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

Rui Ai* Boxiang Lyu[†] Zhaoran Wang[‡] Zhuoran Yang[§] Michael I. Jordan[¶] October 20, 2022

Abstract

We study reserve price optimization in multi-phase second price auctions, where 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 nontruthful bidders who aim to manipulates 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 unknown, nonlinear, and cannot even be directly observed from the environment.

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 $\widetilde{\mathcal{O}}(H^{5/2}\sqrt{K})$ revenue regret when the market noise is known and $\widetilde{\mathcal{O}}(H^3\sqrt{K})$ revenue regret when the noise is unknown with no assumptions on bidders' truthfulness.

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 is known ever since the seminal work of Myerson (1981), directly applying the result requires 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 temporal dependent dynamics, particularly, incorporating Markov Decision Process (MDP) induced dynamics in the evolution of bidders' preferences, open 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

^{*}Peking University; ruiai@pku.edu.cn

[†]The University of Chicago: blyu@chicagobooth.edu

[‡]Northwestern University; zhaoranwang@gmail.com

[§]Yale University; zhuoran.yang@yale.edu

 $[\]P$ University of California, Berkeley; jordan@cs.berkeley.edu

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 assumption on 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 nontruthful 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 $\widetilde{\mathcal{O}}(K^{2/3})$ guarantee (Amin et al., 2014; Golrezaei et al., 2019) and $\Omega(K^{2/3})$ revenue regret lower bound exists (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 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 $\widetilde{\mathcal{O}}(\sqrt{K})$ 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 $\widetilde{\mathcal{O}}(\sqrt{T})$ regret 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 $\widetilde{\mathcal{O}}(\sqrt{T})$ revenue regret bound. Drutsa (2017, 2020) studies revenue maximization in repeated second-price auction 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 the cannot be applied to our setting.

Among this line of research, Golrezaei et al. (2019) is possibly the closest to our work. The work assumes a linear stochastic contextual bandit setting, where the contexts independent and identically distributed, achieving $\tilde{\mathcal{O}}(1)$ regret when the market noise distribution is known and $\tilde{\mathcal{O}}(K^{2/3})$ when it is unknown and nonparametric. While the $\tilde{\mathcal{O}}(1)$ 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 required in our work and, even in generic linear MDPs, a $\Omega(\sqrt{K})$ regret lower bound exists (Jin et al., 2020). Moreover, our algorithm achieves $\tilde{\mathcal{O}}(\sqrt{K})$ regret when the market noise distribution is not known, only with

mild additional assumptions. 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.

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 $\mathcal{O}(\sqrt{K})$ 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 inspirations from Wang et al. (2020) and Gao et al. (2021), as we discussed previously, these inspirations are needed to 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 $\{1, \ldots, n\}$. 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.

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" 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 $\rho \in \mathbb{R}^N_{\geq 0}$, with ρ_i denoting bidder i's reserve price. The bidder with the highest bid only if he also clear his personal reserve price, i.e., $b_i \geq \rho_i$. If the bidder i receives the item, he pay 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 $q_i = 1$ (bidder i 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 $(S, \Upsilon, H, \mathbb{P}, \{r_i\}_{i=1}^N)$, with the state space denoted by S, seller's item choice space Υ^2 , horizon H, transition kernel $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^M$ where $\mathbb{P}_h : S \times \Upsilon \to \Delta(S)$, and the individual bidders' reward functions $r_i = \{r_{ih}\}_{h=1}^M$ 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 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

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

¹We 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.

²Here we use "item choice" to better illustrate what Υ intuitively represents. The term can be extended to more generic notions of seller's action.

³Wes use the term "reward" to maintain consistency with existing RL literature.

Here, z_{ih} denotes the randomness within bidders' rewards and are 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 $\mu_{i,h}: \mathcal{S} \times \Upsilon \to [0,1]$ 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 $\mathcal{A} = \Upsilon \times \mathbb{R}^N_{\geq 0}$. 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 $\rho \in \mathbb{R}^N_{\geq 0}$. The seller's policy is then $\pi = \{\pi_h\}_{h=1}^H$, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$. 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

$$R_h = \mathbb{E}_{\{z_{ih}\}_{i=1}^N} \left[\sum_{i=1}^N m_{ih} \, \mathbb{1}(m_{ih} \le b_{ih}) \right]$$
 (2.1)

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 $Q_h^{\pi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are then

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

and

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

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 $\mathbb{R}^N_{\geq 0}$ 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, we use Q(s,v) to denote Q-function for simplicity. For any function $f: \mathcal{S} \to \mathbb{R}$, we define the transition operator \mathcal{P} and the Bellman operator \mathcal{B} as

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

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

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

$$Regret(K) = \sum_{k=1}^{K} SubOpt_k(\pi_k).$$
 (2.2)

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 sake of simplicity, we assume γ is common knowledge. Bidder i's utility at step h is given by $(r_{ih}(s_h, \nu_h) - m_{ih}) \mathbb{1}(b_{ih} \geq m_{ih})$, as we note that he only receives nonzero utility upon winning the auction. His objective is to maximize his discounted cumulative utility

Utility_i =
$$\sum_{k=1}^{K} \gamma^k \mathbb{E}_{\pi_k} \left[\sum_{h=1}^{H} (r_{ih}(s_h^k, \nu_h^k) - m_{ih}^k) \mathbb{1}(b_{ih}^k \ge m_{ih}^k) \, | \, s_1^k = x_1 \right].$$

Note that in practical applications, sellers are usually more patient than bidders and discount their future rewards less. Consider sponsored search auction, where the seller usually auctions off large numbers of ad slots every day. Bidders usually urgently need advertisement 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 2.1. Assume that there exists known feature mapping $\phi : \mathcal{S} \times \Upsilon \to \mathbb{R}^d$ 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 $i \in [N]$.

We close off the section by remarking 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).

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 (2.1), 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 regretwhen K is unknown or infinite.

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.

Algorithm 1 Definition of $\pi_{\rm rand}$

- 1: **for** h = 1, ..., 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 infinite.
- 4: end for

A typical algorithm in bandit setting features $\pi_{\rm rand}$ and a sequence of learning periods that double in length (Amin et al., 2014; Golrezaei et al., 2019; Deng et al., 2020). 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, which by extension ensures that the impatient buyers need to wait longer before benefiting from untruthful reporting, deterring them from doing so.

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

- 1: Receives buffer start buffer.s($\widetilde{k}+1$) = k and end buffer.e($\widetilde{k}+1$) = $k + \frac{3 \log k}{\log(1/\gamma)}$.
- 2: Do nothing for all episodes 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 $\pi_{\widetilde{k}+1}$ and Q-function estimate $\widehat{Q}_h^{\pi_{\widetilde{k}+1}}(\cdot,\cdot)$ using Algorithm 4, and then increment buffer period counter $k \leftarrow k + 1$.

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. Inspired by low-switching cost RL literature, we use the smallest eigenvalue of the covariance matrix to determine when to start a new period, ensuring our estimate is sufficiently close to the ground truth. However, the technique leads to yet another challenge as we lose the ability to ensure sufficient delay before the buyers could benefit from untruthful behavior. To combat this, we introduce a novel technique, "buffer period", explicitly focing the bidders to wait before starting a new learning period, thereby decreases the discounted utility the impatient bidders may gain from untruthfulness. More concretely, we describe buffer periods in Algorithm 2.

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

- 1: Initialize policy estimate π_0 , buffer period counter $\widetilde{k}=0$, buffer period starting points $\mathtt{buffer.s}(0)=1$, and buffer period end points buffer.e(0) = 1.
- 2: **for** episodes k = 1, ..., K **do**
- Execute mixture policy $\frac{1}{HK} \circ \pi_{\text{rand}} + (1 \frac{1}{HK}) \circ \pi_{\widetilde{k}}$, collecting outcomes q_{ih}^{τ} and updating covariance
- 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 $h \in [H]$.

 If there exists $h \in [H]$ such that $(\Lambda_h^{\text{buffer.e}(\widetilde{k})})^{-1} \succeq 2(\Lambda_h^k)^{-1}$, schedule a new buffer period starting at buffer.s $(\widetilde{k}+1) = k$ and ending at buffer.e $(\widetilde{k}+1) = k + \frac{3 \log k}{\log(1/\gamma)}$ using Algorithm 2.
- 5: end for

With buffer periods defined, we summarize CLUB's update schedule in Algorithm 3 and include Figure 1 for visual representation. Let $\frac{1}{HK} \circ \pi_{\text{rand}} + (1 - \frac{1}{HK}) \circ \pi_{\widetilde{k}}$ represent a mixture policy combining π_{rand} and $\pi_{\widetilde{k}}$ where for each h, with probability $\frac{1}{HK}$ we act according to π_{rand} and with probability $1 - \frac{1}{HK}$ according to $\pi_{\widetilde{k}}$. For convenience, we assume buffer. $e(\widetilde{k})$ is an integer, as rounding up buffer. $e(\widetilde{k})$ 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.

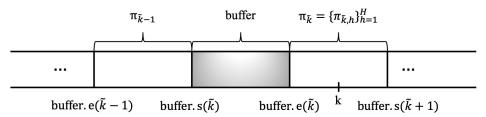


Figure 1: Learning periods and buffer periods.

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. While $\pi_{\rm rand}$ 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 then tackle the third challenge outlined in the abstract: provably efficient reinforcement learning even when the per-step revenue is nonlinear.

Addressing Challenge 3: 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. 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

$$\widehat{\theta}_{ih} = \underset{\|\theta\| \le 2\sqrt{d}}{\operatorname{argmin}} \sum_{\tau=1}^{\text{buffer.e}(\widetilde{k}+1)} (q_{ih}^{\tau} - 1 + F(m_{ih}^{\tau} - 1 - \langle \phi(x_h^{\tau}, v_h^{\tau}), \theta \rangle))^2, \tag{3.1}$$

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 (3.1) 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 $\widehat{Q}_{h}^{\pi_{\widetilde{k}+1}}(\cdot,\cdot)$

```
1: Estimate \widehat{\theta}_{ih} using (3.1) and set \widehat{\mu}_{ih}(\cdot,\cdot) \leftarrow \langle \phi(\cdot,\cdot), \widehat{\theta}_{ih} \rangle for all i,h.

2: Estimate reserve price \widehat{\rho}_{ih}(\cdot,\cdot) = \operatorname{argmax}_{y} y(1 - F(y - 1 - \widehat{\mu}_{ih}(\cdot,\cdot))) for all i,h.

3: Estimate revenue \widehat{R}_{h}(\cdot,\cdot) \leftarrow \mathbb{E}[\max\{\widetilde{b}_{h}^{-}(\cdot,\cdot),\widehat{\rho}_{h}^{+}(\cdot,\cdot)\} \mathbb{1}(\widetilde{b}_{h}^{+}(\cdot,\cdot) \geq \widehat{\rho}_{h}^{+}(\cdot,\cdot))].

4: for h = H, \ldots, 1 do

5: \Lambda_{h} \leftarrow \sum_{\tau=1}^{\operatorname{buffer.e}(\widetilde{k}+1)} \phi(x_{h}^{\tau}, v_{h}^{\tau}) \phi(x_{h}^{\tau}, v_{h}^{\tau})^{T} + \lambda I.

6: \omega_{h} \leftarrow \Lambda_{h}^{-1} \sum_{\tau=1}^{\operatorname{buffer.e}(\widetilde{k}+1)} \phi(x_{h}^{\tau}, v_{h}^{\tau})[\max_{v} \widehat{Q}_{h+1}(x_{h+1}^{\tau}, v)].

7: \widehat{Q}_{h}^{\pi_{k+1}}(\cdot,\cdot) \leftarrow \min\{\omega_{h}^{T}\phi(\cdot,\cdot) + \widehat{R}(\cdot,\cdot) + \operatorname{poly}(\log K) \|\phi(\cdot,\cdot)\|_{\Lambda_{h}^{-1}}, 3H\}.

8: \pi_{\widetilde{k}+1,h}^{v}(\cdot) \leftarrow \operatorname{argmax}_{v} \widehat{Q}_{h}^{\pi_{k+1}}(\cdot,v).

9: \pi_{k+1,h}^{\rho_{i}}(\cdot) \leftarrow \widehat{\rho}_{ih}(\cdot,\pi_{k+1,h}^{v}(\cdot)).

10: end for

11: Return \{\widehat{Q}_{h}^{\pi_{k+1}}(\cdot,\cdot)\}_{h=1}^{H} and \{\pi_{k+1,h}(\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 to two halves: per-step revenue estimation (lines 1 to 3) and transition estimation (lines 4 to 10). In the first half, we use (3.1) 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.

While the rest of Algorithm 4 resembles a typical LSVI-UCB algorithm (Jin et al., 2020), we highlight several key differences. First, we use the plug-in revenue estimate, whereas existing works estimates 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 addressed 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.

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 3.1. 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 3.2. Market noise pdf f is differentiable and its derivative is bounded. That is, there exists a constant L such that $|f'| \leq L$.

Assumption 3.3. Market noise cdf $F(\cdot)$ and $1 - F(\cdot)$ are log-concave.

At a high level, Assumptions 3.1 and 3.2 ensures that the pdf f is generally well-behaved, namely, bounded and smooth. Assumption 3.3 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 truncated Gaussian distribution and uniform distribution (Golrezaei et al., 2019).

We are now ready to state our result. If we set poly(log K) = $C_7 + C_6 H \log^2 K$ in Algorithm 4, where constant C_6 is determined in Theorem A.7 and constant $C_7 = B_8 H^{\frac{3}{2}} \log K$ with constant B_8 determined in Theorem C.13, then we have Theorem 3.4.

Theorem 3.4. Under Theorem 2.1, Theorem 3.1 and Theorem 3.2, for any fixed failure probability $\delta \in (0,1)$, with probability at least $1 - \delta$, Algorithm 3 achieves at most $\widetilde{\mathcal{O}}(\sqrt{H^5K})$ revenue regret, where $\widetilde{\mathcal{O}}(\cdot)$ hides only absolute constants and logarithmic terms.

Proof. See Appendix A 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 $\widetilde{\mathcal{O}}(1)$ revenue regret in the stochastic bandit setting). The closest work we are aware of is Cesa-Bianchi et al. (2014), which obtains a similar $\widetilde{\mathcal{O}}(\sqrt{K})$ 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), incorporates pure exploration rounds to address the challenge, yet necessitates a $\tilde{\mathcal{O}}(K^{2/3})$ revenue regret. In this section, we instead introduce a novel technique dubbed "simulation", eliminates the need for pure exploration rounds and achieves instead a $\tilde{\mathcal{O}}(\sqrt{K})$ 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

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.

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

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

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 is a power of 2, also ensuring that $\widehat{F}(\cdot)$ 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 buffer.s(0) = 1, and buffer period end points buffer.e(0) = 1.
- 2: for episodes $k = 1, \dots, K$ do
- 3: Execute mixture policy $\frac{1}{HK} \circ \pi_{\text{rand}} + (1 \frac{1}{HK}) \circ \pi_{\widetilde{k}}$, collecting outcomes q_{ih}^{τ} and updating covariance 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 $(\Lambda_h^{\mathtt{buffer.e}(\widetilde{k})})^{-1} \succeq 2(\Lambda_h^k)^{-1}$ or $\log_2(k)$ is an integer, schedule a new buffer period starting at $\mathtt{buffer.s}(\widetilde{k}+1) = k$ and ending at $\mathtt{buffer.e}(\widetilde{k}+1) = k + \frac{3\log k}{\log(1/\gamma)}$ using Algorithm 5.
- 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 (3.1) 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 revene 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 $\pi_{\rm rand}$, similar to the technique in Golrezaei et al. (2019). However, doing so incurs $\widetilde{\mathcal{O}}(K^{2/3})$ revenue regret, as $\pi_{\rm rand}$ 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 $\widetilde{q}_{ih}^{\tau} = \mathbb{1}(b_{ih}^{\tau} \geq \widetilde{\rho}_{ih}^{\tau})$, where for each h, τ we select one $i \in [N]$ uniformly at random and then draw $\widetilde{\rho}_{ih}^{\tau}$ from Unif([0,3]). For all $j \neq i$ we set $\widetilde{\rho}_{ih}^{\tau}$ to ∞ . At a high level, $\widetilde{q}_{ih}^{\tau}$ "simulates" executing π_{rand} : holding x_h^{τ} and v_h^{τ} constant, what would be the outcome if we were to

Algorithm 7 Simulation

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

act according to $\pi_{\rm rand}$ instead? As we do not need to execute $\pi_{\rm rand}$, revenue regret can be decreased. Furthermore, \tilde{q}_{ih}^{τ} still enjoys the same 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 $\widehat{F}(\cdot)$ and $F(\cdot)$ decays at a rate of $O(1/\sqrt{K})$. If we simply use Equation (3.1), only replacing $F(\cdot)$ with $\widehat{F}(\cdot)$, the estimation error is roughly on the order of $\widetilde{\mathcal{O}}(\sqrt{\text{buffer.e}(\widetilde{k}+1)})$ which precludes achieving $\widetilde{\mathcal{O}}(\sqrt{K})$ regret. Second, replacing $\widetilde{q}_{ih}^{\tau}$ 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 $\pi_{\rm rand}$. Even when the bidders report truthfully, we cannot guarantee that $\Pr(q_{ih}^{\tau}=1\,|\,x_h^{\tau},v_h^{\tau})$ 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 (4.1).

$$\widehat{\theta}_{ih} = \underset{\|\theta\| \le 2\sqrt{d}}{\operatorname{argmin}} \sum_{\tau=1}^{\operatorname{buffer.e}(\widetilde{k}+1)} (3N\widetilde{q}_{ih}^{\tau} - (1 + \langle \phi(x_h^{\tau}, v_h^{\tau}), \theta \rangle))^2,$$

$$\widehat{F}(z) = \frac{1}{N \operatorname{buffer.e}(\widetilde{k}+1)H} \sum_{i=1}^{N} \sum_{\tau=1}^{\operatorname{buffer.e}(\widetilde{k}+1)} \sum_{h=1}^{H} \mathbb{1}(b_{i\tau h} - 1 - \langle \phi_h^{\tau}, \widehat{\theta}_{ih} \rangle \le z).$$

$$(4.1)$$

We note that we are simply using a histogram to estimate $F(\cdot)$ and, as we have successfully decoupled estimation error of $F(\cdot)$ from that of θ_{ih} , using histogram is sufficient for achieving $\mathcal{O}(\sqrt{K})$ revenue regret. We then introduce Algorithm 8, whose key difference with Algorithm 4 lies in the added uncertainty due to F 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 for 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 third challenge, we build off of the simulation technique and introduce new estimation procedure for jointly estimating $F(\cdot)$ and θ .

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

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

Assumption 4.1. The market noise distribution $F(\cdot)$ belongs to a distribution family \mathcal{F} .

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

```
1: Collect simulation outcome \widetilde{q} using Algorithm 7.

2: Estimate \widehat{\theta}_{ih}, \widehat{F}(\cdot) using (4.1).

3: Estimate \widehat{\mu}_{ih}(\cdot, \cdot) \leftarrow \langle \phi(\cdot, \cdot), \widehat{\theta}_{ih} \rangle.

4: Set reserve price \widehat{\rho}_{ih}(\cdot, \cdot) = \operatorname{argmax}_y y(1 - \widehat{F}(y - 1 - \widehat{\mu}(\cdot, \cdot))).

5: Estimate revenue \widehat{R}_h(\cdot, \cdot) \leftarrow \mathbb{E}[\max\{\widetilde{b}_h^-(\cdot, \cdot), \widehat{\rho}_h^+(\cdot, \cdot)\} \, \mathbb{1}(\widetilde{b}_h^+(\cdot, \cdot) \geq \widehat{\rho}_h^+(\cdot, \cdot))].

6: for h = H, \dots, 1 do

7: \Lambda_h \leftarrow \sum_{\tau=1}^{\operatorname{buffer.e}(\widetilde{k}+1)} \phi(x_h^{\tau}, v_h^{\tau}) \phi(x_h^{\tau}, v_h^{\tau})^T + \lambda I. \triangleright We set \lambda = 1 in this paper.

8: \omega_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{\operatorname{buffer.e}(\widetilde{k}+1)} \phi(x_h^{\tau}, v_h^{\tau}) [\max_a \widehat{Q}_{h+1}(x_{h+1}^{\tau}, a)].

9: \widehat{Q}_h^{\pi_{\widetilde{k}+1}}(\cdot, \cdot) \leftarrow \min\{\omega_h^T \phi(\cdot, \cdot) + \widehat{R}(\cdot, \cdot) + \operatorname{poly}_1(\log K) \| \phi(\cdot, \cdot) \|_{\Lambda_h^{-1}} + \frac{\operatorname{poly}_2(\log K)}{\sqrt{\operatorname{buffer.e}(\widetilde{k}+1)}}, 3H\}

10: \pi_{\widetilde{k}+1,h}^{\mathcal{P}_i}(\cdot) \leftarrow \operatorname{argmax}_v \widehat{Q}_h^{\pi_{\widetilde{k}+1}}(\cdot, v).

11: \pi_{\widetilde{k}+1,h}^{\mathcal{P}_i}(\cdot) \leftarrow \widehat{\rho}_{ih}(\cdot, \pi_{k+1,h}^a(\cdot)).

12: end for

13: Return \{\widehat{Q}_h^{\pi_{\widetilde{k}+1}}(\cdot, \cdot)\}_{h=1}^H and \{\pi_{\widetilde{k}+1,h}(\cdot)\}_{h=1}^H.
```

We further let $\mathcal{N}_{\epsilon}(\mathcal{F})$ be the ϵ -covering number of \mathcal{F} with respect to the metric that $\operatorname{dist}(F,G) = \sup_x |F(x) - G(x)|$. We now have our main theorem when noise distribution is unknown. If we let $\operatorname{poly}_1(\log K) = C_{15} + C_{13}H\log^2 K$ and $\operatorname{poly}_2(\log K) = C_{14}H^2\log^4 K$ in Algorithm 8, where $C_{15} = D_7H^{\frac{3}{2}}$ and the constant D_7 is determined in Theorem D.6, constants C_{13} and C_{14} are determined in Theorem B.7, we would attain the following regret guarantee.

Theorem 4.2. Under Assumptions 2.1, 3.1, 3.2, and 4.1, when $F(\cdot)$ is unknown, for any fixed failure probability $\delta \in (0,1)$, Algorithm 6 achieves at most $\widetilde{\mathcal{O}}(H^3\sqrt{K} + H^{2.5}\sqrt{K\log\mathcal{N}_{1/K}(\mathcal{F})})$ regret with probability at least $1 - \delta$ in the worst case, where $\widetilde{\mathcal{O}}(\cdot)$ hides only absolute constants and logarithmic terms.

Proof. See Appendix B for a detailed proof.

We highlight that when $\mathcal{N}_{1/K}(\mathcal{F})$ is polynomial in 1/K, an implicit assumption found in Kong et al. (2021); Foster et al. (2021); Jin et al. (2021a), Theorem 4.2 shows that Algorithm 6 achieves $\widetilde{\mathcal{O}}(\sqrt{K})$ 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 break the well-known $\Omega(K^{2/3})$ revenue lower bound in Kleinberg and Leighton (2003) with the help of Assumptions 3.1 and 3.2. Nevertheless, as we argued previously, these assumptions are satisfied by widely-used parametric distribution families such as normal distribution and truncated normal distribution (Golrezaei et al., 2019), hence our result still remains broadly applicable.

5 Proof Sketch

Before sketching out the proof techniques, we take a slight detour and discuss the how revenue regret could be decomposed. Recall that $\pi_{\widetilde{k}}$ denotes the optimistic policy estimate maintained from episode $\mathtt{buffer.e}(\widetilde{k})+1$ to $\mathtt{buffer.s}(\widetilde{k}+1)$, namely the estimate from the end of the \widetilde{k} -th buffer period to the start of the $(\widetilde{k}+1)$ -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 $\pi_{\widetilde{k}}$ is executed and some steps where π_{rand} is executed. Let $\mathbb{1}(k \in \text{buffer})$ indicate the event that there

exists some integer \widetilde{k} such that $k \in [\mathtt{buffer.s}(\widetilde{k}), \mathtt{buffer.e}(\widetilde{k})]$, i.e. the episode k is within a buffer period. To better highlight the effect of untruthfulness, we let \widetilde{V} denote the optimistic V-function estimate if all bidders were to report truthfully.

5.1 Regret Decomposition

The regret can be decomposed to the following five parts.

- 1. $\Delta_1 = \sum_{k=1}^K [V_1^{\pi^*}(x_1) \widetilde{V}_1^{\pi_k}(x_1)] \mathbb{1}(\pi_k = \pi_{\widetilde{k}} \text{ and } k \notin \text{buffer})$. 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. $\Delta_2 = \sum_{k=1}^K [V_1^{\pi^*}(x_1) V_1^{\pi_k}(x_1)] \mathbb{1}(k \in \text{buffer})$. The second term comes from the buffer periods, which causes suboptimality as we intentionally delayed the policy update in order to further punish untruthfulness. While conducive to more truthful reports, delayed update schedule induces regret as the policy estimate is stale during these buffer periods.
- 3. $\Delta_3 = \sum_{k=1}^K [V_1^{\pi^*}(x_1) V_1^{\pi_k}(x_1)] \, \mathbb{1}(\text{exists } h \text{ such that } \pi_{k,h} = \pi_{\text{rand}} \text{ and } k \notin \text{buffer}).$ The third term Δ_3 is caused by π_{rand} , as it sets reserve prices and chooses items entirely randomly.
- 4. $\Delta_4 = \sum_{k=1}^K [V_1^{\pi^*}(x_1) V_1^{\pi_k}(x_1)] \, \mathbbm{1}(k \in L \text{ and } k \not\in \text{buffer})$. We only provide intuition behind the term L and defer its precise mathematical definition to (A.1) for when $F(\cdot)$ is known and (B.1) for when $F(\cdot)$ is not. The term L is a collection of episode indices where the bidders' untruthful bids altered the outcome of the multi-phase auction, through either q_{ih} or \widetilde{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. $\Delta_5 = \sum_{k=1}^K [\widetilde{V}_1^{\pi_k}(x_1) V_1^{\pi_k}(x_1)] \mathbb{1}(\pi_k = \pi_{\widetilde{k}} \text{ and } k \notin \text{buffer})$. 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 5.1. With Δ_1 to Δ_5 defined as above, it holds that Regret $\leq \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5$.

Proof. Since our benchmark is the maximized 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 $\Delta_1 + \Delta_5 = \sum_{k=1}^K [V_1^{\pi^*}(x_1) - V_1^{\pi_k}(x_1)] \mathbb{1}(\pi_k = \pi_{\widetilde{k}} \text{ and } k \notin \text{buffer}) \text{ and } \mathbb{1}(k \in \text{buffer}) + \mathbb{1}(\pi_k = \pi_{\widetilde{k}} \text{ and } k \notin \text{buffer}) + \mathbb{1}(k \in L \text{ and } k \notin \text{buffer}) \geq 1$, it holds that

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

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 of 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 force 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 Theorem A.2 in Appendix A.

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 show 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 Theorem A.3 and Theorem B.3.

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. (2020) 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 Theorem A.5 and Theorem A.6. When the market noise distribution is unknown, the step corresponds to Theorem B.4, Theorem B.5, and Theorem B.6.

Step 4: Prove $\widehat{R}(\cdot,\cdot) \approx R(\cdot,\cdot)$ and extend LSVI-UCB to non-linear reward function. By applying Taylor expansion to the revenue function R_h , we relate the accuracy of $\widehat{R}(\cdot,\cdot)$ to accuracy of $\widehat{\theta}_{ih}$, which is shown to be accurate in the previous step. We can then show our policy estimate $\pi_{\widetilde{k}}$ is approximately optimal with standard LSVI-UCB analysis. Steps 3 and 4 then combine to control Δ_1 . The step corresponds to Theorem A.7, Theorem A.9, Theorem B.7, and Theorem B.8.

Step 5: Limit the effects of $\pi_{\rm rand}$ and buffer periods. We finally control the revenue regret due to $\pi_{\rm rand}$ and buffer periods. A key observation is that the number of times in which $\pi_{\rm rand}$ is executed and the length of the buffer periods are all in $\mathcal{O}(1)$ and hence do not harm our regret asymptotically. Consequently, terms Δ_2 and Δ_3 are controlled effectively. For $\pi_{\rm rand}$, the step corresponds to Theorem A.4 and as for buffer periods, it corresponds to Theorem A.1 and Theorem B.2.

6 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 $\widetilde{\mathcal{O}}(H^3\sqrt{K})$ regret, improving upon the $\widetilde{\mathcal{O}}(K^{2/3})$ guarantees in Golrezaei et al. (2019); Amin et al. (2014), 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? We leave these interesting questions as potential next steps.

Acknowledgements

Rui Ai is partially supported by the elite undergraduate training program of School of Mathematical Sciences in Peking University. Part of the work is done when Boxiang Lyu is a student researcher at Google Research, Brain Team. Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports. Zhuoran Yang acknowledges Simons Institute (Theory of Reinforcement Learning).

References

- Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. Advances in neural information processing systems, 24.
- Amin, K., Rostamizadeh, A. and Syed, U. (2013). Learning prices for repeated auctions with strategic buyers. Advances in Neural Information Processing Systems, 26.
- Amin, K., Rostamizadeh, A. and Syed, U. (2014). Repeated Contextual Auctions with Strategic Buyers. In *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc.
- Athey, S. and Segal, I. (2013). An efficient dynamic mechanism. Econometrica, 81 2463–2485.
- Auer, P., Cesa-Bianchi, N. and Gentile, C. (2002). Adaptive and self-confident on-line learning algorithms. Journal of Computer and System Sciences, 64 48–75.
- Bagnoli, M. and Bergstrom, T. (2006). Log-concave probability and its applications. In *Rationality and Equilibrium*. Springer, 217–241.
- Bernstein, S. (1924). On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, **1** 38–49.
- Besson, L. and Kaufmann, E. (2018). What doubling tricks can and can't do for multi-armed bandits. arXiv preprint arXiv:1803.06971.
- Broder, J. and Rusmevichientong, P. (2012). Dynamic pricing under a general parametric choice model. Operations Research, 60 965–980.
- Cesa-Bianchi, N., Gentile, C. and Mansour, Y. (2014). Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory*, **61** 549–564.
- Chu, W., Li, L., Reyzin, L. and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings.
- Deng, Y., Lahaie, S. and Mirrokni, V. (2020). Robust pricing in dynamic mechanism design. In *International Conference on Machine Learning*. PMLR.
- Dolgov, D. A. and Durfee, E. H. (2006). Resource allocation among agents with mdp-induced preferences. Journal of Artificial Intelligence Research, 27 505–549.
- Drutsa, A. (2017). Horizon-independent optimal pricing in repeated auctions with truthful and strategic buyers. In *Proceedings of the 26th International Conference on World Wide Web*.
- Drutsa, A. (2020). Reserve pricing in repeated second-price auctions with strategic bidders. In *International Conference on Machine Learning*. PMLR.
- Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics* 642–669.

- Fan, J., Guo, Y. and Yu, M. (2021). Policy optimization using semiparametric models for dynamic pricing. arXiv preprint arXiv:2109.06368.
- Foster, D. J., Kakade, S. M., Qian, J. and Rakhlin, A. (2021). The statistical complexity of interactive decision making. arXiv preprint arXiv:2112.13487.
- Gao, M., Xie, T., Du, S. S. and Yang, L. F. (2021). A provably efficient algorithm for linear markov decision process with low switching cost. arXiv preprint arXiv:2101.00494.
- Ginsburgh, V. and Van Ours, J. C. (2007). On organizing a sequential auction: results from a natural experiment by christie's. Oxford Economic Papers, **59** 1–15.
- Golrezaei, N., Javanmard, A. and Mirrokni, V. (2019). Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Advances in Neural Information Processing Systems*, **32**.
- Guo, W., Jordan, M. and Vitercik, E. (2022). No-regret learning in partially-informed auctions. In *International Conference on Machine Learning*. PMLR.
- Hoeffding, W. (1994). Probability inequalities for sums of bounded random variables. In *The collected works* of Wassily Hoeffding. Springer, 409–426.
- Javanmard, A. and Nazerzadeh, H. (2019). Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research*, **20** 315–363.
- Javanmard, A., Nazerzadeh, H. and Shao, S. (2020). Multi-product dynamic pricing in high-dimensions with heterogeneous price sensitivity. In 2020 IEEE International Symposium on Information Theory (ISIT). IEEE.
- Jiang, C., Chen, Y., Wang, Q. and Liu, K. R. (2015). Data-driven auction mechanism design in iaas cloud computing. *IEEE Transactions on Services Computing*, **11** 743–756.
- Jin, C., Liu, Q. and Miryoosefi, S. (2021a). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. Advances in neural information processing systems, **34** 13406–13418.
- Jin, C., Yang, Z., Wang, Z. and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*. PMLR.
- Jin, Y., Yang, Z. and Wang, Z. (2021b). Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*. PMLR.
- Jones, C., Menezes, F. and Vella, F. (2004). Auction price anomalies: Evidence from wool auctions in australia. *Economic Record*, **80** 271–288.
- Kanoria, Y. and Nazerzadeh, H. (2020). Dynamic reserve prices for repeated auctions: Learning from bids. arXiv preprint arXiv:2002.07331.
- Keskin, B., Simchi-Levi, D. and Talwai, P. (2021). Dynamic pricing and demand learning on a large network of products: A pac-bayesian approach. arXiv preprint arXiv:2111.00790.
- Kleiber, C. and Kotz, S. (2003). Statistical size distributions in economics and actuarial sciences. John Wiley & Sons.
- Kleinberg, R. and Leighton, T. (2003). 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. IEEE.
- Kong, D., Salakhutdinov, R., Wang, R. and Yang, L. F. (2021). Online sub-sampling for reinforcement learning with general function approximation. arXiv preprint arXiv:2106.07203.

- Lange, K. Y., Johnson, J. W., Wilson, K. and Johnson, W. (2010). Price determinants of ranch horses sold at auction in texas. Tech. rep.
- Lattimore, T. and Szepesvári, C. (2020). Bandit algorithms. Cambridge University Press.
- Li, Y., Wang, Y. and Zhou, Y. (2019). Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*. PMLR.
- Lusht, K. M. (1994). Order and price in a sequential auction. The Journal of Real Estate Finance and Economics, 8 259–266.
- Myerson, R. B. (1981). Optimal auction design. Mathematics of operations research, 6 58–73.
- Nisan, N., Roughgarden, T., Tardos, E. and Vazirani, V. V. (2007). Algorithmic Game Theory. Cambridge University Press.
- Qiang, S. and Bayati, M. (2016). Dynamic pricing with demand covariates. arXiv preprint arXiv:1604.07463.
- Roth, A. E. and Ockenfels, A. (2002). Last-minute bidding and the rules for ending second-price auctions: Evidence from ebay and amazon auctions on the internet. *American economic review*, **92** 1093–1103.
- Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, **35** 395–411.
- Shah, V., Johari, R. and Blanchet, J. (2019). Semi-parametric dynamic contextual pricing. Advances in Neural Information Processing Systems, 32.
- Uehara, M., Zhang, X. and Sun, W. (2021). Representation learning for online and offline rl in low-rank mdps. arXiv preprint arXiv:2110.04652.
- Wang, T., Zhou, D. and Gu, Q. (2021). Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. Advances in Neural Information Processing Systems, 34 13524–13536.
- Wang, Y., Wang, R., Du, S. S. and Krishnamurthy, A. (2020). Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*.
- Yang, L. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*. PMLR.
- Yang, Z., Jin, C., Wang, Z., Wang, M. and Jordan, M. I. (2020). On function approximation in reinforcement learning: optimism in the face of large state spaces. In *Proceedings of the 34th International Conference* on Neural Information Processing Systems.
- Yu, M., Yang, Z. and Fan, J. (2022). Strategic decision-making in the presence of information asymmetry: Provably efficient rl with algorithmic instruments. arXiv preprint arXiv:2208.11040.
- Zanette, A., Lazaric, A., Kochenderfer, M. and Brunskill, E. (2020). Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*. PMLR.
- Zhao, J., Qiu, G., Guan, Z., Zhao, W. and He, X. (2018). Deep reinforcement learning for sponsored search real-time bidding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining.*

A 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 untruthfully bidding. Then, we will show that we are able to estimate unknown parameters accurately. Finally, combing them leads to bounded regret with high probability.

A.1 Useful Lemmas for Proving Theorem 3.4

Now, we begin to prove our conclusions. First of all, we show the following lemma to bound the number of buffers.

Lemma A.1. Under Theorem 2.1 about linear MDP, it holds that the number of episodes of buffer is not larger than $\frac{3HC_2\log^2 K}{\log\frac{1}{\gamma}}$. Then, the number of corresponding steps is not larger than $\frac{3H^2C_2\log^2 K}{\log\frac{1}{\gamma}}$, where C_2 is a constant only depends on d and λ .

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

Lemma A.2. 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 $\frac{C_3H}{K}$.

Then we define L being 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 Theorem A.2. We formalize the definition of L for any given i, h as follows.

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

Lemma A.3. With probability at least $1 - \delta$, it holds that for any given i, h

$$L \le \frac{3HC_2\log^2 K}{\log\frac{1}{\gamma}} + 4C_1C_3H + 8\log(\frac{2NH}{\delta}) \le C_4H\log^2 K,$$

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 A.4. 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 $\widehat{\mu}_{ih}(\cdot,\cdot)$ and $\mu_{ih}(\cdot,\cdot)$ for any bidder i and step h. It holds the following lemma.

Lemma A.5. We use θ_{ih}^* to denote the true parameter and $\widehat{\theta}_{ih}$ to represent the outcome from Equation (3.1) in episode buffer.e(\widetilde{k}). Therefore, under Theorem 3.1 and Theorem 3.2, for any i and h, it holds the following union bound that C_5 is a constant and

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

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

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

Lemma A.6. Conditional on Good Event \mathcal{E} , it holds that

$$|\widehat{\mu}_{ih}^k(\cdot,\cdot) - \mu_{ih}^k(\cdot,\cdot)| \le C_5 \sqrt{H} \log K \|\phi(\cdot,\cdot)\|_{(\Lambda_h^{\mathrm{buffer.e}(\widetilde{k})})^{-1}},$$

where $\mathtt{buffer.e}(\widetilde{k})$ is the last episode using Equation (3.1) before episode k, similarly hereinafter.

Now, we focus on the gap between $R(\cdot,\cdot)$ and the estimate $\widehat{R}(\cdot,\cdot)$. 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 $\widehat{R}_h^k(\cdot,\cdot) = \sum_{i=1}^N \mathbb{E}[\max\{\widehat{r}_{ih}^{k-},\alpha_{ih}^k\} \mathbb{1}(\widehat{r}_{ih}^k \geq \max\{\widehat{r}_{ih}^{k-},\alpha_{ih}^k\})]$. 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 every one bids truthfully based on his valuation. Respectively, $\widehat{R}(\cdot,\cdot)$ is the one we choose reserve price α_{ih}^k with the estimation of $\mu_{ih}^k(\cdot,\cdot)$, i.e., $\widehat{\mu}_{ih}^k(\cdot,\cdot)$.

Lemma A.7. When Theorem A.6 holds, we have

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

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 A.8. In this situation, $R(\cdot, \cdot) = \alpha^*(1 - F(\alpha^* - 1 - \mu(\cdot, \cdot)))$ and $\widehat{R}(\cdot, \cdot) = \alpha(1 - F(\alpha - 1 - \widehat{\mu}(\cdot, \cdot)))$. Therefore,

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

which is consistent with Theorem A.7.

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.

$$\Delta_1 = \sum_{\tau=1}^K [V_1^{\pi^*}(x_1^k) - \widetilde{V}_1^{\pi_{\widetilde{k}}}(x_1^k)] \, \mathbb{1}(k \not\in \mathtt{buffer}).$$

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 R-function and Q-function. Then, we chase down the greedy policy. Therefore, when we take 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 $\widetilde{V}(\cdot)$ from now on.

Then, we have the following lemma.

Lemma A.9. Under Theorem 2.1, Theorem 3.1 and Theorem 3.2, if we set poly(log K) = $(C_7 + C_6 H \log^2 K) \times \|\phi(\cdot,\cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}$ in Algorithm 4, where $C_7 = B_8 H^{\frac{3}{2}} \log K$ and B_8 is determined in Theorem C.13, it holds that with probability at least $1-2\delta$,

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

where C_8 is a constant independent of H and K.

A.2 Proof of Theorem 3.4

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

Regret
$$\leq \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 + \Delta_5$$
.

 Δ_1 is defined in Theorem A.9 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 buffer. With Theorem A.1, 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 $\pi_{\rm rand}$. By applying Theorem A.4, 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 Theorem A.3, we have $\Delta_4 \leq NH(4C_1C_3H + 8\log(\frac{2NH}{\delta}))3H = 3NH^2(4C_1C_3H + 8\log(\frac{2NH}{\delta}))$, with probability at least $1 - \delta$. As we have already considered 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 $\frac{C_3H}{K}$ thanks to Theorem A.2. Therefore, 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 is less than $\frac{C_3H}{K}$ each step. So, it holds that $\Delta_5 \leq C_3H^2$.

When estimating $\widehat{R}(\cdot,\cdot)$, we have at most probability δ not satisfying the inequality in Theorem A.5. Consequently, we set $\delta = \frac{p}{5}$, and it ends our proof.

B Omitted Proof in Section 4

Comparing to Appendix A, this section introduce a well-performed estimator to estimate underlying distribution. With the help of it, we prove corresponding the theorems when the market noise distribution is unknown.

B.1 Useful Lemmas for Proving Theorem 4.2

In order to estimate noise distribution, we have the following lemma (Dvoretzky et al., 1956) to bound the gap between true distribution and empirical distribution. We assume that $\widehat{F}(\cdot)$ and $\widehat{f}(\cdot)$ inherit all the properties of $F(\cdot)$ and $f(\cdot)$, because we can easily use some smooth kernels⁴ to achieve this goal. However, in order to make the paper easy to understand, we do not explicitly write down the choice of smooth kernel.

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

$$\sup_{x} |\widehat{F}_{t}(x) - F(x)| \le \sqrt{\frac{1}{2} \log \frac{2}{\delta}} t^{-\frac{1}{2}}.$$

Now, similar to the methodology in Appendix A, we state the following lemmas parallelly.

Lemma B.2. Under Theorem 2.1 about linear MDP, it holds that the number of episodes of buffer is not larger than $C_9H\log^2 K$. Then, the number of corresponding steps is not larger than $C_9H^2\log^2 K$, 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, L there considers simulation outcome \tilde{q} rather than real outcome q. We formalize the definition of L there as follows and we use $\tilde{\rho}$ to represent reserve price in Algorithm 7.

$$L = \{k : \mathbb{1}(v_{ih}^k > \max\{b_{-ih}^{k+}, \widetilde{\rho}_{ih}^k\}) \neq \mathbb{1}(b_{ih}^k > \max\{b_{-ih}^{k+}, \widetilde{\rho}_{ih}^k\})\}.$$

$$L = \{k : \mathbb{1}(v_{ih}^k > \max\{b_{-ih}^{k+}, \widetilde{\rho}_{ih}^k\}) \neq \mathbb{1}(b_{ih}^k > \max\{b_{-ih}^{k+}, \widetilde{\rho}_{ih}^k\})\}. \tag{B.1}$$

⁴It may introduce a constant 2 when describing the distance of two distributions. However, it doesn't matter as we consider order only.

Lemma B.3. With probability at least $1 - \delta$, it holds that for any given i, h

$$L \le C_9 H \log^2 K + 4C_1 C_3 H + 8 \log(\frac{2NH}{\delta}) \le C_{10} H \log^2 K$$

where C_3 is defined in Theorem A.2 and C_{10} is a constant independent of K and H.

Lemma B.4. We use θ_{ih}^* to denote the true parameter and $\hat{\theta}_{ih}$ to represent the outcome from Equation (4.1) in episode buffer.e(\tilde{k}). Therefore, under Theorem 3.1 and Theorem 3.2, for any i and h, it holds the following union bound that C_{11} is a constant and

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

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

As same as Theorem A.6, we have the following lemma.

Lemma B.5. Conditional on Good Event \mathscr{E} , it holds that

$$|\widehat{\mu}_{ih}^k(\cdot,\cdot) - \mu_{ih}^k(\cdot,\cdot)| \leq C_{11} \sqrt{H} \log K \|\phi(\cdot,\cdot)\|_{(\Lambda_{\operatorname{L}}^{\operatorname{buffer.e}(\bar{k})})^{-1}}.$$

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

Lemma B.6. Conditional on Good Event \mathscr{E} , it holds with probability at least $1-\delta$ that for any x in episode buffer.e(\widetilde{k})

$$\begin{split} |F(x) - \widehat{F}(x)| &\leq \sqrt{\frac{1}{2}\log\frac{2K}{\delta}} (NH \texttt{buffer.e}(\widetilde{k}))^{-\frac{1}{2}} + \frac{C_1C_3H}{K} + \frac{C_9H\log^2K}{\texttt{buffer.e}(\widetilde{k})} \\ &\quad + C_1C_{11}\sqrt{H}\log K \overline{\|\phi(x_h^{\tau}, \upsilon_h^{\tau})\|}_{(\Lambda_h^{\texttt{buffer.e}(\widetilde{k})})^{-1}} \\ &\leq C_{12}\frac{H\log^2K}{\sqrt{\texttt{buffer.e}(\widetilde{k})}}, \end{split}$$

where C_{12} is a constant.

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

Lemma B.7. Conditional on Good Event \mathscr{E} , we have

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

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

We define Δ_1 as the one in Theorem A.9 of Appendix A.

Lemma B.8. Under Theorem 2.1, Theorem 3.1 Theorem 3.2 and Theorem 4.1, if we set $\operatorname{poly}_1(\log K) = C_{15} + C_{13}H \log^2 K$ and $\operatorname{poly}_2(\log K) = C_{14}H^2 \log^4 K$ in Algorithm 8, where $C_{15} = D_7H^{\frac{3}{2}}$ and D_7 is determined in Theorem D.6, it holds that with probability at least $1 - 2\delta$,

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

B.2 Proof of Theorem 4.2

It is similar to proof of Theorem 3.4. The only difference comes from Theorem B.6. The probability of Bad Event \mathscr{E}^c is now less than 6δ . Then, we set $p = \frac{\delta}{6}$ and it ends the proof.

C Auxiliary Lemmas and Proofs in Appendix A

In this section, we prove the lemmas mentioned in Appendix A detailedly. It is organized by the order of lemmas.

C.1 Proof of Theorem A.1

First of all, we have the following lemmas.

Lemma C.1 (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 $A^{-1} \not\prec 2B^{-1}$, we have

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

Lemma C.2 (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

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

Therefore, for $2(\Lambda^{\text{buffer.s}(\widetilde{k}+1)})^{-1} \not\succ (\Lambda^{\text{buffer.e}(\widetilde{k})})^{-1}$ and $\text{buffer.e}(\widetilde{k}+1) \geq \text{buffer.s}(\widetilde{k}+1)$, it holds that $2(\Lambda^{\text{buffer.e}(\widetilde{k}+1)})^{-1} \not\succ (\Lambda^{\text{buffer.e}(\widetilde{k})})^{-1}$. Therefore, $\det \Lambda^{\text{buffer.e}(\widetilde{k}+1)} \geq 2 \det \Lambda^{\text{buffer.e}(\widetilde{k})}$. Then, using Theorem C.1, 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 Theorem C.1, we have that the number of episodes of buffer 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 $K_1 \log K - d \log \lambda \leq C_2 \log 2 \log K$. There-

fore, the total 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.

C.2 Proof of Theorem A.2

Myerson (1981) shows that the optimal strategy for 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

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

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

The inequality holds since that with probability $\frac{1}{KHN}$, the policy will be $\pi_{\rm rand}$ and the bidder is selected, and the total loss is higher than the loss with policy $\pi_{\rm rand}$. With a uniform reserve price, the probability that loss happens is $\frac{\beta}{3K}$. Then, average loss is $\frac{\beta}{2K}$. Since the existence of 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 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.

C.3 Proof of Theorem A.3

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

$$L_i = \{t : t \in [0, K] \text{ and } \mathbb{1}(v_i^t \ge m_i^t) \ne \mathbb{1}(b_i^t \ge m_i^t)\},$$

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, ..., 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, $L_i = L_i^s \bigcup L_i^o$, where $L_i^s = \{t : \mathbb{1}(v_i^t \ge r_i^t) = 1, \mathbb{1}(b_i^t \ge r_i^t) = 0\}$ and $L_i^o = \{t : \mathbb{1}(v_i^t \ge r_i^t) = 0, \mathbb{1}(b_i^t \ge r_i^t) = 1\}$. Finally, we have $S_i^c = \{t : q_i^t = 1 \text{ or } s_i^t \le \alpha\}$. 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 Theorem A.1 and Theorem A.2, we have that if we set $\alpha = C_3 \frac{H}{K}$, it is bounded by $\frac{3HC_2 \log^2 K}{\log \frac{1}{2}}$.

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:

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

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 $\mathbb{E}(\xi_t - \xi_t)$ $\omega_t \mid \mathcal{F}_t = 0.$

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

$$\Pr(|S_i^c \bigcap L_i^s| \ge \frac{1+\iota}{1-\epsilon} \sum_{1}^K \omega_t) \le \exp(-\epsilon\iota \sum_{1}^K \omega_t).$$

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

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

with probability at least $1 - \frac{\delta}{2NH}$. Similarly, we bound $|O_i^c \cap L_i^c|$ with the same bound that

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

with probability at least $1-\frac{\delta}{2NH}$. Then, we set $\alpha=\frac{C_3H}{K}$ and combine the items all to obtain

$$|L_i| \le \frac{3HC_2\log^2 K}{\log\frac{1}{\alpha}} + 4C_1C_3H + 8\log(\frac{2NH}{\delta}),$$

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

$$\mathtt{L} \leq \frac{3HC_2\log^2K}{\log\frac{1}{\gamma}} + 4C_1C_3H + 8\log(\frac{2NH}{\delta}),$$

and it finishes the proof.

C.4Proof of Theorem A.4

We use random variables X_1, \ldots, 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

$$\Pr(\sum_{i=1}^{KH} X_i - KH \frac{1}{KH} \ge t) \le \exp\{\frac{-t^2/2}{(1 - 1/KH) + t/3}\},\,$$

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.

C.5 Proof of Theorem A.5

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 \widetilde{q}_{τ} to represent the outcome that every bidder bids truthfully and \widehat{q}_{τ} to represent the outcome with real bidding. Then $\widehat{\theta}$ and $\widetilde{\theta}$ correspond to $\{\widehat{q}_{\tau}\}$ and $\{\widetilde{q}_{\tau}\}$.

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 C.3. Under Equation (3.1), it holds that

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

where $L \leq C_4 H \log^2 K$ due to Theorem A.3.

C.5.1 Proof of Theorem C.3

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

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

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

$$-2\sum_{\tau} \widehat{(1-q_{\tau})} F(m_{\tau}-1-\langle \phi_{\tau},\theta \rangle) \leq -2\sum_{\tau} \widetilde{(1-q_{\tau})} F(m_{\tau}-1-\langle \phi_{\tau},\theta \rangle) + 2L.$$

for any θ .

Therefore, it holds that

$$\sum_{\tau} (\widetilde{q}_{\tau} - 1 + F(m_{\tau} - 1 - \langle \phi_{\tau}, \theta \rangle))^{2} \le \sum_{\tau} (\widehat{q}_{\tau} - 1 + F(m_{\tau} - 1 - \langle \phi_{\tau}, \theta \rangle))^{2} + 3L, \tag{C.1}$$

for any θ .

Finally, with the optimality of $\widehat{\theta}$ and $\widetilde{\theta}$, it holds that

$$\sum_{\tau} (\widetilde{q}_{\tau} - 1 + F(m_{\tau} - 1 - \langle \phi_{\tau}, \widehat{\theta} \rangle))^{2}$$

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

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

$$\leq \sum_{\tau} (\widetilde{q}_{\tau} - 1 + F(m_{\tau} - 1 - \langle \phi_{\tau}, \widetilde{\theta} \rangle))^{2} + 6L$$

$$\leq \sum_{\tau} (\widetilde{q}_{\tau} - 1 + F(m_{\tau} - 1 - \langle \phi_{\tau}, \widehat{\theta} \rangle))^{2} + 6L.$$

The first and third inequalities holds due to Ineq. (C.1). The second and last inequalities hold because of the optimality of $\widehat{\theta}$ and $\widetilde{\theta}$. Then, it finishes the proof.

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 Theorem C.3, we have

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

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

$$f_{m_{\tau}}(\langle \phi_{\tau}, \widehat{\theta} \rangle) - f_{m_{\tau}}(\langle \phi_{\tau}, \theta^{*} \rangle) = \int_{\langle \phi_{\tau}, \theta^{*} \rangle}^{\langle \phi_{\tau}, \widehat{\theta} \rangle} f'_{m_{\tau}}(s) ds$$

$$= \langle \phi_{\tau}, \widehat{\theta} - \theta^{*} \rangle \int_{0}^{1} f'_{m_{\tau}}(\langle \phi_{\tau}, s\widehat{\theta} + (1 - s)\theta^{*} \rangle) ds$$

$$= \langle \phi_{\tau}, \widehat{\theta} - \theta^{*} \rangle D_{\tau},$$

where $D_{\tau} = \int_{0}^{1} f'_{m_{\tau}}(\langle \phi_{\tau}, s\hat{\theta} + (1-s)\theta^{*} \rangle) ds$. So, it holds that

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

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 Theorem 3.1 and Theorem 3.2, we have $|f''| \leq L$ and $|D_{\tau}| \leq C_1$.

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

$$|\langle \phi_{\tau}, D_{\tau}(\theta, \theta')(\theta - \theta') - D_{\tau}(\widehat{\theta}, \theta^{*})(\widehat{\theta} - \theta^{*})\rangle|$$

$$\leq 2\sqrt{d}|D_{\tau}(\theta, \theta') - D_{\tau}(\widehat{\theta}, \theta^{*})| + |D_{\tau}|(\|\theta - \widehat{\theta}\| + \|\theta' - \theta^{*}\|)$$

$$\leq 2L\sqrt{d}\epsilon + C_{1}\epsilon$$

$$\leq (2L\sqrt{d} + C_{1})\epsilon.$$

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

$$\|\sum_{\tau} \xi_{\tau} \langle \phi_{\tau}, D_{\tau}(\widehat{\theta}, \theta^{*})(\widehat{\theta} - \theta^{*})\| \leq \|\sum_{\tau} \xi_{\tau} \langle \phi_{\tau}, D_{\tau}(\theta, \theta')(\theta - \theta')\| + (2L\sqrt{d} + C_{1}) \text{buffer.e}(\widetilde{k})\epsilon,$$

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

Let's define the following shorthands

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

Therefore, by applying the inequality above, we have

$$V(\phi) \leq V(\widehat{\phi}) + 4C_1\sqrt{d}(2L\sqrt{d} + C_1) \text{buffer.e}(\widetilde{k})\epsilon. \tag{C.2}$$

The inequality holds because of the square difference formula.

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

$$\sqrt{V(\phi)} \le \sqrt{V(\widehat{\phi})} + \sqrt{4C_1\sqrt{d}(2L\sqrt{d} + C_1)} \text{buffer.e}(\widetilde{k})\epsilon. \tag{C.3}$$

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

$$\Pr[|\sum_{\tau} \xi_{\tau} D_{\tau} \langle \phi_{\tau}, \theta - \theta' \rangle| \ge \sqrt{\log \frac{2B_2 H N}{\delta \epsilon^{2d}} V(\phi)}] \le \frac{\delta}{H N}, \tag{C.4}$$

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

Therefore, it holds that

$$\begin{split} V(\widehat{\phi}) \leq & 4C_1\sqrt{d}(2L\sqrt{d} + C_1) \text{buffer.e}(\widetilde{k})\epsilon + V(\phi) \\ \leq & 4C_1\sqrt{d}(2L\sqrt{d} + C_1) \text{buffer.e}(\widetilde{k})\epsilon + 2\sqrt{\log\frac{2B_2HN}{\delta\epsilon^{2d}}}V(\phi) + 6L \\ \leq & 4C_1\sqrt{d}(2L\sqrt{d} + C_1) \text{buffer.e}(\widetilde{k})\epsilon + 2\sqrt{\log\frac{2B_2HN}{\delta\epsilon^{2d}}}\left[\sqrt{V(\widehat{\phi})} + \sqrt{4C_1\sqrt{d}(2L\sqrt{d} + C_1)} \text{buffer.e}(\widetilde{k})\epsilon\right] + 6L \\ = & 4C_1\sqrt{d}(2L\sqrt{d} + C_1) + 2\sqrt{\log\frac{2B_2HN}{\delta}}\left[\sqrt{V(\widehat{\phi})} + \sqrt{4C_1\sqrt{d}(2L\sqrt{d} + C_1)}\right] + 6L \\ \leq & 4C_1\sqrt{d}(2L\sqrt{d} + C_1) + 2\sqrt{\log\frac{2B_2HN}{\delta}} \left[\sqrt{V(\widehat{\phi})} + \sqrt{4C_1\sqrt{d}(2L\sqrt{d} + C_1)}\right] + 6C_4H\log^2K. \end{split}$$

The first inequality holds due to Ineq. (C.2) while the second one holds due to Ineq. (C.4) and Theorem C.3. The third inequality holds because of Ineq. (C.3). The equality holds since we set $\epsilon = \frac{1}{\mathsf{buffer.e}(\widetilde{k})}$. The final inequality holds because of Theorem A.3.

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

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

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

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

It holds since

$$\sqrt{(\widehat{\theta} - \theta^*)^T \Lambda^{\texttt{buffer.e}(\widetilde{k})}(\widehat{\theta} - \theta^*)} \leq \sqrt{(\widehat{\theta} - \theta^*)^T (\sum_{\tau} \phi_{\tau} \phi_{\tau}^T)(\widehat{\theta} - \theta^*)} + \sqrt{(\widehat{\theta} - \theta^*)^T (\lambda I)(\widehat{\theta} - \theta^*)}.$$

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

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

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

which ends the proof.

C.6 Proof of Theorem A.6

Using Cauchy inequality, we have the following statement:

Lemma C.4. It holds that

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

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

Then Theorem C.4 and Theorem A.5 lead to Theorem A.6.

C.7 Proof of Theorem A.7

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

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

$$\begin{split} |\widetilde{R}_h^k(\cdot,\cdot) - \widehat{R}_h^k(\cdot,\cdot)| &\leq \sum_{i=1}^N \mathbb{E}|[\max\{r_{ih}^{k-},\alpha_{ih}^k\} \, \mathbb{1}(r_{ih}^k \geq \max\{r_{ih}^{k-},\alpha_{ih}^k)]| \\ &- [\max\{\widehat{r}_{ih}^{k-},\alpha_{ih}^k\} \, \mathbb{1}(\widehat{r}_{ih}^k \geq \max\{\widehat{r}_{ih}^{k-},\alpha_{ih}^k)]| \\ &\leq \sum_{i=1}^N \Delta_1 + \Delta_2 + \Delta_3 \\ &\leq (1+6C_1)NC_5\sqrt{H} \log K \|\phi(\cdot,\cdot)\|_{(\Lambda_h^{\mathrm{buffer.e}(\bar{k})})^{-1}}, \end{split}$$

where

$$\Delta_{1} = |[\max\{r_{ih}^{k-}, \alpha_{ih}^{k}\} \mathbb{1}(r_{ih}^{k} \ge \max\{r_{ih}^{k-}, \alpha_{ih}^{k}\})] - [\max\{\hat{r}_{ih}^{k-}, \alpha_{ih}^{k}\} \mathbb{1}(r_{ih}^{k} \ge \max\{r_{ih}^{k-}, \alpha_{ih}^{k}\})]|,$$

$$\begin{split} \Delta_2 = & |[\max\{\widehat{r}_{ih}^{k-}, \alpha_{ih}^k\} \operatorname{1\hspace{-.1em}l}(r_{ih}^k \geq \max\{r_{ih}^{k-}, \alpha_{ih}^k\})] \\ & - [\max\{\widehat{r}_{ih}^{k-}, \alpha_{ih}^k\} \operatorname{1\hspace{-.1em}l}(\widehat{r}_{ih}^k \geq \max\{r_{ih}^{k-}, \alpha_{ih}^k\})]| \end{split}$$

and

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

The first inequality holds due to properties of convex functions. The second inequality holds due to 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 $|R_h^k(\cdot,\cdot) - \widetilde{R}_h^k(\cdot,\cdot)|$, 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 C.5 (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 Theorem 3.1, Theorem 3.2 and Theorem 3.3.

Lemma C.6 (Lemma C.4. (Golrezaei et al., 2019)). Under Theorem 3.3, it holds that

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

By applying Theorem C.6, we have

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

The first inequality holds due to Taylor expansion. The second inequality holds due to Theorem C.6, while the third one holds due to Theorem A.6. The last inequality holds since $\|\phi\|_{\Lambda^{-1}} \leq \frac{1}{\lambda}$.

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

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

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

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

and it ends the proof.

C.8 Proof of Theorem A.9

In order to prove Theorem A.9, we have the following lemmas for help.

Lemma C.7. 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

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

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 $||\mathcal{M}_h(\mathcal{S})|| \leq \sqrt{d}$, Theorem C.7 holds.

Lemma C.8. For any $(k,h) \in [K] \times [H]$, the vector $\omega_h^{\text{buffer.e}(\widetilde{k})}$ in Algorithm 4 satisfies:

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

Proof. Since we only update at episode buffer.e(\widetilde{k}), ω_h^k is the same as $\omega_h^{\text{buffer.e}(\widetilde{k})}$. For any vector $\nu \in \mathbb{R}^d$, we have

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

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

Lemma C.9 (Lemma D.1. (Jin et al., 2020)). Let $\Lambda^{\text{buffer.e}(\widetilde{k})} = \lambda I + \sum_{\tau=1}^{\text{buffer.e}(\widetilde{k})} \phi_{\tau} \phi_{\tau}^{T}$ where $\phi_{\tau} \in \mathbb{R}^{d}$ and $\lambda > 0$. Then it holds

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

Thus, with $\|\omega_h^{\text{buffer.e}(\widetilde{k})}\| = \max_{\nu:\|\nu\|=1} |\nu^T \omega_h^{\text{buffer.e}(\widetilde{k})}|$, it ends the proof. In order to prove the next lemma, we introduce two useful lemmas at first.

Lemma C.10. 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 $\Lambda_{\text{buffer.e}(\widetilde{k})} = \lambda I + \sum_{\tau=1}^{\text{buffer.e}(\widetilde{k})} \phi_{\tau}$, then for any δ , with probability at least $1 - \delta$, for any k corresponding to buffer.e(\widetilde{k}) and any $V \in \mathcal{V}$ so that $\sup_{x} |V(x)| \leq 3H$, we have

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

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

Proof. First of all, we have

$$\begin{split} &\| \sum_{\tau=1}^k \phi_{\tau} \{ V(x_{\tau}) - \mathbb{E}[V(x_{\tau}) \, | \, \mathcal{F}_{\tau-1}] \} \|_{\Lambda_{\text{buffer.e}(\tilde{k})}^{-1}}^2 \\ \leq & 2 \times 2 \| \sum_{\tau=1}^k \phi_{\tau} \{ V(x_{\tau}) - \mathbb{E}[V(x_{\tau}) \, | \, \mathcal{F}_{\tau-1}] \} \, \mathbb{1} \{ k \not\in \text{buffer} \} \|_{\Lambda_{k}^{-1}}^2 + 2 \times 3H \frac{1}{\lambda} 3H \frac{3HC_2 \log^2 K}{\log \frac{1}{\gamma}} \\ \leq & 4 \| \sum_{\tau=1}^k \phi_{\tau} \{ V(x_{\tau}) - \mathbb{E}[V(x_{\tau}) \, | \, \mathcal{F}_{\tau-1}] \} \|_{\Lambda_{k}^{-1}}^2 + \frac{54C_2 H^3 \log^2 K}{\lambda \log \frac{1}{\gamma}}. \end{split}$$

Firstly, we have $(a+b)^2 \le 2a^2 + 2b^2$. Then, it holds since we divide the episodes into two parts, the ones in buffer and the ones not. For the ones in buffer, due to the definition of buffer.e(\widetilde{k}), it is easy to prove that

it is smaller than $4\|\sum_{\tau=1}^k \phi_{\tau}\{V(x_{\tau}) - \mathbb{E}[V(x_{\tau}) \mid \mathcal{F}_{\tau-1}]\} \, \mathbb{1}\{k \notin \text{buffer}\}\|_{\Lambda_k^{-1}}^2$. As for the one not in buffer, $\frac{54C_2H^3\log^2K}{\lambda\log\frac{1}{\gamma}}$ is a trivial bound due to Theorem A.1 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.

Lemma C.11. Let \mathcal{V} denote a class of functions mapping from \mathcal{S} to \mathbb{R} with the following parametric form

$$V(\cdot) = \min\{\max_{\sigma} \omega^T \phi(\cdot, \upsilon) + \widehat{R}(\cdot, \upsilon) + \beta \|\phi(\cdot, \upsilon)\|_{\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 $\operatorname{dist}(V, V') = \sup_{x} |V(x) - V'(x)|$. Then, it holds

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

where B_5 is a constant.

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

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

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 $\widehat{R}(x,v)$, we show that it links to $\{\widehat{\theta}_i\}_{i=1}^N$ first. As $\widehat{R}(\cdot,\cdot)$ is function of $\{\widehat{\mu}_i\}_{i=1}^N$ and $F(\cdot)$ is differentiable with $|f| \leq C_1$, it holds that $\frac{\partial \widehat{R}}{\partial \mu_i} \leq B_5$ 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 $\widehat{\mu} = \langle \phi, \widehat{\theta} \rangle$, it holds that

$$\sup_{x,v} |\widehat{R}_{1}(x,v) - \widehat{R}_{2}(x,v)| \leq \sup_{\phi: \|\phi\| \leq 1} \sum_{i=1}^{N} B_{5} |(\widehat{\theta}_{1i} - \widehat{\theta}_{2i})^{T} \phi|$$
$$\leq \sum_{i=1}^{N} B_{5} \|\widehat{\theta}_{1i} - \widehat{\theta}_{2i}\|.$$

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

$$|C_{\widehat{R}}| \le (1 + \frac{8NB_5\sqrt{d}}{\epsilon})^{dN}.$$

Then, it finishes the proof.

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

Lemma C.12. For any δ , with probability at least $1 - \delta$, there exists constants B_6 and B_7 independent of K and H so that

$$\forall (k,h) \in [K] \times [H]: \| \sum_{\tau=1}^{k} \phi_{h}^{\tau} [\widehat{V}_{h+1}^{k}(x_{h+1}^{\tau}) - \mathbb{P} \widehat{V}_{h+1}^{k}(x_{h}^{\tau}, \upsilon_{h}^{\tau})] \|_{(\Lambda_{h}^{\text{buffer.e}(\widetilde{k})})^{-1}}^{2} \leq B_{6}H^{3} \log^{2} K + B_{7}H^{2} \log C_{7}.$$

Proof. Combining Theorem C.8, Theorem C.10 and Theorem C.11, we set $L = 3H\sqrt{\frac{dk}{\lambda}}$. With Algorithm 4, we have $B = C_7 + C_6H\log^2 K$. Then we have

$$\begin{split} &\| \sum_{\tau=1}^k \phi_h^{\tau} [\widehat{V}_{h+1}^k(x_{h+1}^{\tau}) - \mathbb{P} \widehat{V}_{h+1}^k(x_h^{\tau}, \upsilon_h^{\tau})] \|_{(\Lambda_h^{\text{buffer.e}(\widehat{k})})^{-1}}^2 \\ \leq & \frac{54C_2H^3 \log^2 K}{\lambda \log \frac{1}{\gamma}} + 72dH^2 \log \frac{k+\lambda}{\lambda} + 144H^2 d \log (1 + \frac{24H}{\epsilon} \sqrt{\frac{dk}{\lambda}}) + 144H^2 \log \frac{KH}{\delta} \\ & + 144H^2 d^2 \log [1 + \frac{32\sqrt{d}(C_7 + C_6H \log^2 K)^2}{\lambda \epsilon^2}] + 144H^2 dN \log (1 + \frac{8NB_5\sqrt{d}}{\epsilon}) + \frac{32k^2\epsilon^2}{\lambda}. \end{split}$$

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.

Now, let's show the determination of C_7 .

Lemma C.13. 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 \mathscr{E} , i.e., all inequalities hold, we have for all $(x, v, h, k) \in \mathcal{S} \times \Upsilon \times [H] \times [K]$ that:

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

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

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

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

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

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

where

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

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

Firstly, it holds

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

The first inequality holds due to Cauchy inequality and $\Lambda_{\mathtt{buffer.e}(\widetilde{k})} \succeq \lambda I$. The second inequality holds due to Theorem C.7.

Secondly, it holds that

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

It holds because of Theorem C.12.

Lastly, we have

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

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

With the bounds in hand, we have $\Delta_k^h(\cdot,\cdot) \leq (3H\sqrt{d\lambda} + \sqrt{B_6H^3\log^2K + B_7H^2\log C_7} + 3H\sqrt{d\lambda} + C_6H\log^2K)\|\phi(\cdot,\cdot)\|_{(\Lambda_h^{\text{buffer.e}(\tilde{k})})^{-1}}$. Then, it is obviously that there exists a constant B_8 , so that $B_8H^{\frac{3}{2}}\log K \geq 3H\sqrt{d\lambda} + \sqrt{B_6H^3\log^2K + B_7H^2\log C_7} + 3H\sqrt{d\lambda}$ and it finishes the proof.

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

Lemma C.14. Under the setting of Theorem 3.4, on the Good Event \mathscr{E} , it holds that for any $(x, v, h, k) \in \mathcal{S} \times \Upsilon \times [H] \times [K]$,

$$\widehat{Q}_h^k(x,\upsilon) \le Q_h^{\pi^*}(x,\upsilon).$$

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

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

due to Theorem C.13. Therefore, we have

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

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 Theorem C.13, it holds that

$$|[\widehat{R}_{h}^{k} + \langle \phi, \omega_{h}^{k} \rangle - Q_{h}^{\pi^{*}} - \mathbb{P}_{h}(V_{h+1}^{k} - V_{h+1}^{\pi^{*}})](\cdot, \cdot)| \leq (C_{7} + C_{6}H \log^{2}K) \|\phi(\cdot, \cdot)\|_{(\Lambda_{h}^{\text{buffer.e}(\widetilde{k})})^{-1}}.$$

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

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

which ends the proof.

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

Lemma C.15. Let $\delta_h^k = V_h^k(x_h^k) - V_h^{\pi_{\widetilde{k}}}(x_h^k)$ and $\xi_{h+1}^k = \mathbb{E}[\delta_{h+1}^k \mid x_h^k, v_h^k] - \delta_{h+1}^k$. Then conditional on Good Event \mathscr{E} , it holds that for any $(k,h) \in [K] \times [H]$,

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

Proof. Due to Theorem C.13, it holds that

$$\widehat{Q}_h^k(\cdot,\cdot) - Q_h^{\pi_{\widetilde{k}}}(\cdot,\cdot) \leq \mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi_{\widetilde{k}}})(\cdot,\cdot) + 2(C_7 + C_6H\log^2K)\|\phi(\cdot,\cdot)\|_{(\Lambda_{\iota}^{\mathrm{buffer.e}(\widetilde{k})})^{-1}}.$$

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

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

Then, it ends the proof.

With these preparations, we begin to prove Theorem A.9.

Using notations in Theorem C.15, it holds that conditional on Good Event \mathscr{E}

$$\begin{split} &\Delta_{1} = \sum_{\tau=1}^{K} [V_{1}^{\pi^{*}}(x_{1}^{k}) - V_{1}^{\pi_{\widetilde{k}}}(x_{1}^{k})] \, \mathbb{1}(k \not\in \text{buffer}) \\ &\leq \sum_{\tau=1}^{K} \delta_{1}^{k} \, \mathbb{1}(k \not\in \text{buffer}) \\ &\leq \sum_{\tau=1}^{K} \sum_{h=1}^{H} \xi_{h}^{k} + 2(C_{7} + C_{6}H \log^{2}K) \|\phi(\cdot, \cdot)\|_{(\Lambda_{h}^{\text{buffer.e}(\widetilde{k})})^{-1}} \, \mathbb{1}(k \not\in \text{buffer}) \\ &\leq \sum_{\tau=1}^{K} \sum_{h=1}^{H} \xi_{h}^{k} + 2\sqrt{2}(C_{7} + C_{6}H \log^{2}K) \|\phi(\cdot, \cdot)\|_{(\Lambda_{h}^{k})^{-1}} \, \mathbb{1}(k \not\in \text{buffer}) \\ &\leq \sum_{\tau=1}^{K} \sum_{h=1}^{H} \xi_{h}^{k} + 2\sqrt{2}(C_{7} + C_{6}H \log^{2}K) \|\phi(\cdot, \cdot)\|_{(\Lambda_{h}^{k})^{-1}}. \end{split}$$

The first inequality holds due to Theorem C.14, while the second one holds due to Theorem C.15. 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

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

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

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

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

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

Then with Cauchy inequality, we have

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

Finally, combining the two terms and we have

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

and it finishes our proof.

D Auxiliary Lemmas and Proofs in Appendix B

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

D.1 Proof of Theorem B.2

In Algorithm 6, there are two types of {buffer.e(\widetilde{k})}. The number of {buffer.e(\widetilde{k})} satisfying $2(\Lambda_h^k)^{-1} \not\succ$ $(\Lambda_h^{\mathtt{buffer.e}(\widetilde{k})})^{-1}$ is smaller than $\frac{3C_2H\log^2K}{\log\frac{1}{\gamma}}$ due to Theorem A.1. The number of $\{\mathtt{buffer.e}(\widetilde{k})\}$ when \log_2k is an integer is smaller than $[\log_2 K] + 1$. Combining the two parts finishes the proof П

D.2Proof of Theorem B.3

Since we have buffer period, the bound of the size of overbid or underbid is as same as the situation when market noise distribution is known. Then, recall that the proof of Theorem A.3 is conditional on reserve price and others' bid, it doesn't matter whether we consider q or \widetilde{q} because the only difference between them is the way generating reserve has been π_0 . Conditional on reserve, the proof of Theorem A.3 still holds on regarding to \widetilde{q} .

With the same methodology in Theorem A.3, we have the lemma due to Theorem B.2.

D.3Proof of Theorem B.4

Similar to the proof of Theorem A.5, 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 (4.1). We just need to prove that $\mathbb{E}[\widetilde{q} - \frac{1}{3N}(1 + \langle \phi_{\tau}, \theta \rangle)] = 0$ if bidders bid truthfully. If $\widetilde{q}_{ih}^{\tau} = 1$, 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 Theorem A.5, we prove Theorem B.4.

D.4 Proof of Theorem B.6

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

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

However, bidders may overbid or underbid for less than $\frac{C_3H}{K}$ due to Theorem A.2 and the estimation of μ has error. Therefore, the c.d.f that $\widehat{F}(\cdot)$ estimates is not the same as $F(\cdot)$. Since $|f(\cdot)| \leq C_1$, the difference because of overbid or underbid is smaller than $\frac{C_1C_3H}{K}$. Then, due to Theorem B.5, the difference because of error in μ is smaller than

$$C_1C_{11}\sqrt{H}\log K\frac{\sum_{h=1}^{H}\sum_{\tau=1}^{\mathrm{buffer.e}(\widetilde{k})}\|\phi(x_h^{\tau},\upsilon_h^{\tau})\|_{(\Lambda_h^{\mathrm{buffer.e}(\widetilde{k})})^{-1}}}{H\mathrm{buffer.e}(\widetilde{k})}\leq C_1C_{11}\sqrt{H}\log K\frac{\sqrt{d}}{\sqrt{\mathrm{buffer.e}(\widetilde{k})}}.$$

The inequality holds since we have the mean value inequality and Theorem C.9.

Since the number of episodes in buffer for each buyer i is no larger than $C_9H\log^2 K$, it holds that

$$\begin{split} |F(\cdot) - \widehat{F}(\cdot)| \leq & \sqrt{\frac{1}{2}\log\frac{2K}{\delta}} (NH \texttt{buffer.e}(\widetilde{k}))^{-\frac{1}{2}} + \frac{C_1C_3H}{K} + \frac{C_9H\log^2K}{\texttt{buffer.e}(\widetilde{k})} \\ & + C_1C_{11}\sqrt{H}\log K \frac{\sqrt{d}}{\sqrt{\texttt{buffer.e}(\widetilde{k})}}. \end{split}$$

Because the number of episode we run Equation (4.1) is smaller than K, then the total probability happening Bad Event \mathscr{E}^c is smaller than δ . Then, it ends the proof.

Proof of Theorem B.7 D.5

In order to prove Theorem B.7, we introduce the following lemma first.

Lemma D.1. Under assumption Theorem 3.2, when Theorem B.6 holds, using histogram method to estimate p.d.f $f(\cdot)$ leads to the following bound that for any x

$$|f(x) - \widehat{f}(x)| \le D_1 \frac{\sqrt{H} \log K}{\operatorname{buffer.e}(\widetilde{k})^{\frac{1}{4}}},$$

where D_1 is a constant.

Proof of Theorem D.1

With Theorem B.6 in hand, we divide [-1,1] into 2M parts denoted by $\{-M,\ldots,0,\ldots,M-1\}$ uniformly, then we have

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

where $x \in (\frac{i}{M}, \frac{i+1}{M}]$. Under assumption Theorem 3.2, it holds that

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

Therefore, it holds that

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

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

Therefore, unlike Theorem C.6, we have the following lemma.

Lemma D.2. Under Theorem 3.3, it holds that

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

where D_2 is a constant.

D.5.2 Proof of Theorem D.2

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 optimal reserve price while $\hat{\alpha}$ to denote the reserve price we use with $\hat{F}(\cdot)$ and $\hat{f}(\cdot)$. Also, we use $\widetilde{\alpha}$ to denote reserve price corresponding to $\widehat{\mu}$, $F(\cdot)$ and $f(\cdot)$.

Theorem C.6 shows that $|\widetilde{\alpha} - \alpha^*| \leq |\langle \phi_h^k, \theta_{ih} - \theta_{ih} \rangle|$

To bound $|\widetilde{\alpha} - \widehat{\alpha}|$, we have

$$\begin{split} |\frac{1-F(\cdot)}{f(\cdot)} - \frac{1-\widehat{F}(\cdot)}{\widehat{f}(\cdot)}| &\leq |\frac{1-F(\cdot)}{f(\cdot)} - \frac{1-\widehat{F}(\cdot)}{f(\cdot)}| + |\frac{1-\widehat{F}(\cdot)}{f(\cdot)} - \frac{1-\widehat{F}(\cdot)}{\widehat{f}(\cdot)}| \\ &\leq \frac{C_{12}H\log^2K}{c_1\sqrt{\mathtt{buffer.e}(\widetilde{k})}} + \frac{D_1\sqrt{H}\log K}{c_1^2\mathtt{buffer.e}(\widetilde{k})^{\frac{1}{4}}}. \end{split}$$

The first inequality holds due to triangle inequality. The second inequality holds due to Theorem 3.1, Theorem B.6 and Theorem D.1.

Theorem B.6 and Theorem D.1. Then, we will show that $\phi'(\cdot) \leq 1$. It holds that $\phi(x) = x - \frac{1 - F(x)}{f(x)} = x + \frac{1}{\log'(1 - F(x))}$. Under Theorem 3.3, it holds that $1 - F(\cdot)$ is log-concave implying $\log'(1 - F(\cdot))$ is decreasing. Therefore, $\phi'(x) \geq 1$. Therefore, we have $|\phi(\widehat{\alpha}) - \widehat{\phi}(\widehat{\alpha})| \leq \frac{C_{12}H\log^2K}{c_1\sqrt{\text{buffer.e}(\widetilde{k})}} + \frac{D_1\sqrt{H}\log K}{c_1^2\text{buffer.e}(\widetilde{k})^{\frac{1}{4}}}$ and $\phi(\widehat{\alpha}) = \widehat{\phi}(\widehat{\alpha})$. Then, it holds that

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

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

Then, it ends our proof.

Now, we are ready to prove Theorem B.7. Using notations in Theorem A.7, we use another factor F to show that we use $F(\cdot)$ and $f(\cdot)$ in the function while factor F to denote the use of $F(\cdot)$ and $f(\cdot)$.

With the same methodology in Theorem A.7, it holds that

$$\begin{split} |R_h^k(\cdot,\cdot,F) - \widehat{R}_h^k(\cdot,\cdot,F)| \leq & [(1+6C_1)C_{11}\sqrt{H}\log K]N\|\phi(\cdot,\cdot)\|_{(\Lambda_h^{\mathrm{buffer.e}(\widetilde{k})})^{-1}} \\ & + \frac{NB_4}{2}[2(|\langle\phi_h^k,\theta_{ih}-\widehat{\theta}_{ih}\rangle|)^2 + 2(\frac{D_2H\log^2K}{\mathrm{buffer.e}(\widetilde{k})}^{\frac{1}{4}})^2] \\ \leq & D_3H\log^2K\|\phi(\cdot,\cdot)\|_{(\Lambda_h^{\mathrm{buffer.e}(\widetilde{k})})^{-1}} + D_4H^2\log^4K\frac{1}{\sqrt{\mathrm{buffer.e}(\widetilde{k})}}, \end{split}$$

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

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

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

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

Combining the two parts, it holds

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

which ends the proof.

D.6 Proof of Theorem B.8

Now, we introduce some lemmas in parallel in order to prove Theorem B.8.

Lemma D.3. 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 $\Lambda_{\text{buffer.e}(\widetilde{k})} = \lambda I + \sum_{\tau=1}^{\text{buffer.e}(\widetilde{k})} \phi_{\tau}$, then for any δ , with probability at least $1-\delta$, for any k corresponding to buffer.e (\widetilde{k}) and any $V \in \mathcal{V}$ so that $\sup_{x} |V(x)| \leq 3H$, we have

$$\| \sum_{\tau=1}^{k} \phi_{\tau} \{ V(x_{\tau}) - \mathbb{E}[V(x_{\tau}) \mid \mathcal{F}_{\tau-1}] \} \|_{\Lambda_{\text{buffer.e}(\tilde{k})}^{-1}}^{2} \leq \frac{54C_{9}H^{3} \log^{2} K}{\lambda \log \frac{1}{\gamma}} + \frac{32k^{2}\epsilon^{2}}{\lambda} + \frac{32k^{2}\epsilon^{2}}{\lambda} + \frac{144H^{2}[\frac{d}{2} \log \frac{k+\lambda}{\lambda} + \log \frac{\mathcal{N}_{\epsilon}}{\delta}],$$

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

Lemma D.4. Let \mathcal{V} denote a class of functions mapping from \mathcal{S} to \mathbb{R} with the following parametric form

$$V(\cdot) = \min\{\max_{a} \omega^{T} \phi(\cdot, \upsilon) + \widehat{R}(\cdot, \upsilon) + \beta \|\phi(\cdot, \upsilon)\|_{\Lambda^{-1}} + A, 3H\},$$

where $\|\omega\| \leq L$, $\beta \in [0, B]$, $A = \frac{C_{14}H^2 \log^4 K}{\sqrt{\text{buffer.e}(\widetilde{k})}}$ 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

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

where B_5 is a constant.

D.6.1 Proof of Theorem D.4

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

$$\sup_{x,v} |\widehat{R}_{1}(x,v) - \widehat{R}_{2}(x,v)| = \sup_{x,v} |\widehat{R}_{1}(x,v,\widehat{F}_{1}) - \widehat{R}_{2}(x,v,\widehat{F}_{2})|
\leq \sup_{x,v} |\widehat{R}_{1}(x,v,\widehat{F}_{1}) - \widehat{R}_{2}(x,v,\widehat{F}_{1})|
+ \sup_{x,v} |\widehat{R}_{2}(x,v,\widehat{F}_{1}) - \widehat{R}_{2}(x,v,\widehat{F}_{2})|.$$

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

Like the proof of Theorem C.11, we simply use $\frac{\epsilon}{4NB_5}$ -ball to cover $\hat{\theta}_i$, and it holds that

$$|C_{\widehat{\theta}}| \le (1 + \frac{16NB_5\sqrt{d}}{\epsilon})^{dN}.$$

Conditional on ω , A and $\{\widehat{\theta}_i\}_{i=1}^N$, with Theorem B.7, we know that in order to satisfy $\sup_{x,v} |\widehat{R}(x,v,\widehat{F}) - \widehat{R}(x,v,F)| \le \frac{\epsilon}{4}$, what we need is $\|\widehat{F} - F\|_{\infty} \le \frac{\epsilon}{12N^2}$. Then, it ends the proof.

Then, it holds the following lemma.

Lemma D.5. For any δ , with probability at least $1 - \delta$, there exists constants B_6 and B_7 independent of K and H so that

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

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

Proof. Similar to the proof of Theorem C.12, 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 $\lambda=1$, $B=C_{15}+C_{13}H\log^2K$ and $\epsilon=\frac{dH}{k}$. With Theorem 4.1, we finishes our proof.

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

Lemma D.6. There exist $D_7 \sim \widetilde{\mathcal{O}}(1)$ so that $C_{15} = D_7 H^{\frac{3}{2}}$, and for any fixed policy π , on Good Event \mathscr{E} , i.e., all inequalities hold, we have for all $(x, v, h, k) \in \mathcal{S} \times \Upsilon \times [H] \times [K]$ that:

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

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

Proof. The proof of Theorem D.6 is the same as proof of Theorem C.13. Let's show the determination of D_7 in parallel. With Theorem D.5 in hand, it holds that

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

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

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 Theorem D.6 and Theorem C.14.

Lemma D.7. Let $\delta_h^k = V_h^k(x_h^k) - V_h^{\pi_{\tilde{k}}}(x_h^k)$ and $\xi_{h+1}^k = \mathbb{E}[\delta_{h+1}^k \mid x_h^k, v_h^k] - \delta_{h+1}^k$. Then conditional on Good Event \mathscr{E} , it holds that for any $(k,h) \in [K] \times [H]$,

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

Now, we are ready to prove Theorem B.8. Similar to the proof of Theorem A.9, it holds that

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

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

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

Therefore, it holds that

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

which ends the proof.