

A Reinforcement Learning Approach in Multi-Phase Second-Price Auction Design

Rui Ai

RUIAI@MIT.EDU

*Institute for Data, Systems, and Society
Massachusetts Institute of Technology
Cambridge, MA 02139, USA*

Boxiang Lyu

BLYU@CHICAGOBOOTH.EDU

*Booth School of Business
The University of Chicago
Chicago, IL 60637, USA*

Zhaoran Wang

ZHAORANWANG@GMAIL.COM

*Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208, USA*

Zhuoran Yang

ZHUORAN.YANG@YALE.EDU

*Department of Statistics and Data Science
Yale University
New Haven, CT 06511, USA*

Michael I. Jordan

JORDAN@CS.BERKELEY.EDU

*Department of EECS, Department of Statistics
University of California
Berkeley, CA 94720, USA*

Editor: Tor Lattimore

Abstract

We study reserve price optimization in multi-phase second price auctions, where the seller’s prior actions affect the bidders’ later valuations through a Markov Decision Process (MDP). Compared to the bandit setting in existing works, the setting in ours involves three challenges. First, from the seller’s perspective, we need to efficiently explore the environment in the presence of potentially untruthful bidders who aim to manipulate the seller’s policy. Second, we want to minimize the seller’s revenue regret when the market noise distribution is unknown. Third, the seller’s per-step revenue is an unknown, nonlinear random variable, and cannot even be directly observed from the environment but realized values.

We propose a mechanism addressing all three challenges. To address the first challenge, we use a combination of a new technique named “buffer periods” and inspirations from Reinforcement Learning (RL) with low switching cost to limit bidders’ surplus from untruthful bidding, thereby incentivizing approximately truthful bidding. The second one is tackled by a novel algorithm that removes the need for pure exploration when the market noise distribution is unknown. The third challenge is resolved by an extension of LSVI-UCB, where we use the auction’s underlying structure to control the uncertainty of the revenue function. The three techniques culminate in the Contextual-LSVI-UCB-Buffer (CLUB) algorithm which achieves $\tilde{O}(H^{5/2}\sqrt{K})$ revenue regret, where K is the number of episodes and

H is the length of each episode, when the market noise is known and $\tilde{O}(H^3\sqrt{K})$ revenue regret when the noise is unknown with no assumptions on bidders’ truthfulness.

Keywords: Mechanism Design, Second Price Auction, Reserve Price Optimization, Reinforcement Learning

1. Introduction

Second price auction with reserve prices is one of the most popular auctions both in theory (Nisan et al., 2007) and in practice (Roth and Ockenfels, 2002). While closed-form expressions for the optimal reserve price have been known ever since the seminal work of Myerson (1981), directly applying the result requires that population information, such as the bidders’ valuations’ distribution, is known a priori. Various attempts have been made to weaken the assumption, with one of the most prominent lines of literature being reserve price optimization for repeated auctions in the contextual bandit setting (Amin et al., 2014; Golrezaei et al., 2019; Javanmard and Nazerzadeh, 2019; Deng et al., 2020).

A limitation of existing works lies in the bandit assumption. Indeed, while reserve price optimization is already challenging as-is, allowing the auction to be both contextual and introducing temporally dependent dynamics, particularly, incorporating Markov Decision Process (MDP) induced dynamics in the evolution of bidders’ preferences, opens up a wider range of problems for studying. For example, Dolgov and Durfee (2006) studies optimal auction under the setting and developed novel resource allocation mechanisms, Jiang et al. (2015) leverages both MDP and auctions to better analyze resource allocation in IaaS cloud computing, and Zhao et al. (2018) uses deep Reinforcement Learning (RL) to study sponsored search auctions. We refer interested readers to Athey and Segal (2013) for more motivating examples. A question naturally arises: is it possible to optimize reserve prices when bidders’ preferences evolve according to MDPs?

In this article, we provide an affirmative answer. Our work assumes that the state of the auction is affected by the state and the seller’s action in the preceding step. To facilitate interpretation, we refer to the seller’s action in this context as “item choice”: bidders’ later preferences could be affected by the types of items sold in previous rounds, a phenomenon well-documented by empirical works in auctions (Lusht, 1994; Jones et al., 2004; Lange et al., 2010; Ginsburgh and Van Ours, 2007).

As is the case in many real-world problems, we assume that the underlying transition dynamics and the bidder’s valuations are both unknown. We further emphasize that we do not make any truthfulness assumptions about the bidders, allowing them to be strategic with their reporting. Under such a challenging setting, our goal is to learn the optimal policy of the seller in the unknown environment, in the presence of untruthful bidders.

Our Contributions. We begin by summarizing the three key challenges we face. First, bidders have the incentive to report their valuation untruthfully, in hopes of manipulating the seller’s learned policy, through either overbidding or underbidding, making it difficult to estimate their true preferences and the underlying MDP dynamics. Existing works such as Amin et al. (2014); Golrezaei et al. (2019); Deng et al. (2020) do not apply due to technical challenges unique to MDP. Second, when the market noise distribution is unknown, even in the bandit setting existing literature often only obtains $\tilde{O}(K^{2/3})$ guarantee (Amin et al., 2014; Golrezaei et al., 2019) and $\Omega(K^{2/3})$ revenue regret lower bound exists in the worst

case (Kleinberg and Leighton, 2003). Third, the seller’s reward function, namely revenue, is unknown, nonlinear, and can not be directly observed from the bidders’ submitted bids, and LSVI-UCB (Jin et al., 2020) cannot be directly applied.

We are able to address all three challenges with the CLUB algorithm. Motivated by the ever increasing learning periods in existing works (Amin et al., 2014; Golrezaei et al., 2019; Deng et al., 2020), our work further draws inspiration from RL with low switching cost (Wang et al., 2021) and proposes a novel concept dubbed “buffer periods” to ensure that the bidders are sufficiently truthful. Additionally, we feature a novel algorithm we dub “simulation” which, combined with a novel proof technique leveraging the Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956), yields $\tilde{O}(\sqrt{K})$ revenue regret under only mild additional assumptions. Finally, by exploiting the mathematical properties of the revenue function, our work provides a provably efficient RL algorithm for when the reward function is nonlinear.

1.1 Related Works

We summarize below two lines of existing literature pertinent to our work.

Reserve Price Optimization. There is a vast amount of literature on price estimation (Cesa-Bianchi et al., 2014; Qiang and Bayati, 2016; Shah et al., 2019; Drutsa, 2020; Kanoria and Nazerzadeh, 2020; Keskin et al., 2021; Guo et al., 2022). Deng et al. (2020) considers a model where buyers and sellers are equipped with different discount rates, proposing a robust mechanism for revenue maximization in contextual auctions. Javanmard et al. (2020) proposes an algorithm with $\tilde{O}(\sqrt{T})$ regret ($T = KH$ in our paper) while Fan et al. (2021) achieves sublinear regret in a more complex setting. Cesa-Bianchi et al. (2014) studies reserve price optimization in non-contextual second price auctions, obtaining $\tilde{O}(\sqrt{T})$ revenue regret bound. Drutsa (2017, 2020) studies revenue maximization in repeated second-price auctions with one or multiple bidders, proposing an algorithm with a $\mathcal{O}(\log \log T)$ worst-case regret bound. However, their setting is non-contextual, and they cannot be applied to our setting.

Among this line of research, Golrezaei et al. (2019, 2023) are possibly the closest to our work. Golrezaei et al. (2019) assumes a linear stochastic contextual bandit setting, where the contexts are independent and identically distributed, achieving $\tilde{O}(1)$ regret when the market noise distribution is known and $\tilde{O}(K^{2/3})$ when it is unknown and nonparametric. While the $\tilde{O}(1)$ regret under known market noise distribution seems to be better than our bound, we emphasize that their stochastic bandit setting does not require exploration over the action space but needed in our work and, even in generic linear MDPs, a $\Omega(\sqrt{K})$ regret lower bound exists (Jin et al., 2020). For unknown distribution, there’s another difference that they consider a time-varying model while we focus on dealing with the underlying fixed MDP. Though the difficulty of these tasks is hard to compare directly, Amin et al. (2014) considers a non-parametric but fixed distribution setting and suffers $\tilde{O}(K^{2/3})$ regret, which may hint at the main difficulty comes from a non-parametric rather than time-varying setting. We delay more discussion about concrete techniques in Golrezaei et al. (2019) in Section A. Lastly, as we discussed previously, the approaches in Golrezaei et al. (2019) cannot be directly applied in the MDP setting, necessitating our novel algorithmic structure.

At the same time with our paper, [Golrezaei et al. \(2023\)](#) considers another pricing problem with non-parametric noise, achieving $\tilde{O}(\sqrt{T})$ regret. However, they only set a reserve price for all bidders, while we customize reserve prices for each bidder to attain more revenue. On the one hand, the seller will achieve more revenue by setting different reserve prices for different bidders, which is in line with the goal of the seller because there are fewer corresponding constraints. On the other hand, in the real world, it is more common to set up personalized reserve prices in the online advertisement market, like price discrimination ([Paes Leme et al., 2016](#); [Wu et al., 2019](#)). Additionally, [Golrezaei et al. \(2023\)](#) is in the scope of contextual bandits and is a special case of our MDP setting. Pricing in contextual bandit settings is much easier than MDP because i.i.d. context will form a positive definite covariance matrix, and linear regression works well. But in MDP, features depend on action and are absolutely not i.i.d. Without the positive definite assumption, algorithms designed for contextual bandits lose effectiveness, and we need innovative algorithms to incorporate pricing and complex information structures.

RL with Linear Function Approximation. Linear contextual bandit is a popular model for online decision making ([Rusmevichientong and Tsitsiklis, 2010](#); [Abbasi-Yadkori et al., 2011](#); [Chu et al., 2011](#); [Li et al., 2019](#); [Lattimore and Szepesvári, 2020](#)) that has also been extensively studied from the auction design perspective ([Amin et al., 2014](#); [Golrezaei et al., 2019](#)). Its dynamic counterpart, Linear MDP, remains popular in the analysis of provably efficient RL ([Yang and Wang, 2019](#); [Jin et al., 2020, 2021b](#); [Yang et al., 2020](#); [Zanette et al., 2020](#); [Jin et al., 2021a](#); [Uehara et al., 2021](#); [Yu et al., 2022](#); [Wang et al., 2021](#); [Gao et al., 2021](#)). In particular, [Jin et al. \(2020\)](#) is one of the first papers to introduce the concept, proposing a provably efficient RL algorithm with $\tilde{O}(\sqrt{K})$ regret. [Jin et al. \(2021b\)](#) generalizes the idea to offline RL.

While we use linear function approximation, the seller’s per-step reward function, revenue, is non-linear. Our work also features novel per-step optimization problems to combat effects from untruthful reporting. While our work draws inspiration from [Wang et al. \(2020b\)](#) and [Gao et al. \(2021\)](#), as we discussed previously, these inspirations are needed for obtaining high-quality estimates when the bidders are untruthful. Thus, our work differs significantly from prior works on linear MDPs.

Notations. For any positive integer n we let $[n]$ denote the set $\{1, \dots, n\}$. For any set A we let $\Delta(A)$ denote the set of probability measures over A . For sets A, B , we let $A \times B$ be the Cartesian product of the two. During the whole paper, we use $k \in [K]$ to refer to an episode and $h \in [H]$ to refer to a horizon. In addition, we use \tilde{k} to refer to a buffer period associated with the k -th episode.

2. Preliminaries

We consider a repeated (lazy) multi-phase second-price auction with personalized reserve prices. Particularly, we assume that there are N rational bidders, indexed by $[N]$, and one seller participating in the auction. For ease of presentation, we use “he” to refer to a specific bidder and “she” to the seller.

Second Price Auction with Personalized Reserve Prices. We begin by describing a single round of the auction. Each bidder $i \in [N]$ submits some bid $b_i \in \mathbb{R}_{\geq 0}$ and the seller determines the personalized reserve prices for the bidders in the form of reserve price vector

$\rho \in \mathbb{R}_{\geq 0}^N$, with ρ_i denoting bidder i 's reserve price. The bidder with the highest bid only wins if he also clears his personal reserve price, i.e., $b_i \geq \rho_i$. If the bidder i receives the item, he pays the seller the maximum of his personalized reserve and the second highest bid, namely $\max\{\rho_i, \max_{j \neq i} b_j\}$, which we dub m_i for simplicity. When the bidder with the highest bid fails to clear his personalized reserve price, the auction fails, the seller gains zero, and the item remains unsold. In summary, bidder i receives the item if and only if $b_i \geq m_i$ and the price he pays is m_i . For any round of auction, we let $q_i = \mathbb{1}(\text{bidder } i \text{ receives the item})$ indicate whether bidder i received the item or not. For the sake of convenience, throughout the paper, we assume that there are no ties in the submitted bids.

A Multi-Phase Second Price Auction. We now characterize the dynamics of the multi-phase auction setting we study. Assume that the transition dynamic between rounds can be modeled as an episodic Markov Decision Process (MDP)¹. A multi-phase second price auction with personalized reserves is parameterized as $(\mathcal{S}, \Upsilon, H, \mathbb{P}, \{r_i\}_{i=1}^N)$, with the state space denoted by \mathcal{S} , seller's item choice space Υ ², horizon H , transition kernel $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ where $\mathbb{P}_h : \mathcal{S} \times \Upsilon \rightarrow \Delta(\mathcal{S})$, and the individual bidders' reward functions $r_i = \{r_{ih}\}_{h=1}^H$ for all $i \in [N]$. The choice of item $v \in \Upsilon$ affects the bidders' rewards as well as the transition.

The interaction between the bidders and the seller is then defined as follows. We assume without loss of generality that the state at the initial step is fixed at some $x_1 \in \mathcal{S}$. For each $h \in [H]$, the seller and the bidders engage in a single round of the second-price auction. Given the seller's item choice at step h , v_h , nature transitions to the next state according to the transition kernel \mathbb{P}_h .

Bidder Rewards. We assume that for each bidder $i \in [N]$ at time $h \in [H]$, his reward³ depends on both the state x and item being auctioned off at that round $v \in \Upsilon$, which we formalize as

$$r_{ih}(x, v) = 1 + \mu_{ih}(x, v) + z_{ih}, \text{ where } z_{ih} \stackrel{\text{i.i.d.}}{\sim} F.$$

Here, z_{ih} denotes the randomness within bidders' rewards and is drawn i.i.d. from the market noise distribution $F(\cdot)$. We assume that $F(\cdot)$ is supported on $[-1, 1]$ and has mean 0. Let $\mu_{i,h} : \mathcal{S} \times \Upsilon \rightarrow [0, 1]$ denote the conditional expectation of the reward less one, where the constant is added to ensure $r_{ih}(x, v) \in [0, 3]$.

Policies and Value Functions. Before we describe the seller's policy, we first discuss the action space $\mathcal{A} = \Upsilon \times \mathbb{R}_{\geq 0}^N$. At each $h \in [H]$, the seller chooses some action $a_h = (v_h, \rho_h)$, comprising of item choice $v \in \Upsilon$ and reserve price vector $\rho \in \mathbb{R}_{\geq 0}^N$. The seller's policy is then $\pi = \{\pi_h\}_{h=1}^H$, where $\pi_h : \mathcal{S} \rightarrow \Delta(\mathcal{A})$. We let π^v and π^ρ denote the marginal item choice and reserve price policies, respectively. Recall that the seller garners revenue only when the item is sold to a bidder. At each $h \in [H]$, her per-step expected revenue is then

$$R_h = \mathbb{E}_{\{z_{ih}\}_{i=1}^N} \left[\sum_{i=1}^N m_{ih} \mathbb{1}(m_{ih} \leq b_{ih}) \right] \quad (1)$$

-
1. We can easily extend our setting to that of an infinite-horizon MDP by improperly learning the process as an episodic one. Here we focus on the finite-horizon case purely for simplicity of presentation.
 2. Here we use "item choice" to better illustrate what Υ intuitively represents. The term can be extended to more generic notions of the seller's action.
 3. We use the term "reward" to maintain consistency with existing RL literature.

as we recall that $m_{ih} = \max\{\rho_{ih}, \max_{j \neq i} b_{jh}\}$ and bidder i pays the seller m_{ih} if and only if $b_{ih} \geq m_{ih}$. The value function (V-function) of the seller's revenue for any policy π and the action-value function (Q-function) is $Q_h^\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ are then

$$V_h^\pi(x) = \mathbb{E}_\pi \left[\sum_{h'=h}^H R_{h'}(x_{h'}, a_{h'}) \mid x_h = x \right]$$

and

$$Q_h^\pi(x, a) = \mathbb{E}_\pi \left[\sum_{h'=h}^H R_{h'}(x_{h'}, a_{h'}) \mid x_h = x, a_h = a \right],$$

respectively.

Since the bidder reward only depends on state x and the choice of item v instead of reserve ρ , we have a family of mappings from $\mathcal{S} \times \Upsilon$ to $\mathbb{R}_{\geq 0}^N$ that determines ρ . Therefore, with a slight abuse of notation, we can rewrite our Q-function as $Q(x, a) = Q(x, (v, \rho(x, v)))$, restricting the role of setting reserve prices using such mappings without loss of generality. From now on, we use $Q(x, v)$ to denote the Q-function for simplicity. For any function $f : \mathcal{S} \rightarrow \mathbb{R}$, we define the transition operator \mathcal{P} and the Bellman operator \mathcal{B} as

$$(\mathcal{P}_h f)(x, a) = \mathbb{E}[f(x_{h+1}) \mid x_h = x, a_h = a], \quad (\mathcal{B}_h f)(x, a) = \mathbb{E}[R_h(x_h, a_h)] + (\mathbb{P}_h f)(x, a),$$

respectively. Finally, we let π^* denote the optimal policy when the bidders' reward functions, the MDP's underlying transition, and the market noise distribution are all known to the seller. We remark that when these parameters are known, second price auctions with personalized reserve prices are inherently incentive compatible and rational bidders will bid truthfully.

Performance Metric. The revenue suboptimality for each episode $k \in [K]$ is

$$\text{SubOpt}_k(\pi_k) = V_1^{\pi^*}(x_1) - V_1^{\pi_k}(x_1),$$

with π_k being the strategy used in episode k . Our evaluation metric is then the revenue regret attained over K episodes, namely

$$\text{Regret}(K) = \sum_{k=1}^K \text{SubOpt}_k(\pi_k). \quad (2)$$

Impatient Utility-Maximizing Bidders. We assume the bidders are equipped with some discount rate $\gamma \in (0, 1)$ while the seller's reward is not discounted. For the sake of simplicity, we assume γ is common knowledge. [Drutsa \(2020\)](#) consider a scenario where γ is unknown but with a strictly less than one upper bound. We highlight that it also works with our CLUB algorithm as long as we replace γ with its upper bound. We can have regret bounds with the same order because we adopt more conservative estimators, and buyers won't violate as much as the corresponding results of γ . Then all results in our paper hold up to some changes of absolute constants. Rational bidder i 's utility at step h is given by $(r_{ih}(x_h, v_h) - m_{ih}) \mathbb{1}(b_{ih} \geq m_{ih})$, as we note that he only receives nonzero utility upon winning the auction. His objective is to maximize his discounted cumulative utility

$$\text{Utility}_i = \sum_{k=1}^K \gamma^k \mathbb{E}_{\pi_k} \left[\sum_{h=1}^H (r_{ih}(x_h^k, v_h^k) - m_{ih}^k) \mathbb{1}(b_{ih}^k \geq m_{ih}^k) \mid x_1^k = x_1 \right].$$

Note that in practical applications, sellers are usually more patient than bidders and discount their future rewards less. Consider a sponsored search auction, where the seller usually auctions off large numbers of ad slots every day. Bidders usually urgently need advertisements and value future rewards less. On the other hand, the seller is not especially concerned with slight decreases in immediate rewards. We refer the readers to [Drutsa \(2017\)](#); [Golrezaei et al. \(2019\)](#) for a more detailed discussion on the economic justifications of the assumption and emphasize that the assumption is necessary, as [Amin et al. \(2013\)](#) shows that when the bidders are as patient as the seller, achieving sub-linear revenue regret is impossible.

Linear Markov Decision Process. As a concrete setting, we study linear function approximation.

Assumption 1. *Assume that there exists known feature mapping $\phi : \mathcal{S} \times \Upsilon \rightarrow \mathbb{R}^d$ such that there exist d -dimension unknown (signed) measures \mathcal{M}_h over \mathcal{S} and unknown vectors $\{\theta_{ih}\}_{i=1}^N \in \mathbb{R}^d$ that satisfy*

$$\mathbb{P}_h(x'|x, v) = \langle \phi(x, v), \mathcal{M}_h(x') \rangle, \mu_{ih}(x, v) = \langle \phi(x, v), \theta_{ih} \rangle$$

for all $(x, v, x') \in \mathcal{S} \times \Upsilon \times \mathcal{S}$, $i \in [N]$, and $h \in [H]$. Without loss of generality, we assume that $\|\phi(x, v)\| \leq 1$ for all $(x, v) \in \mathcal{S} \times \Upsilon$, $\|\mathcal{M}_h(\mathcal{S})\| \leq \sqrt{d}$, and $\|\theta_{ih}\| \leq \sqrt{d}$ for all $h \in [H]$ and $i \in [N]$.

There're some scenarios in reality that mapping $\phi(\cdot, \cdot)$ is public knowledge like representing the order of items. However, for unknown mapping ([Lattimore et al., 2020](#)), there are some ways to pre-train features using a reproducing kernel Hilbert space, neural networks, or the Knowledge Discovery in Databases (KDD) method ([Lange and Riedmiller, 2010](#); [Claessens et al., 2016](#); [Wang et al., 2020a](#)). Utilizing these, we can obtain a working feature representation in practice.

We remark that while the transition kernel \mathbb{P}_h and the bidders' individual expected reward functions $\{\mu_i\}_{i=1}^N$ are linear, the seller's objective, revenue, is not linear, differentiating our work from typical linear MDP literature (see [Yang and Wang \(2019\)](#); [Jin et al. \(2020\)](#) for representative works).

Motivations for the MDP Model . We close off the section by providing some practical applications of our MDP model. The core of our setting is to study what will happen when selling heterogeneous goods and how the order, part of the state, will affect the revenue. We provide three real-world scenarios to motivate this phenomenon.

- **(Online Advertisement)** Google sells lots of advertising positions every day, while buyers face budget constraints. In the early rounds, since buyers have more budget left, they are usually eager to bid higher and have a stronger willingness to pay. Therefore, Google may want to sell the most valuable position at first so that buyers have the ability to pay higher acceptable prices and avoid being underbid and unsold.
- **(Antique Auction)** For traditional auction design, the prior auctions may affect the latter auctions. For instance, consider when Sotheby's wants to sell several antiques. The order of selling is of significance, and that's the reason why Sotheby's needs to sell a few other pieces to warm up before selling the final flagship piece. The order, part of

the state, influences people’s valuation and, consequently, total revenue. For example, if Sotheby’s wishes to auction a valuable Chinese ancient artifact, they would auction some related artifacts during the warm-up session to enhance buyers’ expectations.

- **(Automobile Sales Market)** The last example is on the market of cars. If one buyer wants to buy a sedan from General Motors, recommending Chevrolet first or Cadillac first will influence his preference for the course. If he sees Chevrolet first, he may think Cadillac is too expensive. However, if he sees Cadillac first, he may think Chevrolet lacks a sense of experiential quality. To achieve maximum profitability, General Motors carefully arranges the recommended order. In a broader sense, they meticulously design the sequence in which cars appear in advertisements.

All in all, contextual bandits lack the ability to depict such kinds of problems. We need to use MDP to model these issues.

3. Known Market Noise Distribution

We remind the readers of our three main challenges, with the first challenge being exploring the environment, even when the bidders submit their bids potentially untruthfully. The second challenge emerges only when the market noise distribution is unknown, and we defer its resolution to Section 4. The third challenge is performing provably efficient RL even when the seller’s per-step revenue, detailed in Equation (1), is nonlinear and not directly observable.

In this section, we present a version of CLUB when the market noise distribution is known. We assume for convenience that K is known, as we can use the doubling trick (see Auer et al. (2002) and Besson and Kaufmann (2018) for discussions) to achieve the same order of regret when K is unknown or infinite. Since we can utilize the doubling trick to partition K into at most $\lceil \log_2 K \rceil + 1$, adding corresponding regret will lead to a regret bound of the same order up to some logarithmic terms.

3.1 CLUB Algorithm When $F(\cdot)$ is Known

We start with the first challenge, which we address by a collection of algorithms that successfully induce approximately truthful bids from the bidders.

Addressing Challenge 1: Untruthfulness. To curb the sellers’ untruthfulness, we need to punish such behavior, achieved through a random pricing policy in the form of Algorithm 1. For each $h \in [H]$, π_{rand} randomly chooses an item and a bidder, offering him the item with a reserve price drawn uniformly at random. The bidder’s utility decreases whenever he reports untruthfully, risking either not receiving the item when he underbids or overpaying for an item when he overbids. Combining lazy updates (see Algorithm 2), we can ensure approximate truthfulness because with the discount rate being less than one, the benefit the bidder gains from misreporting the bids will decay as the timestep increases. However, since we consider a multi-phase auction design, there remains some nuisance introduced by MDP. For instance, there is no guarantee of a positive definite covariance matrix, and it’s challenging to give a low regret union bound. In other words, $\sum \phi\phi^T$ might have some zero eigenvalues. We will see how to solve them in the following paragraphs.

Algorithm 1 Definition of π_{rand}

- 1: **for** $h = 1, \dots, H$ **do**
 - 2: Randomly chooses an item $v_h \in \Upsilon_h$.
 - 3: Choose a bidder $i \in [N]$ uniformly at random and offer him the item with reserve price $\rho_{ih} \sim \text{Unif}([0, 3])$. Set other bidders' reserve prices to infinity.
 - 4: **end for**
-

We further introduce a novel technique, “buffer period”, which explicitly forces the bidders to wait before starting a new learning period, thereby decreasing the discounted utility the impatient bidders may gain from untruthfulness. Indeed, a typical algorithm in the bandit setting only features π_{rand} and a sequence of learning periods that double in length (Amin et al., 2014; Golrezaei et al., 2019; Deng et al., 2020). In the bandit setting, data collected in all previous periods is used to update the policy at the end of each period. The increasingly lengthy periods ensure that the seller switches policy less frequently, ensuring that the impatient buyers need to wait longer before benefiting from untruthful reporting, deterring them from doing so. Unfortunately, the same technique does not work for MDPs, as the rate at which the smallest eigenvalue of the covariance matrix estimate grows cannot be determined, and we cannot ensure our estimate of the underlying environment is not “stale” when we double the length of the periods.

Algorithm 2 Buffer Period with Known $F(\cdot)$

- 1: Receives buffer start $\text{buffer.s}(\tilde{k} + 1) = k$ and end $\text{buffer.e}(\tilde{k} + 1) = k + \frac{3 \log K}{\log(1/\gamma)}$.
 - 2: Do nothing for all episodes $\text{buffer.s}(\tilde{k} + 1) \leq k < \text{buffer.e}(\tilde{k} + 1)$, i.e., do nothing during the buffer period before the end.
 - 3: At the end of the buffer period, update policy estimate $\pi_{\tilde{k}+1}$ and Q-function estimate $\hat{Q}_h^{\pi_{\tilde{k}+1}}(\cdot, \cdot)$ using Algorithm 4, and then increment buffer period counter $\tilde{k} \leftarrow \tilde{k} + 1$.
-

While we can mimic the aforementioned bandit algorithms by drawing inspiration from low-switching cost RL literature, we cannot guarantee that the periods are sufficiently long without buffer periods. Indeed, we can use the smallest eigenvalue of the covariance matrix to determine when to start a new period. However, it is impossible to determine a priori the rate at which the smallest eigenvalue grows. Buffer periods ensure that each period is sufficiently long, deferring any utility gain from untruthful reporting. Combined with the bidders' discount rate, a combination of π_{rand} and buffer periods ensures that the bidders behave approximately truthfully. The technique is detailed in Algorithm 2.

With buffer periods defined, we summarize CLUB's update schedule in Algorithm 3 and include Figure 1 for visual representation. Let $\frac{1}{HK} \circ \pi_{\text{rand}} + (1 - \frac{1}{HK}) \circ \pi_{\tilde{k}}$ represent a mixture policy combining π_{rand} and $\pi_{\tilde{k}}$ where for each h , with probability $\frac{1}{HK}$ we act according to π_{rand} and with probability $1 - \frac{1}{HK}$ according to $\pi_{\tilde{k}}$. For convenience, we assume $\text{buffer.e}(\tilde{k})$ is an integer, as rounding up $\text{buffer.e}(\tilde{k})$ does not affect asymptotic regret. Unlike a typical low switching cost RL algorithm, Algorithm 4 further delays updating for $\frac{3 \log K}{\log(1/\gamma)}$ episodes after the switching criterion in Line 4 is satisfied.

Algorithm 3 Contextual-LSVI-UCB-Buffer (CLUB) with Known F

- 1: Initialize policy estimate π_0 , buffer period counter $\tilde{k} = 0$, buffer period starting points $\mathbf{buffer.s}(0) = 1$, and buffer period end points $\mathbf{buffer.e}(0) = 1$.
 - 2: **for** episodes $k = 1, \dots, K$ **do**
 - 3: Execute mixture policy $\frac{1}{HK} \circ \pi_{\text{rand}} + (1 - \frac{1}{HK}) \circ \pi_{\tilde{k}}$, collecting outcomes q_{ih}^τ and updating matrices $\Lambda_h^k \leftarrow \sum_{\tau=1}^k \phi(x_h^\tau, v_h^\tau) \phi(x_h^\tau, v_h^\tau)^T + I$ for all $h \in [H]$.
 - 4: If there exists $h \in [H]$ such that $(\Lambda_h^{\mathbf{buffer.e}(\tilde{k})})^{-1} \not\leq 2(\Lambda_h^k)^{-1}$, schedule a new buffer period starting at $\mathbf{buffer.s}(\tilde{k} + 1) = k$ and ending at $\mathbf{buffer.e}(\tilde{k} + 1) = k + \frac{3 \log K}{\log(1/\gamma)}$ using Algorithm 2, and set $k \leftarrow \mathbf{buffer.e}(\tilde{k} + 1)$.
 - 5: **end for**
-

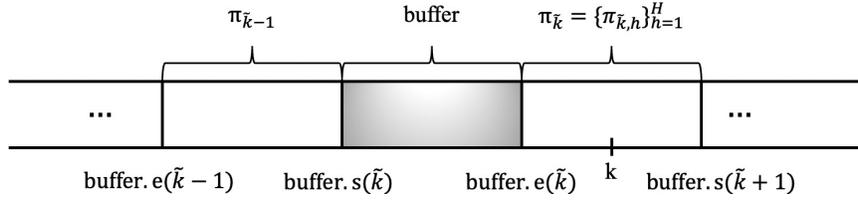


Figure 1: Learning periods and buffer periods: $\mathbf{buffer.s}(\cdot)$ and $\mathbf{buffer.e}(\cdot)$ represent the start point and the end point of a buffer, respectively. Episode k lays between $\mathbf{buffer.e}(\tilde{k})$ and $\mathbf{buffer.s}(\tilde{k} + 1)$ and the length of each buffer is $\frac{3 \log K}{\log(1/\gamma)}$.

The mixture policy sufficiently punishes untruthfulness. Combined with buffer periods (Algorithm 2) and the update schedule (Line 4), Algorithm 3 also limits the discounted utility bidders gain from untruthfulness, thereby curbing excessive overbidding and/or underbidding. Line 4 represents a kind of lazy update. We only calculate the new Q-function when at least one eigenvalue decays by half, restricting the total number of updates and benefiting the construction of high probability union regret bounds. At the same time, we wait for the length of buffer periods before updating to motivate truthful bidding. While π_{rand} is suboptimal, the mixture policy ensures that it is not executed too many times, reducing its damage to revenue.

With the techniques discussed above, namely Algorithms 1, 2, and 3, we now have sufficiently addressed our first challenge, obtaining approximately truthful reports in the face of strategic bidders. We now turn to tackling the third challenge outlined in the abstract: provably efficient reinforcement learning even when the per-step revenue is nonlinear.

Addressing Challenges 2 and 3: Regret Minimization and Nonlinear Revenue.

Having shown that our algorithm punishes untruthful behavior, we begin by showing that the resulting reports are sufficiently truthful for obtaining accurate parameter estimates. It's still quite intricate as regret depends on both the state, action related to the transition kernel, and reserve prices. Traditional point estimation with uncertainty quantity is not enough since we need to not only combine the structure of the underlying MDP and coordinate with buffer periods, the so-called lazy updates, but also consider the small proportion of untruthful bids. Whereas LSVI-UCB directly learns from empirical rewards, here we use indicators q_{ih}^k , which we recall is one if bidder i receives the item at episode k step h and

zero otherwise. As we cannot guarantee that the empirical covariance matrix is positive definite, existing techniques in [Amin et al. \(2014\)](#); [Golrezaei et al. \(2019\)](#) cannot be applied. We instead have

$$\hat{\theta}_{ih} = \arg \min_{\|\theta\| \leq 2\sqrt{d}} \sum_{\tau=1}^{\text{buffer.e}(\bar{k}+1)} (q_{ih}^\tau - 1 + F(m_{ih}^\tau - 1 - \langle \phi(x_h^\tau, v_h^\tau), \theta \rangle))^2, \quad (3)$$

where ρ_{ih}^τ is agent i 's reserve price and $m_{ih}^\tau = \max\{\max_{j \neq i} b_{ih}^\tau, \rho_{ih}^\tau\}$. Equation (3) is justified by the observation that, assuming that he bids truthfully, bidder i wins the auction with probability $1 - F(m_{ih}^\tau - 1 - \langle \phi(x_h^\tau, v_h^\tau), \theta \rangle)$, conditioned on x_h^τ, v_h^τ , and m_{ih}^τ . Controlling the uncertainty around $\hat{\theta}_{ih}$ then resembles controlling the uncertainty of a generalized linear model with $F(\cdot)$ being the link function. As bidders need to overbid or underbid significantly to alter the outcome of the auction, $\hat{\theta}_{ih}$ is less susceptible to untruthfulness.

While we use a typical linear function approximation assumption, the seller's revenue function R_h is not linear, and we cannot directly apply existing approaches. We instead directly estimate R_h and link our uncertainty on the seller's revenue to the typical linear MDP uncertainty quantifier, summarized Algorithm 4.

Algorithm 4 Estimation of $\hat{Q}_h^{\pi_{k+1}}(\cdot, \cdot)$

- 1: Estimate $\hat{\theta}_{ih}$ using Equation (3) and set $\hat{\mu}_{ih}(\cdot, \cdot) \leftarrow \langle \phi(\cdot, \cdot), \hat{\theta}_{ih} \rangle$ for all i, h .
 - 2: Estimate reserve price $\hat{\rho}_{ih}(\cdot, \cdot) = \arg \max_y y(1 - F(y - 1 - \hat{\mu}_{ih}(\cdot, \cdot)))$ for all i, h .
 - 3: Estimate revenue $\hat{R}_h(\cdot, \cdot) \leftarrow \mathbb{E}[\max\{\tilde{b}_h^-(\cdot, \cdot), \hat{\rho}_h^+(\cdot, \cdot)\} \mathbf{1}(\tilde{b}_h^+(\cdot, \cdot) \geq \hat{\rho}_h^+(\cdot, \cdot))]$.
 - 4: **for** $h = H, \dots, 1$ **do** ▷ Estimate Q-function and optimal policy.
 - 5: $\Lambda_h \leftarrow \sum_{\tau=1}^{\text{buffer.e}(\bar{k}+1)} \phi(x_h^\tau, v_h^\tau) \phi(x_h^\tau, v_h^\tau)^T + \lambda I$. ▷ We set $\lambda = 1$ in this paper.
 - 6: $\omega_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{\text{buffer.e}(\bar{k}+1)} \phi(x_h^\tau, v_h^\tau) [\max_v \hat{Q}_{h+1}(x_{h+1}^\tau, v)]$.
 - 7: $\hat{Q}_h^{\pi_{k+1}}(\cdot, \cdot) \leftarrow \min\{\omega_h^T \phi(\cdot, \cdot) + \hat{R}(\cdot, \cdot) + \text{poly}(\log K) \|\phi(\cdot, \cdot)\|_{\Lambda_h^{-1}}, 3H\}$.
 - 8: $\pi_{k+1,h}^v(\cdot) \leftarrow \arg \max_v \hat{Q}_h^{\pi_{k+1}}(\cdot, v)$.
 - 9: $\pi_{k+1,h}^{\rho^i}(\cdot) \leftarrow \hat{\rho}_{ih}(\cdot, \pi_{k+1,h}^v(\cdot))$.
 - 10: **end for**
 - 11: Return $\{\hat{Q}_h^{\pi_{k+1}}(\cdot, \cdot)\}_{h=1}^H$ and $\{\pi_{k+1,h}^v(\cdot)\}_{h=1}^H$.
-

We let \tilde{b}^+ and ρ^+ denote the highest truthful bid and the highest reserve price, respectively. Similarly, let \tilde{b}^- and ρ^- denote the second-highest. Algorithm 4 estimates the Q-function optimistically by dividing the problem into two halves: per-step revenue estimation (Lines 1 to 3) and transition estimation (Lines 4 to 10). In the first half, we use Equation (3) to estimate all θ_{ih} , which in turn gives estimates for bidders' rewards in the form of $\hat{\mu}_{ih}$. We then feed the reward function estimates to Line 2, yielding an estimate for the optimal reserve price. With Algorithms 1, 2, and 3, the effects of untruthful reports are controlled, and we can ensure that the revenue estimate is sufficiently close to the ground truth. With ρ_{ih} estimated, we then obtain revenue estimates for all states and item choices via Line 3. Consequently, we decide both nearly optimal reserve prices and the order of items, addressing the second challenge of regret minimization.

While the rest of Algorithm 4 resembles a typical LSIV-UCB algorithm (Jin et al., 2020), we highlight several key differences. First, we use the plug-in revenue estimate, whereas existing works estimate the Q-function with the empirically observed rewards. To accommodate the plug-in estimate, here ω_h estimates $\mathbb{P}_h V_{h+1}$, the transition operator applied to the V-function, as opposed to $\mathbb{B}_h V_{h+1}$, which uses the Bellman evaluation operator instead. Lastly, in Line 7 we link the uncertainty of revenue to the uncertainty bonus typically seen in linear MDPs, thereby obtaining an optimistic estimate of the Q-function induced by revenue. We conjecture the transition estimation procedure in Algorithm 4 can be changed to other suitable online RL algorithms under other function approximation assumptions.

In summary, in this section, we address the first and third challenges. The first challenge is addressed mainly by a novel technique dubbed “buffer periods” and the third one through nontrivial extensions to the LSIV-UCB framework. By combining the loss from incentivizing a truthful mechanism and learning the underlying model to set reserve prices, we get the final Algorithm 3, which explores efficiently and achieves the following regret upper bound, and then addresses the second challenge of regret minimization.

3.2 Regret Bound When $F(\cdot)$ is Known

We introduce the following assumptions before we bound the regret. These regularity assumptions are commonly found in economics literature (Kleiber and Kotz, 2003; Bagnoli and Bergstrom, 2006).

Assumption 2. *Market noise pdf f is bounded, i.e. there exist constants c_1, C_1 such that $c_1 \leq f \leq C_1$.*

Assumption 3. *Market noise pdf f is differentiable and its derivative is bounded. That is, there exists a constant L such that $|f'| \leq L$.*

Assumption 4. *Market noise cdf $F(\cdot)$ and $1 - F(\cdot)$ are log-concave.*

At a high level, Assumptions 2 and 3 ensure that the pdf f is generally well-behaved, namely, bounded and smooth. Assumption 4 is a popular assumption in economics that ensures the validity of the Myerson lemma (Myerson, 1981; Kleiber and Kotz, 2003; Bagnoli and Bergstrom, 2006). We further remark that these assumptions are mild and are satisfied by commonly used distributions such as the truncated Gaussian distribution and the uniform distribution (Golrezaei et al., 2019).

Remark 5. *We note that Assumption 4 is in fact made redundant by Assumption 2 because we have a quite “smooth” distribution with bounded differential. Then, once we have a good estimation for the parameters, “smooth” $F(\cdot)$ leads to a good estimation of the reward function. Nevertheless, we retain this assumption as it streamlines our proof by avoiding discussion of market stability with multi-optimal reserve prices and getting bogged down in tedious regret decomposition.*

We are now ready to state our results. If we set $\text{poly}(\log K) = C_7 + C_6 H \log^2 K$ in Algorithm 4, where constant C_6 is determined in Lemma 16 and constant $C_7 = B_8 H^{\frac{3}{2}} \log K$ with constant B_8 determined in Lemma 39, then we have Theorem 6.

Theorem 6. *Under Assumption 1, 2, 3 and 4, for any fixed failure probability $\delta \in (0, 1)$, with probability at least $1 - \delta$, Algorithm 3 achieves at most $\tilde{O}(\sqrt{H^5 K})$ revenue regret, where $\tilde{O}(\cdot)$ hides only absolute constants and logarithmic terms.*

Proof See Section B for a detailed proof. ■

As we discussed previously, when $H = 1$, our result cannot be compared to existing works that focus on the stochastic bandit setting due to our need to explore the action space Υ (see Broder and Rusmevichientong (2012); Drutsa (2020, 2017); Golrezaei et al. (2019) for works that achieves $\tilde{O}(1)$ revenue regret in the stochastic bandit setting). The closest work we are aware of is Cesa-Bianchi et al. (2014), which obtains a similar $\tilde{O}(\sqrt{K})$ regret in the adversarial multi-armed bandit setting, matched by our bounds.

4. Unknown Market Noise Distribution

We now discuss when the market noise distribution is unknown. Recall from previous discussions that our second challenge lies in minimizing revenue regret when the market noise distribution is unknown. Existing techniques, similar to the one in Golrezaei et al. (2019), incorporate pure exploration rounds to address the challenge, yet necessitate a $\tilde{O}(K^{2/3})$ revenue regret. In this section, we instead introduce a novel technique dubbed “simulation”, which eliminates the need for pure exploration rounds and achieves instead a $\tilde{O}(\sqrt{K})$ regret. While the first and third challenges have been previously addressed, the approaches in Section 3 also require careful adjustments, as the unknown market noise distribution makes a direct application of these approaches impossible. We detail our techniques and procedures in the rest of this section.

4.1 CLUB Algorithm When $F(\cdot)$ is Unknown

Similarly, there are three steps to do auction design when $F(\cdot)$ is unknown. First, we leverage Algorithm 1 and Algorithm 5 to motivate an approximately truthful mechanism. Second, we utilize Algorithm 8 in coordination with the newly proposed Algorithm 7 to estimate the underlying MDP and set reserve prices. We motivate truthfulness through buffer periods and quantify the uncertainty by constructing corresponding ellipsoid bounds. Finally, we add up all these uncertainties and minimize regret with high probability.

Addressing Challenge 1: Untruthfulness. When the market noise distribution is unknown, the techniques used in Section 3 cannot be applied directly, necessitating careful adaptations. We summarize the changes to these techniques, beginning by introducing Algorithm 5, the counterpart to Algorithm 2, for when $F(\cdot)$ is unknown. The key difference lies in the optimization subroutine called in Line 3, which is required for addressing the third challenge when the market noise distribution $F(\cdot)$ is unknown.

We then discuss Algorithm 6, a close variant of Algorithm 3, whose biggest change lies in the update schedule in Line 4. Algorithm 3 maintains only an accurate estimate of the underlying MDP, achieved with a low switching cost style update schedule, which in turn deters untruthful bidding. On the other hand, Algorithm 6 needs accurate estimates of both the MDP and the market noise distribution $F(\cdot)$. We force additional updates whenever k

Algorithm 5 Buffer Period with Unknown $F(\cdot)$

- 1: Receives buffer start $\mathbf{buffer.s}(\tilde{k} + 1) = k$ and end $\mathbf{buffer.e}(\tilde{k} + 1) = k + \frac{3 \log K}{\log(1/\gamma)}$.
 - 2: Do nothing for all episodes $\mathbf{buffer.s}(\tilde{k} + 1) \leq k < \mathbf{buffer.e}(\tilde{k} + 1)$, i.e., do nothing during the buffer period before the end.
 - 3: At the end of the buffer period, update policy estimate $\pi_{\tilde{k}+1}$ and Q-function estimate $\hat{Q}_h^{\pi_{\tilde{k}+1}}(\cdot, \cdot)$ using Algorithm 8, and then increment buffer period counter $\tilde{k} \leftarrow \tilde{k} + 1$.
-

is a power of 2, also ensuring that $\hat{F}(\cdot)$ is close to $F(\cdot)$. As the number of updates remains in $\mathcal{O}(\log K)$, the extraneous updates do not affect the regret asymptotically.

Algorithm 6 Contextual-LSVI-UCB-Buffer (CLUB) with Unknown F

- 1: Initialize policy estimate π_0 , buffer period counter $\tilde{k} = 0$, buffer period starting points $\mathbf{buffer.s}(0) = 1$, and buffer period end points $\mathbf{buffer.e}(0) = 1$.
 - 2: **for** episodes $k = 1, \dots, K$ **do**
 - 3: Execute mixture policy $\frac{1}{HK} \circ \pi_{\text{rand}} + (1 - \frac{1}{HK}) \circ \pi_{\tilde{k}}$, collecting outcomes q_{ih}^τ and updating matrices $\Lambda_h^k \leftarrow \sum_{\tau=1}^k \phi(x_h^\tau, v_h^\tau) \phi(x_h^\tau, v_h^\tau)^T + I$ for all $h \in [H]$.
 - 4: If there exists $h \in [H]$ such that $(\Lambda_h^{\mathbf{buffer.e}(\tilde{k})})^{-1} \not\leq 2(\Lambda_h^k)^{-1}$ or $\log_2(k)$ is an integer, schedule a new buffer period starting at $\mathbf{buffer.s}(\tilde{k} + 1) = k$ and ending at $\mathbf{buffer.e}(\tilde{k} + 1) = k + \frac{3 \log K}{\log(1/\gamma)}$ using Algorithm 5, and set $k \leftarrow \mathbf{buffer.e}(\tilde{k} + 1)$.
 - 5: **end for**
-

Similar to Section 3, these techniques, namely the buffer periods and the update schedule, ensure that the impatient bidders are sufficiently truthful. However, for estimating θ_{ih} , as we do not know $F(\cdot)$, the optimization problem in Equation (3) no longer applies. Fortunately, we know that whenever π_{rand} is executed, assuming the bidders are truthful, $\Pr(q_i^\tau = 1) = \frac{1}{3N}(2 - \langle \phi(x_h^\tau, v_h^\tau), \theta \rangle)$ conditioned on x_h^τ, v_h^τ , as the bidder i and the reserve price ρ_{ih}^τ are drawn uniformly at random. Leveraging this observation, we quickly realize that we can simply use the outcomes from when π_{rand} is executed to estimate the bidders' rewards, even when $F(\cdot)$ is unknown. Unfortunately, using the observation naively introduces the second challenge: minimizing revenue regret when $F(\cdot)$ is unknown.

Addressing Challenge 2: Regret Minimization. An intuitive way to incorporate the previous observation is to simply perform pure exploration rounds with π_{rand} , similar to the technique in Golrezaei et al. (2019). However, doing so incurs $\tilde{\mathcal{O}}(K^{2/3})$ revenue regret, as π_{rand} does not set the reserve prices optimally and we are not exploring and exploiting simultaneously. To balance exploration and exploitation, we propose a new technique that we dub “simulation”, which allows us to continue exploiting with the mixture policy.

Here we introduce a new random variable $\tilde{q}_{ih}^\tau = \mathbb{1}(b_{ih}^\tau \geq \tilde{\rho}_{ih}^\tau)$, where for each h, τ we select one $i \in [N]$ uniformly at random and then draw $\tilde{\rho}_{ih}^\tau$ from $\text{Unif}([0, 3])$. For all $j \neq i$ we set $\tilde{\rho}_{jh}^\tau$ to ∞ . At a high level, \tilde{q}_{ih}^τ “simulates” executing π_{rand} : holding x_h^τ and v_h^τ constant, what would be the outcome if we were to act according to π_{rand} instead? As we do not need to execute π_{rand} , revenue regret can be decreased. Furthermore, \tilde{q}_{ih}^τ still enjoys the same

Algorithm 7 Simulation

- 1: **for** $h = 1, \dots, H$ and $\tau = 1, \dots, K$ **do**
 - 2: Generate virtual reserve prices $\tilde{\rho}_{ih}^\tau$ by selecting one bidder $i \in [N]$ uniformly at random. Let $\tilde{\rho}_{ih}^\tau \sim \text{Unif}([0, 3])$ and set all other reserve prices to infinity, i.e. $\tilde{\rho}_{jh}^\tau = \infty$ for all $j \neq i$.
 - 3: Use real bidding data b_{ih}^τ and simulated reserve prices $\tilde{\rho}_{ih}^\tau$ to simulate outcome \tilde{q}_{ih}^τ for all $i \in [N]$, namely set $b_{ih}^\tau = \mathbb{1}(b_{ih}^\tau \geq \tilde{\rho}_{ih}^\tau)$ for all $i \in [N]$.
 - 4: **end for**
 - 5: Return the simulated outcomes $\{\tilde{q}_{ih}^k\}$.
-

resilience towards untruthful reporting that q_{ih}^τ does. Indeed, when the bidder overbid or underbid by a small amount, the number of times \tilde{q}_{ih}^τ changes could be controlled effectively.

More technically, Algorithm 7 is critical for two reasons. First, the difference between $\hat{F}(\cdot)$ and $F(\cdot)$ decays at a rate of $O(1/\sqrt{K})$. If we simply use Equation (3), only replacing $F(\cdot)$ with $\hat{F}(\cdot)$, the estimation error is roughly on the order of $\tilde{O}(\sqrt{\text{buffer.e}(\tilde{k} + 1)})$ which precludes achieving $\tilde{O}(\sqrt{K})$ regret. Second, replacing \tilde{q}_{ih}^τ with q_{ih}^τ does not work, as we need to de-bias the estimator when we switch from $F(\cdot)$ to the uniform distribution induced by π_{rand} . Even when the bidders report truthfully, we cannot guarantee that $\Pr(q_{ih}^\tau = 1 \mid x_h^\tau, v_h^\tau)$ could be related to $\frac{1}{3N}(1 + \langle \phi(x_h^\tau, v_h^\tau), \theta_{ih} \rangle)$. Consequently, it would be hard to ensure that when all bidders are truthful, the estimator $\hat{\theta}_{ih}^\tau$ would converge to θ_{ih} .

Addressing Challenge 3: Nonlinear Revenue. With the first challenge addressed by carefully adjusting techniques in Section 3 and the second by the simulation technique detailed in Algorithm 7, we now discuss the third challenge: provably efficient reinforcement learning when the revenue is nonlinear and $F(\cdot)$ is unknown. We start with summarizing how we simultaneously estimate θ_{ih} and $F(\cdot)$ in the form of Equation (4).

$$\begin{aligned} \hat{\theta}_{ih} &= \arg \min_{\|\theta\| \leq 2\sqrt{d}} \sum_{\tau=1}^{\text{buffer.e}(\tilde{k}+1)} (3N\tilde{q}_{ih}^\tau - (1 + \langle \phi(x_h^\tau, v_h^\tau), \theta \rangle))^2, \\ \hat{F}(z) &= \frac{1}{N\text{buffer.e}(\tilde{k} + 1)H} \sum_{i=1}^N \sum_{\tau=1}^{\text{buffer.e}(\tilde{k}+1)} \sum_{h=1}^H \mathbb{1}(b_{i\tau h} - 1 - \langle \phi_h^\tau, \hat{\theta}_{ih} \rangle \leq z). \end{aligned} \tag{4}$$

We note that we are simply using a histogram to estimate $F(\cdot)$ and, as we have successfully decoupled the estimation error of $F(\cdot)$ from that of θ_{ih} , using histogram is sufficient for achieving $\tilde{O}(\sqrt{K})$ revenue regret. We then introduce Algorithm 8, whose key difference with Algorithm 4 lies in the added uncertainty due to $\hat{F}(\cdot)$ and the inclusion of the simulation subroutine. Similar to Section 3, the procedure then provides us with sufficiently accurate policy and Q-function estimates, resolving our third and final challenge.

In summary, we have addressed all three challenges when the market noise distribution is unknown. The first challenge is resolved by carefully adjusting the techniques introduced in Section 3, ensuring that they are still valid when $F(\cdot)$ is unknown. For the second challenge we feature a novel technique dubbed ‘simulation’ that allows us to ‘simulate’ pure exploration rounds without actually executing them, reducing revenue regret. For the

Algorithm 8 Estimation of $\hat{Q}_h^{\pi_{\tilde{k}+1}}(\cdot, \cdot)$ with Unknown $F(\cdot)$

- 1: Collect simulation outcome \tilde{q} using Algorithm 7.
 - 2: Estimate $\hat{\theta}_{ih}, \hat{F}(\cdot)$ using Equation (4).
 - 3: Estimate $\hat{\mu}_{ih}(\cdot, \cdot) \leftarrow \langle \phi(\cdot, \cdot), \hat{\theta}_{ih} \rangle$.
 - 4: Set reserve price $\hat{\rho}_{ih}(\cdot, \cdot) = \arg \max_y y(1 - \hat{F}(y - 1 - \hat{\mu}(\cdot, \cdot)))$.
 - 5: Estimate revenue $\hat{R}_h(\cdot, \cdot) \leftarrow \mathbb{E}[\max\{\tilde{b}_h^-(\cdot, \cdot), \hat{\rho}_h^+(\cdot, \cdot)\} \mathbf{1}(\tilde{b}_h^+(\cdot, \cdot) \geq \hat{\rho}_h^+(\cdot, \cdot))]$.
 - 6: **for** $h = H, \dots, 1$ **do** ▷ Estimate Q -function and optimal policy.
 - 7: $\Lambda_h \leftarrow \sum_{\tau=1}^{\text{buffer.e}(\tilde{k}+1)} \phi(x_h^\tau, v_h^\tau) \phi(x_h^\tau, v_h^\tau)^T + \lambda I$. ▷ We set $\lambda = 1$ in this paper.
 - 8: $\omega_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{\text{buffer.e}(\tilde{k}+1)} \phi(x_h^\tau, v_h^\tau) [\max_a \hat{Q}_{h+1}(x_{h+1}^\tau, a)]$.
 - 9: $\hat{Q}_h^{\pi_{\tilde{k}+1}}(\cdot, \cdot) \leftarrow \min\{\omega_h^T \phi(\cdot, \cdot) + \hat{R}(\cdot, \cdot) + \text{poly}_1(\log K) \|\phi(\cdot, \cdot)\|_{\Lambda_h^{-1}} + \frac{\text{poly}_2(\log K)}{\sqrt{\text{buffer.e}(\tilde{k}+1)}}, 3H\}$
 - 10: $\pi_{\tilde{k}+1,h}^v(\cdot) \leftarrow \arg \max_v \hat{Q}_h^{\pi_{\tilde{k}+1}}(\cdot, v)$.
 - 11: $\pi_{\tilde{k}+1,h}^{\rho_i}(\cdot) \leftarrow \hat{\rho}_{ih}(\cdot, \pi_{\tilde{k}+1,h}^a(\cdot))$.
 - 12: **end for**
 - 13: **Return** $\{\hat{Q}_h^{\pi_{\tilde{k}+1}}(\cdot, \cdot)\}_{h=1}^H$ and $\{\pi_{\tilde{k}+1,h}(\cdot)\}_{h=1}^H$.
-

third challenge, we build off of the simulation technique and introduce a new estimation procedure for jointly estimating $F(\cdot)$ and θ .

4.2 Regret Bound of CLUB Algorithm When $F(\cdot)$ is Unknown

We now argue that Algorithm 6 achieves $\tilde{O}(\sqrt{K})$ regret. We begin with a slight detour, making a basic assumption on the hypothesis class for $F(\cdot)$.

Assumption 7. *The market noise distribution $F(\cdot)$ belongs to a distribution family \mathcal{F} .*

We further let $\mathcal{N}_\epsilon(\mathcal{F})$ be the ϵ -covering number of \mathcal{F} with respect to the metric that $\text{dist}(F, G) = \sup_x |F(x) - G(x)|$. We now have our main theorem when the noise distribution is unknown. If we let $\text{poly}_1(\log K) = C_{15} + C_{13}H \log^2 K$ and $\text{poly}_2(\log K) = C_{14}H^2 \log^4 K$ in Algorithm 8, where $C_{15} = D_7 H^{\frac{3}{2}}$ and the constant D_7 is determined in Lemma 47, constants C_{13} and C_{14} are determined in Lemma 24, we would attain the following regret guarantee.

Theorem 8. *Under Assumptions 1, 2, 3, 4 and 7, when $F(\cdot)$ is unknown, for any fixed failure probability $\delta \in (0, 1)$, Algorithm 6 achieves at most $\tilde{O}(H^3 \sqrt{K} + H^{2.5} \sqrt{K \log \mathcal{N}_{1/K}(\mathcal{F})})$ regret with probability at least $1 - \delta$ in the worst case, where $\tilde{O}(\cdot)$ hides only absolute constants and logarithmic terms.*

Proof See Section C for a detailed proof. ■

We highlight that when $\mathcal{N}_{1/K}(\mathcal{F})$ is polynomial in K , an implicit assumption found in Kong et al. (2021); Foster et al. (2021); Jin et al. (2021a), Theorem 8 shows that Algorithm 6 achieves $\tilde{O}(\sqrt{K})$ regret, improving over revenue regret guarantees found in Amin et al. (2014); Golrezaei et al. (2019) with only mild additional assumptions on the nonparametric

hypothesis class \mathcal{F} . Our result is able to beat the well-known $\Omega(K^{2/3})$ revenue lower bound in Kleinberg and Leighton (2003) with the help of Assumptions 2 and 3 for similar but not totally the same scenarios to be fair. Nevertheless, as argued previously, these assumptions are satisfied by widely-used parametric distribution families such as the uniform distribution and the truncated normal distribution (Golrezaei et al., 2019), hence our result still remains broadly applicable. The way Kleinberg and Leighton (2003) constructs regret lower bound is to find a special case containing no information. As they say “the expected revenue per buyer is a constant independent of the offer price outside the interval of good prices”, it provides nothing useful for learning. However, with Assumption 2, it guarantees the information in each exploration and partial out this extreme situation.

Finally, we highlight that both bounds in Sections 3 and 4 match corresponding lower bounds with respect to K . From the $\Omega(\sqrt{K})$ lower bound in Jin et al. (2020), we directly know that results in Theorems 6 and 8 match corresponding regret lower bounds, as the problem in Jin et al. (2020) is a subproblem of our problem.

5. Proof Sketch

Before sketching out the proof techniques, we take a slight detour and discuss how revenue regret could be decomposed. Recall that $\pi_{\tilde{k}}$ denotes the optimistic policy estimate maintained from episode $\mathbf{buffer.e}(\tilde{k}) + 1$ to $\mathbf{buffer.s}(\tilde{k} + 1)$, namely the estimate from the end of the \tilde{k} -th buffer period to the start of the $(\tilde{k} + 1)$ -th. We also recall that π^* is the optimal item choice and pricing policy when the seller knows the bidders’ reward functions, the transition kernel \mathbb{P} and the market noise distribution $F(\cdot)$ beforehand, and we use V^{π^*} to denote the revenue’s V-function for the optimal policy π^* .

We now introduce several new notations that will be used in the rest of the section. We use π_k to denote the policy executed at episode k . Intuitively, the policy π_k consists of some steps in which the corresponding $\pi_{\tilde{k}}$ is executed and some steps where π_{rand} is executed. Let $\mathbb{1}(k \in \mathbf{buffer})$ indicate the event that there exists some integer \tilde{k} such that $k \in [\mathbf{buffer.s}(\tilde{k}), \mathbf{buffer.e}(\tilde{k})]$, i.e. the episode k is within a buffer period. To better highlight the effect of untruthfulness, we let \tilde{V} denote the optimistic V-function estimate if all bidders were to report truthfully.

5.1 Regret Decomposition

The regret can be decomposed into the following five parts.

1. $\Delta_1 = \sum_{k=1}^K [V_1^{\pi^*}(x_1) - \tilde{V}_1^{\pi_k}(x_1)] \mathbb{1}(\pi_k \neq \pi_{\tilde{k}} \text{ and } k \notin \mathbf{buffer})$. The term Δ_1 is due to the seller not knowing the bidders’ reward functions and the underlying transition dynamics of the MDP. The term is nonzero even if we were to assume that all bidders report truthfully due to the uncertainty of the environment.
2. $\Delta_2 = \sum_{k=1}^K [V_1^{\pi^*}(x_1) - V_1^{\pi_k}(x_1)] \mathbb{1}(k \in \mathbf{buffer})$. The second term comes from the buffer periods, which cause suboptimality as we intentionally delay the policy update in order to further punish untruthfulness. While conducive to more truthful reports, a delayed update schedule induces regret as the policy estimate is stale during these buffer periods.

3. $\Delta_3 = \sum_{k=1}^K [V_1^{\pi^*}(x_1) - V_1^{\pi^k}(x_1)] \mathbf{1}(\text{exists } h \text{ such that } \pi_{k,h} = \pi_{\text{rand}} \text{ and } k \notin \text{buffer})$. The third term Δ_3 is caused by π_{rand} , as it sets reserve prices and chooses items entirely randomly.
4. $\Delta_4 = \sum_{k=1}^K [V_1^{\pi^*}(x_1) - V_1^{\pi^k}(x_1)] \mathbf{1}(k \in \text{L and } k \notin \text{buffer})$. We only provide intuition behind the term L and defer its precise mathematical definition to Equation (5) for when $F(\cdot)$ is known and Equation (6) for when $F(\cdot)$ is not. The term L is a collection of episode indices where the bidders' untruthful bids alter the outcome of the multi-phase auction, through either q_{ih} or \tilde{q}_{ih} . At a high level, while we could measure the revenue suboptimality of the selected reserve prices if the bidders are truthful, the seller's revenue could be harmed arbitrarily by bidders who underbid/overbid so much that the auction's outcome itself is altered. The term Δ_4 then measures the effect of the changed outcomes due to untruthful bidding.
5. $\Delta_5 = \sum_{k=1}^K [\tilde{V}_1^{\pi^k}(x_1) - V_1^{\pi^k}(x_1)] \mathbf{1}(\pi_k = \pi_{\tilde{k}} \text{ and } k \notin \text{buffer})$. Compared to Δ_4 , which measures the effect of changed outcomes due to untruthfulness, Δ_5 measures the effect of changed bids due to untruthfulness. Intuitively, a bidder who overbids/underbids a small amount would not affect the auction's outcome, but could change the amount the seller charges slightly. We measure the effect with Δ_5 .

With easy algebra calculation, we have the following proposition.

Proposition 9. *With Δ_1 to Δ_5 defined as above, it holds that $\text{Regret} \leq \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5$.*

Proof Since our benchmark is the maximum revenue when everything is common knowledge, it holds that $V_1^{\pi^*}(x_1) \geq V_1^{\pi^k}(x_1)$ at any time. It is because that V^{π^*} is no less than the revenue achieved when existing hidden information with any policy due to its optimality.

Since $\Delta_1 + \Delta_5 = \sum_{k=1}^K [V_1^{\pi^*}(x_1) - V_1^{\pi^k}(x_1)] \mathbf{1}(\pi_k = \pi_{\tilde{k}} \text{ and } k \notin \text{buffer}) + \mathbf{1}(k \in \text{buffer}) + \mathbf{1}(\pi_k = \pi_{\tilde{k}} \text{ and } k \notin \text{buffer}) + \mathbf{1}(\text{exists } h \text{ such that } \pi_{k,h} = \pi_{\text{rand}} \text{ and } k \notin \text{buffer}) + \mathbf{1}(k \in \text{L and } k \notin \text{buffer}) \geq 1$, it holds that

$$\text{Regret} = \sum_{k=1}^K V_1^{\pi^*}(x_1) - V_1^{\pi^k}(x_1) \leq \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5,$$

which ends the proof. ■

5.2 Proof Techniques

With the sources of revenue regret sketched out, we summarize the high-level intuition behind our proof, which mainly comprises the following steps.

Step 1: Limit the magnitude of untruthful reporting. As we discussed in Section 3, reducing the frequency at which we update the policies and including the buffer periods forces bidders to wait before they can gain from untruthful reporting. When the bidders are impatient, the amount they can gain from untruthful reports is then upper-bounded. With the help of π_{rand} , we are also always punishing the bidders for untruthful reports.

Combining the two halves, we can control the total amount by which bidders overbid or underbid, as overbidding or underbidding too much would decrease their utilities. Moreover, by directly controlling the “amount” of overbidding and underbidding, we are able to upper bound Δ_5 , the part of the revenue regret due to untruthfulness. The step corresponds to Lemma 11 in Section B.

Step 2: Control the number of times q_{ih}^k change due to untruthfulness. Since we are using q_{ih}^k , as opposed to b_{ih}^k , to learn the bidder’s reward functions, to ensure the estimates’ accuracy, we only need to show that the q_{ih}^k ’s are close to their values when the bidders are truthful. As q_{ih}^k ’s are outcomes of an auction, bidders need to overbid or underbid by a significant amount in order to alter q_{ih}^k . Combined with the previous step, we can limit the number of times q_{ih}^k is altered due to untruthful behavior. With the number of changes controlled, we can also control Δ_4 . The step corresponds to Lemma 12 and Lemma 20.

Step 3: Prove the estimates of personal parameters and noise distribution are good. Having shown that the bidders provide us with sufficiently truthful reports, we connect our work to RL with *generalized* linear function approximation in Wang et al. (2020b) to show $\hat{\theta}_{ih}$ is sufficiently accurate and apply the Dvoretzky–Kiefer–Wolfowitz inequality to show $\hat{F}(\cdot)$ is sufficiently accurate (Dvoretzky et al., 1956). When the market noise distribution is known, the step corresponds to Lemma 14 and Lemma 15. When the market noise distribution is unknown, the step corresponds to Lemma 21, Lemma 22 and Lemma 23.

Step 4: Prove $\hat{R}(\cdot, \cdot) \approx R(\cdot, \cdot)$ and extend LSVI-UCB to non-linear reward function. By applying Taylor expansion to the revenue function R_h , we relate the accuracy of $\hat{R}(\cdot, \cdot)$ to the accuracy of $\hat{\theta}_{ih}$, which is shown to be accurate in the previous step. We can then show our policy estimate $\pi_{\bar{k}}$ is approximately optimal with standard LSVI-UCB analysis. Steps 3 and 4 then combine to control Δ_1 . The step corresponds to Lemma 16, Lemma 17, Lemma 24 and Lemma 25.

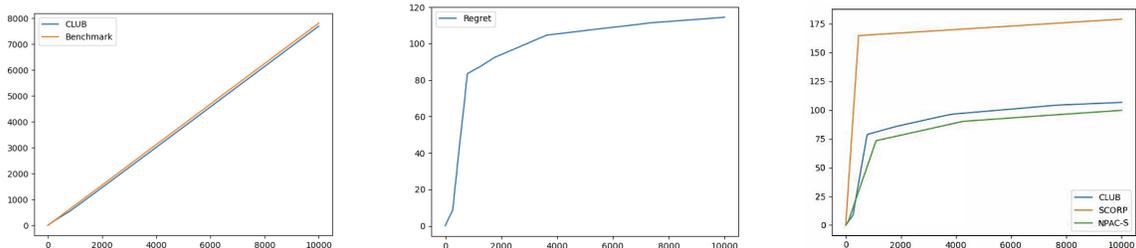
Step 5: Limit the effects of π_{rand} and buffer periods. We finally control the revenue regret due to π_{rand} and buffer periods. A key observation is that the number of times in which π_{rand} is executed and the length of the buffer periods are all in $\tilde{O}(1)$ and hence do not harm our regret asymptotically. Consequently, terms Δ_2 and Δ_3 are controlled effectively. For π_{rand} , the step corresponds to Lemma 13, and as for buffer periods, it corresponds to Lemma 10 and Lemma 19.

6. Numerical Experiments

Here, we present numerical simulations to compare the performance of Algorithm 6 with several baseline policies in different settings¹. To be specific, we compare the performances of CLUB (i.e. Algorithm 6), SCORP (Golrezaei et al., 2019) and NPAC-S (Golrezaei et al., 2023) in contextual bandit settings (i.e. $H = 1$) and the performances of CLUB and NPAC-S in MDP settings. In all experiments, we assume that the noise distribution $F(\cdot)$

1. Code is available at https://github.com/Air-8/SPA_CLUB.

is unknown. The numerical experiment written in Python 3.10.9 runs on a laptop with an Apple M2 CPU. All three algorithms use less than 30 seconds to calculate 10000 episodes, which shows their practicality in reality. We delay more details and further robustness experiments in Section F.



(a) The performance of CLUB against the benchmark.

(b) The regret accumulation of CLUB.

(c) The average performances of three algorithms.

Figure 2: Experiment results for the contextual bandit setting: Figure 2a compares the revenue achieved by CLUB and benchmark (the maximum revenue when everything is common knowledge), showing CLUB obtains more than 98% revenue. Figure 2b shows the sublinear regret associated with our CLUB algorithm as the curve trend is below linear. Figure 2c exhibits that CLUB is comparable with NPAC-S, overwhelming SCORP.

In contextual bandit setting, we set $K = 10000$, $\gamma = 0.9$ for each setting and repeat the procedure for $n = 30$ trails for each algorithm. We show results in Figure 2. Figures 2a and 2b show results in one trial, where we find that CLUB can obtain more than 98% revenue compared with the benchmark, in which the underlying model is common knowledge. At the same time, Figure 2b testifies $\tilde{O}(\sqrt{K})$ -shaped regret. In Figure 2c, we show the average regrets among all 30 trials of these three different algorithms. The average regrets in 30 trials are 106.62, 178.96 and 99.69 respectively. Because the scale of regret depends on the specific instances, and most of the variance comes from randomness in the instances themselves, we report the regret of each trial and the number of winning times in Table 1, which better reflects the performance of the algorithms. As for the number of winning times, CLUB wins 15 times while NPAC-S wins 14 times. SCORP only wins once. Therefore, we conclude that the performances of CLUB and NPAC-S are comparable, overwhelming the performance of SCORP. Since SCORP doesn't work well even in contextual bandit settings, we only compare CLUB and NPAC-S under MDP.

In the MDP setting, we incorporate $K = 10000$, $H = 2$, $\gamma = 0.9$ and also conduct $n = 30$ trails for both two algorithms. We show the corresponding results in Figure 3. Our CLUB can obtain more than 98% revenue (c.f. Figure 3a) against the benchmark which highlights its great performance. In Figure 3b, it's clear to see the $\tilde{O}(\sqrt{K})$ -shaped regret. Among all 30 trails, CLUB wins all 30 times. As for average regrets, it's 203.07 for CLUB and 756.31 for NPAC-S. Therefore, we can conclude that under the MDP setting, CLUB sufficiently works better than NPAC-S. Together with experiments under contextual bandits, our experiments are in favor of CLUB algorithms, which show the importance of our newly proposed techniques.

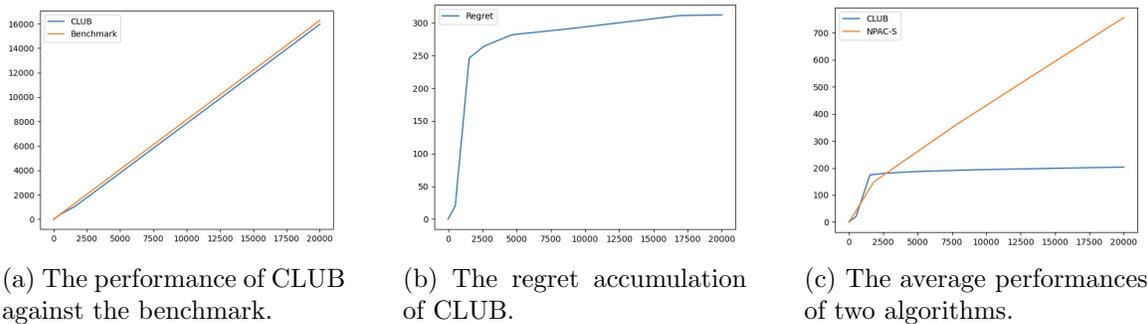


Figure 3: Experiment results for the MDP setting: Figure 3a compares the revenue achieved by CLUB and benchmark (the maximum revenue when everything is common knowledge), showing CLUB obtains more than 98% revenue. Figure 3b shows the sublinear regret associated with our CLUB algorithm, as the curve trend is below linear. Figure 3c exhibits that compared with NPAC-S, CLUB has less regret, testifying to its optimality.

7. Conclusion and Discussion

In this paper, we propose a multi-phase second-price auction mechanism based on reinforcement learning. We highlight that when market noise distribution is unknown, our algorithm achieves $\tilde{O}(H^3\sqrt{K})$ regret, improving upon the $\tilde{O}(K^{2/3})$ guarantees in [Amin et al. \(2014\)](#); [Golrezaei et al. \(2019\)](#), using a new method to deal with unknown distribution. Our work is also the first to introduce the notion of “buffer periods”, a concept crucial to bringing existing techniques in the bandit setting to the more general MDP setting.

Questions raise themselves for future explorations. Is it possible to further generalize our results to RL with general function approximation under bounded Bellman Eluder dimensions ([Jin et al., 2021a](#))? Can we optimize the dependence on horizon H and feature dimension d ? Is it possible to consider a continuous representation of buyers and form a mean-field game? We leave these interesting questions as potential next steps.

Acknowledgments

The authors thank the anonymous Reviewers and the AE for their very constructive comments. Part of the work was done when Rui Ai was an undergraduate at Peking University, and he was partially supported by the elite undergraduate training program of School of Mathematical Sciences at Peking University. Zhaoran Wang acknowledges National Science Foundation (Awards 2235451, 2225087, 2211210, CAREER-2048075, 2015568, 2008827, 1934931/2216970), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, Two Sigma, and Google for their support. Zhuoran Yang acknowledges Simons Institute (Theory of Reinforcement Learning).

Appendix A. Detailed Comparison with Golrezaei et al. (2019)

There are three different models and corresponding algorithms named CORP, CORP-II, and SCORP, respectively, in Golrezaei et al. (2019). We compare them with our model one by one.

CORP considers a contextual bandit setting with known noise distribution achieving $\tilde{O}(1)$ regret. However, as we mentioned before, accommodating underlying MDP, $\Omega(\sqrt{K})$ regret lower bound is inevitable. In Section 3, we propose our optimal CLUB algorithm matching the lower bound.

CORP-II considers an unknown but parametric noise distribution and achieves $\tilde{O}(\sqrt{K})$ regret. However, in Section 4, we consider an unknown and non-parametric noise distribution. Therefore, compared with the setting for CORP-II, our model is strictly much harder for the following two-fold reasons. We need to consider the extra MDP structure and non-parametric noise distribution. Moreover, since CORP-II doesn't have enough horizons to explore, it doesn't work well under our MDP setting and cannot achieve its original $\tilde{O}(\sqrt{K})$ regret.

SCORP considers time-varying and non-parametric noise distribution achieving $\tilde{O}(K^{2/3})$ regret. We share the similarity that both of these settings can't be parameterized, which means that we lose the opportunity to utilize some concentration inequalities directly, and we need to bypass these obstacles to achieve sublinear regrets. There are two main differences between our model and SCORP. First, the underlying MDP makes our problem harder than that of SCORP. Second, we consider a fixed noise distribution and use a different benchmark, making these two models not directly comparable. As a result, our algorithm achieves $\tilde{O}(\sqrt{K})$ regret with mild additional assumptions on the shape of $F(\cdot)$ (c.f. Assumption 2 and 3). Although it is hard to compare the difficulties between our setting and SCORP in strict order, we believe they have a similar degree of difficulty. As we mentioned in Section 1.1, the work (Amin et al., 2014) explores a scenario with a non-parametric yet fixed distribution setting, experiencing a regret of $\tilde{O}(K^{2/3})$. This observation suggests that the primary challenge might arise from the non-parametric nature of the problem, as opposed to the time-varying setting. Moreover, we should highlight that an $\tilde{O}(K^{2/3})$ regret is inevitable even though the distribution is fixed corresponding to a saddle point for SCORP, as they spend "too many" episodes to explore while we don't "waste" time to do pure exploration so that balance the exploration-exploitation tradeoff better and achieve better regret bounds. Objectively, our method will suffer $\Omega(K^{2/3})$ regret lower bound for a time-varying model, and it's of independent interest for future research.

Appendix B. Omitted Proof in Section 3

In this section, we show some useful lemmas in order to prove theorems in Section 3. We organize the section as follows. Firstly, we introduce lemmas to bound the effect of untruthful bidding. Then, we will show that we are able to estimate unknown parameters accurately. Finally, combining them leads to bounded regret with high probability.

B.1 Useful Lemmas for Proving Theorem 6

Now, we begin to prove our conclusions. First of all, we show the following lemma to bound the number of buffers. Buffer episodes represent those episodes in buffer periods.

Lemma 10. *Under Assumption 1 about linear MDP, it holds that the number of buffer episodes is not larger than $\frac{3HC_2 \log^2 K}{\log \frac{1}{\gamma}}$. Then, the number of corresponding steps is not larger than $\frac{3H^2 C_2 \log^2 K}{\log \frac{1}{\gamma}}$, where C_2 is a constant that only depends on d and λ .*

Because of the existence of a buffer, the bidder will not overbid or underbid a lot in the other episodes. Then, we have the following lemma.

Lemma 11. *Apart from the buffer periods, a rational bidder won't overbid or underbid for more than $\frac{3H\sqrt{2N}}{K\sqrt{1-\gamma}}$, denoted by $\frac{C_3 H}{K}$.*

Then we define L as the number of steps the bidder doesn't bid his true value and change the outcome of the auction. Then, it holds the following lemma with the help of Lemma 11. We formalize the definition of L for any given i, h as follows.

$$L = \{k : \mathbb{1}(v_{ih}^k w > \max\{b_{-ih}^{k+}, \rho_{ih}^k\}) \neq \mathbb{1}(b_{ih}^k > \max\{b_{-ih}^{k+}, \rho_{ih}^k\})\}. \quad (5)$$

Lemma 12. *With probability at least $1 - \delta$, it holds that for any given i, h*

$$L \leq \frac{3HC_2 \log^2 K}{\log \frac{1}{\gamma}} + 4C_1 C_3 H + 8 \log\left(\frac{2NH}{\delta}\right) \leq C_4 H \log^2 K,$$

where C_4 is a constant independent of K and H .

Now, we bound the number of steps we use π_{rand} instead of $\pi_{\tilde{k}}$. Especially, we regard π_{rand} as the policy used in the situation that happens with probability $\frac{1}{KH}$.

Lemma 13. *With probability at least $1 - \delta$, the number of steps using π_{rand} is smaller than $\max\{4, 1 + \frac{4}{3} \log \frac{1}{\delta}\}$.*

Now, we will show the wedge between $\hat{\mu}_{ih}(\cdot, \cdot)$ and $\mu_{ih}(\cdot, \cdot)$ for any bidder i and step h . It holds the following lemma.

Lemma 14. *We use θ_{ih}^* to denote the true parameter and $\hat{\theta}_{ih}$ to represent the outcome from Equation (3) in episode $\text{buffer.e}(\tilde{k})$. Therefore, under Assumption 2 and Assumption 3, for any i and h , it holds the following union bound that C_5 is a constant and*

$$\sqrt{(\hat{\theta}_{ih} - \theta_{ih}^*)^T \Lambda^{\text{buffer.e}(\tilde{k})} (\hat{\theta}_{ih} - \theta_{ih}^*)} \leq C_5 \sqrt{H} \log K,$$

with probability at least $1 - \delta$, conditional on Good Event \mathcal{E} .

Then, we are ready to have the bound for $\hat{\mu}$. It holds the following lemma:

Lemma 15. *Conditional on Good Event \mathcal{E} , it holds that*

$$|\hat{\mu}_{ih}^k(\cdot, \cdot) - \mu_{ih}^k(\cdot, \cdot)| \leq C_5 \sqrt{H} \log K \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}},$$

where $\text{buffer.e}(\tilde{k})$ is the last episode using Equation (3) before episode k , similarly hereinafter.

Now, we focus on the gap between $R(\cdot, \cdot)$ and the estimate $\hat{R}(\cdot, \cdot)$. We are ready to show the following lemma.

First of all, we introduce some notations. $R_h^k(\cdot, \cdot) = \sum_{i=1}^N \mathbb{E}[\max\{r_{ih}^{k-}, \alpha_{ih}^{k*}\} \mathbb{1}(r_{ih}^k \geq \max\{r_{ih}^{k-}, \alpha_{ih}^{k*}\})]$ and $\hat{R}_h^k(\cdot, \cdot) = \sum_{i=1}^N \mathbb{E}[\max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\} \mathbb{1}(\hat{r}_{ih}^k \geq \max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\})]$. In short, $R(\cdot, \cdot)$ is the expectation of revenue if we choose the optimal reserve price α_{ih}^{k*} for every bidder based on the knowledge of $\mu_{ih}^k(\cdot, \cdot)$ and everyone bids truthfully based on his valuation. Respectively, $\hat{R}(\cdot, \cdot)$ is the one we choose reserve price α_{ih}^k with the estimation of $\mu_{ih}^k(\cdot, \cdot)$, i.e., $\hat{\mu}_{ih}^k(\cdot, \cdot)$.

Lemma 16. *When Lemma 15 holds, we have*

$$|R_h^k(\cdot, \cdot) - \hat{R}_h^k(\cdot, \cdot)| \leq C_6 H \log^2 K \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}},$$

where C_6 is a constant independent of K and H .

Let's have an example when $N = 1$, i.e., there is only one bidder.

Example 1. *In this situation, $R(\cdot, \cdot) = \alpha^*(1 - F(\alpha^* - 1 - \mu(\cdot, \cdot)))$ and $\hat{R}(\cdot, \cdot) = \alpha(1 - F(\alpha - 1 - \hat{\mu}(\cdot, \cdot)))$. Therefore,*

$$|R(\cdot, \cdot) - \hat{R}(\cdot, \cdot)| \leq (6C_1 + 1)C_5 \sqrt{H} \log K \|\phi(\cdot, \cdot)\|_{\Lambda^{-1}},$$

which is consistent with Lemma 16.

Now, we focus on the regret not in buffer caused by Algorithm 4, denoted by Δ_1 . In order to facilitate the understanding, we rewrite the definition of Δ_1 explicitly as follows.

$$\Delta_1 = \sum_{\tau=1}^K [V_1^{\pi^*}(x_1^k) - \tilde{V}_1^{\pi_{\tilde{k}}}(x_1^k)] \mathbb{1}(k \notin \text{buffer}).$$

Let's revisit our thought of bounding regret. We use empirical data to estimate unknown parameters, and then we assume that bidders will bid truthfully to construct the estimation of the R-function and Q-function. Then, we chase down the greedy policy. Therefore, when we take the expectation operator, we assume truthful bidding. Since Δ_5 is easy to bound, we focus on how to bound Δ_1 . With a little abuse of notation, we will use $V(\cdot)$ to replace $\tilde{V}(\cdot)$ from now on.

Then, we have the following lemma.

Lemma 17. *Under Assumption 1, Assumption 2 and Assumption 3, if we set $\text{poly}(\log K) = (C_7 + C_6 H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}$ in Algorithm 4, where $C_7 = B_8 H^{\frac{3}{2}} \log K$ and B_8 is determined in Lemma 39, it holds that with probability at least $1 - 2\delta$,*

$$\Delta_1 \leq C_8 \sqrt{H^5 K \log^5 K},$$

where C_8 is a constant independent of H and K .

B.2 Proof of Theorem 6

Let's make a decomposition of the regret at first. It holds that

$$\text{Regret} \leq \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5.$$

Δ_1 is defined in Lemma 17 and with probability at least $1 - 2\delta$, $\Delta_1 \leq C_8 \sqrt{H^5 K \log^5 K}$.

Δ_2 comes from the use of a buffer. With Lemma 10, it holds that $\Delta_2 \leq 3H \frac{3HC_2 \log^2 K}{\log \frac{1}{\gamma}}$.

Δ_3 comes from the use of policy π_{rand} . By applying Lemma 13, it holds that $\Delta_3 \leq 3H \max\{4, 1 + \frac{4}{3} \log \frac{1}{\delta}\}$ with probability at least $1 - \delta$.

Δ_4 comes from the consequence from the existence of L. Due to Lemma 12, we have $\Delta_4 \leq NH(4C_1 C_3 H + 8 \log(\frac{2NH}{\delta}))3H = 3NH^2(4C_1 C_3 H + 8 \log(\frac{2NH}{\delta}))$, with probability at least $1 - \delta$. As we have already considered the loss from buffer in Δ_2 , there is no need for us to consider it in Δ_4 .

Δ_5 comes from the difference between the expectation of revenue when buyers bid truthfully and the actual expectation of revenue when buyers overbid or underbid, but it does not change the outcome. Since we already consider the loss from buffer, the size of overbid or underbid we should think about is less than $\frac{C_3 H}{K}$ thanks to Lemma 11. Therefore, the difference between the expectation of revenue when buyers bid truthfully and the actual expectation of revenue when buyers overbid or underbid without changing the outcome is less than $\frac{C_3 H}{K}$ each step. So, it holds that $\Delta_5 \leq C_3 H^2$.

When estimating $\hat{R}(\cdot, \cdot)$, we have at most probability δ not satisfying the inequality in Lemma 14.

Consequently, we set $\delta = \frac{\rho}{5}$, and it ends our proof. \square

Appendix C. Omitted Proof in Section 4

Compared to Section B, this section introduces a well-performed estimator to estimate the underlying distribution. With its help, we prove corresponding theorems when the market noise distribution is unknown.

C.1 Useful Lemmas for Proving Theorem 8

In order to estimate noise distribution, we have the following lemma (Dvoretzky et al., 1956) to bound the gap between the true distribution and the empirical distribution. We assume that $\hat{F}(\cdot)$ and $\hat{f}(\cdot)$ inherit all the properties of $F(\cdot)$ and $f(\cdot)$, because we can easily use some smooth kernels¹ to achieve this goal. However, in order to make the paper easy to understand, we do not explicitly write down the choice of a smooth kernel.

Lemma 18. *Given $t \in \mathbb{N}$, let m_1, m_2, \dots, m_t be real-valued independent and identically distributed random variables with cumulative distribution function $F(\cdot)$. Let $\hat{F}_t(\cdot)$ denote the associated empirical distribution function defined by $\hat{F}_t(x) = \frac{1}{t} \sum_{i=1}^t \mathbf{1}_{\{m_i \leq x\}}$ where $x \in \mathbb{R}$. Then with probability $1 - \delta$, it holds*

$$\sup_x |\hat{F}_t(x) - F(x)| \leq \sqrt{\frac{1}{2} \log \frac{2}{\delta}} t^{-\frac{1}{2}}.$$

1. It may introduce a constant 2 when describing the distance of two distributions. However, it doesn't matter as we consider order only.

Now, similar to the methodology in Section B, we state the following lemmas in parallel.

Lemma 19. *Under Assumption 1 about linear MDP, it holds that the number of buffer episodes is not larger than $C_9 H \log^2 K$. Then, the number of corresponding steps is not larger than $C_9 H^2 \log^2 K$, where C_9 is a constant that only depends on d and λ .*

Recall that when market noise distribution is unknown, we implement Algorithm 7 to generate \tilde{q} and we use \tilde{q} to estimate θ instead of q . Therefore, \mathbf{L} there considers simulation outcome \tilde{q} rather than real outcome q . We formalize the definition of \mathbf{L} there as follows, and we use $\tilde{\rho}$ to represent the reserve price in Algorithm 7.

$$\begin{aligned} \mathbf{L} &= \{k : \mathbb{1}(v_{ih}^k > \max\{b_{-ih}^{k+}, \tilde{\rho}_{ih}^k\}) \neq \mathbb{1}(b_{ih}^k > \max\{b_{-ih}^{k+}, \tilde{\rho}_{ih}^k\})\}. \\ \mathbf{L} &= \{k : \mathbb{1}(v_{ih}^k > \max\{b_{-ih}^{k+}, \tilde{\rho}_{ih}^k\}) \neq \mathbb{1}(b_{ih}^k > \max\{b_{-ih}^{k+}, \tilde{\rho}_{ih}^k\})\}. \end{aligned} \quad (6)$$

Lemma 20. *With probability at least $1 - \delta$, it holds that for any given i, h*

$$\mathbf{L} \leq C_9 H \log^2 K + 4C_1 C_3 H + 8 \log\left(\frac{2NH}{\delta}\right) \leq C_{10} H \log^2 K,$$

where C_3 is defined in Lemma 11 and C_{10} is a constant independent of K and H .

Lemma 21. *We use θ_{ih}^* to denote the true parameter and $\hat{\theta}_{ih}$ to represent the outcome from Equation (4) in episode $\text{buffer.e}(\tilde{k})$. Therefore, under Assumption 2 and Assumption 3, for any i and h , it holds the following union bound that C_{11} is a constant and*

$$\sqrt{(\hat{\theta}_{ih} - \theta_{ih}^*)^T \Lambda^{\text{buffer.e}(\tilde{k})} (\hat{\theta}_{ih} - \theta_{ih}^*)} \leq C_{11} \sqrt{H} \log K,$$

with probability at least $1 - \delta$, conditional on Good Event \mathcal{E} .

Same as Lemma 15, we have the following lemma.

Lemma 22. *Conditional on Good Event \mathcal{E} , it holds that*

$$|\mu_{ih}^k(\cdot, \cdot) - \mu_{ih}^k(\cdot, \cdot)| \leq C_{11} \sqrt{H} \log K \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}.$$

Now, we introduce a lemma bounding the gap between the noise distribution $F(\cdot)$ and $\hat{F}(\cdot)$.

Lemma 23. *Conditional on Good Event \mathcal{E} , it holds with probability at least $1 - \delta$ that for any x in episode $\text{buffer.e}(\tilde{k})$*

$$\begin{aligned} |F(x) - \hat{F}(x)| &\leq \sqrt{\frac{1}{2} \log \frac{2K}{\delta}} (NH \text{buffer.e}(\tilde{k}))^{-\frac{1}{2}} + \frac{C_1 C_3 H}{K} + \frac{C_9 H \log^2 K}{\text{buffer.e}(\tilde{k})} \\ &\quad + C_1 C_{11} \sqrt{H} \log K \|\phi(x_h^\tau, v_h^\tau)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}} \\ &\leq C_{12} \frac{H \log^2 K}{\sqrt{\text{buffer.e}(\tilde{k})}}, \end{aligned}$$

where C_{12} is a constant.

Now, we begin to bound the wedge of $R(\cdot, \cdot)$ and $\hat{R}(\cdot, \cdot)$ corresponding to $\hat{F}(\cdot)$. It holds the following lemma.

Lemma 24. *Conditional on Good Event \mathcal{E} , we have*

$$|R_h^k(\cdot, \cdot) - \hat{R}_h^k(\cdot, \cdot)| \leq C_{13}H \log^2 K \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}} + C_{14} \frac{H^2 \log^4 K}{\sqrt{\text{buffer.e}(\tilde{k})}},$$

where C_{13} and C_{14} are constants independent of K and H .

We define Δ_1 as the one in Lemma 17 of Section B.

Lemma 25. *Under Assumption 1, Assumption 2 Assumption 3 and Assumption 7, if we set $\text{poly}_1(\log K) = C_{15} + C_{13}H \log^2 K$ and $\text{poly}_2(\log K) = C_{14}H^2 \log^4 K$ in Algorithm 8, where $C_{15} = D_7 H^{\frac{3}{2}}$ and D_7 is determined in Lemma 47, it holds that with probability at least $1 - 2\delta$,*

$$\Delta_1 \lesssim \tilde{\mathcal{O}}(H^3 \sqrt{K}).$$

C.2 Proof of Theorem 8

It is similar to the proof of Theorem 6. The only difference comes from Lemma 23. The probability of Bad Event \mathcal{E}^c is now less than 6δ . Then, we set $\delta = \frac{\rho}{6}$ and it ends the proof. \square

Appendix D. Auxiliary Lemmas and Proofs in Section B

In this section, we prove the lemmas mentioned in Section B in detail. It is organized by the order of lemmas.

D.1 Proof of Lemma 10

First of all, we have the following lemmas.

Lemma 26 (Lemma 2, (Gao et al., 2021)). *Assume $m \leq n$, $A = \sum_{\tau=1}^m \phi_\tau \phi_\tau^T + \lambda I$. $B = \sum_{\tau=1}^n \phi_\tau \phi_\tau^T + \lambda I$, where ϕ_τ is abridge for $\phi(x_\tau, v_\tau)$, similarly hereinafter. Then if $A^{-1} \not\leq 2B^{-1}$, we have*

$$\log \det B \geq \log \det A + \log 2.$$

Lemma 27 (Lemma 1, (Gao et al., 2021)). *Since $\|\phi_\tau\| \leq 1$. Let $A = \sum_{\tau=1}^K \phi_\tau \phi_\tau^T + \lambda I$, then we have*

$$\log \det A \leq d \log d + d \log(K + \lambda) \leq K_1 \log K.$$

Therefore, for $2(\Lambda^{\text{buffer.s}(\tilde{k}+1)})^{-1} \not\leq (\Lambda^{\text{buffer.e}(\tilde{k})})^{-1}$ and $\text{buffer.e}(\tilde{k}+1) \geq \text{buffer.s}(\tilde{k}+1)$, it holds that $2(\Lambda^{\text{buffer.e}(\tilde{k}+1)})^{-1} \not\leq (\Lambda^{\text{buffer.e}(\tilde{k})})^{-1}$. Therefore, $\det \Lambda^{\text{buffer.e}(\tilde{k}+1)} \geq 2 \det \Lambda^{\text{buffer.e}(\tilde{k})}$. Then, using Lemma 26, we know that for any h and k , it holds $\log \det \Lambda_h^k \leq K_1 \log K$. We have $\log \det \Lambda_h^0 = d \log \lambda$. Combining Lemma 26, we have that the number of buffer episodes for any h is not larger than $\frac{3 \log K}{\log \frac{1}{\gamma}} \frac{K_1 \log K - d \log \lambda}{\log 2}$. Then, there is a constant

C_2 satisfying $K_1 \log K - d \log \lambda \leq C_2 \log 2 \log K$. Therefore, the total number of episodes in buffer is not larger than $\frac{3HC_2 \log^2 K}{\log \frac{1}{\gamma}}$. For the number of total steps, it is obvious that it is smaller than H times the number of episodes. Then, it ends the proof. \square

D.2 Proof of Lemma 11

Myerson (1981) shows that the optimal strategy for a one-round second-price auction is to bid truthfully. Therefore, if a bidder overbids or underbids for more than $\frac{3H\sqrt{2N}}{K\sqrt{1-\gamma}}$, his loss holds that

$$\text{Loss} \geq \frac{1}{NHK} \frac{\beta}{2K} \frac{1}{3} \frac{\beta}{K} = \frac{3H}{K^3(1-\gamma)},$$

where $\beta = \frac{3H\sqrt{2N}}{\sqrt{1-\gamma}}$.

The inequality holds since with probability $\frac{1}{KHN}$, the policy will be π_{rand} and the bidder is selected, and the total loss is higher than the loss with policy π_{rand} . With a uniform reserve price, the probability that a loss happens is $\frac{\beta}{3K}$. Then, average loss is $\frac{\beta}{2K}$. Due to the existence of a buffer, the overbid or underbid can only make an influence on policy $t = \frac{3 \log K}{\log \frac{1}{\gamma}}$ episodes later. Because of the existence of a discount rate, an upper bound of revenue for each buyer after t episodes is $\frac{\gamma^t}{1-\gamma} 3H = \frac{3H}{K^3(1-\gamma)}$.

Therefore, with the assumption that buyers are all rational, it finishes the proof. \square

The proof also illuminates why we need to keep π_{rand} in every episode instead of only utilizing buffer periods to explore. We run π_{rand} so that each untruthful bid will suffer an immediate loss. Note that there are buffer periods before the policies are updated. Therefore, any possible reward from manipulating the policy estimate is delayed by the length of the buffer period. Consequently, rational bidders would reduce their extent of untruthful bidding, as they need to offset the immediate losses incurred by π_{rand} . If we only run π_{rand} in the buffer period as we hope to explore, we cannot bound the extent of untruthful bidding in other horizons. Then, there is no guarantee for regret bounds.

D.3 Proof of Lemma 12

For convenience, similar to Golrezaei et al. (2019), we define

$$L_i = \{t : t \in [0, K] \text{ and } \mathbf{1}(v_i^t \geq m_i^t) \neq \mathbf{1}(b_i^t \geq m_i^t)\},$$

for each buyer i .

We define $o_i^t = (b_i^t - v_i^t)_+$ and $s_i^t = (v_i^t - b_i^t)_+$, where $t = 1, \dots, K$ given h . When we can determine the subscript through the context, we omit the subscript h for convenience.

Then we define q_i^t , which is a binary variable. It equals one if buyer i wins and zero if loses. Therefore, we have $S_i = \{t : t \in [1, K], q_i^t = 0 \text{ and } s_i^t \geq \alpha\}$ and $O_i = \{t : q_i^t = 1 \text{ and } o_i^t \geq \alpha\}$. As a result, $L_i = L_i^s \cup L_i^o$, where $L_i^s = \{t : \mathbf{1}(v_i^t \geq r_i^t) = 1, \mathbf{1}(b_i^t \geq r_i^t) = 0\}$ and $L_i^o = \{t : \mathbf{1}(v_i^t \geq r_i^t) = 0, \mathbf{1}(b_i^t \geq r_i^t) = 1\}$. Finally, we have $S_i^c = \{t : q_i^t = 1 \text{ or } s_i^t \leq \alpha\}$. So, $|L_i^s| = |S_i \cap L_i^s| + |S_i^c \cap L_i^s|$.

To bound $|(S_i \cap L_i^s) \cup (O_i \cap L_i^o)|$: using Lemma 10 and Lemma 11, we have that if we set $\alpha = C_3 \frac{H}{K}$, it is bounded by $\frac{3HC_2 \log^2 K}{\log \frac{1}{\gamma}}$.

To bound $|S_i^c \cap L_i^s|$: it means that underbid changes the outcome, and the level of underbid is smaller than α . Since $|f| \leq C_1$, it holds for origin x :

$$\Pr(t \in S_i^c \cap L_i^s | \mathcal{F}_t) \leq \int_x^{x+\alpha} f(z) dz \leq C_1 \alpha.$$

Let's define $\xi_t = \mathbb{1}(t \in S_i^c \cap L_i^s)$ while $\omega_t = \Pr(t \in S_i^c \cap L_i^s | \mathcal{F}_t)$. Then $|S_i^c \cap L_i^s| = \sum_{t=1}^K \xi_t$ and $\mathbb{E}(\xi_t - \omega_t | \mathcal{F}_t) = 0$.

Using Azuma-Hoeffding inequality (Hoeffding, 1994), it holds that

$$\Pr(|S_i^c \cap L_i^s| \geq \frac{1+\iota}{1-\epsilon} \sum_1^K \omega_t) \leq \exp(-\epsilon \iota \sum_1^K \omega_t).$$

Let $A = \sum_1^K \omega_t \leq KC_1 \alpha$, $\epsilon = \frac{1}{2}$ and $\iota = \frac{2}{A} \log(\frac{2NH}{\delta})$, we have

$$|S_i^c \cap L_i^s| \leq 2(1+\iota)A \leq 2KC_1 \alpha + 4 \log(\frac{2NH}{\delta}),$$

with probability at least $1 - \frac{\delta}{2NH}$.

Similarly, we bound $|O_i^c \cap L_i^o|$ with the same bound that

$$|O_i^c \cap L_i^o| \leq 2KC_1 \alpha + 4 \log(\frac{2NH}{\delta}),$$

with probability at least $1 - \frac{\delta}{2NH}$.

Then, we set $\alpha = \frac{C_3 H}{K}$ and combine the items all to obtain

$$|L_i| \leq \frac{3HC_2 \log^2 K}{\log \frac{1}{\gamma}} + 4C_1 C_3 H + 8 \log(\frac{2NH}{\delta}),$$

with probability at least $1 - \frac{\delta}{NH}$.

With the same methodology, we obtain the union bound for any given i and h with probability at least $1 - \delta$ that

$$L \leq \frac{3HC_2 \log^2 K}{\log \frac{1}{\gamma}} + 4C_1 C_3 H + 8 \log(\frac{2NH}{\delta}),$$

and it finishes the proof. \square

D.4 Proof of Lemma 13

We use random variables X_1, \dots, X_{KH} to represent whether π_{rand} is used. If we choose policy π_{rand} , then $X = 1$, or $X = 0$ otherwise.

Using Bernstein inequalities (Bernstein, 1924), it holds that

$$\Pr(\sum_{i=1}^{KH} X_i - KH \frac{1}{KH} \geq t) \leq \exp\left\{\frac{-t^2/2}{(1 - 1/KH) + t/3}\right\},$$

since $X - \frac{1}{KH}$ has mean zero and $\text{var}(X) = \frac{1}{KH}(1 - \frac{1}{KH})$.

Therefore, set $t = \max\{3, \frac{4}{3} \log \frac{1}{\delta}\}$, the right side is smaller than δ and it finishes the proof. \square

D.5 Proof of Lemma 14

First of all, we omit subscripts i and h for convenience, and we will get the union bound in the end.

Then, we introduce some notations. We use \tilde{q}_τ to represent the outcome that every bidder bids truthfully and \hat{q}_τ to represent the outcome with real bidding. Then $\hat{\theta}$ and $\tilde{\theta}$ correspond to $\{\hat{q}_\tau\}$ and $\{\tilde{q}_\tau\}$.

Now, we focus on buyer i and step h , so we omit subscripts i and h from now on. We have the following lemma at first:

Lemma 28. *Under Equation (3), it holds that*

$$\sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} (\tilde{q}_\tau - 1 + F(m_\tau - 1 - \langle \phi_\tau, \hat{\theta} \rangle))^2 \leq \sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} (\tilde{q}_\tau - 1 + F(m_\tau - 1 - \langle \phi_\tau, \theta^* \rangle))^2 + 6L,$$

where $L \leq C_4 H \log^2 K$ due to Lemma 12.

D.5.1 PROOF OF LEMMA 28

Since there are at most L steps that overbid or underbid changes the outcome, \hat{q}_τ and \tilde{q}_τ differ in at most L different points. Since \tilde{q}_τ and \hat{q}_τ belong to $\{0, 1\}$, we have

$$\sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} (\hat{q}_\tau - 1)^2 \leq \sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} (\tilde{q}_\tau - 1)^2 + L.$$

Then, since $F(\cdot) \in [0, 1]$, it holds that

$$-2 \sum_{\tau} (1 - \hat{q}_\tau) F(m_\tau - 1 - \langle \phi_\tau, \theta \rangle) \leq -2 \sum_{\tau} (1 - \tilde{q}_\tau) F(m_\tau - 1 - \langle \phi_\tau, \theta \rangle) + 2L.$$

for any θ .

Therefore, it holds that

$$\sum_{\tau} (\tilde{q}_\tau - 1 + F(m_\tau - 1 - \langle \phi_\tau, \theta \rangle))^2 \leq \sum_{\tau} (\hat{q}_\tau - 1 + F(m_\tau - 1 - \langle \phi_\tau, \theta \rangle))^2 + 3L, \quad (7)$$

for any θ .

Finally, with the optimality of $\hat{\theta}$ and $\tilde{\theta}$, it holds that

$$\begin{aligned} & \sum_{\tau} (\tilde{q}_\tau - 1 + F(m_\tau - 1 - \langle \phi_\tau, \hat{\theta} \rangle))^2 \\ & \leq \sum_{\tau} (\hat{q}_\tau - 1 + F(m_\tau - 1 - \langle \phi_\tau, \hat{\theta} \rangle))^2 + 3L \\ & \leq \sum_{\tau} (\hat{q}_\tau - 1 + F(m_\tau - 1 - \langle \phi_\tau, \tilde{\theta} \rangle))^2 + 3L \\ & \leq \sum_{\tau} (\tilde{q}_\tau - 1 + F(m_\tau - 1 - \langle \phi_\tau, \tilde{\theta} \rangle))^2 + 6L \\ & \leq \sum_{\tau} (\tilde{q}_\tau - 1 + F(m_\tau - 1 - \langle \phi_\tau, \theta^* \rangle))^2 + 6L. \end{aligned}$$

The first and third inequalities hold due to Ineq. (7). The second and last inequalities hold because of the optimality of $\hat{\theta}$ and $\tilde{\theta}$. Then, it finishes the proof. \square

Then we use $f_{m_\tau}(\langle \phi_\tau, \theta \rangle)$ to represent $F(m_\tau - 1 - \langle \phi_\tau, \theta \rangle)$ in shorthand.

Therefore, with Lemma 28, we have

$$\sum_{\tau} [f_{m_\tau}(\langle \phi_\tau, \hat{\theta} \rangle) - f_{m_\tau}(\langle \phi_\tau, \theta^* \rangle)] \leq 2 \left| \sum_{\tau} \xi_\tau (f_{m_\tau}(\langle \phi_\tau, \hat{\theta} \rangle) - f_{m_\tau}(\langle \phi_\tau, \theta^* \rangle)) \right| + 6L,$$

where $\xi_\tau = (1 - \tilde{q}_\tau) - f_{m_\tau}(\langle \phi_\tau, \theta^* \rangle)$. The inequality holds because of simple rearrangement.

Then, we have

$$\begin{aligned} f_{m_\tau}(\langle \phi_\tau, \hat{\theta} \rangle) - f_{m_\tau}(\langle \phi_\tau, \theta^* \rangle) &= \int_{\langle \phi_\tau, \theta^* \rangle}^{\langle \phi_\tau, \hat{\theta} \rangle} f'_{m_\tau}(s) ds \\ &= \langle \phi_\tau, \hat{\theta} - \theta^* \rangle \int_0^1 f'_{m_\tau}(\langle \phi_\tau, s\hat{\theta} + (1-s)\theta^* \rangle) ds \\ &= \langle \phi_\tau, \hat{\theta} - \theta^* \rangle D_\tau, \end{aligned}$$

where $D_\tau = \int_0^1 f'_{m_\tau}(\langle \phi_\tau, s\hat{\theta} + (1-s)\theta^* \rangle) ds$.

So, it holds that

$$\sum_{\tau} D_\tau^2 (\langle \phi_\tau, \hat{\theta} - \theta^* \rangle)^2 \leq 2 \left| \sum_{\tau} \xi_\tau D_\tau \langle \phi_\tau, \hat{\theta} - \theta^* \rangle \right| + 6L.$$

Since $\|\theta\| \leq \sqrt{d}$, we use V_ϵ which is a set of ball with radius ϵ to cover $\mathcal{B}(0, \sqrt{d}) \times \mathcal{B}(0, \sqrt{d})$. Then, the cardinality of V_ϵ is smaller than $B_1 (\frac{\sqrt{d}}{\epsilon})^{2d} = \frac{B_2}{\epsilon^{2d}}$, where B_1 and B_2 are constants only depending on dimension d . Thanks to Assumption 2 and Assumption 3, we have $|f''| \leq L$ and $|D_\tau| \leq C_1$.

Therefore, for any $(\hat{\theta}, \theta^*)$, there exists (θ, θ') , which is the center of a ball in V_ϵ , so that $\|(\hat{\theta}, \theta^*) - (\theta, \theta')\| \leq \epsilon$. In this way, it holds that

$$\begin{aligned} &|\langle \phi_\tau, D_\tau(\theta, \theta')(\theta - \theta') - D_\tau(\hat{\theta}, \theta^*)(\hat{\theta} - \theta^*) \rangle| \\ &\leq 2\sqrt{d} |D_\tau(\theta, \theta') - D_\tau(\hat{\theta}, \theta^*)| + |D_\tau| (\|\theta - \hat{\theta}\| + \|\theta' - \theta^*\|) \\ &\leq 2L\sqrt{d}\epsilon + C_1\epsilon \\ &\leq (2L\sqrt{d} + C_1)\epsilon. \end{aligned}$$

The first inequality holds since $\|\theta\| \leq \sqrt{d}$. The second inequality holds since $|f''| \leq L$ and $|D_\tau| \leq C_1$.

Therefore, it holds that

$$\left\| \sum_{\tau} \xi_\tau \langle \phi_\tau, D_\tau(\hat{\theta}, \theta^*)(\hat{\theta} - \theta^*) \rangle \right\| \leq \left\| \sum_{\tau} \xi_\tau \langle \phi_\tau, D_\tau(\theta, \theta')(\theta - \theta') \rangle \right\| + (2L\sqrt{d} + C_1) \text{buffer.e}(\tilde{k})\epsilon,$$

since $|\xi_\tau| \leq 1$.

Let's define the following shorthands

$$V(\phi) = \sum_{\tau} \langle \phi_\tau, D_\tau(\theta - \theta') \rangle^2,$$

$$V(\hat{\phi}) = \sum_{\tau} \langle \phi_{\tau}, D_{\tau}(\hat{\theta} - \theta^*) \rangle^2.$$

Therefore, by applying the inequality above, we have

$$V(\phi) \leq V(\hat{\phi}) + 4C_1\sqrt{d}(2L\sqrt{d} + C_1)\mathbf{buffer.e}(\tilde{k})\epsilon. \quad (8)$$

The inequality holds because of the square difference formula.

Since for positive number a , b and c , if $a \leq b + c$, then $\sqrt{a} \leq \sqrt{b} + \sqrt{c}$. So, it holds that

$$\sqrt{V(\phi)} \leq \sqrt{V(\hat{\phi})} + \sqrt{4C_1\sqrt{d}(2L\sqrt{d} + C_1)\mathbf{buffer.e}(\tilde{k})\epsilon}. \quad (9)$$

Since θ^* is the true parameter and $\xi_{\tau} = (1 - \tilde{q}_{\tau}) - f_{m_{\tau}}(\langle \phi_{\tau}, \theta^* \rangle)$ which is determined by truthful bid, it holds $\mathbb{E}(\xi_{\tau} | \phi_{1:\tau}, \xi_{1:\tau-1}) = 0$ whose value is determined by z_{τ} only. Due to Azuma-Hoeffding inequality (Hoeffding, 1994), it holds that

$$\Pr\left[\left|\sum_{\tau} \xi_{\tau} D_{\tau} \langle \phi_{\tau}, \theta - \theta^* \rangle\right| \geq \sqrt{\log \frac{2B_2HN}{\delta\epsilon^{2d}} V(\phi)}\right] \leq \frac{\delta}{HN}, \quad (10)$$

for any (θ, θ') with probability at least $1 - \frac{\delta}{HN}$.

Therefore, it holds that

$$\begin{aligned} V(\hat{\phi}) &\leq 4C_1\sqrt{d}(2L\sqrt{d} + C_1)\mathbf{buffer.e}(\tilde{k})\epsilon + V(\phi) \\ &\leq 4C_1\sqrt{d}(2L\sqrt{d} + C_1)\mathbf{buffer.e}(\tilde{k})\epsilon + 2\sqrt{\log \frac{2B_2HN}{\delta\epsilon^{2d}} V(\phi)} + 6L \\ &\leq 4C_1\sqrt{d}(2L\sqrt{d} + C_1)\mathbf{buffer.e}(\tilde{k})\epsilon + 2\sqrt{\log \frac{2B_2HN}{\delta\epsilon^{2d}} [\sqrt{V(\hat{\phi})} \\ &\quad + \sqrt{4C_1\sqrt{d}(2L\sqrt{d} + C_1)\mathbf{buffer.e}(\tilde{k})\epsilon}]} + 6L \\ &= 4C_1\sqrt{d}(2L\sqrt{d} + C_1) + 2\sqrt{\log \frac{2B_2HN\mathbf{buffer.e}(\tilde{k})^{2d}}{\delta}} [\sqrt{V(\hat{\phi})} \\ &\quad + \sqrt{4C_1\sqrt{d}(2L\sqrt{d} + C_1)}] + 6L \\ &\leq 4C_1\sqrt{d}(2L\sqrt{d} + C_1) + 2\sqrt{\log \frac{2B_2HN\mathbf{buffer.e}(\tilde{k})^{2d}}{\delta}} [\sqrt{V(\hat{\phi})} \\ &\quad + \sqrt{4C_1\sqrt{d}(2L\sqrt{d} + C_1)}] + 6C_4H \log^2 K. \end{aligned}$$

The first inequality holds due to Ineq. (8) while the second one holds due to Ineq. (10) and Lemma 28. The third inequality holds because of Ineq. (9). The equality holds since we set $\epsilon = \frac{1}{\mathbf{buffer.e}(\tilde{k})}$. The final inequality holds because of Lemma 12.

Finally, applying the root formula of the quadratic equation, it is obvious that there exists a constant $B_3 > 0$ such that $V(\hat{\phi}) \leq B_3H \log^2 K$.

Similar to Wang et al. (2020b), we have

$$\sqrt{(\hat{\theta}_{ih} - \theta_{ih}^*)^T \Lambda \mathbf{buffer.e}(\tilde{k}) (\hat{\theta}_{ih} - \theta_{ih}^*)} \leq c_1^{-1} \sqrt{V(\hat{\phi})} + 2\sqrt{d\lambda},$$

for any i and h with probability at least $1 - \delta$.

It holds since

$$\sqrt{(\hat{\theta} - \theta^*)^T \Lambda^{\text{buffer.e}(\bar{k})} (\hat{\theta} - \theta^*)} \leq \sqrt{(\hat{\theta} - \theta^*)^T \left(\sum_{\tau} \phi_{\tau} \phi_{\tau}^T \right) (\hat{\theta} - \theta^*)} + \sqrt{(\hat{\theta} - \theta^*)^T (\lambda I) (\hat{\theta} - \theta^*)}.$$

Then, we have $D_{\tau}^2 \geq c_1^2$ and $\|(\hat{\theta}_{ih} - \theta_{ih}^*)\|_{\lambda I} \leq 2\sqrt{d\lambda}$.

In the end, we find that there exists a constant C_5 that satisfies

$$\sqrt{(\hat{\theta}_{ih} - \theta_{ih}^*)^T \Lambda^{\text{buffer.e}(\bar{k})} (\hat{\theta}_{ih} - \theta_{ih}^*)} \leq C_5 \sqrt{H} \log K,$$

which ends the proof. \square

D.6 Proof of Lemma 15

Using the Cauchy inequality, we have the following statement:

Lemma 29. *It holds that*

$$|\langle \phi(x, v), \hat{\theta} - \theta \rangle| \leq \sqrt{(\hat{\theta} - \theta)^T \Lambda (\hat{\theta} - \theta)} \|\phi(x, v)\|_{\Lambda^{-1}}.$$

Specially, taking $\Lambda = \Lambda_h^{\text{buffer.e}(\bar{k})} = \sum_{\tau=1}^{\text{buffer.e}(\bar{k})} \phi(x_h^{\tau}, v_h^{\tau}) \phi(x_h^{\tau}, v_h^{\tau})^T + \lambda I$, the inequality holds.

Then Lemma 29 and Lemma 14 lead to Lemma 15. \square

D.7 Proof of Lemma 16

Firstly, we define $\tilde{R}_h^k(\cdot, \cdot) = \sum_{i=1}^N \mathbb{E}[\max\{r_{ih}^{k-}, \alpha_{ih}^k\} \mathbf{1}(r_{ih}^k \geq \max\{r_{ih}^{k-}, \alpha_{ih}^k\})]$. Then, $|R_h^k(\cdot, \cdot) - \hat{R}_h^k(\cdot, \cdot)| \leq |R_h^k(\cdot, \cdot) - \tilde{R}_h^k(\cdot, \cdot)| + |\tilde{R}_h^k(\cdot, \cdot) - \hat{R}_h^k(\cdot, \cdot)|$.

To bound $|\tilde{R}_h^k(\cdot, \cdot) - \hat{R}_h^k(\cdot, \cdot)|$, we have

$$\begin{aligned} |\tilde{R}_h^k(\cdot, \cdot) - \hat{R}_h^k(\cdot, \cdot)| &\leq \sum_{i=1}^N \mathbb{E}[|\max\{r_{ih}^{k-}, \alpha_{ih}^k\} \mathbf{1}(r_{ih}^k \geq \max\{r_{ih}^{k-}, \alpha_{ih}^k\}) \\ &\quad - [\max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\} \mathbf{1}(\hat{r}_{ih}^k \geq \max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\})]|] \\ &\leq \sum_{i=1}^N \Delta_1 + \Delta_2 + \Delta_3 \\ &\leq (1 + 6C_1) N C_5 \sqrt{H} \log K \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\bar{k})})^{-1}}, \end{aligned}$$

where

$$\begin{aligned} \Delta_1 &= |\max\{r_{ih}^{k-}, \alpha_{ih}^k\} \mathbf{1}(r_{ih}^k \geq \max\{r_{ih}^{k-}, \alpha_{ih}^k\}) \\ &\quad - [\max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\} \mathbf{1}(\hat{r}_{ih}^k \geq \max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\})]|, \end{aligned}$$

$$\begin{aligned} \Delta_2 = & |[\max\{r_{ih}^{k-}, \alpha_{ih}^k\} \mathbb{1}(r_{ih}^k \geq \max\{r_{ih}^{k-}, \alpha_{ih}^k\})] \\ & - [\max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\} \mathbb{1}(\hat{r}_{ih}^k \geq \max\{r_{ih}^{k-}, \alpha_{ih}^k\})]| \end{aligned}$$

and

$$\begin{aligned} \Delta_3 = & |[\max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\} \mathbb{1}(\hat{r}_{ih}^k \geq \max\{r_{ih}^{k-}, \alpha_{ih}^k\})] \\ & - [\max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\} \mathbb{1}(\hat{r}_{ih}^k \geq \max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\})]|. \end{aligned}$$

The first inequality holds due to the properties of convex functions. The second inequality holds due to the triangle inequality. The third inequality holds since $\Delta_1 \leq |\max\{r_{ih}^{k-}, \alpha_{ih}^k\} - \max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\}| \leq |r - \hat{r}|$, $\Delta_2 \leq 3C_1|r - \hat{r}|$ and $\Delta_3 \leq 3C_1|r - \hat{r}|$. The reason why $\Delta_2 \leq 3C_1|r - \hat{r}|$ is $\max\{\hat{r}, \alpha\} \leq 3$ and $\mathbb{E}|\mathbb{1}(r_{ih}^k \geq \max\{r_{ih}^{k-}, \alpha_{ih}^k\}) - \mathbb{1}(\hat{r}_{ih}^k \geq \max\{r_{ih}^{k-}, \alpha_{ih}^k\})| \leq C_1|r - \hat{r}|$.

To bound $|R_h^k(\cdot, \cdot) - \tilde{R}_h^k(\cdot, \cdot)|$, we have the following lemmas. We define $W_{ih}^k(\alpha) = \mathbb{E}[\max\{v_{ih}^{k-}, \alpha\} \mathbb{1}(v_{ih}^k \geq \max\{v_{ih}^{k-}, \alpha\}) | \phi_h^k]$ at first.

Lemma 30 (Lemma C.3. (Golrezaei et al., 2019)). *Since α_{ih}^{k*} is determined by Myerson Lemma (Myerson, 1981), we have $W_{ih}^k(\alpha_{ih}^{k*}) = 0$. Furthermore, there exists a constant B_4 that for any α between α_{ih}^k and α_{ih}^{k*} , we have $|W_{ih}^k(\alpha)| \leq B_4$ for any i and h , under assumption Assumption 2, Assumption 3 and Assumption 4.*

Lemma 31 (Lemma C.4. (Golrezaei et al., 2019)). *Under Assumption 4, it holds that*

$$|\alpha_{ih}^{k*} - \alpha_{ih}^k| \leq |\langle \phi_h^k, \theta_{ih} - \hat{\theta}_{ih} \rangle|.$$

By applying Lemma 31, we have

$$\begin{aligned} |R_h^k(\cdot, \cdot) - \tilde{R}_h^k(\cdot, \cdot)| & \leq \sum_{i=1}^N \frac{B_4}{2} (\alpha_{ih}^{k*} - \alpha_{ih}^k)^2 \\ & \leq N \frac{B_4}{2} (\langle \phi_h^k, \theta_{ih} - \hat{\theta}_{ih} \rangle)^2 \\ & \leq N \frac{B_4}{2} C_5^2 H \log^2 K \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\bar{k})})^{-1}}^2 \\ & \leq N \frac{B_4}{2} C_5^2 H \log^2 K \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\bar{k})})^{-1}} \frac{1}{\sqrt{\lambda}}. \end{aligned}$$

The first inequality holds due to the Taylor expansion. The second inequality holds due to Lemma 31, while the third one holds due to Lemma 15. The last inequality holds since $\|\phi\|_{\Lambda^{-1}} \leq \frac{1}{\lambda}$.

Remark 32. *Without Assumption 4, we can get the last inequality from the integral form of $R(\cdot, \cdot)$. For example, when $N = 1$, it holds that $R(\cdot, \cdot) = \alpha(1 - F(\alpha - 1 - \langle \phi, \theta \rangle))$. Then, $|R - \tilde{R}| \leq 3C_1|\langle \theta - \hat{\theta}, \phi \rangle|$ due to Assumption 2. It shows that Assumption 4 is actually redundant as Assumption 2 exists.*

Combining the differences $|\tilde{R}_h^k(\cdot, \cdot) - \hat{R}_h^k(\cdot, \cdot)|$ and $|R_h^k(\cdot, \cdot) - \tilde{R}_h^k(\cdot, \cdot)|$, it holds that

$$|R_h^k(\cdot, \cdot) - \hat{R}_h^k(\cdot, \cdot)| \leq [(1 + 6C_1)C_5\sqrt{H} \log K + \frac{B_4}{2\sqrt{\lambda}}C_5^2H \log^2 K]N\|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}.$$

Therefore, there exists a constant C_6 which is independent of H and K , satisfying

$$|R_h^k(\cdot, \cdot) - \hat{R}_h^k(\cdot, \cdot)| \leq C_6H \log^2 K\|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}},$$

and it ends the proof. \square

D.8 Proof of Lemma 17

In order to prove Lemma 17, we have the following lemmas for help.

Lemma 33. *For any fixed policy π , let $\{\omega_h^\pi\}_{h \in [H]}$ be the corresponding vectors such that $Q_h^\pi(\cdot, \cdot) = R(\cdot, \cdot) + \langle \phi(\cdot, \cdot), \omega_h^\pi \rangle$ for any h . Then, it holds that*

$$\|\omega_h^\pi\| \leq 3H\sqrt{d},$$

for any h .

Proof Since it holds

$$Q_h^\pi(\cdot, \cdot) = (R + \mathbb{P}_h V_{h+1}^\pi)(\cdot, \cdot),$$

and the linearity of MDP, we have

$$\omega_h^\pi = \int V_{h+1}^\pi(\cdot) d\mathcal{M}_h(\cdot).$$

Therefore, considering $|V| \leq 3H$ and $\|\mathcal{M}_h(\mathcal{S})\| \leq \sqrt{d}$, Lemma 33 holds. \blacksquare

Lemma 34. *For any $(k, h) \in [K] \times [H]$, the vector $\omega_h^{\text{buffer.e}(\tilde{k})}$ in Algorithm 4 satisfies:*

$$\|\omega_h^{\text{buffer.e}(\tilde{k})}\| = \|\omega_h^k\| \leq 3H\sqrt{\frac{d\text{buffer.e}(\tilde{k})}{\lambda}} \leq 3H\sqrt{\frac{dk}{\lambda}}.$$

Proof Since we only update at episode $\text{buffer.e}(\tilde{k})$, ω_h^k is the same as $\omega_h^{\text{buffer.e}(\tilde{k})}$.

For any vector $\nu \in \mathbb{R}^d$, we have

$$\begin{aligned} |\nu^T \omega_h^{\text{buffer.e}(\tilde{k})}| &= |\nu^T (\Lambda_k^{\text{buffer.e}(\tilde{k})})^{-1} \sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} \phi_h^\tau \max_a Q_{h+1}(\cdot, \cdot)| \\ &\leq \sum_{\tau} 3H |\nu^T (\Lambda_k^{\text{buffer.e}(\tilde{k})})^{-1} \phi_h^\tau| \\ &\leq 3H \sqrt{[\sum_{\tau} \nu^T (\Lambda_k^{\text{buffer.e}(\tilde{k})})^{-1} \nu][\sum_{\tau} (\phi_h^\tau)^T (\Lambda_k^{\text{buffer.e}(\tilde{k})})^{-1} \phi_h^\tau]} \\ &\leq 3H \|\nu\| \sqrt{\frac{d\text{buffer.e}(\tilde{k})}{\lambda}}. \end{aligned}$$

The first inequality holds since $Q \leq 3H$, while the second inequality holds due to the Cauchy inequality. The third inequality holds since $(\Lambda_k^{\text{buffer.e}(\tilde{k})})^{-1} \preceq \frac{1}{\lambda}I$ and the following lemma. ■

Lemma 35 (Lemma D.1. (Jin et al., 2020)). *Let $\Lambda^{\text{buffer.e}(\tilde{k})} = \lambda I + \sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} \phi_\tau \phi_\tau^T$ where $\phi_\tau \in \mathbb{R}^d$ and $\lambda > 0$. Then it holds*

$$\sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} \phi_\tau^T (\Lambda^{\text{buffer.e}(\tilde{k})})^{-1} \phi_\tau \leq d.$$

Thus, with $\|\omega_h^{\text{buffer.e}(\tilde{k})}\| = \max_{\nu: \|\nu\|=1} |\nu^T \omega_h^{\text{buffer.e}(\tilde{k})}|$, it ends the proof. □

In order to prove the next lemma, we introduce two useful lemmas first.

Lemma 36. *For any given h , suppose $\{x_\tau\}_{\tau=1}^\infty$ being a stochastic process on state space \mathcal{S} with corresponding filtration $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$. Let $\{\phi_\tau\}_{\tau=1}^\infty$ be an \mathbb{R}^d -valued stochastic process when $\phi_\tau \in \mathcal{F}_{\tau-1}$. Since $\|\phi_\tau\| \leq 1$ and $\Lambda_{\text{buffer.e}(\tilde{k})} = \lambda I + \sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} \phi_\tau \phi_\tau^T$, then for any δ , with probability at least $1 - \delta$, for any k corresponding to $\text{buffer.e}(\tilde{k})$ and any $V \in \mathcal{V}$ so that $\sup_x |V(x)| \leq 3H$, we have*

$$\begin{aligned} \left\| \sum_{\tau=1}^k \phi_\tau \{V(x_\tau) - \mathbb{E}[V(x_\tau) | \mathcal{F}_{\tau-1}]\} \right\|_{\Lambda_{\text{buffer.e}(\tilde{k})}^{-1}}^2 &\leq \frac{54C_2 H^3 \log^2 K}{\lambda \log \frac{1}{\gamma}} + \frac{32k^2 \epsilon^2}{\lambda} \\ &\quad + 144H^2 \left[\frac{d}{2} \log \frac{k + \lambda}{\lambda} + \log \frac{\mathcal{N}_\epsilon}{\delta} \right], \end{aligned}$$

where \mathcal{N}_ϵ is the ϵ -covering number of \mathcal{V} with respect to the distance $\text{dist}(V, V') = \sup_x (V(x) - V'(x))$.

Proof First of all, we have

$$\begin{aligned} &\left\| \sum_{\tau=1}^k \phi_\tau \{V(x_\tau) - \mathbb{E}[V(x_\tau) | \mathcal{F}_{\tau-1}]\} \right\|_{\Lambda_{\text{buffer.e}(\tilde{k})}^{-1}}^2 \\ &\leq 2 \times 2 \left\| \sum_{\tau=1}^k \phi_\tau \{V(x_\tau) - \mathbb{E}[V(x_\tau) | \mathcal{F}_{\tau-1}]\} \mathbf{1}\{k \notin \text{buffer}\} \right\|_{\Lambda_k^{-1}}^2 + 2 \times 3H \frac{1}{\lambda} 3H \frac{3HC_2 \log^2 K}{\log \frac{1}{\gamma}} \\ &\leq 4 \left\| \sum_{\tau=1}^k \phi_\tau \{V(x_\tau) - \mathbb{E}[V(x_\tau) | \mathcal{F}_{\tau-1}]\} \right\|_{\Lambda_k^{-1}}^2 + \frac{54C_2 H^3 \log^2 K}{\lambda \log \frac{1}{\gamma}}. \end{aligned}$$

Firstly, we have $(a + b)^2 \leq 2a^2 + 2b^2$. Then, it holds since we divide the episodes into two parts: the ones in the buffer and the ones not. For the ones in buffer, due to the definition of $\text{buffer.e}(\tilde{k})$, it is easy to prove that it is smaller than $4 \left\| \sum_{\tau=1}^k \phi_\tau \{V(x_\tau) - \mathbb{E}[V(x_\tau) | \mathcal{F}_{\tau-1}]\} \mathbf{1}\{k \notin \text{buffer}\} \right\|_{\Lambda_k^{-1}}^2$. As for the one not in buffer, $\frac{54C_2 H^3 \log^2 K}{\lambda \log \frac{1}{\gamma}}$ is a trivial bound due to Lemma 10 and $V(\cdot) \leq 3H$.

Therefore, with Lemma D.4. in [Jin et al. \(2020\)](#), we simply replace its H with our upper bound of $V(\cdot)$, i.e., $3H$, and it finishes our proof. \blacksquare

Lemma 37. *Let \mathcal{V} denote a class of functions mapping from \mathcal{S} to \mathbb{R} with the following parametric form*

$$V(\cdot) = \min\{\max_a \omega^T \phi(\cdot, v) + \hat{R}(\cdot, v) + \beta \|\phi(\cdot, v)\|_{\Lambda^{-1}}, 3H\},$$

where $\|\omega\| \leq L$, $\beta \in [0, B]$ and the minimum eigenvalue satisfies $\lambda_{\min}(\Lambda) \geq \lambda$. Suppose $\|\phi(\cdot, \cdot)\| \leq 1$ and let \mathcal{N}_ϵ be the ϵ -covering number of \mathcal{V} with respect to the distance $\text{dist}(V, V') = \sup_x |V(x) - V'(x)|$. Then, it holds

$$\log \mathcal{N}_\epsilon \leq d \log\left(1 + \frac{8L}{\epsilon}\right) + d^2 \log\left(1 + \frac{32\sqrt{d}B^2}{\lambda\epsilon^2}\right) + dN \log\left(1 + \frac{8NB_5\sqrt{d}}{\epsilon}\right),$$

where B_5 is a constant.

Proof Due to Lemma D.6. in [Jin et al. \(2020\)](#), it holds that

$$\text{dist}(V_1, V_2) \leq \|\omega_1 - \omega_2\| + \sqrt{\|A_1 - A_2\|_F} + \sup_{x,v} |\hat{R}_1(x, v) - \hat{R}_2(x, v)|,$$

where $A = \beta^2 \Lambda^{-1}$. Let C_ω be an $\frac{\epsilon}{4}$ -cover of $\{\omega \in \mathbb{R}^d \mid \|\omega\| \leq L\}$, and then it holds $|C_\omega| \leq (1 + \frac{8L}{\epsilon})^d$. Similarly, for $\frac{\epsilon^2}{16}$ -cover for $\{A\}$, we have $|C_A| \leq [1 + \frac{32B^2\sqrt{d}}{\lambda\epsilon^2}]^{d^2}$.

Now, in order to bound the covering number corresponding to $\hat{R}(x, v)$, we show that it links to $\{\hat{\theta}_i\}_{i=1}^N$ first. As $\hat{R}(\cdot, \cdot)$ is function of $\{\hat{\mu}_i\}_{i=1}^N$ and $F(\cdot)$ is differentiable with $|f| \leq C_1$, it holds that $\frac{\partial \hat{R}}{\partial \mu_i} \leq B_5$ for any i , where B_5 is a constant. B_5 is bounded since $\mu_i \in [0, 1]$ and the interval $[0, 1]$ is compact. Therefore, since $\hat{\mu} = \langle \phi, \hat{\theta} \rangle$, it holds that

$$\begin{aligned} \sup_{x,v} |\hat{R}_1(x, v) - \hat{R}_2(x, v)| &\leq \sup_{\phi: \|\phi\| \leq 1} \sum_{i=1}^N B_5 |(\hat{\theta}_{1i} - \hat{\theta}_{2i})^T \phi| \\ &\leq \sum_{i=1}^N B_5 \|\hat{\theta}_{1i} - \hat{\theta}_{2i}\|. \end{aligned}$$

Therefore, it holds that combining $\frac{\epsilon}{2NB_5}$ -cover for $\hat{\theta}_i$,

$$|C_{\hat{R}}| \leq \left(1 + \frac{8NB_5\sqrt{d}}{\epsilon}\right)^{dN}.$$

Then, it finishes the proof. \blacksquare

Now, with lemmas prepared, we have the following lemma.

Lemma 38. For any δ , with probability at least $1 - \delta$, there exists constants B_6 and B_7 independent of K and H so that

$$\forall (k, h) \in [K] \times [H] : \left\| \sum_{\tau=1}^k \phi_h^\tau [\hat{V}_{h+1}^k(x_{h+1}^\tau) - \mathbb{P} \hat{V}_{h+1}^k(x_h^\tau, v_h^\tau)] \right\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}^2 \leq B_6 H^3 \log^2 K + B_7 H^2 \log C_7.$$

Proof Combining Lemma 34, Lemma 36 and Lemma 37, we set $L = 3H\sqrt{\frac{dk}{\lambda}}$. With Algorithm 4, we have $B = C_7 + C_6 H \log^2 K$. Then we have

$$\begin{aligned} & \left\| \sum_{\tau=1}^k \phi_h^\tau [\hat{V}_{h+1}^k(x_{h+1}^\tau) - \mathbb{P} \hat{V}_{h+1}^k(x_h^\tau, v_h^\tau)] \right\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}^2 \\ & \leq \frac{54C_2 H^3 \log^2 K}{\lambda \log \frac{1}{\gamma}} + 72dH^2 \log \frac{k+\lambda}{\lambda} + 144H^2 d \log \left(1 + \frac{24H}{\epsilon} \sqrt{\frac{dk}{\lambda}}\right) + 144H^2 \log \frac{KH}{\delta} \\ & \quad + 144H^2 d^2 \log \left[1 + \frac{32\sqrt{d}(C_7 + C_6 H \log^2 K)^2}{\lambda \epsilon^2}\right] + 144H^2 dN \log \left(1 + \frac{8NB_5\sqrt{d}}{\epsilon}\right) + \frac{32k^2\epsilon^2}{\lambda}. \end{aligned}$$

Therefore, by setting $\lambda = 1$ and $\epsilon = \frac{dH}{k}$, then we have the right side of the inequality is $\mathcal{O}(H^3 \log^2 K + H^2 \log C_7)$, and it finishes our proof. \blacksquare

Now, let's show the determination of C_7 .

Lemma 39. There exist a constant B_8 so that $C_7 = B_8 H^{\frac{3}{2}} \log K$, and for any fixed policy π , on Good Event \mathcal{E} , i.e., all inequalities hold, we have for all $(x, v, h, k) \in \mathcal{S} \times \Upsilon \times [H] \times [K]$ that:

$$\langle \phi(\cdot, \cdot), \omega_h^k \rangle + \hat{R}_h^k(\cdot, \cdot) - Q_h^\pi(\cdot, \cdot) = \mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi)(\cdot, \cdot) + \Delta_h^k(\cdot, \cdot),$$

where $\Delta_h^k(\cdot, \cdot) \leq (C_7 + C_6 H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}$.

Proof Due to Bellman equation, we know that for any $(x, v, h) \in \mathcal{S} \times \Upsilon \times [H]$, it holds

$$Q_h^\pi(\cdot, \cdot) = R_h(\cdot, \cdot) + \langle \phi(\cdot, \cdot), \omega_h^\pi \rangle = (R_h + \mathbb{P}_h V_{h+1}^\pi)(\cdot, \cdot).$$

Therefore, it gives

$$\langle \phi(\cdot, \cdot), \omega_h^k \rangle + \hat{R}_h^k(\cdot, \cdot) - Q_h^\pi(\cdot, \cdot) = \langle \phi(\cdot, \cdot), \omega_h^k - \omega_h^\pi \rangle + (\hat{R}_h^k - R_h)(\cdot, \cdot).$$

Then, since $\omega_h^k = \omega_h^{\text{buffer.e}(\tilde{k})}$, it holds that

$$\begin{aligned} \omega_h^k - \omega_h^\pi &= (\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1} \sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} \phi_h^\tau \hat{V}_{h+1}^k(x_{h+1}^\tau) - \omega_h^\pi \\ &= (\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1} \left\{ -\lambda \omega_h^\pi + \sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} \phi_h^\tau [\hat{V}_{h+1}^k(x_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^\pi(x_h^\tau, v_h^\tau)] \right\} \\ &= \delta_1 + \delta_2 + \delta_3, \end{aligned}$$

where

$$\begin{aligned}\delta_1 &= -\lambda(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}w_h^\pi, \\ \delta_2 &= (\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1} \sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} \phi_h^\tau [\hat{V}_{h+1}^k(x_{h+1}^\tau) - \mathbb{P}_h \hat{V}_{h+1}^k(x_h^\tau, v_h^\tau)], \\ \delta_3 &= (\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1} \sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} \phi_h^\tau \mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi)(x_h^\tau, v_h^\tau).\end{aligned}$$

Then, we begin to bound items corresponding to δ_1 , δ_2 , and δ_3 individually. Firstly, it holds

$$\begin{aligned}|\langle \phi(\cdot, \cdot), \delta_1 \rangle| &\leq \sqrt{\lambda} \|w_h^\pi\| \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}} \\ &\leq 3H\sqrt{d\lambda} \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}.\end{aligned}$$

The first inequality holds due to Cauchy inequality and $\Lambda_{\text{buffer.e}(\tilde{k})} \succeq \lambda I$. The second inequality holds due to Lemma 33.

Secondly, it holds that

$$|\langle \phi(\cdot, \cdot), \delta_2 \rangle| \leq \sqrt{B_6 H^3 \log^2 K + B_7 H^2 \log C_7} \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}.$$

It holds because of Lemma 38.

Lastly, we have

$$\begin{aligned}\langle \phi(\cdot, \cdot), \delta_3 \rangle &= \langle \phi(\cdot, \cdot), (\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1} \sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} \phi_h^\tau \mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi)(x_h^\tau, v_h^\tau) \rangle \\ &= \langle \phi(\cdot, \cdot), (\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1} \sum_1^{\text{buffer.e}(\tilde{k})} \phi_h^\tau (\phi_h^\tau)^T \int (\hat{V}_{h+1}^k - V_{h+1}^\pi)(x') d\mathcal{M}_h(x') \rangle \\ &= \langle \phi(\cdot, \cdot), \int (\hat{V}_{h+1}^k - V_{h+1}^\pi)(x') d\mathcal{M}_h(x') \rangle - \lambda \langle \phi(\cdot, \cdot), \int (\hat{V}_{h+1}^k - V_{h+1}^\pi) d\mathcal{M}_h \rangle \\ &= \mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi)(\cdot, \cdot) - \lambda \langle \phi(\cdot, \cdot), (\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1} \int (\hat{V}_{h+1}^k - V_{h+1}^\pi)(x') d\mathcal{M}_h(x') \rangle \\ &\leq \mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi)(\cdot, \cdot) + 3H\sqrt{d\lambda} \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}.\end{aligned}$$

The second and fourth equations hold due to the definition of the operator \mathbb{P}_h . The third equation holds due to simple algebraic arrangement. The inequality holds due to Cauchy inequality, $V(\cdot) \leq 3H$ and $\Lambda_{\text{buffer.e}(\tilde{k})} \succeq \lambda I$.

With the bounds in hand, we have $\Delta_k^h(\cdot, \cdot) \leq (3H\sqrt{d\lambda} + \sqrt{B_6 H^3 \log^2 K + B_7 H^2 \log C_7} + 3H\sqrt{d\lambda} + C_6 H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}$. Then, it is obviously that there exists a constant B_8 , so that $B_8 H^{\frac{3}{2}} \log K \geq 3H\sqrt{d\lambda} + \sqrt{B_6 H^3 \log^2 K + B_7 H^2 \log C_7} + 3H\sqrt{d\lambda}$ and it finishes the proof. \blacksquare

Now, we are ready to show the reason why we chose such a bonus. We have the following lemma.

Lemma 40. *Under the setting of Theorem 6, on the Good Event \mathcal{E} , it holds that for any $(x, v, h, k) \in \mathcal{S} \times \Upsilon \times [H] \times [K]$,*

$$\hat{Q}_h^k(x, v) \leq Q_h^{\pi^*}(x, v).$$

Proof We will prove this lemma by induction.

First of all, for the last step H , since the value function is zero at $H + 1$, we have

$$|\hat{R}_H^k(\cdot, \cdot) + \langle \phi(\cdot, \cdot), \omega_H^k \rangle - Q_H^{\pi^*}(\cdot, \cdot)| \leq (C_7 + C_6 H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_H^{\text{buffer.e}(\bar{k})})^{-1}}$$

due to Lemma 39. Therefore, we have

$$Q_H^{\pi^*}(\cdot, \cdot) \leq \min\{\hat{R}_H^k(\cdot, \cdot) + \langle \phi(\cdot, \cdot), \omega_H^k \rangle + (C_7 + C_6 H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_H^{\text{buffer.e}(\bar{k})})^{-1}}, 3H\},$$

and we use $Q_H^k(\cdot, \cdot)$ to represent the right side.

Now, supposing the statement holds at step $h + 1$, then for step h , with Lemma 39, it holds that

$$|[\hat{R}_h^k + \langle \phi, \omega_h^k \rangle - Q_h^{\pi^*} - \mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^*})](\cdot, \cdot)| \leq (C_7 + C_6 H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\bar{k})})^{-1}}.$$

By the induction assumption that $\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^*})(\cdot, \cdot) \geq 0$, it holds that

$$Q_h^{\pi^*}(\cdot, \cdot) \leq \min\{\hat{R}_h^k(\cdot, \cdot) + \langle \phi(\cdot, \cdot), \omega_h^k \rangle + (C_7 + C_6 H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\bar{k})})^{-1}}, 3H\} = Q_h^k(\cdot, \cdot),$$

which ends the proof. \blacksquare

Then, we have the following lemma about a recursive formula from $\delta_h^k = V_h^k(x_h^k) - V_h^{\pi_{\bar{k}}}(x_h^k)$.

Lemma 41. *Let $\delta_h^k = V_h^k(x_h^k) - V_h^{\pi_{\bar{k}}}(x_h^k)$ and $\xi_{h+1}^k = \mathbb{E}[\delta_{h+1}^k | x_h^k, v_h^k] - \delta_{h+1}^k$. Then conditional on Good Event \mathcal{E} , it holds that for any $(k, h) \in [K] \times [H]$,*

$$\delta_h^k \leq \delta_{h+1}^k + \xi_{h+1}^k + 2(C_7 + C_6 H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\bar{k})})^{-1}}.$$

Proof Due to Lemma 39, it holds that

$$\hat{Q}_h^k(\cdot, \cdot) - Q_h^{\pi_{\bar{k}}}(\cdot, \cdot) \leq \mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi_{\bar{k}}})(\cdot, \cdot) + 2(C_7 + C_6 H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\bar{k})})^{-1}}.$$

Then, since $\pi_{\bar{k}} = \pi_{\text{buffer.e}(\bar{k})}$ is the greedy policy before mixture at episode k by Algorithm 4, we have

$$\delta_h^k = Q_h^k(x_h^k, v_h^k) - Q_h^{\pi_{\bar{k}}}(x_h^k, v_h^k).$$

Then, it ends the proof. \square

With these preparations, we begin to prove Lemma 17.

Using notations in Lemma 41, it holds that conditional on Good Event \mathcal{E}

$$\begin{aligned}
 \Delta_1 &= \sum_{\tau=1}^K [V_1^{\pi^*}(x_1^k) - V_1^{\pi^k}(x_1^k)] \mathbf{1}(k \notin \text{buffer}) \\
 &\leq \sum_{\tau=1}^K \delta_1^k \mathbf{1}(k \notin \text{buffer}) \\
 &\leq \sum_{\tau=1}^K \sum_{h=1}^H \xi_h^k + 2(C_7 + C_6 H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\bar{k})})^{-1}} \mathbf{1}(k \notin \text{buffer}) \\
 &\leq \sum_{\tau=1}^K \sum_{h=1}^H \xi_h^k + 2\sqrt{2}(C_7 + C_6 H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_h^k)^{-1}} \mathbf{1}(k \notin \text{buffer}) \\
 &\leq \sum_{\tau=1}^K \sum_{h=1}^H \xi_h^k + 2\sqrt{2}(C_7 + C_6 H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_h^k)^{-1}}.
 \end{aligned}$$

The first inequality holds due to Lemma 40, while the second one holds due to Lemma 41. The third inequality holds due to the process of Algorithm 3, while the last one is trivial.

For the first term, since the computation of $\hat{V}_h^k(\cdot)$ is independent of the new observation x_h^k at episode k , we obtain that $\{\xi_h^k\}$ is a martingale difference sequence satisfying $|\xi_h^k| \leq 3H$ for all (k, h) . Therefore, with Azuma-Hoeffding inequality (Hoeffding, 1994), it holds

$$\Pr\left(\sum_{\tau=1}^K \sum_{h=1}^H \xi_h^k \geq \epsilon\right) \geq \exp\left(-\frac{\epsilon^2}{18KH^3}\right).$$

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

$$\sum_{\tau=1}^K \sum_{h=1}^H \xi_h^k \leq \sqrt{18KH^3 \log \frac{1}{\delta}}.$$

For the second term, thanks to Abbasi-Yadkori et al. (2011), it holds that

$$\sum_{\tau=1}^K (\phi_h^\tau)^T (\Lambda_h^\tau)^{-1} \phi_h^\tau \leq 2d \log \frac{\lambda + \tau}{\lambda}.$$

Then, with the Cauchy inequality, we have

$$\sum_{\tau=1}^K \sum_{h=1}^H \|\phi_h^\tau\|_{(\Lambda_h^\tau)^{-1}} \leq \sum_{h=1}^H \sqrt{K} \left[\sum_{\tau=1}^K (\phi_h^\tau)^T (\Lambda_h^\tau)^{-1} \phi_h^\tau \right]^{\frac{1}{2}} \leq H \sqrt{2dK \log \frac{\lambda + K}{\lambda}}.$$

Finally, combining the two terms, we have

$$\begin{aligned}
 \Delta_1 &\leq \sqrt{18KH^3 \log \frac{1}{\delta}} + 2\sqrt{2}(C_7 + C_6 H \log^2 K) H \sqrt{2dK \log \frac{\lambda + K}{\lambda}} \\
 &\leq C_8 H^{2.5} \sqrt{K \log^5 K},
 \end{aligned}$$

and it finishes our proof. ■

Appendix E. Auxiliary Lemmas and Proofs in Section C

In this section, we provide proof of lemmas in Section C in detail. We organize this section in the order of lemmas.

E.1 Proof of Lemma 19

In Algorithm 6, there are two types of $\{\mathbf{buffer.e}(\tilde{k})\}$. The number of $\{\mathbf{buffer.e}(\tilde{k})\}$ satisfying $2(\Lambda_h^k)^{-1} \not\leq (\Lambda_h^{\mathbf{buffer.e}(\tilde{k})})^{-1}$ is smaller than $\frac{3C_2H\log^2 K}{\log \frac{1}{\gamma}}$ due to Lemma 10. The number of $\{\mathbf{buffer.e}(\tilde{k})\}$ when $\log_2 k$ is an integer is smaller than $\lceil \log_2 K \rceil + 1$. Combining the two parts finishes the proof. \square

E.2 Proof of Lemma 20

Since we have a buffer period, the upper bound of the size of overbid or underbid is the same as the situation when the market noise distribution is known. Then, recall that the proof of Lemma 12 is conditional on reserve price and others' bid, it doesn't matter whether we consider q or \tilde{q} because the only difference between them is the way generating reserve has become π_0 . Conditional on reserve, the proof of Lemma 12 still holds regarding \tilde{q} .

With the same methodology in Lemma 12, we have the lemma due to Lemma 19. \square

E.3 Proof of Lemma 21

Similar to the proof of Lemma 14, we replace $1 - F(m_\tau - 1 - \langle \phi_\tau, \theta \rangle)$ by $\frac{1}{3N}(1 + \langle \phi_\tau, \theta \rangle)$ to form Equation (4). We just need to prove that $\mathbb{E}[\tilde{q} - \frac{1}{3N}(1 + \langle \phi_\tau, \theta \rangle)] = 0$ if bidders bid truthfully. If $\tilde{q}_{ih}^\tau = 1$, it satisfies that we choose i using π_0 with reserve price ρ_i and $1 + \langle \phi_\tau, \theta \rangle + z \geq \rho_i$. With some conditional probability calculation, the probability is $\frac{1}{3N}(1 + \langle \phi_\tau, \theta \rangle)$.

Therefore, by simply setting $c_1 = C_1 = \frac{1}{3N}$ in Lemma 14, we prove Lemma 21. \square

We now discuss the intuition behind the estimator of θ in Equation (4). It is a constrained form of ridge regression. First, let's clarify the rationale behind the correct choice of the loss function. In comparison to Equation (3), where the noise distribution F is known, our current scenario lacks this knowledge. Consequently, we cannot construct a random variable $q - 1 + F$, prompting the need to identify a new zero-mean random variable for estimating θ . This is why simulation is employed to generate \tilde{q} as distinct from q . We have $\mathbb{E}[3N\tilde{q} - (1 + \langle \phi_\tau, \theta \rangle)] = 0$. Thus, \tilde{q} facilitates the estimation of θ using L_2 -norm constrained ridge via Equation (4), while q persists in exploiting to prevent excessive regret. Second, instead of resorting to ordinary ridge regression, we constrain the norm of the parameters. This choice is driven by the confined space, as outlined in Assumption 1.

E.4 Proof of Lemma 23

In order to estimate $F(\cdot)$ precisely. We need to bound two-fold errors. First, we need to bound errors coming from randomness. Second, we need to bound errors from untruthful bidding.

First of all, if every buyer bids truthfully, then with Lemma 18, it holds with probability at least $1 - \frac{\delta}{K}$ for each update that

$$|F(\cdot) - \hat{F}(\cdot)| \leq \sqrt{\frac{1}{2} \log \frac{2K}{\delta}} (NH \text{buffer.e}(\tilde{k}))^{-\frac{1}{2}}.$$

However, bidders may overbid or underbid for less than $\frac{C_3 H}{K}$ due to Lemma 11 and the estimation of μ has an error. Therefore, the c.d.f that $\hat{F}(\cdot)$ estimates is not the same as $F(\cdot)$. Since $|f(\cdot)| \leq C_1$, the difference because of overbid or underbid is smaller than $\frac{C_1 C_3 H}{K}$. Then, due to Lemma 22, the difference because of error in μ is smaller than

$$C_1 C_{11} \sqrt{H} \log K \frac{\sum_{h=1}^H \sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} \|\phi(x_h^\tau, v_h^\tau)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}}{H \text{buffer.e}(\tilde{k})} \leq C_1 C_{11} \sqrt{H} \log K \frac{\sqrt{d}}{\sqrt{\text{buffer.e}(\tilde{k})}}.$$

The inequality holds since we have the mean value inequality and Lemma 35.

Since the number of episodes in buffer for each buyer i is no larger than $C_9 H \log^2 K$, it holds that

$$\begin{aligned} |F(\cdot) - \hat{F}(\cdot)| &\leq \sqrt{\frac{1}{2} \log \frac{2K}{\delta}} (NH \text{buffer.e}(\tilde{k}))^{-\frac{1}{2}} + \frac{C_1 C_3 H}{K} + \frac{C_9 H \log^2 K}{\text{buffer.e}(\tilde{k})} \\ &\quad + C_1 C_{11} \sqrt{H} \log K \frac{\sqrt{d}}{\sqrt{\text{buffer.e}(\tilde{k})}}. \end{aligned}$$

Because the number of episodes we run Equation (4) is smaller than K , then the total probability of happening Bad Event \mathcal{E}^c is smaller than δ . Then, it ends the proof. \square

E.5 Proof of Lemma 24

In order to prove Lemma 24, we introduce the following lemma first.

Lemma 42. *Under assumption Assumption 3, when Lemma 23 holds, using histogram method to estimate p.d.f $f(\cdot)$ leads to the following bound that for any x*

$$|f(x) - \hat{f}(x)| \leq D_1 \frac{\sqrt{H} \log K}{\text{buffer.e}(\tilde{k})^{\frac{1}{4}}},$$

where D_1 is a constant.

E.5.1 PROOF OF LEMMA 42

With Lemma 23 in hand, we divide $[-1, 1]$ into $2M$ parts denoted by $\{-M, \dots, 0, \dots, M-1\}$ uniformly, then we have

$$\hat{f}(x) = M \left[\hat{F}\left(\frac{i+1}{M}\right) - \hat{F}\left(\frac{i}{M}\right) \right],$$

where $x \in \left(\frac{i}{M}, \frac{i+1}{M}\right]$.

Under assumption Assumption 3, it holds that

$$|f(x) - M[F(\frac{i+1}{M}) - F(\frac{i}{M})]| \leq \frac{L}{M}.$$

Therefore, it holds that

$$|f(x) - \hat{f}(x)| \leq 2MC_{12} \frac{H \log^2 K}{\sqrt{\text{buffer.e}(\tilde{k})}} + \frac{L}{M}.$$

By setting $M = \frac{\text{buffer.e}(\tilde{k})^{\frac{1}{4}}}{\sqrt{H \log K}}$, we finish our proof. \square

Therefore, unlike Lemma 31, we have the following lemma.

Lemma 43. *Under Assumption 4, it holds that*

$$|\alpha_{ih}^{k*} - \alpha_{ih}^k| \leq |\langle \phi_h^k, \theta_{ih} - \hat{\theta}_{ih} \rangle| + \frac{D_2 H \log^2 K}{\text{buffer.e}(\tilde{k})^{\frac{1}{4}}},$$

where D_2 is a constant.

E.5.2 PROOF OF LEMMA 43

Myerson (1981) shows that the optimal reserve price satisfies

$$\alpha = 1 + \mu(\cdot, \cdot) + \phi^{-1}(-1 - \mu(\cdot, \cdot)),$$

where $\phi(x) = x - \frac{1-F(x)}{f(x)}$ is virtual valuation function.

We use α^* to denote the optimal reserve price, while $\hat{\alpha}$ denotes the reserve price we use with $\hat{F}(\cdot)$ and $\hat{f}(\cdot)$. Also, we use $\tilde{\alpha}$ to denote reserve price corresponding to $\hat{\mu}$, $F(\cdot)$ and $f(\cdot)$.

Lemma 31 shows that $|\tilde{\alpha} - \alpha^*| \leq |\langle \phi_h^k, \theta_{ih} - \hat{\theta}_{ih} \rangle|$.

To bound $|\tilde{\alpha} - \hat{\alpha}|$, we have

$$\begin{aligned} \left| \frac{1-F(\cdot)}{f(\cdot)} - \frac{1-\hat{F}(\cdot)}{\hat{f}(\cdot)} \right| &\leq \left| \frac{1-F(\cdot)}{f(\cdot)} - \frac{1-\hat{F}(\cdot)}{f(\cdot)} \right| + \left| \frac{1-\hat{F}(\cdot)}{f(\cdot)} - \frac{1-\hat{F}(\cdot)}{\hat{f}(\cdot)} \right| \\ &\leq \frac{C_{12} H \log^2 K}{c_1 \sqrt{\text{buffer.e}(\tilde{k})}} + \frac{D_1 \sqrt{H} \log K}{c_1^2 \text{buffer.e}(\tilde{k})^{\frac{1}{4}}}. \end{aligned}$$

The first inequality holds due to the triangle inequality. The second inequality holds due to Assumption 2, Lemma 23 and Lemma 42.

Then, we will show that $\phi'(\cdot) \geq 1$.

It holds that $\phi(x) = x - \frac{1-F(x)}{f(x)} = x + \frac{1}{\log'(1-F(x))}$. Under Assumption 4, it holds that $1-F(\cdot)$ is log-concave implying $\log'(1-F(\cdot))$ is decreasing. Therefore, $\phi'(x) \geq 1$.

Therefore, we have $|\phi(\hat{\alpha}) - \hat{\phi}(\hat{\alpha})| \leq \frac{C_{12} H \log^2 K}{c_1 \sqrt{\text{buffer.e}(\tilde{k})}} + \frac{D_1 \sqrt{H} \log K}{c_1^2 \text{buffer.e}(\tilde{k})^{\frac{1}{4}}}$ and $\phi(\tilde{\alpha}) = \hat{\phi}(\hat{\alpha})$. Then, it holds that

$$|\hat{\alpha} - \tilde{\alpha}| \leq \frac{C_{12} H \log^2 K}{c_1 \sqrt{\text{buffer.e}(\tilde{k})}} + \frac{D_1 \sqrt{H} \log K}{c_1^2 \text{buffer.e}(\tilde{k})^{\frac{1}{4}}},$$

because $\phi'(\cdot) \geq 1$.

Then, it ends our proof. \square

Now, we are ready to prove Lemma 24. Using notations in Lemma 16, we use another factor F to show that we use $F(\cdot)$ and $f(\cdot)$ in the function, while factor \hat{F} to denote the use of $\hat{F}(\cdot)$ and $\hat{f}(\cdot)$.

With the same methodology in Lemma 16, it holds that

$$\begin{aligned} |R_h^k(\cdot, \cdot, F) - \hat{R}_h^k(\cdot, \cdot, F)| &\leq [(1 + 6C_1)C_{11}\sqrt{H} \log K]N \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}} \\ &\quad + \frac{NB_4}{2} [2(|\langle \phi_h^k, \theta_{ih} - \hat{\theta}_{ih} \rangle|)^2 + 2(\frac{D_2 H \log^2 K}{\text{buffer.e}(\tilde{k})^{\frac{1}{4}}})^2] \\ &\leq D_3 H \log^2 K \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}} + D_4 H^2 \log^4 K \frac{1}{\sqrt{\text{buffer.e}(\tilde{k})}}, \end{aligned}$$

where D_3 and D_4 are two constants. The first inequality holds since $(a + b)^2 \leq 2(a^2 + b^2)$. The second inequality holds by rearrangement.

Then, we will bound $|\hat{R}_h^k(\cdot, \cdot, F) - \hat{R}_h^k(\cdot, \cdot, \hat{F})|$.

Since $\hat{R}_h^k(\cdot, \cdot, F) = \sum_{i=1}^N \mathbb{E}_F[\max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\} \mathbf{1}(\hat{r}_{ih}^k \geq \max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\})]$ and $\hat{R}_h^k(\cdot, \cdot, \hat{F}) = \sum_{i=1}^N \mathbb{E}_{\hat{F}}[\max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\} \mathbf{1}(\hat{r}_{ih}^k \geq \max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^k\})]$, we have that the difference of expected revenue about each buyer is smaller than $3NC_{12} \frac{H \log^2 K}{\sqrt{\text{buffer.e}(\tilde{k})}}$. It comes from the fact that the expected revenue depends on an N -fold integral with respect to the random variable $\{z_{ih}^k\}_{i=1}^N$. Since $\int x(dF - dF') = -\int (F - F')dx \leq 3\|F - F'\|_\infty \leq 3C_{12} \frac{H \log^2 K}{\sqrt{\text{buffer.e}(\tilde{k})}}$, each integral has error less than $3C_{12} \frac{H \log^2 K}{\sqrt{\text{buffer.e}(\tilde{k})}}$. With N buyers in total, it holds that

$$|\hat{R}_h^k(\cdot, \cdot, F) - \hat{R}_h^k(\cdot, \cdot, \hat{F})| \leq 3N^2 C_{12} \frac{H \log^2 K}{\sqrt{\text{buffer.e}(\tilde{k})}}.$$

Combining the two parts, it holds

$$\begin{aligned} |R_h^k(\cdot, \cdot) - \hat{R}_h^k(\cdot, \cdot)| &= |R_h^k(\cdot, \cdot, F) - \hat{R}_h^k(\cdot, \cdot, \hat{F})| \\ &\leq C_{13} H \log^2 K \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}} + \frac{C_{14} H^2 \log^4 K}{\sqrt{\text{buffer.e}(\tilde{k})}}, \end{aligned}$$

which ends the proof. Similarly, we can use Assumption 2 to achieve parallel results without Assumption 4 as Remark 32 says. \square

E.6 Proof of Lemma 25

Now, we introduce some lemmas in parallel in order to prove Lemma 25.

Lemma 44. *For any given h omitted for convenience, suppose $\{x_\tau\}_{\tau=1}^\infty$ being a stochastic process on state space \mathcal{S} with corresponding filtration $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$. Let $\{\phi_\tau\}_{\tau=1}^\infty$ be an \mathbb{R}^d -valued stochastic process when $\phi_\tau \in \mathcal{F}_{\tau-1}$. Since $\|\phi_\tau\| \leq 1$ and $\Lambda_{\text{buffer.e}(\tilde{k})} = \lambda I +$*

$\sum_{\tau=1}^{\text{buffer.e}(\tilde{k})} \phi_\tau \phi_\tau^T$, then for any δ , with probability at least $1 - \delta$, for any k corresponding to $\text{buffer.e}(\tilde{k})$ and any $V \in \mathcal{V}$ so that $\sup_x |V(x)| \leq 3H$, we have

$$\begin{aligned} \left\| \sum_{\tau=1}^k \phi_\tau \{V(x_\tau) - \mathbb{E}[V(x_\tau) | \mathcal{F}_{\tau-1}]\} \right\|_{\Lambda^{-1}}^2 &\leq \frac{54C_9 H^3 \log^2 K}{\lambda \log \frac{1}{\gamma}} + \frac{32k^2 \epsilon^2}{\lambda} \\ &\quad + 144H^2 \left[\frac{d}{2} \log \frac{k + \lambda}{\lambda} + \log \frac{\mathcal{N}_\epsilon}{\delta} \right], \end{aligned}$$

where \mathcal{N}_ϵ is the ϵ -covering number of \mathcal{V} with respect to the distance $\text{dist}(V, V') = \sup_x (V(x) - V'(x))$.

Lemma 45. Let \mathcal{V} denote a class of functions mapping from \mathcal{S} to \mathbb{R} with the following parametric form

$$V(\cdot) = \min_a \{ \max \omega^T \phi(\cdot, v) + \hat{R}(\cdot, v) + \beta \|\phi(\cdot, v)\|_{\Lambda^{-1}} + A, 3H \},$$

where $\|\omega\| \leq L$, $\beta \in [0, B]$, $A = \frac{C_{14} H^2 \log^4 K}{\sqrt{\text{buffer.e}(\tilde{k})}}$ in episode k and the minimum eigenvalue satisfies $\lambda_{\min}(\Lambda) \geq \lambda$. Suppose $\|\phi(\cdot, \cdot)\| \leq 1$ and let \mathcal{N}_ϵ be the ϵ -covering number of \mathcal{V} with respect to the distance $\text{dist}(V, V') = \sup_x |V(x) - V'(x)|$. Then, it holds

$$\log \mathcal{N}_\epsilon \leq d \log \left(1 + \frac{8L}{\epsilon} \right) + d^2 \log \left(1 + \frac{32\sqrt{d}B^2}{\lambda\epsilon^2} \right) + dN \log \left(1 + \frac{16NB_5\sqrt{d}}{\epsilon} \right) + \log \mathcal{N}_{\frac{\epsilon}{12N^2}}(\mathcal{F}),$$

where B_5 is a constant.

E.6.1 PROOF OF LEMMA 45

When $F(\cdot)$ is unknown, it holds that

$$\begin{aligned} \sup_{x,v} |\hat{R}_1(x, v) - \hat{R}_2(x, v)| &= \sup_{x,v} |\hat{R}_1(x, v, \hat{F}_1) - \hat{R}_2(x, v, \hat{F}_2)| \\ &\leq \sup_{x,v} |\hat{R}_1(x, v, \hat{F}_1) - \hat{R}_2(x, v, \hat{F}_1)| \\ &\quad + \sup_{x,v} |\hat{R}_2(x, v, \hat{F}_1) - \hat{R}_2(x, v, \hat{F}_2)|. \end{aligned}$$

Then, we use $C_{\hat{\theta}}$ to denote the cardinality of the balls corresponding to $\hat{\theta}$ and $C_{\mathcal{F}}$ to denote the cardinality of the balls corresponding to \mathcal{F} .

Like the proof of Lemma 37, we simply use $\frac{\epsilon}{4NB_5}$ -ball to cover $\hat{\theta}_i$, and it holds that

$$|C_{\hat{\theta}}| \leq \left(1 + \frac{16NB_5\sqrt{d}}{\epsilon} \right)^{dN}.$$

Conditional on ω , A and $\{\hat{\theta}_i\}_{i=1}^N$, with Lemma 24, we know that in order to satisfy $\sup_{x,v} |\hat{R}(x, v, \hat{F}) - \hat{R}(x, v, F)| \leq \frac{\epsilon}{4}$, what we need is $\|\hat{F} - F\|_\infty \leq \frac{\epsilon}{12N^2}$. Then, it ends the proof. \square

Then, it holds the following lemma.

Lemma 46. For any δ , with probability at least $1 - \delta$, there exists constants B_6 and B_7 independent of K and H so that

$$\forall (k, h) \in [K] \times [H] : \left\| \sum_{\tau=1}^k \phi_h^\tau [\hat{V}_{h+1}^k(x_{h+1}^\tau) - \mathbb{P} \hat{V}_{h+1}^k(x_h^\tau, v_h^\tau)] \right\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}^2 \leq D_5 H^3 + D_6 H^2 \log C_{15},$$

where $D_5 \sim \tilde{\mathcal{O}}(1)$ omitting $\log K$ and D_6 is a constant.

Proof Similar to the proof of Lemma 38, we just replace \mathcal{N}_ϵ by $d \log(1 + \frac{8L}{\epsilon}) + d^2 \log(1 + \frac{32\sqrt{d}B^2}{\lambda\epsilon^2}) + dN \log(1 + \frac{16NB_5\sqrt{d}}{\epsilon}) + \log \mathcal{N}_{\frac{\epsilon}{12N^2}}(\mathcal{F})$. Then, we set $\lambda = 1$, $B = C_{15} + C_{13}H \log^2 K$ and $\epsilon = \frac{dH}{k}$. With Assumption 7, we finish our proof. \blacksquare

Now, let's show the determination of C_{15} .

Lemma 47. There exist $D_7 \sim \tilde{\mathcal{O}}(1)$ so that $C_{15} = D_7 H^{\frac{3}{2}}$, and for any fixed policy π , on Good Event \mathcal{E} , i.e., all inequalities hold, we have for all $(x, v, h, k) \in \mathcal{S} \times \Upsilon \times [H] \times [K]$ that:

$$\langle \phi(\cdot, \cdot), \omega_h^k \rangle + \hat{R}_h^k(\cdot, \cdot) - Q_h^\pi(\cdot, \cdot) = \mathbb{P}_h(\hat{V}_{h+1}^k - V_{h+1}^\pi)(\cdot, \cdot) + \Delta_h^k(\cdot, \cdot),$$

where $\Delta_h^k(\cdot, \cdot) \leq (C_{15} + C_{13}H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}} + C_{14} \frac{H^2 \log^4 K}{\sqrt{\text{buffer.e}(\tilde{k})}}$.

Proof The proof of Lemma 47 is the same as proof of Lemma 39. Let's show the determination of D_7 in parallel. With Lemma 46 in hand, it holds that

$$D_7 H^{\frac{3}{2}} \geq 3H\sqrt{d\lambda} + \sqrt{D_5 H^3 + D_6 H^2 \log C_{15}} + 3H\sqrt{d\lambda}.$$

Then, it is easy to see the existence of D_7 where $D_7 \sim \tilde{\mathcal{O}}(1)$. \square \blacksquare

Also, we have the following lemma about the recursive formula from $\delta_h^k = V_h^k(x_h^k) - V_h^{\pi_{\tilde{k}}}(x_h^k)$. It holds due to Lemma 47 and Lemma 40.

Lemma 48. Let $\delta_h^k = V_h^k(x_h^k) - V_h^{\pi_{\tilde{k}}}(x_h^k)$ and $\xi_{h+1}^k = \mathbb{E}[\delta_{h+1}^k | x_h^k, v_h^k] - \delta_{h+1}^k$. Then conditional on Good Event \mathcal{E} , it holds that for any $(k, h) \in [K] \times [H]$,

$$\delta_h^k \leq \delta_{h+1}^k + \xi_{h+1}^k + 2(C_{15} + C_{13}H \log^2 K) \|\phi(\cdot, \cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}} + 2C_{14} \frac{H^2 \log^4 K}{\sqrt{\text{buffer.e}(\tilde{k})}}.$$

Now, we are ready to prove Lemma 25.

Similar to the proof of Lemma 17, it holds that

$$\Delta_1 \lesssim \tilde{\mathcal{O}}(\sqrt{H^5 K}) + \sum_{k=1}^K \sum_{h=1}^H 2C_{14} \frac{H^2 \log^4 K}{\sqrt{\text{buffer.e}(\tilde{k})}}.$$

Due to Algorithm 6, we have $k \leq 2\text{buffer.e}(\tilde{k})$. Therefore, it holds that

$$\sum_{k=1}^K \frac{1}{\sqrt{\text{buffer.e}(\tilde{k})}} \leq \sum_{k=1}^K \frac{\sqrt{2}}{\sqrt{k}} \leq 2\sqrt{2K}.$$

Therefore, it holds that

$$\Delta_1 \lesssim \tilde{O}(\sqrt{H^5 K}) + \tilde{O}(H^3 \sqrt{K}),$$

which ends the proof. □

Appendix F. Detailed Results of Numerical Experiments

In this section, we give some details about our numerical experiments.

Trail \ Regret	CLUB	SCORP	NPAC-S
1	57.20	170.77	131.41
2	139.75	230.29	113.23
3	58.01	189.06	41.46
4	238.57	168.39	54.59
5	79.43	161.72	59.99
6	171.67	211.33	53.72
7	52.24	204.67	185.61
8	59.40	185.07	135.82
9	228.57	176.15	37.69
10	150.11	181.72	91.58
11	80.74	197.85	123.08
12	179.27	167.39	239.79
13	37.25	186.11	56.14
14	83.27	168.86	240.07
15	54.92	163.89	219.48
16	72.72	175.39	86.02
17	56.35	174.99	35.80
18	55.40	178.67	52.55
19	34.40	170.65	70.55
20	15.57	160.40	169.44
21	95.18	164.27	171.89
22	324.05	176.15	24.25
23	184.31	174.79	30.46
24	41.43	174.32	64.36
25	51.32	171.11	89.65
26	30.47	177.63	191.52
27	30.46	178.80	58.29
28	367.42	182.17	84.62
29	54.69	171.78	44.27
30	114.49	174.32	33.29

Table 1: Regrets of three different algorithms in each trail.

In the contextual bandits setting, we show the total regrets of three different algorithms (i.e., CLUB, SCORP and NPAC-S) in all 30 trails in the following table. Among all 30 trials, CLUB has the lowest regret in 15 trials, while NPAC-S does in 14 trials. SCORP only wins in the twelfth trial. For their average regrets, it's 106.62 for CLUB, 178.96 for SCORP, and 99.69 for NPAC-S. Therefore, we conclude that for contextual bandit settings, the performances of CLUB and NPAC-S are comparable, overwhelming the performance of SCORP sufficiently.

For the implementation details, we assume $N = 1$, and there are two different contexts, both appearing with probability 0.5. Besides, we assume $\theta = [0, 4, 0.6]^T$ and underlying noise distribution is $\text{Unif}([-1, 1])$. In order to distinguish these strategies, we constrain that bids must be a multiple of 0.01. To simulate strategic bidders, we use Lemma 11. Once it's in the buffer period, we assume bidders randomly bid. However, if not, we assume bidders bid their value plus a random noise with scale $\frac{C_3}{K}$. For NPAC-S, we use similar ways to simulate strategic behaviors. However, for SCORP, we stated before that it uses too many episodes to explore; we loosen its constraints and assume truthful bidding. Although we only consider an upper bound for its performance, SCORP still performs worse than CLUB and NPAC-S. So, we only compare CLUB and NPAC-S in MDP settings. To solve Equation (4), we seek help from `scipy.optimize` package. Actually, most of the running time is spent on solving Equation (4). We believe we can reduce our running time by using other commercial optimization solvers.

In the MDP setting, we show the total regrets of CLUB and NPAC-S in the 30 trials in Table 2. Among all 30 trials, CLUB wins NPAC-S every time. The average of CLUB is 203.07, overwhelming the corresponding 756.31 for NPAC-S. As a result, it shows that CLUB has better performance against NPAC-S in the MDP setting.

Trail\Regret	CLUB	NPAC-S	Trail\Regret	CLUB	NPAC-S
1	111.12	719.32	16	202.51	843.94
2	86.96	744.47	17	24.77	699.18
3	369.94	694.44	18	262.83	709.15
4	78.32	1204.41	19	505.96	802.21
5	586.62	660.06	20	163.90	696.09
6	46.89	647.03	21	33.60	653.59
7	303.41	695.98	22	156.05	872.66
8	61.22	698.99	23	46.15	746.18
9	281.11	686.92	24	388.10	781.76
10	40.48	742.37	25	160.19	699.93
11	125.29	790.36	26	552.07	732.08
12	140.18	744.64	27	89.34	734.74
13	516.48	855.74	28	112.73	702.72
14	55.23	660.48	29	191.32	663.03
15	87.22	1002.02	30	311.99	804.94

Table 2: Regrets of two different algorithms in each trail.

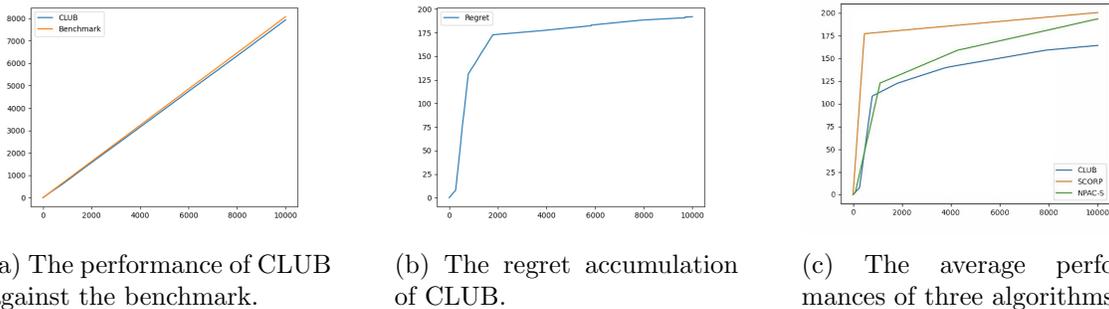


Figure 4: Experiment results for the contextual bandit setting under truncated Gaussian noise distribution.

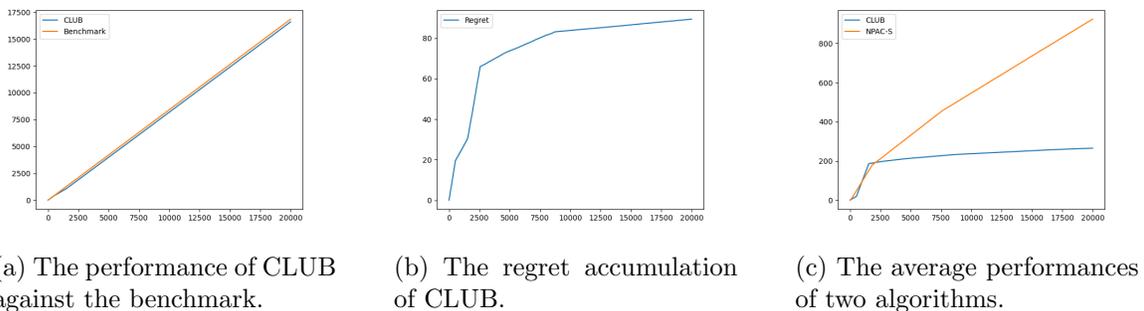


Figure 5: Experiment results for the MDP setting under truncated Gaussian noise distribution.

For the detailed setting of MDP and the implementation, we consider the situation that $H = 2$. We state different settings than the ones in contextual bandits as follows. The action space contains two actions. The first action will lead to the first context with probability 1, and the second action will lead to the second context in the next phase. In our MDP setting, we only discount once every episode, which means two phases. Therefore, we set the discount rate to be $\sqrt{\gamma}$ for NPAC-S. It is a more conservative situation and will decrease the extent of untruthful bidding for NPAC-S. At the same time, we assume NPAC-S will choose actions randomly. For our CLUB algorithm, we construct a 4-dimensional feature space to capture the structure of the underlying MDP. Additionally, instead of selecting δ , we set $\text{poly}_1(\cdot) = H \log^2(K)$ and $\text{poly}_2(\cdot) = H^2 \log^4(K)$, which decides a unique probability to break our PAC-learning bounds.

To examine the robustness of our algorithm and its potential for real-world applications, we consider different market noise distributions. Specifically, we replace $F(\cdot)$ from a uniform distribution with a normal distribution $\mathcal{N}(0, 1)$ truncated to the interval $[-1, 1]$, while keeping all other settings unchanged. For the contextual bandit setup (c.f. Figure 4), we also report one representative instance in Figure 4b. We find that under truncated Gaussian noise, our algorithm CLUB outperforms both SCORP and NPAC-S, with average 164.09, 200.29 and 193.15, respectively, demonstrating the strong performance of CLUB under different noise distributions. For the MDP setting (c.f. Figure 5), our CLUB algorithm is still

significantly plausible compared to NPAC-S. The average regret of CLUB is only 265.59, while NPAC-S incurs 923.95. It shows that our CLUB algorithm can effectively handle different types of bidders in real-world environments.

To sum up, the performance of CLUB and NPAC-S is comparable in contextual bandit settings, overwhelming the performance of SCORP sufficiently. As for MDP setting, CLUB is the only one to achieve sublinear regret bounds in both theory and practice. Therefore, CLUB captures the underlying information structures precisely and depicts a practical way in dynamic mechanism design.

References

- Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.
- Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning prices for repeated auctions with strategic buyers. *Advances in Neural Information Processing Systems*, 26, 2013.
- Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Repeated Contextual Auctions with Strategic Buyers. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.
- Susan Athey and Ilya Segal. An efficient dynamic mechanism. *Econometrica*, 81(6):2463–2485, 2013.
- Peter Auer, Nicolo Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.
- Mark Bagnoli and Ted Bergstrom. Log-concave probability and its applications. In *Rationality and Equilibrium*, pages 217–241. Springer, 2006.
- Sergei Bernstein. On a modification of chebyshev’s inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.
- Lilian Besson and Emilie Kaufmann. What doubling tricks can and can’t do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.
- Josef Broder and Paat Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.
- Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory*, 61(1):549–564, 2014.
- Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

- Bert J Claessens, Peter Vrancx, and Frederik Ruelens. Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control. *IEEE Transactions on Smart Grid*, 9(4):3259–3269, 2016.
- Yuan Deng, Sébastien Lahaie, and Vahab Mirrokni. Robust pricing in dynamic mechanism design. In *International Conference on Machine Learning*, pages 2494–2503. PMLR, 2020.
- Dmitri A Dolgov and Edmund H Durfee. Resource allocation among agents with mdp-induced preferences. *Journal of Artificial Intelligence Research*, 27:505–549, 2006.
- Alexey Drutsa. Horizon-independent optimal pricing in repeated auctions with truthful and strategic buyers. In *Proceedings of the 26th International Conference on World Wide Web*, pages 33–42, 2017.
- Alexey Drutsa. Reserve pricing in repeated second-price auctions with strategic bidders. In *International Conference on Machine Learning*, pages 2678–2689. PMLR, 2020.
- Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.
- Jianqing Fan, Yongyi Guo, and Mengxin Yu. Policy optimization using semiparametric models for dynamic pricing. *arXiv preprint arXiv:2109.06368*, 2021.
- Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.
- Minbo Gao, Tianle Xie, Simon S Du, and Lin F Yang. A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.
- Victor Ginsburgh and Jan C Van Ours. On organizing a sequential auction: results from a natural experiment by christie’s. *Oxford Economic Papers*, 59(1):1–15, 2007.
- Negin Golrezaei, Adel Javanmard, and Vahab Mirrokni. Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Advances in Neural Information Processing Systems*, 32, 2019.
- Negin Golrezaei, Patrick Jaillet, and Jason Cheuk Nam Liang. Incentive-aware contextual pricing with non-parametric market noise. In *International Conference on Artificial Intelligence and Statistics*, pages 9331–9361. PMLR, 2023.
- Wenshuo Guo, Michael Jordan, and Ellen Vitercik. No-regret learning in partially-informed auctions. In *International Conference on Machine Learning*, pages 8039–8055. PMLR, 2022.
- Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.
- Adel Javanmard and Hamid Nazerzadeh. Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research*, 20(1):315–363, 2019.

- Adel Javanmard, Hamid Nazerzadeh, and Simeng Shao. Multi-product dynamic pricing in high-dimensions with heterogeneous price sensitivity. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2652–2657. IEEE, 2020.
- Chunxiao Jiang, Yan Chen, Qi Wang, and KJ Ray Liu. Data-driven auction mechanism design in iaas cloud computing. *IEEE Transactions on Services Computing*, 11(5):743–756, 2015.
- 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, 2020.
- Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021a.
- Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021b.
- Chris Jones, Flavio Menezes, and Francis Vella. Auction price anomalies: Evidence from wool auctions in australia. *Economic Record*, 80(250):271–288, 2004.
- Yash Kanoria and Hamid Nazerzadeh. Dynamic reserve prices for repeated auctions: Learning from bids. *arXiv preprint arXiv:2002.07331*, 2020.
- Bora Keskin, David Simchi-Levi, and Prem Talwai. Dynamic pricing and demand learning on a large network of products: A pac-bayesian approach. *arXiv preprint arXiv:2111.00790*, 2021.
- Christian Kleiber and Samuel Kotz. *Statistical size distributions in economics and actuarial sciences*. John Wiley & Sons, 2003.
- Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE, 2003.
- Dingwen Kong, Ruslan Salakhutdinov, Ruosong Wang, and Lin F Yang. Online sub-sampling for reinforcement learning with general function approximation. *arXiv preprint arXiv:2106.07203*, 2021.
- Kelly Y Lange, Jeffrey W Johnson, Kris Wilson, and Wesley Johnson. Price determinants of ranch horses sold at auction in texas. Technical report, 2010.
- Sascha Lange and Martin Riedmiller. Deep auto-encoder neural networks in reinforcement learning. In *The 2010 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2010.
- Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

- Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.
- Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, pages 2173–2174. PMLR, 2019.
- Kenneth M Lusht. Order and price in a sequential auction. *The Journal of Real Estate Finance and Economics*, 8(3):259–266, 1994.
- Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.
- Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007.
- Renato Paes Leme, Martin Pal, and Sergei Vassilvitskii. A field guide to personalized reserve prices. In *Proceedings of the 25th international conference on world wide web*, pages 1093–1102, 2016.
- Sheng Qiang and Mohsen Bayati. Dynamic pricing with demand covariates. *arXiv preprint arXiv:1604.07463*, 2016.
- Alvin E Roth and Axel Ockenfels. Last-minute bidding and the rules for ending second-price auctions: Evidence from ebay and amazon auctions on the internet. *American economic review*, 92(4):1093–1103, 2002.
- Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.
- Virag Shah, Ramesh Johari, and Jose Blanchet. Semi-parametric dynamic contextual pricing. *Advances in Neural Information Processing Systems*, 32, 2019.
- Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.
- Hanrui Wang, Kuan Wang, Jiacheng Yang, Linxiao Shen, Nan Sun, Hae-Seung Lee, and Song Han. Gcn-rl circuit designer: Transferable transistor sizing with graph neural networks and reinforcement learning. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2020a.
- Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *Advances in Neural Information Processing Systems*, 34:13524–13536, 2021.
- Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2020b.

- Xiaohong Wu, Yonggen Gu, Jie Tao, Guoqiang Li, Jingti Han, and Naixue Xiong. An effective data-driven cloud resource procurement scheme with personalized reserve prices. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(8):4693–4705, 2019.
- 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.
- 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. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 13903–13916, 2020.
- Mengxin Yu, Zhuoran Yang, and Jianqing Fan. Strategic decision-making in the presence of information asymmetry: Provably efficient rl with algorithmic instruments. *arXiv preprint arXiv:2208.11040*, 2022.
- Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.
- Jun Zhao, Guang Qiu, Ziyu Guan, Wei Zhao, and Xiaofei He. Deep reinforcement learning for sponsored search real-time bidding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1021–1030, 2018.