{n+2}\}$, and action space $\mathcal{A} = \{b_1, b_2, \cdots, b_A\}$ with $|\mathcal{A}| = A \geq n+2$. We let the initial state be fixed as x_0. For the transition kernel, at the first step $h = 1$, we set

$$\begin{aligned}&\mathcal{P}_1(x_i|x_0, b_i) = 1, \quad \text{for all } i \in \{1, 2, \cdots, n+1\},\\ &\mathcal{P}_1(x_{n+2}|x_0, b_i) = 1 \quad \text{for all } i \in \{n+2, \cdots, A\}.\end{aligned}$$

Meanwhile, at any subsequent step $h \in \{2, \cdots, H\}$, we set

$$\mathcal{P}_h(x_i|x_i, a) = 1, \quad \text{for all } a \in \mathcal{A},$$

i.e., state $\{x_i\}_{i=1}^{n+2}$ are absorbing states. For the reward function, we let $\mathrm{Ber}(p)$ denote a Bernoulli random variable with success probability p and set

$$\begin{aligned}&r_{0h}(s,a) = 0, \quad \text{for all } (h,s,a) \in \{1,\cdots,H\} \times \mathcal{S} \times \mathcal{A},\\ &r_{i1}(x_0,a) = 0, \quad \text{for all } (i,a) \in [n+2] \times \mathcal{A},\\ &r_{jh}(x_i,a) \sim \mathrm{Ber}(1/2), \quad \text{for all } j \neq i \text{ and } (i,h,a) \in [n] \times \{2,\cdots,H\} \times \mathcal{A},\\ &r_{ih}(x_i,a) = 0, \quad \text{for all } (i,h,a) \in [n] \times \{2,\cdots,H\} \times \mathcal{A},\\ &r_{jh}(x_{n+1},a) \sim \mathrm{Ber}(1/2), \quad \text{for all } (j,h,a) \in [n] \times \{2,\cdots,H\} \times \mathcal{A},\\ &r_{jh}(x_{n+2},a) \sim \mathrm{Ber}(1/8), \quad \text{for all } (j,h,a) \in [n] \times \{2,\cdots,H\} \times \mathcal{A},\end{aligned} \tag{71}$$

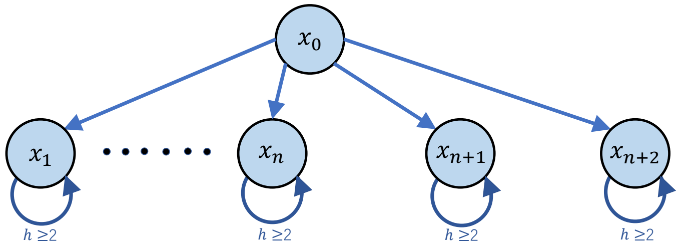

Figure 1: An illustration of the episodic MDPs $\mathcal{M}_0, \mathcal{M}_1$ with the state space $\mathcal{S} = \{x_0, x_1, \cdots, x_{n+2}\}$ and action space $\mathcal{A} = \{b_j\}_{j=1}^{A}$. Here we fix the initial state as $x_1 = x_0$, where the agent takes the action $a \in \mathcal{A}$ and transitions into the second state $s_2 \in \{x_1, \cdots, x_{n+2}\}$. In both MDPs, we have the same transition kernel. At the first step $h = 1$, the transition kernel satisfies $\mathcal{P}_1(x_i|x_0, b_i) = 1$ for all $i \in \{1, 2, \cdots, n+1\}$ and $\mathcal{P}_1(x_{n+2}|x_0, b_i) = 1$ for all $i \in \{n+2, \cdots, A\}$. Also, $x_1, x_2, sx_{n+2} \in \mathcal{S}$ are the absorbing states. The reward functions for $\mathcal{M}_0, \mathcal{M}_1$ are showed as in Equations (71) and (72).

which means the seller's reward is always 0. Please see Figure 1 for an illustration of the construction.

Note that $\mathcal{M}_0$ is a linear MDP with the dimension $d = n + 2$. We set the corresponding feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ as

$$\begin{aligned}
&\phi(x_0, b_i) = e_i, \text{for all } i = 1, 2, \cdots, n+1,\\
&\phi(x_0, b_i) = e_{n+2}, \text{for all } i = n+2, \cdots, A,\\
&\phi(x_i, b_j) = e_i, \text{for all } i = 1, 2, \cdots, n+1 \text{ and } j \in [A],\\
&\phi(x_i, b_j) = e_{n+2}, \text{for all } i = n+2, \cdots, A \text{ and } j \in [A],
\end{aligned}$$

where $\{e_j\}$ are the canonical basis of $\mathbb{R}^{n+2}$. Additionally, if the seller transitions to state x_{h+1}, the sum of agents' utilities will be the largest. We can also obtain the following results about problem θ_0 directly,

$$\begin{aligned}
&V_1^{\pi_*}(x_0; R) = Q_1(x_0, b_{n+1}; R) = \frac{1}{2}n(H-1),\\
&V_1^{\pi_*^{-i}}(x_0; R^{-i}) = Q_1\big(x_0, b_i; R^{-i}\big) = \frac{1}{2}(n-1)(H-1),\\
&\sum_{i=1}^{n} V_1^{\pi_*^{-i}}(x_0; R^{-i}) = \frac{1}{2}n(n-1)(H-1).
\end{aligned}$$

For the rest of this section, we slightly abuse the notation and drop the superscript from the Q-function, as the Q-functions of the different policies we consider are determined by the actions taken by these policies at the first step.

The second problem, i.e., θ_1, with the underlying MDP $\mathcal{M}_1$ is nearly the same as θ_0 but differs in reward functions at state x_i for $i \in [n]$. Then, we define θ_1 as

$$\begin{aligned}
&r_{jh}(x_i, a) \sim \mathrm{Ber}(1/2 + \delta), \quad \text{for all } j \neq i \text{ and } (i, h, a) \in [n] \times \{2, \cdots, H\} \times \mathcal{A},\\
&r_{ih}(x_i, a) = 0, \quad \text{for all } (i, h, a) \in [n] \times \{2, \cdots, H\} \times \mathcal{A}.
\end{aligned} \tag{72}$$

Here we set $\delta \in (0, 1/(2n-2))$. The problem θ_1 shares the same feature maps ϕ and the transition parameters μ with problem θ_0. And the difference lies in the reward parameters. Please see figure 1 for an illustration. Then, we can obtain the following inequalities for problem θ_1,

$$\begin{aligned}
&V_1^{\pi_*}(x_0; R) = Q_1(x_0, b_{n+1}; R) = \frac{1}{2}n(H-1),\\
&V_1^{\pi_*^{-i}}(x_0; R^{-i}) = Q_1\big(x_0, b_i; R^{-i}\big) = \Big(\frac{1}{2}+\delta\Big)(n-1)(H-1),\\
&\sum_{i=1}^{n} V_1^{\pi_*^{-i}}(x_0; R^{-i}) = \Big(\frac{1}{2}+\delta\Big)n(n-1)(H-1).
\end{aligned}$$

Specifically, we denote $S_T(\theta_0)$ and $S_T(\theta_1)$ as the S_T under problems θ_0 and θ_1 respectively. The expectations and probabilities corresponding to problem θ_i will be denoted as $\mathbb{E}_{\theta_i}$ and $\mathbf{Pr}_{\theta_i}$ respectively. Let $N_k(a) = \sum_{\tau=1}^{k} \mathbb{I}\{(a_1^\tau = a)\}$ denote the number of times that the seller takes action a at the first step in the initial k rounds. Here we rewrite the lower bound of the welfare regret in problem $\theta \in \{\theta_1, \theta_2\}$ as

$$\begin{aligned}
\mathbb{E}_\theta[\mathrm{Reg}_T^W] &= \sum_{j=1, j\neq n+1}^{n+2} \big(Q_1(x_0, b_{n+1}; R) - Q_1(x_0, b_j; R)\big)\mathbb{E}_\theta[N_K(b_j)]\\
&\geq \sum_{j=1}^{n} \big(Q_1(x_0, b_{n+1}; R) - Q_1(x_0, b_j; R)\big)\mathbb{E}_\theta[N_K(b_j)].
\end{aligned}$$

Observing that $Q_1(x_0, b_{n+1}; R) - Q_1(x_0, b_j; R) = (H-1)/2$ in problem θ_0, and that $|Z_T|$ is at least $n(n-1)(H-1)/2$ when $Y_T > [n^2/2 - n/2 + n(n-1)\delta/2](H-1)$, we get the following lower bound of $\mathbb{E}_{\theta_0}[S_T(\theta_0)]$ as

$$\begin{aligned}
&\mathbb{E}_{\theta_0}[S_T(\theta_0)]\\
&\geq \frac{2}{5}n\mathrm{Reg}_T^W + \frac{1}{5}T|Z_T|\\
&\geq \frac{2}{5}n\sum_{j=1}^{n}\frac{H-1}{2}\mathbb{E}_{\theta_0}[N_K(b_j)] + \frac{T}{5}\frac{n(n-1)(H-1)\delta}{2}\mathbf{Pr}_{\theta_0}\Big(\underbrace{Y_T > \Big[\frac{n^2}{2} - \frac{n}{2} + \frac{n(n-1)\delta}{2}\Big](H-1)}_{\text{event } E}\Big)\\
&\geq \frac{n(H-1)}{10}\Big[\sum_{j=1}^{n} 2\mathbb{E}_{\theta_0}[N_K(b_j)] + T(n-1)\delta\mathbf{Pr}_{\theta_0}(E)\Big]. \qquad (73)
\end{aligned}$$

In problem θ_1, we have $|Z_T|$ is at least $n(n-1)(H-1)/2$ when $Y_T \leq [n^2/2 - n/2 + n(n-1)\delta/2](H-1)$. We drop the welfare regret, which is positive, in the analysis and use the above statement regarding Y_T under the event E^c in problem θ_1 to obtain

$$\mathbb{E}_{\theta_1}[S_T(\theta_1)] \geq \frac{n(H-1)}{10}T(n-1)\delta\mathbf{Pr}_{\theta_1}(E^c). \qquad (74)$$

Applying Lemma E.3 to $\mathbf{Pr}_{\theta_0}(E) + \mathbf{Pr}_{\theta_1}(E^c)$, we have

$$\mathbf{Pr}_{\theta_0}(E) + \mathbf{Pr}_{\theta_1}(E^c) \geq \frac{1}{2}\exp(-\mathrm{KL}(\mathbf{Pr}_{\theta_0}^T || \mathbf{Pr}_{\theta_1}^T)),$$

where we slightly abuse the notation and let $\mathbf{Pr}_{\theta_0}^T$ and $\mathbf{Pr}_{\theta_1}^T$ denote the probability distribution of the observed rewards up to time T in problem θ_0 and θ_1 respectively. We also notice that if the seller takes action b_{n+1}, b_{n+2} at the first step, then $\mathbf{Pr}_{\theta_0}^T = \mathbf{Pr}_{\theta_1}^T$. If the seller take action b_i for $i \in \{1, 2, sn\}$ in the first step, then the reward distributions of agent i are the same in both θ_0 and θ_1.However, for other agents $j \neq i$, the KL divergence between the corresponding distributions in the two problems is $-\log(1-4\delta^2)(H-1)$ since the rewards are mutually independent and the KL divergence between $\mathrm{Ber}(1/2)$ and $\mathrm{Ber}(1/2+\delta)$ is $-\log(1-4\delta^2)$. Then we have

$$\mathrm{KL}(\mathbf{Pr}_{\theta_0}^T || \mathbf{Pr}_{\theta_1}^T) = -(n-1)(H-1)\log(1-4\delta^2)\sum_{j=1}^{n} \mathbb{E}_{\theta_0}[N_K(b_j)]. \tag{75}$$

By combining Equations (73), (74),(70), and (75), we obtain the lower bound for $\mathbb{E}_{\theta_0}[S_T(\theta_0)] + \mathbb{E}_{\theta_1}[S_T(\theta_1)]$ as

$$\begin{aligned}
&\mathbb{E}_{\theta_0}[S_T(\theta_0)] + \mathbb{E}_{\theta_1}[S_T(\theta_1)] \\
&\quad \geq \frac{n(H-1)}{10}\Big[\sum_{j=1}^{n} 2\mathbb{E}_{\theta_0}[N_K(b_j)] + T(n-1)\delta\big(\mathbf{Pr}_{\theta_0}(E) + \mathbf{Pr}_{\theta_1}(E^c)\big)\Big] \\
&\quad \geq \frac{n(H-1)}{10}\Big[2\sum_{j=1}^{n} \mathbb{E}_{\theta_0}[N_K(b_j)] \\
&\qquad + \frac{1}{2}T(n-1)\delta\exp\Big((n-1)(H-1)\log(1-4\delta^2)\sum_{j=1}^{n}\mathbb{E}_{\theta_0}[N_K(b_j)]\Big)\Big] \\
&\quad \geq \frac{n(H-1)}{10}\min\Big\{\underbrace{2x + \frac{1}{2}T(n-1)\delta\exp\Big((n-1)(H-1)\log(1-4\delta^2)x\Big)}_{:=f(x)}\Big\},
\end{aligned}$$

where we combine Equation (73) and (74) in the first inequality, and the second inequality is by Equation (70) and Equation (75). For the last step we substitute $\sum_{j=1}^{n} \mathbb{E}_{\theta_0}[N_K(b_j)]$ by x and turn to find the minimum value of the function $f(x)$. Then, we have

$$x_0 = \frac{-1}{(n-1)(H-1)\log(1-4\delta^2)}\log\Big(\frac{-T(n-1)^2(H-1)\delta\log(1-4\delta^2)}{4}\Big)$$

as the minimum of $f(x)$. Thus, we have

$$\begin{aligned}
\mathbb{E}_{\theta_0}[S_T(\theta_0)] + \mathbb{E}_{\theta_1}[S_T(\theta_1)] &\geq \frac{n(H-1)}{10}2x_0 \\
&\geq \frac{-1}{5\log(1-4\delta^2)}\log\Big(\frac{-T(n-1)^2(H-1)\delta\log(1-4\delta^2)}{4}\Big).
\end{aligned} \tag{76}$$

Using the basic inequality $x/(1+x) \leq \log(1+x) \leq x$ for $x > -1$, we have

$$-4\delta^2 \geq \log(1-4\delta^2) \geq \frac{-4\delta^2}{1-4\delta^2} \geq -8\delta^2,$$

when $0 \le \delta^2 \le 1/8$. Combining Equation (76) and the above inequality, we obtain

$$\mathbb{E}_{\theta_0}[S_T(\theta_0)] + \mathbb{E}_{\theta_1}[S_T(\theta_1)] \ge \frac{-1}{5(-8\delta^2)} \log\left(\frac{-T(n-1)^2(H-1)\delta(-4\delta^2)}{4}\right)$$
$$= \frac{1}{40\delta^2} \log\Big(T(n-1)^2(H-1)\delta^3\Big).$$

Finally, we choose $\delta = \Big(1/\big(T(n-1)^2(H-1)\big)\Big)^{1/3}$ to obtain the lower bound

$$\frac{1}{2}\big(\mathbb{E}_{\theta_0}[S_T(\theta_0)] + \mathbb{E}_{\theta_1}[S_T(\theta_1)]\big) \ge cn^{4/3}H^{2/3}T^{2/3},$$

for some absolute constant c. Here $\delta \in (0, 1/(2n-2))$ is satisfied when $T \ge 8(n-1)/(H-1)$ and $\delta^2 \in (0, 1/8)$ is satisfied when $n \ge 3$. Observing that

$$\sup_{\theta\in\Theta} \mathbb{E}[S_T(\theta)] \ge \max\{\big(\mathbb{E}_{\theta_0}[S_T(\theta_0)] + \mathbb{E}_{\theta_1}[S_T(\theta_1)]\big)\} \ge \frac{1}{2}\big(\mathbb{E}_{\theta_0}[S_T(\theta_0)] + \mathbb{E}_{\theta_1}[S_T(\theta_1)]\big)$$

we have the conclusion that

$$\inf_{Alg}\sup_{\Theta}\ \mathbb{E}\left[\max\big(n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\big)\right] \ge \Omega(n^{4/3}H^{2/3}T^{2/3}).$$

On the other hand, noting that $\max\big(n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\big) \ge n\mathrm{Reg}_T^W$ always holds, we have

$$\max\big(n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T}\big) \ge n\mathrm{Reg}_T^W = n\left[TV_1^*(x_1; R) - \sum_{t=1}^T V_1^{\pi^t}(x_1; R)\right], \tag{77}$$

where we recall $V_1^*(x_1; r) := \max_\pi V^\pi(x_1; r)$ for any reward function r. Since $R = \sum_{i=0}^n r_i$, we consider a simple hard instance that $r_1 = r_2 = s = r_n = r'$ and $r_0 = R_{\max} \times r'$, where $r' : \mathcal{S}\times\mathcal{A} \mapsto [0,1]$ is some reward function. In other words, here we consider an instance with the same reward function for all $r_i, 1 \le i \le n$, and r_0 is simply the same reward function scaled by $R_{\max}$. Under this setting, by (77), we have

$$\max\big(n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T})\big] \ge n\left[TV_1^*(x_1; R) - \sum_{t=1}^T V_1^{\pi^t}(x_1; R)\right]$$
$$= n(n + R_{\max})\left[TV_1^*(x_1; r') - \sum_{t=1}^T V_1^{\pi^t}(x_1; r')\right].$$

The above inequality implies that the lower bound of $\max\big(n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T})\big]$ can be further lower bounded by the lower bound of the regret for linear MDPs of dimension d with rewards in $[0,1]$. Theorem 1 in Zhou et al. (2020b) shows that for any algorithm, if $d \ge 4$ and $T \ge 64(d-3)^2H$, then there exists at least one linear MDP instance that incurs regret at least $\Omega(d\sqrt{HT})$. Therefore, we can further obtain that under the same assumptions, the minimax lower bound for $\max\big(n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T})]$ is at least $\Omega\big(n(n+R_{\max})d\sqrt{HT}\big)$, i.e.,

$$\inf_{\mathsf{Alg}} \sup_{\Theta} \mathbb{E}\Big[\max\big(n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T})\Big] \geq \Omega\bigg(n(n+R_{\max})d\sqrt{HT}\bigg).$$

Combining the above results together, we have the following lower bound as

$$\inf_{\mathsf{Alg}} \sup_{\Theta} \mathbb{E}\Big[\max\big(n\mathrm{Reg}_T^W, \mathrm{Reg}_T^\sharp, \mathrm{Reg}_{0T})\Big] \geq \Omega\bigg(n^{4/3}H^{2/3}T^{2/3} + n(n+R_{\max})d\sqrt{HT}\bigg).$$

This concludes the proof of Theorem 4.3. ■

Appendix F. Other Supporting Lemmas

The following lemma from Abbasi-Yadkori et al. (2011) establishes the concentration of self-normalized processes.

Lemma F.1 (Concentration of Self-Normalized Processes) *Let $\{\mathcal{F}_t\}_{t=0}^\infty$ be a filtration and $\{\epsilon_t\}_{t=1}^\infty$ be an $\mathbb{R}$-valued stochastic process such that ϵ_t is $\mathcal{F}_t$-measurable for all $t \geq 1$. Moreover, suppose that conditioning on $\mathcal{F}_{t-1}$, ϵ_t is a zero-mean and σ-sub-Gaussian random variable for all $t \geq 1$, that is,*

$$\mathbb{E}[\epsilon_t \,|\, \mathcal{F}_{t-1}] = 0, \qquad \mathbb{E}\big[\exp(\lambda\epsilon_t) \,\big|\, \mathcal{F}_{t-1}\big] \leq \exp(\lambda^2\sigma^2/2), \qquad \forall \lambda \in \mathbb{R}.$$

Meanwhile, let $\{\phi_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that ϕ_t is $\mathcal{F}_{t-1}$-measurable for all $t \geq 1$. Also, let $M_0 \in \mathbb{R}^{d\times d}$ be a deterministic positive-definite matrix and

$$M_t = M_0 + \sum_{s=1}^t \phi_s\phi_s^\top$$

for all $t \geq 1$. For all $\delta > 0$, it holds that

$$\Big\|\sum_{s=1}^t \phi_s\epsilon_s\Big\|_{M_t^{-1}}^2 \leq 2\sigma^2 \log\Big(\frac{\det(M_t)^{1/2}\det(M_0)^{-1/2}}{\delta}\Big)$$

for all $t \geq 1$ with probability at least $1-\delta$.

Lemma F.2 (Abbasi-Yadkori et al. (2011)) *Let $\{\boldsymbol{\phi}_t\}_{t\geq 0}$ be a bounded sequence in $\mathbb{R}^d$ satisfying $\sup_{t\geq 0}\|\boldsymbol{\phi}_t\| \leq 1$. Let $\Lambda_0 \in \mathbb{R}^{d\times d}$ be a positive definite matrix. For any $t \geq 0$, we define $\Lambda_t = \Lambda_0 + \sum_{j=1}^t \boldsymbol{\phi}_j^\top\boldsymbol{\phi}_j$. Then, if the smallest eigenvalue of Λ_0 satisfies $\lambda_{\min}(\Lambda_0) \geq 1$, we have*

$$\log\bigg[\frac{\det(\Lambda_t)}{\det(\Lambda_0)}\bigg] \leq \sum_{j=1}^t \boldsymbol{\phi}_j^\top\Lambda_{j-1}^{-1}\boldsymbol{\phi}_j \leq 2\log\bigg[\frac{\det(\Lambda_t)}{\det(\Lambda_0)}\bigg].$$

The following lemma from Cai et al. (2019) depicts the difference between an estimated value function and the value function under a certain policy.

Lemma F.3 (Extended Value Difference (Cai et al., 2019)) *Let $\pi = \{\pi_h\}_{h=1}^H$ and $\pi' = \{\pi'_h\}_{h=1}^H$ be any two policies and let $\{\hat{Q}_h\}_{h=1}^H$ be any estimated Q-functions. For all $h \in [H]$, we define the estimated value function $\hat{V}_h \colon \mathcal{S} \mapsto \mathbb{R}$ by setting $\hat{V}_h(x) = \langle \hat{Q}_h(x,), \pi_h(\,|\, x)\rangle_{\mathcal{A}}$ for all $x \in \mathcal{S}$. For all $x \in \mathcal{S}$, we have*

$$\begin{aligned}\widehat{V}_1(x) - V_1^{\pi'}(x) &= \sum_{h=1}^{H} \mathbb{E}_{\pi'}\big[\langle \hat{Q}_h(x_h,), \pi_h(\,|\, x_h) - \pi'_h(\,|\, x_h)\rangle_{\mathcal{A}} \,\big|\, x_1 = x\big] \\ &\quad + \sum_{h=1}^{H} \mathbb{E}_{\pi'}\big[\hat{Q}_h(x_h, a_h) - (\mathbb{B}_h \widehat{V}_{h+1})(x_h, a_h) \,\big|\, x_1 = x\big],\end{aligned}$$

where $\mathbb{E}_{\pi'}$ is taken with respect to the trajectory generated by π', while $\mathbb{B}_h$ is the Bellman operator defined in Equation (44).

The following lemma controls the norms of the w's generated by either Algorithm 3 or Algorithm 4 and is used heavily for the concentration analysis.

Lemma F.4 (Bounded Weights of Value Functions (Jin et al., 2020c)) *Let $V_{\max} > 0$ be an absolute constant. For any function $V : \mathcal{S} \to [0, V_{\max}], h \in [H]$, and $(\mathfrak{R}, \pi) \in \{(R, \widehat{\pi}), (\widetilde{R}, \widetilde{\pi}_t^{\ddagger})\} \cup \{(r_i + \widetilde{R}^{-i}, \widetilde{\pi}_t^{\dagger i}), (R^{-i}, *), (\widetilde{R}^{-i}, \dagger), (\widetilde{R}^{-i}, \ddagger), (R^{-i}, \widehat{\pi}^t), (\widetilde{R}^{-i}, \widetilde{\pi}_t^{\dagger i}), (\widetilde{R}^{-i}, \widetilde{\pi}_t^{\ddagger})\}_{i=1}^n$, we have*

$$\|w_h\| \leq \|\theta_h\| + V_{\max}\sqrt{d}, \qquad \|\hat{w}_h^{t,\pi}\|, \|\breve{w}_h^{t,\pi}\| \leq (n + R_{\max})H\sqrt{Kd/\lambda},$$

where $\hat{w}_h^{t,\pi}, \breve{w}_h^{t,\pi}$ are the linear weights associated with the pair $(\mathfrak{R}, \pi)$, w_h parameterizes $(\mathbb{B}_h V)(, ; \mathfrak{R})$, and θ_h parameterizes $\mathfrak{R}$.

Proof Observe that in our setting, the absolute value of the empirical observations of $(\mathbb{B}_h V)(, ; \mathfrak{R})$ is instead $|\mathfrak{R}_h^\tau + \widehat{V}_{h+1}^{t,\pi}(; \mathfrak{R})|$, which is upper bounded by $2(n + R_{\max})H$. Rescaling the Lemma B.1 of Jin et al. (2020c) completes the proof. ■

Lemma F.5 *For all $h \in [H]$ and all $\varepsilon > 0$,we have*

$$\log |\mathcal{N}_h(\varepsilon; L, B, \lambda)| \leq d \log(1 + 4L/\varepsilon) + d^2 \log\big(1 + 8d^{1/2}B^2/(\varepsilon^2\lambda)\big),$$

where the function class

$$\begin{aligned}&\mathcal{V}_h(L, B, \lambda) = \big\{V_h(x; \theta, \beta, \Sigma) \colon \mathcal{S} \to [0, (n + R_{\max})H] \text{ with } \|\theta\| \leq L, \beta \in [0, B], \Sigma \succeq \lambda I\big\} \\ &\text{with } \ V_h(x; \theta, \beta, \Sigma) = \max_{a \in \mathcal{A}} \Big\{\min\big\{\phi(x,a)^\top \theta + \beta\sqrt{\phi(x,a)^\top \Sigma^{-1} \phi(x,a)}, (n + R_{\max})H\big\}\Big\}\end{aligned}$$

and $\mathcal{N}_h(\varepsilon; L, B, \lambda)$ is the ε-cover of $\mathcal{V}_h(L, B, \lambda)$ with respect to the distance $\mathrm{dist}(V, V') = \sup_{x \in \mathcal{S}} \big\|V(x) - V'(x)\big\|$.

Proof See Lemma D.6 in Jin et al. (2020b) for a detailed proof. ■

Lemma F.6 (Concentration of Self-Normalized Processes) *Let $V\colon \mathcal{S}\mapsto[0,(n+R_{\max})(H-1)]$ be any fixed function. For any $h\in[H], p\in(0,1)$, and reward function r, we have*

$$\mathbf{Pr}\Big(\Big\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\epsilon_h^\tau(V;r)\Big\|^2_{(\Lambda_h^t)^{-1}} > (n+R_{\max})^2H^2\big(2\log(1/p)+d\log(1+K/\lambda)\big)\Big)\le p.$$

Proof For the fixed $h\in[H]$ and all $\tau\in\{0,s,K\}$, we define the σ-algebra

$$\mathcal{F}_{h,\tau}=\sigma\big(\{(x_h^j,a_h^j,x_{h+1}^j)\}_{j=1}^{\tau}\cup(x_h^{(\tau+1)\wedge K},a_h^{(\tau+1)\wedge K})\big),$$

where $\sigma(\cdot)$ denotes the σ-algebra generated by a set of random variables and $(\tau+1)\wedge K$ denotes $\min\{\tau+1,K\}$. For all $\tau\in[K]$, we have $\phi(x_h^\tau,a_h^\tau)\in\mathcal{F}_{h,\tau-1}$, as (x_h^τ,a_h^τ) is $\mathcal{F}_{h,\tau-1}$-measurable. Also, for the fixed function $V\colon\mathcal{S}\mapsto[0,(n+R_{\max})(H-1)]$ and all $\tau\in[K]$, we have

$$\epsilon_h^\tau(V;r)=r_h^\tau+V(x_{h+1}^\tau;r)-(\mathbb{B}_hV)(x_h^\tau,a_h^\tau;r)\in\mathcal{F}_{h,\tau},$$

as $(x_h^\tau,a_h^\tau,x_{h+1}^\tau)$ is $\mathcal{F}_{h,\tau}$-measurable. Hence, $\{\epsilon_h^\tau(V)\}_{\tau=1}^K$ is a stochastic process adapted to the filtration $\{\mathcal{F}_{h,\tau}\}_{\tau=0}^K$. Furthermore, we have

$$\begin{aligned}\mathbb{E}\big[\epsilon_h^\tau(V;r)\,\big|\,\mathcal{F}_{h,\tau-1}\big]&=\mathbb{E}\big[r_h^\tau+V(x_{h+1}^\tau;r)\,\big|\,\{(x_h^j,a_h^j,x_{h+1}^j)\}_{j=1}^{\tau-1},(x_h^\tau,a_h^\tau)\big]-(\mathbb{B}_hV)(x_h^\tau,a_h^\tau;r)\\&=\mathbb{E}\big[r_h^\tau+V(s_{h+1})\,\big|\,s_h=x_h^\tau,a_h=a_h^\tau\big]-(\mathbb{B}_hV)(x_h^\tau,a_h^\tau;r)=0,\end{aligned}$$

where the first step is because $(\mathbb{B}_hV)(x_h^\tau,a_h^\tau;r)$ is $\mathcal{F}_{h,\tau-1}$-measurable and the second step follows from the Markov property of the process. Moreover, as $(\mathbb{B}_hV)(x_h^\tau,a_h^\tau;r)\in[0,(n+R_{\max})H]$, we have $|\epsilon_h^\tau(V;r)|\le(n+R_{\max})H$. Hence, the random variable $\epsilon_h^\tau(V;r)$ defined in Equation (49) is mean-zero and $(n+R_{\max})H$-sub-Gaussian conditioning on $\mathcal{F}_{h,\tau-1}$.

Invoke Lemma F.1 with $M_0=\lambda I$ and $M_k=\lambda I+\sum_{\tau=1}^k\phi(x_h^\tau,a_h^\tau)\ \phi(x_h^\tau,a_h^\tau)^\top$ for all $k\in[K]$. We then know that

$$\mathbf{Pr}\Big(\Big\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\epsilon_h^\tau(V;r)\Big\|^2_{(\Lambda_h^t)^{-1}} > 2(n+R_{\max})^2H^2\log\Big(\frac{\det(\Lambda_h^t)^{1/2}}{p\det(\lambda I)^{1/2}}\Big)\Big)\le p \tag{78}$$

for all $p\in(0,1)$. Here, we use the fact that $\Lambda_h^t=M_k$. To upper bound $\det(\Lambda_h^t)^{1/2}$, we first notice that

$$\|\Lambda_h^t\|_{\mathrm{op}}=\Big\|\lambda I+\sum_{\tau=1}^K\phi(x_h^\tau,a_h^\tau)\phi(x_h^\tau,a_h^\tau)^\top\Big\|_{\mathrm{op}}\le\lambda+\sum_{\tau=1}^K\|\phi(x_h^\tau,a_h^\tau)\phi(x_h^\tau,a_h^\tau)^\top\|_{\mathrm{op}}\le\lambda+K,$$

where the first inequality follows from the triangle inequality of operator norm and the second inequality follows from the fact that $\|\phi(x,a)\|\le1$ for all $(x,a)\in\mathcal{S}\times\mathcal{A}$ by our assumption. This implies $\det(\Lambda_h^t)\le(\lambda+K)^d$. Combining with the fact that $\det(\lambda I)=\lambda^d$ and Equation (78), we have

$$\mathbf{Pr}\Big(\Big\|\sum_{\tau=1}^{K}\phi(x_h^\tau,a_h^\tau)\epsilon_h^\tau(V;r)\Big\|^2_{(\Lambda_h^t)^{-1}} > (n+R_{\max})^2H^2\big(2\log(1/p)+d\log(1+K/\lambda)\big)\Big)\le p.$$

Therefore, we conclude the proof of Lemma F.6. ■

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320, 2011.

Susan Athey and Ilya Segal. An efficient dynamic mechanism. *Econometrica*, 81(6): 2463–2485, 2013.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pages 551–560. PMLR, 2020.

Abhishek Bapna and Thomas A Weber. Efficient dynamic allocation with uncertain valuations. *Available at SSRN 874770*, 2005.

Jorge Barrera and Alfredo Garcia. Dynamic incentives for congestion control. *IEEE Transactions on Automatic Control*, 60(2):299–310, 2014.

Arman Kiani Bejestani and Anuradha Annaswamy. A dynamic mechanism for wholesale energy market: Stability and robustness. *IEEE Transactions on Smart Grid*, 5(6): 2877–2888, 2014.

Dirk Bergemann and Alessandro Pavan. Introduction to symposium on dynamic contracts and mechanism design. *Journal of Economic Theory*, 159:679–701, 2015.

Dirk Bergemann and Juuso Välimäki. Efficient dynamic auctions. Technical report, Cowles Foundation for Research in Economics, Yale University, 2006.

Dirk Bergemann and Juuso Välimäki. The dynamic pivot mechanism. *Econometrica*, 78(2): 771–789, 2010.

Dirk Bergemann and Juuso Välimäki. Dynamic mechanism design: An introduction. *Journal of Economic Literature*, 57(2):235–74, 2019.

Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Ruggiero Cavallo. Efficiency and redistribution in dynamic mechanism design. In *Proceedings of the 9th ACM conference on Electronic commerce*, pages 220–229, 2008.

Ruggiero Cavallo. Mechanism design for dynamic settings. *ACM SIGecom Exchanges*, 8(2): 1–5, 2009.

Ruggiero Cavallo, David C Parkes, and Satinder Singh. Efficient mechanisms with dynamic populations and dynamic types. *Harvard University Technical Report*, 2009.

M Keith Chen and Michael Sheldon. Dynamic pricing in a labor market: Surge pricing and flexible work on the Uber platform. *Ec*, 16:455, 2016.

Xiaoyu Chen, Jiachen Hu, Lin F Yang, and Liwei Wang. Near-optimal reward-free exploration for linear mixture MDPs with plug-in solver. *arXiv preprint arXiv:2110.03244*, 2021.

Edward H Clarke. Multipart pricing of public goods. *Public choice*, pages 17–33, 1971.

Claude d'Aspremont and Louis-André Gérard-Varet. Incentives and incomplete information. *Journal of Public economics*, 11(1):25–45, 1979.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Alessandro Epasto, Mohammad Mahdian, Vahab Mirrokni, and Song Zuo. Incentive-aware learning for large markets. In *Proceedings of the 2018 World Wide Web Conference*, pages 1369–1378, 2018.

Eric J Friedman and David C Parkes. Pricing WiFi at Starbucks: issues in online mechanism design. In *Proceedings of the 4th ACM conference on Electronic commerce*, pages 240–241, 2003.

Theodore Groves. Efficient collective choice when compensation is possible. *The Review of Economic Studies*, 46(2):227–241, 1979.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020b.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? *arXiv preprint arXiv:2012.15085*, 2020c.

Kirthevasan Kandasamy, Joseph E Gonzalez, Michael I Jordan, and Ion Stoica. Vcg mechanism design with unknown agent values under stochastic bandit feedback. *arXiv preprint arXiv:2004.08924*, 2020.

Anna R Karlin and Yuval Peres. *Game theory, alive*, volume 101. American Mathematical Soc., 2017.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pages 865–891. PMLR, 2021.

Dingwen Kong, Ruslan Salakhutdinov, Ruosong Wang, and Lin F Yang. Online subsampling for reinforcement learning with general function approximation. *arXiv preprint arXiv:2106.07203*, 2021.

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

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

Boxiang Lyu, Zhaoran Wang, Mladen Kolar, and Zhuoran Yang. Pessimism meets vcg: Learning dynamic mechanism design via offline reinforcement learning. *arXiv preprint arXiv:2205.02450*, 2022.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pages 7599–7608. PMLR, 2021.

Sobhan Miryoosefi and Chi Jin. A simple reward-free approach to constrained reinforcement learning. *arXiv preprint arXiv:2107.05216*, 2021.

Roger B Myerson. Mechanism design. In *Allocation, Information and Markets*, pages 191–206. Springer, 1989.

Hamid Nazerzadeh, Amin Saberi, and Rakesh Vohra. Dynamic cost-per-action mechanisms and applications to online advertising. In *Proceedings of the 17th international conference on World Wide Web*, pages 179–188, 2008.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic Game Theory.* Cambridge University Press, 2007.

David C Parkes. Online mechanisms. In N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, editors, *Algorithmic Game Theory*, pages 411–439. Cambridge University Press, 2007.

David C Parkes and Satinder Singh. An mdp-based approach to online mechanism design. *Advances in neural information processing systems*, 16, 2003.

David C Parkes, Satinder Singh, and Dimah Yanovsky. Approximately efficient online mechanism design. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, pages 1049–1056, 2004.

Alessandro Pavan, Ilya R Segal, and Juuso Toikka. Dynamic mechanism design: Incentive compatibility, profit maximization and information disclosure. *Profit Maximization and Information Disclosure (May 1, 2009)*, 2009.

Alessandro Pavan, Ilya Segal, and Juuso Toikka. Dynamic mechanism design: A Myersonian approach. *Econometrica*, 82(2):601–653, 2014.

Shuang Qiu, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. On reward-free rl with kernel and neural function approximations: Single-agent mdp and markov game. In *International Conference on Machine Learning*, pages 8737–8747. PMLR, 2021.

Max Simchowitz and Aleksandrs Slivkins. Exploration and incentives in reinforcement learning. *Operations Research*, 2023.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

Andrew Wagenmaker, Yifang Chen, Max Simchowitz, Simon S Du, and Kevin Jamieson. Reward-free RL is no harder than reward-aware RL in linear Markov decision processes. *arXiv preprint arXiv:2201.11206*, 2022.

R. Wang, S. S. Du, L. F. Yang, and R. Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.12274*, 2020.

Jingfeng Wu, Lin Yang, et al. Accommodating picky customers: Regret bound and exploration complexity for multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Lin Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael Jordan. Provably efficient reinforcement learning with kernel and neural function approximations. *Advances in Neural Information Processing Systems*, 33:13903–13916, 2020a.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces. *arXiv preprint arXiv:2011.04622*, 2020b.

Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. *Advances in Neural Information Processing Systems*, 33:11756–11766, 2020.

Zihan Zhang, Simon Du, and Xiangyang Ji. Near optimal reward-free reinforcement learning. In *International Conference on Machine Learning*, pages 12402–12412. PMLR, 2021.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pages 11492–11502. PMLR, 2020a.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021a.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted MDPs with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021b.

Huozhi Zhou, Jinglin Chen, Lav R Varshney, and Ashish Jagmohan. Nonstationary reinforcement learning with linear function approximation. *arXiv preprint arXiv:2010.04244*, 2020b.