THE UNIVERSITY OF CHICAGO

OPTIMAL MECHANISM DESIGN IN SEQUENTIAL DECISION MAKING PROCESSES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
BOXIANG LYU

CHICAGO, ILLINOIS
MARCH 2024

To those who have helped me. None of this would be possible without you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Similar to how Princess Carolyn dreams about her descendant in the episode "Ruthie" on *Bojack Horseman*, I have been dreaming about completing the thesis to cope with stress for a long time. It seems almost surreal that I am actually completing this thesis.

I must thank my wonderful advisor Mladen Kolar, who gave me opportunity to come to Chicago, and offered me the freedom to pursue every project, seek out every collaboration that interests me. Without his gentle guidance, there is no chance for me to become the researcher that I am today. I sincerely thank Sanmi Koyejo, who believed in my ability when I had little to show for it, offering constant encouragement. I thank Haifeng Xu and Rad Niazadeh for taking the time to serve on my dissertation committee.

I had the amazing opportunity to work with many brilliant researchers both in academia and in industry during my PhD study. When I applied for PhD, I dreamed of working with some of the smartest, most hardworking people in the world. Thank you all for making this dream come true. I am grateful for Zhuoran Yang and Zhaoran Wang, who introduced me to theoretical reinforcement learning and its applications in dynamic mechanism design; Mike Mozer, Phil Long, and Zhe Feng for an amazing summer at Google Research; Xiang Zhou and Kyle Xu for the supportive environment at Point72. I would also like to thank my collaborators Shuang Qiu, Boxin Zhao, Dake Zhang, Pedro Cisneros-Velarde, Zachary Robertson, Rui Ai, among others.

I deeply thank all my friends and family who have supported me. To you I dedicate this thesis, and without you it would be impossible for me to be here right now. Huge thanks to Zixu Lu, Maoyuan Song, Ziniu Wu, Yiqin Xu, Bo Gao, Rujia Yang, Qiang Chen, Hao Qiu, Han Zheng, Tinglei Zhang, and many others. I apologize for my terrible memory if you do not see your name on this list. My deepest gratitude goes to my parents Ling Li and Yongning Lv (equal contribution, alphabetical order) for their unconditional love and support, and my fiancée Kaili Shan for shining a light on my life.

ABSTRACT

This dissertation focuses on efficiently learning the optimal dynamic mechanism when the agents' valuations can be characterized by a Markov Decision Process (MDP). In Chapter 1, we provide a high-level motivation for our problem setting, discussing the motivations for combining theoretical Reinforcement Learning (RL) with dynamic mechanism design. In subsequent chapters, we provide three representative problems at the intersection of these two fields. These chapters vary in terms of the difficulty of designing the optimal mechanism, the generality of function approximation, and the RL setup considered, providing a high-level overview of recent advances along this interdisciplinary research direction.

In particular, in Chapter 2, we show how the revenue maximizing, incentive compatible, and ex-post individually rational mechanism can be learned computationally and sample efficiently, when the single buyer's type distribution is governed by a tabular MDP. In Chapter 3, we provide an online learning algorithm that learns the optimal second-price auction with reserve prices with $\widetilde{\mathcal{O}}(\sqrt{T})$ regret, even if the participating buyers behave strategically. In Chapter 4, we show that the welfare-maximizing, ex-ante incentive compatible, and ex-ante individually rational mechanism can be obtained under general function approximation setting using only a pre-collected data set, with no additional interactions with the environment.

CHAPTER 1

A GENERAL INTRODUCTION

Sequential decision making processes, in particular Markov Decision Processes (MDPs), have long attracted researchers for their ability to capture time-based and state-based dynamics in human decision making [Puterman, 1990, White, 1993]. In recent years, development of computational and theoretical methods leads to an unprecedented amount of interest in the topic, as we can now efficiently learn MDPs both in theory and in practice [Auer et al., 2008, Kakade, 2001, Li, 2017, Sewak, 2019]. These new tools gave rise to the vibrant field of Reinforcement Learning (RL), which has seen great empirical success in games (the colloquial sense) such as Go [Chen, 2016a, Silver et al., 2017, 2016], Atari games [Mnih et al., 2013], and StarCraft II [Vinyals et al., 2019]. RL has also been applied in high risk real-world control problems, ranging from self-driving cars [Liang et al., 2018, Spielberg et al., 2019] to even nuclear fusion [Degrave et al., 2022]. Even ChatGPT, arguably the most popular machine learning project in early 2023, uses RL under the hood, as it incorporates human feedback in training via reinforcement learning with human feedback [Choi et al., 2023, Guo et al., 2023].

It is then no wonder that MDP and RL have also been used in real-world economics problems such as tax policies [Zheng et al., 2020, 2021] and business applications such as bargaining on eBay [Green and Plunkett, 2022]. Indeed, the state-based and time-based structure that MDP captures better reflect the sequential nature of human behavior in real life, where earlier actions and earlier history affect agents' later decision making and shaping the agents earlier history would then undoubtedly affect their later behavior.

Of course, computer scientists are far, far from the only people who have made this basic observation about human decision making. In economics and operations research, the field of dynamic mechanism design studies allocating and pricing goods when agents' types evolve over time [Bergemann and Välimäki, 2019, Gallien, 2006, Doepke and Townsend,

2006, Kakade et al., 2013], with widespread applications ranging from sponsored search auctions [Mirrokni et al., 2018, Shen et al., 2020] to pricing Wi-Fi at Starbucks [Parkes and Singh, 2003, Friedman and Parkes, 2003].

Applications of dynamic mechanism design often depend heavily on the assumption that the transition dynamic of the agents' types is known beforehand, yet the assumption can be unrealistic in practice. The problem is further complicated by the nature of MDP, where transition probabilities depend on the action taken. While observational studies can be used to efficiently estimate type distributions in the stationary setting or the Markov chain setting, a reasonable justification for assuming the type distributions are known a priori, for MDPs simply observing the outcomes of arbitrarily chosen actions is insufficient, a well-known result that highlights the difficulty of RL compared to the stationary setting [Osband and Van Roy, 2016].

It is then natural to bring together dynamic mechanism design and RL by using techniques from the latter to learn the optimal dynamic mechanisms for MDPs. To our dismay, existing results are limited, with existing works that are either lacking in theory or not fully indicative the nuances of dynamic mechanism design.

Due to the limits of existing works, the dissertation focuses on answering the following central question

Can we provide an efficient algorithm for learning the "optimal" dynamic mechanism in the single agent setting, where the agent's type evolve according to an MDP?

The remaining chapters are devoted to three different variants of the problem, organized as follows. More specifically, in Chapter 2, we define optimal as the revenue-maximizing, dynamic incentive compatible (IC), and ex-post individually rational (IR) mechanism. We focus on tabular MDPs, where action is defined as the allocation the buyer receives, state is some feature variable dubbed public context, and the buyer's private types is drawn from a distribution parameterized by the public context. The learning setup is PAC RL, with

the goal of outputting a near-optimal mechanism using only polynomially many interactions with the environment, with high probability. We emphasize that the term PAC RL is a bit of a misnomer: in PAC learning, the samples are i.i.d. generated from some distribution, whereas typically PAC RL allows the learner to use different policies to generate sample trajectories from the environment (see Dann et al. [2018], for instance). Nevertheless, we use the term to differentiate it from online RL, discussed in the sequel. The chapter is adapted from a currently unpublished manuscript and the result of a collaboration with Haifeng Xu and Song Zuo.

In Chapter 3, we define optimal as the revenue-maximizing multi-phase second price auction with reserve prices. We focus on linear MDPs, where action is decomposed into two components, with the first being the reserve price, and the second being the type of the item being sold at that particular step. State is again some feature variable that parameterizes the buyers' private types. The learning setup is online RL, where we aim to minimize regret throughout the interactions with strategic agents. The chapter is adapted from Ai et al. [2022], a collaboration with Rui Ai, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan.

In Chapter 4, we define optimal as the welfare maximizing, incentive compatible, and ex-ante IR mechanism. We focus on a general function approximation setting, where we only assume that the relevant value functions can be approximated sufficiently well by some function class. Particularly, we assume that the participating agents' types are fixed, but their valuation functions depend on some arbitrary state and action that evolve according to an MDP. The setup considered is offline RL, where we only assume the existence of an pre-collected dataset, and aims to recover a near-optimal mechanism with no further interactions with the environment. The chapter is adapted from Lyu et al. [2022b], a collaboration with Zhaoran Wang, Mladen Kolar, and Zhuoran Yang.

We hope that the three chapters can highlight the number of possible open questions in the intersection between RL and dynamic mechanism design. Due to the changing notions

of optimality, function approximation assumption, and learning setup, it is hard to provide a concrete mathematical overview of these problems, encompassing all three chapters. Nevertheless, we highlight three important threads connecting these chapters, and hope that the discussion demonstrates the inherent cohesiveness of the chapters.

Complexity of mechanism design. From front to back, the optimal mechanism becomes easier and easier to characterize. The most complicated dynamic mechanism is constructed in Chapter 2, requiring a novel revelation-style argument for describing the optimal mechanism. In Chapter 3, mechanism design is simplified by the fact that we only focus on the optimal multi-phase second price auction with reserve prices, which is IC and IR by well-known results, with the key challenge being simultaneously learning (1) the best policy for choosing which item to sell at each step, conditioned on the state, and (2) the optimal reserve price for each state and item sold, under the condition that the buyers are strategic and may bid untruthfully. The dynamic VCG mechanism being learned in Chapter 4 is the easiest to characterize, depending only on the welfare-maximizing policy and, for each agent, a price to be paid.

Generality of function approximation. From front to back, we allow the underlying MDP to be more general. In Chapter 2, we only focus on tabular MDPs, requiring that both the state and action spaces are finite. In Chapter 3, the assumption is weakened, as we assume a linear MDP, which loosely translates to the assumption that the transition dynamics and the valuation distributions are linear in some feature vector for the state and action. In Chapter 4, the assumption is further weakened, and we only assume the existence of some arbitrary function class that can sufficiently accurately describe the value functions.

Difficulty of exploration exploitation trade-off. From front to back, we gradually reduce the emphasis on exploration and focus more on exploitation. The PAC-RL setting

in Chapter 2 does not consider exploitation at all, as the only goal is to output a near-optimal mechanism at the end of algorithm. We do note it is nearly trivial to obtain an online RL algorithm by adapting the PAC RL one we obtained using the explore-then-commit framework discussed in [Lyu et al., 2022a], albeit with a looser regret bound. For the online RL setting in Chapter 3, we focus on obtaining the tightest regret bound possible, necessitating various techniques employed to better balance exploration and exploitation, discussed in the sequel. Finally, for Chapter 4, we forgo exploration altogether and instead focus on how to best exploit a pre-collected dataset, with no additional interactions with the environment.

We hope that our discussion can provide general insights for future researchers working on the intersection of RL and mechanism design. By "tuning" these three factors contributing to the overall hardness of the problem, researchers can either make existing results more broadly applicable and thereby interesting, or develop novel results under a simplified setting for a seemingly impossible problem. With the discussion in mind, we proceed with the rest of the dissertation.

CHAPTER 2

PAC RL FOR REVENUE MAXIMIZING DYNAMIC MECHANISM IN TABULAR MDPS

2.1 Introduction

How should a monopolistic seller (she) sell an item to a buyer (he), when the buyer's later demands are affected by the earlier allocations he receives? For example, consider an airline trying to price tickets to popular travel destinations over the course of a year. If the airline set the prices too low early in the year, passengers may splurge more initially but will be less likely to travel to expensive destinations towards the end of the year: their demand for travel has been fulfilled early on. As an another example, consider a new restaurant pricing a prix-fixe menu, i.e. one where a multi-course meal is offered at a pre-determined price, for the year when the restaurant opens. It may be more favorable for the restaurateur to lower the prices first to promote its brand, and then raise these prices in order to increase total revenue.[1]

A blessing of modern technology is the subjective data that can be used to gauge consumer demand. Instead of having to survey its customers on how likely they are to travel, the airline may use data from search engines to accurately estimate consumer demand level. The restaurateur may use the number of social media followers or its ratings on review platforms to gauge valuation distribution. The existence of such data ensures that, while the buyer's later valuation distribution are affected by earlier allocations, it is possible to accurately estimate these distributions using publicly available information. An approach these sellers could take is to first understand how their pricing strategy alters the evolution of these observable public data, which we call *public contexts*, and then estimate the valuation distribution conditioned on these contexts.

1. This can be done via discount, rather than truly changing the price values on the menu.

Nevertheless, the key challenge behind both settings is the fact that later public context and valuation distributions can both depend on earlier allocations. It is not always clear whether lowering prices earlier on increases or decreases later demand. Unlike seasonal trends in demands that can be estimated accurately by simply recording the valuation distributions in each time period, it is imperative for the seller to also explore different mechanisms. Should she set the prices moderately high initially to ensure that there will be sufficient demand later on? Or should she lower the prices to drum up demand? These questions cannot be answered by using only observational data and often require actual experimentation with different strategies, naturally leading to *reinforcement learning*.

The challenges of learning optimal mechanisms naturally arise in large-scale Internet applications as well. For instance, consider the *spot instance* [2] pricing problem faced by Amazon Web Services (AWS) [Agmon Ben-Yehuda et al., 2013], where dynamic pricing has been used. Holding substituting services' prices fixed, we may view AWS as a monopolistic seller of cloud computing instances to a group of buyers. The company can gauge the ground-truth valuation distribution by examining the current demand level for spot instances. Increasing allocation by lowering prices may attract more customers, but those attracted may be less willing-to-pay, as it is possible for them to be only interested in paying low prices. Reducing allocation, on the other hand, could drive away customers that are willing to pay more for earlier access to these spot instances, leading them to favor competing services. Fortunately, these changes in buyers' valuations can be estimated via user bids and demand level. Using these publicly observable features as the public context, the seller can model buyer valuation's evolution via a MDP framework, faithfully representing the effect of its allocation policy on next-step valuation distribution.

A different setting that nevertheless shares similar underlying problem structure is *online*

2. Spot instances are idle cloud computing resources that not currently used by any users. These instances are priced dynamically by AWS based on predicted market demand, often at a considerably lower price but with lower priority when other contracted demands come [Agmon Ben-Yehuda et al., 2013, George et al., 2019, Baughman et al., 2019].

advertising where an ad exchange platform sells advertising opportunities to an advertiser. In this case, both the advertiser's targetting user population and demographic information about each Internet user are public and dynamically changing contexts. They help the platform to estimate the advertiser's willingness to pay and thus to better price the advertiser. The evolution of the advertiser's valuation distribution then naturally follows the following MDP structure. If he is allocated the ad spot, the user population he has already captured is updated. This may facilitate the advertiser's value to transit to a different one due to his potentially switching target of audience. This contributes to the new public context about the advertiser, to whom the platform proceeds to sell the next ad opportunity. In such applications, our MDP formulation is crucial for capturing such transition dynamics, as a context variable is necessary for recording the audience that the advertiser has already reached, and the public context's evolution is affected by both the current public context and the allocation the advertiser receives.

Our contributions. Motivated by the applications above, we formalize and study a dynamic mechanism design problem where the buyer's valuation distributions is described by a Markov Decision Process (MDP), with past allocation outcomes as the "decision". We aim to find the direct, incentive-compatible (IC), and ex-post individually rational (ex-post IR) dynamic mechanism with maximal revenue. To address the practical challenge that the seller may not know in advance how allocations affect later valuations, we also study the natural machine learning problem on how to efficiently learn an optimal mechanism through interacting with the environment. Towards that end, we present three major findings, as detailed below.

First, we introduce a natural dynamic design problem in an MDP environment and show a revelation-principle-style characterization. That is, it is without loss of generality to consider a family of mechanisms we call *Augmented Bank Account Mechanism* (ABAM). ABAM significantly generalizes the previously known bank account mechanism (BAM) [Mirrokni

et al., 2016a, 2020]. Specifically, BAMs were developed for situations where past trajectory does not affect future environment parameters such as buyer value distribution. Hence, they only need to track buyer's total surplus thus far (i.e., the "bank account"), and does not need to track past item allocations to the buyer neither the public context. However, in our MDP setup, past allocations affect both future context and buyer valuations. Therefore, to optimize aggregated utility and also to account for buyer's incentive, our mechanism need to additional track past allocations, hence the name "augmented" bank account mechanism. Notably, this augmentation is not a trivial modification — it makes both the revenue maximization and enforcing dynamic incentive compatibility much more challenging. To address these challenges, the ABAM has to carefully design the payment rule that can accurately account for the impact of untruthful reporting on both current-step utility and expected future utility. Built upon these changes to the mechanism, we prove that ABAMs can achieve optimal revenue in our setting. On the technical side, core to this proof is a non-trivial generalization of the family of "symmetric mechanisms", first introduced in in [Mirrokni et al., 2020], to the novel and more challenging MDP setting. The generalization is meticulously constructed to keep track of the effects of earlier allocations, while also taking care avoiding violating the timing of interactions in our setting.

Second, built upon the revelation principle above, we develop an algorithm to compute an (additive) ϵ-optimal mechanism in time polynomial in the input size and $1/\epsilon$ when the MDP's transition probabilities are known beforehand.[3] This result requires us to go significantly beyond earlier results for settings with independent valuations across time (e.g., [Ashlagi et al., 2023, Mirrokni et al., 2016b]). For these settings, the key idea shared by previous approaches is to show that the seller's expected continuation revenue is a *concave function* of the buyer's expected continuation utility. This crucial property then allows them to construct a polynomially-sized piece-wise linear approximation to the expected continuation revenue.

3. This is also called a Fully Polynomial Time Approximation Scheme (FPTAS).

Then near-optimal mechanism can by found via dynamic programming. Unfortunately, such approach fails in our MDP setting because the seller's expected continuation revenue here depends on not only buyer's expected continuation utility but also the allocation probability of each type. A similar approach of discretization as in Mirrokni et al. [2016b] will lead to an approximated continuation revenue function that is the product of a piecewise linear function in expected utility and a linear function in the allocation rule. However, since both expected utility and allocation rule are decision variables, the resulting approximation revenue function is a piece-wise *bilinear* function, making it intractable to optimize. To bypass these challenges, we resort to a refined analysis of the properties of continuation revenue in our setting. Specifically, employing a generalized variant of the envelope theorem by Milgrom and Segal [2002], we are able to show that the continuation revenue is sufficiently smooth in the allocation level. Using the observation, our proposed algorithm simultaneously discretize the allocation level and continuation revenue, using only polynomially many pieces.

Finally, we study the situation without knowing the underlying MDP in advance. Leveraging the above computational algorithm, we develop an efficient reinforcement learning algorithm to learn an ϵ-optimal, approximately IC, and approximately ex-post IR mechanism, using polynomially many samples in polynomial time.[4] We stress that, whereas typical RL approaches look to learn a near-optimal Markovian policy, the optimal mechanism in our setting is non-Markovian. In particular, the allocation, per-step payment, and spend rules of an ABAM are all non-Markovian, requiring special care when designing a learning algorithm. Moreover, we also need to guarantee that the learned non-Markovian spend rule yields a mechanism that is approximately IC and approximately ex-post IR. To address these challenges, we carefully relax the dynamic program for solving the optimal mechanism, which allows us to isolate the estimation errors to the spend rules alone. As approximate IC and

4. We are aware that for static setups of mechanism design, there is a quite general black-box reduction from ϵ-BIC mechanism to an exactly BIC mechanism modulo negligible revenue loss Cai et al. [2021]. However, it appears unclear whether such a reduction exits for our dynamic mechanism design setup, which is an intriguing open problem.

approximate ex-post IR guarantees all require on high-accuracy estimates of the spend rule, we draw inspiration from "reward-free reinforcement learning" to ensure that the spend rule is estimated sufficiently well with polynomially many samples [Jin et al., 2020a]. Specifically, we are able to show that, loosely speaking, so long as there exists some policy that can reach the context with non-negligible probability, then the corresponding spend rule will be estimated well, despite its non-Markovian nature. Careful analysis of the error terms shows that the constraint violation at each public context is small, as long as there exists some dynamic mechanism that can easily reach the context. Conversely, we show that IC and ex-post IR are not violated significantly, unless the context occurs rarely under all possible dynamic mechanisms. Specifically for the approximate-IC guarantee, our results imply that the amount a buyer may gain from untruthful reporting is bounded *uniformly* for all bidding policies, and the amount is independent of either the learned mechanism or the optimal mechanism itself.

Related Works. Our work subscribes to the rich line of research on optimal dynamic mechanism design [Krähmer and Strausz, 2015, Mirrokni et al., 2016b, Papadimitriou et al., 2016, Deng et al., 2019, Mirrokni et al., 2020, Bergemann and Välimäki, 2010, Kakade et al., 2013, Kanoria and Nazerzadeh, 2014, Athey and Segal, 2013]. Particularly, [Mirrokni et al., 2016a, 2020, Ashlagi et al., 2023] are among the first works to pursue the revenue-maximizing, IC, and ex-post IR mechanism in dynamic mechanism design, focusing on the setting where the buyer's valuation distribution may change over time. However, in their models, the buyer's value is not affected by his previous allocations. Following this line of work, Deng et al. [2021] offer a more general treatment of the problem, allowing additional constraints such as the buyer's budget. However, their results do not imply computationally efficient algorithms for computing near-optimal mechanisms in our setting neither show how such a mechanism could be efficiently learned. In the same vein, Pavan et al. [2014] characterize the first-order optimality conditions for a very general setup of dynamic mechanism design. However, their

results are not computational neither learning-theoretic. Another loosely related line of work studies "Markovian buyers" [Battaglini, 2005, Ouyang et al., 2015, Garrett, 2016], where the buyer's valuation is assumed to evolve according to a *Markov process*, as opposed to a *Markov Decision Process*. Compared to our setting, this line of work omits the challenges caused by the dependence of valuation distribution on earlier allocations, and often do not consider efficient computing or learning algorithms. We refer interested readers to [Bergemann and Välimäki, 2019] for additional discussions on related concepts in dynamic mechanism design.

Our work is also related to, but drastically different from, the line of recent work combining reinforcement learning (RL) with dynamic mechanism or information design [Wu et al., 2022, Ai et al., 2022, Lyu et al., 2022b, Min et al., 2022, Liu et al., 2022, Zu et al., 2021, Mansour et al., 2022, Simchowitz and Slivkins, 2023]. Of these works, some features a "factorized" action space, where one component of the seller's action affects only the later valuation distributions, but not the current step's IC constraints, and the other component focuses on only the current step's IC constraints, without considering the later steps' valuation distribution [Ai et al., 2022, Min et al., 2022]. Some assumes that a new participant arrives at every step of the underlying MDP and leaves immediately after [Wu et al., 2022]. Both setups significantly reduce the difficulty of designing an optimal IC mechanism, leading to easier-to-characterize optimal mechanisms that are also easier to learn via RL. On the other hand, while works such as [Mansour et al., 2022, Simchowitz and Slivkins, 2023] do not feature similar simplifying assumptions, their setting is drastically different from ours. In our setting, the learner is a monopolistic seller whose aim is to maximize revenue. In this line of work, however, the learner is simply trying to learn the optimal policy in some MDP, whose challenge lies in incentivizing agents to explore via strategically revealing information to these agents. While they provide efficient learning algorithms, these algorithms shed little light on how a revenue-maximizing mechanism could be recovered.

Finally, our techniques for learning optimal mechanisms is inspired by the line of work

on reward-free RL, pioneered by Jin et al. [2020a], with later works improving sample-complexity [Zhang et al., 2021b, Li et al., 2023] or extending the results into more general function approximation settings [Wang et al., 2020b, Wagenmaker et al., 2022, Zhang et al., 2021a]. While we utilize the design idea of Jin et al. [2020a], it is far from being straightforward to convert reward-free exploration guarantees to approximate IC guarantees, a technical contribution of our work.

Notations and Terminology. We use $\|\cdot\|_\infty$ to denote the ℓ_∞-norm. For two non-negative functions f, g, we say $f(n) = \mathcal{O}(g(n))$ if there exists some $c > 0$, $n_0 > 0$ such that $f(n) \leqslant cg(n)$ for all $n \geqslant n_0$. We use $\widetilde{\mathcal{O}}$ when the logarithmic factors are ignored. Finally, we say a mechanism is ϵ-*optimal* if the expected revenue it achieves is at least $\mathrm{OPT} - \epsilon$, where OPT is the maximum expected revenue that can be achieved by any exactly IC, IR, and direct mechanism.

2.2 Preliminaries

We use the tuple $(\mathcal{S}, \Theta, \mathcal{X}, H, \mathcal{P})$ to describe the MDP that governs the evolution of the buyer's valuation distribution. Particularly, $\mathcal{S}$ is the space for the public context at each step, Θ is the buyer's type space, $\mathcal{X} = \{0, 1\}$ is the seller's allocation space, where $x_h = 0$ means that the seller does not sell the item to the buyer at step h and $x_h = 1$ means the seller does sell the item. We use H to denote the horizon. Let $\mathcal{P}^{\mathcal{S}} = \{\mathcal{S} \times \mathcal{X} \to \Delta(\mathcal{S})\}^H$ and $\mathcal{P}^{\Theta} = \{\mathcal{S} \to \Delta(\Theta)\}^H$ denote the transition kernel, where for each $h \in [H]$, $\mathcal{P}_h^{\mathcal{S}}(\cdot \mid s_h, x_h)$ is the distribution over the public context s_{h+1} at $h + 1$, conditioned on the public context s_h and realized allocation x_h at step h. The buyer then draws his private type θ_{h+1} from a distribution dependent on s_{h+1}, i.e. $\theta_{h+1} \sim \mathcal{P}_{h+1}^{\Theta}(\cdot|s_{h+1})$. For convenience, we let $\mathcal{P} = \{\mathcal{P}^{\mathcal{S}}, \mathcal{P}^{\Theta}\}$ and assume without loss of generality that s_1 is deterministic, that is, the initial context is always fixed at some s_1.

We focus on direct mechanisms and let the buyer's report space be Θ. The interaction

Figure 2.1: Visualization of the Transition Dynamics. Dashed lines denote correlation and solid arrows denote conditional independence.

between the buyer and the seller can then be summarized as follows. When $h = 1$, the initial public context s_1 is realized and the buyer receives his private type θ_1. For all $h \in [H]$, the public context realizes at $s_h \in \mathcal{S}$ and the buyer receives his type $\theta_h \in \Theta$. The buyer then reports, potentially untruthfully, to the seller his type as $\widehat{\theta}_h \in \Theta$. The seller, having observed the public context s_h and received the reported type $\widehat{\theta}_h$, chooses her action $x_h \sim \text{Ber}(\chi_h(\cdot))$ and price $p_h = \psi_h(\cdot)$ according to some mechanism $\mathcal{M}(\chi, \psi)$. The buyer realizes his instantaneous utility and the next step's context is given by $s_{h+1} \sim \mathcal{P}_h^{\mathcal{S}}(\cdot|s_h, x_h)$.

Before we characterize the mechanism, we define the history available to the seller at each step $h \in [H]$, which we denote by $\widehat{\eta}_{(1,h-1)}$ to reflect the fact that the history is dependent on the reported type $\widehat{\theta}$ rather than the actual private type itself. Particularly, we let

$$\widehat{\eta}_h = (s_h, \widehat{\theta}_h, x_h), \text{ for all } h \in [H],$$

and use $\mathcal{H}_h$ to denote the space of all histories with length h. Additionally, we let $\mathcal{H} = \cup_{h=1}^{H} \mathcal{H}_h$. Under the definition, the tuple $(\widehat{\eta}_{(1,h-1)}, s_h)$ captures all the publicly available "historical" variables, save for the buyer's current step's report, that the buyer may include as inputs to the mechanism $\chi_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h)$ and $\psi_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h)$. For convenience, we let $\eta_{(1,h-1)}$ denote the history when the buyer is truthful in the first $h-1$ steps. At each step h, the buyer's utility is $u_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h; \theta_h) = \theta_h \mathbb{E}[x_h] - p_h = \theta_h \chi_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h) - \psi_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h)$, where the expectation is taken over the randomness of x_h, the current step's allocation.

The buyer's expected continuation utility for any arbitrary bidding policy b is then

$$\overline{U}_h^{b(h+1,H)}(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h) = \mathop{\mathbb{E}}_{\widehat{\eta}_{(h+1,H)}^{b(h+1,H)}, x_h} \left[\sum_{\tau=h+1}^{H} u_\tau(\widehat{\eta}_{(1,\tau-1)}^{b(h+1,H)}, s_\tau, b_\tau(\widehat{\eta}_{(1,\tau-1)}^{b(h+1,H)}, s_\tau, \theta_\tau); \theta_\tau) \right],$$

where we let $\widehat{\eta}^b$ denote the history under a specific bidding policy b. We highlight the fact that $\overline{U}$ also takes expectation over the current step's allocation outcome x_h, as the buyer cannot observe the realized outcome when attempting to maximize his future expected utility at each step. For convenience we let $\overline{U}$ denote the expected future utility when the buyer reports truthfully.

For any mechanism $\mathcal{M}$ we let $\textsc{Utl}(\mathcal{M})$ denote the expected episodic utility attained under the mechanism and $\textsc{Utl}(\mathcal{M}|\eta_{(h,h')})$ denote the expected episodic utility conditioned on the history $\eta_{(h,h')}$, both under the assumption that the buyer is truthful. Particularly

$$\textsc{Utl}(\mathcal{M}|\eta_{(h,h')}) = \mathbb{E}_{\eta'_{(1,h-1)}, s'_h} \left[\sum_{\tau=1}^{H} u_\tau(\eta'_{(1,\tau-1)}, s'_\tau, \theta'_\tau; \theta'_\tau) | (s', \theta', x')_{(h,h')} = (s, \theta, x)_{(h,h')} \right].$$

Let $\textsc{Utl}(\mathcal{M}|\eta_{(h,h'-1)}, s_{h'}, \theta_{h'})$ be analogously defined whenever we wishes to take into consideration the realized public context and private type at step h'. We note that

$$\textsc{Utl}(\mathcal{M}|\eta_{(h,h'-1)}, s_{h'}, \theta_{h'}) = \mathbb{E}_{x_{h'}} \left[\textsc{Utl}(\mathcal{M}|\eta_{(h,h')}) \right].$$

With these definitions in mind, we now formalize two key desiderata that an ideal mechanism should satisfy: dynamic incentive compatibility and ex-post individual rationality.

Dynamic Incentive Compatibility. Our definition of dynamic incentive compatibility is directly adapted from [Mirrokni et al., 2020], where the seller's goal is to incentivize the buyer to report truthfully regardless of his prior history. As the buyer does not know

the realizations of future public contexts nor private types, he considers his current-step utility u in addition to the continuation utility $\bar{U}$. We formalize the definition below and refer interested readers to [Mirrokni et al., 2016a, 2020] for additional discussions on the constraint.

Definition 2.2.1 (Dynamic Incentive Compatibility)**.** *We say a mechanism $\mathcal{M}$ is dynamic incentive compatible (IC) if for any step $h \in [H]$, history $\widehat{\eta}_{(1,h-1)}$, context s_h, deviating report $\widehat{\theta}_h$, and bidding strategy in subsequent steps $b_{(h+1,H)}$, we have*

$$\begin{aligned} &u_h(\widehat{\eta}_{(1,h-1)}, s_h, \theta_h; \theta_h) + \overline{U}_h(\widehat{\eta}_{(1,h-1)}, s_h, \theta_h) \\ &\quad \geqslant u_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h; \theta_h) + \overline{U}_h^{b_{(h+1,H)}}(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h). \end{aligned} \tag{2.2.1}$$

We also introduce stage-IC, a notion that is equivalent to IC but is easier to work with in a dynamic mechanism design setting. The key distinction between the two definitions is that the former takes into consideration all possible continuation bidding strategies, whereas the latter restricts our focus to the setting where the buyer deviates at the current step, but reports truthfully in all ensuing steps.

$$\begin{aligned} \textbf{Stage-IC} \quad &u_h(\widehat{\eta}_{(1,h-1)}, s_h, \theta_h; \theta_h) + \overline{U}_h(\widehat{\eta}_{(1,h-1)}, s_h, \theta_h) \\ &\quad \geqslant u_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h; \theta_h) + \overline{U}_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h). \end{aligned} \tag{2.2.2}$$

Despite the differences in appearances, stage-IC and IC are in fact equivalent.

Lemma 2.2.2. *IC, as defined by Definition 2.2.1, is equivalent to stage-IC* (2.2.2)*.*

Ex-Post Individual Rationality. Inspired by prior literature, we focus on ex-post participation constraints, where the buyer's episodic utility is non-negative for every realization of public context and buyer types. In particular, we impose the following ex-post individual

rationality constraint on the mechanism, where for any truthfully reported history we have

$$\textbf{IR} \quad \sum_{h=1}^{H} u_h(\eta_{(1,h-1)}, s_h, \theta_h; \theta_h) \geqslant 0, \tag{2.2.3}$$

namely, buyer's who participate truthfully should almost surely receive non-negative utility at the end of the mechanism.

Reinforcement Learning Preliminaries. We now introduce concepts in reinforcement learning that will be useful when we discuss our learning algorithm. Let $\pi_{(1,H)} : \{\mathcal{S} \to \Delta(\mathcal{X})\}^H$ denote a (Markovian) type-agnostic policy. That is, for any h, s, $\pi_h(s)$ outputs a distribution over the allocations in $\mathcal{X}$ when the public context reached at step h is s, not affected by either $\eta_{(1,h-1)}$ or θ_h. Let $\Pr_h^\pi(s)$ denote the probability that the context s is reached by π at step h. Although type-agnostic policies seem much more restrictive than the allocation rule in general mechanism design, it can be used to generate any distribution over the contexts for any allocation rule χ and bidding policy b. We formalize the statement as follows.

Lemma 2.2.3. *For any allocation rule $\chi_{(1,H)}$ and bidding policy $b_{(1,H)}$, there exists a type-agnostic policy $\pi_{(1,H)}$ such that for all h, s*

$$\Pr_h^\pi(s) = \Pr_{\widehat{\eta}_{(1,H)} \sim \chi_{(1,H)}, b_{(1,H)}}(s_h = s),$$

where the probability on the right hand side is calculated with respect to the distribution over the reported history $\widehat{\eta}_{(1,H)}$.

A key implication of the lemma is that we can use a Markovian policy to fully explore the environment, covering all possible distributions that could be covered by possibly non-Markovian dynamic mechanisms, even when the buyer is not truthful.

2.3 The Augmented Bank Account Mechanism and its Characterization

Before we formally introduce the family of augmented bank account mechanisms, we begin with a short detour to better understand the constraints posed by incentive compatibility, turning our focus back to (2.2.2). Plugging in the definition of $u_h(\cdot,\cdot;\theta_h)$, we rewrite the equation as

$$\begin{aligned}\theta_h\chi_h(\widehat{\eta}_{(1,h-1)},s_h,\theta_h)-\psi_h(\widehat{\eta}_{(1,h-1)},s_h,\theta_h)+\overline{U}_h(\widehat{\eta}_{(1,h-1)},s_h,\theta_h)\\ \geqslant\theta_h\chi_h(\widehat{\eta}_{(1,h-1)},s_h,\widehat{\theta}_h)-\psi_h(\widehat{\eta}_{(1,h-1)},s_h,\widehat{\theta}_h)+\overline{U}_h(\widehat{\eta}_{(1,h-1)},s_h,\widehat{\theta}_h),\end{aligned}$$

which in turn rearranges to

$$\begin{aligned}\psi_h(\widehat{\eta}_{(1,h-1)},s_h,\theta_h)-\psi_h(\widehat{\eta}_{(1,h-1)},s_h,\widehat{\theta}_h)\leqslant&\theta_h\chi_h(\widehat{\eta}_{(1,h-1)},s_h,\theta_h)-\theta_h\chi_h(\widehat{\eta}_{(1,h-1)},s_h,\widehat{\theta}_h)\\ &+\overline{U}_h(\widehat{\eta}_{(1,h-1)},s_h,\theta_h)-\overline{U}_h(\widehat{\eta}_{(1,h-1)},s_h,\widehat{\theta}_h).\end{aligned}\tag{2.3.1}$$

The expression, at a high level, highlights the fact that the pricing rule for any dynamic IC mechanism must roughly account for two terms, the instantaneous utility $\theta_h\chi_h(\cdot,\cdot,\cdot)$ and the expected future utility $\overline{U}_h(\cdot,\cdot,\cdot)$. More specifically, for each step $h\in[H]$, the difference in payments levied for type θ_h and type $\widehat{\theta}_h$ should reflect both the change in the buyer's instantaneous utility and the buyer's future expected utility.

The inequality (2.3.1) reminds us of bank account mechanisms introduced by Mirrokni et al. [2016a,b]. A defining feature of this family of mechanisms is the notion of bank account "balance", which keeps track of the buyer's expected episodic conditioned on the history observed so far. Via the "balance" term, bank account mechanisms are able to measure the change in expected future utility $\bar{U}_h$ as the history evolves, thereby ensuring IC.

In the MDP setting, however, it is not sufficient to keep track of the bank account balances, as the distribution over types is affected by the previous step's allocation due to the Markovian transition kernel $\mathcal{P}^{\mathcal{S}}$. To resolve the issue, we propose the following family of mechanisms called Augmented Bank Account Mechanisms (ABAMs), where we *augment* balances with the previous step's history.

Definition 2.3.1. *An augmented bank account mechanism $\mathcal{B}$ is parameterized by the tuple $\langle \xi_{(1,H)}, \varphi_{(1,H)}, \texttt{bal}_{(1,H)}, \delta_{(1,H)}, \sigma_{(1,H)} \rangle$ where for each $h \in [H]$*

- *allocation rule $\xi_h : \mathbb{R}_+ \times \mathcal{H}_1 \times \mathcal{S} \times \Theta \to [0,1]$ maps balance, previous step's one step history, current public context, and current type to an allocation probability,*
- *payment rule $\varphi_h : \mathbb{R}_+ \times \mathcal{H}_1 \times \mathcal{S} \times \Theta \to \mathbb{R}_+$ maps balance, previous step's one step history, current public context, and current type to step payment,*
- *balance function $\texttt{bal}_h : \mathcal{H}_{h-1} \times \mathcal{S} \times \Theta \to \mathbb{R}_+$ is defined recursively by the following equation,*

$$\forall \eta_{(1,h-1)}, s_h, \theta_h,\ \texttt{bal}_{h+1} = \texttt{bal}_h - \sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h) + \delta_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h)$$

  where $\texttt{bal}_1 = 0$,

- *deposit rule $\delta_h : \mathbb{R}_+ \times \mathcal{H}_1 \times \mathcal{S} \times \Theta \to \mathbb{R}_+$ maps balance, previous step's one step history, current public context, and current type non-negative real that is added to the current balance.*
- *spend rule $\sigma_h : \mathbb{R}_+ \times \mathcal{H}_1 \times \mathcal{S} \to \mathbb{R}$ maps balance, previous step's one step history, and the current public context to a real number no greater than the current balance,*

where we recall $\mathcal{H}_1$ is the space of all possible tuples $(s_\tau, \theta_\tau, x_\tau)$ for any arbitrary $\tau \in [H]$.

As the type distribution at each step relies on the previous step's allocation, only by keeping track of previous step's history can we accurately calculate the next step's type distribution. The proposed mechanism can be viewed as a specialized instance of Lossless History Compression mechanism proposed by Deng et al. [2021], where the history is compressed to balance and previous step's history. However, we note that the LHC framework is too general and lacking in computational guarantees. Moreover, although Deng et al. [2021] provide a convincing argument extending their results to our setting, we are not aware of mathematically rigorous proofs for why LHCs extend to the case where later type distributions depend on earlier allocations.

As a shorthand, we use

$$\widehat{u}(\texttt{bal}_h, \widehat{\eta}_{h-1}, s_h, \widehat{\theta}_h; \theta_h) = \xi_h(\texttt{bal}_h, \widehat{\eta}_{h-1}, s_h, \widehat{\theta}_h)\theta_h - \varphi_h(\texttt{bal}_h, \widehat{\eta}_{h-1}, s_h, \widehat{\theta}_h)$$

to denote the per-step utility of the augmented bank account mechanism when the buyer's true type is θ_h and reports instead $\widehat{\theta}_h$, differentiating it from the instantaneous utility u_h. At each step, the mechanism charges the buyer

$$p_h = \varphi_h(\texttt{bal}_h, \widehat{\eta}_{h-1}, s_h, \widehat{\theta}_h) + \sigma_h(\texttt{bal}_h, \widehat{\eta}_{h-1}, s_h),$$

according to the balance, the previous step's history, and the current step's public context and reported type.

Before proceeding further, we highlight the importance of and explain the intuition behind the spending rule σ_h in our setting. Consider a simple two-stage setting. The seller's allocation rule in the first stage affects the distribution of s_2, which is observed only at the next stage. The terms $s_1, \widehat{\theta}_1$, when combined with the seller's allocation rule, yield the conditional distribution of s_2, much like the type distribution in a traditional one-step mechanism design problem. The seller can then account for the information rent of s_2 via

σ_2, and the allocation and payment rules ξ, φ only need to account for the uncertainty surrounding θ_2. More succinctly, σ_h accounts for the uncertainty in s_h, and ξ, φ accounts for the uncertainty of θ_h after observing s_h.

We now show how this property simplifies the analysis of IC conditions.

Lemma 2.3.2. *An augmented bank account mechanism $\mathcal{B}$ is IC if for all $h \in [H]$, $\texttt{bal}_h$, η_{h-1},s_h, and pair of types θ_h, θ'_h*

$$\widehat{u}(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) \geqslant \widehat{u}(\texttt{bal}_h, \eta_{h-1}, s_h, \theta'_h; \theta_h)$$

and for any $\texttt{bal}_h, \widehat{\eta}_{h-1}$ and $\texttt{bal}'_h, \widehat{\eta}'_{h-1}$, we have

$$\begin{aligned}
&\mathbb{E}_{x_{h-1}, s_h | s_{h-1}, \widehat{\theta}_{h-1}} [\sigma_h(\texttt{bal}_h, \widehat{\eta}_{h-1}, s_h)] - \mathbb{E}_{x'_{h-1}, s'_h | s'_{h-1}, \widehat{\theta}'_{h-1}} [\sigma_h(\texttt{bal}'_h, \widehat{\eta}'_{h-1}, s'_h)] \\
&\quad = \mathbb{E}_{x_{h-1}, s_h, \theta_h | s_{h-1}, \widehat{\theta}_{h-1}} \left[\widehat{u}(\texttt{bal}_h, \widehat{\eta}_{h-1}, s_h, \theta_h; \theta_h)\right] \\
&\qquad - \mathbb{E}_{x'_{h-1}, s'_h, \theta'_h | s'_{h-1}, \widehat{\theta}'_{h-1}} \left[\widehat{u}(\texttt{bal}'_h, \widehat{\eta}'_{h-1}, s'_h, \theta'_h; \theta'_h)\right].
\end{aligned}$$

Similarly, the deposit rule δ_h streamlines the analysis of the ex-post IR condition.

Lemma 2.3.3. *An augmented bank account mechanism $\mathcal{B}$ is ex-post IR if for all $h \in [H]$, $\texttt{bal}_h$, η_{h-1}, s_h, θ_h, and θ'_h*

$$\widehat{u}(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) \geqslant \delta_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h).$$

Our goal is to show that it is without loss of generality to restrict our focus to augmented bank account mechanisms. Observe that in Definition 2.3.1, the tuple $\texttt{bal}_h, s_{h-1}, \theta_{h-1}$ almost ubiquitous. As it turns out, some notion of equivalence is implied by the tuple, and we focus on mechanisms that, loosely speaking, treat all histories with length h by only looking at the tuple $(\texttt{bal}_h, s_{h-1}, \theta_{h-1})$ with an appropriately constructed way to update

balances. Particularly, we have the following definitions, extending their counterparts found in earlier works into our novel MDP setting.

Definition 2.3.4 (History Equivalence)**.** *Given any direct mechanism $\mathcal{M}$, equivalence relation between histories, $(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) \sim (\eta'_{(1,h-2)}, s'_{h-1}, \theta'_{h-1})$, is defined as*

$$
\begin{aligned}
&(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) \sim (\eta'_{(1,h-2)}, s'_{h-1}, \theta'_{h-1}) \\
&\quad \iff \begin{cases} \textsc{Utl}(\mathcal{M}|\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = \textsc{Utl}(\mathcal{M}|\eta'_{(1,h-2)}, s'_{h-1}, \theta'_{h-1}), \\ s_{h-1} = s'_{h-1}, \theta_{h-1} = \theta'_{h-1}. \end{cases}
\end{aligned}
$$

History equivalence allows us to focus a specific family of dynamic mechanisms. Specifically, one that treats all equivalent histories as if they were the same. At a high level, because equivalent histories share the same public context and reported type, if they receive the same allocation level at step $h-1$, the resulting public context and type distributions at step h will also be the same. Moreover, as these histories have the same expected episodic utility, using the same submechanism starting from step h yields the same utility for both histories. We formalize the definition of such mechanisms as follows.

Definition 2.3.5 (Symmetric Mechanism)**.** *We say a mechanism $\mathcal{M}$ is symmetric if for pairs of equivalent histories the corresponding submechanisms are identical. That is, if $(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) \sim (\eta'_{(1,h-2)}, s'_{h-1}, \theta'_{h-1})$, then for all possible x_{h-1}, $\tau \in [H]$, and history from step h to $\tau - 1$ we have*

$$
\begin{cases}
\chi_\tau((\eta_{(1,h-2)}, (s_{h-1}, \theta_{h-1}, x_{h-1}), \eta_{(h,\tau-1)}), s_\tau, \theta_\tau) \\
\qquad = \chi_\tau((\eta'_{(1,h-2)}, (s'_{h-1}, \theta'_{h-1}, x_{h-1}), \eta_{(h,\tau-1)}), s_\tau, \theta_\tau) \text{ for all } (s_\tau, \theta_\tau) \in \mathcal{S} \times \Theta, \\
\psi_\tau((\eta_{(1,h-2)}, (s_{h-1}, \theta_{h-1}, x_{h-1}), \eta_{(h,\tau-1)}), s_\tau, \theta_\tau) \\
\qquad = \psi_\tau((\eta'_{(1,h-2)}, (s'_{h-1}, \theta'_{h-1}, x_{h-1}), \eta_{(h,\tau-1)}), s_\tau, \theta_\tau) \text{ for all } (s_\tau, \theta_\tau) \in \mathcal{S} \times \Theta.
\end{cases}
\tag{2.3.2}
$$

Symmetric mechanisms are much less complicated than the general class of all possible dynamic mechanisms. Rather than keeping track of all possible values, we can select one "representative" value for each $(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1})$. By Definition 2.3.5, symmetric mechanisms function the same for all histories equivalent to the selected representative, significantly reducing the policy's complexity. As it turns out, in the MDP setting it is again without loss of generality to consider only symmetric mechanisms, which we formalize below.

Lemma 2.3.6 (Symmetrization)**.** *For direct, IC, and ex-post IR dynamic mechanism $\mathcal{M}$, there is a symmetric, IC, and ex-post IR dynamic mechanism $\mathcal{M}^{\text{symmetric}}$ such that*

$$\textsc{Utl}(\mathcal{M}) = \textsc{Utl}(\mathcal{M}^{\text{symmetric}}), \qquad \textsc{Rev}(\mathcal{M}) \leqslant \textsc{Rev}(\mathcal{M}^{\text{symmetric}}).$$

Moreover, if $\mathcal{M}$ is deterministic, $\mathcal{M}^{\text{symmetric}}$ is also deterministic.

A key challenge behind Lemma 2.3.6 is, again, the fact that the buyer's valuation distribution relies on the allocation he receives in the previous round. Unlike earlier works [Mirrokni et al., 2016b, 2020], the definitions of history equivalence and symmetric mechanisms both need to include the public context and the buyer's reported type at step $h-1$: both variables are needed for calculating the type distribution at step h, and only tracking the expected utilities is no longer sufficient.

We now focus on a specific subset of mechanisms within the family of ABAMs. Inspired by [Mirrokni et al., 2016a], we call this kind of mechanism *core Augmented Bank Account Mechanism*, or core ABAM for short.

Definition 2.3.7. *Let $g_h : \mathcal{H} \times \mathcal{S} \times \Theta \to \mathbb{R}$ be a function mapping a history of length $h-1$, a public context, and a type report to a real number for all $h \in [H] \cup \{0\}$ and $y_h : \mathcal{H} \times \mathcal{S} \times \Theta \to \Delta\mathcal{X}$ one that maps from a history of length $h-1$, the current public context, and the current reported type to an allocation.*

Consider the following construction of an augmented bank account mechanism based on functions $g = \{g_h\}_{h=0}^{H}$ *and* $y = \{y_h\}_{h=1}^{H}$*, which we denote as* $B^{g,y}$*.*

- $\texttt{bal}_{h+1}(\eta_{(1,h-1)}, s_h, \theta_h) = g_h(\eta_{(1,h-1)}, s_h, \theta_h) - C_h$,
- $\xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) = y_h(\eta'_{(1,h-1)}, s_h, \theta_h)$*, where* $\eta'_{(1,h-1)}$ *is some arbitrary history such that* $\texttt{bal}_h(\eta'_{(1,h-2)}, s'_{h-1}, \theta'_{h-1}) = \texttt{bal}_h$ *and* $\eta'_{h-1} = \eta_{h-1}$.
- $\varphi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) = \xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h)\theta_h - \int_0^{\theta_h} \xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta)d\theta$,
- $\delta_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) = \xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h)\theta_h - \varphi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h)$,
- $\sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h) = \texttt{bal}_h + \delta_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) - \texttt{bal}_{h+1}$,

where the constants C_h*,* C_H *satisfy the following constraints to ensure the balance is non-negative.*

$$\begin{cases} C_h \leqslant \inf_{\eta_{(1,h-1)}, s_h, \theta_h} g_h(\eta_{(1,h-1)}, s_h, \theta_h) \text{ for all } h \in [H-1], \\ C_H \leqslant \min\{0, \inf_{\eta_{(1,H-1)}, s_H, \theta_H} g_H(\eta_{(1,H-1)}, s_H, \theta_H)\}. \end{cases}$$

When the construction is well-defined, we call the resulting mechanism a core ABAM.

Core ABAM differs from earlier works due to its need to track the previous step's history. Without such information, the seller cannot exactly calculate the buyer's type distribution at the next step, which is affected by the allocation rule the seller selects in the current step, making optimal mechanism design impossible. We now show that for each symmetric direct mechanism, one can construct a core ABAM with the same overall outcome in the form of the following lemma.

Lemma 2.3.8. *For any symmetric direct, IC, and IR mechanism* $\mathcal{M} = (\chi, \psi)$*,* $B^{g,y}$ *is a core ABAM if for all* h *and all histories of length* h

$$g_h(\eta_{(1,h-1)}, s_h, \theta_h) = \mathbb{E}[\textsc{Utl}(\mathcal{M}|\eta_{(1,h-1)}, s_h, \theta_h)].$$

Moreover, $B^{g,y}$ satisfies

$$\textsc{Utl}(B^{g,y}) = \textsc{Utl}(\mathcal{M}) = \mathbb{E}_{\theta_1}[g_1(\varnothing, s_1, \theta_1)], \qquad \textsc{Rev}(B^{g,y}) = \textsc{Rev}(\mathcal{M}),$$

and is IC and IR.

Proof Sketch of Lemma 2.3.8. The full proof is deferred to Appendix 2.7.3. The key component is to utilize that different histories can be mapped to a representative one and the subsequent submechanisms are the same under symmetric mechanisms.

While our construction and proof draws inspiration from earlier results in [Mirrokni et al., 2016a], a crucial yet nuanced difference between core ABAMs and existing methods lies in the input space to core ABAMs, unique to our MDP setting. Observe that core ABAMs make use of $(\texttt{bal}_h, \eta_{h-1})$ as opposed to $(\texttt{bal}_h, s_{h-1}, \theta_{h-1})$. In other words, core ABAMs also keep track of the *realized* allocation, in addition to the public context and reported type. Indeed, it is possible for the general family of symmetric mechanisms to also depend on the realized allocation x_{h-1} at the previous step, and omitting the variable could limit the generality of core ABAMs. As $x_{h-1} \in \mathcal{X} = \{0, 1\}$, the inclusion of the variable only scale the number of possible inputs by a constant factor, and will not incur significant computational costs. □

Having shown that it is without loss of generality to consider only core ABAMs, we conclude this section by characterizing such mechanisms.

Theorem 2.3.9. *A function g is* consistent *if for all $h \in [H]$, histories $\eta_{(1,h-1)}$,*

$$g_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) - \mathbb{E}_{x_{h-1}, s_h, \theta_h}[g_h(\eta_{(1,h-1)}, s_h, \theta_h)]$$

is an absolute constant. A function g is symmetric *if for all $h \in [H]$, histories $\eta'_{(1,h-2)}$, $\eta'_{(1,h-2)}$, if $g_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = g_{h-1}(\eta'_{(1,h-2)}, s_{h-1}, \theta_{h-1})$ for some s_{h-1}, θ_{h-1},*

then for all x_{h-1}, s_h, θ_h

$$g_h(\eta_{(1,h-1)}, s_h, \theta_h) = g_h((\eta'_{(1,h-2)}, \eta_{h-1}), s_h, \theta_h),$$

where we recall $\eta_{h-1} = (s_{h-1}, \theta_{h-1}, x_{h-1})$ *and* $\eta_{(1,h-1)} = (\eta_{(1,h-2)}, \eta_{h-1})$. *The mechanism* $\mathcal{B}^{g,y}$ *is a core ABAM if and only if for any* $h \in [H]$

- y_h *is a sub-gradient of* g_h *with respect to* θ_h, *with range being* $\Delta(\mathcal{X})$. *That is, for all* h *and* $(\eta_{(1,h-1)}, s_h, \theta_h)$, *we have* $y_h(\eta_{(1,h-1)}, s_h, \theta_h) = \frac{\partial}{\partial \theta_h} g_h(\eta_{(1,h-1)}, s_h, \theta_h) \in [0, 1]$.
- g_h *is consistent and symmetric, convex in* θ_h, *and weakly increasing in* θ_h *for all* $h, \eta_{(1,h-1)}, s_h$.

Detailed proof is deferred to Appendix 2.7.3. The result is key to designing an efficient algorithm for computing an ϵ-optimal mechanism, one that we discuss in detail in the sequel.

2.4 Efficient Computation of an ϵ-Optimal Dynamic Mechanism

We now discuss how the optimal mechanism can be approximated when $\mathcal{S}$ and Θ are finite. We recall from Theorem 2.3.9 and Lemma 2.3.8 that it is without loss of generality to assume that $g(\eta) = \textsc{Utl}(\mathcal{M}|\eta)$ where $\mathcal{M}$ is the Augmented Bank Account Mechanism induced by g. We may use (β_h, η_{h-1}) to represent $\eta_{(1,h-1)}$, which is justified by Lemma 2.3.6, from which we know is without loss of generality to consider symmetric mechanisms. Finally, we assume without loss of generality that the lowest type in Θ is 0, namely $0 \in \Theta$.

As revenue is always the difference between welfare and utility, our objective reduces to maximizing the difference between welfare and balance. We use $\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1})$ to capture the difference between the mechanism's welfare from step h to H minus the final

balance. Specifically, for all $1 < h \leqslant H$, we let

$$\begin{aligned}
&\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \\
&\quad = \max_{\substack{g,y,\varphi \\ g_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = \beta_h}} \mathop{\mathbb{E}}_{x_{h-1}, \eta_{(h,H)}} \left[\sum_{\tau=h}^{H} y_\tau(\beta_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau)\theta_\tau - \texttt{bal}_{H+1} \right].
\end{aligned}$$

For the special cases where $h = H + 1$ or $h = 1$, we have

$$\begin{aligned}
&\Psi_{H+1}(\beta_{H+1}, s_H, \theta_H; y_H) = -\beta_{H+1}, \\
&\Psi_1(\beta_1, \varnothing, \varnothing; \varnothing) = \max_{\substack{g,y,\varphi \\ g_0(\varnothing) = \beta_1}} \mathop{\mathbb{E}}_{\eta_{(1,H)}} \left[\sum_{\tau=1}^{H} y_\tau(\beta_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau)\theta_\tau - \texttt{bal}_{H+1} \right].
\end{aligned} \tag{2.4.1}$$

In this case, we use β as a stand-in for expected episodic utility. The term $y_{h-1} \in [0, 1]$ denotes the allocation probability (i.e., the probability that $x_h = 1$) assigned to the the history $(\eta_{(1,h-1)}, s_h, \theta_h)$ and is needed for capturing the public context distribution at step h.

The function Ψ_h can be computed from Ψ_{h+1} via the program in Table 2.1. Particularly, for any $\beta_h \geqslant 0, s_{h-1} \in \mathcal{S}, \theta_{h-1} \in \mathcal{S}$, and $y_{h-1} = y_{h-1}(\beta_{h-1}, \eta_{h-2}, s_{h-1}, \theta_{h-1}) \in [0, 1]$, the optimum of the program equals to $\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1})$. Because the buyer's types are now discrete, the characterization provided in Definition 2.3.1 no longer pinpoints a unique per-step payment rule, and we seemingly need to optimize over the per-step payment rule φ as well. Nevertheless, as we show in Lemma 2.7.3, the program described in Table 2.1 is correct, as φ is determined by y_h and g_h.

Lemma 2.4.1. *Maximum revenue is given by* $\max_{\beta_1 \geqslant 0} \Psi_1(\beta_1, \varnothing, \varnothing; \varnothing)$*, where the function* $\Psi_1(\beta_1, \varnothing, \varnothing; \varnothing)$ *is obtained by recursively solving for* $\psi_H, \ldots, \Psi_1$ *using the program in Table 2.1, with* Ψ_{H+1} *given by* (2.4.1)*.*

Detailed proof is deferred to Appendix 2.7.4. The proof makes use of the characterization

$$\max_{g_h, y_h} \mathop{\mathbb{E}}_{x_{h-1}, s_h, \theta_h \,|\, y_{h-1}} \Big[y_h(\beta_h, \eta_{h-1}, s_h, \theta_h)\theta_h + \Psi_{h+1}(\beta_{h+1}(\beta_h, \eta_{h-1}, s_h, \theta_h), s_h, \theta_h; y_h(\beta_h, \eta_{h-1}, s_h, \theta_h)) \Big]$$

(2.4.2)

$$\text{s.t. } \widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta_h) = y_h(\beta_h, \eta_{h-1}, s_h, \theta')\theta_h - \varphi_h(\beta_h, \eta_{h-1}, s_h, \theta'), \tag{2.4.3}$$

$$\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h) \geqslant \widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta_h), \tag{2.4.4}$$

$$\begin{aligned} \beta_{h+1}(\beta_h, \eta_{h-1}, s_h, \theta_h) &= g_h(\eta_{(1,h-1)}, s_h, \theta_h) \\ &= \beta_h + u_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h) + \overline{U}_h(\beta_h, \eta_{h-1}, s_h, \theta_h) \\ &\quad - \mathbb{E}_{x_{h-1}, s_h, \theta' \,|\, s_{h-1}, y_{h-1}}[(u_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta') + \overline{U}_h(\beta_h, \eta_{h-1}, s_h, \theta'))] \\ &= \beta_h + \widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h) \\ &\quad - \mathbb{E}_{x_{h-1}, s_h, \theta' \,|\, s_{h-1}, y_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta')], \end{aligned} \tag{2.4.5}$$

$$g_h(\eta_{(1,h-1)}, s_h, \theta_h) \geqslant 0. \tag{2.4.6}$$

Table 2.1: Dynamic Programming for Ψ_h when $\mathcal{P}$ is Known

of core ABAMs, provided in Theorem 2.3.9. Indeed, notice the various constraints in Table 2.4.2, where (2.4.4) and (2.4.4) ensures that y_h is a subgradient of g_h and (2.4.5) ensures that g_h satisfies the requisite conditions discussed in Theorem 2.3.9.

Unfortunately, naively solving the program from H to 1 is not computationally efficient, as the β_h and y_{h-1} are both continuous. FPTAS obtained by earlier works such as [Mirrokni et al., 2016a] does not apply, either, as each step's type distribution relies on previous step's allocation rule, making it much more challenging to use piece-wise linear functions to approximate Ψ_{h+1}. It is unclear if we could approximate Ψ_{h+1} sufficiently well using only polynomially many pieces, as its domain now contains two continuous variables. Moreover, as there are multiple possible distributions over types at each step, due to the range of values that the public context s_h may take, it is more challenging to control the propagation of computation errors. Nevertheless, as we show in Theorem 2.4.2, there exists a polynomial-

time algorithm that returns an (additive) ϵ-optimal mechanism, enabled by careful analysis of the properties of Ψ_h.

Theorem 2.4.2. *For any $\epsilon > 0$, there is an algorithm that computes an (additive) ϵ-optimal mechanism in $\widetilde{\mathcal{O}}(\text{poly}(1/\epsilon, N))$ time, with N being the input size of the problem.*

Proof. Our proof is largely comprised of three steps. First, we carefully uncover a crucial property of Ψ_h, showing that it is in fact H-Lipschitz in the allocation level y_{h-1}. Second, we use the property to design a piece-wise linear additive approximation to Ψ_h, using only polynomially many pieces. Third, we use the piece-wise linear approximation to design an efficient algorithm that solves for an ϵ-optimal mechanism.

Step 1. We begin by showing that the function Ψ_h is Lipschitz in y_{h-1}.

Proposition 2.4.3 (Lipschitz)**.** *The function $\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1})$ is H-Lipschitz in y_{h-1} for any $h \in [H]$, $s_{h-1} \in \mathcal{S}$, and $\theta_{h-1} \in \Theta$.*

Proof. Let $\beta_h, s_{h-1}, \theta_{h-1}$, and y_{h-1} be arbitrary and fixed. The feasible region of the program in Table 2.1 is clearly compact due to the linearity of the constraint. The objective function in the program is also continuous, and its derivative in y_{h-1} is clearly continuous by linearity of expectation. As such, by Theorem 2.7.13, a variant of envelope theorem proven by Milgrom and Segal [2002], we have

$$\begin{aligned}
&\frac{\partial}{\partial y_{h-1}} \Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \\
&\quad = \mathbb{E}[y_h^*(\beta_h, \eta_{h-1}, s_h, \theta_h)\theta_h \mid y_{h-1}] \\
&\qquad + \mathbb{E}[\Psi_{h+1}(\beta_{h+1}^*(\beta_h, \eta_{h-1}, s_h, \theta_h), s_h, \theta_h; y_h^*(\beta_h, \eta_{h-1}, s_h, \theta_h) \mid y_{h-1}],
\end{aligned}$$

with y_h^*, β_h^* denoting the solution to the program in Table 2.1 for the specific choice of $\beta_h, s_{h-1}, \theta_{h-1}$, and y_{h-1}. Because $\theta \in \Theta \subseteq [0, 1]$, it is bounded by 1, and therefore Ψ_{h+1} is

also bounded. Consequently, we have

$$\begin{aligned}
&\left|\frac{\partial}{\partial y_{h-1}}\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1})\right| \\
&\quad = \left|\mathbb{E}[y_h^*(\beta_h, \eta_{h-1}, s_h, \theta_h)\theta_h \,|\, y_{h-1}]\right| \\
&\qquad + \left|\mathbb{E}[\Psi_{h+1}(\beta_{h+1}^*(\beta_h, \eta_{h-1}, s_h, \theta_h, s_h, \theta_h; y_h^*(\beta_h, \eta_{h-1}, s_h, \theta_h)) \,|\, y_{h-1}]\right| \\
&\quad \leqslant H,
\end{aligned}$$

completing the proof. □

Step 2. Using the previous fact, we then show that the function Ψ_h can be approximated using a piece-wise linear function with polynomially many pieces, despite the fact that it depends on two continuous variables in β_h and y_{h-1}.

Corollary 2.4.4. *For any $h \in [H]$, $s_{h-1} \in \mathcal{S}$, and $\theta_{h-1} \in \Theta$, the function $\Psi_h(\cdot, s_{h-1}, \theta_{h-1}; \cdot)$ can be additively κ-approximated by two functions $\Psi_h^{\triangle}$ and $\Psi_h^{\triangledown}$. More specifically, the following holds for all $(\beta_h, y_{h-1}) \in [0, \sum_{\tau=h}^{H} \max_{s\in\mathcal{S}} \mathbb{E}_{\theta\sim\mathcal{P}_\tau^\Theta(\cdot\,|\,s)}[\theta]] \times [0, 1]$*

$$\begin{aligned}
&\Psi_h^{\triangledown}(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \leqslant \Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \leqslant \Psi_h^{\triangle}(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}), \\
&\Psi_h^{\triangle}(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) - \Psi_h^{\triangledown}(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \leqslant \kappa.
\end{aligned}$$

Moreover, the functions $\Psi_h^{\triangledown}$ and $\Psi_h^{\triangle}$ are piece-wise linear, have at most $\mathcal{O}\left(N^2/\kappa^2\right)$ pieces, and can be constructed using $\mathcal{O}\left(N^2/\kappa^2\right)$ calls to an evaluation oracle for Ψ_h, with N being the input size.

Proof Sketch of Corollary 2.4.4. Detailed proof is deferred to Appendix 2.7.4. Our first step is to show that for any s_{h-1}, θ_{h-1}, and y_{h-1}, there exists a piece-wise linear approximation to $\Psi_h(\cdot, s_{h-1}, \theta_{h-1}; y_{h-1})$ with at most polynomially many pieces, using the technique found in [Mirrokni et al., 2016a]. The specific technique we use is detailed in Lemma 2.7.12 for

completeness.

We emphasize that while the construction of the piece-wise linear approximation directly uses prior results, showing that the conditions of Lemma 2.7.12 hold for all s_{h-1}, θ_{h-1}, and y_{h-1} requires significantly different arguments than those found in [Mirrokni et al., 2016a]. Specifically, in our setting the next-step context's distribution relies on the current step's allocation rule. This makes finding a suitable upper bound for the function $\Psi_h(\cdot, s_{h-1}, \theta_{h-1}; y_{h-1})$ more challenging: we need to control the impact of y_h on both Ψ_{h+1} itself and the distribution over s_{h+1} it induces. As a result, in Appendix 2.7.4 we feature a series of algebraic manipulations that differ significantly from those found in [Mirrokni et al., 2016a] in order to ensure that their approximation scheme remains valid.

Having shown that Ψ_h is H-Lipschitz in y_{h-1} (Proposition 2.4.3) and that there exists piece-wise linear approximations to $\Psi_h(\cdot, s_{h-1}, \theta_{h-1}; y_{h-1})$ for all possible values of y_{h-1}, the construction of the approximation scheme is straightforward: we simply find $\mathcal{O}(H/\epsilon)$ points over the interval y_{h-1}, and then find a piece-wise linear approximation of $\Psi_h(\cdot, s_{h-1}, \theta_{h-1}; y_{h-1})$ for each y_{h-1} on the grid. Combining Lemma 2.7.12 with Proposition 2.4.3 completes the proof. □

Step 3. Finally, we show that we can use the piece-wise linear approximation discussed in Corollary 2.4.4 to design an efficient algorithm. Specifically, we show that the approximation error in Ψ_h can be controlled via induction. Let $\Psi^{\bigtriangledown}_{h+1}, \Psi^{\bigtriangleup}_{h+1}$ be an additive κ-approximation of Ψ_{h+1}. With a slight abuse of notation, let $\Psi_h(\cdot, s_{h-1}, \theta_{h-1}; \cdot, \bigtriangledown)$, $\Psi_h(\cdot, s_{h-1}, \theta_{h-1}; \cdot, \bigtriangleup)$ denote the solutions to the optimization program in Algorithm 2.1 when all Ψ_{h+1}'s are replaced by their respective additive κ-approximations. For all $\beta_h, s_{h-1}, \theta_{h-1}$, and y_{h-1}, we have

$$\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}, \bigtriangledown) \leqslant \Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \leqslant \Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}, \bigtriangleup),$$

and

$$\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) - \Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}, \bigtriangledown) \leqslant \kappa,$$
$$\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}, \bigtriangleup) - \Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \leqslant \kappa.$$

By Corollary 2.4.4, a polynomially-sized piece-wise linear additive 2κ-approximation can also be constructed for Ψ_h: rather than directly querying $\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1})$, we instead solve for $\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}, \bigtriangledown)$, and use it to construct the piece-wise linear approximation. The corresponding upper bound is constructed similarly.

Setting $\kappa = \epsilon/2H$ and inducting from $h = H + 1$ to 1, noting that Ψ_{H+1} itself is already linear, shows that we can construct a polynomially-sized piece-wise linear additive ϵ-approximation to Ψ_1 in $\mathcal{O}(\text{poly}(1/\epsilon, N))$ time, with N being the problem's input size. Recalling the correctness of the dynamic program from Lemma 2.4.1 completes the proof.

□

We remark that the generalized version of Envelope theorem is crucial to the proof: typically, Envelope theorem requires the solution of the program to be continuously differentiable in the variable of interest, y_{h-1}. Showing the property holds for the one in Table 2.4.2 requires a precise characterization of its optimal value's relationship to y_{h-1}, which has not been characterized even in the easier independent valuation setting discussed in earlier works [Ashlagi et al., 2023, Mirrokni et al., 2016b].

2.5 Pricing Unknown Buyers via Reinforcement Learning

We first describe our learning setup. Assume that the seller only knows the context space $\mathcal{S}$, the type space Θ, as well as the horizon H. The transition probabilities $\mathcal{P}^{\mathcal{S}}$ and $\mathcal{P}^{\Theta}$ are unknown and need to be recovered from repeatedly interacting with the environment over a number of episodes. The interaction can be described as follows.

1. A new buyer arrives at the beginning of each episode and stays until the end of the episode.

2. For $h = 1, \ldots, H$, the seller executes some dynamic mechanism, records the public context, reported private type, and realized allocation at each step.

3. The buyer leaves. A new episode begins.

Recalling Definition 2.2.1, we know that ensuring IC in a dynamic mechanisms requires that the improvement in continuation utility is bounded for any potentially untruthful bidding policy. Unfortunately, it is in general impossible to satisfy the constraint exactly when the transition probabilities $\mathcal{P}$ are unknown, as estimation error in transition probabilities preclude exactly estimating $\bar{U}$. As such, we define the following notion of approximate IC.

Definition 2.5.1 (Approximate Dynamic Incentive Compatibility)**.** *For any mechanism, we say it is ζ-approximately IC at step h for context s if it satisfies for all earlier reported history $\widehat{\eta}_{(1,h-1)}$, reported type $\widehat{\theta}_h$, actual type θ_h, and future bidding policy $b_{(h+1,H)}$ that*

$$\begin{aligned} &u_h(\widehat{\eta}_{(1,h-1)}, s, \widehat{\theta}_h; \theta_h) + \overline{U}_h^{b_{(h+1,H)}}(\widehat{\eta}_{(1,h-1)}, s, \widehat{\theta}_h) \\ &\quad - (u_h(\widehat{\eta}_{(1,h-1)}, s_h, \theta_h; \theta_h) + \overline{U}_h(\widehat{\eta}_{(1,h-1)}, s_h, \theta_h)) \leqslant \zeta. \end{aligned}$$

As we will show in the sequel, the estimation error is manifested in errors in the spend rule. While Lemma 2.3.3 seemingly implies that an ABAM is ex-post IR as long as the amount deposited is no greater than the per-step utility, an amount that can be exactly evaluated when allocation and per-step payment rules are given, the result implicitly relies on the fact that the spend rules are accurately calculated, as statistical errors may cause the spend rule to exceed the current balance. As a result, we also propose the following relaxation of ex-post IR.

Definition 2.5.2 (Approximate Ex-post Individual Rationality)**.** *For any mechanism, we say it is ζ-approximately ex-post IR for the history $\eta_{(1,H)}$ if*

$$-\sum_{h=1}^{H} u_h(\eta_{(1,h-1}, s_h, \theta_h; \theta_h)) \leqslant \zeta.$$

The seller's objective is to output an ϵ-optimal mechanism with probability at least $1-\delta$, where $\epsilon > 0, \delta \in (0,1)$, using as few interactions with the environment as possible. Moreover, the mechanism needs to be approximately IC and approximately ex-post IR.

Unlike earlier RL literature, even those combining dynamic mechanism design with RL, the near-optimal mechanism being learned in our setting is *non-Markovian*. Indeed, while Lemma 2.3.8 seems to convert all possible non-Markovian mechanisms into a Markovian core ABAM, we emphasize that the balance $\texttt{bal}_h$ is a quantity that depends on $\eta_{(1,h-1)}$. Via balance, core ABAMs are non-Markovian, which explains its generality depicted in Lemma 2.3.8, but causes these mechanisms to be much harder to learn.

In addition, one major challenges to learning the near-optimal mechanism is the approximate IC guarantee. By Lemma 2.3.9, we know that obtaining sufficiently good guarantees on IC requires sufficiently good estimates of the distribution over s_h, θ_h conditioned on the previous step's realized allocation x_{h-1} and public context s_{h-1}. Such guarantees are, however, impossible, unless we can guarantee that the context s_{h-1} and the allocation x_{h-1} are covered by the data-collecting algorithm. Moreover, to ensure IC, we need to ensure that the continuation utilities $\bar{U}$ cannot be improved significantly by *any* bidding policy, even including those that are non-Markovian in nature.

To resolve the challenges, we propose Algorithm 1, utilizing a reward-free exploration procedure to carefully generate a dataset that, loosely speaking, covers well all public contexts that can be reached. The dataset is then used to construct empirical estimates of the transition probabilities. Particularly, for the collected dataset $\mathcal{D}$, we first calculate the

Algorithm 1 Reinforcement Learning via Reward-Free Exploration

Input: Accuracy $\epsilon > 0$, failure probability $\delta \in (0,1)$.

1: Collect a set of $\widetilde{\mathcal{O}}(H^5|\mathcal{S}|^2/\epsilon^2)$ trajectories $\mathcal{D} = \{(s_h, \theta_h, x_h, s_{h+1})\}$ via REWARD-FREE RL-EXPLORE, using only type-agnostic policies to interact with the environment for up to $\widetilde{\mathcal{O}}(H^5|\mathcal{S}|^2/\epsilon^2 + H^7|\mathcal{S}|^4/\epsilon)$ episodes. Do not charge the buyers anything.

2: Update empirical counts using (2.5.1) for all $(s, \theta, x, s', h) \in \mathcal{S} \times \Theta \times \{0,1\} \times \mathcal{S} \times [H]$.

3: **for** all $(s, \theta, x, s', h) \in \mathcal{S} \times \Theta \times \{0,1\} \times \mathcal{S} \times [H]$ **do**

4: $\quad \widehat{\mathcal{P}}_h^{\mathcal{S}}(s'|s,x) \leftarrow \frac{N_h(s,x,s')}{N_h(s,x)}$ if $N_h(s,x) > 0$, else $\widehat{\mathcal{P}}_h^{\mathcal{S}}(\cdot|s,x) \leftarrow \frac{1}{|\mathcal{S}|}$.

5: $\quad \widehat{\mathcal{P}}_h^{\Theta}(\theta|s) \leftarrow \frac{N_h(s,\theta)}{N_h(s)}$ if $N_h(s) > 0$, else $\widehat{\mathcal{P}}_h^{\Theta}(\cdot|s) \leftarrow \frac{1}{|\Theta|}$.

6: **end for**

7: Solve the relaxed program in Table 2.2 using $\widehat{\mathcal{P}}_h = \{\widehat{\mathcal{P}}_h^{\mathcal{S}}, \widehat{\mathcal{P}}_h^{\Theta}\}$ as transition probabilities.

Output: Estimated mechanism $\widehat{\mathcal{B}} = \mathcal{B}^{\widehat{g},\widehat{y},\widehat{\varphi}}$.

number of occurrences for all (s, θ, x, s', h) as follows

$$
\begin{aligned}
&N_h(s,\theta,x,s') \leftarrow \sum_{(s_h,\theta_h,x_h,s_{h+1})\in\mathcal{D}} \mathbb{1}\{s_h = s, \theta_h = \theta, x_h = x, s_{h+1} = s'\} \\
&N_h(s,x,s') \leftarrow \sum_{\theta\in\Theta} N_h(s,\theta,x,s'), \qquad N_h(s,x) \leftarrow \sum_{s'\in\mathcal{S}} N_h(s,x,s') \\
&N_h(s,\theta) \leftarrow \sum_{x\in\{0,1\},s'\in\mathcal{S}} N_h(s,\theta,x,s'), \qquad N_h(s) \leftarrow \sum_{x\in\{0,1\},s'\in\mathcal{S}} N_h(s,\theta,x,s').
\end{aligned}
\tag{2.5.1}
$$

These counts are then used to construct the estimated transition probabilities $\widehat{\mathcal{P}}_h^{\mathcal{S}}$ and $\widehat{\mathcal{P}}_h^{\Theta}$. We let $\widehat{\mathcal{P}} = \{(\widehat{\mathcal{P}}_h^{\mathcal{S}}, \widehat{\mathcal{P}}_h^{\Theta})\}_{h=1}^H$ for convenience and, from now on, use $\widehat{\mathbb{E}}$ to denote expectation taken over the estimated probabilities. Inspired by the "optimism in the face of uncertainty" principle in RL, we feature a relaxed dynamic program for calculating Ψ_h, given in Table 2.2.

The defining feature of the relaxed program is the relaxation of the spend rule estimate, constructed in (2.5.2). Observe that there are two places in the program in Table 2.2 that uses the empirical estimate: the objective function, and the expected per-step utility $\widehat{\mathbb{E}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta')]$. Controlling the effect that estimation error has on the objective function is relatively easy, and we focus on how Table 2.2 controls for the estimation error

$$
\begin{aligned}
\max_{\widehat{y}_h, \widehat{\varphi}_h, \widehat{\sigma}_h} \; & \widehat{\mathbb{E}}_{x_{h-1}, s_h, \theta_h \,|\, \widehat{y}_{h-1}} \Big[\widehat{y}_h(\beta_h, \eta_{h-1}, s_h, \theta_h)\theta_h \\
& \qquad + \Psi_{h+1}(\beta_{h+1}(\beta_h, \eta_{h-1}, s_h, \theta_h), s_h, \theta_h; \widehat{y}_h(\beta_h, \eta_{h-1}, s_h, \theta_h)) \Big] \\
\text{s.t. } & \widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta_h) = \widehat{y}_h(\beta_h, \eta_{h-1}, s_h, \theta')\theta_h - \widehat{\varphi}_h(\beta_h, \eta_{h-1}, s_h, \theta'), \\
& \widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h) \geqslant \widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta_h), \\
& \left| \widehat{\sigma}_h(\beta_h, \eta_{h-1}, s_h) - \widehat{\mathbb{E}}_{x_{h-1}, s_h, \theta' \,|\, s_{h-1}, \widehat{y}_{h-1}} [\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta')] \right| \\
& \qquad \leqslant cH|\mathcal{S}||\Theta| \sqrt{\log(cH|\mathcal{S}||\Theta|)/\delta} \\
& \qquad\quad \times ((N_{h-1}(s_{h-1}, 0))^{-1/2} + (N_{h-1}(s_{h-1}, 1))^{-1/2}), \\
& \beta_{h+1}(\beta_h, \eta_{h-1}, s_h, \theta_h) = \widehat{g}_h(\eta_{(1,h-1)}, s_h, \theta_h) \\
& \qquad = \beta_h + \widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h) - \widehat{\sigma}_h(\beta_h, \eta_{h-1}, s_h), \\
& \widehat{g}_h(\eta_{(1,h-1)}, s_h, \theta_h) \\
& \qquad \geqslant -cH|\mathcal{S}||\Theta| \sqrt{\log(cH|\mathcal{S}||\Theta|)/\delta} \\
& \qquad\quad \times ((N_{h-1}(s_{h-1}, 0))^{-1/2} + (N_{h-1}(s_{h-1}, 1))^{-1/2}).
\end{aligned}
\tag{2.5.2}
$$

Table 2.2: Relaxed Program for Ψ_h when $\mathcal{P}$ is Learned

in $\widehat{\mathbb{E}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta')]$.

As the spend rule is now estimated via $\widehat{\mathbb{E}}$, the expectation taken over the empirical transition probability estimates $\widehat{\mathcal{P}}$, we allow the spend rule estimate to deviate from the estimated expected per-step utility, in order to ensure that valid ABAMs are feasible despite the estimation errors. Such relaxation leverages the "optimism in the face of uncertainty" principle (see for instance [Neu and Pike-Burke, 2020]), as we can ensure that while the estimated mechanism $\widehat{\mathcal{B}}$ may not be exactly IC, it achieves higher estimated revenue than the optimal mechanism itself, as the optimal mechanism is feasible under the relaxed program. Of course, under the relaxation, we are no longer directly solving the program over (g, y), but over (y, φ, σ), i.e. the allocation, payment, and spend rules. Doing so helps us isolate the estimation error to the estimated spend rule $\widehat{\sigma}_h$ only, which remains the same for different θ_h. By isolating the statistical errors to the spend rule, which is in itself an estimate of expected

utility, we can control the estimation errors in a non-Markovian dynamic mechanism to a term that is "Markovian", depending only how well we estimate the transition probabilities in $\mathcal{P}$. The construction enables us to use the properties of reward-free exploration, explained in greater detail in Appendix 2.7.1, to efficiently recover a near-optimal non-Markovian dynamic mechanism.

Finally, we stress that Algorithm 1 is IC and ex-post IR throughout the learning process. As Reward-Free RL-Explore interacts with the environment only via type-agnostic policies π and the seller does not charge the buyers anything, these buyers' reporting policies affect neither their received allocations nor the amount they pay. As each buyer stays for only one episode, he also has no incentive to alter the type-agnostic policy used during exploration, ensuring that the learning algorithm itself is IC. The fact that the seller does not charge the buyer anything directly ensures that Algorithm 1 is ex-post IR.

With our learning algorithm defined, we then introduce the key result of this section, that is, the guarantee that a near-optimal mechanism can be learned in polynomial-time using only polynomially many samples.

Theorem 2.5.3. *There exists a polynomial-time algorithm such that, for any $\epsilon > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$, the algorithm outputs a dynamic mechanism satisfying the following.*

- *(ϵ-optimal.) Assuming the buyer reports truthfully, the learned mechanism is $(H+1)\epsilon$-optimal.*
- *(Approximate IC.) The learned mechanism is $\mathcal{O}\left(\frac{|\mathcal{S}|^{1/2}|\Theta|\epsilon}{\max_\pi \Pr_h^\pi(s)}\right)$ approximately IC at step h for context s for all $s \in \mathcal{S}, h \in [H]$.*
- *(Approximate ex-post IR.) The learned mechanism is $\mathcal{O}\left(\sum_{h=1}^{H} \frac{|\mathcal{S}|^{1/2}|\Theta|\epsilon}{\max_\pi \Pr_h^\pi(s_h)}\right)$ approximately ex-post IR for any history $\eta_{(1,H)}$, where $s_{(1,H)}$ are the public contexts included in $\eta_{(1,H)}$.*

Moreover, the algorithm requires at most $\widetilde{\mathcal{O}}(H^5|\mathcal{S}|^2/\epsilon^2 + H^7|\mathcal{S}|^4/\epsilon)$ *episodes of interaction.*

Detailed proof of Theorem 2.5.3 is deferred to Appendix 2.7.5. Notably, the approximate IC and approximate ex-post IR guarantees in Theorem 2.5.3 do not depend on either the optimal mechanism or the estimated mechanism. The term $\max_\pi \Pr_h^\pi(s)$ depends only on the underlying ground-truth transition probabilities $\mathcal{P}$, and measures how likely the context s is to be reached under *any possible* type-agnostic policy: as long as the context can be reached by some type-agnostic policy, then $\max_\pi \Pr_h^\pi(s)$ is large, and the IC and ex-post IR violations are small. Recalling Lemma 2.2.3, Theorem 2.5.3 implies that it is either unlikely to reach the context s at step h by *any dynamic mechanism*, or the amount the buyer can gain from *any bidding policy* is small, offering a strong deterrence to untruthfulness. The approximate ex-post IR guarantee complements the result, further showing that truthful buyer will not be much worse off from participating: unless some rare state is reached, the buyer's ex-post utility will not be too low.

2.6 Conclusion

We introduce a novel problem setting where the buyer's valuation distributions may change according to the allocations that he receives. We first derive a concrete family of mechanisms capable of achieving the optimal revenue under the settings. We then show that (additively) ϵ-optimal mechanisms can be calculated efficiently when the seller knows a priori how the buyer's valuation distribution changes according to the allocations she chooses. Leveraging the efficient algorithm for computing the near-optimal mechanism, we further show that such an mechanism can also be learned efficiently in polynomial time and using only polynomially many samples.

We believe our results pave the way for numerous important future directions. For instance, would it be possible to extend our results to the dynamic auction setting, similar to how [Mirrokni et al., 2016b] extends the results in [Mirrokni et al., 2016a]? Although

it is trivial to adapt our results to a no-regret learning algorithm, it it possible to devise a learning algorithm that recovers an *exactly* IC, *exactly* ex-post IR dynamic mechanism, under suitable additional assumptions on the transition probabilities $\mathcal{P}$? Moreover, can we incorporate additional constraints such as buyer budget into consideration? Lastly, again inspired by the AWS spot instance pricing problem [Baughman et al., 2019], the seller may also prefer a more stable dynamic mechanism, in which the average allocation level does not change significantly from step to step. Can our results be extended to this setting as well?

2.7 Technical Details

2.7.1 Detailed Description of Reward-Free RL-Explore

In this section we discuss the key intuition behind the Reward-Free RL-Explore procedure introduced by Jin et al. [2020a] and the theoretical properties of the algorithm that we use in order to prove Theorem 2.5.3. For brevity, we only provide a high-level description of the procedure as well as some intuitive arguments for why the algorithm is valid. We refer interested readers to the original paper for an in-depth description of the algorithm. Reward-Free RL-Explore utilizes a two-stage approach and can be described as follows.

1. For each h, s, perform the following procedure. Let the reward function $r_h(s_h, \cdot) = 1$ if $s_h = s$, and set the reward to 0 for all other h and contexts. Run an RL algorithm for N_0 episodes to find a policy that approximately maximizes the reward. Record the policy for all h, s and store them in a policy set.

2. Collect and return a set of N trajectories obtained by by first sampling one policy from the set of policies given by the first step, and then executing the policy.

Intuitively, if $\max_\pi \Pr_h^\pi(s)$ is sufficiently large, that is, if there is some policy that can reach the public context s in step h with sufficiently high probability, then step 1 of the algorithm

will also find a policy that reaches s at step h with sufficiently high probability. By executing these policies uniformly at random in step 2, the sampling distribution of $\mathcal{D}$, the dataset returned by the algorithm, covers well all (s, h) that can be reached by any policy. We formalize the intuition as follows.

Theorem 2.7.1 (Restatement of Theorem 3.3 in [Jin et al., 2020a]). *There exists absolute constant $c > 0$ such that for any $\epsilon > 0$ and $\delta \in (0, 1)$, if $N_0 \geqslant c|\mathcal{S}|^3 H^6 (\log(c|\mathcal{S}|^2 H^3/(\epsilon\delta)))^3/\epsilon$, then with probability at least $1 - \delta$,* Reward-Free RL-Explore *will return a dataset $\mathcal{D}$ consisting of N trajectories which are i.i.d. sampled from a distribution $\mu_{(1,H)} \in \Delta(\mathcal{S} \times \Theta \times \mathcal{X})^H$ satisfying*

$$\forall\, s, h \text{ where } \max_{\pi} \Pr_h^{\pi}(s) \geqslant \frac{\epsilon}{2|\mathcal{S}|H^2},\ \text{we have } \max_{x,\pi} \frac{\Pr_h^{\pi}(s_h, x)}{\mu_h(s_h, x)} \leqslant 4|\mathcal{S}|H. \tag{2.7.1}$$

Proof. The theorem is a direct application of Theorem 3.3 in [Jin et al., 2020a], noting that there are exactly 2 possible actions in our problem setting. □

We then introduce the concept of V-functions as follows. For any bounded function $f : \{\mathcal{S} \times \{0, 1\} \to [0, 1]\}^H$, which maps the public context and received allocation at step h to a number in $[0, 1]$, and type-agnostic policy $\pi : \{\mathcal{S} \to \Delta(\{0, 1\})\}^H$, we define the V-function as

$$V_h^{\pi,f}(s) = \mathbb{E}_{s_{(h,H)}, x_{(h,H)} \,|\, \pi} \left[\sum_{h'=h}^{H} f_{h'}(s_{h'}, x_{h'}) \,|\, s_h = s \right].$$

Additionally, our results will also use the following result on evaluation of V-functions.

Lemma 2.7.2 (Restatement of Lemma 3.6 in [Jin et al., 2020a]). *There exists absolute constant $c > 0$, for any $\epsilon > 0, \delta \in (0, 1)$, assume dataset $\mathcal{D}$ has N i.i.d. samples from distribution $\mu_{(1,H)}$ which satisfies (2.7.1), if $N \geqslant cH^5|\mathcal{S}|^2 \log(c|\mathcal{S}|H/(\delta\epsilon))/\epsilon^2$, then with probability at least $1 - \delta$, for any bounded function $f : \{\mathcal{S} \times \{0, 1\} \to [0, 1]\}^H$ and any type*

agnostic policy π, we have

$$|V_1^{\pi,f}(s_1) - \widehat{V}_1^{\pi,f}(s_1)| \leqslant \epsilon,$$

where $\widehat{V}$ is the estimated value function under transition probability $\widehat{\mathcal{P}}^{\mathcal{S}}$.

Proof. The lemma, again, is by direct application of Lemma 3.6 in [Jin et al., 2020a], noting that there are exactly 2 possible actions in our problem setting. □

2.7.2 Omitted Proofs in Section 2.2

We include below the full proofs of the statements in 2.2.

Proof of Lemma 2.2.2

As IC takes into consideration all possible bidding strategies, reporting truthfully in the ensuing steps is also covered by the definition. Therefore, IC implies stage IC.

We then show that stage IC implies IC. Let $\mathcal{M}$ denote an arbitrary mechanism that is stage IC. Let b denote an arbitrary and fixed reporting strategy and let h be arbitrary and fixed. By (2.2.2), the definition of stage IC, and taking the expectation over θ_{h+1}, we know that for any history of length h, $(\widehat{\eta}_{(1,h)})$, and any context at step $h+1$, s_{h+1}, we have

$$\begin{aligned}
&\mathbb{E}_{\theta_{h+1}}\left[u_{h+1}(\widehat{\eta}_{(1,h)}, s_{h+1}, \theta_{h+1}; \theta_{h+1}) + \overline{U}_{h+1}(\widehat{\eta}_{(1,h)}, s_{h+1}, \theta_{h+1})\right] \\
&\quad \geqslant \mathbb{E}_{\theta_{h+1}}[u_{h+1}(\widehat{\eta}_{(1,h)}, s_{h+1}, b_{h+1}(\widehat{\eta}_{(1,h)}, s_{h+1}, \theta_{h+1}); \theta_{h+1}) \\
&\qquad\qquad + \overline{U}_{h+1}(\widehat{\eta}_{(1,h)}, s_{h+1}, b_{h+1}(\widehat{\eta}_{(1,h)}, s_{h+1}, \theta_{h+1}))].
\end{aligned} \tag{2.7.2}$$

Equivalently, for all histories $(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h)$, taking the expectation over x_h and realizations of s_{h+1} shows that

$$\overline{U}_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h) \geqslant \overline{U}_h^{b_{h+1}}(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h),$$

where with a slight abuse of notation we let b_{h+1} denote the bidding strategy where the buyer bids according to b at the $(h+1)$-th step and then bids truthfully in the remaining steps. Repeat the same argument but replacing step $(h+1)$ with $(h+2)$, and integrating over θ_{h+2} instead, we know that

$$\overline{U}_h^{b_{h+1}}(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h) \geqslant \overline{U}_h^{b_{(h+1,h+2)}}(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h).$$

Repeating the argument until we reach the final step H shows that

$$\overline{U}_h^{b_{h+1}}(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h) \geqslant \overline{U}_h^{b_{(h+1,H)}}(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h).$$

In other words, stage IC implies the following inequality by applying the stage IC definition for steps $h+1, \ldots, H$

$$\begin{aligned}
&u_h(\widehat{\eta}_{(1,h-1)}, s_h, \theta_h; \theta_h) + \overline{U}_h(\widehat{\eta}_{(1,h-1)}, s_h, \theta_h) \\
&\quad \geqslant u_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h; \theta_h) + \overline{U}_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h) \\
&\quad \geqslant u_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h; \theta_h) + \overline{U}_h^{b_{h+1}}(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h) \\
&\qquad\qquad \vdots \\
&\quad \geqslant u_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h; \theta_h) + \overline{U}_h^{b_{(h+1,H)}}(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h),
\end{aligned}$$

for any reported type at step h, $\widehat{\theta}_h$. The result is then equivalent to Definition 2.2.1, concluding the proof. □

Proof of Lemma 2.2.3

We prove the claim by construction. For all h, s let

$$\pi_h(s) = \mathbb{E}_{\widehat{\eta}_{(1,H)} \sim b_{(1,H)}}[\chi_h(\widehat{\eta}_{(1,h-1)}, s_h, \widehat{\theta}_h) \,|\, s_h = s].$$

By definition of conditional random variables, the expression on the right hand side is a valid function of s for each h. Additionally, as $\chi_{(1,H)}$'s range is $\Delta(\mathcal{X})$, the range of the term on the right hand side is also $\Delta(\mathcal{X})$, ensuring that for all h, π_h is a function mapping $\mathcal{S}$ to $\Delta(\mathcal{X})$.

We then show the equation holds by induction from $h = 1$ to H. The base case trivially holds as there is no history prior to s_1. We then show that the marginal distribution over (s_{h-1}, x_{h-1}) is the same under both π and the distribution induced by $\chi_{(1,H)}, b_{(1,H)}$, assuming the marginal distribution over s_{h-1} is the same. By construction

$$\pi_{h-1}(s_{h-1}) = \mathbb{E}_{\widehat{\eta}_{(1,H)} \sim b_{(1,H)}}[\chi_{h-1}(\widehat{\eta}_{(1,h-2)}, s_{h-1}, \widehat{\theta}_{h-1}) \,|\, s_{h-1} = s_{h-1}],$$

therefore when conditioned on any s_{h-1}, the distribution over x_{h-1} is the same. By inductive hypothesis, the marginal distribution over (s_{h-1}, x_{h-1}) is also the same. The definition of the transition kernel $\mathcal{P}^{\mathcal{S}}_{h-1}$ completes the proof. □

2.7.3 Omitted Proofs in Section 2.3

Throughout this section, we use the following shorthand notations for convenience

$$\begin{aligned}
&\texttt{bal}_\tau = \texttt{bal}_\tau(\eta_{(1,\tau-2)}, s_{\tau-1}, \theta_{\tau-1}),\ \texttt{bal}^{(h)}_\tau = \texttt{bal}_\tau(\eta^{(h)}_{(1,\tau-2)}, s^{(h)}_{\tau-1}, \theta^{(h)}_{\tau-1}),\\
&\sigma_\tau = \sigma_\tau(\texttt{bal}_\tau, \eta_{\tau-1}, s_\tau),\ \sigma^{(h)}_\tau = \sigma_\tau(\texttt{bal}^{(h)}_\tau, \eta^{(h)}_{\tau-1}, s^{(h)}_\tau),
\end{aligned}$$

where $\theta^{(h)}_\tau$ has the same distribution as θ_τ for all $\tau < h$, equals to a potential untruthful report $\widehat{\theta}_h$ when $\tau = h$, and has the same distribution as that of the private type at step τ if the report at step h were to be changed to some arbitrary $\widehat{\theta}_h$ instead. Additionally, $\eta^{(h)}_\tau$ denotes the history at step h if the reported type at step h were changed instead and $s^{(h)}_\tau$ is similarly defined.

With shorthand notations defined, we proceed with the proofs.

Proof of Lemma 2.3.2

Let $\mathcal{B}$ be an arbitrary and fixed ABAM satisfying the conditions in Lemma 2.3.2. By Lemma 2.2.2, we know that it suffices to show that any ABAM satisfying the desired properties also satisfy stage IC. By construction of the per-step payment rule of ABAMs, we have

$$u_h(\eta_{(1,h-1)}, s_h, \theta_h; \theta_h) = \widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) - \sigma_h$$

for all $h \in [H]$. The bidder's instantaneous and continuation utility if he were to report $\widehat{\theta}_h$ instead at step h, but truthful from step 1 to $h-1$, is

$$\begin{aligned}
&u_h(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h; \theta_h) + \overline{U}_h(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h) \\
&\quad = \widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \widehat{\theta}_h; \theta_h) - \sigma_h \\
&\qquad + \mathbb{E}_{x_h^{(h)}, \eta_{(h+1,H)}^{(h)}} \left[\sum_{\tau=h+1}^{H} (\widehat{u}_\tau(\texttt{bal}_\tau^{(h)}, \eta_{\tau-1}^{(h)}, s_\tau^{(h)}, \theta_\tau^{(h)}; \theta_\tau^{(h)}) - \sigma_\tau^{(h)}) \right] \\
&\quad = \widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \widehat{\theta}_h; \theta_h) - \sigma_h \\
&\qquad + \sum_{\tau=h+1}^{H} \mathbb{E}_{x_h^{(h)}, \eta_{(h+1,\tau-1)}^{(h)}, s_\tau^{(h)}, \theta_\tau^{(h)}} \left[\widehat{u}_\tau(\texttt{bal}_\tau^{(h)}, \eta_{\tau-1}^{(h)}, s_\tau^{(h)}, \theta_\tau^{(h)}; \theta_\tau^{(h)}) - \sigma_\tau^{(h)} \right].
\end{aligned}$$

Similarly, if the buyer were to report truthfully at step h, his instantaneous and continuation utility sum to

$$\begin{aligned}
&u_h(\eta_{(1,h-1)}, s_h, \theta_h; \theta_h) + \overline{U}_h(\eta_{(1,h-1)}, s_h, \theta_h) \\
&\quad = \widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) - \sigma_h \\
&\qquad + \mathbb{E}_{x_h, \eta_{(h+1,H)}} \left[\sum_{\tau=h+1}^{H} \widehat{u}_\tau(\texttt{bal}_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau; \theta_\tau) - \sigma_\tau \right] \\
&\quad = \widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) - \sigma_h \\
&\qquad + \sum_{\tau=h+1}^{H} \mathbb{E}_{x_h, \eta_{(h+1,\tau-1)}, s_\tau, \theta_\tau} \left[\widehat{u}_\tau(\texttt{bal}_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau; \theta_\tau) - \sigma_\tau \right].
\end{aligned}$$

The difference between the two is

$$\begin{aligned}
&u_h(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h; \theta_h) + \overline{U}_h(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h) \\
&\qquad - (u_h(\eta_{(1,h-1)}, s_h, \theta_h; \theta_h) + \overline{U}_h(\eta_{(1,h-1)}, s_h, \theta_h)) \\
&\quad = \widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \widehat{\theta}_h; \theta_h) - \sigma_h - (\widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) - \sigma_h) \\
&\qquad + \sum_{\tau=h+1}^{H} \Bigg(\mathbb{E}_{x_h^{(h)}, \eta_{(h+1,\tau-1)}^{(h)}, s_\tau^{(h)}, \theta_\tau^{(h)}} \left[\widehat{u}_\tau(\texttt{bal}_\tau^{(h)}, \eta_{\tau-1}^{(h)}, s_\tau^{(h)}, \theta_\tau^{(h)}; \theta_\tau^{(h)}) - \sigma_\tau^{(h)} \right] \\
&\qquad\qquad - \mathbb{E}_{x_h, \eta_{(h+1,\tau-1)}, s_\tau, \theta_\tau} \left[\widehat{u}_\tau(\texttt{bal}_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau; \theta_\tau) - \sigma_\tau \right] \Bigg).
\end{aligned}$$

We know by the conditions for $\widehat{u}$ that

$$\widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \widehat{\theta}_h; \theta_h) \leqslant \widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h).$$

For the second difference term, consider an arbitrary and fixed $\tau \geqslant h+1$ and the joint distribution over $(x_h^{(h)}, \eta_{(h+1,\tau-1)}^{(h)}, s_{\tau-1}^{(h)}, \theta_{\tau-1}^{(h)})$ and $(x_h, \eta_{(h+1,\tau-2)}, s_{\tau-1}, \theta_{\tau-1})$. Due to the Markovian transition kernel, the pair is independent conditioned on $(\eta_{(1,h-1)}, s_h, \theta_h, \widehat{\theta}_h)$ and

we know

$$\begin{aligned}
&\mathbb{E}_{x_h^{(h)},\eta_{(h+1,\tau-1)}^{(h)},s_\tau^{(h)},\theta_\tau^{(h)}}\left[\widehat{u}_\tau(\texttt{bal}_\tau^{(h)},\eta_{\tau-1}^{(h)},s_\tau^{(h)},\theta_\tau^{(h)};\theta_\tau^{(h)})-\sigma_\tau^{(h)}\right]\\
&\quad-\mathbb{E}_{x_h,\eta_{(h+1,\tau-1)},s_\tau,\theta_\tau}\left[\widehat{u}_\tau(\texttt{bal}_\tau,\eta_{\tau-1},s_\tau,\theta_\tau;\theta_\tau)-\sigma_\tau\right]\\
&=\mathbb{E}\Bigg[\mathbb{E}_{x_{\tau-1}^{(h)},s_\tau^{(h)},\theta_\tau^{(h)}|s_{\tau-1}^{(h)},\theta_{\tau-1}^{(h)}}\left[\widehat{u}_\tau(\texttt{bal}_\tau^{(h)},\eta_{\tau-1}^{(h)},s_\tau^{(h)},\theta_\tau^{(h)};\theta_\tau^{(h)})-\sigma_\tau^{(h)}\right]\\
&\qquad\qquad-\mathbb{E}_{x_{\tau-1},s_\tau,\theta_\tau|s_{\tau-1},\theta_{\tau-1}}\left[\widehat{u}_\tau(\texttt{bal}_\tau,\eta_{\tau-1},s_\tau,\theta_\tau;\theta_\tau)-\sigma_\tau\right]\Bigg],
\end{aligned}$$

where the expectation on the outside is taken with respect to the joint distribution over the pair $(x_h^{(h)},\eta_{(h+1,\tau-2)}^{(h)},s_{\tau-1}^{(h)},\theta_{\tau-1}^{(h)})$ and $(x_h,\eta_{(h+1,\tau-2)},s_{\tau-1},\theta_{\tau-1})$. For the equation, we used the fact that the transition dynamics of the buyer's private type is Markovian and thus the distribution over $x_{\tau-1}^{(h)},s_\tau^{(h)},\theta_\tau^{(h)}$ depend only on $s_{\tau-1}^{(h)},\theta_{\tau-1}^{(h)}$ (and the same holds for the distribution when the buyer is truthful). We know that for any $(x_h^{(h)},\eta_{(h+1,\tau-1)}^{(h)},s_{\tau-1}^{(h)},\theta_{\tau-1}^{(h)})$ and $(x_h,\eta_{(h+1,\tau-2)},s_{\tau-1},\theta_{\tau-1})$

$$\begin{aligned}
&\mathbb{E}_{x_h^{(h)},\eta_{(h+1,\tau-1)}^{(h)},s_\tau^{(h)},\theta_\tau^{(h)}}\left[\widehat{u}_\tau(\texttt{bal}_\tau^{(h)},\eta_{\tau-1}^{(h)},s_\tau^{(h)},\theta_\tau^{(h)};\theta_\tau^{(h)})-\sigma_\tau^{(h)}\right]\\
&\quad-\mathbb{E}_{x_h,\eta_{(h+1,\tau-1)},s_\tau,\theta_\tau}\left[\widehat{u}_\tau(\texttt{bal}_\tau,\eta_{\tau-1},s_\tau,\theta_\tau;\theta_\tau)-\sigma_\tau\right]\\
&=\mathbb{E}_{x_{\tau-1}^{(h)},s_\tau^{(h)},\theta_\tau^{(h)}|s_{\tau-1}^{(h)},\theta_{\tau-1}^{(h)}}[\sigma_\tau^{(h)}]-\mathbb{E}_{x_{\tau-1},s_\tau,\theta_\tau|s_{\tau-1},\theta_{\tau-1}}[\sigma_\tau]\\
&\quad-\Bigg(\mathbb{E}_{x_{\tau-1}^{(h)},s_\tau^{(h)},\theta_\tau^{(h)}|s_{\tau-1}^{(h)},\theta_{\tau-1}^{(h)}}\left[\widehat{u}_\tau(\texttt{bal}_\tau^{(h)},\eta_{\tau-1}^{(h)},s_\tau^{(h)},\theta_\tau^{(h)};\theta_\tau^{(h)})-\sigma_\tau^{(h)}\right]\\
&\qquad\qquad-\mathbb{E}_{x_{\tau-1},s_\tau,\theta_\tau|s_{\tau-1},\theta_{\tau-1}}\left[\widehat{u}_\tau(\texttt{bal}_\tau,\eta_{\tau-1},s_\tau,\theta_\tau;\theta_\tau)-\sigma_\tau\right]\Bigg).
\end{aligned}$$

Recalling the statement of Lemma 2.3.2

$$\begin{aligned}&\mathbb{E}_{x_h^{(h)},\eta_{(h+1,\tau-1)}^{(h)},s_\tau^{(h)},\theta_\tau^{(h)}}\left[\widehat{u}_\tau(\mathtt{bal}_\tau^{(h)},\eta_{\tau-1}^{(h)},s_\tau^{(h)},\theta_\tau^{(h)};\theta_\tau^{(h)})-\sigma_\tau^{(h)}\right]\\&\quad=\mathbb{E}_{x_h,\eta_{(h+1,\tau-1)},s_\tau,\theta_\tau}\left[\widehat{u}_\tau(\mathtt{bal}_\tau,\eta_{\tau-1},s_\tau,\theta_\tau;\theta_\tau)-\sigma_\tau\right]\end{aligned}$$

Applying the equality from $\tau = h+1,\ldots,H$,

$$\overline{U}_h(\eta_{(1,h-1)},s_h,\widehat{\theta}_h)-\overline{U}_h(\eta_{(1,h-1)},s_h,\theta_h)=0,$$

showing

$$\begin{aligned}&u_h(\eta_{(1,h-1)},s_h,\widehat{\theta}_h;\theta_h)+\overline{U}_h(\eta_{(1,h-1)},s_h,\widehat{\theta}_h)\\&\qquad\leqslant(u_h(\eta_{(1,h-1)},s_h,\theta_h;\theta_h)+\overline{U}_h(\eta_{(1,h-1)},s_h,\theta_h)).\end{aligned}$$

The inequality immediately implies stage IC, completing the proof. □

Proof of Lemma 2.3.3

By construction of the update formula for $\mathtt{bal}_h$, we know that

$$\sum_{h=1}^{H}\delta_h(\mathtt{bal}_h,\eta_{h-1},s_h,\theta_h)=\mathtt{bal}_{H+1}-\mathtt{bal}_1+\sum_{h=1}^{H}\sigma_h(\mathtt{bal}_h,\eta_{h-1},s_h).$$

When the condition in the statement of Lemma 2.3.3 holds for all h, summing from $h = 1$ to $h = H$ obtains

$$\begin{aligned}\sum_{h=1}^{H} \widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) &\geqslant \sum_{h=1}^{H} \delta_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) \\ &= \texttt{bal}_{H+1} - \texttt{bal}_1 + \sum_{h=1}^{H} \sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h) \\ &\geqslant \sum_{h=1}^{H} \sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h),\end{aligned}$$

where the inequality uses the fact that $\texttt{bal}_1 = 0$ and that $\texttt{bal}_{H+1} \geqslant 0$. By construction of the payment rule, we know $u_h(\eta_{(1,h-1)}, s_h, \theta_h; \theta_h) = \widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) - \sigma_h$. Therefore,

$$\sum_{h=1}^{H} u_h(\eta_{(1,h-1)}, s_h, \theta_h; \theta_h) = \sum_{h=1}^{H} \widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) - \sum_{h=1}^{H} \sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h) \geqslant 0,$$

which is exactly the definition of ex post IR. □

Proof of Lemma 2.3.6

We first argue that it is without loss of generality to assume that for any $h \in [H-1]$ and $(\eta_{(1,h-1)}, s_h, \theta_h)$ the expected utility is 0, namely

$$u_h(\eta_{(1,h-1)}, s_h, \theta_h; \theta_h) = 0. \tag{2.7.3}$$

For any dynamic mechanism $\mathcal{M} = (\chi, \psi)$, consider the following mechanism

$\mathcal{M}' = (\chi, \psi')$, where

$$\psi'(\eta_{(1,h-1)}, s_h, \theta_h) = \begin{cases} \chi(\eta_{(1,h-1)}, s_h, \theta_h)\theta_h & h \in [H-1], \\ \chi(\eta_{(1,h-1)}, s_h, \theta_h)\theta_h - \textsc{Utl}(\mathcal{M}|\eta_{(1,h-1)}, s_h, \theta_h) & h = H. \end{cases}$$

As the mechanisms $\mathcal{M}'$ and $\mathcal{M}$ have the same allocation policy, the distributions over the space of possible histories are the same between the two mechanisms. Moreover, the two mechanism levy the same amount of payment over the entire episode and would lead to the same episodic utility and revenue.

We then follow the outline in [Mirrokni et al., 2016a] and proceed by inducting on step h from $h = 1$ to $h = H$. For each $h \in [H]$, our goal is to show that it is possible to construct a mechanism that yields the same buyer utility, higher seller revenue, and is symmetric in the first h steps. We begin with the base case.

Base Case. When $h = 1$, the statement is trivially true, as there is no history at the start of each episode and hence the submechanisms are the same.

Inductive Hypothesis. Let $1 \leqslant h \leqslant H$ be arbitrary and fixed. Assume that for any dynamic mechanism $\mathcal{M}$, there exists some mechanism $\mathcal{M}'$ that is symmetric over equivalent histories up to length $h - 1$. Moreover, $\mathcal{M}'$ yields the same expected utility as $\mathcal{M}$ and at least the same amount of seller utility.

Inductive Case. Let $\mathcal{M} = (\chi, \psi)$ denote a mechanism that satisfies the inductive hypothesis, i.e. is symmetric up to equivalent histories of length $h - 1$. Our goal is to design a mechanism that is symmetric up to equivalent histories of length h and yields the same utility and revenue as $\mathcal{M}$.

Step 1: Constructiing the Mechanism. Consider the following construction of $\mathcal{M}' = (\chi', \psi')$

- For all $\tau \leqslant h$, the payment rules and allocation rules of $\mathcal{M}'$ remain the same as those of $\mathcal{M}$, namely for all $\eta_{(1,h-1)}, s_h, \theta_h$

$$\chi'_\tau(\eta_{(1,h-1)}, s_h, \theta_h) = \chi_\tau(\eta_{(1,h-1)}, s_h, \theta_h), \quad \psi'_\tau(\eta_{(1,h-1)}, s_h, \theta_h) = \psi_\tau(\eta_{(1,h-1)}, s_h, \theta_h).$$

We stress that under such construction, for any history $(\eta_{(1,h-1)}, s_h, \theta_h)$, the distribution over x_h is the same for both $\mathcal{M}$ and $\mathcal{M}'$.

- For any $\tau \geqslant h+1$, we first define the concept "representative history" using Definition 2.3.4, partitioning the space of histories based on the equivalence relationship. Particularly,

$$\begin{aligned}\mathcal{H}_{(\eta_{(1,h-1)},s_h,\theta_h)} = \{(\eta'_{(1,h-1)}, s'_h, \theta'_h) \in \mathcal{H}_{h-1} \times \mathcal{S} \times \Theta : \\ \textsc{Utl}(\mathcal{M}|\eta_{(1,h-1)}, s_h, \theta_h) = \textsc{Utl}(\mathcal{M}|\eta'_{(1,h-1)}, s'_h, \theta'_h), \\ s_h = s'_h, \theta_h = \theta'_h\}.\end{aligned}$$

The representative history for each partitioned set is the history with the highest expected seller revenue

$$\begin{aligned}&\eta^*(\mathcal{H}_{(\eta_{(1,h-1)},s_h,\theta_h)}) \\ &\quad = \underset{(\eta'_{(1,h-1)},s'_h,\theta'_h)}{\operatorname{argmax}} \mathbb{E}_{x'_h,\eta_{(h+1,H)}}\left[\sum_{\tau=h+1}^{H} \psi_\tau((\eta'_{(1,h)}, \eta_{(h+1,\tau-1)}), s_\tau, \theta_\tau)\right],\end{aligned}$$

picking among the set of histories a representative with the highest expected future revenue under $\mathcal{M}$[5]. Here with a slight abuse of notation we let $\eta^*(\mathcal{H}_{(\eta_{(1,h-1)},s_h,\theta_h)}) \in \mathcal{H}_{h-1} \times \mathcal{S} \times \Theta$ denote the history, context, type tuple over which the equivalence relationship is defined. We finally formally introduce the allocation policy and payment

5. The maximization procedure here is for simplicity of presentation. Equivalently, we can also pick, for each set of equivalent histories, a unique history that yields expected continuation revenue no less than the average taken over the set of equivalent histories.

rule, where for all $\tau \geqslant h+1$

$$\chi'_\tau(\eta_{(1,\tau-1)}, s_\tau, \theta_\tau) = \chi_\tau((\eta^*(\mathcal{H}_{(\eta_{(1,h-1)}, s_h, \theta_h)}), x_h, \eta_{(h+1,\tau-1)}), s_\tau, \theta_\tau),$$
$$\psi'_\tau(\eta_{(1,\tau-1)}, s_\tau, \theta_\tau) = \psi_\tau((\eta^*(\mathcal{H}_{(\eta_{(1,h-1)}, s_h, \theta_h)}), x_h, \eta_{(h+1,\tau-1)}), s_\tau, \theta_\tau).$$

In other words, the mechanism first takes $(\eta_{(1,h-1)}, s_h, \theta_h)$ and finds an equivalent history with the highest expected continuation revenue (finding the "representative history"). The ensuing mechanism then acts as if first h steps' history (excluding the realized allocation at step h) was instead changed to $\eta^*(\mathcal{H}_{(\eta_{(1,h-1)}, s_h, \theta_h)})$.

Our goal then reduces to showing that the constructed mechanism $\mathcal{M}'$ satisfies the following two properties.

1. $\mathcal{M}'$ is symmetric for any equivalent histories of length at most h, namely for all $\nu \leqslant h$ and two equivalent histories of length ν, $(\eta_{(1,\nu-1)}, s_\nu, \theta_\nu) \sim (\eta'_{(1,\nu-1)}, s'_\nu, \theta'_\nu)$, we have for all $\nu+1 \leqslant \tau \leqslant H$, x_ν, $\eta_{(\nu+1,H)}$, and all s_τ, θ_τ that

$$\begin{cases} \chi'_\tau(\eta_{(1,\tau-1)}, s_\tau, \theta_\tau) = \chi'_\tau((\eta'_{(1,\nu-1)}, \eta_{(\nu,\tau-1)}), s_\tau, \theta_\tau) \\ \psi'_\tau(\eta_{(1,\tau-1)}, s_\tau, \theta_\tau) = \psi'_\tau((\eta'_{(1,\nu-1)}, \eta_{(\nu,\tau-1)}), s_\tau, \theta_\tau) \end{cases},$$

where we recall that $\eta_\nu, (s_\nu, \theta_\nu, x_\nu) = (s'_\nu, \theta'_\nu, x_\nu)$ by definition of equivalence.

2. $\textsc{Utl}(\mathcal{M}') = \textsc{Utl}(\mathcal{M})$ and $\textsc{Rev}(\mathcal{M}') \geqslant \textsc{Rev}(\mathcal{M})$.

For ease of presentation, we prove the two properties in a switched order, beginning with property 2 and ending with property 1. We proceed as follows.

Proof of Property 2. For any $h < H$ and $\eta_{(1,h-1)}, s_h, \theta_h$, by definition of expected episodic

utility UTL, we have

$$\begin{aligned}
&\textsc{Utl}(\mathcal{M}'|\eta_{(1,h-1)}, s_h, \theta_h) \\
&\quad = \mathop{\mathbb{E}}_{x_h, \eta_{(h+1,H)} \sim \mathcal{M}'} \left[\sum_{\tau=1}^{H} u_\tau^{\mathcal{M}'}(\eta_{(1,\tau-1)}, s_\tau, \theta_\tau; \theta_\tau) | \eta_{(1,h-1)}, s_h, \theta_h \right],
\end{aligned}$$

where with a slight abuse of notation let $u^{\mathcal{M}'}$ denote the utility at a particular step under $\mathcal{M}'$. By linearity of expectation and the construction of $\mathcal{M}'$, we know

$$\begin{aligned}
&\textsc{Utl}(\mathcal{M}'|\eta_{(1,h-1)}, s_h, \theta_h) \\
&\quad = \sum_{\tau=1}^{h} u_\tau(\eta_{(1,\tau-1)}, s_\tau, \theta_\tau; \theta_\tau) + \mathop{\mathbb{E}}_{x_h, \eta_{(h+1,H)} \sim \mathcal{M}'} \left[\sum_{\tau=h+1}^{H} u_\tau^{\mathcal{M}'}(\eta_{(1,\tau-1)}, s_\tau, \theta_\tau; \theta_\tau) \right] \\
&\quad = \sum_{\tau=1}^{h} u_\tau(\eta_{(1,\tau-1)}, s_\tau, \theta_\tau; \theta_\tau) + \overline{U}_h^{\mathcal{M}'}(\eta_{(1,h-1)}, s_h, \theta_h),
\end{aligned} \tag{2.7.4}$$

where the first equation comes from the fact that in the first h steps, $\mathcal{M}'$ and $\mathcal{M}$ have the same allocation and payment rules, and for the second line we, similar to $u^{\mathcal{M}'}$, abuse the notation and let $\overline{U}^{\mathcal{M}'}$ denote the expected continuation utility under $\mathcal{M}'$.

Recall that it is without loss of generality to assume that at all steps, save for the H-th, the utility is exactly zero. We can simplify (2.7.4) as

$$\textsc{Utl}(\mathcal{M}'|\eta_{(1,h-1)}, s_h, \theta_h) = \overline{U}_h^{\mathcal{M}'}(\eta_{(1,h-1)}, s_h, \theta_h).$$

By construction of $\mathcal{M}'$,

$$\begin{aligned}\overline{U}_h^{\mathcal{M}'}(\eta_{(1,h-1)}, s_h, \theta_h) &= \mathbb{E}_{x_h,\eta_{(h+1,H)}\sim\mathcal{M}'}\left[\sum_{\tau=h+1}^{H} u^{\mathcal{M}'}(\eta_{(1,\tau-1)}, s_\tau, \theta_\tau)\right] \\ &= \mathbb{E}_{x_h,\eta_{(h+1,H)}\sim\mathcal{M}'}\left[\sum_{\tau=h+1}^{H} u^{\mathcal{M}'}((\eta^*(\mathcal{H}_{(\eta_{(1,h-1)},s_h,\theta_h)}), x_h, \eta_{(h+1,\tau-1)}), s_\tau, \theta_\tau)\right] \\ &= \mathbb{E}_{x_h,\eta_{(h+1,H)}\sim\mathcal{M}}\left[\sum_{\tau=h+1}^{H} u^{\mathcal{M}}((\eta^*(\mathcal{H}_{(\eta_{(1,h-1)},s_h,\theta_h)}), x_h, \eta_{(h+1,\tau-1)}), s_\tau, \theta_\tau)\right],\end{aligned}$$

where the third equality comes from the fact that $\mathcal{M}'$ reduces to $\mathcal{M}$ under the representative history and the distribution over x_h is the same under both $\mathcal{M}$ and $\mathcal{M}'$. Letting $\overline{U}_h(\cdot,\cdot)$ denote the expected continuation utility under $\mathcal{M}$, we have

$$\begin{aligned}&\mathbb{E}_{x_h,\eta_{(h+1,H)}\sim\mathcal{M}}\left[\sum_{\tau=h+1}^{H} u^{\mathcal{M}}((\eta^*(\mathcal{H}_{(\eta_{(1,h-1)},s_h,\theta_h)}), x_h, \eta_{(h+1,\tau-1)}), s_\tau, \theta_\tau)\right] \\ &\quad = \overline{U}_h^{\mathcal{M}}(\eta^*(\mathcal{H}_{(\eta_{(1,h-1)},s_h,\theta_h)})).\end{aligned}$$

We also note that

$$\textsc{Utl}(\mathcal{M}|\eta_{(1,h-1)}, s_h, \theta_h) = \overline{U}_h^{\mathcal{M}}(\eta_{(1,h-1)}, s_h, \theta_h) = \overline{U}_h^{\mathcal{M}}(\eta^*(\mathcal{H}_{(\eta_{(1,h-1)},s_h,\theta_h)})),$$

where first equality is by similar reasoning as (2.7.4) and the second the definition of equivalent history. Therefore

$$\begin{aligned}\textsc{Utl}(\mathcal{M}'|\eta_{(1,h-1)}, s_h, \theta_h) &= \overline{U}_h^{\mathcal{M}'}(\eta_{(1,h-1)}, s_h, \theta_h) \\ &= \overline{U}_h^{\mathcal{M}}(\eta^*(\mathcal{H}_{(\eta_{(1,h-1)},s_h,\theta_h)})) \\ &= \textsc{Utl}(\mathcal{M}|\eta_{(1,h-1)}, s_h, \theta_h).\end{aligned} \tag{2.7.5}$$

Moreover, as the mechanism $\mathcal{M}$ and $\mathcal{M}'$ have the same allocation rule for steps $1, \ldots, h-1$,

the distribution over $(\eta_{(1,h-1)}, s_h, \theta_h)$ is the same for both mechanisms. Integrating over the tuple then gives us

$$\textsc{Utl}(\mathcal{M}') = \textsc{Utl}(\mathcal{M}').$$

Our attention then turns to showing that $\mathcal{M}'$ yields no less revenue than $\mathcal{M}$. Recalling how the representative history is selected, for any $\eta_{(1,h)}$ the revenue $\mathcal{M}'$ achieves starting from step $h+1$ satisfies

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_h,\eta_{(h+1,H)}}\left[\sum_{\tau=h+1}^{H}\psi'_\tau(\eta_{(1,\tau-1)}, s_\tau, \theta_\tau)\right] \\
&\quad= \mathop{\mathbb{E}}_{x_h,\eta_{(h+1,H)}}\left[\sum_{\tau=h+1}^{H}\psi'_\tau((\eta^*(\mathcal{H}_{(\eta_{(1,h-1)},s_h,\theta_h)}), x_h, \eta_{(h+1,\tau-1)}), s_\tau, \theta_\tau)\right] \\
&\quad\geqslant \mathop{\mathbb{E}}_{x_h,\eta_{(h+1,H)}}\left[\sum_{\tau=h+1}^{H}\psi_\tau(\eta_{(1,\tau)}, s_\tau, \theta_\tau)\right],
\end{aligned}$$

where we note that conditioned on the representative history the distribution over future types remain the same for both mechanisms. The equation highlights that, similar what we have shown for utility, the expected continuation revenue of any history of length h under $\mathcal{M}'$ can be re-written as the expected continuation revenue of its representative history under $\mathcal{M}$. By definition of representative history, any history's expected future revenue is no greater than that of its representative history.

By construction of $\mathcal{M}'$, all histories of lengths h yield the same revenue under $\mathcal{M}$ and $\mathcal{M}'$ in the first h steps, as both the allocation and the payment rules are the same. Therefore, we have

$$\textsc{Rev}(\mathcal{M}') \geqslant \textsc{Rev}(\mathcal{M}).$$

Proof of Property 1. Note that the constructions of χ' and ψ', as well as their desired properties, have the same mathematical form. Therefore, for sake of brevity we only prove the property for χ', as the proof for ψ' can be obtained by simply swapping the two symbols.

Let τ denote an arbitrary time step and ν the length of some history. We recall that, for property 1 to hold, we need to show that for all equivalent histories of length $\nu \leqslant h$, the resulting submechanism induced by $\mathcal{M}'$ is the same for all $H \geqslant \tau > \nu$. We then divide the problem into three cases.

1. When $\tau \leqslant h$. Since $\tau > \nu$, in this case we know that $\nu < \tau \leqslant h$, or in other words, $\nu \leqslant h - 1$. By inductive hypothesis, we know that the mechanism is symmetric w.r.t. equivalent histories of length ν. Since $\chi'_{(1,h)}(\cdot,\cdot,\cdot) = \chi_{(1,h)}(\cdot,\cdot,\cdot)$ by construction of $\mathcal{M}'$, the claim trivially holds in this case, as $\mathcal{M}'$ reduces to $\mathcal{M}$ for all $\tau \leqslant h$.

2. When $\tau > h$ and $\nu < h$. We first show for all $x_\nu, \eta_{(\nu+1,h-1)}, s_h$, and θ_h that

$$\begin{aligned}&(\eta_{(1,\nu-1)}, s_\nu, \theta_\nu) \sim_{\mathcal{M}'} (\eta'_{(1,\nu-1)}, s'_\nu, \theta'_\nu)\\ &\quad\Rightarrow (\eta_{(1,h-1)}, s_h, \theta_h) \sim_{\mathcal{M}'} ((\eta'_{(1,\nu-1)}, \eta_{(\nu,h-1)}), s_h, \theta_h),\end{aligned}$$

   where we slightly abuse the notation and let $\sim_{\mathcal{M}'}$ denote equivalence under $\mathcal{M}'$. Here we recall by Definition 2.3.4 that $\eta_\nu = (s_\nu, \theta_\nu, x_\nu) = (s'_\nu, \theta'_\nu, x_\nu)$. Intuitively, we want to show the two histories of length ν under $\mathcal{M}'$ implies that the induced histories of length h are also equivalent under $\mathcal{M}'$. The observation allows us to without loss of generality consider only equivalent histories of length h, which we discuss immediately after addressing this case.

   Recall from (2.7.5) that

$$\textsc{Utl}(\mathcal{M}'|\eta_{(1,h-1)}, s_h, \theta_h) = \textsc{Utl}(\mathcal{M}|\eta_{(1,h-1)}, s_h, \theta_h).$$

   As $\mathcal{M}$ and $\mathcal{M}'$ share the same allocation rule up to step h, $s_\nu = s'_\nu$, and $\theta_\nu = \theta'_\nu$, the distribution over $(x_\nu, \eta_{(\nu+1,h-1)}, s_h, \theta_h)$ is the same under $\mathcal{M}$ and $\mathcal{M}'$. Integrating

over the distribution gives us

$$\textsc{Utl}(\mathcal{M}'|\eta_{(1,\nu-1)}, s_\nu, \theta_\nu) = \textsc{Utl}(\mathcal{M}|\eta_{(1,\nu-1)}, s_\nu, \theta_\nu).$$

In other words, if $\eta_{(1,\nu)} \sim_{\mathcal{M}'} \eta'_{(1,\nu)}$, then $\eta_{(1,\nu)} \sim_{\mathcal{M}} \eta'_{(1,\nu)}$, where $\sim_{\mathcal{M}}$ denotes the equivalence relationship under $\mathcal{M}$.

As $\nu < h$, $\nu \leqslant h-1$. By the inductive hypothesis on $\mathcal{M}'$ the submechanisms following step ν are the same under both histories. Crucially, for the same $(x_\nu, \eta_{(\nu+1,h-1)}, s_h, \theta_h)$, we have

$$\textsc{Utl}(\mathcal{M}'|\eta_{(1,h-1)}, s_h, \theta_h) = \textsc{Utl}(\mathcal{M}'|(\eta'_{(1,\nu-1)}, \eta_{(\nu+1,h-1)}), s_h, \theta_h),$$

where again we recall that $\eta_\nu = (s_\nu, \theta_\nu, x_\nu) = (s'_\nu, \theta'_\nu, x_\nu)$. The equation thus shows that $(\eta_{(1,h-1)}, s_h, \theta_h) \sim_{\mathcal{M}'} ((\eta'_{(1,\nu-1)}, \eta_{(\nu,h-1)}), s_h, \theta_h)$ for all $(x_\nu, \eta_{(\nu+1,h-1)}, s_h, \theta_h)$ as the two histories trivially have the same context and type at step h. Therefore we know

$$\begin{aligned}&(\eta_{(1,\nu-1)}, s_\nu, \theta_\nu) \sim_{\mathcal{M}'} (\eta'_{(1,\nu-1)}, s'_\nu, \theta'_\nu)\\ &\quad\Rightarrow (\eta_{(1,h-1)}, s_h, \theta_h) \sim_{\mathcal{M}'} ((\eta'_{(1,\nu-1)}, \eta_{(\nu,h-1)}), s_h, \theta_h).\end{aligned}$$

3. When $\tau > h$ and $\nu = h$. Since the two histories have length h and are equivalent under $\mathcal{M}'$, again invoking (2.7.5), we know that they are also equivalent under $\mathcal{M}$. Consequently, they are mapped to the same partition $\mathcal{H}$.

   Since the representative history is unique, the two equivalent histories are mapped to the same representative history and would hence have the same submechanism at all steps $\tau > h$.

Combining the three cases, we know for any h, we can use a mechanism symmetric with

respect to equivalent histories up to length $h-1$ to construct $\mathcal{M}'$, a mechanism that is symmetric with respect to equivalent histories up to length h.

By mathematical induction, we can then recursively construct a symmetric mechanism with the same buyer utility and at least the same amount of seller revenue for any dynamic mechanism, completing the proof.

□

Proof of Lemma 2.3.8

We begin by outlining the structure of our proof. The proof consists of two parts, where we first show that $\mathcal{B}^{g,y}$ is a valid augmented ABAM and subsequently we show that $\mathcal{B}^{g,y}$ is IC and IR.

Part 1: $\mathcal{B}^{g,y}$ is a valid ABAM. We first outline the requirements for $\mathcal{B}^{g,y}$ to be a well-defined ABAM.

1. The functions involved in $\mathcal{B}^{g,y}$ are defined on their domains,
2. The bank account payment rule φ_h and bank account deposit rule δ_h are always non-negative.
3. The spending rule $\sigma_h(\cdot,\cdot,\cdot)$ is independent with θ_h and does not exceed $\texttt{bal}_h$.
4. The spending rule $\sigma_h(\cdot,\cdot,\cdot)$ and the deposit rule $\delta_h(\cdot,\cdot,\cdot,\cdot)$ lead to the correct next step balance $\texttt{bal}_{h+1}$.

We then prove these properties one by one. By Lemma 2.3.6, we assume without loss of generality that $\mathcal{M}$, the mechanism from which $\mathcal{B}^{g,y}$ is constructed, is a symmetric mechanism.

1. We first show that ξ_h, φ_h, d_h are valid functions. For the bank account allocation ξ_h, we first note that if two histories have the same balance, then they have the same expected conditional episodic utility and are hence equivalent. More concretely, let

$(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1})$ and $(\eta'_{(1,h-2)}, s'_{h-1}, \theta'_{h-1})$ be a pair of equivalent histories. Since $\mathcal{M}$ is symmetric, there exist two histories $(\eta^{\dagger}_{(1,h-2)}, s^{\dagger}_{h-1}, \theta^{\dagger}_{h-1}), (\eta^{\ddagger}_{(1,h-2)}, s^{\ddagger}_{h-1}, \theta^{\ddagger}_{h-1})$ such that

$$\xi_h(\texttt{bal}_h(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}), \eta_{h-1}, s_h, \theta_h) = y_h((\eta^{\dagger}_{(1,h-2)}, s^{\dagger}_{h-1}, \theta^{\dagger}_{h-1}), \eta_{h-1}, s_h, \theta_h),$$
$$\xi_h(\texttt{bal}_h(\eta'_{(1,h-2)}, s'_{h-1}, \theta'_{h-1}), \eta_{h-1}, s_h, \theta_h) = y_h((\eta^{\dagger}_{(1,h-2)}, s^{\dagger}_{h-1}, \theta^{\dagger}_{h-1}), \eta_{h-1}, s_h, \theta_h).$$

where

$$(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) \sim (\eta^{\dagger}_{(1,h-2)}, s^{\dagger}_{h-1}, \theta^{\dagger}_{h-1}), \tag{2.7.6}$$
$$(\eta'_{(1,h-2)}, s'_{h-1}, \theta'_{h-1}) \sim (\eta^{\ddagger}_{(1,h-2)}, s^{\ddagger}_{h-1}, \theta^{\ddagger}_{h-1}). \tag{2.7.7}$$

By Definition 2.3.4 we have

$$\begin{cases} \texttt{bal}_h(\eta^{\dagger}_{(1,h-2)}, s^{\dagger}_{h-1}, \theta^{\dagger}_{h-1}) = \texttt{bal}_h(\eta^{\ddagger}_{(1,h-2)}, s^{\ddagger}_{h-1}, \theta^{\ddagger}_{h-1}) \\ s^{\dagger}_{h-1} = s^{\ddagger}_{h-1}, \theta^{\dagger}_{h-1} = \theta^{\ddagger}_{h-1} \end{cases},$$

showing that $(\eta^{\dagger}_{(1,h-2)}, s^{\dagger}_{h-1}, \theta^{\dagger}_{h-1}) \sim (\eta^{\ddagger}_{(1,h-2)}, s^{\ddagger}_{h-1}, \theta^{\ddagger}_{h-1})$.

As $\mathcal{M}$ is symmetric, we know that ξ is a well-defined function with a unique output for all possible inputs, as pairs of histories with the same η_{h-1} and balance are assigned to the same allocation rule under y. As the bank account payment rule φ_h and the bank account deposit rule δ_h are all constructed off of ξ_h, we also ensure that for the same bank account balance $\texttt{bal}_h$, η_{h-1}, s_h, and private type θ_h, the functions yield unique outputs and are valid functions. Moreover, note that $\texttt{bal}_h(\cdot,\cdot,\cdot)$ is always non-negative due to C_h. The functions $\xi_h, \varphi_h, \delta_h, s_h$ are then well-defined on their respective domains.

2. We show the bank account payment rule φ_h and bank account deposit rule δ_h are always non-negative. By construction of the bank account allocation rule, ξ, we have

$$\xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) = \chi_h(\eta_{(1,h-1)}, s_h, \theta_h),$$

where, due to symmetry of $\mathcal{M}$, it is without loss of generality to consider one arbitrary history $\eta_{(1,h-1)}$ that yields some $\texttt{bal}_h$ with the $h-1$-th step's history being η_{h-1}. As the underlying mechanism is IC, the allocation rule χ_h, and by extension ξ_h, must be increasing in θ_h.

We then turn our attention back to φ_h and have

$$\varphi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) = \xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h)\theta_h - \int_0^{\theta_h} \xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta)d\theta \geqslant 0.$$

Similarly, $\delta_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) = \int_0^{\theta_h} \xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta)d\theta \geqslant 0$.

3. We show that spending rule $\sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h)$ is independent of θ_h and does not exceed $\texttt{bal}_h$ regardless of the realization of s_h. By construction of σ_h,

$$\begin{aligned}
\frac{\partial \sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h)}{\partial \theta_h} &= \frac{\partial \texttt{bal}_h}{\partial \theta_h} + \frac{\partial(\xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h)\theta_h - \varphi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h))}{\partial \theta_h} \\
&\quad - \frac{\partial(\text{UTL}(\mathcal{M}|\eta_{(1,h-1)}, s_h, \theta_h) - C_h)}{\partial \theta_h} \\
&= \frac{\partial \int_0^{\theta_h} \xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta)d\theta}{\partial \theta_h} - \frac{\partial \text{UTL}(\mathcal{M}|\eta_{(1,h-1)}, s_h, \theta_h)}{\partial \theta_h} \\
&= \xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) - \frac{\partial \text{UTL}(\mathcal{M}|\eta_{(1,h-1)}, s_h, \theta_h)}{\partial \theta_h},
\end{aligned}$$

where for the first equality we use the definition of σ_h and g, the second the definitions of $\texttt{bal}_h, \varphi_h, C_h$, and the third the fundamental theorem of calculus.

Focusing on the second term, we have

$$\begin{aligned}\frac{\partial \textsc{Utl}(\mathcal{M}|\eta_{(1,h-1)}, s_h, \theta_h)}{\partial \theta_h} &= \frac{\partial \sum_{\tau=1}^{h} u_\tau(\eta_{(1,\tau-1)}, s_\tau, \theta_\tau; \theta_\tau) + \bar{U}_h(\eta_{(1,h-1)}, s_h, \theta_h)}{\partial \theta_h} \\ &= \frac{\partial u_h(\eta_{(1,h-1)}, s_h, \theta_h; \theta_h) + \bar{U}_h(\eta_{(1,h-1)}, s_h, \theta_h)}{\partial \theta_h} \\ &= \chi_h(\eta_{(1,h-1)}, s_h, \theta_h),\end{aligned}$$

where the last equation is by Envelope theorem and the incentive compatibility of $\mathcal{M}$. Plugging the result back, we have

$$\frac{\partial \sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h)}{\partial \theta_h} = \xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) - \chi_h(\eta_{(1,h-1)}, s_h, \theta) = 0$$

by construction of the bank account allocation rule ξ_h and the fact that $\mathcal{M}$ is symmetric.

Due to the independence between θ_h and σ_h, we may assume $\theta_h = 0$, in which case

$$\begin{aligned}\sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h) &= \texttt{bal}_h + \delta_h(\texttt{bal}_h, \eta_{h-1}, s_h, 0) - \texttt{bal}_{h+1}(\texttt{bal}_h, \eta_{h-1}, s_h, 0) \\ &= \texttt{bal}_h - \texttt{bal}_{h+1}(\texttt{bal}_h, \eta_{h-1}, s_h, 0) \leqslant \texttt{bal}_h,\end{aligned}$$

where for second equality we used the construction of δ_h and the fact that balances are nonnegative.

4. Plugging in the definitions of the respective terms, we quickly note that the spending rule $\sigma_h(\cdot,\cdot,\cdot)$ and the deposit rule $\delta_h(\cdot,\cdot,\cdot,\cdot)$ lead the correct next step balance $\texttt{bal}_{h+1}$.

Part 2: $\mathcal{B}^{g,y}$ is IC and IR. We begin with showing $\mathcal{B}^{g,y}$ is IC. By construction of the in-stage payment rule φ, it is easy to verify that the in-stage utility $\widehat{u}$ is maximized when the buyer reports truthfully.

Our goal is then showing that the differences in spending rules capture the differences in

expected in-stage utility. Condition on any arbitrary history $(\eta_{(1,h-1)}, s_h, \theta_h)$, rewriting the update rule for `bal` gives us

$$\begin{aligned}
&\sigma_h(\texttt{bal}_h(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}), \eta_{h-1}, s_h) \\
&\quad = \textsc{Utl}(\mathcal{M}|\eta_{(1,h-1)}, s_h, \theta_h) - \textsc{Utl}(\mathcal{M}|\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) \\
&\qquad + \widehat{u}(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) - C_{h-1} + C_h,
\end{aligned}$$

which holds true for all possible realizations of the current private type θ_h as σ_h is not affected by θ_h. We then integrate the right-hand side over the distribution of x_{h-1}, s_h, θ_h when conditioned on s_{h-1}, θ_{h-1} to obtain

$$\begin{aligned}
&\mathbb{E}_{x_{h-1}, s_h \,|\, s_{h-1}, \theta_{h-1}}[\sigma_h(\texttt{bal}_h(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}), \eta_{h-1}, s_h)] \\
&\quad = \textsc{Utl}(\mathcal{M}|\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) - \textsc{Utl}(\mathcal{M}|\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) \\
&\qquad + \mathbb{E}_{x_{h-1}, s_h, \theta_h}[\widehat{u}(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h)|s_{h-1}, \theta_{h-1}] - C_{h-1} + C_h \\
&\quad = \mathbb{E}_{x_{h-1}, s_\theta, \theta_h}[\widehat{u}(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h)|s_{h-1}, \theta_{h-1}] - C_{h-1} + C_h.
\end{aligned}$$

Consequently the conditions of Lemma 2.3.2 hold for both $\widehat{u}, \sigma$, and $\mathcal{B}^{g,y}$ is incentive compatible.

Examining the deposit rule shows that the ABAM satisfies the requirement outlined in Lemma 2.3.3, showing that $\mathcal{B}^{g,y}$ is also ex post IR. □

Proof of Theorem 2.3.9

Sufficiency. We first show that the conditions given in the statement of Theorem 2.3.9 is *sufficient* for $\mathcal{B}^{g,y}$ to be an ABAM. The proof largely follows that of Lemma 2.3.8 and we summarize below the key differences.

- We know by the construction of $\texttt{bal}_h$ that for any s_{h-1}, θ_{h-1} and pair of histories

$\eta_{(1,h-2)}, \eta'_{(1,h-2)}$ such that $\texttt{bal}_h(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = \texttt{bal}_h(\eta'_{(1,h-2)}, s_{h-1}, \theta_{h-1})$, we would also have

$$g_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = g_{h-1}(\eta'_{(1,h-2)}, s_{h-1}, \theta_{h-1}).$$

By definition of symmetry, for any x_{h-1}, s_h, θ_h we have

$$g_h((\eta_{(1,h-2)}, \eta_{h-1}), s_h, \theta_h) = g_h((\eta'_{(1,h-2)}, \eta_{h-1}), s_h, \theta_h)$$

where we recall that $\eta_{(h-1)} = (s_{h-1}, \theta_{h-1}, x_{h-1})$. Since y_h is the sub-gradient of g_h with respect to θ_h

$$\begin{aligned} y_h((\eta_{(1,h-2)}, \eta_{h-1}), s_h, \theta_h) &= \frac{\partial g_h(\eta_{(1,h-1)}, s_h, \theta_h)}{\partial \theta_h} \\ &= \frac{\partial g_h((\eta'_{(1,h-2)}, \eta_{h-1}), s_h, \theta_h)}{\partial \theta_h} \\ &= y_h((\eta'_{(1,h-2)}, \eta_{h-1}), s_h, \theta_h) \end{aligned}$$

for all possible values of η_{h-1}, s_h. Therefore, we know that $\xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h)$ is a well-defined function, as any pair of history of length $h-2$ mapping to the same balance leads to the same allocation for any fixed $(\eta_{h-1}, s_h, \theta_h)$. As $y_h(\eta_{(1,h-1)}, s_h, \theta_h)$ is a valid allocation rule, the function ξ_h also maps to a valid allocation policy. Moreover ξ_h is nonstochastic whenever y_h is.

- $\xi_h, \varphi_h, \delta_h, \sigma_h$ are defined on their domains and φ_h, δ_h are always non-negative for the same reason as in Appendix 2.7.3.

- By construction of the spending rule and the deposit rule, we know for any $h \in [H]$,

$\eta_{h-1}, s_h \in \mathcal{H}$, and balance $\texttt{bal}$

$$\begin{aligned}\sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h) &= \texttt{bal}_h + \delta_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) - \texttt{bal}_{h+1} \\ &= g_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) - C_{h-1} + \int_0^{\theta_h} \xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta) d\theta \\ &\quad - g_h(\eta_{(1,h-1)}, s_h, \theta_h) + C_h.\end{aligned}$$

Taking partial derivative with respect to θ_h shows that

$$\frac{\partial \sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h)}{\partial \theta_h} = y_h(\eta_{(1,h-1)}, s_h, \theta_h) - \frac{\partial g_h(\eta_{(1,h-1)}, s_h, \theta_h)}{\partial \theta_h} = 0,$$

where the second equality comes from the fact that y is the sub-gradient of g. We can then conclude that the construction of σ yields a valid function of $\texttt{bal}$ and s.

- Following the proof of Lemma 2.3.8 we know that $\sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h) \leqslant \texttt{bal}_h$.

- Since $\xi_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h)$, given by y_h, is the subgradient of $g_h(\eta_{(1,h-1)}, s_h, \theta_h)$ with respect to θ_h, by Envelope theorem and the construction of φ_h, for all $\texttt{bal}_h, \eta_{h-1}$, s_h, θ_h, the term $\widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta'_h; \theta_h)$ is maximized when $\theta'_h = \theta_h$.

  Furthermore, as $g_h(\eta_{(1,h-1)}, s_h, \theta_h) = \textsc{Utl}(\mathcal{M}|\eta_{(1,h-1)}, s_h, \theta_h)$, using a similar argument as the proof of Lemma 2.3.8 we can show that σ_h satisfies the conditions in Lemma 2.3.2 and the resulting mechanism is IC.

- Using a similar argument as the proof of Lemma 2.3.8 shows that the mechanism $\mathcal{B}^{g,y}$ is IR.

Necessity. We now show that the conditions given in the statement of Theorem 2.3.9 is *necessary* for $\mathcal{B}^{g,y}$ to be an ABAM, namely, any ABAM $\mathcal{B}^{g,y}$ must satisfy the conditions listed in Theorem 2.3.9.

- Since an ABAM satisfies IC, by construction of the payment rule and the spending rule we have for any $\eta_{(1,h-1)}, s_h, \theta_h$

$$\frac{\partial \sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h)}{\partial \theta_h} = y_h(\eta_{(1,h-1)}, s_h, \theta_h) - \frac{\partial g_{h-1}(\eta_{(1,h-1)}, s_h, \theta_h)}{\partial \theta_h}.$$

  Since the spend rule is not affected by θ_h by construction, the derivative must be zero and we have $y_h(\eta_{(1,h-1)}, s_h, \theta_h) = \frac{\partial g_{h-1}(\eta_{(1,h-1)}, s_h, \theta_h)}{\partial \theta_h}$.

- Since the spend rule $\sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h)$ is a valid function, for any pair of histories $\eta_{(1,h-2)}, \eta'_{(1,h-2)}$ such that

$$\texttt{bal}_h(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = \texttt{bal}_h(\eta'_{(1,h-2)}, s_{h-1}, \theta_{h-1})$$

  we also have

$$\sigma_h(\texttt{bal}_h(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}), \eta_{h-1}, s_h) = \sigma_h(\texttt{bal}_h(\eta'_{(1,h-2)}, s_{h-1}, \theta_{h-1}), \eta_{h-1}, s_h).$$

  Consequently, for any θ_h we have

$$\begin{aligned}
&\texttt{bal}_h(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) + \delta_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) - \texttt{bal}_{h+1}(\eta_{(1,h-1)}, s_h, \theta_h) \\
&\quad = \texttt{bal}_h(\eta'_{(1,h-2)}, s_{h-1}, \theta_{h-1}) + \delta_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) \\
&\qquad - \texttt{bal}_{h+1}((\eta'_{(1,h-2)}, \eta_{h-1}), s_h, \theta_h).
\end{aligned}$$

  As δ and $\texttt{bal}_h$ cancel out, we know that the next step balances are equal, namely, the balances $\texttt{bal}_{h+1}(\eta_{(1,h-1)}, s_h, \theta_h)$ and $\texttt{bal}_{h+1}((\eta'_{(1,h-2)}, \eta_{h-1}), s_h, \theta_h)$ are equal. Recalling how the balance function $\texttt{bal}$ is constructed from g proves that the function g must be symmetric.

- The construction of ψ_h implies that the functions ψ_h and y_h must have the same range.

Thus the range of y_h must be $\Delta(\mathcal{X})$ as well.

- By IC and the Envelope theorem y_h must be weakly increasing in θ_h and its integral, g_h, must be convex and increasing in θ_h.
- Again by IC, the term $\sigma_h(\texttt{bal}_h, \eta_{h-1}, s_h) - \mathbb{E}[\widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h)]$ must be a constant for any fixed $\texttt{bal}_h, \eta_{h-1}, s_h$. Recalling the recursive construction of $\texttt{bal}_{h+1}$ from $\texttt{bal}_h, \sigma_h$, and δ_h shows that g is consistent.

□

2.7.4 Omitted Proofs in Section 2.4

Throughout the section, we assume the type space Θ is discrete. Let $Q = |\Theta|$ denote the number of distinct types and θ^q denote the q-th smallest type. More specifically, we have $\theta^1 \leqslant \ldots \leqslant \theta^Q$. Furthermore, we recall it is without the loss of generality to assume that the lowest type $\theta^1 = 0$.

Proof of Lemma 2.4.1

By definition, we know

$$\begin{aligned}
\max_{\beta_1 \geqslant 0} \Psi_1(\beta_1, \varnothing, \varnothing; \varnothing) &= \max_{\beta_1 \geqslant 0} \max_{\substack{g,y,\varphi \\ g_0(\varnothing)=\beta_1}} \mathop{\mathbb{E}}_{\eta_{(1,H)}} \left[\sum_{\tau=1}^{H} y_\tau(\texttt{bal}_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau)\theta_\tau - \texttt{bal}_{H+1} \right] \\
&= \max_{\beta_1 \geqslant 0} \max_{\mathcal{B}:\textsc{Utl}(\mathcal{B})=\beta_1} \mathop{\mathbb{E}}_{\eta_{(1,H)}} \left[\sum_{\tau=1}^{H} y_\tau(\texttt{bal}_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau)\theta_\tau - \texttt{bal}_{H+1} \right] \\
&= \max_{\mathcal{B}} \mathop{\mathbb{E}}_{\eta_{(1,H)}} \left[\sum_{\tau=1}^{H} (\varphi_\tau(\texttt{bal}_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau) + \sigma_\tau(\texttt{bal}_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau)) - \texttt{bal}_1 \right] \\
&= \max_{\mathcal{B}} \mathop{\mathbb{E}}_{\eta_{(1,H)}} \left[\sum_{\tau=1}^{H} (\varphi_\tau(\texttt{bal}_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau) + \sigma_\tau(\texttt{bal}_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau)) \right],
\end{aligned}$$

which is exactly the amount of revenue the seller generates, comprising of the per-step payment φ_τ and spend rule σ_τ at each step $\tau \in [H]$. For the second equality, we recall Lemma 2.3.8 the assumption that g is UTL. Here $\mathcal{B}$ represents a generic ABAM, as we have shown it is without the loss of generality to consider only such mechanisms. For the third equality, we recall the construction of the balances in Definition 2.3.7. The fourth equality comes from telescoping the sum and the last equality the initial condition that $\texttt{bal}_1 = 0$, again given by Definition 2.3.7.

Our next step is to show that the dynamic programming approach in Table 2.1 is correct.

Lemma 2.7.3. *The program in Table 2.1 correctly returns $\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1})$ as long as Ψ_{h+1} is given.*

Proof. For brevity, for any arbitrary random variable X, we let $\mathbb{E}[X|y_{h-1}]$ be the shorthand for

$$\mathbb{E}[X|y_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = y_{h-1}],$$

the expected value of X conditioned on the event that the allocation probability at step $h-1$ is set to y_{h-1} for the history $(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1})$. Additionally, we let $\beta_h = g_h(\eta_{(1,h-1)}, s_{h-1}, \theta_{h-1})$ denote the balance that the history is mapped to. By definition,

we know

$$
\begin{aligned}
&\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \\
&\quad = \max_{\substack{g,y,\varphi \\ g_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = \beta_h}} \mathop{\mathbb{E}}_{\eta_{(h,H)}} \left[\sum_{\tau=h}^{H} y_\tau(\beta_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau)\theta_\tau - \texttt{bal}_{H+1} \,|\, y_{h-1} \right] \\
&\quad = \max_{\substack{g,y,\varphi \\ g_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = \beta_h}} \Bigg(\mathop{\mathbb{E}}_{\eta_{(h,H)}} \left[y_h(\beta_h, \eta_{h-1}, s_h, \theta_h)\theta_h \,|\, y_{h-1} \right] \\
&\qquad\qquad + \mathop{\mathbb{E}}_{\eta_{(h,H)}} \left[\sum_{\tau=h+1}^{H} y_\tau(\beta_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau)\theta_\tau - \texttt{bal}_{H+1} \,|\, y_{h-1} \right] \Bigg) \\
&\quad = \max_{\substack{g_{(h,H)}, y_{(h,H)}, \varphi_{(h,H)} \\ g_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = \beta_h}} \Bigg(\mathbb{E}_{\theta_h} \left[y_h(\beta_h, \eta_{h-1}, s_h, \theta_h)\theta_h \,|\, y_{h-1} \right] \\
&\qquad\qquad + \mathop{\mathbb{E}}_{\eta_{(h,H)}} \left[\sum_{\tau=h+1}^{H} y_\tau(\beta_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau)\theta_\tau - \texttt{bal}_{H+1} \,|\, y_{h-1} \right] \Bigg)
\end{aligned}
$$

where we recall that it is without the loss of generality to assume that $g(\eta) = \textsc{Utl}(\mathcal{M}|\eta)$. For the second equality, we observe that the objective and constraints are independent of $(g_{(1,h-1)}, y_{(1,h-1)}, \varphi_{(1,h-1)})$. We can further rewrite Ψ_h as

$$
\begin{aligned}
&\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \\
&\quad = \max_{\substack{g_h, y_h, \varphi_h \\ g_{(h+1,H)}, y_{(h+1,H)}, \varphi_{(h+1,H)} \\ g_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = \beta_h}} \Bigg(\mathbb{E}_{x_{h-1}, s_h, \theta_h} \left[y_h(\beta_h, \eta_{h-1}, s_h, \theta_h)\theta_h \,|\, y_{h-1} \right] \\
&\qquad\qquad + \mathop{\mathbb{E}}_{\eta_{(h,H)}} \left[\sum_{\tau=h+1}^{H} y_\tau(\beta_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau)\theta_\tau - \texttt{bal}_{H+1} \,|\, s_h \right] \Bigg) \\
&\quad = \max_{\substack{g_h, y_h, \varphi_h \\ g_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = \beta_h}} \Bigg(\mathbb{E}_{\theta_h} \left[y_h(\beta_h, \eta_{h-1}, s_h, \theta_h)\theta_h \,|\, s_h \right] \\
&\qquad\qquad + \max_{\substack{g,y,\varphi \\ g_h(\eta_{(1,h-1)}, s_h, \theta_h) = \beta_{h+1}}} \mathop{\mathbb{E}}_{\eta_{(h,H)}} \left[\sum_{\tau=h+1}^{H} y_\tau(\beta_\tau, \eta_{\tau-1}, s_\tau, \theta_\tau)\theta_\tau - \texttt{bal}_{H+1} \right] \Bigg) \\
&\quad = \max_{\substack{g_h, y_h, \varphi_h \\ g_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = \beta_h}} \Bigg(\mathbb{E}_{\theta_h} \left[y_h(\beta_h, \eta_{h-1}, s_h, \theta_h)\theta_h \,|\, s_h \right] \\
&\qquad\qquad + \mathbb{E}_{\theta_h} \left[\Psi_{h+1}(\beta_{h+1}, s_h, \theta_h; y_h) \right] \Bigg),
\end{aligned}
$$

where for the second equality, we note that $g_{(h+1,H)}, y_{(h+1,H)}, \varphi_{(h+1,H)}$ affect only the second term in the objective function. For the third equality, we recall the definition of Ψ_{h+1}. Observe that the expression matches exactly the objective defined in equation (2.4.2).

We then note that, by construction of the ABAM, we must have

$$\overline{U}_h(\beta_h, \eta_{h-1}, s_h, \theta_h) - \mathbb{E}_{x_{h-1}, s_h, \theta' \,|\, y_{h-1}}[\overline{U}_h(\beta_h, \eta_{h-1}, s_h, \theta')] = 0.$$

due to the construction of the spend rule σ_τ for $\tau \geqslant h+1$, ensuring that the balance update rule given in (2.4.5) is correct. Additionally, by (2.4.5), we know φ_h can be determined once g_h, y_h are given. Since

$$g_h(\eta_{(1,h-1)}, s_h, \theta_h) = \beta_h + \widehat{u}(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h) - \mathbb{E}_{x_{h-1}, s_h, \theta' \,|\, y_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta')]$$

and we have assumed without the loss of generality that $0 \in \Theta$, we have

$$g_h(\eta_{(1,h-1)}, s_h, 0) = \beta_h - \mathbb{E}_{x_{h-1}, s_h, \theta' \,|\, y_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta')].$$

Consequently, recalling (2.4.3)

$$\begin{aligned}
\varphi_h(\beta_h, \eta_{h-1}, s_h, \theta_h) &= \beta_h + y_h(\beta_h, \eta_{h-1}, s_h, \theta_h)\theta_h + g_h(\eta_{(1,h-1)}, s_h, 0) \\
&\quad - \beta_h - g_h(\eta_{(1,h-1)}, s_h, \theta_h) \\
&= y_h(\beta_h, \eta_{h-1}, s_h, \theta_h)\theta_h + g_h(\eta_{(1,h-1)}, s_h, 0) - g_h(\eta_{(1,h-1)}, s_h, \theta_h).
\end{aligned}$$

As such, it suffices to maximize over only g_h, y_h in the program in Table 2.1.

We finally argue that the resulting mechanism is a valid ABAM. By definition of Ψ_{h+1}, we know the conditions in Theorem 2.3.9 is satisfied by $g_{(h+1,H)}$ and $y_{(h+1,H)}$. Moreover,

- Equation (2.4.4) ensures that $g_h(\eta_{(1,h-1)}, s_h, \theta_h)$ is convex in h.

- Equation (2.4.5) ensures that $g_h(\eta_{(1,h-1)}, s_h, \theta_h)$ is consistent, because for any $\eta_{(1,h-1)}$

$$\mathbb{E}_{x_{h-1},s_h,\theta_h}[g_h(\eta_{(1,h-1)}, s_h, \theta_h)] - g_{h-1}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}) = \beta_h - \beta_h = 0.$$

- The assumption that g_h is expected episodic utility and Lemma 2.3.6 jointly imply that g_h is symmetric.

- Equations (2.4.3) and (2.4.5) jointly imply y_h is a subgradient of $g_h(\eta_{(1,h-1)}, s_h, \theta_h)$.

- Finally, (2.4.6) ensures that the next step balance is non-negative and hence the inputs to Ψ_{h+1} are within the function's domain.

As such, whenever Ψ_{h+1} is given, we can use the program in Table 2.1 to solve for Ψ_h, completing the proof. □

Recursively apply Lemma 2.7.3 and we know that we can solve for Ψ_1 via the program in Table 2.1, obtaining a valid ABAM in the process. □

Proof of Corollary 2.4.4

Inspired by Mirrokni et al. [2016a], we then show that Ψ_h can be approximated in β_h as soon as y_{h-1} is fixed. As the function is defined over two continuous variables, we first show that it is concave in both β_h and y_{h-1} as follows.

Proposition 2.7.4 (Concavity)**.** *The function $\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1})$ is jointly concave in $(\beta_h, y_{h-1}) \in \mathbb{R}_+ \times [0, 1]$ for any $h \in [H]$, $s_{h-1} \in \mathcal{S}$, and $\theta_{h-1} \in \Theta$.*

We then characterize Ψ at extreme choices of balances. Intuitively, when balance is too low, the buyer achieves low episodic utility, and there is relatively little excess utility for the seller to extract at later iterations. When balance is too high, however, seller promises buyer too much utility and, due to the fact that the maximum welfare is bounded, the excess

promised utility could in turn hurt revenue. Below we provide quantitative analysis of Ψ at two extremes. We begin with the following result for when balance is zero.

Proposition 2.7.5 (Value at Low Balance)**.** *For any $h \in [H]$, $s_{h-1} \in \mathcal{S}, \theta_{h-1} \in \Theta$, and $y_{h-1} \in [0,1]$, we have*

$$\begin{aligned}\Psi_h(0, s_{h-1}, \theta_{h-1}; y_{h-1}) &= \mathbb{E}[\Psi_{h+1}(0, s_h, \theta_h; 0)] \\ &+ \mathbb{E}_{x_{h-1}, s_h}[\mathcal{P}_h^{\Theta}(\theta^Q \mid s_h) \max\{0, \theta^Q + \Psi_{h+1}(0, s_h, \theta_h; 1) - \Psi_{h+1}(0, s_h, \theta_h; 0)\}].\end{aligned}$$

Moreover, for any β_h

$$\begin{aligned}\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \leqslant &\Psi_h(0, s_{h-1}, \theta_{h-1}; y_{h-1}) \\ &+ H\left(\sum_{\tau=h}^{H} \max_{s \in \mathcal{S}, i<Q} \frac{\mathcal{P}_\tau^{\Theta}(\theta^i|s)}{\mathcal{P}_\tau^{\Theta}(\theta^{i+1}|s)(\theta^{i+1} - \theta^i)}\right)\beta_h,\end{aligned}$$

where we recall θ^q is the q-th highest-type in Θ.

The result for high balance is more involved, and requires defining the function MaxWel. Particularly, for any history $\eta_{(h,h')}$, the function

$$\textsc{MaxWel}_{h'}(\eta_{(h,h')}) = \max_{\mathcal{M}} \mathbb{E}_{\eta_{(1,H)}}\left[\sum_{\tau=h'}^{H} (y_\tau(\eta_{(1,\tau-1)}, s_\tau, \theta_\tau)\theta_\tau) \,|\, \eta_{(h,h')}\right] \tag{2.7.8}$$

is the maximum continuation welfare when conditioned on the history $\eta_{(h,h')}$. The functions $\textsc{MaxWel}_h(\eta_{(1,h-1)}, s_h)$ and $\textsc{MaxWel}_h(\eta_{(1,h-1)}, s_h, \theta_h)$ are similarly defined.

We stress that in the MDP setting, the welfare maximizing policy is not necessarily the one that always gives the item to buyers without charging anything. The distribution over later contexts, and by extension the buyer's future types, are affected by the allocation rule. As such, it is possible that withholding the item from the buyer may increase his future types, which can be exploited by the welfare maximizing policy. As such, it is necessary for

us to introduce the notion of $\textsc{MaxWel}_h$ as we cannot simply replace it with the expected sum of all future types.

Proposition 2.7.6 (Value at High Balance)**.** *For any arbitrary and fixed h, $\eta_{(1,h-2)}$, s_{h-1}, θ_{h-1}, and y_{h-1}, the following holds*

1. *for all $\beta_h \geqslant 0$, $\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \leqslant \mathbb{E}_{x_{h-1}}[\textsc{MaxWel}_h(\eta_{(1,h-1)}) \,|\, y_{h-1}] - \beta_h$,*
2. *for all $\beta_h \geqslant \sum_{\tau=h}^{H} \max_{s\in\mathcal{S}} \mathbb{E}_{\theta\sim\mathcal{P}_\tau^\Theta(\cdot\,|\,s)}[\theta]$,*

$$\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) = \mathbb{E}_{x_{h-1}}[\textsc{MaxWel}_h(\eta_{(1,h-1)}) \,|\, y_{h-1}] - \beta_h.$$

With these properties in mind, we are now ready to state the proof itself.

Proof of Corollary 2.4.4. We prove the claim by construction and begin by introducing our proposed discretization scheme.

1. Select $\frac{2H}{\kappa}$ evenly spaced points on the interval $[0, 1]$. Let $\mathcal{Y}^\dagger$ denote the set of points selected.
2. For each $y^\dagger \in \mathcal{Y}^\dagger$, use the discretization scheme in Lemma 2.7.12 to construct a multiplicative $\frac{\kappa}{2H}$-approximation for the function $\Psi_h(\cdot, s_{h-1}, \theta_{h-1}; y^\dagger)$. We then receive functions $\Psi_h^\bigtriangledown$ and $\Psi_h^\bigtriangleup$ which are piece-wise linear in β_h.
3. For all (β_h, y_{h-1}), let $y^\ddagger \in \mathcal{Y}^\dagger$ be the point that y_{h-1} is the closest to. Set

$$\Psi_h^\bigtriangledown(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) = \Psi_h^\bigtriangledown(\beta_h, s_{h-1}, \theta_{h-1}; y^\ddagger) - \frac{\kappa}{2}$$
$$\Psi_h^\bigtriangleup(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) = \Psi_h^\bigtriangleup(\beta_h, s_{h-1}, \theta_{h-1}; y^\ddagger) + \frac{\kappa}{2}.$$

By Propositions 2.7.4, 2.7.5, 2.7.6, we know that for all fixed $y_{h-1} \in [0, 1]$, the function $\Psi_h(\cdot, s_{h-1}, \theta_{h-1}; y_{h-1})$ satisfies the conditions in Lemma 2.7.12. As a result, Step 2 of the proposed discretization scheme is valid.

We then show that the functions $\Psi_h^{\triangledown}, \Psi_h^{\triangle}$ are an additive κ-approximation. Let (β_h, y_{h-1}) be arbitrary and fixed, and let $y^{\ddagger} \in \mathcal{Y}^{\dagger}$ be the closest point to y_{h-1}. From Proposition 2.4.3 and the number of evenly spaced points in $\mathcal{Y}^{\dagger}$ we know

$$|\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) - \Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y^{\ddagger})| \leqslant \frac{\kappa}{2}.$$

Because the types θ are bounded, Ψ_h is uniformly bounded by H, and applying Lemma 2.7.12 shows that $\Psi_h^{\triangledown}$ and $\Psi_h^{\triangle}$ is an additive κ-approximation.

Finally we control the number of pieces. Again by Lemma 2.7.12, we know that constructing $\Psi_h^{\triangledown}, \Psi_h^{\triangle}$ requires $\mathcal{O}(N^2/\kappa^2)$ calls to the evaluation oracle and both $\Psi_h^{\triangledown}, \Psi_h^{\triangle}$ have $\mathcal{O}\left(N^2/\kappa^2\right)$ pieces, completing the proof. □

Proof of Proposition 2.7.4

Let $\mathcal{M}^{(1)} = \mathcal{B}^{(g^{(1)}, y^{(1)})}, \mathcal{M}^{(2)} = \mathcal{B}^{(g^{(2)}, y^{(2)})}$ denote two arbitrary and fixed ABAMs. We use $(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1})$ to denote some arbitrary and fixed history. With a slight abuse of notation, we overload the notation and let $y^{(1)}$ and $y^{(2)}$ denote the allocation levels for the history given by the two mechanisms. That is

$$y^{(1)} = y^{(1)}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}), \qquad y^{(2)} = y^{(2)}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1}).$$

Assume without loss of generality that $g^{(1)}(\cdot) = \textsc{Utl}(\mathcal{M}^{(1)}|\cdot), g^{(2)}(\cdot) = \textsc{Utl}(\mathcal{M}^{(2)}|\cdot)$.

Let $c \in [0, 1]$ be an arbitrary and fixed constant, $\beta_h^{(1)} = g_h^{(1)}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1})$, and $\beta_h^{(2)} = g_h^{(2)}(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1})$. Consider the mechanism formed by mixing $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$, whose allocation rule is $y(\cdot, \cdot, \cdot) = cy^{(1)}(\cdot, \cdot, \cdot) + (1 - c)y^{(2)}(\cdot, \cdot, \cdot)$. With a slight abuse of notation, let

$$y = cy^{(1)} + (1 - c)y^{(2)},$$

the allocation level the mixture mechanism chooses for the history $(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1})$. By linearity of expectation, the balance corresponding to the history $(\eta_{(1,h-2)}, s_{h-1}, \theta_{h-1})$ is $c\beta_h^{(1)} + (1-c)_h^{(2)}$.

While the mixture mechanism is not necessarily an ABAM, applying Lemma 2.3.6 and Lemma 2.3.8, there must exist an ABAM with revenue no less than the interpolation of revenue from $\mathcal{M}^{(1)}$ and revenue from $\mathcal{M}^{(2)}$. As $\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1})$ takes the maximum over all ABAMs, we must have

$$\begin{aligned}&\Psi_h(c\beta_h^{(1)} + (1-c)\beta_h^{(2)}, s_{h-1}, \theta_{h-1}; cy_{h-1}^{(1)} + (1-c)y_{h-1}^{(2)}) \\ &\quad \geqslant c\Psi_h(\beta_h^{(1)}, s_{h-1}, \theta_{h-1}; y_{h-1}^{(1)}) + (1-c)\Psi_h(\beta_h^{(2)}, s_{h-1}, \theta_{h-1}; y_{h-1}^{(2)}),\end{aligned}$$

completing the proof.

□

Proof of Proposition 2.7.5

Let h, s_{h-1}, θ_{h-1}, and y_{h-1} be arbitrary and fixed. For brevity, we drop them from the notation during the proof. By (2.4.5) and (2.4.6), when $\beta_h = 0$, recalling that the lowest type in Θ is 0, we must have

$$g_h(\beta_h, \eta_{h-1}, 0) = 0 - \mathbb{E}_{x_{h-1}, s_h, \theta' \,|\, y_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta')] \geqslant 0,$$

which implies $\mathbb{E}_{x_{h-1}, s_h, \theta' \,|\, y_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta')] = 0$.

By non-negativity of $\widehat{u}_h$, we know $\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h) = 0$ and $g_h(\beta_h, \eta_{h-1}, s_h, \theta_h) = 0$ for all x_{h-1}, s_h, θ_h. Focusing on some fixed x_{h-1}, s_h, we know $y_h(0, \eta_{h-1}, s_h, \theta^q) = 0$ for all $q < Q$, that is, the allocation probability is zero for all types save for the highest one.

Moreover, in this case the part of the objective function in (2.4.2) reduces to

$$y_h(\beta_h, \eta_{h-1}, s_h, \theta^Q)\theta^Q + \Psi_{h+1}(0, s_h, \theta^Q; y_h(\beta_h, \eta_{h-1}, s_h, \theta^Q)).$$

By the Markovian transition kernel and the fact that y_h encodes the probability that $x_h = 1$, we have

$$\begin{aligned}
&\Psi_{h+1}(0, s_h, \theta^Q; y_h(\beta_h, \eta_{h-1}, s_h, \theta^Q)) \\
&\quad = y_h(\beta_h, \eta_{h-1}, s_h, \theta^Q)\Psi_{h+1}(0, s_h, \theta^Q; 1) + (1 - y_h(\beta_h, \eta_{h-1}, s_h, \theta^Q))\Psi_{h+1}(0, s_h, \theta^Q; 0).
\end{aligned}$$

The objective function is thus now linear in y_h. Moreover, as $g_h(\beta_h, \eta_{h-1}, s_h, \theta_h) = 0$ for all x_{h-1}, s_h, θ_h, the optimization program in Algorithm 2.1 is now reduced to a linear program that can be solved exactly by hand, whose solution is

$$y_h(\beta_h, \eta_{h-1}, s_h, \theta^Q) = \mathbb{1}\{\theta^Q + \Psi_{h+1}(0, s_h, \theta_h; 1) \geqslant \Psi_{h+1}(0, s_h, \theta_h; 0)\}$$

and we have

$$\begin{aligned}
\Psi_h(0, s_{h-1}, \theta_{h-1}; y_{h-1}) &= \mathbb{E}[\Psi_{h+1}(0, s_h, \theta_h; 0)] \\
&\quad + \mathbb{E}_{x_{h-1}, s_h}[\mathcal{P}_h^{\Theta}(\theta^Q \mid s_h) \max\{0, \theta^Q + \Psi_{h+1}(0, s_h, \theta_h; 1) - \Psi_{h+1}(0, s_h, \theta_h; 0)\}].
\end{aligned}$$

We now focus on the second equation and prove it by inducting on h from $H + 1$ to h. The base case for when $h = H + 1$ is clearly true. We then focus on some arbitrary and fixed β_h, η_{h-1}, s_h and assume that the inequality holds for Ψ_{h+1}. For brevity, we drop β_h, η_{h-1}, s_h from our notation in our ensuing discussion.

By (2.4.4), we know for two adjacent types θ^q, θ^{q+1}

$$\widehat{u}_h(\theta^q; \theta^{q+1}) = y_h(\theta^q)\theta^{q+1} - \varphi_h(\theta^q) \leqslant \widehat{u}_h(\theta^{q+1}; \theta^{q+1}).$$

Subtracting $\widehat{u}_h(\theta^q; \theta^q) = y_h(\theta^q)\theta^q - \varphi_h(\theta^q)$ from both sides gives us

$$y_h(\theta^q)(\theta^{q+1} - \theta^q) \leqslant \widehat{u}_h(\theta^{q+1}; \theta^{q+1}) - \widehat{u}_h(\theta^q; \theta^q) \leqslant \widehat{u}_h(\theta^{q+1}; \theta^{q+1}),$$

where the second inequality comes from the fact that $\widehat{u}_h(\theta^q; \theta^q) \geqslant 0$ for all $q \in [Q]$. Recalling that θ^q is the q-th largest type, we know $\theta^{q+1} - \theta^q > 0$ and dividing both sides by the term gives us $y_h(\theta^q) \leqslant \frac{1}{\theta^{q+1} - \theta^q} \widehat{u}_h(\theta^{q+1}; \theta^{q+1})$.

Therefore for any $q < Q$ and valid $y_h(\theta^q)$, we must have

$$\begin{aligned}
&y_h(\theta^q)\theta^q + \Psi_{h+1}(\beta_{h+1}(\theta^q), \theta^q; y_h(\theta^q)) \\
&\qquad - H \left(\sum_{\tau=h+1}^{H} \max_{s \in \mathcal{S}, i<Q} \frac{\mathcal{P}_\tau^\Theta(\theta^i|s)}{\mathcal{P}_\tau^\Theta(\theta^{i+1}|s)(\theta^{i+1} - \theta^i)} \right) \beta_{h+1}(\theta^q) \\
&\overset{\text{(i)}}{\leqslant} y_h(\theta^q)\theta^q + \Psi_{h+1}(0, \theta^q; y_h(\theta^q)) \\
&\overset{\text{(ii)}}{=} y_h(\theta^q)\theta^q + y_h(\theta^q)\Psi_{h+1}(0, \theta^q; 1) + (1 - y_h(\theta^q))\Psi_{h+1}(0, \theta^q; 0) \\
&= y_h(\theta^q)\left(\theta^q + \Psi_{h+1}(0, \theta^q; 1) - \Psi_{h+1}(0, \theta^q; 0)\right) + \Psi_{h+1}(0, \theta^q; 0) \\
&\overset{\text{(iii)}}{\leqslant} y_h(\theta^q)(H - h + 1) + \Psi_{h+1}(0, \theta^q; 0) \\
&\leqslant \max_{i<Q} \frac{H - h + 1}{\theta^{i+1} - \theta^i} \widehat{u}(\theta^{q+1}; \theta^{q+1}) + \Psi_{h+1}(0, \theta^q; 0),
\end{aligned}$$

where (i) comes from the inductive hypothesis on Ψ_{h+1}, (ii) from the fact that $y_h(\theta^q)$ itself is the allocation probability and the definition of the transition probabilities $\mathcal{P}_h^\mathcal{S}$, and (iii) from applying a naive upper bound on the value of Ψ_{h+1} and that $\Psi_\tau(0, s_{\tau-1}, \theta_{\tau-1}; y_{\tau-1}) \geqslant 0$ for all τ. The naive upper bound holds by noting that $\theta \leqslant 1$ for all $\theta \in \Theta$, and is easily be proven by induction, using the formula for $\Psi_\tau(0, s_{\tau-1}, \theta_{\tau-1}; y_{\tau-1})$ that we have established

in the earlier part of the proof.

For the highest type θ^Q, we have

$$\begin{aligned}
&y_h(\theta^Q)\theta^Q + \Psi_{h+1}(\beta_{h+1}(\theta^Q), \theta^Q; y_h(\theta^Q)) \\
&\quad \leqslant y_h(\theta^Q)\theta^Q + \Psi_{h+1}(0, \theta^Q; y_h(\theta^Q)) \\
&\qquad + H\left(\sum_{\tau=h+1}^{H} \max_{s\in\mathcal{S}, i<Q} \frac{\mathcal{P}_\tau^\Theta(\theta^i|s)}{\mathcal{P}_\tau^\Theta(\theta^{i+1}|s)(\theta^{i+1}-\theta^i)}\right)\beta_{h+1}(\theta^Q) \\
&\quad \leqslant \max_{y_h(\theta^Q)\in[0,1]} \{y_h(\theta^Q)\theta^Q + \Psi_{h+1}(0, \theta^Q; y_h(\theta^Q))\} \\
&\qquad + H\left(\sum_{\tau=h+1}^{H} \max_{s\in\mathcal{S}, i<Q} \frac{\mathcal{P}_\tau^\Theta(\theta^i|s)}{\mathcal{P}_\tau^\Theta(\theta^{i+1}|s)(\theta^{i+1}-\theta^i)}\right)\beta_{h+1}(\theta^Q).
\end{aligned}$$

Taking expectation over x_{h-1}, s_h, θ_h conditioned on s_{h-1}, y_{h-1}, we know

$$\begin{aligned}
&\mathbb{E}[y_h(\theta^q) + \Psi_{h+1}(\beta_{h+1}(\theta^q), \theta^q; y_h(\theta^q))] \\
&\qquad - H\left(\sum_{\tau=h+1}^{H} \max_{s\in\mathcal{S}, i<Q} \frac{\mathcal{P}_\tau^\Theta(\theta^i|s)}{\mathcal{P}_\tau^\Theta(\theta^{i+1}|s)(\theta^{i+1}-\theta^i)}\right)\mathbb{E}[\beta_{h+1}(x_{h-1}, s_h, \theta^q)] \\
&\quad \leqslant \mathbb{E}\Bigg[\mathcal{P}_h^\Theta(\theta^Q \mid s_h) \max_{y_h(\theta^Q)\in[0,1]} \{y_h(\theta^Q)\theta^Q + \Psi_{h+1}(0, \theta^Q; y_h(\theta^Q))\} \\
&\qquad\qquad + \sum_{q=1}^{Q-1} \mathcal{P}_h^\Theta(\theta^q \mid s_h)\Psi_{h+1}(0, \theta^q; 0)\Bigg] \\
&\qquad + \mathbb{E}\left[\left(\max_{i<Q} \frac{H-h+1}{(\theta^{i+1}-\theta^i)}\right)\sum_{i=1}^{Q-1} \mathcal{P}_h^\Theta(\theta^i \mid s_h)\widehat{u}(\theta^{i+1}; \theta^{i+1})\right] \\
&\quad = \Psi_h(0, s_{h-1}, \theta_{h-1}; y_{h-1}) + \mathbb{E}\left[\left(\max_{i<Q} \frac{H-h+1}{(\theta^{i+1}-\theta^i)}\right)\sum_{i=1}^{Q-1} \mathcal{P}_h^\Theta(\theta^i \mid s_h)\widehat{u}(\theta^{i+1}; \theta^{i+1})\right],
\end{aligned}$$

where the equality again comes from the expression for $\Psi_h(0, s_{h-1}, \theta_{h-1}; y_{h-1})$ that we have

just developed. Meanwhile, we note that for any x_{h-1} and s_h

$$\begin{aligned}
&\left(\max_{i<Q} \frac{H-h+1}{(\theta^{i+1}-\theta^i)}\right) \sum_{i=1}^{Q-1} \mathcal{P}_h^\Theta(\theta^i \mid s_h)\widehat{u}(\theta^{i+1};\theta^{i+1}) \\
&\quad \leqslant \left(\max_{i<Q} \frac{\mathcal{P}_h^\Theta(\theta^i \mid s_h)(H-h+1)}{\mathcal{P}_h^\Theta(\theta^{i+1} \mid s_h)(\theta^{i+1}-\theta^i)}\right) \sum_{i=1}^{Q-1} \mathcal{P}_h^\Theta(\theta^{i+1} \mid s_h)\widehat{u}(\theta^{i+1};\theta^{i+1}) \\
&\quad = \left(\max_{i<Q} \frac{\mathcal{P}_h^\Theta(\theta^i \mid s_h)(H-h+1)}{\mathcal{P}_h^\Theta(\theta^{i+1} \mid s_h)(\theta^{i+1}-\theta^i)}\right) \mathbb{E}_{\theta_h \mid s_h}[\widehat{u}(\theta_h;\theta_h)],
\end{aligned}$$

where for the equality we recall that we assumed without the loss of generality that the lowest type is 0, which corresponds to a per-step utility of 0. Consequently

$$\begin{aligned}
&\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \\
&\quad \leqslant \Psi_h(0, s_{h-1}, \theta_{h-1}; y_{h-1}) + \left(\max_{i<Q} \frac{\mathcal{P}_h^\Theta(\theta^i \mid s_h)(H-h+1)}{\mathcal{P}_h^\Theta(\theta^{i+1} \mid s_h)(\theta^{i+1}-\theta^i)}\right) \mathbb{E}_{x_{h-1},s_h,\theta_h}[\widehat{u}(\theta_h;\theta_h)] \\
&\qquad + H\left(\sum_{\tau=h+1}^{H} \max_{s\in\mathcal{S},i<Q} \frac{\mathcal{P}_\tau^\Theta(\theta^i|s)}{\mathcal{P}_\tau^\Theta(\theta^{i+1}|s)(\theta^{i+1}-\theta^i)}\right) \mathbb{E}_{x_{h-1},s_h,\theta_h}[\beta_{h+1}(\beta_h, x_{h-1}, s_h, \theta^q)] \\
&\quad \leqslant \Psi_h(0, s_{h-1}, \theta_{h-1}; y_{h-1}) + H\left(\sum_{\tau=h}^{H} \max_{s\in\mathcal{S},i<Q} \frac{\mathcal{P}_\tau^\Theta(\theta^i|s)}{\mathcal{P}_\tau^\Theta(\theta^{i+1}|s)(\theta^{i+1}-\theta^i)}\right) \beta_h,
\end{aligned}$$

where the second inequality comes from (2.4.5). □

Proof of Proposition 2.7.6

We prove both claims by induction. The base case for when $h = H+1$ trivially holds. For the rest of the proof, we assume both claims hold for $h+1$, all $\beta_{h+1} \geqslant 0, s_h, \theta_h$, and y_h. For convenience, similar to the proof of Proposition 2.7.5, for now we focus on a specific $\beta_h, s_{h-1}, \theta_{h-1}$, and y_{h-1} and drop them from our notation unless otherwise specified.

We begin with the first claim. Observe that

$$\begin{aligned}
&\mathbb{E}_{x_{h-1},s_h,\theta_h}[y_h(x_{h-1},s_h,\theta_h)\theta_h + \Psi_{h+1}(\beta_{h+1}(x_{h-1},s_h,\theta_h),s_h,\theta_h;y_h(x_{h-1},s_h,\theta_h))] \\
&\quad\leqslant \mathbb{E}_{x_{h-1},s_h,\theta_h}[y_h(x_{h-1},s_h,\theta_h)\theta_h \\
&\qquad\qquad + \mathbb{E}_{x_h}[\textsc{MaxWel}_{h+1}(\eta_{(1,h)}) \,|\, y_h(x_{h-1},s_h,\theta_h)] - \beta_{h+1}(x_{h-1},s_h,\theta_h)] \\
&\quad\leqslant \mathbb{E}_{x_{h-1}}[\textsc{MaxWel}_h(\eta_{(1,h-1)} \,|\, y_{h-1}) - \beta_{h+1}(x_{h-1},s_h,\theta_h)] \\
&\quad= \mathbb{E}_{x_{h-1}}[\textsc{MaxWel}_h(\eta_{(1,h-1)}) \,|\, y_{h-1}] - \beta_h,
\end{aligned}$$

where the inequality holds by our inductive hypothesis, the second by a one-step expansion of $\textsc{MaxWel}_h(\eta_{(1,h-1)} \,|\, y_{h-1})$ using its definition in (2.7.8), and the last equation the constraint in (2.4.5). Noting that the upper bound holds for (2.4.2), the objective function of the optimization program in Algorithm 2.1, completes the proof.

We then show a corresponding lower bound, and these two bounds meet for all sufficiently large β_h. Easy to see that when $\beta_h \geqslant \sum_{\tau=h}^{H} \max_{s\in\mathcal{S}} \mathbb{E}_{\theta\sim\mathcal{P}_\tau^\Theta(\cdot\,|\,s)}[\theta]$, a feasible solution to the optimization program in Algorithm 2.1 is to set y_h, g_h to the sub-mechanism of the welfare-maximizing mechanism at step h. By (2.4.5), we know that under this choice,

$$\begin{aligned}
\beta_{h+1}(x_{h-1},s_h,\theta_h) &\geqslant \sum_{\tau=h}^{H} \max_{s\in\mathcal{S}} \mathbb{E}_{\theta\sim\mathcal{P}_\tau^\Theta(\cdot\,|\,s)}[\theta] - \max_{s\in\mathcal{S}} \mathbb{E}_{\theta\sim\mathcal{P}_h^\Theta(\cdot\,|\,s)}[\theta] \\
&\geqslant \sum_{\tau=h+1}^{H} \max_{s\in\mathcal{S}} \mathbb{E}_{\theta\sim\mathcal{P}_\tau^\Theta(\cdot\,|\,s)}[\theta],
\end{aligned}$$

noting that $\widehat{u}_h$ is non-negative and naively bounding $\mathbb{E}_{x_{h-1},s_h,\theta_h}[\widehat{u}_h(x_{h-1},s_h,\theta_h;\theta_h)]$. Again using a one-step expansion of $\textsc{MaxWel}_h(\eta_{(1,h-1)} \,|\, y_{h-1})$ and that the sub-mechanism of the welfare-maximizing mechanism is not necessarily optimal, we know

$$\Psi_h(\beta_h, s_{h-1}, \theta_{h-1}; y_{h-1}) \geqslant \mathbb{E}_{x_{h-1}}[\textsc{MaxWel}_h(\eta_{(1,h-1)}) \,|\, y_{h-1}] - \beta_h$$

for all $\beta_h \geqslant \sum_{\tau=h}^{H} \max_{s\in\mathcal{S}} \mathbb{E}_{\theta\sim\mathcal{P}_\tau^\Theta(\cdot\,|\,s)}[\theta]$. Combining the lower bound with the upper bound above completes the proof. □

2.7.5 Ommited Proofs in Section 2.5

Proof of Theorem 2.5.3

We first introduce useful intermediate results that streamlines the proof of Theorem 2.5.3, beginning with the following bound on the expectation of the evaluation error of spend and payment rules, when the expectation is taken over $\widehat{\mathcal{P}}$ rather than $\mathcal{P}$. Proofs are deferred to Appendix 2.7.5 unless stated otherwise.

Lemma 2.7.7. *With probability at least $1-\delta$, the following holds simultaneously all ABAMs $\mathcal{B}^{g,y}$*

$$\left|\mathbb{E}_{\eta_{(1,H)}}\left[\sum_{h=1}^{H}\sigma_h(\beta_h,\eta_{h-1},s_h)\right]-\widehat{\mathbb{E}}_{\eta_{(1,H)}}\left[\sum_{h=1}^{H}\sigma_h(\beta_h,\eta_{h-1},s_h)\right]\right|\leqslant H\epsilon,$$
$$\left|\mathbb{E}_{\eta_{(1,H)}}\left[\sum_{h=1}^{H}\varphi_h(\beta_h,\eta_{h-1},s_h,\theta_h)\right]-\widehat{\mathbb{E}}_{\eta_{(1,H)}}\left[\sum_{h=1}^{H}\varphi_h(\beta_h,\eta_{h-1},s_h,\theta_h)\right]\right|\leqslant \epsilon,$$

where we recall $\widehat{\mathbb{E}}$ is taken with respect to the estimated transition probabilities $\widehat{\mathcal{P}}$.

The lemma relies on properties of Reward-Free RL-Explore, which we discuss in detail in Appendix 2.7.1. At a high level, for *any* Markovian function that depends only on the public context s, Reward-Free RL-Explore ensures that its expected value is estimated accurately, when the expectation is taken over the public context distribution induced by *any* type-agnostic policy π. Such strong guarantees are crucial for controlling the estimation error of non-Markovian functions σ, φ, as we take the "worst possible" realized values of $\eta_{(1,h-1)}$ for any π, σ, φ, taken with respect to the estimation error in $\mathbb{E}_{\eta_h}$, and transform the non-Markovian functions σ, φ to Markovian ones. The errors can then be

controlled via properties of REWARD-FREE RL-EXPLORE, noting that the guarantees hold simultaneously for all Markovian functions and policies.

The lemma then implies that the estimated revenue of any given mechanism can be controlled, which we formalize as follows.

Corollary 2.7.8 (Revenue Estimate Error)**.** *For any (possibly non-core) ABAM $\mathcal{B}$, letting* REV *denote its expected revenue under the ground-truth transition probabilities $\mathcal{P}$ and* $\widehat{\text{REV}}$ *the estimated revenue computed using the estimated transition probabilities $\widehat{\mathcal{P}}$, we have*

$$|\text{REV}(\mathcal{B}) - \widehat{\text{REV}}(\mathcal{B})| \leqslant (H+1)\epsilon.$$

Proof. Revenues from ABAMs only come from the payment and spend rules. Lemma 2.7.7 directly bounds the estimation error in revenue, completing the proof. □

We now turn our attention to the estimation procedure, specifically the relaxed program in Table 2.2. We first obtain the following lemma.

Lemma 2.7.9 (Expected Per-Step Utility Error)**.** *There exists some absolute constant $c > 0$ such that, with probability at least $1-\delta$, the following inequality holds simultaneously for all $h \in [H], \beta_h \geqslant 0, s_{h-1} \in \mathcal{S}, \theta_{h-1} \in \Theta,$ and $y_{h-1} \in [0,1]$*

$$\begin{aligned}
&|\mathbb{E}_{x_{h-1},s_h,\theta' \,|\, s_{h-1},y_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)] \\
&\qquad - \widehat{\mathbb{E}}_{x_{h-1},s_h,\theta_h \,|\, s_{h-1},y_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)]| \\
&\qquad \leqslant cH|\mathcal{S}||\Theta|\sqrt{\log(cH|S||\Theta|/\delta)}\left((N_{h-1}(s_{h-1},0))^{-1/2} + (N_{h-1}(s_{h-1},1))^{-1/2}\right).
\end{aligned}$$

Proof Sketch of Lemma 2.7.9. Estimation error in $\widehat{\mathbb{E}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta'; \theta')]$ can be controlled via bounding estimation error in $\widehat{\mathcal{P}}^{\mathcal{S}}_{h-1}$ and $\widehat{\mathcal{P}}^{\Theta}_{h}$ using Dvoretzky–Kiefer–Wolfowitz inequality [Massart, 1990]. We note that extra care is taken to relate the sum of the estimation errors in $\widehat{\mathcal{P}}^{\Theta}_{h}(\theta' \,|\, s_h)$ over all s_h, θ' back to $N_{h-1}(s_{h-1}, 0)$ and $N_{h-1}(s_{h-1}, 1)$, the number of

times that s_{h-1} has been visited. □

The following Corollary then immediately shows that the relaxation taken by optimization program in Table 2.2 is correct, that is, it ensures that the optimal ABAM remains feasible under the relaxed program.

Corollary 2.7.10 (Optimistic Estimation)**.** *With probability at least $1-\delta$, the optimal ABAM $\mathcal{B}^* = \mathcal{B}^{g^*,y^*}$ is a feasible solution to the relaxed optimization program in Table 2.2.*

Proof. As the mechanism $\mathcal{B}^*$ is a valid ABAM, by correctness of the program in Table 2.1 (Lemma 2.4.1) and the upper bound on estimation error of $\widehat{g}$ (Lemma 2.7.9), with probability at least $1-\delta$, the mechanism $\mathcal{B}^*$ yields a feasible solution to the relaxed program in Table 2.2. □

Moreover, also via Lemma 2.7.9, we can control the amount of estimation error in the spend rules. We first let $\bar{\sigma}_h(\beta_h, \eta_{h-1}, s_h) = \mathbb{E}_{x_{h-1},s_h,\theta_h}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h)]$ denote the "ground-truth" spend rule for the estimated mechanism $\widehat{\mathcal{B}}$, calculated using *exactly* the underlying transitions probabilities $\mathcal{P}$, rather than estimated values in $\widehat{\mathcal{P}}$. While estimation errors in spend rule incur errors in β_h, we recall that β_h is merely a function of $\eta_{(1,h-1)}$ and let $\widehat{\sigma}_h(\eta_{(1,h-1)}, s_h) = \widehat{\sigma}_h(\beta_h, \eta_{h-1}, s_h)$, where β_h is updated according to the estimated spend rule $\widehat{\sigma}$. Additionally, let $\bar{\sigma}_h(\eta_{(1,h-1)}, s_h) = \bar{\sigma}_h(\beta_h, \eta_{h-1}, s_h)$ where β_h is instead updated according to $\bar{\sigma}$. We then have the following.

Corollary 2.7.11. *For any estimated spend rule $\widehat{\sigma}$ with probability at least $1-\delta$ we have*

$$\max_{\pi,\theta_{(1,H)},x_{(1,H)}} \mathbb{E}_{s_{(1,H)}\,|\,\pi}\left[\sum_{h=1}^{H} |\widehat{\sigma}_h(\eta_{(1,h-1)}, s_h) - \bar{\sigma}_h(\eta_{(1,h-1)}, s_h)|\right] \leqslant 5|\Theta|\sqrt{c|\mathcal{S}|}\epsilon.$$

Proof Sketch of Corollary 2.7.11. We prove the claim by combining properties of Reward-Free RL-Explore with Lemma 2.7.9. Specifically, we divide all possible values of s_{h-1} into two categories: those that can be easily reached by some arbitrary type-agnostic policy

at step $h-1$, and those that are hard to reach by any type-agnostic policy at step $h-1$. For the former, we bound the expected error via Lemma 2.7.9 and multiplicative Chernoff bounds, as Reward-Free RL-Explore ensures $N_{h-1}(s_{h-1}, x_{h-1})$ is sufficiently large for both $x_{h-1} = 0$ and $x_{h-1} = 1$. For the latter, we combine a naive upper bound on the estimation error with the fact that s_{h-1} is hard to reach by any policy. □

With the auxiliary results in mind, we are now ready to state our proof.

Proof of Theorem 2.5.3. We divide the proof into three parts.

ϵ-optimal. By Corollary 2.7.10, the optimal mechanism $\mathcal{B}^*$ is feasible under the relaxed program in Table 2.2 with probability at least $1-\delta$. Conditioned on the event, as $\mathcal{B}^*$ is not necessarily the optimal solution, letting $\widehat{\mathcal{B}}$ denote the output to Algorithm 1, we have

$$\widehat{\textsc{Rev}}(\mathcal{B}^*) \leqslant \widehat{\textsc{Rev}}(\widehat{\mathcal{B}}),$$

where we note that the objective function of the program in Table 2.2 is exactly $\widehat{\textsc{Rev}}$ following the same logic in Lemma 2.4.1. Therefore we know with probability at least $1-\delta$

$$\begin{aligned}\textsc{Rev}(\mathcal{B}^*) - \textsc{Rev}\widehat{\mathcal{B}} &= \textsc{Rev}(\mathcal{B}^*) - \widehat{\textsc{Rev}}(\mathcal{B}^*) + \widehat{\textsc{Rev}}(\mathcal{B}^*) - \widehat{\textsc{Rev}}(\widehat{\mathcal{B}}) \leqslant \textsc{Rev}(\mathcal{B}^*) - \widehat{\textsc{Rev}}(\mathcal{B}^*) \\ &\leqslant (H+1)\epsilon,\end{aligned}$$

where the last inequality is by Lemma 2.7.7 and observing that the revenue comes from only the spend and payment rules. Following the same technique as in Theorem 2.4.2 shows that an ϵ-optimal and feasible solution to the relaxed program in Table 2.2 can be solved in $\widetilde{\mathcal{O}}(\text{poly}(1/\epsilon, N))$ time. We thus complete the proof by noting the additivity of the suboptimality term.

Approximate IC. Conditioned on any $\eta_{(1,h-1)}$, for any s_h and $\widehat{\theta}_h$ we know that the difference between the expected continuation utility under $\widehat{\mathcal{B}}$ and $\bar{\mathcal{B}} = \bar{\mathcal{B}}^{\widehat{y},\widehat{\varphi},\bar{\sigma}}$, i.e. the mechanism

parameterized by the learned allocation rule, learned per-step payment rule, and spend-rule calculated using the ground-truth transition probabilities, is at most

$$
\begin{aligned}
&|\bar{U}_h^{\widehat{\mathcal{B}}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h) - \bar{U}_h^{\bar{\mathcal{B}}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h)| \\
&\quad = \left| \sum_{\tau=h+1}^{H} \mathbb{E}_{x_h, \eta_{(h+1,\tau-1)}, s_\tau, \theta_\tau}[\widehat{\sigma}_\tau - \bar{\sigma}_\tau] \right| \leqslant \sum_{\tau=h+1}^{H} \mathbb{E}_{x_h, \eta_{(h+1,\tau-1)}, s_\tau, \theta_\tau}[|\widehat{\sigma}_\tau - \bar{\sigma}_\tau|].
\end{aligned}
$$

For any potentially untruthful bidding policy $b_{(h,H)}$, consider the following type-agnostic policy π^0.

- From steps 1 to $h-1$, use the type-agnostic policy that maximizes $\Pr_h^\pi(s_h)$, that is, use $\operatorname{argmax}_\pi \Pr_h^\pi(s_h)$.
- From steps h to H, if the type at step h is s_h, then use the type-agnostic policy π that generates the same marginal distribution over $s_{(h+1,H)}$ as $\widehat{y}_{(h,H)}$ when the balance at the h-th step is $\texttt{bal}_h$. Use some arbitrary policy otherwise. Such a policy exists by Lemma 2.2.3.

As π^0 induces the same marginal distribution over $s_{(h+1,H)}$ as $\widehat{\mathcal{B}}$ and $\bar{\mathcal{B}}$, combined with Corollary 2.7.11, the following holds with probability at least $1-\delta$

$$
\begin{aligned}
&|\bar{U}_h^{\widehat{\mathcal{B}}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h) - \bar{U}_h^{\bar{\mathcal{B}}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h)| \\
&\quad \leqslant \max_{\texttt{bal}_{(h,H)}, \theta_{(h,H)}, x_{(h,H)}} \mathbb{E}_{\pi_0}\left[\sum_{\tau=h+1}^{H} |\widehat{\sigma}_\tau - \bar{\sigma}_\tau| \,|\, s_h = s_h \right] \\
&\quad \leqslant \frac{1}{\Pr_h^{\pi_0}(s_h)} \max_{\texttt{bal}_{(1,H)}, \theta_{(1,H)}, x_{(1,H)}} \mathbb{E}_{\pi_0}\left[\sum_{\tau=1}^{H} |\widehat{\sigma}_\tau - \bar{\sigma}_\tau| \right] \leqslant \frac{5c^{1/2}|\mathcal{S}|^{1/2}|\Theta|\epsilon}{\max_\pi \Pr_h^\pi(s_h)},
\end{aligned}
$$

where the last line also uses the definition of π^0.

Noting that $\bar{\mathcal{B}}$ itself is stage-IC and therefore IC by Lemma 2.3.2, we also know that

$$\bar{U}_h^{\bar{\mathcal{B}}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h) \geqslant \bar{U}_h^{\bar{\mathcal{B}}, b_{(h,H)}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h),$$

where $b_{(h,H)}$ is some potentially untruthful bidding policy. Following the same procedure as above, where we now construct some π^1 as follows

- From steps 1 to $h-1$, use the type-agnostic policy that maximizes $\Pr_h^\pi(s_h)$, that is, use $\operatorname{argmax}_\pi \Pr_h^\pi(s_h)$.
- From steps h to H, if the type at step h is s_h, then use the type-agnostic policy π that generates the same marginal distribution over $s_{(h+1,H)}$ as $\widehat{y}_{(h,H)}$ when the balance at the h-th step is $\texttt{bal}_h$ and the buyer reports according to the bidding policy $b_{(h,H)}$. Use some arbitrary policy otherwise. Such a policy exists by Lemma 2.2.3.

Via π^1, the difference between $\bar{U}_h^{\widehat{\mathcal{B}}, b_{(h,H)}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h)$ and $\widehat{U}_h^{\bar{\mathcal{B}}, b_{(h,H)}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h)$ can also be controlled. Therefore we know the following holds with probability at least $1-\delta$ for any potentially untruthful bidding policy

$$\begin{aligned}
&\bar{U}_h^{\widehat{\mathcal{B}}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h) - \bar{U}_h^{\widehat{\mathcal{B}}, b_{(h,H)}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h) \\
&\quad \geqslant -|\bar{U}_h^{\widehat{\mathcal{B}}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h) - \bar{U}_h^{\bar{\mathcal{B}}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h)| \\
&\qquad - |\bar{U}_h^{\widehat{\mathcal{B}}, b_{(h,H)}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h) - \bar{U}_h^{\bar{\mathcal{B}}, b_{(h,H)}}(\eta_{(1,h-1)}, s_h, \widehat{\theta}_h)| \\
&\quad \geqslant -\frac{10}{\max_\pi \Pr_h^\pi(s_h)} c^{1/2} |\mathcal{S}|^{1/2} |\Theta| \epsilon,
\end{aligned}$$

completing the proof for approximate IC.

Approximate ex-post IR. The proof is largely the same as the one provided for approximate IC. Note that the induced $\delta_h = \widehat{u}_h$ and does not need to be estimated. Notice that the mechanism $\bar{\mathcal{B}}$ is also exactly ex-post IR, as $\bar{\sigma}_h$ is exactly calculated according to $\mathcal{P}$, and

therefore $\bar{\mathcal{B}}$ is a core ABAM. As a result, we know that

$$\sum_{h=1}^{H} \widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) \geqslant \sum_{h=1}^{H} \widehat{\delta}(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h) \geqslant \sum_{h=1}^{H} \bar{\sigma}_h(\texttt{bal}_h, \eta_{h-1}, s_h),$$

where the exact derivation follows the proof of Lemma 2.3.3, provided in Appendix 2.5.2. Here we also use the fact that the allocation, per-step payment, and deposit rules are the same for $\widehat{\mathcal{B}}$ and $\bar{\mathcal{B}}$.

We then let $\eta_{(1,H)}$ be arbitrary and fixed. For each h, s_h, similar to the proof for approximate IC, we can construct some type-agnostic policy that maximizes the probability of reaching s_h at step h. As Corollary 2.7.11 holds uniformly over all policies, using the fact that absolute values are non-negative, we know that

$$|\bar{\sigma}_h(\texttt{bal}_h, \eta_{h-1}, s_h) - \widehat{\sigma}_h(\texttt{bal}_h, \eta_{h-1}, s_h)| \leqslant \frac{5}{\max_\pi \Pr_h^\pi(s_h)} c^{1/2} |\mathcal{S}|^{1/2} |\Theta| \epsilon.$$

Therefore for the mechanism $\widehat{\mathcal{B}}$

$$\begin{aligned}\sum_{h=1}^{H} u_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) &= \sum_{h=1}^{H} \widehat{u}_h(\texttt{bal}_h, \eta_{h-1}, s_h, \theta_h; \theta_h) - \sum_{h=1}^{H} \widehat{\sigma}_h(\texttt{bal}_h, \eta_{h-1}, s_h) \\ &\geqslant -\sum_{h=1}^{H} |\bar{\sigma}_h(\texttt{bal}_h, \eta_{h-1}, s_h) - \widehat{\sigma}_h(\texttt{bal}_h, \eta_{h-1}, s_h)| = -\sum_{h=1}^{H} \frac{5c^{1/2}|\mathcal{S}|^{1/2}|\Theta|\epsilon}{\max_\pi \Pr_h^\pi(s_h)},\end{aligned}$$

completing the proof. □

Proof of Lemma 2.7.7

As the proof of the two inequalities are largely the same, we focus on showing the first inequality holds, as proving the second inequality only needs repeating the arguments used for proving the first.

We show that the estimation error of $\mathbb{E}_{\eta_{(1,H)}}\left[\sum_{h=1}^{H} \sigma_h(\beta_h, \eta_{h-1}, s_h)\right]$ can be bounded

by estimation errors of some V-function, which can in turn be controlled by Lemma 2.7.2. Observe that

$$\left|\mathbb{E}_{\eta_{(1,H)}}\left[\sum_{h=1}^{H}\sigma_h(\beta_h,\eta_{h-1},s_h)\right]-\widehat{\mathbb{E}}_{\eta_{(1,H)}}\left[\sum_{h=1}^{H}\sigma_h(\beta_h,\eta_{h-1},s_h)\right]\right|$$
$$\leqslant \max_{\beta_{(1,H)},\eta_{(1,H)}}\left|\mathbb{E}_{s_{(h,H)}}\left[\sum_{h=1}^{H}\sigma_h(\beta_h,\eta_{h-1},s_h)\right]-\widehat{\mathbb{E}}_{s_{(h,H)}}\left[\sum_{h=1}^{H}\sigma_h(\beta_h,\eta_{h-1},s_h)\right]\right|.$$

Because β_h is a function of $\eta_{(1,h-1)}$, there are only finitely many possible choices for $\beta_{(1,H)}$, and there must be some $\beta^*_{(1,H)}$ that maximizes the expression above. Similarly we can find some $\eta^*_{(1,H)}$. We note that we do not require these values to be valid. In other words, we do not require

$$\beta_{*h+1}=g_h(\eta^*_{(1,h-1)},s_h,\theta^*_h)$$

nor do we require $\eta^*_h=(s_h,\theta^*_h,x^*_h)$ for some θ^*,x^*_h. The goal of $\beta^*_{(1,H)},\eta^*_{(1,H)}$ is merely to find a pair of sufficiently adversarial β and histories that maximizes the estimation error.

As a result, plugging $\beta^*_{(1,H)},\eta^*_{(1,H)}$ back into the inequality gives us

$$\left|\mathbb{E}_{\eta_{(1,H)}}\left[\sum_{h=1}^{H}\sigma_h(\beta_h,\eta_{h-1},s_h)\right]-\widehat{\mathbb{E}}_{\eta_{(1,H)}}\left[\sum_{h=1}^{H}\sigma_h(\beta_h,\eta_{h-1},s_h)\right]\right|$$
$$\leqslant\left|\mathbb{E}_{s_{(h,H)}}\left[\sum_{h=1}^{H}\sigma_h(\beta^*_h,\eta^*_{h-1},s_h)\right]-\widehat{\mathbb{E}}_{s_{(h,H)}}\left[\sum_{h=1}^{H}\sigma_h(\beta^*_h,\eta^*_{h-1},s_h)\right]\right|$$
$$\leqslant\max_{\pi}\left|\mathbb{E}_{s_1\sim\mathcal{P}^{\mathcal{S}}_0}[V_1^{\pi,\sigma_h(\beta^*_h,\eta^*_{h-1},\cdot)}(s_1)]-\mathbb{E}_{s_1\sim\mathcal{P}^{\mathcal{S}}_0}[\widehat{V}_1^{\pi,\sigma_h(\beta^*_h,\eta^*_{h-1},\cdot)}(s_1)]\right|\leqslant H\epsilon,$$

where the last inequality is by Lemma 2.7.2, noting that $\sigma_h(\beta^*_h,\eta^*_{h-1},\cdot)$'s range is in $[0,H]$ and the lemma holds *uniformly* over all such functions. The bound for cumulative estimation errors in the payment rule holds similarly, by picking an adversarial combination of $\beta_{(1,H)},\eta_{(1,H)},\theta_{(1,H)}$, ignoring whether or not such a combination is possible under the ABAM $\mathcal{B}^{g,y}$. □

Proof of Lemma 2.7.9

Focus on an arbitrary and fixed history $\eta_{(1,h-1)}$. Taking the expectation over s_h, θ_h, we have

$$\begin{aligned}
&|\mathbb{E}_{s_h,\theta_h \,|\, s_{h-1},x_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)] - \widehat{\mathbb{E}}_{s_h,\theta_h \,|\, s_{h-1},x_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)]| \\
&\quad \leqslant \| \Pr(x_{h-1}, s_h, \theta_h \,|\, s_{h-1}, y_{h-1}) - \widehat{\Pr}(x_{h-1}, s_h, \theta_h \,|\, s_{h-1}, y_{h-1})\|_1 \\
&\quad = \sum_{s_h,\theta_h} |\Pr(x_{h-1}, s_h, \theta_h \,|\, s_{h-1}, y_{h-1}) - \widehat{\Pr}(x_{h-1}, s_h, \theta_h \,|\, s_{h-1}, y_{h-1})|.
\end{aligned}$$

As the inequality above uses only the fact that $\widehat{u}_h$ is in the interval $[0, 1]$, it holds for all possible values of β_h and η_{h-1}, as the function remains bounded regardless of the values of these parameters. Expanding the probabilities shows for any s_h, θ_h

$$\begin{aligned}
&|\Pr(x_{h-1}, s_h, \theta_h \,|\, s_{h-1}, y_{h-1}) - \widehat{\Pr}(x_{h-1}, s_h, \theta_h \,|\, s_{h-1}, y_{h-1})| \\
&\quad \leqslant |\mathcal{P}_h^{\mathcal{S}}(s_h \,|\, s_{h-1}, x_{h-1})\mathcal{P}_h^{\mathcal{S}}(\theta_h \,|\, s_h) - \widehat{\mathcal{P}}_h^{\mathcal{S}}(s_h \,|\, s_{h-1}, x_{h-1})\widehat{\mathcal{P}}_h^{\mathcal{S}}(\theta_h \,|\, s_h)| \\
&\quad \leqslant \mathcal{P}_h^{\mathcal{S}}(\theta_h \,|\, s_h)|\mathcal{P}_h^{\mathcal{S}}(s_h \,|\, s_{h-1}, x_{h-1}) - \widehat{\mathcal{P}}_h^{\mathcal{S}}(s_h \,|\, s_{h-1}, x_{h-1})| \\
&\qquad + \widehat{\mathcal{P}}_h^{\mathcal{S}}(s_h \,|\, s_{h-1}, x_{h-1})|\mathcal{P}_h^{\mathcal{S}}(\theta_h \,|\, s_h) - \widehat{\mathcal{P}}_h^{\mathcal{S}}(\theta_h \,|\, s_h)|.
\end{aligned}$$

By Dvoretzky–Kiefer–Wolfowitz inequality we know for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for a *specific* choice of h, s_{h-1}, x_{h-1} and *all* choices of β_h, θ_{h-1}

$$\begin{aligned}
&|\mathcal{P}_h^{\mathcal{S}}(s_h \,|\, s_{h-1}, x_{h-1})\mathcal{P}_h^{\mathcal{S}}(\theta_h \,|\, s_h) - \widehat{\mathcal{P}}_h^{\mathcal{S}}(s_h \,|\, s_{h-1}, x_{h-1})\widehat{\mathcal{P}}_h^{\mathcal{S}}(\theta_h \,|\, s_h)| \\
&\quad \leqslant \mathcal{P}_h^{\mathcal{S}}(\theta_h \,|\, s_h)\sqrt{\frac{\log(4/\delta)}{2N_{h-1}(s_{h-1}, x_{h-1})}} + \widehat{\mathcal{P}}_h^{\mathcal{S}}(s_h \,|\, s_{h-1}, x_{h-1})\sqrt{\frac{\log(4/\delta)}{2N_h(s_h)}}.
\end{aligned}$$

Summing over s_h, θ_h gives us

$$|\mathbb{E}_{s_h,\theta_h \,|\, s_{h-1},x_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)] - \widehat{\mathbb{E}}_{s_h,\theta_h \,|\, s_{h-1},x_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)]|$$
$$\leqslant |\mathcal{S}|\sqrt{\frac{\log(4|\mathcal{S}||\Theta|/\delta)}{2N_{h-1}(s_{h-1}, x_{h-1})}} + |\Theta| \sum_{s_h} \widehat{\mathcal{P}}_h^{\mathcal{S}}(s_h \,|\, s_{h-1}, x_{h-1}) \sqrt{\frac{\log(4|\mathcal{S}||\Theta|/\delta)}{2N_h(s_h)}}.$$

Because

$$\widehat{\mathcal{P}}_h^{\mathcal{S}}(s_h \,|\, s_{h-1}, x_{h-1}) = \frac{N_{h-1}(s_{h-1}, x_{h-1}, s_h)}{N_{h-1}(s_{h-1}, x_{h-1})}$$

and clearly $N_h(s_h) \geqslant N_{h-1}(s_{h-1}, x_{h-1}, s_h)$, by Cauchy-Schwarz inequality

$$\sum_{s_h} \widehat{\mathcal{P}}_h^{\mathcal{S}}(s_h \,|\, s_{h-1}, x_{h-1}) \sqrt{\frac{\log(4|\mathcal{S}||\Theta|/\delta)}{2N_h(s_h)}} \leqslant \sqrt{\sum_{s_h} \widehat{\mathcal{P}}_h^{\mathcal{S}}(s_h \,|\, s_{h-1}, x_{h-1}) \frac{\log(4|\mathcal{S}||\Theta|/\delta)}{2N_h(s_h)}}$$
$$\leqslant \sqrt{\sum_{s_h} \frac{N_{h-1}(s_{h-1}, x_{h-1}, s_h)}{N_{h-1}(s_{h-1}, x_{h-1})} \frac{\log(4|\mathcal{S}||\Theta|/\delta)}{2N_{h-1}(s_{h-1}, x_{h-1}, s_h)}} = \sqrt{\frac{|\mathcal{S}| \log(4|\mathcal{S}||\Theta|/\delta)}{2N_{h-1}(s_{h-1}, x_{h-1})}}.$$

Consequently for a specific choice of h, s_{h-1}, x_{h-1} and all β_h, θ_{h-1}, with probability at least $1 - \delta$,

$$|\mathbb{E}_{s_h,\theta_h \,|\, s_{h-1},x_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)] - \widehat{\mathbb{E}}_{s_h,\theta_h \,|\, s_{h-1},x_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)]|$$
$$\leqslant |\mathcal{S}|\sqrt{\frac{\log(4|\mathcal{S}||\Theta|/\delta)}{2N_{h-1}(s_{h-1}, x_{h-1})}} + |\Theta|\sqrt{\frac{|\mathcal{S}| \log(4|\mathcal{S}||\Theta|/\delta)}{2N_{h-1}(s_{h-1}, x_{h-1})}}$$
$$\leqslant 2|\mathcal{S}||\Theta|\sqrt{\frac{\log(4|\mathcal{S}||\Theta|/\delta)}{N_{h-1}(s_{h-1}, x_{h-1})}}.$$

We then take a union bound over all possible values that $\mathcal{P} = \{\mathcal{P}^{\mathcal{S}}, \mathcal{P}^{\Theta}\}$ may take. For any $\epsilon_0 > 0$ the $\|\cdot\|_\infty$-covering number of $\Delta(\Theta)$ is upper bounded by $(1/\epsilon_0)^{|\Theta|}$. Similarly, the $\|\cdot\|_\infty$-covering number of $\Delta(|\mathcal{S}|)$ is no greater than $(1/\epsilon_0)^{|\mathcal{S}|}$. Via a simple discretization and covering-based argument, we know by union bound that there is some constant c such that

the following holds for all possible values of h, s_h, x_{h-1} and all possible $\mathcal{P}$ with probability at least $1-\delta$

$$\begin{aligned}
&|\mathbb{E}_{s_h,\theta_h \,|\, s_{h-1},x_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)] - \widehat{\mathbb{E}}_{s_h,\theta_h \,|\, s_{h-1},x_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)]| \\
&\quad \leqslant cH|\mathcal{S}||\Theta|\sqrt{\frac{\log(cH|\mathcal{S}||\Theta|/\delta)}{N_{h-1}(s_{h-1}, x_{h-1})}}.
\end{aligned}$$

When the bound above holds for all $h, \beta_h, s_{h-1}, \theta_{h-1}, x_{h-1}$ and all $\mathcal{P}$, for all choices of $y_{h-1} \in [0,1]$ we have

$$\begin{aligned}
&|\mathbb{E}_{x_{h-1},s_h,\theta_h \,|\, s_{h-1},y_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)] \\
&\qquad - \widehat{\mathbb{E}}_{x_{h-1},s_h,\theta_h \,|\, s_{h-1},y_{h-1}}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)]| \\
&\quad \leqslant |\mathbb{E}_{s_h,\theta_h \,|\, s_{h-1},x_{h-1}=1}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)] \\
&\qquad - \widehat{\mathbb{E}}_{s_h,\theta_h \,|\, s_{h-1},x_{h-1}=1}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)]| \\
&\qquad + |\mathbb{E}_{s_h,\theta_h \,|\, s_{h-1},x_{h-1}=0}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)] \\
&\qquad\quad - \widehat{\mathbb{E}}_{s_h,\theta_h \,|\, s_{h-1},x_{h-1}=0}[\widehat{u}_h(\beta_h, \eta_{h-1}, s_h, \theta_h; \theta_h)]|,
\end{aligned}$$

where the inequality again holds by Hölder's inequality. Reusing the bound developed for all x_{h-1} completes the proof. □

Proof of Corollary 2.7.11

Let π be arbitrary and fixed. Focus on an arbitrary and fixed $h \in [H]$, and let

$$\begin{aligned}
&\eta^*_{(1,h-1)}, x^*_h, \theta^*_h, s^*_{h+1} \\
&\quad = \operatorname*{argmax}_{\beta_{h+1},x_h,\theta_h,s_{h+1}} \mathbb{E}_{s_h \,|\, \pi}[|\widehat{\sigma}_{h+1}((\eta_{(1,h-1)}, (s_h, \theta_h, x_h)), s_{h+1}) \\
&\qquad\qquad - \bar{\sigma}_h((\eta_{(1,h-1)}, (s_h, \theta_h, x_h)), s_{h+1})|],
\end{aligned}$$

As a result, we know for all $\eta_{(1,h-1)}, x_h, \theta_h s_{h+1}$ and all π

$$\begin{aligned}
&\operatorname*{argmax}_{\beta_{h+1}, x_h, \theta_h, s_{h+1}} \mathbb{E}_{s_h \mid \pi}[|\widehat{\sigma}_{h+1}((\eta_{(1,h-1)}, (s_h, \theta_h, x_h)), s_{h+1}) \\
&\qquad\qquad - \bar{\sigma}_h((\eta_{(1,h-1)}, (s_h, \theta_h, x_h)), s_{h+1})|] \\
&\leqslant \mathbb{E}_{s_h \mid \pi}[|\widehat{\sigma}_{h+1}(\beta^*_{h+1}, (s_h, \theta^*_h, x^*_h), s^*_{h+1}) - \bar{\sigma}_{h+1}(\beta^*_{h+1}, (s_h, \theta^*_h, x^*_h), s^*_{h+1})|],
\end{aligned}$$

Note that the function whose expectation we take only takes s_h as an argument. By Lemma 2.2.3, we let π be the task-agnostic policy that generates the same marginal distribution over s_h. We also let

$$\mathcal{S}_h^+ = \left\{ s \in \mathcal{S} : \max_{\pi} \Pr_h^{\pi}(s) \geqslant \frac{\epsilon}{2|\mathcal{S}|H^2} \right\},$$

and have

$$
\begin{aligned}
&\mathbb{E}_{\eta_{(1,H)}\sim\widehat{\mathcal{B}}}[|\widehat{\sigma}_{h+1}(\eta_{(1,h)}, s_{h+1}) - \bar{\sigma}_{h+1}(\eta_{(1,h)}, s_{h+1})|] \\
&\quad = \sum_{s_h\in\mathcal{S}_h^+} [|\widehat{\sigma}_{h+1}((\eta^*_{(1,h-1)}, (s_h, \theta^*_h, x^*_h)), s^*_{h+1}) \\
&\qquad\qquad - \bar{\sigma}_{h+1}((\eta^*_{(1,h-1)}, (s_h, \theta^*_h, x^*_h)), s^*_{h+1})|] \Pr^\pi_h(s_h) \\
&\quad + \sum_{s_h\notin\mathcal{S}_h^+} [|\widehat{\sigma}_{h+1}((\eta^*_{(1,h-1)}, (s_h, \theta^*_h, x^*_h)), s^*_{h+1}) \\
&\qquad\qquad - \bar{\sigma}_{h+1}((\eta^*_{(1,h-1)}, (s_h, \theta^*_h, x^*_h)), s^*_{h+1})|] \Pr^\pi_h(s_h) \\
&\quad \leqslant \sum_{s_h\in\mathcal{S}_h^+} |\widehat{\sigma}_{h+1}((\eta^*_{(1,h-1)}, (s_h, \theta^*_h, x^*_h)), s^*_{h+1}) \\
&\qquad\qquad - \bar{\sigma}_{h+1}((\eta^*_{(1,h-1)}, (s_h, \theta^*_h, x^*_h)), s^*_{h+1})| \Pr^\pi_h(s_h) + \frac{\epsilon}{2H^2} \\
&\quad \leqslant \Bigg(\sum_{s_h\in\mathcal{S}_h^+} |\widehat{\sigma}_{h+1}((\eta^*_{(1,h-1)}, (s_h, \theta^*_h, x^*_h)), s^*_{h+1}) \\
&\qquad\qquad - \bar{\sigma}_{h+1}((\eta^*_{(1,h-1)}, (s_h, \theta^*_h, x^*_h)), s^*_{h+1})|^2 \Pr^\pi_h(s_h) \Bigg)^{1/2} + \frac{\epsilon}{2H^2}
\end{aligned}
$$

where the second-to-last line is by definition of $\mathcal{S}_h^+$ and the last by Cauchy-Schwarz inequality. Noting that

$$
\widehat{\sigma}_{h+1}((\eta^*_{(1,h-1)}, (s_h, \theta^*_h, x^*_h)), s^*_{h+1}) - \bar{\sigma}_{h+1}((\eta^*_{(1,h-1)}, (s_h, \theta^*_h, x^*_h)), s^*_{h+1})
$$

is entirely induced by estimation error in expected per-step utility, applying Lemma 2.7.9, with probability at least $1-\delta$ we have

$$\begin{aligned}
&\mathbb{E}_{\eta_{(1,H)}\sim\widehat{\mathcal{B}}}[|\widehat{\sigma}_{h+1}(\beta_{h+1},\eta_h,s_{h+1})-\bar{\sigma}_{h+1}(\beta_{h+1},\eta_h,s_{h+1})|] \\
&\quad\leqslant cH|\mathcal{S}||\Theta|\sqrt{\iota}\sqrt{\sum_{s_h\in\mathcal{S}_h^+}((N_h(s_h,0))^{-1/2}+(N_h(s_h,1))^{-1/2})^2\Pr_h^\pi(s_h)}+\frac{\epsilon}{2H^2} \\
&\quad\leqslant cH|\mathcal{S}||\Theta|\sqrt{2\iota}\sqrt{\sum_{s_h\in\mathcal{S}_h^+}(N_h(s_h,0)^{-1}+N_h(s_h,1)^{-1})\Pr_h^\pi(s_h)}+\frac{\epsilon}{2H^2},
\end{aligned} \tag{2.7.9}$$

where the second inequality uses the fact that $(a+b)^2\leqslant 2(a^2+b^2)$. By Theorem 2.7.1

$$\mu_h(s_h,0)\geqslant\frac{1}{4|\mathcal{S}|H}\Pr_h^\pi(s_h),\qquad \mu_h(s_h,1)\geqslant\frac{1}{4|\mathcal{S}|H}\Pr_h^\pi(s_h).$$

Therefore, for any $s_h\in\mathcal{S}_h^+, x_h\in\{0,1\}$, we know

$$N_h(s_h,x)^{-1}\Pr_h^\pi(s_h)\leqslant 4|\mathcal{S}|H\frac{\mu_h(s_h,x)}{N_h(s_h,x)}.$$

The claim then holds by applying multiplicative Chernoff bound and recalling our choice of N. More specifically, we know that $N\geqslant c\frac{H^5|\mathcal{S}|^2}{\epsilon^2}\log\left(\frac{c|\mathcal{S}|H}{\delta\epsilon}\right)$. Therefore, for any h,s_h,x_h

$$\mathbb{E}_{\mu_h(s_h,x_h)}[N_h(s_h,x_h)]=\mu_h(s_h,x_h)N\geqslant\frac{cH^3|\mathcal{S}|}{\epsilon}\log\left(\frac{c|\mathcal{S}|H}{\delta\epsilon}\right).$$

Taking a union bound over all h, s_h, and $x_h \in [0,1]$, we know by one-sided multiplicative Chernoff bound that

$$\begin{aligned}
&\Pr\left(\exists(h, s_h, x_h) \text{ s.t. } \frac{\mu_h(s_h, x_h)}{N_h(s_h, x_h)} \geqslant 4N^{-1}\right) \\
&\quad = \Pr(\exists(h, s_h, x_h) \text{ s.t. } N_h(s_h, x_h) \leqslant \mu_h(s_h, x_h)N/4) \\
&\quad \leqslant 2|\mathcal{S}|H \exp\left(-\frac{9cH^3|\mathcal{S}|}{32\epsilon}\log\left(\frac{c|\mathcal{S}|H}{\delta\epsilon}\right)\right) \\
&\quad \leqslant \delta \exp\left(-\frac{9cH^3|\mathcal{S}|}{32\epsilon}\log\left(\frac{1}{\epsilon}\right)\right) \leqslant \delta,
\end{aligned}$$

where the last line holds as long as c, $|\mathcal{S}|$, H are sufficiently large.

By the union bound, we then know with probability at least $1-\delta$, we have for all h, s_h, x_h that

$$N_h(s_h, x_h)^{-1} \Pr_h^{\pi}(s_h) \leqslant 8|\mathcal{S}|HN^{-1} \leqslant \frac{8\epsilon^2}{cH^4|\mathcal{S}|\log(c|\mathcal{S}|H/(\delta\epsilon))},$$

and plugging the bound back into (2.7.9) shows that with probability at least $1-\delta$

$$\begin{aligned}
&\mathbb{E}_{\eta_{(1,H)}}[|\widehat{\sigma}_{h+1}(\beta_{h+1}, \eta_h, s_{h+1}) - \bar{\sigma}_{h+1}(\beta_{h+1}, \eta_h, s_{h+1})|] \\
&\quad \leqslant cH|\mathcal{S}||\Theta|\sqrt{2\iota} \sum_{s_h \in \mathcal{S}_h^+, x_h \in \{0,1\}} \sqrt{N_h(s_h, x_h)^{-1} \Pr_h^{\pi}(s_h)} + \frac{\epsilon}{2H^2} \\
&\quad \leqslant \frac{4|\Theta|\epsilon}{H}\sqrt{\frac{c|\mathcal{S}|}{\log(c|\mathcal{S}|H/(\delta\epsilon))}} + \frac{\epsilon}{2H^2} \leqslant \frac{4|\Theta|\epsilon\sqrt{c|\mathcal{S}|}}{H} + \frac{\epsilon}{2H^2}.
\end{aligned}$$

Finally, summing over all H and noting that the claim holds for arbitrary π completes the proof. □

2.7.6 Auxiliary Results

Lemma 2.7.12 (Piece-wise Linear Approximation, Lemma A.6 in [Mirrokni et al., 2016a])**.** *For any concave function f defined on interval $[a, b]$, if*

- *$f(a)$ and $f(b)$ are given;*
- *there exists $|\beta_a| \leqslant +\infty$ and $|\beta_b| \leqslant +\infty$ such that*

$$f(\xi) \leqslant \beta_a(\xi - a) + f(a),\ f(\xi) \leqslant \beta_b(\xi - b) + f(b);$$

then for any $\kappa > 0$, a pair of lower and upper bounds, $f^{\triangle}$ and $f^{\triangledown}$ can be computed via $O(n)$ queries to the evaluation oracle of f, such that

- *let $\beta = (f(b) - f(a))/(b - a)$, then*

$$n \leqslant \frac{4}{\kappa} + \log \frac{(\beta_a - \beta_b)^2}{(\beta_a - \beta)(\beta_b - b)};$$

- *both of $f^{\triangle}$ and $f^{\triangledown}$ are concave, continuous, and piece-wise linear and have at most $O(n)$ pieces;*
- *the gap between $f^{\triangle}$ and $f^{\triangledown}$ is $\left(\max_{\zeta} f(\zeta) - \min\{f(a), f(b)\}\right)\kappa$.*

Theorem 2.7.13 (Envelope Theorem, Corollary 4 in [Milgrom and Segal, 2002])**.** *Suppose that X is a nonempty compact space, $f(x,t)$ is upper semicontinuous in X, and $\frac{\partial}{\partial t}f(x,t)$ is continuous in (x,t). Let $V(t) = \sup_{x \in X} f(x,t)$. Then*

1. *V is absolutely continuous and $V(t) = V(0) + \int_0^t \frac{\partial}{\partial t}f(x^*(s), s)ds$.*
2. *$V'(t+) = \max_{x \in X^*(t)} \frac{\partial}{\partial t}f(x,t)$ for any $0 \leqslant t < 1$ and $V'(t-) = \min_{x \in X^*(t)} \frac{\partial}{\partial t}f(x,t)$ for any $0 < t \leqslant 1$.*
3. *V is differentiable at a given $t \in (0,1)$ if and only if $\{\frac{\partial}{\partial t}f(x,t) \,|\, x \in X^*(t)\}$ is a singleton, and in that case $V'(t) = \frac{\partial}{\partial t}f(x,t)$ for all $x \in X^*(t)$.*

CHAPTER 3

ONLINE RL FOR REVENUE MAXIMIZING SECOND PRICE AUCTIONS IN MDPS WITH LINEAR FUNCTION APPROXIMATION

3.1 Introduction

Second price auction with reserve prices is one of the most popular auctions both in theory [Nisan et al., 2007] and in practice [Roth and Ockenfels, 2002]. While closed form expressions for the optimal reserve price have been known ever since the seminal work of Myerson [1981], directly applying the result requires 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, opens up a wider range of problems for studying. For example, Dolgov and Durfee [2006] studies optimal auction under the setting and developed novel resource allocation mechanisms, Jiang et al. [2015] leverages both MDP and auctions to better analyze resource allocation in IaaS cloud computing, and Zhao et al. [2018] uses deep Reinforcement Learning (RL) to study sponsored search auctions. We refer interested readers to Athey and Segal [2013] for more motivating examples. A question naturally arises: is it possible to optimize reserve prices when bidders' preferences evolve according to MDPs?

In this article, we provide an affirmative answer. Our work assumes that the state of the

auction is affected by the state and the seller's action in the preceding step. To facilitate interpretation, we refer to the seller's action in this context as "item choice": bidders' later preferences could be affected by the types of items sold in previous rounds, a phenomenon well-documented by empirical works in auctions [Lusht, 1994, Jones et al., 2004, Lange et al., 2010, Ginsburgh and Van Ours, 2007].

As is the case in many real-world problems, we assume that the underlying transition dynamics and the bidder's valuations are both unknown. We further emphasize that we do not make any truthfulness assumptions 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 in the worst case [Kleinberg and Leighton, 2003]. Third, the seller's reward function, namely revenue, is unknown, nonlinear, and can not be directly observed from the bidders' submitted bids and LSVI-UCB cannot be directly applied.

We are able to address all three challenges with the CLUB algortihm. 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.

3.1.1 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, 2014, Keskin et al., 2021, Guo et al., 2022a]. 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 auctions with one or multiple bidders, proposing an algorithm with a $\mathcal{O}(\log\log T)$ worst-case regret bound. However, their setting is non-contextual and they cannot be applied to our setting.

Among this line of research, Golrezaei et al. [2019, 2023] are possibly the closest to our work. Golrezaei et al. [2019] assumes a linear stochastic contextual bandit setting, where the contexts are independent and identically distributed, achieving $\widetilde{\mathcal{O}}(1)$ regret when the market noise distribution is known and $\widetilde{\mathcal{O}}(K^{2/3})$ when it is unknown and nonparametric. While the $\widetilde{\mathcal{O}}(1)$ 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., 2020b]. For unknown distribution, there's another difference that they

consider a time-varying model while we focus on dealing with underlying MDP but fixed. Though the difficulty of these tasks is hard to compare directly, Amin et al. [2014] considers a non-parametric but fixed distribution setting and suffers $\widetilde{O}(K^{2/3})$ regret which may hint at the main difficulty comes from a non-parametric rather than time-varying setting. We delay more discussion about concrete techniques in Golrezaei et al. [2019] in Section 3.6.1. Lastly, as we discussed previously, the approaches in Golrezaei et al. [2019] cannot be directly applied in the MDP setting, necessitating our novel algorithmic structure.

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

RL with Linear Function Approximation. Linear contextual bandit is a popular model for online decision making [Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011, Chu et al., 2011, Li et al., 2019, Lattimore and Szepesvári, 2020] that has also been extensively studied from the auction design perspective [Amin et al., 2014, Golrezaei et al., 2019]. Its dynamic counterpart, Linear MDP, remains popular in the analysis of provably

efficient RL [Yang and Wang, 2019, Jin et al., 2020b, 2021b, Yang and Wang, 2020, Zanette et al., 2020a, 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. [2020b] is one of the first papers to introduce the concept, proposing a provably efficient RL algorithm with $\widetilde{\mathcal{O}}(\sqrt{K})$ 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 inspiration from Wang et al. [2020c] 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. Throughout the whole paper, we use $k \in [K]$ to refer to an episode and $h \in [H]$ to refer to a horizon. In addition, we use $\widetilde{k}$ to refer to a buffer period associated with the k-th episode.

3.2 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 $b_i \in \mathbb{R}_{\geqslant 0}$ and the seller determines the personalized reserve prices for the bidders in the form of reserve price vector $\rho \in \mathbb{R}^N_{\geqslant 0}$, with ρ_i denoting bidder i's reserve price. The bidder with the highest bid only wins

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

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

Motivations for the MDP Model. The core of our setting is to study what will happen when selling heterogeneous goods. We provide three real-world scenarios to motivate this phenomenon.

1. We can easily extend our setting to that of an infinite-horizon MDP by improperly learning the process as an episodic one. Here we focus on the finite-horizon case purely for simplicity of presentation.

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

- **(Online Advertisement)** Google sells lots of advertising positions every day while buyers face budget constraints. In the early rounds, since buyers have more budget left, they are usually eager to bid higher and have a stronger willingness to pay. Therefore, Google may want to sell the most valuable position at first so that buyers have the ability to pay higher acceptable prices and avoid being underbid and unsold.

- **(Antique Auction)** For traditional auction design, the prior auctions may affect the latter auctions. For instance, consider when Sotheby's wants to sell several antiques. The order of selling is of significance and that's the reason why Sotheby's needs to sell a few other pieces to warm up before selling the final flagship piece. The order influences people's valuation and consequently, total revenue. For example, if Sotheby's wishes to auction a valuable Chinese ancient artifact, they would auction some related artifacts during the warm-up session to enhance buyers' expectations.

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

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

Bidder Rewards. We assume that for each bidder $i \in [N]$ at time $h \in [H]$, his reward[3] depends on both the state x and item being auctioned off at that round $\upsilon \in \Upsilon$, which we

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

formalize as

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

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 $r_{ih}(x, \upsilon) \in [0, 3]$.

Policies and Value Functions. Before we describe the seller's policy, we first discuss the action space $\mathcal{A} = \Upsilon \times \mathbb{R}^N_{\geqslant 0}$. At each $h \in [H]$, the seller chooses some action $a_h = (\upsilon_h, \rho_h)$, comprising of item choice $\upsilon \in \Upsilon$ and reserve price vector $\rho \in \mathbb{R}^N_{\geqslant 0}$. The seller's policy is then $\pi = \{\pi_h\}_{h=1}^H$, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$. We let π^υ 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

$$R_h = \mathbb{E}_{\{z_{ih}\}_{i=1}^N} \left[\sum_{i=1}^N m_{ih} \, \mathbb{1}(m_{ih} \leqslant b_{ih}) \right] \tag{3.2.1}$$

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 $b_{ih} \geqslant m_{ih}$. 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 υ instead of reserve ρ, we have a family of mappings from $\mathcal{S} \times \Upsilon$ to $\mathbb{R}^N_{\geqslant 0}$ that determines ρ. Therefore,

with a slight abuse of notation, we can rewrite our Q-function as $Q(x,a) = Q(x,(\upsilon,\rho(x,\upsilon)))$, restricting the role of setting reserve prices using such mappings without loss of generality. From now on, 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

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

respectively. Finally, we let $\pi^\star$ 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

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

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

$$\mathrm{Regret}(K) = \sum_{k=1}^{K} \mathrm{SubOpt}_k(\pi_k). \tag{3.2.2}$$

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 the sake of simplicity, we assume γ is common knowledge. Drutsa [2020] consider a scenario where γ is unknown but with a strictly less than one upper bound. We highlight that it works with our CLUB algorithm as well as long as we replace γ with its upper bound. We can have regret bounds with the same order because we adopt more conservative estimators and

buyers won't violate as much as the corresponding results of γ. Then all results in our paper hold up to some changes of absolute constants. Bidder i's utility at step h is given by $(r_{ih}(s_h, \nu_h) - m_{ih})\,\mathbb{1}(b_{ih} \geqslant m_{ih})$, as we note that he only receives nonzero utility upon winning the auction. His objective is to maximize his discounted cumulative utility

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

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

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

Assumption 3.2.1. *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*

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

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

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

We 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. [2020b] for representative works).

3.3 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 3.4. The third challenge is performing provably efficient RL even when the seller's per-step revenue, detailed in (3.2.1), 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 regret when K is unknown or infinite. Since we can utilize the doubling trick to partition K into at most $[\log_2 K] + 1$, adding corresponding regret will lead to a regret bound of the same order up to some logarithmic terms.

3.3.1 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 2. For each $h \in [H]$, π_{rand} randomly chooses an item and a bidder, offering him the item with a reserve price drawn uniformly at random. The bidder's utility decreases whenever he reports untruthfully, risking either not receiving the item when he underbids, or overpaying for an item when he overbids. Combining lazy updates (see Algorithm 3), we can ensure approximate truthfulness because with the discount rate being less than one, the benefit the bidder gains from misreporting the bids will decay as timestep increases. However, since we consider multi-phase auction design, it remains some nuisance introduced by MDP. For instance, there is no guarantee of a positive definite covariance matrix and it's challenging to give a low regret union bound. We will see how to solve them in the following paragraphs.

Algorithm 2 Definition of π_{rand}

1: **for** $h = 1, \ldots, H$ **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**

We further introduce a novel technique, "buffer period", which explicitly forces the bidders to wait before starting a new learning period, thereby decreases the discounted utility the impatient bidders may gain from untruthfulness. Indeed, a typical algorithm in bandit setting only features π_{rand} and a sequence of learning periods that double in length [Amin et al., 2014, Golrezaei et al., 2019, Deng et al., 2020]. In the bandit setting, data collected in all previous periods is used to update the policy at the end of each period. The increasingly lengthy periods ensure that the seller switches policy less frequently, ensuring that the impatient

buyers need to wait longer before benefiting from untruthful reporting, deterring them from doing so. Unfortunately, the same technique does not work for MDPs, as the rate at which the smallest eigenvalue of the covariance matrix estimate grows cannot be determined and we cannot ensure our estimate of the underlying environment is not "stale" when we double the length of the periods.

Algorithm 3 Buffer Period with Known $F(\cdot)$

1: Receives buffer start $\texttt{buffer.s}(\widetilde{k}+1) = k$ and end $\texttt{buffer.e}(\widetilde{k}+1) = k + \frac{3\log K}{\log(1/\gamma)}$.
2: Do nothing for all episodes $\texttt{buffer.s}(\widetilde{k}+1) \leqslant k < \texttt{buffer.e}(\widetilde{k}+1)$, 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 5, and then increment buffer period counter $\widetilde{k} \leftarrow \widetilde{k}+1$.

While we can mimic the aforementioned bandit algorithms by drawing inspiration from low-switching cost RL literature, we cannot guarantee that the periods are sufficiently long without buffer periods. Indeed, we can use the smallest eigenvalue of the covariance matrix to determine when to start a new period. However, it is impossible to determine a priori the rate at which the smallest eigenvalue grows. Buffer periods ensure that each period is sufficiently long, deferring any utility gain from untruthful reporting. Combined with the bidders' discount rate, a combination of π_{rand} and buffer periods ensure that the bidders behave approximately truthfully. The technique is detailed in Algorithm 3.

With buffer periods defined, we summarize CLUB's update schedule in Algorithm 4 and include Figure 3.1 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 $\texttt{buffer.e}(\widetilde{k})$ is an integer, as rounding up $\texttt{buffer.e}(\widetilde{k})$ does not affect asymptotic regret. Unlike a typical low switching cost RL algorithm, Algorithm 5 further delays updating for $\frac{3\log K}{\log(1/\gamma)}$ episodes after the switching criterion in line 4 is satisfied.

The mixture policy sufficiently punishes untruthfulness. Combined with buffer periods

Algorithm 4 Contextual-LSVI-UCB-Buffer (CLUB) with Known F

1: Initialize policy estimate π_0, buffer period counter $\widetilde{k} = 0$, buffer period starting points $\texttt{buffer.s}(0) = 1$, and buffer period end points $\texttt{buffer.e}(0) = 1$.
2: **for** episodes $k = 1, \ldots, K$ **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 $(\Lambda_h^{\texttt{buffer.e}(\widetilde{k})})^{-1} \not\leq 2(\Lambda_h^k)^{-1}$, schedule a new buffer period starting at $\texttt{buffer.s}(\widetilde{k}+1) = k$ and ending at $\texttt{buffer.e}(\widetilde{k}+1) = k + \frac{3\log K}{\log(1/\gamma)}$ using Algorithm 3, and set $k \leftarrow \texttt{buffer.e}(\widetilde{k}+1)$.
5: **end for**

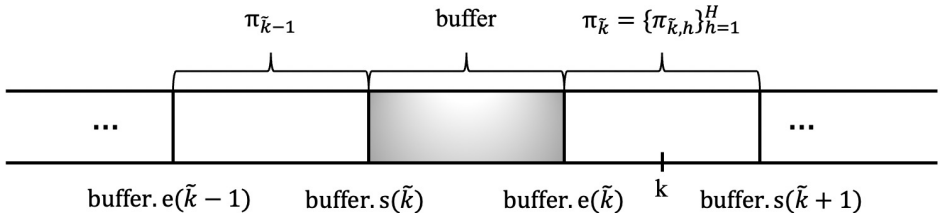


Figure 3.1: Learning periods and buffer periods: $\texttt{buffer.s}(\cdot)$ and $\texttt{buffer.e}(\cdot)$ represent the start point and the end point of a buffer respectively. Episode k lays between $\texttt{buffer.e}(\widetilde{k})$ and $\texttt{buffer.s}(\widetilde{k}+1)$ and the length of each buffer is $\frac{3\log K}{\log(1/\gamma)}$.

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

With the techniques discussed above, namely Algorithms 2, 3, and 4, 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 Challenges 2 and 3: Regret Minimization and Nonlinear Revenue. Having shown that our algorithm punishes untruthful behavior, we begin by showing that the resulting reports are sufficiently truthful for obtaining accurate parameter estimates. It's still quite intricate as regret depends on both state, action related to transition kernel and reserve prices. Traditional point estimation with uncertainty quantity is not enough since we need to not only combine the structure of underlying MDP and coordinate with buffer periods, the so-called lazy updates, but also consider the small proportion of untruthful bids. Whereas LSVI-UCB directly learns from empirical rewards, here we use indicators q_{ih}^k, which we recall is one if bidder i receives the item at episode k step h and zero otherwise. As we cannot guarantee that the empirical covariance matrix is positive definite, existing techniques in Amin et al. [2014], Golrezaei et al. [2019] cannot be applied. We instead have

$$\widehat{\theta}_{ih} = \underset{\|\theta\| \leqslant 2\sqrt{d}}{\operatorname{argmin}} \sum_{\tau=1}^{\texttt{buffer.e}(\widetilde{k}+1)} (q_{ih}^{\tau} - 1 + F(m_{ih}^{\tau} - 1 - \langle \phi(x_h^{\tau}, \upsilon_h^{\tau}), \theta \rangle))^2, \tag{3.3.1}$$

where ρ_{ih}^{τ} is agent i's reserve price and $m_{ih}^{\tau} = \max\{\max_{j \neq i} b_{ih}^{\tau}, \rho_{ih}^{\tau}\}$. (3.3.1) is justified by the observation that, assuming that he bids truthfully, bidder i wins the auction with probability $1 - F(m_{ih}^{\tau} - 1 - \langle \phi(x_h^{\tau}, \upsilon_h^{\tau}), \theta \rangle)$, conditioned on $x_h^{\tau}, \upsilon_h^{\tau}$, and m_{ih}^{τ}. Controlling the uncertainty around $\widehat{\theta}_{ih}$ 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, $\widehat{\theta}_{ih}$ 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 5.

We let $\widetilde{b}^+$ and ρ^+ denote the highest truthful bid and the highest reserve price, respectively. Similarly, let $\widetilde{b}^-$ and ρ^- denote the second-highest. Algorithm 5 estimates the

Algorithm 5 Estimation of $\widehat{Q}_h^{\pi_{\widetilde{k}+1}}(\cdot,\cdot)$

1: Estimate $\widehat{\theta}_{ih}$ using (3.3.1) 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 $\widehat{R}_h(\cdot,\cdot) \leftarrow \mathbb{E}[\max\{\widetilde{b}_h^-(\cdot,\cdot), \widehat{\rho}_h^+(\cdot,\cdot)\}\, \mathbb{1}(\widetilde{b}_h^+(\cdot,\cdot) \geqslant \widehat{\rho}_h^+(\cdot,\cdot))]$.
4: **for** $h = H, \ldots, 1$ **do**
5: $\quad \Lambda_h \leftarrow \sum_{\tau=1}^{\text{buffer.e}(\widetilde{k}+1)} \phi(x_h^\tau, v_h^\tau)\phi(x_h^\tau, v_h^\tau)^T + \lambda I$.
6: $\quad \omega_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{\text{buffer.e}(\widetilde{k}+1)} \phi(x_h^\tau, v_h^\tau)[\max_v \widehat{Q}_{h+1}(x_{h+1}^\tau, v)]$.
7: $\quad \widehat{Q}_h^{\pi_{\widetilde{k}+1}}(\cdot,\cdot) \leftarrow \min\{\omega_h^T \phi(\cdot,\cdot) + \widehat{R}(\cdot,\cdot) + \text{poly}(\log K)\|\phi(\cdot,\cdot)\|_{\Lambda_h^{-1}}, 3H\}$.
8: $\quad \pi^v_{\widetilde{k}+1,h}(\cdot) \leftarrow \operatorname{argmax}_v \widehat{Q}_h^{\pi_{\widetilde{k}+1}}(\cdot, v)$.
9: $\quad \pi^{\rho_i}_{\widetilde{k}+1,h}(\cdot) \leftarrow \widehat{\rho}_{ih}(\cdot, \pi^v_{\widetilde{k}+1,h}(\cdot))$.
10: **end for**
11: Return $\{\widehat{Q}_h^{\pi_{\widetilde{k}+1}}(\cdot,\cdot)\}_{h=1}^H$ and $\{\pi_{\widetilde{k}+1,h}(\cdot)\}_{h=1}^H$.

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 (3.3.1) to estimate all θ_{ih}, which in turn gives estimates for bidders' rewards in the form of $\widehat{\mu}_{ih}$. We then feed the reward function estimates to line 2, yielding an estimate for the optimal reserve price. With Algorithms 2, 3, and 4, the effects of untruthful reports are controlled, and we can ensure that the revenue estimate is sufficiently close to the ground truth. With ρ_{ih} estimated, we then obtain revenue estimates for all states and item choices via line 3. Consequently, we decide both nearly optimal reserve prices and the order of items, addressing the second challenge of regret minimization.

While the rest of Algorithm 5 resembles a typical LSVI-UCB algorithm [Jin et al., 2020b], we highlight several key differences. First, we use the plug-in revenue estimate, whereas existing works estimate the Q-function with the empirically observed rewards. To accommodate the plug-in estimate, here ω_h estimates $\mathbb{P}_h V_{h+1}$, the transition operator applied to the V-function, as opposed to $\mathbb{B}_h V_{h+1}$, which uses the Bellman evaluation operator instead. Lastly, in line 7 we link the uncertainty of revenue to the uncertainty bonus typically seen in linear MDPs, thereby obtaining an optimistic estimate of the Q-function induced by revenue. We

conjecture the transition estimation procedure 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. By combining the loss from incentivizing truthful mechanism and learning underlining model to set reserve prices, we get the final Algorithm 4, which explores efficiently and achieves the following regret upper bound, and then addresses the second challenge of regret minimization.

3.3.2 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 3.3.1. *Market noise pdf f is bounded, i.e. there exist constants c_1, C_1 such that $c_1 \leqslant f \leqslant C_1$.*

Assumption 3.3.2. *Market noise pdf f is differentiable and its derivative is bounded. That is, there exists a constant L such that $|f'| \leqslant L$.*

Assumption 3.3.3. *Market noise cdf $F(\cdot)$ and $1 - F(\cdot)$ are log-concave.*

At a high level, Assumptions 3.3.1 and 3.3.2 ensure that the pdf f is generally well-behaved, namely, bounded and smooth. Assumption 3.3.3 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].

Remark 3.3.4. *We note that Assumption 3.3.3 is in fact made redundant by Assumption 3.3.1 because we have a quite "smooth" distribution with bounded differential. Then, once we have a good estimation for the parameters, "smooth" $F(\cdot)$ leads to a good estimation of the reward function. Nevertheless, we retain this assumption as it streamlines our proof by avoiding discussion of market stability with multi-optimal reserve prices and getting bogged down in tedious regret decomposition.*

We are now ready to state our results. If we set $\text{poly}(\log K) = C_7 + C_6 H \log^2 K$ in Algorithm 5, where constant C_6 is determined in Theorem 3.6.7 and constant $C_7 = B_8 H^{\frac{3}{2}} \log K$ with constant B_8 determined in Theorem 3.6.31, then we have Theorem 3.3.5.

Theorem 3.3.5. *Under Theorem 3.2.1, 3.3.1, 3.3.2 and 3.3.3, for any fixed failure probability $\delta \in (0, 1)$, with probability at least $1 - \delta$, Algorithm 4 achieves at most $\widetilde{\mathcal{O}}(\sqrt{H^5 K})$ revenue regret, where $\widetilde{\mathcal{O}}(\cdot)$ hides only absolute constants and logarithmic terms.*

Proof. See Section 3.6.2 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.

3.4 Unknown Market Noise Distribution

We now discuss when the market noise distribution is unknown. Recall from previous discussions that our second challenge lies in minimizing revenue regret when the market noise distribution is unknown. Existing techniques, similar to the one in Golrezaei et al. [2019], incorporate pure exploration rounds to address the challenge, yet necessitates a $\widetilde{\mathcal{O}}(K^{2/3})$

revenue regret. In this section, we instead introduce a novel technique dubbed "simulation", which eliminates the need for pure exploration rounds and achieves instead a $\widetilde{\mathcal{O}}(\sqrt{K})$ regret. While the first and third challenges have been previously addressed, the approaches in Section 3.3 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.

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

Similarly, there are three steps to do auction design when $F(\cdot)$ is unknown. First, we leverage Algorithm 2 and Algorithm 6 to motivate an approximately truthful mechanism. Second, we utilize Algorithm 9 in coordination with newly proposed Algorithm 8 to estimate underlying MDP and set reserve prices. We motivate truthfulness through buffer periods and quantify the uncertainty by constructing corresponding ellipsoid bounds. Finally, we add up all these uncertainties and minimize regret with high probability.

Addressing Challenge 1: Untruthfulness. When the market noise distribution is unknown, the techniques used in Section 3.3 cannot be applied directly, necessitating careful adaptations. We summarize the changes to these techniques, beginning by introducing Algorithm 6, the counterpart to Algorithm 3, 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 6 Buffer Period with Unknown $F(\cdot)$

1: Receives buffer start $\texttt{buffer.s}(\widetilde{k}+1) = k$ and end $\texttt{buffer.e}(\widetilde{k}+1) = k + \frac{3\log K}{\log(1/\gamma)}$.
2: Do nothing for all episodes $\texttt{buffer.s}(\widetilde{k}+1) \leqslant k < \texttt{buffer.e}(\widetilde{k}+1)$, 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 9, and then increment buffer period counter $\widetilde{k} \leftarrow \widetilde{k}+1$.

We then discuss Algorithm 7, a close variant of Algorithm 4, whose biggest change lies

in the update schedule in line 4. Algorithm 4 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 7 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 7 Contextual-LSVI-UCB-Buffer (CLUB) with Unknown F

1: Initialize policy estimate π_0, buffer period counter $\widetilde{k} = 0$, buffer period starting points $\texttt{buffer.s}(0) = 1$, and buffer period end points $\texttt{buffer.e}(0) = 1$.
2: **for** episodes $k = 1, \ldots, K$ **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 $(\Lambda_h^{\texttt{buffer.e}(\widetilde{k})})^{-1} \not\leq 2(\Lambda_h^k)^{-1}$ or $\log_2(k)$ is an integer, schedule a new buffer period starting at $\texttt{buffer.s}(\widetilde{k}+1) = k$ and ending at $\texttt{buffer.e}(\widetilde{k}+1) = k + \frac{3\log K}{\log(1/\gamma)}$ using Algorithm 6, and set $k \leftarrow \texttt{buffer.e}(\widetilde{k}+1)$.
5: **end for**

Similar to Section 3.3, 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 (3.3.1) 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 revenue regret when $F(\cdot)$ is unknown.

Addressing Challenge 2: Regret Minimization. An intuitive way to incorporate the previous observation is to simply perform pure exploration rounds with π_{rand}, similar to the technique in Golrezaei et al. [2019]. However, doing so incurs $\widetilde{\mathcal{O}}(K^{2/3})$ revenue regret,

as π_{rand} does not set the reserve prices optimally and we are not exploring and exploiting simultaneously. To balance exploration and exploitation, we propose a new technique that we dub "simulation", which allows us to continue exploiting with the mixture policy.

Algorithm 8 Simulation

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

Here we introduce a new random variable $\widetilde{q}_{ih}^{\tau} = \mathbb{1}(b_{ih}^{\tau} \geqslant \widetilde{\rho}_{ih}^{\tau})$, where for each h, τ we select one $i \in [N]$ uniformly at random and then draw $\widetilde{\rho}_{ih}^{\tau}$ from $\text{Unif}([0, 3])$. 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 υ_h^{τ} constant, what would be the outcome if we were to act according to π_{rand} instead? As we do not need to execute π_{rand}, revenue regret can be decreased. Furthermore, $\widetilde{q}_{ih}^{\tau}$ 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 $\widetilde{q}_{ih}^{\tau}$ changes could be controlled effectively.

More technically, Algorithm 8 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 (3.3.1), 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 π_{rand}. Even when the bidders report truthfully, we cannot guarantee that $\Pr(q_{ih}^{\tau} = 1 \,|\, x_h^{\tau}, \upsilon_h^{\tau})$ could be related to $\frac{1}{3N}(1 + \langle \phi(x_h^{\tau}, \upsilon_h^{\tau}), \theta_{ih} \rangle)$. Consequently, it would be hard to ensure that when all bidders are truthful, the estimator $\widehat{\theta}_{ih}^{\tau}$ would converge to θ_{ih}.

Addressing Challenge 3: Nonlinear Revenue. With the first challenge addressed by carefully adjusting techniques in Section 3.3 and the second by the simulation technique

detailed in Algorithm 8, 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 (3.4.1).

$$\begin{aligned}\widehat{\theta}_{ih} &= \operatorname*{argmin}_{\|\theta\|\leqslant 2\sqrt{d}} \sum_{\tau=1}^{\texttt{buffer.e}(\widetilde{k}+1)} (3N\widetilde{q}_{ih}^{\tau} - (1+\langle\phi(x_h^\tau, v_h^\tau), \theta\rangle))^2, \\ \widehat{F}(z) &= \frac{1}{N\texttt{buffer.e}(\widetilde{k}+1)H} \sum_{i=1}^{N} \sum_{\tau=1}^{\texttt{buffer.e}(\widetilde{k}+1)} \sum_{h=1}^{H} \mathbb{1}(b_{i\tau h} - 1 - \langle\phi_h^\tau, \widehat{\theta}_{ih}\rangle \leqslant z).\end{aligned} \tag{3.4.1}$$

We note that we are simply using a histogram to estimate $F(\cdot)$ and, as we have successfully decoupled the estimation error of $F(\cdot)$ from that of θ_{ih}, using histogram is sufficient for achieving $\widetilde{\mathcal{O}}(\sqrt{K})$ revenue regret. We then introduce Algorithm 9, whose key difference with Algorithm 5 lies in the added uncertainty due to $\widehat{F}$ and the inclusion of the simulation subroutine. Similar to Section 3.3, the procedure then provides us with sufficiently accurate policy and Q-function estimates, resolving our third and final challenge.

Algorithm 9 Estimation of $\widehat{Q}_h^{\pi_{\widetilde{k}+1}}(\cdot,\cdot)$ with Unknown $F(\cdot)$

Collect simulation outcome $\widetilde{q}$ using Algorithm 8.
Estimate $\widehat{\theta}_{ih}, \widehat{F}(\cdot)$ using (3.4.1).
Estimate $\widehat{\mu}_{ih}(\cdot,\cdot) \leftarrow \langle\phi(\cdot,\cdot), \widehat{\theta}_{ih}\rangle$.
Set reserve price $\widehat{\rho}_{ih}(\cdot,\cdot) = \operatorname{argmax}_y y(1 - \widehat{F}(y - 1 - \widehat{\mu}(\cdot,\cdot)))$.
Estimate revenue $\widehat{R}_h(\cdot,\cdot) \leftarrow \mathbb{E}[\max\{\widetilde{b}_h^-(\cdot,\cdot), \widehat{\rho}_h^+(\cdot,\cdot)\}\,\mathbb{1}(\widetilde{b}_h^+(\cdot,\cdot) \geqslant \widehat{\rho}_h^+(\cdot,\cdot))]$.
for $h = H, \ldots, 1$ **do**
    $\Lambda_h \leftarrow \sum_{\tau=1}^{\texttt{buffer.e}(\widetilde{k}+1)} \phi(x_h^\tau, v_h^\tau)\phi(x_h^\tau, v_h^\tau)^T + \lambda I$. ▷ We set $\lambda = 1$ in this paper.
    $\omega_h \leftarrow \Lambda_h^{-1} \sum_{\tau=1}^{\texttt{buffer.e}(\widetilde{k}+1)} \phi(x_h^\tau, v_h^\tau)[\max_a \widehat{Q}_{h+1}(x_{h+1}^\tau, a)]$.
    $\widehat{Q}_h^{\pi_{\widetilde{k}+1}}(\cdot,\cdot) \leftarrow \min\{\omega_h^T\phi(\cdot,\cdot) + \widehat{R}(\cdot,\cdot) + \mathrm{poly}_1(\log K)\|\phi(\cdot,\cdot)\|_{\Lambda_h^{-1}} + \frac{\mathrm{poly}_2(\log K)}{\sqrt{\texttt{buffer.e}(\widetilde{k}+1)}}, 3H\}$
    $\pi^{v}_{\widetilde{k}+1,h}(\cdot) \leftarrow \operatorname{argmax}_v \widehat{Q}_h^{\pi_{\widetilde{k}+1}}(\cdot, v)$.
    $\pi^{\rho_i}_{\widetilde{k}+1,h}(\cdot) \leftarrow \widehat{\rho}_{ih}(\cdot, \pi^{a}_{\widetilde{k}+1,h}(\cdot))$.
end for
Return $\{\widehat{Q}_h^{\pi_{\widetilde{k}+1}}(\cdot,\cdot)\}_{h=1}^H$ and $\{\pi_{\widetilde{k}+1,h}(\cdot)\}_{h=1}^H$.

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 3.3, 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.

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

We now argue that Algorithm 7 achieves $\widetilde{\mathcal{O}}(\sqrt{K})$ regret. We begin with a slight detour, making a basic assumption on the hypothesis class for $F(\cdot)$.

Assumption 3.4.1. *The market noise distribution $F(\cdot)$ belongs to a distribution family $\mathcal{F}$.*

We further let $\mathcal{N}_\epsilon(\mathcal{F})$ be the ϵ-covering number of $\mathcal{F}$ with respect to the metric that $\text{dist}(F, G) = \sup_x |F(x) - G(x)|$. We now have our main theorem when noise distribution is unknown. If we let $\text{poly}_1(\log K) = C_{15} + C_{13} H \log^2 K$ and $\text{poly}_2(\log K) = C_{14} H^2 \log^4 K$ in Algorithm 9, where $C_{15} = D_7 H^{\frac{3}{2}}$ and the constant D_7 is determined in Theorem 3.6.39, constants C_{13} and C_{14} are determined in Theorem 3.6.16, we would attain the following regret guarantee.

Theorem 3.4.2. *Under Assumptions 3.2.1, 3.3.1, 3.3.2, 3.3.3 and 3.4.1, when $F(\cdot)$ is unknown, for any fixed failure probability $\delta \in (0, 1)$, Algorithm 7 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 Section 3.6.3 for a detailed proof. □

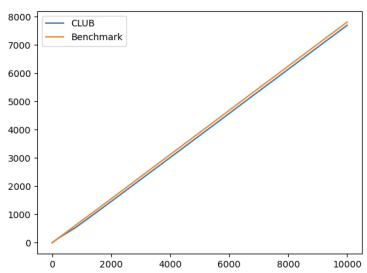
We highlight that when $\mathcal{N}_{1/K}(\mathcal{F})$ is polynomial in $1/K$, an implicit assumption found in Kong et al. [2021], Foster et al. [2021], Jin et al. [2021a], Theorem 3.4.2 shows that Al-

gorithm 7 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 beat the well-known $\Omega(K^{2/3})$ revenue lower bound in Kleinberg and Leighton [2003] with the help of Assumptions 3.3.1 and 3.3.2 for similar but not totally same scenarios to be fair. 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. The way Kleinberg and Leighton [2003] constructs regret lower bound is to find a special case containing no information. As they say "the expected revenue per buyer is a constant independent of the offer price outside the interval of good prices", it provides nothing useful for learning. However, with Theorem 3.3.1, it guarantees the information in each exploration and partial out this extreme situation.

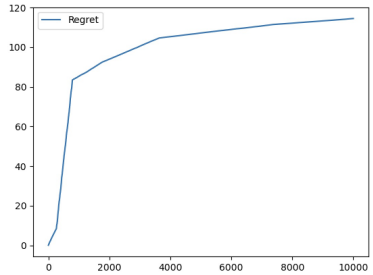
Finally, we highlight that both bounds in Sections 3.3 and 3.4 match corresponding lower bounds with respect to K. From the $\Omega(\sqrt{K})$ lower bound in Jin et al. [2020b], we directly know that results in Theorems 3.3.5 and 3.4.2 match corresponding regret lower bounds as the problem in Jin et al. [2020b] is a subproblem of our problem.

3.5 Numerical Experiments

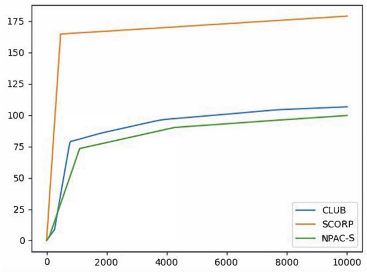
Here, we present numerical simulations to compare the performance of Algorithm 7 with several baseline policies in different settings. To be specific, we compare the performances of CLUB (i.e. Algorithm 7), SCORP [Golrezaei et al., 2019] and NPAC-S [Golrezaei et al., 2023] in contextual bandit settings (i.e. $H = 1$) and the performances of CLUB and NPAC-S in MDP settings. In all experiments, we assume that the noise distribution $F(\cdot)$ is unknown. The numerical experiment written in Python 3.10.9 runs on a laptop with an Apple M2 CPU. All three algorithms use less than 30 seconds to calculate 10000 episodes which shows their practicability in reality. We delay more details in Section 3.6.6.



(a) The performance of CLUB against the benchmark.



(b) The regret accumulation of CLUB.



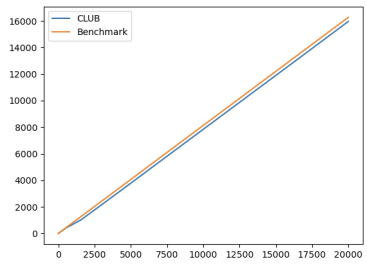
(c) The average performances of three algorithms.

Figure 3.2: Experiment results for contextual bandit settings: Figure 3.2a compares the revenue achieved by CLUB and benchmark, showing CLUB obtains more than 98% revenue. Figure 3.2b shows the sublinear regret associated with our CLUB algorithm as the curve trend is below linear. Figure 3.2c exhibits that CLUB is comparable with NPAC-S, overwhelming SCORP.

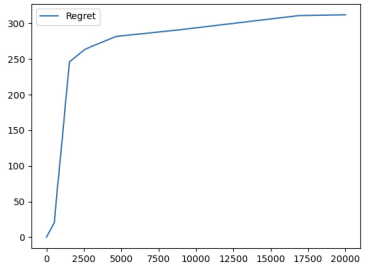
In contextual bandit setting, we set $K = 10000$, $\gamma = 0.9$ for each setting and repeat the procedure for $n = 30$ trails for each algorithm. We show results in Figure 3.2. Figures 3.2a and 3.2b show results in one trial, where we find that CLUB can obtain more than 98% revenue compared with the benchmark, where the underlying model is common knowledge. At the same time, Figure 3.2b testifies $\widetilde{O}(\sqrt{K})$-shaped regret. In Figure 3.2c, we show the average regrets among all 30 trials of these three different algorithms. The average regrets in 30 trials are 106.62, 178.96 and 99.69 respectively. As for the number of winning times, CLUB wins 15 times while NPAC-S wins 14 times. SCORP only wins once. Therefore, we conclude the performances of CLUB and NPAC-S are comparable, overwhelming the performance of SCORP. Since SCORP doesn't work well even in contextual bandit settings, we only compare CLUB and NPAC-S under MDP.

In the MDP setting, we also incorporate $K = 10000$, $H = 2$, $\gamma = 0.9$ and conduct $n = 30$ trails for both two algorithms. We show the corresponding results in Figure 3.3. Our CLUB can obtain more than 98% revenue (c.f. Figure 3.3a) against the benchmark which highlights its great performance. In Figure 3.3b, it's clear to see the $\widetilde{O}(\sqrt{K})$-shaped regret. Among all 30 trails, CLUB wins all 30 times. As for average regrets, it's 203.07 for

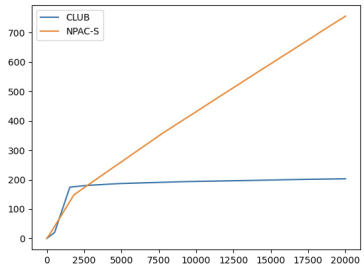
CLUB and 756.31 for NPAC-S. Therefore, we can conclude that under the MDP setting, CLUB sufficiently works better than NPAC-S. Together with experiments under contextual bandits, our experiments are in favor of CLUB algorithms which shows the importance of our newly proposed techniques.



(a) The performance of CLUB against the benchmark.



(b) The regret accumulation of CLUB.



(c) The average performances of two algorithms.

Figure 3.3: Experiment results for MDP settings: Figure 3.3a compares the revenue achieved by CLUB and benchmark, showing CLUB obtains more than 98% revenue. Figure 3.3b shows the sublinear regret associated with our CLUB algorithm as the curve trend is below linear. Figure 3.3c exhibits that compared with NPAC-S, CLUB has less regret testifying its optimality.

3.6 Technical Details

3.6.1 Detailed Comparison with Golrezaei et al. [2019]

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

CORP considers a contextual bandit setting with known noise distribution achieving $\widetilde{O}(1)$ regret. However, as we mentioned before, accommodating underlying MDP, $\Omega(\sqrt{K})$ regret lower bound is inevitable. In Section 3.3, we propose our optimal CLUB algorithm matching the lower bound.

CORP-II considers an unknown but parametric noise distribution and achieves $\widetilde{O}(\sqrt{K})$

regret. However, in Section 3.4, we consider an unknown and non-parametric noise distribution. Therefore, compared with the setting for CORP-II, our model is strictly much harder for the following two-fold reasons. We need to consider extra MDP and non-parametric noise distribution. Moreover, since CORP-II don't have enough horizons to explore, it doesn't work well under our MDP setting and cannot achieve its original $\widetilde{O}(\sqrt{K})$ regret.

SCORP considers time-varying and non-parametric noise distribution achieving $\widetilde{O}(K^{2/3})$ regret. We share the similarity that both of these settings can't be parameterized, which means that we lose the opportunity to utilize some concentration inequalities directly and we need to bypass these obstacles to achieve sublinear regrets. There are two main differences between our model and SCORP. First, the underlying MDP makes our problem harder than the one of SCORP. Second, we consider fixed noise distribution and use a different benchmark making these two models not comparable directly. As a result, our algorithm achieves $\widetilde{\mathcal{O}}(\sqrt{K})$ regret with mild additional assumptions on the shape of $F(\cdot)$ (c.f. Theorem 3.3.1 and 3.3.2). Although it is hard to compare the difficulties between our setting and SCORP in strict order, we believe they have a similar degree of difficulty. As we mentioned in Section 3.1.1, the work [Amin et al., 2014] explores a scenario with a non-parametric yet fixed distribution setting, experiencing a regret of $\widetilde{O}(K^{2/3})$. This observation suggests that the primary challenge might arise from the non-parametric nature of the problem, as opposed to the time-varying setting. Moreover, we should highlight that an $\widetilde{O}(K^{2/3})$ regret is inevitable even though the distribution is fixed corresponding to a saddle point for SCORP, as they spend "too many" episodes to explore while we don't "waste" time to do pure exploration so that balance the exploration-exploitation tradeoff better and achieve better regret bounds. Objectively, our method will suffer $\Omega(K^{2/3})$ regret lower bound for a time-varying model and it's of independent interest for future research.

3.6.2 Omitted Proof in Section 3.3

In this section, we show some useful lemmas in order to prove theorems in Section 3.3. We organize the section as follows. Firstly, we introduce lemmas to bound the effect of untruthful bidding. Then, we will show that we are able to estimate unknown parameters accurately. Finally, combining them leads to bounded regret with high probability.

Useful Lemmas for Proving Theorem 3.3.5

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

Lemma 3.6.1. *Under Theorem 3.2.1 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 3.6.2. *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 $\mathtt{L}$ 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 3.6.2. We formalize the definition of $\mathtt{L}$ for any given i, h as follows.

$$\mathtt{L} = \{k : \mathbb{1}(v_{ih}^k w > \max\{b_{-ih}^{k+}, \rho_{ih}^k\}) \neq \mathbb{1}(b_{ih}^k > \max\{b_{-ih}^{k+}, \rho_{ih}^k\})\}. \tag{3.6.1}$$

Lemma 3.6.3. *With probability at least* $1-\delta$*, it holds that for any given* i*,* h

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

where C_4 is a constant independent of K and H.

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

Lemma 3.6.4. *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 3.6.5. *We use θ^*_{ih} to denote the true parameter and $\widehat{\theta}_{ih}$ to represent the outcome from Equation* (3.3.1) *in episode* $\texttt{buffer.e}(\widetilde{k})$. *Therefore, under Theorem 3.3.1 and Theorem 3.3.2, for any i and h, it holds the following union bound that C_5 is a constant and*

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

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

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

Lemma 3.6.6. *Conditional on Good Event $\mathcal{E}$, it holds that*

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

where $\texttt{buffer.e}(\widetilde{k})$ *is the last episode using Equation* (3.3.1) *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. $R^k_h(\cdot,\cdot) = \sum_{i=1}^N \mathbb{E}[\max\{r^{k-}_{ih}, \alpha^{k*}_{ih}\}\mathbb{1}(r^k_{ih} \geqslant \max\{r^{k-}_{ih}, \alpha^{k*}_{ih}\})]$ and $\widehat{R}^k_h(\cdot,\cdot) = \sum_{i=1}^N \mathbb{E}[\max\{\widehat{r}^{k-}_{ih}, \alpha^{k}_{ih}\}\mathbb{1}(\widehat{r}^k_{ih} \geqslant \max\{\widehat{r}^{k-}_{ih}, \alpha^{k}_{ih}\})]$. In short,

$R(\cdot,\cdot)$ is the expectation of revenue if we choose the optimal reserve price α_{ih}^{k*} for every bidder based on the knowledge of $\mu_{ih}^{k}(\cdot,\cdot)$ and everyone bids truthfully based on his valuation. Respectively, $\widehat{R}(\cdot,\cdot)$ 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 3.6.7. *When Theorem 3.6.6 holds, we have*

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

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 3.6.8. *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,*

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

which is consistent with Theorem 3.6.7.

Now, we focus on the regret not in buffer caused by Algorithm 5, denoted by Δ_1. In order to facilitate the understanding, we rewrite the definition of Δ_1 explicitly as follows.

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

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 3.6.9. *Under Theorem 3.2.1, Theorem 3.3.1 and Theorem 3.3.2, if in Algorithm 5 we set* $\mathrm{poly}(\log K) = (C_7 + C_6 H \log^2 K)\|\phi(\cdot,\cdot)\|_{(\Lambda_h^{\texttt{buffer.e}(\widetilde{k})})^{-1}}$ *, where* $C_7 = B_8 H^{\frac{3}{2}} \log K$ *and* B_8 *is determined in Theorem 3.6.31, it holds that with probability at least* $1 - 2\delta$*,*

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

where C_8 *is a constant independent of* H *and* K*.*

Proof of Theorem 3.3.5

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

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

Δ_1 is defined in Theorem 3.6.9 and with probability at least $1-2\delta$, $\Delta_1 \leqslant C_8\sqrt{H^5 K \log^5 K}$. Δ_2 comes from the use of buffer. With Theorem 3.6.1, it holds that $\Delta_2 \leqslant 3H\frac{3HC_2 \log^2 K}{\log \frac{1}{\gamma}}$. Δ_3 comes from the use of policy π_{rand}. By applying Theorem 3.6.4, it holds that $\Delta_3 \leqslant 3H \max\{4, 1 + \frac{4}{3}\log\frac{1}{\delta}\}$ with probability at least $1-\delta$.

Δ_4 comes from the consequence from the existence of `L`. Due to Theorem 3.6.3, we have $\Delta_4 \leqslant NH(4C_1C_3H + 8\log(\frac{2NH}{\delta}))3H = 3NH^2(4C_1C_3H + 8\log(\frac{2NH}{\delta}))$, 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 $\frac{C_3 H}{K}$ thanks to Theorem 3.6.2. 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 $\frac{C_3 H}{K}$ each step. So, it holds that $\Delta_5 \leqslant C_3 H^2$.

When estimating $\widehat{R}(\cdot,\cdot)$, we have at most probability δ not satisfying the inequality in Theorem 3.6.5.

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

3.6.3 Omitted Proof in Section 3.4

Compared to Section 3.6.2, this section introduces 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.

Useful Lemmas for Proving Theorem 3.4.2

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 kernels[4] 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 3.6.10. *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 $\widehat{F}_t(x) = \frac{1}{t}\sum_{i=1}^{t} \mathbf{1}_{\{m_i \leqslant x\}}$ where $x \in \mathbb{R}$. Then with probability $1-\delta$, it holds*

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

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

Now, similar to the methodology in Section 3.6.2, we state the following lemmas parallelly.

Lemma 3.6.11. *Under Theorem 3.2.1 about linear MDP, it holds that the number of episodes of buffer is not larger than $C_9 H \log^2 K$. Then, the number of corresponding steps is not larger than $C_9 H^2 \log^2 K$, where C_9 is a constant that only depends on d and λ.*

Recall that when market noise distribution is unknown, we implement Algorithm 8 to generate $\widetilde{q}$ and we use $\widetilde{q}$ to estimate θ instead of q. Therefore, $\mathtt{L}$ there considers simulation outcome $\widetilde{q}$ rather than real outcome q. We formalize the definition of $\mathtt{L}$ there as follows and we use $\widetilde{\rho}$ to represent reserve price in Algorithm 8.

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

$$\mathtt{L} = \{k : \mathbb{1}(v_{ih}^k > \max\{b_{-ih}^{k+}, \widetilde{\rho}_{ih}^k\}) \neq \mathbb{1}(b_{ih}^k > \max\{b_{-ih}^{k+}, \widetilde{\rho}_{ih}^k\})\}. \tag{3.6.2}$$

Lemma 3.6.12. *With probability at least $1-\delta$, it holds that for any given i, h*

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

where C_3 is defined in Theorem 3.6.2 and C_{10} is a constant independent of K and H.

Lemma 3.6.13. *We use θ_{ih}^* to denote the true parameter and $\widehat{\theta}_{ih}$ to represent the outcome from Equation* (3.4.1) *in episode* $\mathtt{buffer.e}(\widetilde{k})$. *Therefore, under Theorem 3.3.1 and Theorem 3.3.2, for any i and h, it holds the following union bound that C_{11} is a constant and*

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

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

As same as Theorem 3.6.6, we have the following lemma.

Lemma 3.6.14. *Conditional on Good Event $\mathcal{E}$, it holds that*

$$|\widehat{\mu}_{ih}^k(\cdot,\cdot) - \mu_{ih}^k(\cdot,\cdot)| \leqslant C_{11}\sqrt{H}\log K\|\phi(\cdot,\cdot)\|_{(\Lambda_h^{\texttt{buffer.e}(\widetilde{k})})^{-1}}.$$

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

Lemma 3.6.15. *Conditional on Good Event $\mathcal{E}$, it holds with probability at least $1-\delta$ that for any x in episode* $\texttt{buffer.e}(\widetilde{k})$

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

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 3.6.16. *Conditional on Good Event $\mathcal{E}$, we have*

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

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

We define Δ_1 as the one in Theorem 3.6.9 of Section 3.6.2.

Lemma 3.6.17. *Under Theorem 3.2.1, Theorem 3.3.1 Theorem 3.3.2 and Theorem 3.4.1, if we set* $\mathrm{poly}_1(\log K) = C_{15} + C_{13}H\log^2 K$ *and* $\mathrm{poly}_2(\log K) = C_{14}H^2\log^4 K$ *in Algorithm 9,*

where $C_{15} = D_7 H^{\frac{3}{2}}$ *and* D_7 *is determined in Theorem 3.6.39, it holds that with probability at least* $1 - 2\delta$*,*

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

Proof of Theorem 3.4.2

It is similar to the proof of Theorem 3.3.5. The only difference comes from Theorem 3.6.15. The probability of Bad Event $\mathcal{E}^c$ is now less than 6δ. Then, we set $\delta = \frac{p}{6}$ and it ends the proof. □

3.6.4 Auxiliary Lemmas and Proofs in Section 3.6.2

In this section, we prove the lemmas mentioned in Section 3.6.2 detailedly. It is organized by the order of lemmas.

Proof of Theorem 3.6.1

First of all, we have the following lemmas.

Lemma 3.6.18 (Lemma 2, [Gao et al., 2021])**.** *Assume* $m \leqslant n$*,* $A = \sum_{\tau=1}^{m} \phi_\tau \phi_\tau^T + \lambda I$*.* $B = \sum_{\tau=1}^{n} \phi_\tau \phi_\tau^T + \lambda I$*, where* ϕ_τ *is abridge for* $\phi(x_\tau, \upsilon_\tau)$*, similarly hereinafter. Then if* $A^{-1} \not\leqslant 2B^{-1}$*, we have*

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

Lemma 3.6.19 (Lemma 1, [Gao et al., 2021])**.** *Since* $\|\phi_\tau\| \leqslant 1$*. Let* $A = \sum_{\tau=1}^{K} \phi_\tau \phi_\tau^T + \lambda I$*, then we have*

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

Therefore, if we have $2(\Lambda^{\texttt{buffer.s}(\widetilde{k}+1)})^{-1} \not\geqslant (\Lambda^{\texttt{buffer.e}(\widetilde{k})})^{-1}$, as well as $\texttt{buffer.e}(\widetilde{k}+1) \geqslant \texttt{buffer.s}(\widetilde{k}+1)$, it holds that $2(\Lambda^{\texttt{buffer.e}(\widetilde{k}+1)})^{-1} \not\geqslant (\Lambda^{\texttt{buffer.e}(\widetilde{k})})^{-1}$. We can then

control the matrices' determinants and have $\det \Lambda^{\texttt{buffer.e}(\widetilde{k}+1)} \geqslant 2 \det \Lambda^{\texttt{buffer.e}(\widetilde{k})}$. Then, using Theorem 3.6.18, we know that for any h and k, it holds $\log \det \Lambda_h^k \leqslant K_1 \log K$. We have $\log \det \Lambda_h^0 = d \log \lambda$. Combining Theorem 3.6.18, 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 $K_1 \log K - d \log \lambda \leqslant C_2 \log 2 \log K$. Therefore, 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. □

Proof of Theorem 3.6.2

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

$$\texttt{L}oss \geqslant \frac{1}{NHK} \frac{\beta}{2K} \frac{1}{3} \frac{\beta}{K} = \frac{3H}{K^3(1-\gamma)},$$

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

The inequality holds since that with probability $\frac{1}{KHN}$, the policy will be π_{rand} and the bidder is selected, and the total loss is higher than the loss with policy π_{rand}. With a uniform reserve price, the probability that 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. □

Proof of Theorem 3.6.3

For convenience, similar to Golrezaei et al. [2019], we define

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

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 $t = 1, \ldots, K$ 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 $S_i = \{t : t \in [1, K], q_i^t = 0 \textit{ and } s_i^t \geqslant \alpha\}$ and $O_i = \{t : q_i^t = 1 \textit{ and } o_i^t \geqslant \alpha\}$. As a result, $L_i = L_i^s \bigcup L_i^o$, where $L_i^s = \{t : \mathbb{1}(v_i^t \geqslant r_i^t) = 1, \mathbb{1}(b_i^t \geqslant r_i^t) = 0\}$ and $L_i^o = \{t : \mathbb{1}(v_i^t \geqslant r_i^t) = 0, \mathbb{1}(b_i^t \geqslant r_i^t) = 1\}$. Finally, we have $S_i^c = \{t : q_i^t = 1 \textit{ or } s_i^t \leqslant \alpha\}$. So, $|L_i^s| = |S_i \bigcap L_i^s| + |S_i^c \bigcap L_i^s|$.

To bound $|(S_i \bigcap L_i^s) \bigcup (O_i \bigcap L_i^o)|$: using Theorem 3.6.1 and Theorem 3.6.2, we have that if we set $\alpha = C_3 \frac{H}{K}$, it is bounded by $\frac{3HC_2 \log^2 K}{\log \frac{1}{\gamma}}$.

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

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

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

Using Azuma-Hoeffding inequality [Hoeffding, 1994], it holds that

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

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

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

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

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

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

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

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

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

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

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

and it finishes the proof. □

Proof of Theorem 3.6.4

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

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

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

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

Proof of Theorem 3.6.5

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 $\widetilde{q}_\tau$ to represent the outcome that every bidder bids truthfully and $\widehat{q}_\tau$ 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 3.6.20. *Under Equation* (3.3.1)*, it holds that*

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

where $\texttt{L} \leqslant C_4 H\log^2 K$ *due to Theorem 3.6.3.*

Proof of Theorem 3.6.20 Since there are at most L steps that overbid or underbid changes the outcome, $\widehat{q}_\tau$ and $\widetilde{q}_\tau$ differ in at most L different points. Since $\widetilde{q}_\tau$ and $\widehat{q}_\tau$ belong to $\{0,1\}$, we have

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

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

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

for any θ.

Therefore, it holds that

$$\sum_{\tau}(\widetilde{q}_\tau - 1 + F(m_\tau - 1 - \langle \phi_\tau, \theta\rangle))^2 \leqslant \sum_{\tau}(\widehat{q}_\tau - 1 + F(m_\tau - 1 - \langle \phi_\tau, \theta\rangle))^2 + 3\mathtt{L}, \qquad (3.6.3)$$

for any θ.

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

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

The first and third inequalities hold due to (3.6.3). 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 3.6.20, we have

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

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

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

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

So, it holds that

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

Since $\|\theta\| \leqslant \sqrt{d}$, 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 3.3.1 and Theorem 3.3.2, we have $|f''| \leqslant L$ and $|D_\tau| \leqslant C_1$.

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

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

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

Therefore, it holds that

$$\|\sum_{\tau}\xi_{\tau}\langle\phi_{\tau}, D_{\tau}(\widehat{\theta},\theta^{*})(\widehat{\theta}-\theta^{*})\| \leqslant \|\sum_{\tau}\xi_{\tau}\langle\phi_{\tau}, D_{\tau}(\theta,\theta')(\theta-\theta')\| + (2L\sqrt{d}+C_1)\texttt{buffer.e}(\widetilde{k})\epsilon,$$

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

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

$$V(\phi) \leqslant V(\widehat{\phi}) + 4C_1\sqrt{d}(2L\sqrt{d}+C_1)\texttt{buffer.e}(\widetilde{k})\epsilon. \tag{3.6.4}$$

The inequality holds because of the square difference formula.

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

$$\sqrt{V(\phi)} \leqslant \sqrt{V(\widehat{\phi})} + \sqrt{4C_1\sqrt{d}(2L\sqrt{d}+C_1)\texttt{buffer.e}(\widetilde{k})\epsilon}. \tag{3.6.5}$$

Since θ^{*} is the true parameter and $\xi_{\tau} = (1-\widetilde{q}_{\tau}) - f_{m_{\tau}}(\langle\phi_{\tau}, \theta^{*}\rangle)$ 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

$$\Pr[|\sum_{\tau}\xi_{\tau}D_{\tau}\langle\phi_{\tau}, \theta-\theta'\rangle| \geqslant \sqrt{\log\frac{2B_2HN}{\delta\epsilon^{2d}}V(\phi)}] \leqslant \frac{\delta}{HN}, \tag{3.6.6}$$

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

Therefore, it holds that

$$
\begin{aligned}
V(\widehat{\phi}) \leqslant & 4C_1\sqrt{d}(2L\sqrt{d}+C_1)\texttt{buffer.e}(\widetilde{k})\epsilon + V(\phi) \\
\leqslant & 4C_1\sqrt{d}(2L\sqrt{d}+C_1)\texttt{buffer.e}(\widetilde{k})\epsilon + 2\sqrt{\log\frac{2B_2HN}{\delta\epsilon^{2d}}V(\phi)} + 6\mathtt{L} \\
\leqslant & 4C_1\sqrt{d}(2L\sqrt{d}+C_1)\texttt{buffer.e}(\widetilde{k})\epsilon + 2\sqrt{\log\frac{2B_2HN}{\delta\epsilon^{2d}}}\big[\sqrt{V(\widehat{\phi})} \\
& + \sqrt{4C_1\sqrt{d}(2L\sqrt{d}+C_1)\texttt{buffer.e}(\widetilde{k})\epsilon}\big] + 6\mathtt{L} \\
= & 4C_1\sqrt{d}(2L\sqrt{d}+C_1) + 2\sqrt{\log\frac{2B_2HN\texttt{buffer.e}(\widetilde{k})^{2d}}{\delta}}\big[\sqrt{V(\widehat{\phi})} \\
& + \sqrt{4C_1\sqrt{d}(2L\sqrt{d}+C_1)}\big] + 6\mathtt{L} \\
\leqslant & 4C_1\sqrt{d}(2L\sqrt{d}+C_1) + 2\sqrt{\log\frac{2B_2HN\texttt{buffer.e}(\widetilde{k})^{2d}}{\delta}}\big[\sqrt{V(\widehat{\phi})} \\
& + \sqrt{4C_1\sqrt{d}(2L\sqrt{d}+C_1)}\big] + 6C_4H\log^2 K.
\end{aligned}
$$

The first inequality holds due to (3.6.4) while the second one holds due to (3.6.6) and (3.6.20). The third inequality holds because of (3.6.5). The equality holds since we set $\epsilon = \frac{1}{\texttt{buffer.e}(\widetilde{k})}$. The final inequality holds because of Theorem 3.6.3.

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

Similar to Wang et al. [2020c], we have

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

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

It holds since

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

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

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

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

which ends the proof. □

Proof of Theorem 3.6.6

Using Cauchy inequality, we have the following statement:

Lemma 3.6.21. *It holds that*

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

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

Then Theorem 3.6.21 and Theorem 3.6.5 lead to Theorem 3.6.6. □

Proof of Theorem 3.6.7

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

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

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

where

$$\begin{aligned}
\Delta_1 = & |[\max\{r_{ih}^{k-}, \alpha_{ih}^k\}\, \mathbb{1}(r_{ih}^k \geqslant \max\{r_{ih}^{k-}, \alpha_{ih}^k\})] \\
& - [\max\{\widehat{r}_{ih}^{k-}, \alpha_{ih}^k\}\, \mathbb{1}(r_{ih}^k \geqslant \max\{r_{ih}^{k-}, \alpha_{ih}^k\})]|,
\end{aligned}$$

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

and

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

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

To bound $|R_h^k(\cdot,\cdot) - \widetilde{R}_h^k(\cdot,\cdot)|$, we have the following lemmas. We define $W_{ih}^k(\alpha) =$

$\mathbb{E}[\max\{v_{ih}^{k-}, \alpha\}\, \mathbb{1}(v_{ih}^{k} \geqslant \max\{v_{ih}^{k-}, \alpha\}) \,|\, \phi_h^k]$ at first.

Lemma 3.6.22 (Lemma C.3. [Golrezaei et al., 2019]). *Since α_{ih}^{k*} is determined by Myerson Lemma [Myerson, 1981], we have $W_{ih}^{\prime k}(\alpha_{ih}^{k*}) = 0$. Furthermore, there exists a constant B_4 that for any α between α_{ih}^{k} and α_{ih}^{k*}, we have $|W_{ih}^{\prime\prime k}(\alpha)| \leqslant B_4$ for any i and h, under assumption Theorem 3.3.1, Theorem 3.3.2 and Theorem 3.3.3.*

Lemma 3.6.23 (Lemma C.4. [Golrezaei et al., 2019]). *Under Theorem 3.3.3, it holds that*

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

By applying Theorem 3.6.23, we have

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

The first inequality holds due to Taylor expansion. The second inequality holds due to Theorem 3.6.23, while the third one holds due to Theorem 3.6.6. The last inequality holds since $\|\phi\|_{\Lambda^{-1}} \leqslant \frac{1}{\lambda}$.

Remark 3.6.24. *Without Theorem 3.3.3, we can get the last inequality from the integral form of $R(\cdot,\cdot)$. For example, when $N = 1$, it holds that $R(\cdot,\cdot) = \alpha(1 - F(\alpha - 1 - \langle \phi, \theta \rangle))$. Then, $|R - \widetilde{R}| \leqslant 3C_1 |\langle \theta - \widehat{\theta}, \phi \rangle|$ due to Theorem 3.3.1. It shows that Theorem 3.3.3 is actually redundant as Theorem 3.3.1 exists.*

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

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

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

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

and it ends the proof. □

Proof of Theorem 3.6.9

In order to prove Theorem 3.6.9, we have the following lemmas for help.

Lemma 3.6.25. *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*

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

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 $|V| \leqslant 3H$ and $\|\mathcal{M}_h(\mathcal{S})\| \leqslant \sqrt{d}$, Theorem 3.6.25 holds. □

Lemma 3.6.26. *For any $(k,h) \in [K] \times [H]$, the vector $\omega_h^{\texttt{buffer.e}(\widetilde{k})}$ in Algorithm 5 satisfies:*

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

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

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

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

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

□

Lemma 3.6.27 (Lemma D.1. [Jin et al., 2020b])**.** *Let $\Lambda^{\texttt{buffer.e}(\widetilde{k})} = \lambda I + \sum_{\tau=1}^{\texttt{buffer.e}(\widetilde{k})} \phi_\tau \phi_\tau^T$ where $\phi_\tau \in \mathbb{R}^d$ and $\lambda > 0$. Then it holds*

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

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

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

Lemma 3.6.28. *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 $\|\phi_\tau\| \leqslant 1$ and $\Lambda_{\texttt{buffer.e}(\widetilde{k})} = \lambda I + \sum_{\tau=1}^{\texttt{buffer.e}(\widetilde{k})} \phi_\tau \phi_\tau^T$, then for any δ, with probability at least $1-\delta$, for any k corresponding to $\texttt{buffer.e}(\widetilde{k})$ and any $V \in \mathcal{V}$ so that $\sup_x |V(x)| \leqslant 3H$, we have

$$\begin{aligned}\|\sum_{\tau=1}^{k} \phi_\tau\{V(x_\tau) - \mathbb{E}[V(x_\tau)\,|\,\mathcal{F}_{\tau-1}]\}\|^2_{\Lambda^{-1}_{\texttt{buffer.e}(\widetilde{k})}} \leqslant & \frac{54C_2H^3\log^2 K}{\lambda \log\frac{1}{\gamma}} + \frac{32k^2\epsilon^2}{\lambda} \\ & + 144H^2[\frac{d}{2}\log\frac{k+\lambda}{\lambda} + \log\frac{\mathcal{N}_\epsilon}{\delta}],\end{aligned}$$

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

Proof. First of all, we have

$$\begin{aligned}&\|\sum_{\tau=1}^{k} \phi_\tau\{V(x_\tau) - \mathbb{E}[V(x_\tau)\,|\,\mathcal{F}_{\tau-1}]\}\|^2_{\Lambda^{-1}_{\texttt{buffer.e}(\widetilde{k})}} \\ \leqslant & 2\times 2\|\sum_{\tau=1}^{k} \phi_\tau\{V(x_\tau) - \mathbb{E}[V(x_\tau)\,|\,\mathcal{F}_{\tau-1}]\}\,\mathbb{1}\{k \notin \texttt{buffer}\}\|^2_{\Lambda_k^{-1}} + 2\times 3H\frac{1}{\lambda}3H\frac{3HC_2\log^2 K}{\log\frac{1}{\gamma}} \\ \leqslant & 4\|\sum_{\tau=1}^{k} \phi_\tau\{V(x_\tau) - \mathbb{E}[V(x_\tau)\,|\,\mathcal{F}_{\tau-1}]\}\|^2_{\Lambda_k^{-1}} + \frac{54C_2H^3\log^2 K}{\lambda\log\frac{1}{\gamma}}.\end{aligned}$$

Firstly, we have $(a+b)^2 \leqslant 2a^2 + 2b^2$. 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 $\texttt{buffer.e}(\widetilde{k})$, it is easy to prove that it is smaller than $4\|\sum_{\tau=1}^{k} \phi_\tau\{V(x_\tau) - \mathbb{E}[V(x_\tau)\,|\,\mathcal{F}_{\tau-1}]\}\,\mathbb{1}\{k \notin \texttt{buffer}\}\|^2_{\Lambda_k^{-1}}$. As for the one not in buffer, $\frac{54C_2H^3\log^2 K}{\lambda\log\frac{1}{\gamma}}$ is a trivial bound due to Theorem 3.6.1 and $V(\cdot) \leqslant 3H$.

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

Lemma 3.6.29. *Let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following*

parametric form

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

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

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

where B_5 *is a constant.*

Proof. Due to Lemma D.6. in Jin et al. [2020b], it holds that

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

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

Now, in order to bound the covering number corresponding to $\widehat{R}(x, \upsilon)$, 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 $|f| \leqslant C_1$, it holds that $\frac{\partial \widehat{R}}{\partial \mu_i} \leqslant B_5$ 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

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

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

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

Then, it finishes the proof. □

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

Lemma 3.6.30. *For any δ, with probability at least $1-\delta$, there exists constants B_6 and B_7 independent of K and H so that*

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

Proof. Combining Theorem 3.6.26, Theorem 3.6.28 and Theorem 3.6.29, we set $L = 3H\sqrt{\frac{dk}{\lambda}}$. With Algorithm 5, we have $B = C_7 + C_6 H \log^2 K$. Then we have

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

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

Now, let's show the determination of C_7.

Lemma 3.6.31. *There exist a constant B_8 so that $C_7 = B_8 H^{\frac{3}{2}} \log K$, and for any fixed policy π, on Good Event $\mathcal{E}$, i.e., all inequalities hold, we have for all $(x, \upsilon, h, k) \in \mathcal{S} \times \Upsilon \times$*

$[H] \times [K]$ *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 $\Delta_h^k(\cdot,\cdot) \leqslant (C_7 + C_6 H \log^2 K)\|\phi(\cdot,\cdot)\|_{(\Lambda_h^{\texttt{buffer.e}(\widetilde{k})})^{-1}}$.

Proof. Due to Bellman equation, we know that for any $(x, \upsilon, h) \in \mathcal{S} \times \Upsilon \times [H]$, 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 $\omega_h^k = \omega_h^{\texttt{buffer.e}(\widetilde{k})}$, it holds that

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

where

$$\delta_1 = -\lambda (\Lambda_h^{\texttt{buffer.e}(\widetilde{k})})^{-1} w_h^\pi,$$

$$\delta_2 = (\Lambda_h^{\texttt{buffer.e}(\widetilde{k})})^{-1} \sum_{\tau=1}^{\texttt{buffer.e}(\widetilde{k})} \phi_h^\tau [\widehat{V}_{h+1}^k(x_{h+1}^\tau) - \mathbb{P}_h \widehat{V}_{h+1}^k(x_h^\tau, \upsilon_h^\tau)],$$

$$\delta_3 = (\Lambda_h^{\texttt{buffer.e}(\widetilde{k})})^{-1} \sum_{\tau=1}^{\texttt{buffer.e}(\widetilde{k})} \phi_h^\tau \mathbb{P}_h (\widehat{V}_{h+1}^k - V_{h+1}^\pi)(x_h^\tau, \upsilon_h^\tau).$$

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

Firstly, it holds

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

The first inequality holds due to Cauchy inequality and $\Lambda_{\texttt{buffer.e}(\widetilde{k})} \geq \lambda I$. The second inequality holds due to Theorem 3.6.25.

Secondly, it holds that

$$|\langle\phi(\cdot,\cdot),\delta_2\rangle| \leqslant \sqrt{B_6H^3\log^2 K + B_7H^2\log C_7}\|\phi(\cdot,\cdot)\|_{(\Lambda_h^{\texttt{buffer.e}(\widetilde{k})})^{-1}}.$$

It holds because of Theorem 3.6.30.

Lastly, we have

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

The second and fourth equations hold due to the definition of the operator $\mathbb{P}_h$. The third equation holds due to simple algebra arrangement. The inequality holds due to Cauchy

inequality, $V(\cdot) \leqslant 3H$ and $\Lambda_{\texttt{buffer.e}(\widetilde{k})} \succeq \lambda I$.

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

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

Lemma 3.6.32. *Under the setting of Theorem 3.3.5, on the Good Event $\mathcal{E}$, it holds that for any $(x, \upsilon, h, k) \in \mathcal{S} \times \Upsilon \times [H] \times [K]$,*

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

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

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

due to Theorem 3.6.31. Therefore, we have

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

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 3.6.31, it holds that

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

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

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

which ends the proof. □

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

Lemma 3.6.33. *Let $\delta_h^k = V_h^k(x_h^k) - V_h^{\pi_{\widetilde{k}}}(x_h^k)$ and $\xi_{h+1}^k = \mathbb{E}[\delta_{h+1}^k \,|\, x_h^k, \upsilon_h^k] - \delta_{h+1}^k$. Then conditional on Good Event $\mathcal{E}$, it holds that for any $(k,h) \in [K] \times [H]$,*

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

Proof. Due to Theorem 3.6.31, it holds that

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

Then, since $\pi_{\widetilde{k}} = \pi_{\texttt{buffer.e}(\widetilde{k})}$ is the greedy policy before mixture at episode k by Algorithm 5, we have

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

Then, it ends the proof. □

With these preparations, we begin to prove Theorem 3.6.9.

Using notations in Theorem 3.6.33, it holds that conditional on Good Event $\mathcal{E}$

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

The first inequality holds due to Theorem 3.6.32, while the second one holds due to Theorem 3.6.33. The third inequality holds due to the process of Algorithm 4, while the last one is trivial.

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

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

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

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

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

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

Then with Cauchy inequality, we have

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

Finally, combining the two terms and we have

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

and it finishes our proof. □

3.6.5 Auxiliary Lemmas and Proofs in Section 3.6.3

In this section, we provide proof of lemmas in Section 3.6.3 in detail. We organize this section in the order of lemmas.

Proof of Theorem 3.6.11

In Algorithm 7, there are two types of $\{\texttt{buffer.e}(\widetilde{k})\}$. The number of $\{\texttt{buffer.e}(\widetilde{k})\}$ satisfying $2(\Lambda_h^k)^{-1} \not\preceq (\Lambda_h^{\texttt{buffer.e}(\widetilde{k})})^{-1}$ is smaller than $\frac{3C_2 H\log^2 K}{\log\frac{1}{\gamma}}$ due to Theorem 3.6.1. The number of $\{\texttt{buffer.e}(\widetilde{k})\}$ when $\log_2 k$ is an integer is smaller than $[\log_2 K]+1$. Combining the two parts finishes the proof. □

Proof of Theorem 3.6.12

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 3.6.3 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 3.6.3 still holds on regarding to $\widetilde{q}$.

With the same methodology in Theorem 3.6.3, we have the lemma due to Theorem 3.6.11.

□

Proof of Theorem 3.6.13

Similar to the proof of Theorem 3.6.5, 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 (3.4.1). 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 $1 + \langle \phi_\tau, \theta \rangle + z \geqslant \rho_i$. 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 3.6.5, we prove Theorem 3.6.13.

□

Proof of Theorem 3.6.15

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

First of all, if every buyer bids truthfully, then with Theorem 3.6.10, it holds with probability at least $1 - \frac{\delta}{K}$ for each update that

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

However, bidders may overbid or underbid for less than $\frac{C_3 H}{K}$ due to Theorem 3.6.2 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 $|f(\cdot)| \leqslant C_1$, the difference because of overbid or underbid is smaller than $\frac{C_1 C_3 H}{K}$.

Then, due to Theorem 3.6.14, the difference because of error in μ is smaller than

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

The inequality holds since we have the mean value inequality and Theorem 3.6.27.

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

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

Because the number of episodes we run Equation (3.4.1) is smaller than K, then the total probability happening Bad Event $\mathcal{E}^c$ is smaller than δ. Then, it ends the proof. □

Proof of Theorem 3.6.16

In order to prove Theorem 3.6.16, we introduce the following lemma first.

Lemma 3.6.34. *Under assumption Theorem 3.3.2, when Theorem 3.6.15 holds, using histogram method to estimate p.d.f $f(\cdot)$ leads to the following bound that for any x*

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

where D_1 is a constant.

Proof of Theorem 3.6.34 With Theorem 3.6.15 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 3.3.2, it holds that

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

Therefore, it holds that

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

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

Therefore, unlike Theorem 3.6.23, we have the following lemma.

Lemma 3.6.35. *Under Theorem 3.3.3, it holds that*

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

where D_2 is a constant.

Proof of Theorem 3.6.35 Myerson [1981] shows that the optimal reserve price satisfies

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

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

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

Theorem 3.6.23 shows that $|\widetilde{\alpha} - \alpha^*| \leqslant |\langle \phi_h^k, \theta_{ih} - \widehat{\theta}_{ih} \rangle|$.

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

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

The first inequality holds due to triangle inequality. The second inequality holds due to Theorem 3.3.1, Theorem 3.6.15 and Theorem 3.6.34.

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

It holds that $\phi(x) = x - \frac{1-F(x)}{f(x)} = x + \frac{1}{\log'(1-F(x))}$. Under Theorem 3.3.3, it holds that $1 - F(\cdot)$ is log-concave implying $\log'(1 - F(\cdot))$ is decreasing. Therefore, $\phi'(x) \geqslant 1$.

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

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

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

Then, it ends our proof. □

Now, we are ready to prove Theorem 3.6.16. Using notations in Theorem 3.6.7, we use another factor F to show that we use $F(\cdot)$ and $f(\cdot)$ in the function while factor $\widehat{F}$ to denote the use of $\widehat{F}(\cdot)$ and $\widehat{f}(\cdot)$.

With the same methodology in Theorem 3.6.7, it holds that

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

where D_3 and D_4 are two constants. The first inequality holds since $(a+b)^2 \leqslant 2(a^2+b^2)$. 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 $\widehat{R}_h^k(\cdot,\cdot,F) = \sum_{i=1}^N \mathbb{E}_F[\max\{\widehat{r}_{ih}^{k-},\alpha_{ih}^k\}\mathbb{1}(\widehat{r}_{ih}^k \geqslant \max\{\widehat{r}_{ih}^{k-},\alpha_{ih}^k\})]$ and $\widehat{R}_h^k(\cdot,\cdot,\widehat{F}) = \sum_{i=1}^N \mathbb{E}_{\widehat{F}}[\max\{\widehat{r}_{ih}^{k-},\alpha_{ih}^k\}\mathbb{1}(\widehat{r}_{ih}^k \geqslant \max\{\widehat{r}_{ih}^{k-},\alpha_{ih}^k\})]$, we have that the difference of expected revenue about each buyer is smaller than $3NC_{12}\frac{H\log^2 K}{\sqrt{\texttt{buffer.e}(\widetilde{k})}}$. It comes from that the expected revenue depends on N-fold integral with respect to random variable $\{z_{ih}^k\}_{i=1}^N$. Since $\int x(dF - dF') = -\int(F-F')dx \leqslant 3\|F-F'\|_\infty \leqslant 3C_{12}\frac{H\log^2 K}{\sqrt{\texttt{buffer.e}(\widetilde{k})}}$, each integral has error less than $3C_{12}\frac{H\log^2 K}{\sqrt{\texttt{buffer.e}(\widetilde{k})}}$. With N buyers in total, it holds that

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

Combining the two parts, it holds

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

which ends the proof. Similarly, we can use Theorem 3.3.1 to achieve parallel results without

Theorem 3.3.3 as Theorem 3.6.24 says. □

Proof of Theorem 3.6.17

Now, we introduce some lemmas in parallel in order to prove Theorem 3.6.17.

Lemma 3.6.36. *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 $\|\phi_\tau\| \leqslant 1$ and $\Lambda_{\texttt{buffer.e}(\widetilde{k})} = \lambda I + \sum_{\tau=1}^{\texttt{buffer.e}(\widetilde{k})} \phi_\tau \phi_\tau^T$, then for any δ, with probability at least $1-\delta$, for any k corresponding to $\texttt{buffer.e}(\widetilde{k})$ and any $V \in \mathcal{V}$ so that $\sup_x |V(x)| \leqslant 3H$, we have*

$$\begin{aligned}\|\sum_{\tau=1}^{k} \phi_\tau\{V(x_\tau) - \mathbb{E}[V(x_\tau)\,|\,\mathcal{F}_{\tau-1}]\}\|^2_{\Lambda^{-1}_{\texttt{buffer.e}(\widetilde{k})}} \leqslant & \frac{54C_9H^3\log^2 K}{\lambda\log\frac{1}{\gamma}} + \frac{32k^2\epsilon^2}{\lambda} \\ & + 144H^2[\frac{d}{2}\log\frac{k+\lambda}{\lambda} + \log\frac{\mathcal{N}_\epsilon}{\delta}],\end{aligned}$$

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

Lemma 3.6.37. *Let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following parametric form*

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

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

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

where B_5 is a constant.

Proof of Theorem 3.6.37 When $F(\cdot)$ is unknown, it holds that

$$\begin{aligned}\sup_{x,\upsilon}|\widehat{R}_1(x,\upsilon)-\widehat{R}_2(x,\upsilon)| &= \sup_{x,\upsilon}|\widehat{R}_1(x,\upsilon,\widehat{F}_1)-\widehat{R}_2(x,\upsilon,\widehat{F}_2)| \\ &\leqslant \sup_{x,\upsilon}|\widehat{R}_1(x,\upsilon,\widehat{F}_1)-\widehat{R}_2(x,\upsilon,\widehat{F}_1)| \\ &\quad + \sup_{x,\upsilon}|\widehat{R}_2(x,\upsilon,\widehat{F}_1)-\widehat{R}_2(x,\upsilon,\widehat{F}_2)|.\end{aligned}$$

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 3.6.29, we simply use $\frac{\epsilon}{4NB_5}$-ball to cover $\widehat{\theta}_i$, and it holds that

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

Conditional on ω, A and $\{\widehat{\theta}_i\}_{i=1}^N$, with Theorem 3.6.16, we know that in order to satisfy $\sup_{x,\upsilon}|\widehat{R}(x,\upsilon,\widehat{F})-\widehat{R}(x,\upsilon,F)| \leqslant \frac{\epsilon}{4}$, what we need is $\|\widehat{F}-F\|_\infty \leqslant \frac{\epsilon}{12N^2}$. Then, it ends the proof. □

Then, it holds the following lemma.

Lemma 3.6.38. *For any δ, with probability at least $1-\delta$, there exists constants B_6 and B_7 independent of K and H so that*

$$\begin{aligned}\forall(k,h)\in[K]\times[H]:\ \|\sum_{\tau=1}^{k}\phi_h^\tau[\widehat{V}_{h+1}^k(x_{h+1}^\tau)-\mathbb{P}\widehat{V}_{h+1}^k(x_h^\tau,\upsilon_h^\tau)]\|^2_{(\Lambda_h^{\texttt{buffer.e}(\tilde{k})})^{-1}} \\ \leqslant D_5H^3+D_6H^2\log C_{15},\end{aligned}$$

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

Proof. Similar to the proof of Theorem 3.6.30, we just replace $\mathcal{N}_\epsilon$ by $d\log(1+\frac{8L}{\epsilon})+d^2\log(1+$

$\frac{32\sqrt{d}B^2}{\lambda\epsilon^2}) + dN\log(1+\frac{16NB_5\sqrt{d}}{\epsilon}) + \log\mathcal{N}_{\frac{\epsilon}{12N^2}}(\mathcal{F})$. Then, we set $\lambda = 1$, $B = C_{15} + C_{13}H\log^2 K$ and $\epsilon = \frac{dH}{k}$. With Theorem 3.4.1, we finish our proof. □

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

Lemma 3.6.39. *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 $\mathcal{E}$, i.e., all inequalities hold, we have for all $(x, \upsilon, h, k) \in \mathcal{S} \times \Upsilon \times [H] \times [K]$ 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 $\Delta_h^k(\cdot,\cdot) \leqslant (C_{15} + C_{13}H\log^2 K)\|\phi(\cdot,\cdot)\|_{(\Lambda_h^{\texttt{buffer.e}(\widetilde{k})})^{-1}} + C_{14}\frac{H^2\log^4 K}{\sqrt{\texttt{buffer.e}(\widetilde{k})}}$.

Proof. The proof of Theorem 3.6.39 is the same as proof of Theorem 3.6.31. Let's show the determination of D_7 in parallel. With Theorem 3.6.38 in hand, it holds that

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

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 $\delta_h^k = V_h^k(x_h^k) - V_h^{\pi_{\widetilde{k}}}(x_h^k)$. It holds due to Theorem 3.6.39 and Theorem 3.6.32.

Lemma 3.6.40. *Let $\delta_h^k = V_h^k(x_h^k) - V_h^{\pi_{\widetilde{k}}}(x_h^k)$ and $\xi_{h+1}^k = \mathbb{E}[\delta_{h+1}^k \,|\, x_h^k, \upsilon_h^k] - \delta_{h+1}^k$. Then conditional on Good Event $\mathcal{E}$, it holds that for any $(k,h) \in [K] \times [H]$,*

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

Now, we are ready to prove Theorem 3.6.17.

Similar to the proof of Theorem 3.6.9, it holds that

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

Due to Algorithm 7, we have $k \leqslant 2\texttt{buffer.e}(\widetilde{k})$. Therefore, it holds that

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

Therefore, it holds that

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

which ends the proof. □

3.6.6 Detailed Results of Numerical Experiments

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

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

For the implementation details, we assume $N = 1$ and there are two different contexts both appearing in probability 0.5. Besides, we assume $\theta = [0, 4, 0.6]^T$ and underlying noise distribution is Unif([-1,1]). In order to discrete these strategies, we constrain that bids must be a multiple of 0.01. To simulate strategic bidders, we use Theorem 3.6.2. Once it's in the buffer period, we assume bidders randomly bid. However, if not, we assume bidders bid their value plus a random noise with scale $\frac{C_3}{K}$. For NPAC-S, we use similar ways to simulate strategic behaviors. However, for SCORP, we stated before that it uses too many episodes to

explore, we loosen its constraints and assume truthful bidding. Although we only consider an upper bound for its performance, SCORP still performs worse than CLUB and NPAC-S. So, we only compare CLUB and NPAC-S in MDP settings. To solve Equation (3.4.1), we seek help from `scipy.optimize` package. Actually, most of the running time is spent on solving Equation (3.4.1). We believe we can reduce our running time by using other commercial optimization solvers.

In the MDP setting, we show the total regrets of CLUB and NPAC-S in the 30 trails in Table 3.2. Among all 30 trials, CLUB wins NPAC-S every time. The average of CLUB is 203.07, overwhelming the corresponding 756.31 for NPAC-S. As a result, it shows that CLUB has better performance against NPAC-S in the MDP setting.

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

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

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

Table 3.1: Regrets of three different algorithms in each trail.

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

Table 3.2: Regrets of two different algorithms in each trail.

CHAPTER 4

OFFLINE RL FOR WELFARE MAXIMIZING MECHANISM IN MDPS WITH GENERAL FUNCTION APPROXIMATION

4.1 Introduction

Mechanism design studies how best to allocate goods among rational agents [Maskin, 2008, Myerson, 2008, Roughgarden, 2010]. Dynamic mechanism design focuses on analyzing optimal allocation rules in a changing environment, where demands for goods, the amount of available goods, and their valuations can vary over time [Bergemann and Välimäki, 2019]. Problems ranging from online commerce and electric vehicle charging to pricing Wi-Fi access at Starbucks have been studied under the dynamic mechanism design framework [Gallien, 2006, Gerding et al., 2011, Friedman and Parkes, 2003]. Existing approaches in the literature require knowledge of the problem, such as the evaluation of goods by agents [Bergemann and Välimäki, 2010, Pavan et al., 2014], the transition dynamics of the system [Doepke and Townsend, 2006], or the policy that maximizes social welfare [Parkes and Singh, 2003, Parkes et al., 2004]. Unfortunately, such knowledge is often not available in practice.

A practical approach we take in this chapter is to learn a dynamic mechanism from data using offline Reinforcement Learning (RL). Vickrey-Clarke-Groves (VCG) mechanism provides a blueprint for the design of practical mechanisms in many problems and satisfies crucial mechanisms design desiderata in an extremely general setting [Vickrey, 1961, Clarke, 1971, Groves, 1979]. In this chapter, we approximate the desired VCG mechanism using a priori collected data [Jin et al., 2021b, Xie et al., 2021, Zanette et al., 2021]. We assume that the mechanism designer does not know the utility of the agents or the transition kernel of the states, but has access to an offline data set that contains observed state transitions and utilities [Lange et al., 2012]. The goal of the mechanism designer is to recover the ideal mechanism purely from this data set, without requiring interaction with the agents.

We focus on an adaptation of the classic VCG mechanism to the dynamic setting [Parkes, 2007] and assume that agents' interactions with the seller follow an episodic Markov Decision Process (MDP), where the agents' rewards are state-dependent and evolve over time within each episode. To accommodate the rich class of quasilinear utility functions considered in the economic literature [Bergemann and Välimäki, 2019], we use offline RL with a general function approximation [Xie et al., 2021] to approximate the dynamic VCG mechanism.

Related Works. Parkes and Singh [2003] and Parkes et al. [2004] studied dynamic mechanism design from an MDP perspective. The proposed mechanisms can implement social welfare-maximizing policies in a truth-revealing Bayes-Nash equilibrium both exactly and approximately. Bapna and Weber [2005] studied the dynamic auction setting from a multi-arm bandit perspective. Using the notion of marginal contribution, Bergemann and Välimäki [2006] proposed a dynamic mechanism that is efficient and truth-telling. Pavan et al. [2009] analyzed the first-order conditions of efficient dynamic mechanisms. Athey and Segal [2013] extended both the VCG and AGV mechanisms [d'Aspremont and Gérard-Varet, 1979] to the dynamic regime, obtaining an efficient budget-balanced dynamic mechanism. Kakade et al. [2013] proposed the virtual pivot mechanism that achieves incentive compatibility under a separability condition. See Cavallo [2009], Bergemann and Pavan [2015], and Bergemann and Välimäki [2019] for recent surveys on dynamic mechanism design. This chapter builds on the mechanism in Parkes [2007] and Bergemann and Välimäki [2010], but focuses on learning a mechanism from data rather than designing a mechanism in a known environment.

Only a few recent works have investigated the learning of mechanisms. Kandasamy et al. [2023] provided an algorithm that recovers the VCG mechanism in a stationary multi-arm bandit setting. Cen and Shah [2022], Dai and Jordan [2021], Jagadeesan et al. [2021], and Liu et al. [2021a] studied the recovery of stable matching when the agents' utilities are given by bandit feedback. Balcan et al. [2008] shows that incentive-compatible mechanism design problems can be reduced to a structural risk minimization problem. In contrast, our work

focuses on learning a dynamic mechanism in an offline setting.

This chapter is also related to the literature on offline RL [Yu et al., 2020, Kumar et al., 2020, Liu et al., 2020, Kidambi et al., 2020, Jin et al., 2021b, Xie et al., 2021, Zanette et al., 2021, Yin and Wang, 2021, Uehara and Sun, 2021]. In the context of linear MDPs, Jin et al. [2021b] provided a provably sample-efficient pessimistic value iteration algorithm, while Zanette et al. [2021] used an actor-critic algorithm to further improve the upper bound. Yin and Wang [2021] proposed an instance-optimal method for tabular MDPs. Uehara and Sun [2021] focused on model-based offline RL, while Xie et al. [2021] introduced a pessimistic soft policy iteration algorithm for offline RL with a general function approximation. Compared to Xie et al. [2021], in addition to the social welfare suboptimality, we also provide bounds on both the agents' and the seller's suboptimalities. We also show that our algorithm asymptotically satisfies key mechanism design desiderata, including truthfulness and individual rationality. Finally, we use optimistic and pessimistic estimates to learn the VCG prices, instead of the purely pessimistic approach discussed in Xie et al. [2021]. This difference shows the difference between dynamic VCG and standard MDP. Our work also features a simplified proof of the main technical results in Xie et al. [2021].

Our Contributions. We propose the first offline reinforcement learning algorithm that can learn a dynamic mechanism from any given data set. Additionally, our algorithm does not make any assumption about data coverage and only assumes that the underlying action-value functions are approximately realizable and the function class is approximately complete (see Assumptions 4.2.3 and 4.2.4 for detailed discussions), which makes the algorithm applicable to the wide range of real-world mechanism design problems with quasilinear, potentially non-convex utility functions [Carbajal and Ely, 2013, Bergemann and Välimäki, 2019].

Our work features a soft policy iteration algorithm that allows for both optimistic and pessimistic estimates. When the data set has sufficient coverage of the optimal policy, the value function is realizable, and the function class is complete, our algorithm sublinearly

converges to a mechanism with suboptimality $\mathcal{O}(K^{-1/3})$, matching the rates obtained in Xie et al. [2021], where K denotes the number of trajectories contained in the offline dataset. In addition to suboptimality guarantees, we further show that our algorithm is asymptotically individually rational and truthful with the same $\mathcal{O}(K^{-1/3})$ guarantee.

On the technical side, our work features a simplified theoretical analysis of pessimistic soft policy iteration algorithms [Xie et al., 2021], using an adaptation of the classic tail bound discussed in Györfi et al. [2002]. Moreover, unlike [Xie et al., 2021], our simplified analysis is directly applicable to continuous function classes via a covering-based argument.

Notations. For any positive integer $z \in \mathbb{Z}_{>0}$, let $[z] = \{1, 2, \ldots, z\}$. For any set A, let $\Delta(A)$ be the set of probability distributions supported on A. For two sequences x_n, y_n, we say $x_n = \mathcal{O}(y_n)$ if there exist universal constants $n_0, C > 0$ such that $x_n < Cy_n$ for all $n \geqslant n_0$. We use $\widetilde{\mathcal{O}}(\cdot)$ to denote $\mathcal{O}(\cdot)$ ignoring log factors. Unless stated otherwise, we use $\|\cdot\|$ to denote the ℓ_2-norm

4.2 Background and Preliminaries

In this section, we define the dynamic mechanism and related notions. In addition, we discuss three key mechanism design desiderata and their asymptotic versions. Finally, we introduce the general function approximation regime and related assumptions.

Episodic MDP. Consider an episodic MDP given by $\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \{r_{i,h}\}_{i=0,h=1}^{n,H}\right)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the seller's action space, H is the length of each episode, and $\mathcal{P} = \{\mathcal{P}_h\}_{h=1}^{H}$ is the transition kernel, where $\mathcal{P}_h(s'|s,a)$ denotes the probability that the state $s \in \mathcal{S}$ transitions to the state $s' \in \mathcal{S}$ when the seller chooses the action $a \in \mathcal{A}$ at the h-th step.[1] We assume that $\mathcal{S}, \mathcal{A}$ are both finite but can be arbitrarily large. Let $r_{i,h} : \mathcal{S} \times \mathcal{A} \to [0,1]$ denote the reward function of an agent i at step h and $r_{0,h} : \mathcal{S} \times \mathcal{A} \to [-R_{\max}, -n + R_{\max}]$

1. In mechanism design literature the reward function is often called "value function." We use the tem "reward function" throughout the paper to avoid confusion with state- and action-value functions.

the seller's reward function at step h, which can be negative, as policies can be costly.

A stochastic policy $\pi = \{\pi_h\}_{h=1}^H$ maps the seller's state $\mathcal{S}$ to a distribution over the action space $\mathcal{A}$ at each step h, where $\pi_h(a|s)$ denotes the probability that the seller chooses the action $a \in \mathcal{A}$ when they are in the state $s \in \mathcal{S}$. We use d_π to denote the state-action visitation measure over $\{\mathcal{S} \times \mathcal{A}\}^H$ induced by the policy π and use $\mathbb{E}_\pi$ as a shorthand notation for the expectation taken over the visitation measure.

For any given reward function r and any policy π, the (state-)value function $V_h^\pi(\cdot; r) : \mathcal{S} \to \mathbb{R}$ is defined as $V_h^\pi(x; r) = \mathbb{E}_\pi[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})|s_h = x]$ at each step $h \in [H]$ and the corresponding action-value function (Q-function) $Q_h^\pi(\cdot, \cdot; r) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as $Q_h^\pi(x, a; r) = \mathbb{E}_\pi[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})|s_h = x, a_h = a]$. For any function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, any policy π, and $h \in [H]$, we use the shorthand notation $g(s, \pi_h) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}[g(s, a)]$. We define the policy-specific Bellman evaluation operator at h with respect to reward function r under policy π as

$$(\mathcal{T}_{h,r}^\pi g)(x, a) = r_h(x, a) + \mathbb{E}_\mathcal{P}\left[g(s_{h+1}, \pi_{h+1})|s_h = x, a_h = a\right], \tag{4.2.1}$$

where $\mathbb{E}_\mathcal{P}$ is taken over the randomness in the transition kernel $\mathcal{P}$.

We emphasize that while the problem setting we consider features multiple reward functions and interaction between multiple participants, our setting is not an instance of a Markov game [Littman, 1994] as we allow only the seller to take actions.

Dynamic Mechanism as an MDP. We assume that agents and sellers interact in the following way. Without loss of generality, assume that the seller starts at some fixed state $s_0 \in \mathcal{S}$ when $h = 1$. For each $h \in [H]$, the seller observes its state s and takes some action $a \in \mathcal{A}$. The agent receives the reward $r_{i,h}(s, a)$ and reports to the seller the received reward as $\widetilde{r}_{i,h}(s_h, a_h) \in [0, 1]$, which may be different from the true reward. The seller receives a reward $r_{0,h}(s, a)$ and transitions to some state $s' \sim \mathcal{P}_h(\cdot|s, a)$. At the end of each episode, the seller charges each agent i a price $p_i \in \mathbb{R}$, $i \in [n]$.

We stress the difference between the *reported* reward, $\widetilde{r}_{i,h}$, and the *actual* reward, $r_{i,h}$. The reported reward is equal to $r_{i,h}$ if an agent is truthful but may be given by an arbitrary function $\widetilde{r}_{i,h} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ when the agent is not. In other words, the agent i's reported reward comes from the actual reward function $r_{i,h}$ or some arbitrary reward function $\widetilde{r}_{i,h}$. Our algorithm learns a mechanism via the reported rewards and, under certain assumptions, we can provide guarantees on the actual rewards.

For convenience, let $R = \sum_{i=0}^{n} r_i$ be the sum of true reward functions and $R_{-i} = \sum_{i' \neq i} r_i$ the sum of true reward functions excluding agent i. Let $\widetilde{R}$, $\widetilde{R}_{-i}$ be defined similarly for the reported reward functions. Let $\mathcal{R} = \{R_{-i}\}_{i=1}^{n} \cup \{R\}$ be the set of all true reward functions that we will estimate and $\widetilde{\mathcal{R}}$ be that for the reported reward functions. When all agents are truthful, $\widetilde{\mathcal{R}} = \mathcal{R}$. We also let

$$Q_h^*(\cdot, \cdot; r) = \max_{\pi \in \Pi} Q_h^\pi(\cdot, \cdot; r),\ V_h^*(\cdot; r) = \max_{\pi \in \Pi} V_h^\pi(\cdot; r),$$
$$\pi_r^* = \operatorname*{argmax}_{\pi \in \Pi} V_1^\pi(s_0; r),\ \forall r \in \mathcal{R} \cup \widetilde{\mathcal{R}}.$$

As a shorthand notation, let $\pi^* = \pi^*_R$, $\pi^*_{-i} = \pi^*_{R_{-i}}$, $\widetilde{\pi}^* = \pi^*_{\widetilde{R}}$, and $\widetilde{\pi}^*_{-i} = \pi^*_{\widetilde{R}_{-i}}$. Following Kandasamy et al. [2023], we define the agents' and seller's utilities as follows. For any $i \in [n]$, we define the agent i's utility under policy π, when charged price p_i, as

$$U_i^\pi(p_i) = \mathbb{E}_\pi[\sum_{h=1}^{H} r_{i,h}(s_h, a_h)] - p_i = V_1^\pi(s_0; r_i) - p_i.$$

The seller's utility is similarly defined as

$$U_0^\pi(\{p_i\}_{i=1}^n) = \mathbb{E}_\pi[\sum_{h=1}^{H} r_{0,h}(s_h, a_h)] + \sum_{i=1}^{n} p_i = V_1^\pi(s_0; r_0) + \sum_{i=1}^{n} p_i.$$

The social welfare for any policy $\pi \in \Pi$ is the sum of the utilities, $\sum_{i=0}^{n} \mathbb{E}_\pi[u_i] = V_1^\pi(s_0; R)$, similar to its definition in Bergemann and Välimäki [2010].

4.2.1 A Dynamic VCG Mechanism

We now discuss a dynamic adaptation of the VCG mechanism and three key mechanism design desiderata it satisfies [Nisan et al., 2007]. We begin by introducing the dynamic adaptation of the VCG mechanism.

Definition 4.2.1 (Dynamic VCG Mechanism)**.** *When agents interact according to the aforementioned MDP, assuming the transition kernel $\mathcal{P}$ and the reported reward functions $\{\widetilde{r}_i\}_{i=0}^n$ are known, the VCG mechanism selects $\widetilde{\pi}^*$, the social welfare maximizing policy based on the reported rewards, and charges the agent i price $p_i : \mathcal{S} \to \mathbb{R}$, given by $p_i = V_1^*(s_0; \widetilde{R}_{-i}) - V_1^{\widetilde{\pi}^*}(s_0; \widetilde{R}_{-i})$. More generally, when the mechanism chooses to implement some arbitrary policy π, the VCG price for the agent i is given by*

$$p_i = V_1^*(s_0; \widetilde{R}_{-i}) - V_1^{\pi}(s_0; \widetilde{R}_{-i}). \tag{4.2.2}$$

Observe that when $H = 1$, the dynamic adaptation we propose reduces to exactly the classic VCG mechanism [Nisan et al., 2007].

We highlight the three common mechanism desiderata in the mechanism design literature [Nisan et al., 2007, Bergemann and Välimäki, 2010, Hartline, 2012].

1. *Efficiency:* A mechanism is efficient if it maximizes social welfare when all agents report truthfully.
2. *Individual rationality:* A mechanism is individually rational if it does not charge an agent more than their reported reward, regardless of other agents' behavior. In other words, if an agent reports truthfully, they attain non-negative utility.
3. *Truthfulness:* A mechanism is truthful or (dominant strategy) incentive-compatible if, regardless of the truthfulness of other agents' reports, the agent's utility is maximized when they report their rewards truthfully.

In the MDP setting, the dynamic VCG mechanism simultaneously satisfies all three desiderata.

Proposition 4.2.2. *With $\mathcal{P}$ and the reported rewards $\{\widetilde{r}_i\}_{i=0}^n$ known, choosing $\widetilde{\pi}^*$ and charging p_i for all $i \in [n]$ according to* (4.2.2) *ensures that the mechanism satisfies truthfulness, individual rationality, and efficiency simultaneously.*

Proof. See Appendix 4.5.1 for a detailed proof. □

Performance Metrics. We use the following metrics to evaluate the performance of our estimated mechanism. Let the social welfare suboptimality of an arbitrary policy π be

$$\mathrm{SubOpt}(\pi; s_0) = V_1^*(s_0; R) - V_1^{\pi}(s_0; R). \tag{4.2.3}$$

For any $i \in [n]$, let $p_i^*(s_0) = V_1^*(s_0; R_{-i}) - V_1^{\pi^*}(s_0; R_{-i})$ be the price charged to the agent i by VCG under truthful reporting. We can similarly define the suboptimality with respect to the agents' and the seller's expected utilities. For any $i \in [n]$, the agent i's suboptimality with respect to policy π and price $\{p_i\}_{i=1}^n$ is defined as

$$\mathrm{SubOpt}_i(\pi, \{p_i\}_{i=1}^n; s_0) = U_i^{\pi^*}(p_i^*) - U_i^{\pi}(p_i) = V_1^{\pi^*}(s_0; r_i) - p_i^*(s_0) - V_1^{\pi}(s_0; r_i) + p_i, \tag{4.2.4}$$

and the seller's suboptimality is

$$\begin{aligned} \mathrm{SubOpt}_0(\pi, \{p_i\}_{i=1}^n; s_0) &= U_0^{\pi^*}(\{p_i^*\}_{i=1}^n) - U_0^{\pi}(\{p_i\}_{i=1}^n) \\ &= V_1^{\pi^*}(s_0; r_0) + \sum_{i=1}^n p_i^* - V_1^{\pi}(s_0; r_0) - \sum_{i=1}^n p_i. \end{aligned} \tag{4.2.5}$$

4.2.2 Offline Episodic RL with General Function Approximation

We use offline RL in the general function approximation setting to minimize the aforementioned suboptimalities. Let $\mathcal{D}$ be a precollected data set that contains K trajectories, that is, $\mathcal{D} = \{(x_h^\tau, a_h^\tau, \{\widetilde{r}_{i,h}^\tau\}_{i=1}^n, x_{h+1}^\tau)\}_{h,\tau=1}^{H,K}$. Following the setup in Xie et al. [2021], we consider the i.i.d. data collection regime, where for all $h \in [H]$, $(x_h^\tau, a_h^\tau, x_{h+1}^\tau)_{\tau=1}^K$ is drawn from a distribution μ_h supported on $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$. The distribution μ over $\{\mathcal{S} \times \mathcal{A} \times \mathcal{S}\}^H$ is induced by a behavioral policy used for data collection. We do not make any coverage assumption on μ, similar to the existing literature on offline RL [Jin et al., 2021b, Uehara and Sun, 2021, Zanette et al., 2021].

Consider some general function class $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \ldots \times \mathcal{F}_H$. For each $h \in [H]$, we use some arbitrary yet bounded function class $\mathcal{F}_h \subseteq \mathcal{S} \times \mathcal{A} \to [-(H-h+1)R_{\max}, (H-h+1)R_{\max}]$ to approximate $Q_h^\pi(\cdot, \cdot; r)$ for arbitrary π and $r \in \widetilde{\mathcal{R}}$. For completeness, we let $\mathcal{F}_{H+1} = \{f : f(s,a) = 0 \, \forall (s,a) \in \mathcal{S} \times \mathcal{A}\}$ be the singleton set containing only the degenerate function mapping all inputs to 0.

We make two common assumptions about the expressiveness of the function class $\mathcal{F}$ [Antos et al., 2008, Xie et al., 2021].

Assumption 4.2.3 (Approximate Realizability)**.** *For any $r \in \widetilde{\mathcal{R}}$ and $\pi \in \{\mathcal{S} \to \Delta(\mathcal{A})\}^H$, there exists some $f_r^\pi \in \mathcal{F}$ such that for all $h \in [H]$,*

$$\sup_{\pi' \in \{\mathcal{S} \to \Delta(\mathcal{A})\}^H} \mathbb{E}_{\pi'_h} \left[\| f_{h,r}^\pi(\cdot, \cdot; r) - Q_h^\pi(\cdot, \cdot; r) \|^2 \right] \leqslant \epsilon_{\mathcal{F}}.$$

Intuitively, Assumption 4.2.3 dictates that for all reported reward functions r and all policies π, there exists a function in $\mathcal{F}$ that can approximate Q_r^π sufficiently well.

Assumption 4.2.4 (Approximate Completeness)**.** *For any $h \in [H], r \in \widetilde{\mathcal{R}}$, and $\pi \in \{\mathcal{S} \to \Delta(\mathcal{A})\}^H$, we have*

$$\sup_{f \in \mathcal{F}_{h+1}} \inf_{f' \in \mathcal{F}_h} \mathbb{E}_{\mu_h}[\| f' - \mathcal{T}_{h,r}^\pi f \|^2] \leqslant \epsilon_{\mathcal{F},\mathcal{F}}.$$

Assumption 4.2.4 requires the function class $\mathcal{F}$ to be approximately closed for all reported reward functions and policies. The assumption is prevalent in RL and can be omitted only in rare circumstances [Xie and Jiang, 2021].

A fundamental problem in offline RL is the distribution shift, which occurs when the data generating distribution has only a partial coverage of the policy of interest [Jin et al., 2021b, Zanette et al., 2021]. We address the issue with the help of distribution shift coefficient [Xie et al., 2021].

Definition 4.2.5 (Distribution Shift Coefficient)**.** *Let $C^\pi(\nu)$ be the measure of distribution shift from an arbitrary distribution over $(\mathcal{S} \times \mathcal{A})^H$, denoted ν, to the data distribution μ, when measured under the transition dynamics induced by a policy $\pi \in \{\mathcal{S} \to \Delta(\mathcal{A})\}^H$. In particular,*

$$C^\pi(\nu) = \max_{f^1, f^2 \in \mathcal{F}} \max_{h \in [H]} \max_{r \in \widetilde{\mathcal{R}}} \frac{\mathbb{E}_{\nu_h}[\|f_h^1 - \mathcal{T}_{h,r}^\pi f_{h+1}^2\|^2]}{\mathbb{E}_{\mu_h}[\|f_h^1 - \mathcal{T}_{h,r}^\pi f_{h+1}^2\|^2]}.$$

The coefficient controls how well the Bellman estimation error shifts from one distribution to another for any Bellman transition operator $\mathcal{T}$. For a detailed discussion on how the coefficient generalizes previous measures of distribution shift, please refer to Xie et al. [2021]. As a shorthand notation, when ν is the visitation measure induced by some policy π', we let $C^\pi(\pi') = C^\pi(d_{\pi'}) = C^\pi(\nu)$.

In offline learning, with a finite data set, we can only hope to learn the desired mechanism up to certain statistical error. In particular, we state the approximate versions of the desiderata for finite-sample analysis.

1. *Asymptotic efficiency:* If all agents report truthfully, a mechanism is asymptotically efficient if $\text{SubOpt}(\pi; s_0) \in \mathcal{O}(K^{-\alpha})$ for some $\alpha \in (0, 1)$.

2. *Asymptotic individual rationality:* Let π, p_i be the policy and price chosen by the mechanism when the agent i is truthful. A dynamic mechanism is asymptotically

individually rational if $U_i^{\pi}(p_i) = -\mathcal{O}(K^{-\alpha})$ for some $\alpha \in (0,1)$, regardless of the truthfulness of other agents.

3. *Asymptotic truthfulness:* Let $\widetilde{\pi}, \widetilde{p}_i$ be the policy and price chosen by the mechanism when the agent i is untruthful, and π, p_i those chosen by the mechanism when the agent i is truthful. We say a dynamic mechanism is asymptotically truthful if $U_i^{\widetilde{\pi}}(\widetilde{p}_i) - U_i^{\pi}(p_i) = \mathcal{O}(K^{-\alpha})$ for some $\alpha \in (0,1)$ regardless of the truthfulness of other agents.

As we will see in sequel, we propose a soft policy iteration algorithm that simultaneously satisfies all three criteria above with $\alpha = 1/3$ up to function approximation biases.

4.3 Offline RL for VCG

We develop an algorithm that learns the dynamic VCG mechanism via offline RL. We begin by sketching out a basic outline of our algorithm. Recall the dynamic VCG mechanism given in Definition 4.2.1. At a high level, an algorithm that learns the dynamic VCG mechanism can be summarized as the following procedure.

1. Learn some policy $\breve{\pi}$ such that the social welfare suboptimality $\mathrm{SubOpt}(\breve{\pi}; s_0)$ is small.
2. For all $i \in [n]$, estimate the VCG price p_i, defined in (4.2.2), as $\widehat{p}_i = G_{-i}^{(1)}(s_0) - G_{-i}^{(2)}(s_0)$, where $G_{-i}^{(1)}(s_0)$ estimates $V_1^*(s_0; \widetilde{R}_{-i})$ and $G_{-i}^{(2)}(s_0)$ estimates $V_1^{\breve{\pi}}(s_0; \widetilde{R}_{-i})$.

Step 1 simply minimizes the social welfare suboptimality using offline RL and has been extensively studied in prior literature [Jin et al., 2021b, Zanette et al., 2021, Xie et al., 2021, Uehara and Sun, 2021].

A greater challenge lies in implementing Step 2 and showing that the price estimates, $\{\widehat{p}_i\}_{i=1}^n$, satisfy all three approximate mechanism design desiderata. The estimate $G_{-i}^{(2)}(s_0)$ can be constructed by performing a policy evaluation of the learned policy, $\breve{\pi}$. The construction of $G_{-i}^{(1)}(s_0)$ is more challenging, involving two separate steps: (1) learning a fictitious

policy that approximately maximizes $V_1^{\pi}(s_0; \widetilde{R}_{-i})$ over π from offline data, and (2) performing a policy evaluation of the learned fictitious policy to obtain the estimate of the value function. Consequently, the policy evaluation and policy improvement subroutines are necessary for learning $G_{-i}^{(1)}(s_0)$ and implementing Step 2.

Our challenge is complicated by the fact that a combination of optimism and pessimism is needed for price estimation, whereas the typical offline RL literature only leverages pessimism [Jin et al., 2021b, Uehara and Sun, 2021, Xie et al., 2021]. For example, when $G_{-i}^{(1)}(s_0)$ is a pessimistic estimate of $V_1^*(s_0; \widetilde{R}_{-i})$, the price estimate $\widehat{p}_i$ is a "lower bound," at least in the first term, of the actual price p_i derived in (4.2.2). A lower price estimate would be beneficial to the agent, but would increase the seller's suboptimality since, loosely speaking, the seller is "paying for" the uncertainty in the data set, and the reverse holds when $G_{-i}^{(1)}(s_0)$ is an optimistic estimate. The party burdened with the cost of uncertainty may be different in different settings. When allocating public goods, for instance, the cost of uncertainty should be the seller's burden to better benefit the public [Bergemann and Välimäki, 2019], whereas a company wishing to maximize their profit would prefer having the agents "pay for" uncertainty [Friedman and Parkes, 2003].

To allow for such flexibility, we introduce hyperparameters $\zeta_1, \zeta_2 \in \{\texttt{PES}, \texttt{OPT}\}$, where ζ_1 determines whether $G_{-i}^{(1)}(s_0)$ is a PESsimistic or OPTimistic estimate and ζ_2 does so for $G_{-i}^{(2)}(s_0)$. To highlight the trade-off between agents' and seller's suboptimalities, we focus on the two extreme cases, $(\zeta_1, \zeta_2) = (\texttt{PES}, \texttt{OPT})$ and $(\zeta_1, \zeta_2) = (\texttt{OPT}, \texttt{PES})$, where the former favors the agents and the latter the seller. Depending on the goal of the mechanism designer, different choices of ζ_1, ζ_2 may be selected to favor agents or the seller [Maskin, 2008].

With the crucial challenges identified, we introduce the specific algorithms that we use to implement Steps 1 and 2.

4.3.1 Policy Evaluation and Soft Policy Iteration

We use optimistic and pessimistic variants of soft policy iteration, commonly used for policy improvement [Xie et al., 2021, Cai et al., 2020, Zanette et al., 2021]. At a high level, each iteration of the soft policy iteration consists of two steps: policy evaluation and policy improvement.

We begin by describing our policy evaluation algorithm. The Bellman error can be written as $f_h(s,a) - \mathcal{T}^{\pi}_{h,r} f_{h+1}(s,a)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $h \in [H]$, and the estimate of the action value function $f \in \mathcal{F}$ for policy π and reward r. We construct an empirical estimate of the Bellman error as follows. For any $h \in [H]$, $f, f' \in \mathcal{F}$ and $r \in \widetilde{\mathcal{R}}$, we define $\mathcal{L}_{h,r}(f_h, f'_{h+1}, \pi; \mathcal{D})$ as

$$\mathcal{L}_{h,r}(f_h, f'_{h+1}, \pi; \mathcal{D}) = \frac{1}{K} \sum_{\tau=1}^{K} (f_h(s_h^\tau, a_h^\tau) - r_h(s_h^\tau, a_h^\tau) - f'_{h+1}(s_{h+1}^\tau, \pi_{h+1}))^2,$$

where we slightly abuse the notation and let r_h^τ be the reported rewards $\widetilde{r}_{i,h}^\tau$ summed over i according to the chosen reported reward function $r \in \widetilde{\mathcal{R}}$. Recall that $\widetilde{\mathcal{R}} = \{\widetilde{R}_{-i}\}_{i=1}^n \cup \{\widetilde{R}\}$ is the set of reported reward functions whose action-value functions need to be estimated. The empirical estimate for Bellman error under policy π at step h is then constructed as

$$\mathcal{E}_{h,r}(f, \pi; \mathcal{D}) = \mathcal{L}_{h,r}(f_h, f_{h+1}, \pi; \mathcal{D}) - \min_{g \in \mathcal{F}_h} \mathcal{L}_{h,r}(g, f_{h+1}, \pi; \mathcal{D}). \tag{4.3.1}$$

The goal of the policy evaluation algorithm is to solve the following regularized optimization problems:

$$\begin{aligned} \widehat{Q}_r^\pi &= \operatorname*{argmin}_{f \in \mathcal{F}} -f_1(s_0, \pi) + \lambda \sum_{h=1}^{H} \mathcal{E}_{h,r}(f, \pi; \mathcal{D}), \\ \check{Q}_r^\pi &= \operatorname*{argmin}_{f \in \mathcal{F}} f_1(s_0, \pi) + \lambda \sum_{h=1}^{H} \mathcal{E}_{h,r}(f, \pi; \mathcal{D}), \end{aligned} \tag{4.3.2}$$

thereby obtaining optimistic and pessimistic estimates of $Q^\pi(\cdot, \cdot; r)$ for any policy π and

reward function r. We summarize the procedure in Algorithm 10.

Algorithm 10 Policy Evaluation

Input: Reported reward $r \in \widetilde{\mathcal{R}}$, regularization coefficient λ, policy π, and dataset $\mathcal{D} = \{(x_h^\tau, \omega_h^\tau, \{\widetilde{r}_{i,h}^\tau\}_i^n)\}_{h,\tau=1}^{H,K}$.

1: For all h, τ, calculate r_h^τ as the sum of $\widetilde{r}_{i,h}^\tau$ over i according to the reported reward function r.
2: Obtain the optimistic and pessimistic estimates of Q_r^π using (4.3.2)
3: Return action-value function estimates $\widehat{Q}_r^\pi, \check{Q}_r^\pi$.

Next, we introduce the policy improvement procedure. At each step $t \in [T]$, we use the mirror descent with the Kullback-Leibler (KL) divergence to update the policies for all $(s, a) \in \mathcal{S} \times \mathcal{A}, h \in [H]$. By direct computation, the update rule can be written as

$$\widehat{\pi}_{h,r}^{(t+1)}(a|s) \propto \widehat{\pi}_{h,r}^{(t)}(a|s) \exp\left(\eta \widehat{Q}_{h,r}^{(t)}(s,a)\right), \tag{4.3.3}$$

$$\check{\pi}_{h,r}^{(t+1)}(a|s) \propto \check{\pi}_{h,r}^{(t)}(a|s) \exp\left(\eta \check{Q}_{h,r}^{(t)}(s,a)\right), \tag{4.3.4}$$

where $\widehat{Q}_{h,r}, \check{Q}_{h,r}$ are the action-value function estimates obtained from (4.3.2) [Bubeck et al., 2015, Cai et al., 2020, Xie et al., 2021].

For any set of T policies $\{\pi^{(t)}\}_{t=1}^T$, let $\mathrm{Unif}(\{\pi^{(t)}\}_{t=1}^T)$ be the mixture policy formed by selecting one of $\{\pi^{(t)}\}_{t=1}^T$ uniformly at random. The output of our policy improvement algorithm is then given by $\mathrm{Unif}(\{\widehat{\pi}_r^{(t)}\}_{t=1}^T)$ and $\mathrm{Unif}(\{\check{\pi}_r^{(t)}\}_{t=1}^T)$, that is, the uniform mixture of optimistic and pessimistic policy estimates. We summarize the soft policy iteration algorithm in the form of pseudocode in Algorithm 11.

We defer the pseudocode of our main algorithm to Section 4.5.2 in the form of Algorithm 12, as its construction is apparent given the two key subroutines above.

Algorithm 11 Soft Policy Iteration for Episodic MDPs

Input: Reported reward $r \in \widetilde{\mathcal{R}}$, regularization coefficient λ, number of iterations T, learning rate η , and dataset $\mathcal{D} = \{(x_h^\tau, \omega_h^\tau, \{\widetilde{r}_{i,h}^\tau\}_i^n)\}_{h,\tau=1}^{H,K}$.

1: Initialize optimistic and pessimistic polices, $\widehat{\pi}_r^{(1)}$ and $\breve{\pi}_r^{(1)}$, as the uniform policy.
2: **for** $t = 1, \ldots, T$ **do**
3: Obtain the optimistic and pessimistic estimates of $Q_r^{\widehat{\pi}_r^{(t)}}$ and $Q_r^{\breve{\pi}_r^{(t)}}$ by Algorithm 10.
4: Update policy estimates according to (4.3.3) and (4.3.4).
5: **end for**
6: Let $\widehat{\pi}_r^{\text{out}} = \text{Unif}(\{\widehat{\pi}_r^{(t)}\}_{t=1}^T)$, $\breve{\pi}_r^{\text{out}} = \text{Unif}(\{\breve{\pi}_r^{(t)}\}_{t=1}^T)$.
7: Execute Algorithm 10 to construct optimistic action-value function $\widehat{Q}_r^{\text{out}}$ for $\widehat{\pi}_r^{\text{out}}$ and pessimistic action-value function $\breve{Q}_r^{\text{out}}$ for $\breve{\pi}_r^{\text{out}}$, respectively.
8: **Return** $\{\widehat{\pi}_r^{\text{out}}, \widehat{Q}_r^{\text{out}}\}$ and $\{\breve{\pi}_r^{\text{out}}, \breve{Q}_r^{\text{out}}\}$.

4.4 Main Results

We begin by formally defining the policy class induced by the policy improvement algorithm, Algorithm 11. It is a well-known result that policy iterates induced by mirror descent-style updates in (4.3.3) and (4.3.4) are in the natural policy class attained by soft policy iteration over $\mathcal{F}$ [Cai et al., 2020, Agarwal et al., 2021, Xie et al., 2021, Zanette et al., 2021], given by

$$\Pi_{\text{It}} = \left\{ \pi_h'(\cdot|s) \propto \exp\left(\eta \sum_{t=1}^{T} f_h^t(s, \cdot) \right) : h \in [H], \{f_h^{(t)}\}_{t=1}^T \subseteq \mathcal{F}_h \right\}.$$

Let Π_{SPI} denote the following set of policies

$$\Pi_{\text{SPI}} = \Pi_{\text{It}} \Big\{ \pi : \pi = \text{Unif}(\{\pi^{(t)}\}_{t=1}^T), \{\pi^{(t)}\}_{t=1}^T \subset \Pi_{\text{It}} \Big\}. \tag{4.4.1}$$

Before stating the main result, we introduce an additional notation. The statistical error Err^{stat} denotes

$$\text{Err}^{\text{stat}} = \widetilde{\mathcal{O}}\left(H(HR_{\max})^{5/3} K^{-1/3} \right) + \widetilde{\mathcal{O}}\left(H\left((HR_{\max})^{1/3} \epsilon_{\mathcal{F}}^{1/3} + \sqrt{\epsilon_{\mathcal{F}} + \epsilon_{\mathcal{F},\mathcal{F}}} \right) \right),$$

while the optimization error $\mathrm{Err}^{\mathrm{opt}}$ denotes

$$\mathrm{Err}^{\mathrm{opt}} = \widetilde{\mathcal{O}}\left(H^2 R_{\max}\sqrt{1/T}\right).$$

To differentiate the policies learned under different truthfulness assumptions, let $\breve{\pi} = \breve{\pi}_R^{\mathrm{out}}$ be the policy chosen by the algorithm when all agents are truthful, let $\widetilde{\pi} = \breve{\pi}^{\mathrm{out}}_{r_i+\widetilde{R}_{-i}}$ be the policy chosen when we only assume the agent i is truthful, and let $\breve{\pi}_{\widetilde{R}} = \breve{\pi}^{\mathrm{out}}_{\widetilde{R}}$ be the policy chosen when no agent is truthful. Let $\breve{\pi}^{(t)}, \widetilde{\pi}^{(t)}, \breve{\pi}^{(t)}_{\widetilde{R}}$ be the iterates of Algorithm 11 when learning these policies. Denote the prices charged by $\{\widehat{p}_i\}_{i=1}^n, \{\widetilde{p}_i\}_{i=1}^n$, and $\{\widehat{p}_{i,\widetilde{R}}\}_{i=1}^n$, respectively.

We then summarize the performance of our learned mechanism with asymptotic bounds in Theorem 4.4.1. Theorem 4.5.1 presented in Appendix 4.5.3 provides a more detailed result.

Theorem 4.4.1 (Informal)**.** *With probability at least $1-\delta$, with suitable choices of λ, δ, under Assumptions 4.2.3 and 4.2.4, the following claims hold simultaneously.*

1. *Algorithm 12 returns a mechanism that is asymptotically efficient. More specifically, assuming all agents report truthfully, we have*

$$\mathrm{SubOpt}(\breve{\pi}; s_0) \leqslant \mathrm{Err}^{\mathrm{opt}} + \left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}}(\pi^*)}\right)\mathrm{Err}^{\mathrm{stat}}.$$

2. *Assuming all agents report truthfully, when $(\zeta_1, \zeta_2) = (\texttt{PES}, \texttt{OPT})$, we have*

$$\mathrm{SubOpt}_i(\breve{\pi}, \{\widehat{p}_i\}_{i=1}^n; s_0) \leqslant \mathrm{Err}^{\mathrm{opt}} + \left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}}(\pi^*)}\right)\mathrm{Err}^{\mathrm{stat}}.$$

When $(\zeta_1, \zeta_2) = (\texttt{OPT}, \texttt{PES})$*, we have*

$$\mathrm{SubOpt}_i(\breve{\pi}, \{\widehat{p}_i\}_{i=1}^n; s_0) \leqslant \mathrm{Err}^{\mathrm{opt}} + \mathrm{Err}^{\mathrm{stat}}\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}}(\pi^*)} + \sqrt{C^{\widehat{\pi}_{-i}}(\widehat{\pi}_{-i})} + \sqrt{C^{\breve{\pi}}(\breve{\pi})}\right).$$

3. *Assuming all agents report truthfully, when* $(\zeta_1, \zeta_2) = (\texttt{PES}, \texttt{OPT})$*, we have*

$$\mathrm{SubOpt}_0(\breve{\pi}, \{\widehat{p}_i\}_{i=1}^n; s_0) \leqslant n\mathrm{Err}^{\mathrm{opt}} + \mathrm{Err}^{\mathrm{stat}}\left(\sum_{i=1}^{n}\sqrt{C^{\breve{\pi}_{-i}}(\breve{\pi}_{-i})} + n\sqrt{C^{\breve{\pi}}(\breve{\pi})} + \sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}_{R_{-i}}}(\pi^*_{-i})}\right).$$

and, when $(\zeta_1, \zeta_2) = (\texttt{OPT}, \texttt{PES})$*, we have*

$$\mathrm{SubOpt}_0(\breve{\pi}, \{\widehat{p}_i\}_{i=1}^n; s_0) \leqslant n\mathrm{Err}^{\mathrm{opt}}\mathrm{Err}^{\mathrm{stat}}\left(\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widehat{\pi}^{(t)}_{R_{-i}}}(\widehat{\pi}^{(t)}_{R_{-i}})} + \sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widehat{\pi}^{(t)}_{R_{-i}}}(\pi^*_{-i})}\right).$$

4. *Algorithm 12 returns a mechanism that is asymptotically individually rational. More specifically, even when other agents are untruthful, when* $(\zeta_1, \zeta_2) = (\texttt{PES}, \texttt{OPT})$ *and the agent* i *is truthful, their utility satisfies*

$$U_i^{\widetilde{\pi}}(\widetilde{p}_i) \geqslant -\mathrm{Err}^{\mathrm{opt}} - \mathrm{Err}^{\mathrm{stat}}\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}_{\widetilde{R}_{-i}}}(\widetilde{\pi}^*_{-i})} + \sqrt{C^{\breve{\pi}^{\mathrm{out}}_{\widetilde{R}_{-i}}}(\breve{\pi}^{\mathrm{out}}_{\widetilde{R}_{-i}})} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widetilde{\pi}^{(t)}}(\pi^*_{r_i+\widetilde{R}_{-i}})}\right).$$

and when $(\zeta_1, \zeta_2) = (\texttt{OPT}, \texttt{PES})$ *and the agent* i *is truthful, their utility satisfies*

$$U_i^{\widetilde{\pi}}(\widetilde{p}_i) \geqslant -\operatorname{Err}^{\mathrm{opt}} - \operatorname{Err}^{\mathrm{stat}}\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widetilde{\pi}^{(t)}}(\pi^*_{r_i+\widetilde{R}_{-i}})} + \sqrt{C^{\widehat{\pi}^{(t)}_{\widetilde{R}_{-i}}}(\widetilde{\pi}^*_{-i})} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widehat{\pi}^{(t)}_{\widetilde{R}_{-i}}}(\widehat{\pi}^{(t)}_{\widetilde{R}_{-i}})} + \sqrt{C^{\widetilde{\pi}}(\widetilde{\pi})}\right).$$

5. *Algorithm 12 returns a mechanism that is asymptotically truthful. More specifically, even when all the other agents are untruthful and irrespective of whether the agent* i *is truthful or not, for all* $i \in [n]$ *when* $\zeta_2 = \texttt{OPT}$ *the amount of utility gained by untruthful reporting is upper bounded as*

$$U_i^{\breve{\pi}_{\widetilde{R}}}(\widehat{p}_{i,\widetilde{R}}) - U_i^{\widetilde{\pi}}(\widetilde{p}_i) \leqslant \operatorname{Err}^{\mathrm{opt}} + \operatorname{Err}^{\mathrm{stat}}\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widetilde{\pi}^{(t)}}(\pi^*_{r_i+\widetilde{R}_{-i}})} + \sqrt{C^{\breve{\pi}_{\widetilde{R}}}(\breve{\pi}_{\widetilde{R}}))}\right),$$

and when $\zeta_2 = \texttt{PES}$*, the amount of utility gained by untruthful reporting is upper bounded as*

$$U_i^{\breve{\pi}_{\widetilde{R}}}(\widehat{p}_{i,\widetilde{R}}) - U_i^{\widetilde{\pi}}(\widetilde{p}_i) \leqslant \operatorname{Err}^{\mathrm{opt}} + \operatorname{Err}^{\mathrm{stat}}\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widetilde{\pi}^{(t)}}(\pi^*_{r_i+\widetilde{R}_{-i}})} + \sqrt{C^{\widetilde{\pi}}(\widetilde{\pi}))}\right).$$

Proof. See Section 4.5.3 for a detailed proof. □

We make a few remarks about Theorem 4.4.1.

Dependence on the number of trajectories K. The only term that depends on the number of trajectories K is the statistical error $\operatorname{Err}^{\mathrm{stat}}$ and it decays at the $\widetilde{\mathcal{O}}(K^{-1/3})$ rate, matching the sample complexity of the pessimistic soft policy iteration algorithm [Xie et al., 2021]. When data set has coverage of the optimal policy and no function approximation bias, our algorithm converges sublinearly to a mechanism with suboptimality $\mathcal{O}(K^{1/3})$.

Furthermore, when data set has sufficient coverage over all policies and the function class satisfies Assumptions 4.2.3 and 4.2.4 exactly, our algorithm is asymptotically individually rational and truthful at the same $\mathcal{O}(K^{1/3})$ rate, a result that is not implied by the existing literature on offline RL [Xie et al., 2021, Jin et al., 2021b, Zanette et al., 2021].

Dependence on ζ_1, ζ_2. Observe that ζ_1 and ζ_2 affect the bounds in Theorem 4.4.1 by changing the distribution shift coefficients involved for each suboptimality. The inclusion of optimism in offline RL for mechanism design is crucial, as the optimal individual suboptimality rate is attainable only when $\zeta_1 = \texttt{OPT}$. Different from the existing work on offline RL which extensively uses pessimism, we demonstrate the importance and necessity of optimism when offline RL is used to help design dynamic mechanisms [Xie et al., 2021, Jin et al., 2021a, Zanette et al., 2021].

Dependence on $\mathcal{F}, \Pi_{\text{SPI}}$. The statistical error term Err^{stat} is the only term that depends on $\mathcal{F}, \Pi_{\text{SPI}}$ through the log covering numbers of $\mathcal{F}$ and Π_{SPI}. The covering numbers are formally defined in Appendix 4.5.5 and the theorem's dependence on the covering number is made explicit in the non-asymptotic version, Theorem 4.5.1. We emphasize that our results are directly applicable to general, continuous function classes via a covering-based argument, improving over the results in Xie et al. [2021].

Comparison to related work. While deep RL algorithms such as conservative Q-learning [Kumar et al., 2020], conservative offline model-based policy optimization [Yu et al., 2021], and decision transformer [Chen et al., 2021a] have achieved empirical success on popular offline RL benchmarks, such algorithms rarely have theoretical guarantees without strong coverage assumptions. Within a mechanism design context, such a lack of theoretical guarantees is particularly problematic, as we cannot ensure that the learned mechanism is individually rational or truthful, potentially leading to significant ethical issues when applied to real-world problems. When compared to Xie et al. [2021], our work features a streamlined, simplified theoretical analysis, which we sketch below, that is directly applicable when

both $|\mathcal{F}|$ and $|\Pi|$ are unbounded using a covering-based argument, whereas the convergence bounds in Xie et al. [2021] grows linearly in the term $\sqrt{\frac{\log|\mathcal{F}||\Pi|/\delta}{K}}$ in the general function approximation setting.

4.5 Technical Details

4.5.1 Proof of Proposition 4.2.2

Those familiar with the literature on mechanism design may quickly realize that our price function is derived using the Clarke pivot rule [Nisan et al., 2007]. The result is directly derived from the properties of the VCG mechanism [Nisan et al., 2007, Parkes, 2007, Hartline, 2012]. We include a full proof for completeness.

With $\mathcal{P}$ and $\{\widetilde{r}_i\}_{i=0}^n$ given, the state-value functions $V_h^\pi(s_0, r)$ can be explicitly calculated for all $h \in [H], r \in \widetilde{\mathcal{R}}$. We can then obtain exactly $\widetilde{\pi}^*$ and directly calculate $p_i = V_1^*(s_0, \widetilde{R}_{-i}) - V_1^{\widetilde{\pi}^*}(s_0, \widetilde{R}_{-i})$. Thus, the proposed mechanism is feasible when the rewards and transition kernel are known.

For convenience, let

$$\pi^{(1)} = \pi^*_{r_i + \widetilde{R}_{-i}} = \operatorname*{argmax}_{\pi \in \Pi} V_1^\pi(s_0; r_i + \widetilde{R}_{-i}) \quad \text{and} \quad \pi^{(2)} = \pi^*_{\widetilde{R}} = \operatorname*{argmax}_{\pi \in \Pi} V_1^\pi(s_0; \widetilde{R}),$$

denote the policies chosen by the mechanism when the agent i is truthful and untruthful, respectively, without assumptions on the truthfulness of other agents.

We now show that the three desiderata are satisfied by the mechanism.

1. Efficiency. When the agents report $\{r_i\}_{i=1}^n$ truthfully, the chosen policy π^* maximizes the social welfare and is efficient by definition.

2. Individual rationality. The price charged from the agent i is

$$p_i = V_1^*(s_0; \widetilde{R}_{-i}) - V_1^{\pi^{(2)}}(s_0; \widetilde{R}_{-i}).$$

   Our goal is to then show that $V_1^{\pi^{(2)}}(s_0; \widetilde{r}_i) \geqslant p_i$. That is, the value function of the

reported reward is no less than the price charged. Observe that

$$V_1^{\pi^{(2)}}(s_0; \widetilde{r}_i) - \widetilde{p}_i = V_1^{\pi^{(2)}}(s_0; \widetilde{R}) - V_1^*(s_0; \widetilde{R}_{-i}).$$

Let $\pi_{-i}^{(2)} = \text{argmax}_{\pi \in \Pi} V_1^{\pi}(s_0; \widetilde{R}_{-i})$. Then we know that

$$V_1^{\pi^{(2)}}(s_0; \widetilde{r}_i) - \widetilde{p}_i \geqslant V_1^{\pi_{-i}^{(2)}}(s_0; \widetilde{R}) - V_1^{\pi_{-i}^{(2)}}(s_0; \widetilde{R}_{-i}) = V_1^{\pi_{-i}^{(2)}}(s_0; \widetilde{r}_i) \geqslant 0.$$

3. Truthfulness: If $\widetilde{r}_i = r_i$, that is, the agent i reports truthfully, they attain the following utility

$$\begin{aligned} U_i^{\pi^{(1)}}(p_i) &= V_1^{\pi^{(1)}}(s_0; r_i) - V_1^*(s_0; \widetilde{R}_{-i}) + V_1^{\pi^{(1)}}(s_0; \widetilde{R}_{-i}) \\ &= V_1^{\pi^{(1)}}(s_0; r_i + \widetilde{R}_{-i}) - V_1^*(s_0; \widetilde{R}_{-i}). \end{aligned}$$

When the agent reports some arbitrary $\widetilde{r}_i$, the agent receives instead

$$\begin{aligned} U_i^{\pi^{(2)}}(p_i) &= V_1^{\pi^{(2)}}(s_0; r_i) - V_1^*(s_0; \widetilde{R}_{-i}) + V_1^{\pi^{(2)}}(s_0; \widetilde{R}_{-i}) \\ &= V_1^{\pi^{(2)}}(s_0; r_i + \widetilde{R}_{-i}) - V_1^*(s_0; \widetilde{R}_{-i}). \end{aligned}$$

Since $\pi^{(1)}$ maximizes $V_1^{\pi}(s_0; r_i + \widetilde{R}_{-i})$, $u_i \geqslant \widetilde{u}_i$ regardless of other agents' reported reward $\{\widetilde{r}_j\}_{j \neq i}$ and the mechanism is truthful.

4.5.2 Pseudocode for Offline VCG Learn

Let $\mathcal{N}_\infty(\epsilon, \mathcal{F})$ be the ϵ-covering number of $\mathcal{F}$ with respect to the ℓ_∞-norm, that is, the cardinality of the smallest set of functions $\{f^l\}_{l=1}^{N_L}$ such that for all $f \in \mathcal{F}$ there exists some $l \in [L]$ such that

$$\max_{h \in [H]} \sup_{s \in \mathcal{S}, a \in \mathcal{A}} |f_h^l(s, a) - f_h(s, a)| \leqslant \epsilon.$$

We also let $\mathcal{N}_{\infty,1}(\epsilon, \Pi)$ be the ϵ-covering number of Π with respect to the following norm:

$$\ell_{\infty,1}(\pi - \pi') = \sup_{h\in[H], s\in\mathcal{S}} \sum_{a\in\mathcal{A}} |\pi_h(a|s) - \pi'_h(a|s)|.$$

With the covering numbers defined, we introduce the main algorithm and the parameter choices for the algorithm, which depend on the covering numbers. For the main algorithm, we set

$$\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^2}\right)^{1/3}, \quad \eta = \sqrt{\frac{\log|\mathcal{A}|}{2H^2R_{\max}^2 T}}, \tag{4.5.1}$$

where

$$\epsilon_{\mathrm{S}} = \frac{5136}{K} H^4 R_{\max}^4 \log\left(56nH \cdot \mathcal{N}_\infty\left(\frac{19H^3R_{\max}^3}{K}, \mathcal{F}\right) \cdot \mathcal{N}_{\infty,1}\left(\frac{19H^4R_{\max}^4}{K}, \Pi_{\mathrm{SPI}}\right) \Big/ \delta\right).$$

The pseudocode for our main algorithm can then be summarized as Algorithm 12.

Algorithm 12 Offline VCG Learn

Input: Hyperparameters $\zeta_1, \zeta_2 \in \{\texttt{OPT}, \texttt{PES}\}$, regularization coefficient λ, number of iterations T, learning rate η.

1: Let $\breve{\pi}^{\mathrm{out}}_{\widetilde{R}}$ be the pessimistic policy output of Algorithm 11 with $r = \widetilde{R}$, T, and λ, η set according to (4.5.1).
2: **for** Agent $i = 1, 2, \ldots, n$ **do**
3: Call Algorithm 11 with $r = \widetilde{R}_{-i}$, T, and λ, η set according to (4.5.1).
4: If $\zeta_1 = \texttt{OPT}$, let $G_{-i}^{(1)}(s_0) = \widehat{Q}^{\mathrm{out}}_{1,\widetilde{R}_{-i}}(s_0, \widehat{\pi}^{\mathrm{out}}_{1,\widetilde{R}_{-i}})$. Otherwise let $G_{-i}^{(1)}(s_0) = \breve{Q}^{\mathrm{out}}_{1,\widetilde{R}_{-i}}(s_0, \breve{\pi}^{\mathrm{out}}_{1,\widetilde{R}_{-i}})$.
5: Call Algorithm 10 with $r = \widetilde{R}_{-i}$, $\pi = \breve{\pi}^{\mathrm{out}}_{\widetilde{R}}$, and λ set according to (4.5.1).
6: If $\zeta_2 = \texttt{OPT}$, let $G_{-i}^{(2)}(s_0) = \widehat{Q}^{\breve{\pi}^{\mathrm{out}}_{\widetilde{R}}}_{1,\widetilde{R}_{-i}}(s_0, \breve{\pi}^{\mathrm{out}}_{1,\widetilde{R}})$.
Otherwise let $G_{-i}^{(2)}(s_0) = \breve{Q}^{\breve{\pi}^{\mathrm{out}}_{\widetilde{R}}}_{1,\widetilde{R}_{-i}}(s_0, \breve{\pi}^{\mathrm{out}}_{1,\widetilde{R}})$.
7: Set the estimated price $\widehat{p}_i = G_{-i}^{(1)}(s_0) - G_{-i}^{(2)}(s_0)$.
8: **end for**
9: Return policy $\breve{\pi}^{\mathrm{out}}_{\widetilde{R}}$ and estimated prices $\{\widehat{p}_i\}_{i=1}^n$.

4.5.3 Proof of Theorem 4.4.1

We re-state Theorem 4.4.1 in a finite sample form.

Theorem 4.5.1 (Theorem 4.4.1 restated)**.** *Suppose that λ, η are set according to* (4.5.1) *and Assumptions 4.2.3 and 4.2.4 hold. Then, with probability at least $1-\delta$, the following holds simultaneously.*

1. *Assuming all agents report truthfully, the suboptimality of the output policy $\breve{\pi}$ is bounded as*

$$\begin{aligned}\mathrm{SubOpt}(\breve{\pi}; s_0) \leqslant{}& 2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &+ H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}}(\pi^*)}\right) \\ &\quad\times\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).\end{aligned}$$

2. *Assuming all agents report truthfully, when $(\zeta_1, \zeta_2) = (\texttt{PES}, \texttt{OPT})$, the agent i's suboptimality, for all $i \in [n]$, satisfies*

$$\begin{aligned}\mathrm{SubOpt}_i(\breve{\pi}, \{\widehat{p}_i\}_{i=1}^n; s_0) \leqslant{}& 2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + 3\sqrt{\epsilon_{\mathcal{F}}} + 6(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &+ H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}}(\pi^*)}\right) \\ &\quad\times\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right),\end{aligned}$$

and when $(\zeta_1, \zeta_2) = (\texttt{OPT}, \texttt{PES})$*, the agent* i*'s suboptimality, for all* $i \in [n]$*, satisfies*

$$\begin{aligned}\mathrm{SubOpt}_i(\breve{\pi}, \{\widehat{p}_i\}_{i=1}^n; s_0) \leqslant{}& 2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &+ H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}}(\pi^*)} + \sqrt{C^{\widehat{\pi}_{-i}}(\widehat{\pi}_{-i})} + \sqrt{C^{\breve{\pi}}(\breve{\pi})}\right) \\ &\times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).\end{aligned}$$

3. *Assuming all agents report truthfully, when* $(\zeta_1, \zeta_2) = (\texttt{PES}, \texttt{OPT})$*, the seller's suboptimality satisfies*

$$\begin{aligned}\mathrm{SubOpt}_0(\breve{\pi}, \{\widehat{p}_i\}_{i=1}^n; s_0) \leqslant{}& 2nH^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + n\sqrt{\epsilon_{\mathcal{F}}} \\ &+ 2n(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &+ H\left(\sum_{i=1}^{n}\left(\sqrt{C^{\breve{\pi}_{-i}}(\breve{\pi}_{-i})} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}_{R_{-i}}}(\pi^*_{-i})}\right) + n\sqrt{C^{\breve{\pi}}(\breve{\pi})}\right) \\ &\times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right),\end{aligned}$$

and when $(\zeta_1, \zeta_2) = (\texttt{OPT}, \texttt{PES})$*, the seller's suboptimality satisfies*

$$\begin{aligned}\mathrm{SubOpt}_0(\breve{\pi}, \{\widehat{p}_i\}_{i=1}^n; s_0) \leqslant{}& 2nH^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + 2n\sqrt{\epsilon_{\mathcal{F}}} \\ &+ 4n(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &+ H\left(\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\left(\sqrt{C^{\widehat{\pi}^{(t)}_{R_{-i}}}(\pi^*_{-i})} + \sqrt{C^{\widehat{\pi}^{(t)}_{R_{-i}}}(\widehat{\pi}^{(t)}_{R_{-i}})}\right)\right) \\ &\times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).\end{aligned}$$

4. *(Asymptotic Individual Rationality) When* $(\zeta_1, \zeta_2) = (\texttt{PES}, \texttt{OPT})$ *and the agent* i *is*

truthful, their utility is lower bounded by

$$U_i^{\widetilde{\pi}}(\widetilde{p}_i) \geqslant -4H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} - 3\sqrt{\epsilon_{\mathcal{F}}} - 6(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ - H\left(\frac{1}{T}\sum_{t=1}^{T}\left(\sqrt{C^{\widetilde{\pi}^{(t)}}(\pi^*_{r_i+\widetilde{R}_{-i}})} + \sqrt{C^{\breve{\pi}^{(t)}_{\widetilde{R}_{-i}}}(\widetilde{\pi}^*_{-i})}\right) + \sqrt{C^{\breve{\pi}^{\mathrm{out}}_{\widetilde{R}_{-i}}}(\breve{\pi}^{\mathrm{out}}_{\widetilde{R}_{-i}})}\right) \\ \times\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right),$$

and when $(\zeta_1, \zeta_2) = (\texttt{OPT}, \texttt{PES})$*, their utility is lower bounded by*

$$U_i^{\widetilde{\pi}}(\widetilde{p}_i) \geqslant -4H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} - 2\sqrt{\epsilon_{\mathcal{F}}} - 4(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ - \left(\frac{1}{T}\sum_{t=1}^{T}\left(\sqrt{C^{\widetilde{\pi}^{(t)}}(\pi^*_{r_i+\widetilde{R}_{-i}})} + \sqrt{C^{\widehat{\pi}^{(t)}_{\widetilde{R}_{-i}}}(\widetilde{\pi}^*_{-i})} \right.\right. \\ \left.\left. + \sqrt{C^{\widehat{\pi}^{(t)}_{\widetilde{R}_{-i}}}(\widehat{\pi}^{(t)}_{\widetilde{R}_{-i}})}\right) + \sqrt{C^{\widetilde{\pi}}(\widetilde{\pi})}\right) \\ \times H\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right),$$

even when other agents are untruthful.

5. *(Asymptotic Truthfulness) Even when all the other agents are untruthful and irrespective of whether the agent i is truthful or not, when $\zeta_2 = \texttt{OPT}$, the amount of utility gained by untruthful reporting is upper bounded by*

$$U_i^{\breve{\pi}_{\tilde{R}}}(\widehat{p}_{i,\widetilde{R}}) - U_i^{\widetilde{\pi}}(\widetilde{p}_i) \leqslant 2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + 2\sqrt{\epsilon_{\mathcal{F}}} + 4(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ + H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widetilde{\pi}^{(t)}}(\pi^*_{r_i+\widetilde{R}_{-i}})} + \sqrt{C^{\breve{\pi}_{\tilde{R}}}(\breve{\pi}_{\widetilde{R}})}\right) \\ \times\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right),$$

and when $\zeta_2 = \texttt{PES}$, the amount of utility gained by untruthful reporting is upper

bounded by

$$U_i^{\breve{\pi}_{\widetilde{R}}}(\widehat{p}_{i,\widetilde{R}}) - U_i^{\widetilde{\pi}}(\widetilde{p}_i) \leqslant 2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + 2\sqrt{\epsilon_{\mathcal{F}}} + 4(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3}$$
$$+ H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widetilde{\pi}^{(t)}}(\pi^*_{r_i+\widetilde{R}_{-i}})} + \sqrt{C^{\widetilde{\pi}}(\widetilde{\pi})}\right)$$
$$\times\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).$$

Proof of Theorem 4.5.1. We will make use of the following concentration lemma.

Lemma 4.5.2. *For any* fixed $h \in [H]$, $r \in \widetilde{\mathcal{R}}$, *and any policy class* $\Pi \subset \{\mathcal{S} \to \Delta(\mathcal{A})\}^H$ *we have*

$$\Pr\Big(\exists f, f' \in \mathcal{F}, \pi \in \Pi :$$
$$\left|\mathbb{E}_{\mu_h}\left[\|f_h - \mathcal{T}^{\pi}_{h,r} f'_{h+1}\|^2\right] - \mathcal{L}_{h,r}(f_h, f'_{h+1}, \pi; \mathcal{D}) + \mathcal{L}_{h,r}(\mathcal{T}^{\pi}_{h,r} f'_{h+1}, f'_{h+1}, \pi; \mathcal{D})\right|$$
$$\geqslant \epsilon\left(\alpha + \beta + \mathbb{E}_{\mu_h}\left[\|f_h - \mathcal{T}^{\pi}_{h,r} f'_{h+1}\|^2\right]\right)\Big)$$
$$\leqslant 28\left(\mathcal{N}_\infty\left(\frac{\epsilon\beta}{140HR_{\max}}, \mathcal{F}\right)\right)^2 \mathcal{N}_{\infty,1}\left(\frac{\epsilon\beta}{140H^2R^2_{\max}}, \Pi\right)\exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{214(1+\epsilon)H^4R^4_{\max}}\right).$$

for all $\alpha, \beta > 0$, $0 < \epsilon \leqslant 1/2$.

Proof. See Section 4.5.5 for a detailed proof. □

Our proof hinges upon the occurrence of a "good event" under which the difference between the empirical Bellman error estimator and the Bellman error can be bounded. We formalize the definition of the "good event" below.

Lemma 4.5.3. *For any policy class* $\Pi \subset \{\mathcal{S} \to \Delta(\mathcal{A})\}^H$, *let the "good event"* $\mathcal{G}(\Pi)$ *be defined*

as

$$\mathcal{G}(\Pi) = \big\{\forall\, h \in [H], r \in \widetilde{\mathcal{R}}, \pi \in \Pi, f, f' \in \mathcal{F} : \\ \Big|\mathbb{E}_{\mu_h}[\|f_h - \mathcal{T}^{\pi}_{h,r} f'_{h+1}\|^2] - \mathcal{L}_{h,r}(f_h, f'_{h+1}, \pi; \mathcal{D}) + \mathcal{L}_{h,r}(\mathcal{T}^{\pi}_{h,r} f'_{h+1}, f'_{h+1}, \pi; \mathcal{D})\Big| \\ \leqslant \epsilon_{\mathrm{S}} + \frac{1}{2}\mathbb{E}_{\mu_h}[\|f_h - \mathcal{T}^{\pi}_{h,r} f'_{h+1}\|^2]\big\}, \tag{4.5.2}$$

where

$$\epsilon_{\mathrm{S}} = \frac{5136}{K} H^4 R^4_{\max} \log\left(56nH \cdot \mathcal{N}_\infty\left(\frac{19H^3R^3_{\max}}{K}, \mathcal{F}\right) \cdot \mathcal{N}_{\infty,1}\left(\frac{19H^4R^4_{\max}}{K}, \Pi\right) \Big/ \delta\right). \tag{4.5.3}$$

Then $\mathcal{G}(\Pi)$ occurs with probability at least $1 - \delta$.

Proof. See Section 4.5.5 for a detailed proof. □

On the event $\mathcal{G}(\Pi)$, the best approximations of action-value functions, defined according to Assumption 4.2.3, have small empirical Bellman error estimates.

Corollary 4.5.4. *Let Π be any policy class. Conditioned on the event $\mathcal{G}(\Pi)$, let $f_r^{\pi,*} \in \mathcal{F}$ be the best estimate of $Q^{\pi}_r(\cdot, \cdot; r)$ as defined in Assumption 4.2.3, $\pi \in \Pi$ and $r \in \widetilde{\mathcal{R}}$. Then, for all $h \in [H]$, we have*

$$\mathcal{E}_{h,r}(f_r^{\pi,*}, \pi; \mathcal{D}) \leqslant 2\epsilon_{\mathrm{S}} + 6\epsilon_{\mathcal{F}}.$$

Proof. See Section 4.5.5 for a detailed proof. □

We can also show that any function with sufficiently small empirical Bellman error estimate must also have small Bellman error conditioned on the good event.

Corollary 4.5.5. *Let $\epsilon_0 > 0$ be arbitrary and fixed. For any policy class Π, conditioned on the event $\mathcal{G}(\Pi)$, for all $h \in [H]$, reported reward $r \in \widetilde{\mathcal{R}}, \pi \in \Pi, f \in \mathcal{F}$, if $\mathcal{E}_{h,r}(f, \pi; \mathcal{D}) \leqslant \epsilon_0$,*

then

$$\mathbb{E}_{\mu_h}\left[\|f_h - \mathcal{T}^{\pi}_{h,r} f_{h+1}\|^2\right] \leqslant 2\epsilon_0 + 4\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F},\mathcal{F}}.$$

Proof. See Section 4.5.5 for a detailed proof. □

We introduce the key properties of Algorithms 10 and 11 that we will use. The following lemma states that the outputs of Algorithm 10 are approximately optimistic and pessimistic.

Lemma 4.5.6. *For any $\pi = \{\pi_h\}_{h=1}^H \in \Pi_{\mathrm{SPI}}$, reported reward $r \in \widetilde{\mathcal{R}}$, and λ, conditioned on the event $\mathcal{G}(\Pi_{\mathrm{SPI}})$, the following holds simultaneously for optimistic and pessimistic outputs of Algorithm 10:*

1. $\breve{Q}^{\pi}_{1,r}(s_0,\pi_1) + \lambda \sum_{h=1}^H \mathcal{E}_{h,r}(\breve{Q}^{\pi}_r, \pi; \mathcal{D}) \leqslant Q^{\pi}_1(s_0,\pi_1;r) + \sqrt{\epsilon_{\mathcal{F}}} + 2\lambda H\epsilon_{\mathrm{S}} + 6\lambda H\epsilon_{\mathcal{F}}$;
2. $\widehat{Q}^{\pi}_{1,r}(s_0,\pi_1) - \lambda \sum_{h=1}^H \mathcal{E}_{h,r}(\widehat{Q}^{\pi}_r, \pi; \mathcal{D}) \geqslant Q^{\pi}_1(s_0,\pi_1;r) - \sqrt{\epsilon_{\mathcal{F}}} - 2\lambda H\epsilon_{\mathrm{S}} - 6\lambda H\epsilon_{\mathcal{F}}$.

Proof. See Section 4.5.4 for a detailed proof. □

Additionally, the estimates given by Algorithm 10 are sufficiently good estimates of the ground truth action-value functions.

Lemma 4.5.7. *For any input $\pi = \{\pi_h\}_{h=1}^H \in \Pi_{\mathrm{SPI}}$, reported reward $r \in \widetilde{\mathcal{R}}$, when $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^2}\right)^{1/3}$ and the event $\mathcal{G}(\Pi_{\mathrm{SPI}})$ holds, the outputs of Algorithm 10 satisfy:*

1. $$Q^{\pi}_1(s_0,\pi_1;r) - \breve{Q}^{\pi}_{1,r}(s_0,\pi_1)$$
$$\leqslant H\sqrt{C^{\pi}(\pi)}\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right);$$
2. $$\widehat{Q}^{\pi}_{1,r}(s_0,\pi_1) - Q^{\pi}_1(s_0,\pi_1;r)$$
$$\leqslant H\sqrt{C^{\pi}(\pi)}\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).$$

Proof. See Section 4.5.4 for a detailed proof. □

Finally, we bound the difference between outputs of Algorithm 11 and the true values. More precisely, we characterize the performance of the output policy with respect to *any* comparator policy, not necessarily in the induced policy class Π_{SPI}, and bound the difference between the estimated value function and the true value function of the output policy.

Lemma 4.5.8. *For any comparator policy π (not necessarily in Π_{SPI}), any reported reward function $r \in \widetilde{\mathcal{R}}$, with η set to $\sqrt{\frac{\log|\mathcal{A}|}{2H^2R_{\max}^2T}}$ and λ set to $\left(\frac{R_{\max}}{H^2(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^2}\right)^{1/3}$ in Algorithm 11, the following claims hold conditioned on the event $\mathcal{G}(\Pi_{\mathrm{SPI}})$:*

1. *Let $\breve{Q}_{1,r}^{(t)}$ and $\breve{\pi}_r^{(t)}$ be the pessimistic value function estimate and policy estimate. Then*

$$V_1^{\pi}(s_0;r) - \frac{1}{T}\sum_{t=1}^{T}\breve{Q}_{1,r}^{(t)}(s_0,\breve{\pi}_{1,r}^{(t)}) \leqslant 2H^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}_r^{(t)}}(\pi)}\right) \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}}+12\epsilon_{\mathcal{F}}+3\epsilon_{\mathcal{F},\mathcal{F}}}\right).$$

2. *Let $\widehat{Q}_{1,r}^{(t)}$ and $\widehat{\pi}_r^{(t)}$ be the optimistic value function estimate and policy estimate. Then*

$$V_1^{\pi}(s_0;r) - \frac{1}{T}\sum_{t=1}^{T}\widehat{Q}_{1,r}^{(t)}(s_0,\widehat{\pi}_{1,r}^{(t)}) \leqslant 2H^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widehat{\pi}_r^{(t)}}(\pi)}\right) \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}}+12\epsilon_{\mathcal{F}}+3\epsilon_{\mathcal{F},\mathcal{F}}}\right).$$

Proof. See Section 4.5.4 for a detailed proof. □

We then proceed with the proof as follows. We start by bounding the suboptimality of the output policy, defined according to equation (4.2.3). We then bound the regret of each individual agent and the seller. We follow up with showing that our output asymptotically

satisfies individual rationality. Finally, we prove that our output also asymptotically satisfies truthfulness.

We use the following notation to differentiate the policies and prices learned under different truthfulness assumptions. Let $\breve{\pi} = \breve{\pi}^{\text{out}}_R$ be the policy chosen by the algorithm when all agents are truthful, let $\widetilde{\pi} = \breve{\pi}^{\text{out}}_{r_i + \widetilde{R}_{-i}}$ be the policy chosen when we only assume the agent i is truthful, and finally let $\breve{\pi}_{\widetilde{R}} = \breve{\pi}^{\text{out}}_{\widetilde{R}}$ be the policy chosen when none of the agents are truthful. Let the prices charged by the algorithm be $\{\widehat{p}_i\}_{i=1}^n$, $\{\widetilde{p}_i\}_{i=1}^n$, and $\{\widehat{p}_{i,\widetilde{R}}\}_{i=1}^n$, respectively.

Social Welfare Suboptimality Assuming all agents are truthful, we have $\widetilde{r}_i = r_i$ for all i. Let π^* be the maximizer of $V_1^\pi(s_0; R)$ over π and let $\breve{\pi}_R^{(t)}$ be the pessimistic policy iterate of Algorithm 11. We know that the social welfare suboptimality of $\breve{\pi}$ is

$$\begin{aligned}\text{SubOpt}(\breve{\pi}; s_0) &= V_1^{\pi^*}(s_0; R) - V_1^{\breve{\pi}}(s_0; R) = V_1^{\pi^*}(s_0; R) - \frac{1}{T}\sum_{t=1}^T V_1^{\breve{\pi}_R^{(t)}}(s_0; R) \\ &= \frac{1}{T}\sum_{t=1}^T \left(V_1^{\pi^*}(s_0; R) - Q_1^{\breve{\pi}^{(t)}}(s_0, \breve{\pi}_{1,R}^{(t)}; R)\right),\end{aligned}$$

as we recall that $\breve{\pi}$ is the uniform mixture of policies $\{\breve{\pi}_R^{(t)}\}_{t\in[T]}$. By Lemma 4.5.6, we have

$$\text{SubOpt}(\breve{\pi}; s_0) \leqslant \frac{1}{T}\sum_{t=1}^T \left(V_1^{\pi^*}(s_0; R) - \breve{Q}_{1,R}^{(t)}(s_0, \breve{\pi}_{1,R}^{(t)}; R)\right) + \sqrt{\epsilon_{\mathcal{F}}} + 2\lambda H \epsilon_{\text{S}} + 6\lambda H \epsilon_{\mathcal{F}}, \tag{4.5.4}$$

where $\breve{Q}_R^{(t)}$ is the pessimistic estimate of $Q(\cdot,\cdot;R)$ at the t-th iteration of Algorithm 11. When $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^2}\right)^{1/3}$ and $\eta = \sqrt{\frac{\log|\mathcal{A}|}{2H^2R_{\max}^2T}}$, we apply Lemma 4.5.8 to obtain

$$\begin{aligned}\mathrm{SubOpt}(\breve{\pi};s_0) &\leqslant 2H^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}_R^{(t)}}(\pi^*)}\right)\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}}+12\epsilon_{\mathcal{F}}+3\epsilon_{\mathcal{F},\mathcal{F}}}\right).\end{aligned}$$

Individual Suboptimality Let π^*_{-i} be the maximizer of $V^{\pi}(s_0;R_{-i})$ over π. By Algorithm 12, the price $\widehat{p}_i$ is constructed as

$$\widehat{p}_i = G^{(1)}_{-i}(s_0) - G^{(2)}_{-i}(s_0),$$

where $G^{(1)}_{-i}(s_0)$ is an estimate of $V^{\pi^*_{-i}}(s_0;R_{-i})$ obtained using Algorithm 11 and $G^{(2)}_{-i}(s_0)$ is an estimate of $V^{\breve{\pi}}(s_0;R_{-i})$ for Algorithm 12's output policy, $\breve{\pi}$. This observation will be extensively used in the remainder of the proof.

Assuming all agents are truthful, we have $\widetilde{r}_i = r_i$ for all i. Recalling the construction of $\widehat{p}_i$ in Algorithm 12 line 7 and the definition of $\{p_i^*\}_{i=1}^n$ (see (4.2.2)), we have

$$\begin{aligned}&\mathrm{SubOpt}_i(\breve{\pi},\{\widehat{p}_i\}_{i=1}^n;s_0) \\ &= V_1^{\pi^*}(s_0;r_i) + V_1^{\pi^*}(s_0;R_{-i}) - V_1^{\pi^*_{-i}}(s_0;R_{-i}) - V_1^{\breve{\pi}}(s_0;r_i) + G^{(1)}_{-i}(s_0) - G^{(2)}_{-i}(s_0) \\ &= V_1^{\pi^*}(s_0;R) - V_1^{\pi^*_{-i}}(s_0;R_{-i}) - V_1^{\breve{\pi}}(s_0;r_i) + G^{(1)}_{-i}(s_0) - G^{(2)}_{-i}(s_0) \\ &\leqslant V_1^{\pi^*}(s_0;R) - V_1^{\breve{\pi}}(s_0;R) + \left(G^{(1)}_{-i}(s_0) - V_1^{\pi^*_{-i}}(s_0;R_{-i})\right) + \left(V_1^{\breve{\pi}}(s_0;R_{-i}) - G^{(2)}_{-i}(s_0)\right) \\ &= \mathrm{SubOpt}(\breve{\pi};s_0) + \left(G^{(1)}_{-i}(s_0) - V_1^{\pi^*_{-i}}(s_0;R_{-i})\right) + \left(V_1^{\breve{\pi}}(s_0;R_{-i}) - G^{(2)}_{-i}(s_0)\right).\end{aligned}$$

We have already bounded the first term and now focus on the two latter terms.

We begin by examining $G^{(1)}_{-i}(s_0) - V_1^{\pi^*_{-i}}(s_0;R_{-i})$.

- Suppose $\zeta_1 = \texttt{OPT}$. Since π^*_{-i} maximizes $V_1^{\pi^*_{-i}}(s_0; R_{-i})$ over π, we have

$$G^{(1)}_{-i}(s_0) - V_1^{\pi^*_{-i}}(s_0; R_{-i}) \leqslant G^{(1)}_{-i}(s_0) - V_1^{\widehat{\pi}_{-i}}(s_0; R_{-i}).$$

  Recall that $\widehat{Q}^{\text{out}}_{R_{-i}}$ is the optimistic function estimate from the output of Algorithm 11, which is exactly the output of Algorithm 10 called on the policy returned by Algorithm 11, $\widehat{\pi}_{-i}$. By Lemma 4.5.7, we know that

$$\begin{aligned} &G^{(i)}_{-i}(s_0) - V_1^{\widehat{\pi}_{-i}}(s_0; R_{-i}) \\ &\qquad \leqslant H\sqrt{C^{\widehat{\pi}_{-i}}(\widehat{\pi}_{-i})}\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right). \end{aligned}$$

- Suppose $\zeta_1 = \texttt{PES}$. Since π^*_{-i} maximizes $V_1^{\pi}(s_0; R_{-i})$ over π, we have

$$G^{(1)}_{-i}(s_0) - V_1^{\pi^*_{-i}}(s_0; R_{-i}) \leqslant G^{(1)}_{-i}(s_0) - V_1^{\breve{\pi}_{-i}}(s_0; R_{-i}).$$

  Recall that $G^{(1)}_{-i}(s_0) = \breve{Q}^{\text{out}}_{1,R_{-i}}(s_0, \breve{\pi}_{1,-i})$. From Lemma 4.5.6, we know that if we let $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^2}\right)^{1/3}$, then we have

$$G^{(1)}_{-i}(s_0) - V_1^{\pi^*_{-i}}(s_0; R_{-i}) \leqslant \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3}.$$

We perform a similar analysis for $V_1^{\breve{\pi}}(s_0; R_{-i}) - G^{(2)}_{-i}(s_0)$ and when $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^2}\right)^{1/3}$.

- When $\zeta_2 = \texttt{OPT}$, $V_1^{\breve{\pi}}(s_0; R_{-i}) - G^{(2)}_{-i}(s_0) \leqslant \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3}$ by Lemma 4.5.6.

- When $\zeta_2 = \texttt{PES}$, let $\breve{Q}^{\breve{\pi}}_{R_{-i}}$ be the pessimistic output of Algorithm 10 called on $\breve{\pi}$. By

Lemma 4.5.7, we have

$$V_1^{\breve{\pi}}(s_0; R_{-i}) - G_{-i}^{(2)}(s_0) \leqslant H\sqrt{C^{\breve{\pi}}(\breve{\pi})}\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).$$

Seller Suboptimality We now turn our attention to the sellers' suboptimality. Assuming all agents are truthful, we have $\widetilde{r}_i = r_i$ for all i. Recalling the definition of $\{p_i^*\}_{i=1}^n$ in (4.2.2), we have

$$\begin{aligned}
&\mathrm{SubOpt}_0(\breve{\pi}, \{\widehat{p}_i\}_{i=1}^n; s_0) \\
&= V_1^{\pi^*}(s_0; r_0) - V_1^{\breve{\pi}}(s_0; r_0) + \sum_{i=1}^n \left(\max_{\pi' \in \Pi} V_1^{\pi'}(s_0; R_{-i}) - V_1^{\pi^*}(s_0; R_{-i})\right) - \sum_{i=1}^n \widehat{p}_i \\
&= \sum_{i=1}^n \max_{\pi' \in \Pi} V_1^{\pi'}(s_0; R_{-i}) - (n-1)V_1^{\pi^*}(s_0; R) \\
&\quad - V_1^{\breve{\pi}}(s_0; r_0) - \sum_{i=1}^n G_{-i}^{(1)}(s_0) + \sum_{i=1}^n G_{-i}^{(2)}(s_0) \\
&= \sum_{i=1}^n \left(\max_{\pi' \in \Pi} V_1^{\pi'}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0)\right) - (n-1)V_1^{\pi^*}(s_0; R) \\
&\quad - V_1^{\breve{\pi}}(s_0; r_0) + \sum_{i=1}^n G_{-i}^{(2)}(s_0) \\
&= \sum_{i=1}^n \left(V_1^{\pi^*_{-i}}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0)\right) + (n-1)(V_1^{\breve{\pi}}(s_0; R) - V_1^{\pi^*}(s_0; R)) \\
&\quad + \sum_{i=1}^n \left(G_{-i}^{(2)}(s_0) - V_1^{\breve{\pi}}(s_0, R_{-i})\right) \\
&\leqslant \sum_{i=1}^n \left(V_1^{\pi^*_{-i}}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0)\right) + \sum_{i=1}^n \left(G_{-i}^{(2)}(s_0) - V_1^{\breve{\pi}}(s_0, R_{-i})\right),
\end{aligned} \tag{4.5.5}$$

where the last inequality comes from the fact that π^* is the social welfare-maximizing policy. The two terms can be bounded similarly to bounding the agents' suboptimality. We discuss

the exact bounds for different choices of ζ_1, ζ_2 and $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^2}\right)^{1/3}, \eta = \sqrt{\frac{\log|\mathcal{A}|}{2H^2R_{\max}^2T}}$.

- When $\zeta_1 = \texttt{OPT}$, by Algorithm 12 line 7, we know that for any $i \in [n]$,

$$V_1^{\pi^*_{-i}}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0) = V_1^{\pi^*_{-i}}(s_0; R_{-i}) - \widehat{Q}^{\mathrm{out}}_{1,R_{-i}}(s_0, \widehat{\pi}_{1,-i}).$$

  By Lemma 4.5.8, we know that

$$\begin{aligned} V_1^{\pi^*_{-i}}(s_0; R_{-i}) - \frac{1}{T}\sum_{t=1}^{T} \widehat{Q}^{(t)}_{1,R_{-i}}(s_0, \widehat{\pi}^{(t)}_{1,R_{-i}}) &\leqslant 2H^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} \\ &+ H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widehat{\pi}^{(t)}_{R_{-i}}}(\pi^*_{-i})}\right) \\ &\times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}}+12\epsilon_{\mathcal{F}}+3\epsilon_{\mathcal{F},\mathcal{F}}}\right). \end{aligned}$$

  By Lemma 4.5.7 and recalling that $\widehat{\pi}_{-i}$ is the uniform mixture of $\{\widehat{\pi}^{(t)}_{R_{-i}}\}_{t\in[T]}$, we know that

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T} \widehat{Q}^{(t)}_{1,R_{-i}}(s_0, \widehat{\pi}^{(t)}_{1,R_{-i}}) - V_1^{\widehat{\pi}_{-i}}(s_0; R_{-i}) \\ &= \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{Q}^{(t)}_{1,R_{-i}}(s_0, \widehat{\pi}^{(t)}_{1,R_{-i}}) - V_1^{\widehat{\pi}^{(t)}_{R_{-i}}}(s_0; R_{-i})\right) \\ &\leqslant H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widehat{\pi}^{(t)}_{R_{-i}}}(\widehat{\pi}^{(t)}_{R_{-i}})}\right) \\ &\quad\times\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}}+12\epsilon_{\mathcal{F}}+3\epsilon_{\mathcal{F},\mathcal{F}}}\right). \end{aligned}$$

  Lastly, by Lemma 4.5.6, we also know that

$$V_1^{\widehat{\pi}_{-i}}(s_0; R_{-i}) - \widehat{Q}^{\mathrm{out}}_{1,R_{-i}}(s_0, \widehat{\pi}_{1,-i}) \leqslant \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3}.$$

Summing the three parts tells us that, for all $i \in [n]$, we have

$$\begin{aligned}
&V_1^{\pi^*_{-i}}(s_0; R_{-i}) - G^{(1)}_{-i}(s_0) \\
&\qquad = V_1^{\pi^*_{-i}}(s_0; R_{-i}) - \widehat{Q}^{\mathrm{out}}_{1,R_{-i}}(s_0, \widehat{\pi}_{1,-i}) \\
&\qquad \leqslant 2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\
&\qquad\quad + H\left(\frac{1}{T}\sum_{t=1}^{T}\left(\sqrt{C^{\widehat{\pi}^{(t)}_{R_{-i}}}(\pi^*_{-i})} + \sqrt{C^{\widehat{\pi}^{(t)}_{R_{-i}}}(\widehat{\pi}^{(t)}_{R_{-i}})}\right)\right) \\
&\qquad\qquad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right)
\end{aligned} \tag{4.5.6}$$

and

$$\begin{aligned}
&\sum_{i=1}^{n}\left(V_1^{\pi^*_{-i}}(s_0; R_{-i}) - G^{(1)}_{-i}(s_0)\right) \\
&\qquad \leqslant 2nH^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + n\sqrt{\epsilon_{\mathcal{F}}} + 2n(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\
&\qquad\quad + H\left(\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\left(\sqrt{C^{\widehat{\pi}^{(t)}_{R_{-i}}}(\pi^*_{-i})} + \sqrt{C^{\widehat{\pi}^{(t)}_{R_{-i}}}(\widehat{\pi}^{(t)}_{R_{-i}})}\right)\right) \\
&\qquad\qquad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).
\end{aligned}$$

- When $\zeta_1 = \texttt{PES}$, by Algorithm 12 we know that for any $i \in [n]$,

$$V_1^{\pi^*_{-i}}(s_0; R_{-i}) - G^{(1)}_{-i}(s_0) = V_1^{\pi^*_{-i}}(s_0; R_{-i}) - \breve{Q}^{\mathrm{out}}_{1,R_{-i}}(s_0, \breve{\pi}_{1,-i}).$$

By Lemma 4.5.8, we know that

$$V_1^{\pi^*_{-i}}(s_0; R_{-i}) - \frac{1}{T}\sum_{t=1}^{T} \breve{Q}^{(t)}_{1,R_{-i}}(s_0, \breve{\pi}^{(t)}_{1,R_{-i}}) \leqslant 2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}_{R_{-i}}}(\pi^*_{-i})}\right) \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).$$

By Lemma 4.5.6, we know that

$$\frac{1}{T}\sum_{t=1}^{T} \breve{Q}^{(t)}_{1,R_{-i}}(s_0, \breve{\pi}^{(t)}_{1,R_{-i}}) - V_1^{\breve{\pi}_{-i}}(s_0; R_{-i}) \leqslant \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3}.$$

By Lemma 4.5.7, we further know that

$$V_1^{\breve{\pi}_{-i}}(s_0; R_{-i}) - \breve{Q}^{\mathrm{out}}_{1,R_{-i}}(s_0, \breve{\pi}_{1,-i}) \leqslant H\sqrt{C^{\breve{\pi}_{-i}}(\breve{\pi}_{-i})}\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).$$

Summing the three parts together tells us that, for all $i \in [n]$ and any $C \geqslant 1$, we have

$$\begin{aligned} V_1^{\pi^*_{-i}}(s_0; R_{-i}) - G^{(1)}_{-i}(s_0) &= V_1^{\pi^*_{-i}}(s_0; R_{-i}) - \breve{Q}^{\mathrm{out}}_{1,R_{-i}}(s_0, \breve{\pi}_{1,-i}) \\ &\leqslant 2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H\left(\sqrt{C^{\breve{\pi}_{-i}}(\breve{\pi}_{-i})} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}_{R_{-i}}}(\pi^*_{-i})}\right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right) \end{aligned} \tag{4.5.7}$$

and

$$\begin{aligned}
&\sum_{i=1}^{n}\left(V_1^{\pi^*_{-i}}(s_0;R_{-i}) - G^{(1)}_{-i}(s_0)\right)\\
&\quad\leqslant 2nH^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + n\sqrt{\epsilon_{\mathcal{F}}} + 2n(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3}\\
&\qquad + H\left(\sum_{i=1}^{n}\sqrt{C^{\breve{\pi}}{}_{-i}(\breve{\pi}_{-i})} + \sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}_{R_{-i}}}(\pi^*_{-i})}\right)\\
&\qquad\quad\times\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}}+12\epsilon_{\mathcal{F}}+3\epsilon_{\mathcal{F},\mathcal{F}}}\right).
\end{aligned}$$

- When $\zeta_2 = \texttt{OPT}$, for all $i \in [n]$, let $\breve{Q}^{\breve{\pi}}_{R_{-i}}$ be the pessimistic estimate of $Q^{\breve{\pi}}(\cdot,\cdot;R_{-i})$ returned by Algorithm 10. By Lemma 4.5.7, we know

$$\begin{aligned}
&\sum_{i=1}^{n}\left(G^{(2)}_{-i}(s_0) - V_1^{\breve{\pi}}(s_0,R_{-i})\right)\\
&\qquad\leqslant nH\sqrt{C^{\breve{\pi}}(\breve{\pi})}\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}}+12\epsilon_{\mathcal{F}}+3\epsilon_{\mathcal{F},\mathcal{F}}}\right).
\end{aligned}$$

- When $\zeta_2 = \texttt{PES}$, by Lemma 4.5.6

$$\sum_{i=1}^{n}\left(G^{(2)}_{-i}(s_0) - V_1^{\breve{\pi}}(s_0,R_{-i})\right) \leqslant n\sqrt{\epsilon_{\mathcal{F}}} + 2n(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3}.$$

Plugging in the bound for $\mathrm{SubOpt}(\breve{\pi};s_0)$ completes the proof.

Individual Rationality We show that the utility of any agent i is bounded below. First, assume for convenience that all other agents are truthful and report their true $r_{i',h}$ for $i' \in [n]\backslash i$. Recall that for any price p_i, the agents' expected utility under the chosen policy $\breve{\pi}$ can be written as

$$\mathbb{E}_{d_{\breve{\pi}}}[u_i] = V_1^{\breve{\pi}}(s_0;r_i) - p_i.$$

According to Algorithm 12, we have

$$\begin{aligned}
\mathbb{E}_{\breve{\pi}}[u_i] &= V_1^{\breve{\pi}}(s_0; r_i) - G_{-i}^{(1)}(s_0) + G_{-i}^{(2)}(s_0) \\
&= V_1^{\breve{\pi}}(s_0; r_i) + G_{-i}^{(2)}(s_0) - V^{\pi^*_{-i}}(s_0; R_{-i}) + V^{\pi^*_{-i}}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0) \\
&= (V^{\pi^*}(s_0; R) - V^{\pi^*_{-i}}(s_0; R_{-i})) + V^{\breve{\pi}}(s_0; r_i) + G_{-i}^{(2)}(s_0) - V^{\pi^*}(s_0; R) \\
&\quad + V^{\pi^*_{-i}}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0) \\
&\geqslant V^{\breve{\pi}}(s_0; r_i) + G_{-i}^{(2)}(s_0) - V^{\pi^*}(s_0; R) + V^{\pi^*_{-i}}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0) \\
&= G_{-i}^{(2)}(s_0) - V^{\breve{\pi}}(s_0; R_{-i}) + V^{\breve{\pi}}(s_0; R) - V^{\pi^*}(s_0; R) + V^{\pi^*_{-i}}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0),
\end{aligned} \tag{4.5.8}$$

where the inequality comes from the fact that

$$(V^{\pi^*}(s_0; R) - V^{\pi^*_{-i}}(s_0; R_{-i})) \geqslant (V^{\pi^*_{-i}}(s_0; R) - V^{\pi^*_{-i}}(s_0; R_{-i})) = V^{\pi^*_{-i}}(s_0; r_i)) \geqslant 0,$$

as $r_{i,h} \in [0, 1]$ for all i, h. We already know the lower bounds for $V^{\pi^*_{-i}}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0)$ and $G_{-i}^{(2)}(s_0) - V^{\breve{\pi}}(s_0; R_{-i})$, respectively, when bounding the individual suboptimalities for the agents. Also note that $V^{\breve{\pi}}(s_{0;R}) - V^{\pi^*}(s_0; R) = -\text{SubOpt}(\breve{\pi}; s_0)$ has been bounded when bounding social welfare suboptimality.

Similar to the previous sections, we now discuss the bounds for the different terms under difference choices of ζ_1, ζ_2.

- When $\zeta_1 = \texttt{OPT}$, by equation (4.5.6) we know that

$$\begin{aligned} G_{-i}^{(1)}(s_0) - V_1^{\pi^*_{-i}}(s_0; R_{-i}) \geqslant & -2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} - \sqrt{\epsilon_{\mathcal{F}}} \\ & - 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ & - H\left(\frac{1}{T}\sum_{t=1}^{T}\left(\sqrt{C^{\widehat{\widehat{\pi}}^{(t)}_{R_{-i}}}(\pi^*_{-i})} + \sqrt{C^{\widehat{\widehat{\pi}}^{(t)}_{R_{-i}}}(\widehat{\widehat{\pi}}^{(t)}_{R_{-i}})}\right)\right) \\ & \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right). \end{aligned}$$

- When $\zeta_1 = \texttt{PES}$, by equation (4.5.7) we know that

$$\begin{aligned} G_{-i}^{(1)}(s_0) - V_1^{\pi^*_{-i}}(s_0; R_{-i}) \geqslant & -2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} - \sqrt{\epsilon_{\mathcal{F}}} \\ & - 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ & - H\left(\sqrt{C^{\breve{\pi}_{-i}}(\breve{\pi}_{-i})} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\breve{\pi}}^{(t)}_{R_{-i}}}(\pi^*_{-i})}\right) \\ & \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right). \end{aligned}$$

- When $\zeta_2 = \texttt{OPT}$, by Lemma 4.5.6, we know that

$$G_{-i}^{(2)}(s_0) - V^{\breve{\pi}}(s_0; R_{-i}) \geqslant -\sqrt{\epsilon_{\mathcal{F}}} - 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3}.$$

- When $\zeta_2 = \texttt{PES}$, by Lemma 4.5.7

$$\begin{aligned} & G_{-i}^{(2)}(s_0) - V_1^{\breve{\pi}}(s_0; R_{-i}) \\ & \quad \geqslant -H\sqrt{C^{\breve{\pi}}(\breve{\pi})}\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right). \end{aligned}$$

We now argue that our analysis holds even when the other agents are not truthful. Recall that $\widetilde{\pi}$ is the output policy selected by Algorithm 12 when other agents report $\widetilde{r}_{i'}$ and the

agent i reports truthfully. Observe that here the decomposition in equation (4.5.8) can be written as

$$\begin{aligned}\mathbb{E}_{\widetilde{\pi}}[u_i] \geqslant \widetilde{G}^{(2)}_{-i}(s_0) - V^{\widetilde{\pi}}(s_0; \widetilde{R}_{-i}) + V^{\widetilde{\pi}}(s_0; r_i + \widetilde{R}_{-i}) - V^{\pi^*_{r_i+\widetilde{R}_{-i}}}(s_0; r_i + \widetilde{R}_{-i}) \\ + V^{\widetilde{\pi}^*_{-i}}(s_0; \widetilde{R}_{-i}) - \widetilde{G}^{(1)}_{-i}(s_0),\end{aligned}$$

where we recall that $\widetilde{R}_{-i} = \sum_{i' \neq i} \widetilde{r}_{i'}$, and $\pi^*_{r_i+\widetilde{R}_{-i}}$ and $\widetilde{\pi}^*_{-i}$ maximize $V_1^{\pi}(s_0; r_i + \widetilde{R}_{-i})$ and $V_1^{\pi}(s_0; \widetilde{R}_{-i})$ over π, respectively. We also let $\widetilde{G}^{(1)}_{-i}, \widetilde{G}^{(2)}_{-i}$ be the estimates used in Algorithm 12 line 7 when other agents are reporting untruthfully.

Similar to the previous sections, we bound different terms under difference choices of ζ_1, ζ_2.

- When $\zeta_1 = \texttt{OPT}$, similar to equation (4.5.6), we have

$$\begin{aligned}\widetilde{G}^{(1)}_{-i}(s_0) - V_1^{\widetilde{\pi}^*_{-i}}(s_0; \widetilde{R}_{-i}) \geqslant & -2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} - \sqrt{\epsilon_{\mathcal{F}}} \\ & - 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ & - H\left(\frac{1}{T}\sum_{t=1}^{T}\left(\sqrt{C^{\widehat{\pi}^{(t)}_{\widetilde{R}_{-i}}}(\widetilde{\pi}^*_{-i})} + \sqrt{C^{\widehat{\pi}^{(t)}_{\widetilde{R}_{-i}}}(\widehat{\pi}^{(t)}_{\widetilde{R}_{-i}})}\right)\right) \\ & \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).\end{aligned}$$

- When $\zeta_1 = \texttt{PES}$, similar to equation (4.5.7), we have

$$\begin{aligned}\widetilde{G}^{(1)}_{-i}(s_0) - V_1^{\widetilde{\pi}^*_{-i}}(s_0; \widetilde{R}_{-i}) \geqslant & -2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} - \sqrt{\epsilon_{\mathcal{F}}} \\ & - 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ & - H\left(\sqrt{C^{\breve{\pi}^{\mathrm{out}}_{\widetilde{R}_{-i}}}(\breve{\pi}^{\mathrm{out}}_{\widetilde{R}_{-i}})} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}^{(t)}_{\widetilde{R}_{-i}}}(\widetilde{\pi}^*_{-i})}\right) \\ & \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).\end{aligned}$$

- When $\zeta_2 = \texttt{OPT}$, by Lemma 4.5.6, we know

$$\widetilde{G}^{(2)}_{-i}(s_0) - V^{\widetilde{\pi}}(s_0; \widetilde{R}_{-i}) \geqslant -\sqrt{\epsilon_{\mathcal{F}}} - 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3}.$$

- When $\zeta_2 = \texttt{PES}$, by Lemma 4.5.7

$$\begin{aligned}
&\widetilde{G}^{(2)}_{-i}(s_0) - V_1^{\widetilde{\pi}}(s_0; \widetilde{R}_{-i}) \\
&\quad \geqslant -H\sqrt{C^{\widetilde{\pi}}(\widetilde{\pi})}\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right),
\end{aligned}$$

  where $\widetilde{\pi}$ is the policy that the seller chooses when agent i reports truthfully and the other agents do not.

We finally focus on lower bounding $V^{\widetilde{\pi}}(s_0; r_i + \widetilde{R}_{-i}) - V^{\pi^*_{r_i+\widetilde{R}_{-i}}}(s_0; r_i + \widetilde{R}_{-i})$. Since $\widetilde{\pi}$ is the uniform mixture of $\{\widetilde{\pi}^{(t)}\}_{t\in[T]}$, we have

$$\begin{aligned}
&V_1^{\pi^*_{r_i+\widetilde{R}_{-i}}}(s_0; r_i + \widetilde{R}_{-i}) - V_1^{\widetilde{\pi}}(s_0; r_i + \widetilde{R}_{-i}) \\
&\quad = \frac{1}{T}\sum_{t=1}^{T}\left(V_1^{\pi^*_{r_i+\widetilde{R}_{-i}}}(s_0; r_i + \widetilde{R}_{-i}) - V_1^{\widetilde{\pi}^{(t)}}(s_0; r_i + \widetilde{R}_{-i})\right) \\
&\quad \leqslant \frac{1}{T}\sum_{t=1}^{T}\left(V_1^{\pi^*_{r_i+\widetilde{R}_{-i}}}(s_0; r_i + \widetilde{R}_{-i}) - \breve{Q}^{(t)}_{1,r_i+\widetilde{R}_{-i}}(s_0, \widetilde{\pi}_1^{(t)})\right) \\
&\qquad + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3}
\end{aligned}$$

by Lemma 4.5.6. By Lemma 4.5.8, we know that

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\left(V_1^{\pi^*_{r_i+\widetilde{R}_{-i}}}(s_0; r_i + \widetilde{R}_{-i}) - \breve{Q}^{(t)}_{1,r_i+\widetilde{R}_{-i}}(s_0, \widetilde{\pi}_1^{(t)})\right) &\leqslant 2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} \\
&\quad + H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widetilde{\pi}^{(t)}}(\pi^*_{r_i+\widetilde{R}_{-i}})}\right) \\
&\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).
\end{aligned}$$

Therefore, we have

$$
\begin{aligned}
&V_1^{\pi^*_{r_i+\widetilde{R}_{-i}}}(s_0; r_i + \widetilde{R}_{-i}) - V_1^{\widetilde{\pi}}(s_0; r_i + \widetilde{R}_{-i}) \\
&\quad \leqslant 2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\
&\qquad + H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widetilde{\pi}(t)}(\pi^*_{r_i+\widetilde{R}_{-i}})}\right) \\
&\qquad\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).
\end{aligned}
\tag{4.5.9}
$$

Flipping the signs yields the final bound.

Truthfulness Similar to above and let $\widetilde{r}_{i'}$ be the potentially untruthful reward functions reported by other agents and let $\widetilde{r}_i$ be the untruthful reward function that the agent i may report. Furthermore, let $\widetilde{R}_{-i} = \sum_{i' \neq i} \widetilde{r}_{i'}$ and $\widetilde{R} = \sum_{i=1}^{n} \widetilde{r}_i$.

Let $\widetilde{\pi}$ be the policy chosen by the seller when the agent i is truthful and other agents are possibly non-truthful and $\breve{\pi}_{\widetilde{R}}$ the policy chosen by Algorithm 12 when both the agent i and other agents are non-truthful. The agents' expected utilities for the two cases are

$$
\begin{aligned}
\mathbb{E}_{\widetilde{\pi}}[u_i] &= V_1^{\widetilde{\pi}}(s_0; r_i) + \widetilde{G}_{-i}^{(2)}(s_0) - \widetilde{G}_{-i}^{(1)}(s_0), \\
\mathbb{E}_{d_{\breve{\pi}_{\widetilde{R}}}}[u_i] &= V_1^{\breve{\pi}_{\widetilde{R}}}(s_0; r_i) + \widetilde{G}_{-i}^{(2),\prime}(s_0) - \widetilde{G}_{-i}^{(1),\prime}(s_0),
\end{aligned}
$$

where $\widetilde{G}_{-i}^{(2)}(s_0)$ estimates $V^{\widetilde{\pi}}(s_0; \widetilde{R}_{-i})$ and $\widetilde{G}_{-i}^{(2),\prime}(s_0)$ estimates $V^{\breve{\pi}_{\widetilde{R}}}(s_0; \widetilde{R}_{-i})$.

Observe that both $\widetilde{G}_{-i}^{(1)}(s_0)$ and $\widetilde{G}_{-i}^{(1),\prime}(s_0)$ approximate $V_1^{\widetilde{\pi}^*_{-i}}(s_0; \widetilde{R}_{-i})$ using the same algorithm, Algorithm 11. As the algorithm itself does not contain randomness and $\widetilde{G}_{-i}^{(1)}(s_0)$ and $\widetilde{G}_{-i}^{(1),\prime}(s_0)$ are constructed using the same parameters, the two terms must be equal.

Then we have

$$\begin{aligned}
\mathbb{E}_{\breve{\pi}_{\tilde{R}}}[u_i] - \mathbb{E}_{\widetilde{\pi}}[u_i] &= V_1^{\breve{\pi}_{\tilde{R}}}(s_0; r_i) + \widetilde{G}_{-i}^{(2),\prime}(s_0) - \left(V_1^{\widetilde{\pi}}(s_0; r_i) + \widetilde{G}_{-i}^{(2)}(s_0)\right) \\
&= V_1^{\breve{\pi}_{\tilde{R}}}(s_0; r_i + \widetilde{R}_{-i}) + \widetilde{G}_{-i}^{(2),\prime}(s_0) - V_1^{\breve{\pi}_{\tilde{R}}}(s_0; \widetilde{R}_{-i}) \\
&\quad - \left(V_1^{\widetilde{\pi}}(s_0; r_i + \widetilde{R}_{-i}) + \widetilde{G}_{-i}^{(2)}(s_0) - V_1^{\widetilde{\pi}}(s_0; \widetilde{R}_{-i})\right) \\
&= V_1^{\breve{\pi}_{\tilde{R}}}(s_0; r_i + \widetilde{R}_{-i}) - V_1^{\pi^*_{r_i+\tilde{R}_{-i}}}(s_0; r_i + \widetilde{R}_{-i}) + \widetilde{G}_{-i}^{(2),\prime}(s_0) - V_1^{\breve{\pi}_{\tilde{R}}}(s_0; \widetilde{R}_{-i}) \\
&\quad + V_1^{\pi^*_{r_i+\tilde{R}_{-i}}}(s_0; r_i + \widetilde{R}_{-i}) - V_1^{\widetilde{\pi}}(s_0; r_i + \widetilde{R}_{-i}) + V_1^{\widetilde{\pi}}(s_0; \widetilde{R}_{-i}) - \widetilde{G}_{-i}^{(2)}(s_0),
\end{aligned}$$

where we recall that $\pi^*_{r_i+\widetilde{R}_{-i}}$ is the maximizer of $V_1^{\pi}(s_0; r_i + \widetilde{R}_{-i})$ over π (the social welfare maximizing policy when agent i reports truthfully). We then know that

$$V_1^{\breve{\pi}_{\tilde{R}}}(s_0; r_i + \widetilde{R}_{-i}) - V_1^{\pi^*_{r_i+\tilde{R}_{-i}}}(s_0; r_i + \widetilde{R}_{-i}) \leqslant 0$$

and

$$\begin{aligned}
&\mathbb{E}_{\breve{\pi}_{\tilde{R}}}[u_i] - \mathbb{E}_{\widetilde{\pi}}[u_i] \\
&\quad \leqslant \left(\widetilde{G}_{-i}^{(2),\prime}(s_0) - V_1^{\breve{\pi}_{\tilde{R}}}(s_0; \widetilde{R}_{-i})\right) + \left(V_1^{\pi^*_{r_i+\tilde{R}_{-i}}}(s_0; r_i + \widetilde{R}_{-i}) - V_1^{\widetilde{\pi}}(s_0; r_i + \widetilde{R}_{-i})\right) \\
&\qquad + \left(V_1^{\widetilde{\pi}}(s_0; \widetilde{R}_{-i}) - \widetilde{G}_{-i}^{(2)}(s_0)\right).
\end{aligned}$$

Let us focus on the middle term first. By (4.5.9), we have

$$
\begin{aligned}
&V_1^{\pi^*_{r_i+\widetilde{R}_{-i}}}(s_0; r_i + \widetilde{R}_{-i}) - V_1^{\widetilde{\pi}}(s_0; r_i + \widetilde{R}_{-i}) \\
&\quad \leqslant 2H^2 R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\
&\qquad + H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\widetilde{\pi}^{(t)}}(\pi^*_{r_i+\widetilde{R}_{-i}})}\right) \\
&\qquad\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).
\end{aligned}
$$

We state the results for different values of ζ_2 as the bound no longer depends on ζ_1.

- When $\zeta_2 = \texttt{OPT}$, by Lemma 4.5.6, we have

$$
V_1^{\widetilde{\pi}}(s_0; \widetilde{R}_{-i}) - \widetilde{G}^{(2)}_{-i}(s_0) \leqslant \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3},
$$

  and by Lemma 4.5.7,

$$
\begin{aligned}
&\widetilde{G}^{(2),\prime}_{-i}(s_0) - V_1^{\breve{\pi}_{\widetilde{R}}}(s_0; \widetilde{R}_{-i}) \\
&\qquad \leqslant H\sqrt{C^{\breve{\pi}_{\widetilde{R}}}(\breve{\pi}_{\widetilde{R}})}\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).
\end{aligned}
$$

- When $\zeta_2 = \texttt{PES}$, by Lemma 4.5.7,

$$
\begin{aligned}
&V_1^{\widetilde{\pi}}(s_0; \widetilde{R}_{-i}) - \widetilde{G}^{(2)}_{-i}(s_0) \\
&\qquad \leqslant H\sqrt{C^{\widetilde{\pi}}(\widetilde{\pi})}\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right),
\end{aligned}
$$

  and by Lemma 4.5.6,

$$
\widetilde{G}^{(2),\prime}_{-i}(s_0) - V_1^{\breve{\pi}_{\widetilde{R}}}(s_0; \widetilde{R}_{-i}) \leqslant \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}} + 3\epsilon_{\mathcal{F}})^{1/3}.
$$

Combining the terms completes the proof. □

4.5.4 Supporting Lemmas

In this section, we provide detailed proofs of supporting lemmas used in Section 4.5.3.

Proofs for Algorithm 10

Previous work has shown that the estimate of the value function f^π is the exact value function of an induced MDP that shares the same state space, action space, and transition kernel as $\mathcal{M}$, only with slightly perturbed reward functions [Cai et al., 2020, Uehara and Sun, 2021, Xie et al., 2021, Zanette et al., 2021]. More precisely, let r be the input reward for Algorithm 10, π the input policy, and f^π the output. Let $\mathcal{M}_{f^\pi}$ be the induced MDP. We formally state the result below.

Lemma 4.5.9. *For any input policy π (not necessarily in Π_{SPI}) and input reward function r, Algorithm 10 returns a function f^π such that f^π is the Q-function of the policy π under the induced MDP $\mathcal{M}_{f^\pi}$, given by*

$$\mathcal{M}_{f^\pi} = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r_{f^\pi}), \tag{4.5.10}$$

where $r_{f^\pi,h} = r_h + f_h^\pi - \mathcal{T}_{h,r}^\pi f_{h+1}^\pi$. In other words, $f^\pi(\cdot,\cdot) = Q^\pi(\cdot,\cdot; r_{f^\pi})$.

Proof. See Section C.1 in Zanette et al. [2021] for a detailed proof. □

We immediately have the following corollary.

Corollary 4.5.10. *Let f^π be any one of the two functions returned by Algorithm 10 for any input policy π (not necessarily in Π_{SPI}) and any input reward function r. Then, for all*

$h \in [H]$, we have

$$\left|f_h^\pi(s,a) - Q_h^\pi(s,a;r)\right| \leqslant \sum_{h'=h}^{H} \mathbb{E}_{(S_{h'},A_{h'})\sim\pi|(s,a)}\left[\left|f_h^\pi - \mathcal{T}_{h,r}^\pi f_{h+1}^\pi\right|\right].$$

Proof. By definition of the Q-function, we have

$$\begin{aligned} f_h^\pi(s,a) - Q_h^\pi(s,a;r) &= Q_h^\pi(s,a;r_{f^\pi}) - Q_h^\pi(s,a;r) \\ &= \sum_{h'=h}^{H} \mathbb{E}_{(S_{h'},A_{h'})\sim\pi|(s,a)}[r_h(S_{h'},A_{h'}) - r_{f^\pi,h}(S_{h'},A_{h'})]. \end{aligned}$$

Recalling the definition of r_{f^π} in equation (4.5.10) and using Jensen's inequality concludes the proof. □

We proceed to show that Algorithm 10 is approximately optimistic/pessimistic and bounding the estimation error of its outputs. We begin with the proof of Lemma 4.5.6.

Proof of Lemma 4.5.6. We start by upper bounding two auxiliary terms. Let $f_r^{\pi,*} \in \mathcal{F}$ be the best approximation of $Q^\pi(\cdot,\cdot;r)$, as defined in Assumption 4.2.3. By Jensen's inequality, we have

$$|f_{1,r}^{\pi,*}(s_0,\pi_1) - Q_1^\pi(s_0,\pi_1;r)| \leqslant \mathbb{E}_{a\sim\pi_1(\cdot|s_0)}[|f_{1,r}^{\pi,*}(s_0,\pi_1) - Q_1^\pi(s_0,\pi_1;r)|] \leqslant \sqrt{\epsilon_{\mathcal{F}}}.$$

Additionally, using Lemma 4.5.4 we know that, conditioned on the event $\mathcal{G}(\Pi_{\mathrm{SPI}})$, for all $h \in [H]$ we have $\mathcal{E}_{h,r}(f_r^{\pi,*},\pi;\mathcal{D}) \leqslant 2\epsilon_{\mathrm{S}} + 6\epsilon_{\mathcal{F}}$.

We then consider $\check{Q}_r^\pi$. By (4.3.2), we know that

$$\begin{aligned}
\check{Q}_{1,r}^\pi(s_0,\pi) + \lambda \sum_{h=1}^{H} \mathcal{E}_{h,r}(\check{Q}_r^\pi, \pi; \mathcal{D}) &\leqslant f_{1,r}^{\pi,*}(s_0,\pi) + \lambda \sum_{h=1}^{H} \mathcal{E}_{h,r}(f_r^{\pi,*}, \pi; \mathcal{D}) \\
&\leqslant Q_1^\pi(s_0,\pi;r) + |f_{1,r}^{\pi,*}(s_0,\pi_1) - Q_1^\pi(s_0,\pi_1;r)| + 2\lambda H \epsilon_{\mathrm{S}} + 6\lambda H \epsilon_{\mathcal{F}} \\
&\leqslant Q_1^\pi(s_0,\pi_1;r) + \sqrt{\epsilon_{\mathcal{F}}} + 2\lambda H \epsilon_{\mathrm{S}} + 6\lambda H \epsilon_{\mathcal{F}}.
\end{aligned}$$

Similarly for $\widehat{Q}_r^\pi$, by (4.3.2), we have

$$\begin{aligned}
\widehat{Q}_{1,r}^\pi(s_0,\pi) - \lambda \sum_{h=1}^{H} \mathcal{E}_{h,r}(\widehat{Q}_r^\pi, \pi; \mathcal{D}) &\geqslant f_{1,r}^{\pi,*}(s_0,\pi) - \lambda \sum_{h=1}^{H} \mathcal{E}_{h,r}(f_r^{\pi,*}, \pi; \mathcal{D}) \\
&\geqslant Q_1^\pi(s_0,\pi;r) - |f_{1,r}^{\pi,*}(s_0,\pi_1) - Q_1^\pi(s_0,\pi_1;r)| - 2\lambda H \epsilon_{\mathrm{S}} - 6\lambda H \epsilon_{\mathcal{F}} \\
&\geqslant Q_1^\pi(s_0,\pi_1;r) - \sqrt{\epsilon_{\mathcal{F}}} - 2\lambda H \epsilon_{\mathrm{S}} - 6\lambda H \epsilon_{\mathcal{F}},
\end{aligned}$$

thus completing the proof. □

We prove that the action-value functions returned by Algorithm 10 are sufficiently good estimates.

Proof of Lemma 4.5.7. By Corollary 4.5.10, we have

$$\begin{aligned}
\widehat{Q}_{1,r}^\pi(s_0,\pi_1) - Q_1^\pi(s_0,\pi_1;r) &\leqslant \left| \sum_{h=1}^{H} \mathbb{E}_\pi \left[\widehat{Q}_{h,r}^\pi - \mathcal{T}_{h,r}^\pi \widehat{Q}_{h+1,r}^\pi \right] \right|, \\
Q_1^\pi(s_0,\pi_1;r) - \check{Q}_{1,r}^\pi(s_0,\pi_1) &\leqslant \left| \sum_{h=1}^{H} \mathbb{E}_\pi \left[\check{Q}_{h,r}^\pi - \mathcal{T}_{h,r}^\pi \check{Q}_{h+1,r}^\pi \right] \right|.
\end{aligned}$$

Since the differences share similar forms, we can without loss of generality only consider $\widehat{Q}_r^\pi$.

Recall the definition of $C^\pi(\nu)$, given in Definition 4.2.5. We have

$$\begin{aligned}\left|\sum_{h=1}^{H}\mathbb{E}_\pi\left[\breve{Q}^\pi_{h,r}-\mathcal{T}^\pi_{h,r}\breve{Q}^\pi_{h+1,r}\right]\right| &\leqslant \sum_{h=1}^{H}\mathbb{E}_\pi\left[\left\|\breve{Q}^\pi_{h,r}-\mathcal{T}^\pi_{h,r}\breve{Q}^\pi_{h+1,r}\right\|\right]\\ &\leqslant \sqrt{C^\pi(\pi)}\sum_{h=1}^{H}\mathbb{E}_{\mu_h}\left[\left\|\breve{Q}^\pi_{h,r}-\mathcal{T}^\pi_{h,r}\breve{Q}^\pi_{h+1,r}\right\|\right],\end{aligned} \tag{4.5.11}$$

where the first inequality is by Cauchy-Schwarz, the second inequality by the definition of $C^\pi(\pi)$, which is the shorthand notation for $C^\pi(d_\pi)$. Similar to the proof of Lemma 4.5.6, let $f_r^{\pi,*}$ be the best approximation of $Q^\pi(\cdot,\cdot;r)$ as defined in Assumption 4.2.3. Then

$$\lambda\sum_{h=1}^{H}\mathcal{E}_{h,r}(\breve{Q}^\pi_r,\pi;\mathcal{D})\leqslant f^{\pi,*}_{1,r}(s_0,\pi_1)-\breve{Q}^\pi_{1,r}(s_0,\pi_1)+2\lambda H\epsilon_{\mathrm{S}}+6\lambda H\epsilon_{\mathcal{F}}.$$

Since $f_r^{\pi,*},\breve{Q}^\pi_{1,r}\in\mathcal{F}$, we have $f_r^{\pi,*},\breve{Q}^\pi_{1,r}\in[-HR_{\max},HR_{\max}]$ and thus

$$\sum_{h=1}^{H}\mathcal{E}_{h,r}(\breve{Q}^\pi_r,\pi;\mathcal{D})\leqslant\frac{2HR_{\max}}{\lambda}+2H\epsilon_{\mathrm{S}}+6H\epsilon_{\mathcal{F}}.$$

By Corollary 4.5.5, conditioned on $\mathcal{G}(\Pi_{\mathrm{SPI}})$, we have

$$\begin{aligned}\sum_{h=1}^{H}\mathbb{E}_{\mu_h}\left[\|\breve{Q}^\pi_{h,r}-\mathcal{T}^\pi_{h,r}\breve{Q}^\pi_{h+1,r}\|^2\right]&\leqslant 2\sum_{h=1}^{H}\mathcal{E}_{h,r}(\breve{Q}^\pi_r,\pi;\mathcal{D})+4H\epsilon_{\mathrm{S}}+3H\epsilon_{\mathcal{F},\mathcal{F}}\\ &\leqslant\frac{4HR_{\max}}{\lambda}+8H\epsilon_{\mathrm{S}}+12H\epsilon_{\mathcal{F}}+3H\epsilon_{\mathcal{F},\mathcal{F}}.\end{aligned}$$

Plugging the bound back into (4.5.11) and applying Cauchy-Schwarz inequality gives us

$$\begin{aligned}\left|\sum_{h=1}^{H}\mathbb{E}_\pi\left[\breve{Q}^\pi_{h,r}-\mathcal{T}^\pi_{h,r}\breve{Q}^\pi_{h+1,r}\right]\right|&\leqslant\sqrt{H}\sqrt{C^\pi(\pi)}\sqrt{\frac{4HR_{\max}}{\lambda}+8H\epsilon_{\mathrm{S}}+12H\epsilon_{\mathcal{F}}+3H\epsilon_{\mathcal{F},\mathcal{F}}}\\ &=H\sqrt{C^\pi(\pi)}\sqrt{\frac{4R_{\max}}{\lambda}+8\epsilon_{\mathrm{S}}+12\epsilon_{\mathcal{F}}+3\epsilon_{\mathcal{F},\mathcal{F}}}.\end{aligned}$$

Setting $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^2}\right)^{1/3}$ and using $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ for $a, b \in \mathbb{R}_{\geqslant 0}$ completes the proof. □

Proofs for Algorithm 11

We now turn to analyzing the policies selected in Algorithm 11. In particular, we focus on the mirror descent-style updates given in (4.3.3) and (4.3.4). We start by defining an abstract version of the procedure in Algorithm 11.

Definition 4.5.11. *Consider the following procedure. For any $t \in [T]$:*

1. *Let $f^{(t)} \in \mathcal{F}$ be an arbitrary function in the function class.*
2. *Let $\pi_h^{(t+1)}(a|s) \propto \pi_h^{(t)}(a|s) \exp\left(\eta f_h^{(t)}(s,a)\right)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $h \in [H]$.*

Recall that $\mathbb{E}_{a\in\mathcal{A}}\left[\log \pi_h(a|s)\right] = \sum_{a\in\mathcal{A}} \pi_h(a|s) \log \pi_h(a|s)$ for all π, h, and s. We continue with a standard analysis of the regret of actor-critic algorithms.

Lemma 4.5.12. *For any π (not necessarily in Π_{SPI}), for all $h \in [H]$ and $s \in \mathcal{S}$, setting $\eta = \sqrt{\frac{\log|\mathcal{A}|}{2H^2R_{\max}^2T}}$ in the procedure defined in 4.5.11 ensures that*

$$\sum_{t=1}^{T} \langle \pi_h(\cdot|s) - \pi_h^{(t)}(\cdot|s), f_h^{(t)}(s,\cdot)\rangle \leqslant 2HR_{\max}\sqrt{2T\log|\mathcal{A}|}.$$

Proof. By a direct application of Lemma C.3 of Xie et al. [2021], we know that even for policies not in Π_{SPI} (as we are effectively performing mirror descent over the probability simplex with the KL penalty) we have

$$\begin{aligned}
&\sum_{t=1}^{T} \langle \pi_h(\cdot|s) - \pi_h^{(t)}(\cdot|s), f_h^{(t)}(s,\cdot)\rangle \\
&\qquad \leqslant \sum_{t=1}^{T} \langle \pi_h^{(t+1)} - \pi_h^{(t)}(\cdot|s), f_h^{(t)}(s,\cdot)\rangle - \frac{1}{\eta}\mathbb{E}_{a\sim\pi_h^{(1)}}\left[\log \pi_h^{(1)}(a|s)\right],
\end{aligned}$$

where η is the stepsize. From the proof of Lemma C.4 in Xie et al. [2021], we further note that for any $\pi \in \pi$, $h \in [H]$, $s \in \mathcal{S}$, and $t \in [T]$ we have

$$\langle \pi_h(\cdot|s) - \pi_h^{(t)}(\cdot|s), f_h^{(t)}(s,\cdot)\rangle \leqslant \|f_h^{(t)}(s,\cdot)\|_\infty \sqrt{2\eta \langle \pi_h(\cdot|s) - \pi_h^{(t)}(\cdot|s), f_h^{(t)}(s,\cdot)\rangle}.$$

Because f_h are bounded by $HR_{\max}$, $\langle \pi_h(\cdot|s) - \pi_h^{(t)}(\cdot|s), f_h^{(t)}(s,\cdot)\rangle \leqslant 2\eta H^2 R_{\max}^2$. Following the proof in Section C.1 in Xie et al. [2021] completes our proof. □

With the observations above, we proceed with proving Lemma 4.5.8.

Proof of Lemma 4.5.8. We analyze the pessimistic estimate and note that the analysis is similar for the other part. Let $\breve{\pi}_r^{(t)}$ be the policy iterate of Algorithm 11 and $\breve{Q}_r^{(t)}$ the corresponding value function estimate. We know that

$$\begin{aligned} V_1^\pi(s_0;r) - \frac{1}{T}\sum_{t=1}^T \breve{Q}_{1,r}^{(t)}(s_0, \breve{\pi}_{1,r}^{(t)}) &= \frac{1}{T}\sum_{t=1}^T \left(Q_1^\pi(s_0,\pi_1;r) - \breve{Q}_{1,r}^{(t)}(s_0, \breve{\pi}_{1,r}^{(t)}) \right) \\ &\leqslant \frac{1}{T}\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_\pi \left[\langle \breve{Q}_{h,r}^{(t)}(s_h,\cdot), \pi_h(\cdot|s_h) - \breve{\pi}_{h,r}^{(t)}(\cdot|s_h)\rangle \right] \\ &\quad + \left| \frac{1}{T}\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_\pi \left[\breve{Q}_{h,r}^{(t)} - \mathcal{T}_{h,r}^{\breve{\pi}_r^{(t)}} \breve{Q}_{h+1,r}^{(t)} \right] \right|, \end{aligned}$$

where the inequality is by a standard argument in episodic reinforcement learning (see, for example, Lemma A.1 in Jin et al. [2021b] or Section B.1 in Cai et al. [2020]). By Lemma 4.5.12, we know that when $\eta = \sqrt{\frac{\log|\mathcal{A}|}{2H^2 R_{\max}^2 T}}$, we have

$$\frac{1}{T}\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_\pi \left[\langle \breve{Q}_{h,r}^{(t)}(s_h,\cdot), \pi_h(\cdot|s_h) - \breve{\pi}_{h,r}^{(t)}(\cdot|s_h)\rangle \right] \leqslant 2H^2 R_{\max} \sqrt{\frac{2\log|\mathcal{A}|}{T}}.$$

For all $t \in [T]$, similar to the proof of Lemma 4.5.7, when $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^2}\right)^{1/3}$, we have

$$\left|\sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\breve{Q}_{h,r}^{(t)} - \mathcal{T}_{h,r}^{\breve{\pi}_r^{(t)}} \breve{Q}_{h+1,r}^{(t)}\right]\right| \leqslant H\sqrt{C^{\breve{\pi}_r^{(t)}}(\pi)}\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}}+12\epsilon_{\mathcal{F}}+3\epsilon_{\mathcal{F},\mathcal{F}}}\right).$$

Notice that the distribution shift coefficient is changed from $C^{\pi}(\pi)$ to $C^{\breve{\pi}_r^{(t)}}(\pi)$, as the policy specific Bellman operator $\mathcal{T}$ is now induced by policy $\breve{\pi}_r^{(t)}$ rather than π. Taking the average over t and applying the triangle inequality give us

$$\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi}\left[\breve{Q}_{h,r}^{(t)} - \mathcal{T}_{h,r}^{\breve{\pi}_r^{(t)}} \breve{Q}_{h+1,r}^{(t)}\right]\right| \leqslant H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}_r^{(t)}}(\pi)}\right)\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}}+12\epsilon_{\mathcal{F}}+3\epsilon_{\mathcal{F},\mathcal{F}}}\right).$$

Combining the bounds, we have

$$V_1^{\pi}(s_0;r) - \frac{1}{T}\sum_{t=1}^{T}\breve{Q}_{1,r}^{(t)}(s_0,\breve{\pi}_{1,r}^{(t)}) \leqslant 2H^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + H\left(\frac{1}{T}\sum_{t=1}^{T}\sqrt{C^{\breve{\pi}_r^{(t)}}(\pi)}\right)\left(2(HR_{\max})^{1/3}(\epsilon_{\mathrm{S}}+3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathrm{S}}+12\epsilon_{\mathcal{F}}+3\epsilon_{\mathcal{F},\mathcal{F}}}\right),$$

which completes the proof. □

4.5.5 Concentration Analysis

In this section, we prove the concentration lemmas used in Section 4.5.3.

Proof of Lemma 4.5.2

We start by including a minor adaptation of a useful result from Györfi et al. [2002].

Theorem 4.5.13 (Adaptation of Theorem 11.6 from Györfi et al. [2002])**.** *Let $B \geqslant 1$ and let $\mathcal{G}$ be a class of functions $g : \mathbb{R}^d \to [0, B]$. Let $Z_1, Z_2, \ldots, Z_K$ be i.i.d. $\mathbb{R}^d$-valued random variables. Assume $\alpha > 0$, $0 < \epsilon < 1$, and $K \geqslant 1$. Then*

$$\Pr\left(\sup_{g\in\mathcal{G}} \frac{\frac{1}{K}\sum_{j=1}^{K} g(Z_j) - \mathbb{E}[Z_j]}{\alpha + \frac{1}{K}\sum_{j=1}^{K} g(Z_j) + \mathbb{E}[Z_j]} > \epsilon\right) \leqslant 4\mathcal{N}_\infty\left(\frac{\alpha\epsilon}{5}, \mathcal{G}\right)\exp\left(-\frac{3\epsilon^2\alpha K}{40B}\right).$$

Proof. By Theorem 11.6 from Györfi et al. [2002], we know that

$$\Pr\left(\sup_{g\in\mathcal{G}} \frac{\frac{1}{K}\sum_{j=1}^{K} g(Z_j) - \mathbb{E}[Z_j]}{\alpha + \frac{1}{K}\sum_{j=1}^{K} g(Z_j) + \mathbb{E}[Z_j]} > \epsilon\right) \leqslant 4\mathbb{E}\left[\mathcal{N}_1\left(\frac{\alpha\epsilon}{5}, \mathcal{G}, \{Z_j\}_{j=1}^{K}\right)\right]\exp\left(-\frac{3\epsilon^2\alpha K}{40B}\right),$$

where $\mathcal{N}_1\left(\frac{\alpha\epsilon}{5}, \mathcal{G}, \{Z_j\}_{j=1}^{K}\right)$ is the cardinality of the smallest set of functions $\{g^l\}_{l=1}^{L}$ such that for all $g \in \mathcal{G}$ there exists some $l \in [L]$ where

$$\frac{1}{K}\sum_{j=1}^{K}\left|g(Z_j) - g^l(Z_j)\right| \leqslant \frac{\alpha\epsilon}{5}.$$

See Section 11.4 from Györfi et al. [2002] for a detailed proof of the statement above. We then show that for any $\{Z_j\}_{j=1}^{K}$, $\mathcal{N}_1\left(\frac{\alpha\epsilon}{5}, \mathcal{G}, \{Z_j\}_{j=1}^{K}\right) \leqslant \mathcal{N}_\infty\left(\frac{\alpha\epsilon}{5}, \mathcal{G}\right)$. Let $\{\widetilde{g}^l\}_{l=1}^{L}$ be an $\frac{\alpha\epsilon}{5}$-covering of $\mathcal{G}$ with respect to the ℓ_∞-norm. We then know that for any $g \in \mathcal{G}$, there exists some $l \in [L]$ such that

$$\frac{1}{K}\sum_{j=1}^{K}|g(Z_j) - \widetilde{g}^l(Z_j)| \leqslant \frac{1}{K}\sum_{j=1}^{K}\frac{\alpha\epsilon}{5} = \frac{\alpha\epsilon}{5}.$$

Therefore $\{\widetilde{g}^l\}_{l=1}^{L}$ satisfies the requirement above, concluding our proof. □

Let $h \in [H], r \in \widetilde{\mathcal{R}}$ be arbitrary and fixed. First, we show

$$\begin{aligned}\Pr\Big(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \mathbb{E}_{\mu_h}\left[\|f_h - \mathcal{T}^{\pi}_{h,r} f'_{h+1}\|^2\right] - \mathcal{L}_{h,r}(f_h, f'_{h+1}, \pi; \mathcal{D}) \\ + \mathcal{L}_{h,r}(\mathcal{T}^{\pi}_{h,r} f'_{h+1}, f'_{h+1}, \pi; \mathcal{D}) \geqslant \epsilon\big(\alpha + \beta + \mathbb{E}_{\mu_h}\left[\|f_h - \mathcal{T}^{\pi}_{h,r} f'_{h+1}\|^2\right]\big)\Big) \\ \leqslant 14\left(\mathcal{N}_\infty\left(\frac{\epsilon\beta}{140 H R_{\max}}, \mathcal{F}\right)\right)^2 \mathcal{N}_{\infty,1}\left(\frac{\epsilon\beta}{140 H^2 R^2_{\max}}, \Pi\right) \exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{214(1+\epsilon)H^4 R^4_{\max}}\right).\end{aligned}$$

for all $\alpha, \beta > 0$, $0 < \epsilon \leqslant 1/2$.

Let Z be the random vector $(s_h, a_h, r_h(s_h, a_h), s_{h+1})$ where $(s_h, a_h, s_{h+1}) \sim \mu_h$. Let Z_j be its realization for any $j \in [K]$ drawn independently from $\mathcal{D}_h$. For any $f, f' \in \mathcal{F}$, and $\pi \in \Pi$, we further define the random variable

$$\begin{aligned}g^{\pi}_{f,f'}(Z) = (f_h(s_h, a_h) - r_h - f'_{h+1}(s_{h+1}, \pi_{h+1}))^2 \\ - (\mathcal{T}^{\pi}_{h,r} f'_{h+1}(s_h, a_h) - r_h - f'_{h+1}(s_{h+1}, \pi_{h+1}))^2,\end{aligned}$$

and $g^{\pi}_{f,f'}(Z_j)$ its empirical counterpart evaluated on Z's realization, Z_j. We begin by showing some basic properties of the random variable $g^{\pi}_{f,f'}(Z)$. Recall that by definition of the Bellman evaluation operator

$$\mathcal{T}^{\pi}_{h,r} f'_{h+1}(s_h, a_h) = \mathbb{E}_{\mathcal{P}}\left[r_h + f'_{h+1}(s_{h+1}, \pi_{h+1}) | s_h, a_h\right]. \tag{4.5.12}$$

Since $\mathcal{T}^{\pi}_{h,r} f_{h+1}(s_h, a_h) = \mathbb{E}_{\mu_h}\Big[r_h + f'_{h+1}(s_{h+1}, \pi_{h+1})|s_h, a_h\Big]$, by the law of total probability

$$
\begin{aligned}
&\mathbb{E}_{Z\sim\mu_h}[g^{\pi}_{f,f'}(Z)] \\
&\quad= \mathbb{E}_{s_h,a_h\sim\mu_h}\Big[\mathbb{E}_{s_{h+1}\sim\mu_h|s_h,a_h}[(f_h(s_h,a_h) - r_h - f'_{h+1}(s_{h+1},\pi_{h+1}))^2 - \\
&\qquad\qquad (\mathcal{T}^{\pi}_{h,r} f'_{h+1}(s_h,a_h) - r_h - f'_{h+1}(s_{h+1},\pi_{h+1}))^2|s_h,a_h]\Big] \\
&\quad= \mathbb{E}_{\mu_h}\Big[\mathbb{E}_{s_{h+1}\sim\mu_h|s_h,a_h}[(f_h(s_h,a_h) + \mathcal{T}^{\pi}_{h,r} f'_{h+1}(s_h,a_h) - 2(r_h + f'_{h+1}(s_{h+1},\pi_{h+1})))) \\
&\qquad\qquad \times (f_h(s_h,a_h) - \mathcal{T}^{\pi}_{h,r} f'_{h+1}(s_h,a_h))|s_h,a_h]\Big] \\
&\quad= \mathbb{E}_{\mu_h}\Big[\|f_h(s_h,a_h) - \mathcal{T}^{\pi}_{h,r} f'_{h+1}(s_h,a_h)\|^2\Big].
\end{aligned}
$$

Additionally, recalling that $r_h \in [-R_{\max}, R_{\max}]$, $f'_{h+1} \in [-(H-h)R_{\max}, (H-h)R_{\max}]$, $f_h \in [-(H-h+1)R_{\max}, (H-h+1)R_{\max}]$, we know that $g^{\pi}_{f,f'}(Z) \in [-16H^2R^2_{\max}, 16H^2R^2_{\max}]$. Lastly, notice that

$$
\begin{aligned}
&\mathrm{Var}(g^{\pi}_{f,f'}(Z)) \leqslant \mathbb{E}[(g^{\pi}_{f,f'}(Z))^2] \\
&\quad= \mathbb{E}\Big[\mathbb{E}[(f_h(s_h,a_h) + \mathcal{T}^{\pi}_{h,r} f'_{h+1}(s_h,a_h) - 2(r_h + f'_{h+1}(s_{h+1},\pi_{h+1})))^2 \\
&\qquad\qquad \times (f_h(s_h,a_h) - \mathcal{T}^{\pi}_{h,r} f'_{h+1}(s_h,a_h))^2|s_h,a_h]\Big] \\
&\quad\leqslant \mathbb{E}[16H^2R^2_{\max}(f_h(s_h,a_h) - \mathcal{T}^{\pi}_{h,r} f'_{h+1}(s_h,a_h))^2] = 16H^2R^2_{\max}\mathbb{E}[g^{\pi}_{f,f'}(Z)],
\end{aligned}
\tag{4.5.13}
$$

where for the last inequality we noticed that

$$
f_h(s_h,a_h) + \mathcal{T}^{\pi}_{h,r} f'_{h+1}(s_h,a_h) - 2(r_h + f'_{h+1}(s_{h+1},\pi_{h+1})) \in [-4HR_{\max}, 4HR_{\max}].
$$

Our ensuing proof largely follows the structure of Section 11.5 of Györfi et al. [2002] and we reproduce the proof below for completeness. Let $\alpha, \beta > 0$ and $0 < \epsilon \leqslant \frac{1}{2}$ be arbitrary and fixed constants. We now proceed with the proof.

Symmetrization by Ghost Sample. Consider some $(f_n, f'_n, \pi_n) \in \mathcal{F} \times \mathcal{F} \times \Pi$ depending

on $\{Z_j\}_{j=1}^K$ such that

$$\mathbb{E}[g^{\pi_n}_{f_n,f'_n}(Z)|\{Z_j\}_{j=1}^K] - \frac{1}{K}\sum_{j=1}^K g^{\pi_n}_{f_n,f'_n}(Z_j) \geqslant \epsilon(\alpha+\beta+\mathbb{E}[g^{\pi_n}_{f_n,f'_n}(Z)|\{Z_j\}_{\tau=1}^K]),$$

if such (f_n, f'_n, π_n) exists. If not, choose some arbitrary (f_n, f'_n, π_n). As a shorthand notation, let $g_n = g^{\pi_n}_{f_n,f'_n}$. Finally, introduce ghost samples $\{Z'_j\}_{j=1}^K \sim \mu_h$, drawn i.i.d. from the same distribution as $\{Z_j\}_{j=1}^K$. Recalling that the variance of g_n is bounded by $16\mathbb{E}[g_n(Z)]$, by Chebyshev's inequality we have

$$\begin{aligned}
&\Pr\left(\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K] - \frac{1}{K}\sum_{j=1}^K g_n(Z'_j) \geqslant \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K]|\{Z_j\}_{j=1}^K\right) \\
&\quad\leqslant \frac{\mathrm{Var}(g_n(Z)|\{Z_j\}_{j=1}^K)}{K(\frac{\epsilon}{2}(\alpha+\beta)+\frac{\epsilon}{2}\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K])^2} \\
&\quad\leqslant \frac{16H^2R_{\max}^2\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K]}{K(\frac{\epsilon}{2}(\alpha+\beta)+\frac{\epsilon}{2}\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K])^2} \\
&\quad\leqslant \frac{16H^2R_{\max}^2}{\epsilon^2(\alpha+\beta)K},
\end{aligned}$$

where the last inequality comes from the fact that $\frac{s_0}{(a+s_0)^2} \leqslant \frac{1}{4a}$ for all $s_0 \geqslant 0$ and $a > 0$. Thus, for all $K \geqslant \frac{128H^2R_{\max}^2}{\epsilon^2(\alpha+\beta)}$,

$$\Pr\left(\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K] - \frac{1}{K}\sum_{j=1}^K g_n(Z'_j) \geqslant \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K]|\{Z_j\}_{j=1}^K\right) \leqslant \frac{7}{8}.$$

We then know that

$$\begin{aligned}
&\Pr\Bigg(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \\
&\qquad \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z'_i) - \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z_j) \geqslant \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbb{E}[g^{\pi}_{f_h, f'_{h+1}}(Z)]\Bigg) \\
&\geqslant \Pr\Bigg(\frac{1}{K}\sum_{j=1}^{K} g_n(Z'_i) - \frac{1}{K}\sum_{j=1}^{K} g_n(Z_j) \geqslant \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^{K}]\Bigg) \\
&\geqslant \frac{7}{8}\Pr\Bigg(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \\
&\qquad \mathbb{E}[g^{\pi}_{f_h, f'_{h+1}}(Z)] - \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z_j) \geqslant \epsilon(\alpha+\beta) + \epsilon\mathbb{E}[g^{\pi}_{f_h, f'_{h+1}}(Z)]\Bigg).
\end{aligned}$$

In other words, for $K \geqslant \frac{128H^2R_{\max}^2}{\epsilon^2(\alpha+\beta)}$,

$$\begin{aligned}
&\Pr\Bigg(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \mathbb{E}[g^{\pi}_{f_h, f'_{h+1}}(Z)] - \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z_j) \geqslant \epsilon(\alpha+\beta) + \epsilon\mathbb{E}[g^{\pi}_{f_h, f'_{h+1}}(Z)]\Bigg) \\
&\qquad \leqslant \frac{8}{7}\Pr\Bigg(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z'_j) \\
&\qquad\qquad - \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z_j) \geqslant \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbb{E}[g^{\pi}_{f_h, f'_{h+1}}(Z)]\Bigg). \quad (4.5.14)
\end{aligned}$$

Replacement of Expectation by Empirical Mean of Ghost Sample We begin by

noticing

$$
\begin{aligned}
&\Pr\Big(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \\
&\qquad \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z'_i) - \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z_j) \geqslant \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbb{E}[g^{\pi}_{f_h, f'_{h+1}}(Z)]\Big) \\
&\leqslant \Pr\Big(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \\
&\qquad \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z'_i) - \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z_j) \geqslant \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbb{E}[g^{\pi}_{f_h, f'_{h+1}}(Z)], \\
&\qquad \frac{1}{K}\sum_{j=1}^{K} (g^{\pi}_{f_h, f'_{h+1}})^2(Z'_i) - \mathbb{E}[(g^{\pi}_{f_h, f'_{h+1}})^2(Z)] \leqslant \\
&\qquad\qquad \epsilon\Big(\alpha+\beta + \frac{1}{K}\sum_{j=1}^{K} (g^{\pi}_{f_h, f'_{h+1}})^2(Z_j) + \mathbb{E}[(g^{\pi}_{f_h, f'_{h+1}})^2(Z)]\Big), \\
&\qquad \frac{1}{K}\sum_{j=1}^{K} (g^{\pi}_{f_h, f'_{h+1}})^2(Z'_i) - \mathbb{E}[(g^{\pi}_{f_h, f'_{h+1}})^2(Z)] \leqslant \\
&\qquad\qquad \epsilon\Big(\alpha+\beta + \frac{1}{K}\sum_{j=1}^{K} (g^{\pi}_{f_h, f'_{h+1}})^2(Z'_i) + \mathbb{E}[(g^{\pi}_{f_h, f'_{h+1}})^2(Z)]\Big)\Big) \\
&\quad + 2\Pr\Big(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \\
&\qquad\qquad \frac{\frac{1}{K}\sum_{j=1}^{K} (g^{\pi}_{f_h, f'_{h+1}})^2(Z_j) - \mathbb{E}[(g^{\pi}_{f_h, f'_{h+1}})^2(Z)]}{\alpha+\beta + \frac{1}{K}\sum_{j=1}^{K} (g^{\pi}_{f_h, f'_{h+1}})^2(Z_j) + \mathbb{E}[(g^{\pi}_{f_h, f'_{h+1}})^2(Z)]}\Big).
\end{aligned}
\tag{4.5.15}
$$

Citing Theorem 4.5.13, we may bound the second probability term on the right hand side as

$$
\begin{aligned}
&\Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{\frac{1}{K}\sum_{j=1}^{K} (g^{\pi}_{f_h, f'_{h+1}})^2(Z_j) - \mathbb{E}[(g^{\pi}_{f_h, f'_{h+1}})^2(Z)]}{\left(\alpha+\beta + \frac{1}{K}\sum_{j=1}^{K} (g^{\pi}_{f_h, f'_{h+1}})^2(Z_j) + \mathbb{E}[(g^{\pi}_{f_h, f'_{h+1}})^2(Z)]\right)}\right) \\
&\quad \leqslant 4\mathcal{N}_{\infty}\left(\frac{(\alpha+\beta)\epsilon}{5}, \{g^{\pi}_{f_h, f'_{h+1}} : f, f' \in \mathcal{F}, \pi \in \Pi\}\right) \exp\left(-\frac{3\epsilon^2(\alpha+\beta)K}{40(16H^2R^2_{\max})}\right).
\end{aligned}
$$

For the first probability term, notice that the second event in the conjunction implies

$$(1+\epsilon)\mathbb{E}[(g^{\pi}_{f_h,f'_{h+1}})^2(Z)] \geqslant (1-\epsilon)\frac{1}{K}\sum_{j=1}^{K}(g^{\pi}_{f_h,f'_{h+1}})^2(Z_j) - \epsilon(\alpha+\beta),$$

which is equivalent to

$$\frac{1}{32H^2R_{\max}^2}\mathbb{E}[(g^{\pi}_{f_h,f'_{h+1}})^2(Z)] \geqslant \frac{(1-\epsilon)\frac{1}{K}\sum_{j=1}^{K}(g^{\pi}_{f_h,f'_{h+1}})^2(Z_j)}{32H^2R_{\max}^2(1+\epsilon)} - \frac{(\alpha+\beta)\epsilon}{32H^2R_{\max}^2(1+\epsilon)}.$$

A similar bound may be obtained for the term involving Z'_i. Noticing that by equation (4.5.13), we have $\mathbb{E}[g^{\pi}_{f_h,f'_{h+1}}(Z)] \geqslant \frac{1}{16H^2R_{\max}^2}\mathbb{E}[(g^{\pi}_{f_h,f'_{h+1}})^2(Z)]$, and we know the first probability term in (4.5.15) can be bounded by

$$\Pr\Bigg(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h,f'_{h+1}}(Z'_i) - \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h,f'_{h+1}}(Z_j) \geqslant \frac{\epsilon}{2}(\alpha+\beta) - \frac{2\epsilon^2(\alpha+\beta) + \epsilon(1-\epsilon)\left(\frac{1}{K}\sum_{j=1}^{K}((g^{\pi}_{f_h,f'_{h+1}})^2(Z'_j) + (g^{\pi}_{f_h,f'_{h+1}})^2(Z_j))\right)}{64H^2R_{\max}^2(1+\epsilon)}\Bigg).$$

Additional Randomization by Random Signs Let $\{U_j\}_{j=1}^{K}$ be i.i.d. Rademacher random variables drawn independently from $\{Z_j\}_{j=1}^{K}$ and $\{Z'_j\}_{j=1}^{K}$. Because $\{Z_j\}_{j=1}^{K}$ and $\{Z'_j\}_{j=1}^{K}$

are i.i.d., we know that

$$\Pr\Bigg(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z'_j) - \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z_j) \geqslant \frac{\epsilon}{2}(\alpha+\beta) - \frac{2\epsilon^2(\alpha+\beta) + \epsilon(1-\epsilon)(\frac{1}{K}\sum_{j=1}^{K}((g^{\pi}_{f_h, f'_{h+1}})^2(Z'_j) + (g^{\pi}_{f_h, f'_{h+1}})^2(Z_j)))}{64H^2R^2_{\max}(1+\epsilon)}\Bigg)$$
$$\leqslant 2\Pr\Bigg(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K}\sum_{j=1}^{K}\left|U_j g^{\pi}_{f_h, f'_{h+1}}(Z_j)\right| \geqslant \frac{\epsilon}{4}(\alpha+\beta) - \frac{\epsilon^2(\alpha+\beta)}{64H^2R^2_{\max}(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R^2_{\max}(1+\epsilon)}\frac{1}{K}\sum_{j=1}^{K}((g^{\pi}_{f_h, f'_{h+1}})^2(Z_j))\Bigg). \tag{4.5.16}$$

Conditioning and Covering We then condition the probability on $\{Z_j\}_{j=1}^K$. Fix some $z_1, \ldots, z_K$ and we consider instead

$$\Pr\Bigg\{\exists f, f' \in \mathcal{F}, \pi \in \Pi : \left|\frac{1}{K}\sum_{j=1}^{K} U_j g^{\pi}_{f_h, f'_{h+1}}(z_j)\right| \geqslant \frac{\epsilon(\alpha+\beta)}{4} - \frac{\epsilon^2(\alpha+\beta)}{64H^2R^2_{\max}(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R^2_{\max}(1+\epsilon)}\frac{1}{K}\sum_{j=1}^{K}(g^{\pi}_{f_h, f'_{h+1}})^2(z_j)\Bigg\}.$$

Let $\delta > 0$ and let $\mathcal{G}_\delta$ be an ℓ_∞ δ-cover of $\mathcal{G}_{\mathcal{F},\Pi} = \{g^{\pi}_{f_h, f'_{h+1}} : f, f' \in F, \pi \in \Pi\}$. Fix some $(f, f', \pi) \in \mathcal{F} \times \mathcal{F} \times \Pi$ and there exists some $g \in \mathcal{G}_\delta$ such that $\sup_z |g(z) - g^{\pi}_{f_h, f'_{h+1}}(z)| < \delta$. We then know that

$$\left|\frac{1}{K}\sum_{j=1}^{K} U_j g^{\pi}_{f_h, f'_{h+1}}(z_j)\right| \leqslant \left|\frac{1}{K}\sum_{j=1}^{K} U_j g(z_j)\right| + \delta$$

and

$$\frac{1}{K}\sum_{j=1}^{K}(g^{\pi}_{f_h,f'_{h+1}})^2(z_j) = \frac{1}{K}\sum_{j=1}^{K} g^2(z_j) + \frac{1}{K}\sum_{j=1}^{K}((g^{\pi}_{f_h,f'_{h+1}})^2(z_j) - g^2(z_j))$$
$$\geqslant \frac{1}{K}\sum_{j=1}^{K} g^2(z_j) - 8H^2R^2_{\max}\delta.$$

Set $\delta = \frac{\beta\epsilon}{5}$. Notice that as $HR_{\max} \geqslant 1$, $0 < \epsilon \leqslant \frac{1}{2}$, we have

$$\frac{\epsilon\beta}{4} - \frac{\epsilon^2\beta}{64H^2R^2_{\max}(1+\epsilon)} - \delta - \delta\frac{\epsilon(1-\epsilon)}{8(1+\epsilon)} = \frac{\epsilon\beta}{2} - \frac{\epsilon^2\beta}{64H^2R^2_{\max}(1+\epsilon)} - \frac{\epsilon^2(1-\epsilon)\beta}{40(1+\epsilon)} \geqslant 0.$$

Therefore we have

$$\Pr\Bigg\{\exists f, f' \in \mathcal{F}, \pi \in \Pi : \left|\frac{1}{K}\sum_{j=1}^{K} U_j g^{\pi}_{f_h,f'_{h+1}}(z_j)\right| \geqslant$$
$$\frac{\epsilon(\alpha+\beta)}{4} - \frac{\epsilon^2(\alpha+\beta)}{64H^2R^2_{\max}(1+\epsilon)} + \frac{\epsilon(1-\epsilon)\frac{1}{K}\sum_{j=1}^{K}(g^{\pi}_{f_h,f'_{h+1}})^2(z_j)}{64H^2R^2_{\max}(1+\epsilon)}\Bigg\}$$
$$\leqslant |\mathcal{G}_{\epsilon\beta/5}| \max_{g\in\mathcal{G}_{\epsilon\beta/5}} \Pr\Bigg\{\left|\frac{1}{K}\sum_{j=1}^{K} U_j g(z_j)\right| \geqslant \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{64H^2R^2_{\max}(1+\epsilon)} +$$
$$\frac{\epsilon(1-\epsilon)}{64H^2R^2_{\max}(1+\epsilon)}\frac{1}{K}\sum_{j=1}^{K} g^2(z_j)\Bigg\}. \qquad (4.5.17)$$

We then apply Bernstein's inequality to bound

$$\Pr\Bigg\{\left|\frac{1}{K}\sum_{j=1}^{K} U_j g(z_j)\right| \geqslant \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{64H^2R^2_{\max}(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R^2_{\max}(1+\epsilon)}\frac{1}{K}\sum_{j=1}^{K} g^2(z_j)\Bigg\}$$

for any $g \in \mathcal{G}_{\epsilon\beta/5}$. We begin by relating the variance of $U_j g(z_j)$ with $\frac{1}{K}\sum_{j=1}^{k} g^2(z_j)$. Notice

that as U_j is i.i.d. Rademacher,

$$\frac{1}{K}\sum_{j=1}^{K}\operatorname{Var}(U_j g(z_j)) = \frac{1}{K}\sum_{j=1}^{k} g^2(z_j)\operatorname{Var}(U_i) = \frac{1}{K}\sum_{j=1}^{k} g^2(z_j).$$

Perform a simple change of variable and let $V_j = g(z_j)U_j$. As $g(z_j) \in [-4H^2R_{\max}^2, 4H^2R_{\max}^2]$ for all z_j, we know $|V_j| \leqslant 4H^2R_{\max}^2$. For convenience, further let $A_1 = \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{64H^2R_{\max}^2(1+\epsilon)}$, $A_2 = \frac{\epsilon(1-\epsilon)}{64H^2R_{\max}^2(1+\epsilon)}$, and $\sigma^2 = \frac{1}{K}\sum_{j=1}^{K}\operatorname{Var}(U_j g(z_j)) = \frac{1}{K}\sum_{j=1}^{k} g^2(z_j)$. We then have for any $g \in \mathcal{G}_{\epsilon\beta/5}$

$$\begin{aligned}
&\Pr\left\{\left|\frac{1}{K}\sum_{j=1}^{K} U_j g(z_j)\right| \geqslant \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{64H^2R_{\max}^2(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R_{\max}^2(1+\epsilon)}\frac{1}{K}\sum_{j=1}^{K} g^2(z_j)\right\} \\
&\quad\leqslant 2\exp\left(-\frac{3KA_2}{16H^2R_{\max}^2}\frac{\left(\frac{A_1}{A_2}+\sigma^2\right)^2}{\frac{A_1}{A_2}+\left(1+\frac{3}{8H^2R_{\max}^2A_2}\right)\sigma^2}\right) \\
&\quad\leqslant 2\exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{140H^2R_{\max}^2(1+\epsilon)}\right),
\end{aligned}$$

where the last inequality follows a series of manipulations discussed in greater detail in page 218 of Györfi et al. [2002] that we omit here for brevity. Plugging the result back into equations (4.5.16) and (4.5.17) gives us

$$\begin{aligned}
&\Pr\Bigg(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z'_j) - \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z_j) \geqslant \frac{\epsilon}{2}(\alpha+\beta) - \\
&\qquad\qquad \frac{2\epsilon^2(\alpha+\beta) + \epsilon(1-\epsilon)(\frac{1}{K}\sum_{j=1}^{K}((g^{\pi}_{f_h,f'_{h+1}})^2(Z'_i) + (g^{\pi}_{f_h,f'_{h+1}})^2(Z_j)))}{64H^2R_{\max}^2(1+\epsilon)}\Bigg) \\
&\quad\leqslant 2\mathcal{N}_{\infty}\left(\frac{\epsilon\beta}{5}, \{g^{\pi}_{f_h,f'_{h+1}} : f, f' \in F, \pi \in \Pi\}\right)\exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{140H^2R_{\max}^2(1+\epsilon)}\right).
\end{aligned}$$

Recalling equations (4.5.15) and (4.5.16), we have

$$\begin{aligned}
&\Pr\Bigg(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z'_i) - \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z_j) \\
&\qquad\qquad \geqslant \frac{\epsilon}{2}(\alpha+\beta) + \frac{\epsilon}{2}\mathbb{E}[g^{\pi}_{f_h, f'_{h+1}}(Z)]\Bigg) \\
&\leqslant 4\mathcal{N}_\infty\left(\frac{\epsilon\beta}{5}, \{g^{\pi}_{f_h, f'_{h+1}} : f, f' \in F, \pi \in \Pi\}\right)\exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{140H^2R^2_{\max}(1+\epsilon)}\right) \\
&\quad + 8\mathcal{N}_\infty\left(\frac{(\alpha+\beta)\epsilon}{5}, \{g^{\pi}_{f_h, f'_{h+1}} : f, f' \in \mathcal{F}, \pi \in \Pi\}\right)\exp\left(-\frac{3\epsilon^2(\alpha+\beta)K}{640H^2R^2_{\max}}\right).
\end{aligned}$$

Plugging the result back into equation (4.5.14) and we finally know for $K \geqslant \frac{128H^2R^2_{\max}}{\epsilon^2(\alpha+\beta)}$,

$$\begin{aligned}
&\Pr\Bigg(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \mathbb{E}[g^{\pi}_{f_h, f'_{h+1}}(Z)] - \frac{1}{K}\sum_{j=1}^{K} g^{\pi}_{f_h, f'_{h+1}}(Z_j) \\
&\qquad\qquad \geqslant \epsilon(\alpha+\beta) + \epsilon\mathbb{E}[g^{\pi}_{f_h, f'_{h+1}}(Z)]\Bigg) \\
&\leqslant \frac{32}{7}\mathcal{N}_\infty\left(\frac{\epsilon\beta}{5}, \{g^{\pi}_{f_h, f'_{h+1}} : f, f' \in F, \pi \in \Pi\}\right)\exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{140H^2R^2_{\max}(1+\epsilon)}\right) \\
&\quad + \frac{64}{7}\mathcal{N}_\infty\left(\frac{(\alpha+\beta)\epsilon}{5}, \{g^{\pi}_{f_h, f'_{h+1}} : f, f' \in \mathcal{F}, \pi \in \Pi\}\right)\exp\left(-\frac{3\epsilon^2(\alpha+\beta)K}{640H^2R^2_{\max}}\right) \\
&\leqslant 14\mathcal{N}_\infty\left(\frac{\epsilon\beta}{5}, \{g^{\pi}_{f_h, f'_{h+1}} : f, f' \in F, \pi \in \Pi\}\right)\exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{214(1+\epsilon)H^4R^4_{\max}}\right).
\end{aligned}$$

When $K < \frac{128H^2R^2_{\max}}{\epsilon^2(\alpha+\beta)}$, $\exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{214(1+\epsilon)H^4R^4_{\max}}\right) \geqslant \exp\left(-\frac{128}{214}\right) \geqslant \frac{1}{14}$ and the claim trivially holds.

Bounding the Covering Number. Our final task is bounding

$$\mathcal{N}_\infty\left(\frac{\epsilon\beta}{5}, \{g^{\pi}_{f_h, f'_{h+1}} : f, f' \in F, \pi \in \Pi\}\right)$$

using the covering numbers of Π and $\mathcal{F}$. Let $\mathcal{F}_0$ be a $\frac{\epsilon\beta}{140HR_{\max}}$-covering of $\mathcal{F}$ with respect to ℓ_∞ and Π_0 a $\frac{\epsilon\beta}{140H^2R^2_{\max}}$-covering of Π with respect to $\|\cdot\|_{\infty,1}$. We then know that for any

$f, f' \in \mathcal{F}, \pi \in \Pi$, there exits some $f^\dagger, f^\ddagger \in \mathcal{F}_0, \pi^\dagger \in \Pi_0$ such that

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |f_h(s,a) - f_h^\dagger(s,a)| \leqslant \frac{\epsilon\beta}{140HR_{\max}},$$
$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |f'_{h+1}(s,a) - f_{h+1}^\ddagger(s,a)| \leqslant \frac{\epsilon\beta}{140HR_{\max}},$$
$$\sup_{s\in\mathcal{S}} \int_{a\in\mathcal{A}} |\pi_{h+1}(a|s) - \pi_{h+1}^\dagger(a|s)| \leqslant \frac{\epsilon\beta}{140H^2R_{\max}^2}.$$

Consider any arbitrary $z = (s, a, r, s') \sim \mu_h$. We know that

$$\begin{aligned}
&\left| g^{\pi_{h+1}}_{f_h, f'_{h+1}}(z) - g^{\pi^\dagger_{h+1}}_{f_h^\dagger, f_{h+1}^\ddagger}(z) \right| \\
&\quad\leqslant \left| f_h(s,a) + f_h^\dagger(s,a) - 2r - f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger) \right| \\
&\qquad\times \left| f_h(s,a) - f_h^\dagger(s,a) + f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger) \right| \\
&\qquad+ \left| \mathcal{T}_{h,r}^{\pi_{h+1}} f'_{h+1}(s,a) + \mathcal{T}_{h,r}^{\pi^\dagger_{h+1}} f_{h+1}^\ddagger(s,a) - 2r - f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger) \right| \\
&\qquad\times \left| \mathcal{T}_{h,r}^{\pi_{h+1}} f'_{h+1}(s,a) - \mathcal{T}_{h,r}^{\pi^\dagger_{h+1}} f_{h+1}^\ddagger(s,a) + f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger) \right| \\
&\quad\leqslant 4HR_{\max} \left| f_h(s,a) - f_h^\dagger(s,a) + f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger) \right| \\
&\qquad+ 4HR_{\max} \left| \mathcal{T}_{h,r}^{\pi_{h+1}} f'_{h+1}(s,a) - \mathcal{T}_{h,r}^{\pi^\dagger_{h+1}} f_{h+1}^\ddagger(s,a) + f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger) \right|,
\end{aligned} \tag{4.5.18}$$

where for the last inequality we used the boundedness of functions in $\mathcal{F}_h$ and $\mathcal{F}_{h+1}$. We

then notice that

$$
\begin{aligned}
&\left|f_h(s,a) - f_h^\dagger(s,a) + f'_{h+1}(s',\pi_{h+1}) - f^\ddagger_{h+1}(s',\pi^\dagger_{h+1})\right| \\
&\quad\leqslant |f_h(s,a) - f_h^\dagger(s,a)| + |f'_{h+1}(s',\pi_{h+1}) - f^\ddagger_{h+1}(s',\pi^\dagger_{h+1})| \\
&\quad\leqslant \frac{\epsilon\beta}{140HR_{\max}} + |f'_{h+1}(s',\pi_{h+1}) - f'_{h+1}(s',\pi^\dagger_{h+1})| + |f'_{h+1}(s',\pi^\dagger_{h+1}) - f^\ddagger_{h+1}(s',\pi^\dagger_{h+1})| \\
&\quad\leqslant \frac{\epsilon\beta}{140HR_{\max}} + \|\pi_{h+1} - \pi^\dagger_{h+1}\|_1 \|f'_{h+1}\|_\infty + |f'_{h+1}(s',\pi^\dagger_{h+1}) - f^\ddagger_{h+1}(s',\pi^\dagger_{h+1})| \\
&\quad\leqslant \frac{\epsilon\beta}{140HR_{\max}} + \frac{\epsilon\beta}{140H^2R^2_{\max}} HR_{\max} + |f'_{h+1}(s',\pi^\dagger_{h+1}) - f^\ddagger_{h+1}(s',\pi^\dagger_{h+1})| \\
&\quad\leqslant \frac{\epsilon\beta}{140HR_{\max}} + \frac{\epsilon\beta}{140HR_{\max}} + \mathbb{E}_{a'\sim\pi^\dagger_{h+1}(\cdot|s')}[|f'_{h+1}(s',a') - f^\ddagger_{h+1}(s',a')|] \\
&\quad\leqslant \frac{3\epsilon\beta}{140HR_{\max}},
\end{aligned}
$$

where the third inequality uses Holder's inequality, the fourth definition of Π_0 and boundedness of $\mathcal{F}_h$, the fifth Jensen's inequality, and the last inequality the definition of $\mathcal{F}_0$. Additionally we have

$$
\begin{aligned}
&|\mathcal{T}^{\pi_{h+1}}_{h,r} f'_{h+1}(s,a) - \mathcal{T}^{\pi^\dagger_{h+1}}_{h,r} f^\ddagger_{h+1}(s,a) + f'_{h+1}(s',\pi_{h+1}) - f^\ddagger_{h+1}(s',\pi^\dagger_{h+1})| \\
&\quad\leqslant |\mathcal{T}^{\pi_{h+1}}_{h,r} f'_{h+1}(s,a) - \mathcal{T}^{\pi^\dagger_{h+1}}_{h,r} f^\ddagger_{h+1}(s,a)| + |f'_{h+1}(s',\pi_{h+1}) - f^\ddagger_{h+1}(s',\pi^\dagger_{h+1})| \\
&\quad\leqslant |\mathcal{T}^{\pi_{h+1}}_{h,r} f'_{h+1}(s,a) - \mathcal{T}^{\pi^\dagger_{h+1}}_{h,r} f^\ddagger_{h+1}(s,a)| + \frac{2\epsilon\beta}{140HR_{\max}} \\
&\quad\leqslant \mathbb{E}_{s''\sim\mathcal{P}_h(\cdot|s,a)} |f'_{h+1}(s',\pi_{h+1}) - f^\ddagger_{h+1}(s',\pi^\dagger_{h+1})| + \frac{2\epsilon\beta}{140HR_{\max}} \\
&\quad\leqslant \frac{4\epsilon\beta}{140HR_{\max}},
\end{aligned}
$$

where the second inequality uses the same reasoning as above to bound $|f'_{h+1}(s',\pi_{h+1}) - f^\ddagger_{h+1}(s',\pi^\dagger_{h+1})|$, the third Jensen's inequality, and the last inequality reuses the bound for $|f'_{h+1}(s',\pi_{h+1}) - f^\ddagger_{h+1}(s',\pi^\dagger_{h+1})|$ over arbitrary s'. Plugging these back into equation (4.5.18)

shows

$$\left| g^{\pi_{h+1}}_{f_h, f'_{h+1}}(z) - g^{\pi^\dagger_{h+1}}_{f^\dagger_h, f^\ddagger_{h+1}}(z) \right| \leqslant \frac{7\epsilon\beta}{140HR_{\max}} \times 4HR_{\max} = \frac{\epsilon\beta}{5}.$$

Thus

$$\mathcal{N}_\infty\left(\frac{\epsilon\beta}{5}, \{g^\pi_{f_h, f'_{h+1}} : f, f' \in F, \pi \in \Pi\}\right) \leqslant \left(\mathcal{N}_\infty\left(\frac{\epsilon\beta}{140HR_{\max}}, \mathcal{F}\right)\right)^2 \mathcal{N}_{\infty,1}\left(\frac{\epsilon\beta}{140H^2R^2_{\max}}, \Pi\right),$$

showing one side of the inequality holds.

To show the other side holds, simply replace $g^\pi_{f,f'}(Z)$ with its negative and repeat the analysis above. We then complete the proof by taking a union bound over both halves.

Proofs of "Good Event"

With the help of the previous theorem, we are able to show that $\mathcal{G}(\Pi_{\text{SPI}})$ occurs with high probability.

Proof of Lemma 4.5.3. Taking a union bound over all $h \in [H]$ and reported reward $r \in \widetilde{\mathcal{R}}$ recalling that $|\widetilde{\mathcal{R}}| \leqslant n + 1 \leqslant 2n$, by Lemma 4.5.2, we have

$$\begin{aligned}
&\Pr\Big(\exists h \in [H], r \in \widetilde{\mathcal{R}}, f, f' \in \mathcal{F}, \pi \in \Pi : \\
&\quad \left|\mathbb{E}_{\mu_h}\left[\|f_h - \mathcal{T}^\pi_{h,r} f'_{h+1}\|^2\right] - \mathcal{L}_{h,r}(f_h, f'_{h+1}, \pi; \mathcal{D}) + \mathcal{L}_{h,r}(\mathcal{T}^\pi_{h,r} f'_{h+1}, f'_{h+1}, \pi; \mathcal{D})\right| \\
&\qquad\qquad \geqslant \epsilon\left(\alpha + \beta + \mathbb{E}_{\mu_h}\left[\|f_h - \mathcal{T}^\pi_{h,r} f'_{h+1}\|^2\right]\right)\Big) \\
&\leqslant 56nH\left(\mathcal{N}_\infty\left(\frac{\epsilon\beta}{140HR_{\max}}, \mathcal{F}\right)\right)^2 \mathcal{N}_{\infty,1}\left(\frac{\epsilon\beta}{140H^2R^2_{\max}}, \Pi\right) \\
&\quad \times \exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{214(1+\epsilon)H^4R^4_{\max}}\right).
\end{aligned}$$

Letting $\alpha = \beta$ and $\epsilon = \frac{1}{2}$, setting the right hand side to δ, and solving for α gives us

$$\alpha \leqslant \frac{1}{K} \max \left\{ 5136 H^4 R_{\max}^4, 5136 H^4 R_{\max}^4 \log \frac{56 n H \mathcal{N}_\infty \left(\frac{H R_{\max}}{K}, \mathcal{F} \right) \mathcal{N}_{\infty,1} \left(\frac{1}{K}, \Pi \right)}{\delta} \right\}.$$

As $\log 56 \geqslant 1$, $n, H \geqslant 1$, and $0 < 1 < \delta$, the second term always dominates the first and we can simplify the inequality as

$$\alpha \leqslant \frac{5136 H^4 R_{\max}^4}{K} \log \frac{56 n H \mathcal{N}_\infty \left(\frac{19 H^3 R_{\max}^3}{K}, \mathcal{F} \right) \mathcal{N}_{\infty,1} \left(\frac{19 H^4 R_{\max}^4}{K}, \Pi \right)}{\delta},$$

completing the proof. □

Proof of Corollary 4.5.4. For convenience, let $\widehat{g}_{h,r}^{\pi} = \operatorname{argmin}_{g \in \mathcal{F}_h} \mathcal{L}_{h,r}(g, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D})$. We then know that

$$\begin{aligned}
\mathcal{E}_{h,r}(f_{h,r}^{\pi,*}, \pi; \mathcal{D}) &= \mathcal{L}_{h,r}(f_{h,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) - \mathcal{L}_{h,r}(\widehat{g}_{h,r}^{\pi}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) \\
&= \mathcal{L}_{h,r}(f_{h,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) - \mathcal{L}_{h,r}(\mathcal{T}_{h,r}^{\pi,*} f_{h+1,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) \\
&\quad - \left(\mathcal{L}_{h,r}(\widehat{g}_{h,r}^{\pi}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) - \mathcal{L}_{h,r}(\mathcal{T}_{h,r}^{\pi,*} f_{h+1,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) \right).
\end{aligned}$$

By Lemma 4.5.3, conditionally on the event $\mathcal{G}(\Pi)$ we have the following simultaneously:

$$\begin{aligned}
&\mathcal{L}_{h,r}(f_{h,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) - \mathcal{L}_{h,r}(\mathcal{T}_{h,r}^{\pi,*} f_{h+1,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) \\
&\quad \leqslant \epsilon_{\mathrm{S}} + \frac{3}{2} \mathbb{E}_{\mu_h} \left[\| f_{h,r}^{\pi,*} - \mathcal{T}_{h,r}^{\pi,*} f_{h+1,r}^{\pi,*} \|^2 \right], \\
&- \mathcal{L}_{h,r}(\widehat{g}_{h,r}^{\pi}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) + \mathcal{L}_{h,r}(\mathcal{T}_{h,r}^{\pi,*} f_{h+1,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) \leqslant \epsilon_{\mathrm{S}},
\end{aligned}$$

where the second inequality uses the fact that $\|\cdot\|^2$ is non-negative. Finally, noticing that

$$\begin{aligned}
\mathbb{E}_{\mu_h}\left[\|f^{\pi,*}_{h,r} - \mathcal{T}^{\pi,*}_{h,r} f^{\pi,*}_{h+1,r}\|^2\right] &\leqslant 2\mathbb{E}_{\mu_h}\left[\|f^{\pi,*}_{h,r} - Q^{\pi}_{h}(\cdot,\cdot;r)\|^2\right] \\
&\quad + 2\mathbb{E}_{\mu_h}\left[\|\mathcal{T}^{\pi,*}_{h,r} f^{\pi,*}_{h+1,r} - \mathcal{T}^{\pi,*}_{h,r} Q^{\pi}_{h}(\cdot,\cdot;r)\|^2\right] \\
&\leqslant 2\epsilon_{\mathcal{F}} + 2\mathbb{E}_{\mu'_{h+1}}\left[\|f^{\pi,*}_{h+1,r} - Q^{\pi}_{h+1}(\cdot,\cdot;r)\|^2\right] \\
&\leqslant 4\epsilon_{\mathcal{F}},
\end{aligned}$$

where μ'_{h+1} shares the marginal distribution over $\mathcal{S}$ with μ_{h+1} but the conditional distribution over $\mathcal{A}$ given $s \in \mathcal{S}$ is given by $\pi_{h+1}(\cdot|s)$. The final inequality comes from the fact that μ'_{h+1} is an admissible distribution under Assumption 4.2.3. □

Proof of Corollary 4.5.5. Let $\widehat{g}^{\pi}_{h,r} = \operatorname{argmin}_{g\in\mathcal{F}_h} \mathbb{E}_{\mu_h}[\|g - \mathcal{T}^{\pi}_{h,r} f^{\pi}_{h+1,r}\|^2]$. Recalling the definition of $\mathcal{E}_{h,r}$, we have

$$\begin{aligned}
\mathcal{E}_{h,r}(f^{\pi}_{h,r}, \pi; \mathcal{D}) &= \mathcal{L}_{h,r}(f^{\pi}_{h,r}, f^{\pi}_{h+1,r}, \pi; \mathcal{D}) - \min_{g\in\mathcal{F}_h} \mathcal{L}_{h,r}(g, f^{\pi}_{h+1,r}, \pi; \mathcal{D}) \\
&\geqslant \mathcal{L}_{h,r}(f^{\pi}_{h,r}, f^{\pi}_{h+1,r}, \pi; \mathcal{D}) - \mathcal{L}_{h,r}(\widehat{g}^{\pi}_{h,r}, f^{\pi}_{h+1,r}, \pi; \mathcal{D}) \\
&= \mathcal{L}_{h,r}(f^{\pi}_{h,r}, f^{\pi}_{h+1,r}, \pi; \mathcal{D}) - \mathcal{L}_{h,r}(\mathcal{T}^{\pi}_{h,r} f^{\pi}_{h+1,r}, f^{\pi}_{h+1,r}, \pi; \mathcal{D}) \\
&\quad - \left(\mathcal{L}_{h,r}(\widehat{g}^{\pi}_{h,r}, f^{\pi}_{h+1,r}, \pi; \mathcal{D}) - \mathcal{L}_{h,r}(\mathcal{T}^{\pi}_{h,r} f^{\pi}_{h+1,r}, f^{\pi}_{h+1,r}, \pi; \mathcal{D})\right).
\end{aligned}$$

By Lemma 4.5.3, conditionally on the event $\mathcal{G}(\Pi)$ we have the following:

$$\begin{aligned}
&\mathcal{L}_{h,r}(f^{\pi}_{h,r}, f^{\pi}_{h+1,r}, \pi; \mathcal{D}) - \mathcal{L}_{h,r}(\mathcal{T}^{\pi}_{h,r} f^{\pi}_{h+1,r}, f^{\pi}_{h+1,r}, \pi; \mathcal{D}) \\
&\quad \geqslant -\epsilon_{\mathrm{S}} + \frac{1}{2}\mathbb{E}_{\mu_h}\left[\|f^{\pi}_{h,r} - \mathcal{T}^{\pi}_{h,r} f^{\pi}_{h+1,r}\|^2\right], \\
&- \mathcal{L}_{h,r}(\widehat{g}^{\pi}_{h,r}, f^{\pi}_{h+1,r}, \pi; \mathcal{D}) + \mathcal{L}_{h,r}(\mathcal{T}^{\pi}_{h,r} f^{\pi}_{h+1,r}, f^{\pi}_{h+1,r}, \pi; \mathcal{D}) \\
&\quad \geqslant -\epsilon_{\mathrm{S}} - \frac{3}{2}\mathbb{E}_{\mu_h}\left[\|\widehat{g}^{\pi}_{h,r} - \mathcal{T}^{\pi}_{h,r} f^{\pi}_{h+1,r}\|^2\right].
\end{aligned}$$

Recalling that $\mathcal{E}_{h,r}(f, \pi; \mathcal{D}) \leqslant \epsilon_0$, we have

$$\mathbb{E}_{\mu_H}\left[\|f^{\pi}_{h,r} - \mathcal{T}^{\pi}_{h,r} f^{\pi}_{h+1,r}\|^2\right] \leqslant 4\epsilon_{\mathrm{S}} + 3\mathbb{E}_{\mu_h}\left[\|\widehat{g}^{\pi}_{h,r} - \mathcal{T}^{\pi}_{h,r} h^{\pi}_{h+1,r}\|^2\right] + 2\epsilon_0.$$

We conclude our proof by reminding ourselves of Assumption 4.2.4. □

CHAPTER 5

CONCLUSION

In this thesis, we explored three different variants of the same underlying question: can we use reinforcement learning to learn the optimal dynamic mechanism? As we have shown in these previous chapters, the answer to the question, specifically the resulting learning algorithm, is heavily dependent on the combination of the following factors: the complexity of the mechanism itself, the RL setup being considered, and the function approximation setting being used.

From Chapter 2 to Chapter 3, we see how focusing only on revenue maximizing within a specific subset of dynamic mechanisms significantly simplifies the characterization of the optimal mechanism. On the other hand, we further see how an online learning setup complicates the design of a learning algorithm, especially when we assume the buyers stay throughout the learning process.

From Chapter 3 to Chapter 4, we observe that learning the welfare-maximizing mechanism can be even easier and is equivalent to typical RL setups. On the other hand, Chapter 4 offers key insights on how RL can be used for learning dynamic mechanisms, specifically on the learned policies' uncertainties relate to violations in mechanism design desiderata. These insights are repeatedly used by Chapter 2 and Chapter 3.

As we wrap up this thesis, below we list several promising research directions. Specifically, we focus on those that draw inspiration from the setting in Chapter 2, which is discussed in detail in Section 2.2. We hope that these directions will prove to be useful to future researchers working in the same intersection of RL and dynamic mechanism design.

Enforcing seller constraints. Inspired by [Mirrokni et al., 2020], it may be interesting to see if typical dynamic mechanism design constraints such as buyer budget could be incorporated in our setting. A more compelling class of constraints focuses on the seller, and is

made especially interesting in the MDP setup. For instance, again using AWS spot instance pricing as an example, the company may prefer a more stable allocation level, that is, an allocation policy that has stable expected allocation at each of the H steps.

As another example, we may find ourselves in problem settings where the seller only has a limited amount of inventory. That is, over the course of H steps, she may only be able to allocate up to K items, with $K < H$. Such a constraint could have interesting synergy with the allocation's impact on later type distributions: how should the seller drum up demand when she only has a few items to sell?

Adapting to non-stationarity. When the learning algorithm discussed in 2.4 is deployed in real-life, the underlying MDP being learned can be non-stationary. For instance, recalling the AWS Spot Instance pricing example, the underlying MDP governing buyer's willingness-to-pay can change due to changes in underlying technology or research trends. The increased interest in large language models, for instance, can cause customers to be willing to pay more in general, due to increased demand for computation power.

We conjecture that the procedure described in Algorithm 1 can be combined with the non-stationary online learning framework proposed by Wei and Luo [2021]. As Algorithm 1 is inherently an optimistic one, despite the lack of an explicit uncertainty bonus, it should be fairly amenable to all meta-algorithms that takes an optimistic learning algorithm as an input. We conjecture that such a combination should be straightforward, although it remains interesting to see how the combined algorithm would perform in terms of regret.

Combining Chapter 2 and Chapter 3. Here, we refer to a combination of the setups in Chapter 2 and Chapter 3. Specifically in Chapter 3, we discuss how the order in which a seller sells a collection of items naturally affects later type distributions.

Consider the advertisement platform example provided in Chapter 2. What if the seller is able to both determine which ad spot is sold, in addition to the allocation rule and pricing

rule for the ad spot? Should a seller sell higher-valued ad spots first? Or should she save the best for last? Such a synergy could make our MDP formulation even more realistic, yet we must caution that having multiple items could quickly cause the computational costs to blow up: the procedure discussed in Section 2.4 blows up in complexity as the number of items increases.

REFERENCES

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

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Orna Agmon Ben-Yehuda, Muli Ben-Yehuda, Assaf Schuster, and Dan Tsafrir. Deconstructing Amazon EC2 spot instance pricing. *ACM Transactions on Economics and Computation (TEAC)*, 1(3):1–20, 2013.

Rui Ai, Boxiang Lyu, Zhaoran Wang, Zhuoran Yang, and Michael I Jordan. A reinforcement learning approach in multi-phase second-price auction design. *arXiv preprint arXiv:2210.10278*, 2022.

Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning prices for repeated auctions with strategic buyers. *Advances in Neural Information Processing Systems*, 26, 2013.

Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Repeated contextual auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Itai Ashlagi, Constantinos Daskalakis, and Nima Haghpanah. Sequential Mechanisms with ex-post Participation Guarantees, July 2016.

Itai Ashlagi, Constantinos Daskalakis, and Nima Haghpanah. Sequential mechanisms with ex post individual rationality. *Operations Research*, 71(1):245–258, 2023.

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

Peter Auer, Nicolo Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

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

Mark Bagnoli and Ted Bergstrom. Log-concave probability and its applications. In *Rationality and Equilibrium*, pages 217–241. Springer, 2006.

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

Maria-Florina Balcan, Avrim Blum, Jason D Hartline, and Yishay Mansour. Reducing mechanism design to algorithm design via machine learning. *Journal of Computer and System Sciences*, 74(8):1245–1270, 2008.

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

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

Marco Battaglini. Long-term contracting with Markovian consumers. *American Economic Review*, 95(3):637–658, 2005.

Matt Baughman, Simon Caton, Christian Haas, Ryan Chard, Rich Wolski, Ian Foster, and Kyle Chard. Deconstructing the 2017 changes to AWS spot market pricing. In *Proceedings of the 10th Workshop on Scientific Cloud Computing*, pages 19–26, 2019.

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

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

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

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

Dirk Bergemann and Juuso Välimäki. Dynamic mechanism design: An introduction. *Journal of Economic Literature*, 57(2):235–274, 2019.

Sergei Bernstein. On a modification of Chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Josef Broder and Paat Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Yang Cai, Argyris Oikonomou, Grigoris Velegkas, and Mingfei Zhao. *An Efficient ε-BIC to BIC Transformation and Its Application to Black-Box Reduction in Revenue Maximization*, pages 1337–1356. 2021. doi:10.1137/1.9781611976465.81.

Juan Carlos Carbajal and Jeffrey C Ely. Mechanism design without revenue equivalence. *Journal of Economic Theory*, 148(1):104–133, 2013.

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

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

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

Sarah H Cen and Devavrat Shah. Regret, stability & fairness in matching markets with bandit learners. In *International Conference on Artificial Intelligence and Statistics*, pages 8938–8968. PMLR, 2022.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour. Regret minimization for reserve prices in second-price auctions. *IEEE Transactions on Information Theory*, 61(1):549–564, 2014.

Jim X Chen. The evolution of computing: AlphaGo. *Computing in Science & Engineering*, 18(4):4–7, 2016a.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34, 2021a.

M Keith Chen. Dynamic pricing in a labor market: Surge pricing and flexible work on the Uber platform. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 455–455, 2016b.

Xiaoyu Chen, Jiachen Hu, Lin Yang, and Liwei Wang. Near-optimal reward-free exploration for linear mixture MDPs with plug-in solver. In *International Conference on Learning Representations*, 2021b.

Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. ChatGPT goes to law school. *Available at SSRN*, 2023.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

Pedro Cisneros-Velarde, Boxiang Lyu, Sanmi Koyejo, and Mladen Kolar. One policy is enough: Parallel exploration with a single policy is near-optimal for reward-free reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1965–2001. PMLR, 2023.

Bert J Claessens, Peter Vrancx, and Frederik Ruelens. Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control. *IEEE Transactions on Smart Grid*, 9(4):3259–3269, 2016.

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

Xiaowu Dai and Michael I Jordan. Learning strategies in decentralized matching markets under uncertain preferences. *Journal of Machine Learning Research*, 22(260):1–50, 2021.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient PAC RL with rich observations. *Advances in neural information processing systems*, 31, 2018.

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

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602 (7897):414–419, 2022.

Yuan Deng, Sébastien Lahaie, and Vahab Mirrokni. A robust non-clairvoyant dynamic mechanism for contextual auctions. *Advances in Neural Information Processing Systems*, 32, 2019.

Yuan Deng, Sébastien Lahaie, and Vahab Mirrokni. Robust pricing in dynamic mechanism design. In *International Conference on Machine Learning*, pages 2494–2503. PMLR, 2020.

Yuan Deng, Vahab Mirrokni, and Song Zuo. Non-clairvoyant dynamic mechanism design with budget constraints and beyond. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 369–369, 2021.

Matthias Doepke and Robert M Townsend. Dynamic mechanism design with hidden income and hidden actions. *Journal of Economic Theory*, 126(1):235–285, 2006.

Dmitri A Dolgov and Edmund H Durfee. Resource allocation among agents with MDP-induced preferences. *Journal of Artificial Intelligence Research*, 27:505–549, 2006.

Alexey Drutsa. Horizon-independent optimal pricing in repeated auctions with truthful and strategic buyers. In *Proceedings of the 26th International Conference on World Wide Web*, pages 33–42, 2017.

Alexey Drutsa. Reserve pricing in repeated second-price auctions with strategic bidders. In *International Conference on Machine Learning*, pages 2678–2689. PMLR, 2020.

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

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

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

Jianqing Fan, Yongyi Guo, and Mengxin Yu. Policy optimization using semiparametric models for dynamic pricing. *arXiv preprint arXiv:2109.06368*, 2021.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

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

Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. *The journal of socio-economics*, 40(1):35–42, 2011.

Jérémie Gallien. Dynamic mechanism design for online commerce. *Operations Research*, 54 (2):291–310, 2006.

Minbo Gao, Tianle Xie, Simon S Du, and Lin F Yang. A provably efficient algorithm for linear Markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.

Daniel F Garrett. Intertemporal price discrimination: Dynamic arrivals and changing values. *American Economic Review*, 106(11):3275–3299, 2016.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov Decision Processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR, 2019.

Gareth George, Rich Wolski, Chandra Krintz, and John Brevik. Analyzing AWS spot instance pricing. In *2019 IEEE International Conference on Cloud Engineering (IC2E)*, pages 222–228. IEEE, 2019.

Enrico H Gerding, Valentin Robu, Sebastian Stein, David C Parkes, Alex Rogers, and Nicholas R Jennings. Online mechanism design for electric vehicle charging. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 811–818, 2011.

Victor Ginsburgh and Jan C Van Ours. On organizing a sequential auction: results from a natural experiment by Christie's. *Oxford Economic Papers*, 59(1):1–15, 2007.

Negin Golrezaei, Adel Javanmard, and Vahab Mirrokni. Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Advances in Neural Information Processing Systems*, 32, 2019.

Negin Golrezaei, Patrick Jaillet, and Jason Cheuk Nam Liang. Incentive-aware contextual pricing with non-parametric market noise. In *International Conference on Artificial Intelligence and Statistics*, pages 9331–9361. PMLR, 2023.

Etan A Green and E Barry Plunkett. The science of the deal: Optimal bargaining on ebay using deep reinforcement learning. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 1–27, 2022.

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

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

Wenshuo Guo, Michael Jordan, and Ellen Vitercik. No-regret learning in partially-informed auctions. In *International Conference on Machine Learning*, pages 8039–8055. PMLR, 2022a.

Wenshuo Guo, Kirthevasan Kandasamy, Joseph Gonzalez, Michael Jordan, and Ion Stoica. Learning competitive equilibria in exchange economies with bandit feedback. In *International Conference on Artificial Intelligence and Statistics*, pages 6200–6224. PMLR, 2022b.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

Jason D Hartline. Bayesian mechanism design. *Theoretical Computer Science*, 8(3):143–263, 2012.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

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

Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael Jordan, and Jacob Steinhardt. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34, 2021.

Adel Javanmard and Hamid Nazerzadeh. Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research*, 20(1):315–363, 2019.

Adel Javanmard, Hamid Nazerzadeh, and Simeng Shao. Multi-product dynamic pricing in high-dimensions with heterogeneous price sensitivity. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2652–2657. IEEE, 2020.

Chunxiao Jiang, Yan Chen, Qi Wang, and KJ Ray Liu. Data-driven auction mechanism design in IaaS cloud computing. *IEEE Transactions on Services Computing*, 11(5):743–756, 2015.

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

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

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021a.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021b.

Chris Jones, Flavio Menezes, and Francis Vella. Auction price anomalies: Evidence from wool auctions in australia. *Economic Record*, 80(250):271–288, 2004.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Sham M Kakade, Ilan Lobel, and Hamid Nazerzadeh. Optimal dynamic mechanism design and the virtual-pivot mechanism. *Operations Research*, 61(4):837–854, 2013.

Kirthevasan Kandasamy, Joseph E Gonzalez, Michael I Jordan, and Ion Stoica. VCG mechanism design with unknown agent values under stochastic bandit feedback. *Journal of Machine Learning Research*, 24(53):1–45, 2023.

Yash Kanoria and Hamid Nazerzadeh. Dynamic reserve prices for repeated auctions: Learning from bids. In *Web and Internet Economics: 10th International Conference, WINE 2014, Beijing, China, December 14-17, 2014. Proceedings 10*, pages 232–232. Springer, 2014.

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

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

Bora Keskin, David Simchi-Levi, and Prem Talwai. Dynamic pricing and demand learning on a large network of products: A PAC-Bayesian approach. *arXiv preprint arXiv:2111.00790*, 2021.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823, 2020.

Christian Kleiber and Samuel Kotz. *Statistical size distributions in economics and actuarial sciences*. John Wiley & Sons, 2003.

Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE, 2003.

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

Daniel Krähmer and Roland Strausz. Optimal sales contracts with withdrawal rights. *The Review of Economic Studies*, 82(2):762–790, 2015.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

Kelly Y Lange, Jeffrey W Johnson, Kris Wilson, and Wesley Johnson. Price determinants of ranch horses sold at auction in texas. Technical report, Southern Agricultural Economics Association, 2010.

Sascha Lange and Martin Riedmiller. Deep auto-encoder neural networks in reinforcement learning. In *The 2010 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2010.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

Gen Li, Yuling Yan, Yuxin Chen, and Jianqing Fan. Minimax-optimal reward-agnostic exploration in reinforcement learning. *arXiv preprint arXiv:2304.07278*, 2023.

Yingkai Li, Yining Wang, and Yuan Zhou. Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*, pages 2173–2174. PMLR, 2019.

Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

Xiaodan Liang, Tairui Wang, Luona Yang, and Eric Xing. Cirl: Controllable imitative reinforcement learning for vision-based self-driving. In *Proceedings of the European conference on computer vision (ECCV)*, pages 584–599, 2018.

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

Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22(211):1–34, 2021a.

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

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 33:1264–1274, 2020.

Zhihan Liu, Miao Lu, Zhaoran Wang, Michael Jordan, and Zhuoran Yang. Welfare maximization in competitive equilibrium: Reinforcement learning for Markov exchange economy. In *International Conference on Machine Learning*, pages 13870–13911. PMLR, 2022.

Kenneth M Lusht. Order and price in a sequential auction. *The Journal of Real Estate Finance and Economics*, 8(3):259–266, 1994.

Boxiang Lyu, Qinglin Meng, Shuang Qiu, Zhaoran Wang, Zhuoran Yang, and Michael I Jordan. Learning dynamic mechanisms in unknown environments: A reinforcement learning approach. *arXiv preprint arXiv:2202.12797*, 2022a.

Boxiang Lyu, Zhaoran Wang, Mladen Kolar, and Zhuoran Yang. Pessimism meets VCG: Learning dynamic mechanism design via offline reinforcement learning. In *International Conference on Machine Learning*, pages 14601–14638. PMLR, 2022b.

Ke Ma and PR Kumar. The strategic LQG system: A dynamic stochastic VCG framework for optimal coordination. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 5777–5782. IEEE, 2018.

Yishay Mansour, Alex Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in Bayesian games. *Operations research*, (2):1105–1127, 2022.

Eric S Maskin. Mechanism design: How to implement social goals. *American Economic Review*, 98(3):567–76, 2008.

Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

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

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70 (2):583–601, 2002.

Yifei Min, Tianhao Wang, Ruitu Xu, Zhaoran Wang, Michael Jordan, and Zhuoran Yang. Learn to match with no regret: Reinforcement learning in markov matching markets. *Advances in Neural Information Processing Systems*, 35:19956–19970, 2022.

Vahab Mirrokni, Renato Paes Leme, Pingzhong Tang, and Song Zuo. Optimal dynamic mechanisms with ex-post IR via bank accounts. *arXiv preprint arXiv:1605.08840*, 2016a.

Vahab Mirrokni, Renato Paes Leme, Rita Ren, and Song Zuo. Dynamic mechanism design in the field. In *Proceedings of the 2018 World Wide Web Conference*, pages 1359–1368, 2018.

Vahab Mirrokni, Renato Paes Leme, Pingzhong Tang, and Song Zuo. Non-clairvoyant dynamic mechanism design. *Econometrica*, 88(5):1939–1963, 2020.

Vahab S. Mirrokni, Renato Paes Leme, Pingzhong Tang, and Song Zuo. Dynamic Auctions with Bank Accounts. In *IJCAI*, volume 16, pages 387–393, 2016b.

Sobhan Miryoosefi and Chi Jin. A simple reward-free approach to constrained reinforcement learning. In *International Conference on Machine Learning*, pages 15666–15698. PMLR, 2022.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

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

Roger B Myerson. Perspectives on mechanism design in economic theory. *American Economic Review*, 98(3):586–603, 2008.

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

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1392–1403, 2020.

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

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Yi Ouyang, Hamidreza Tavafoghi, and Demosthenis Teneketzis. Dynamic oligopoly games with private Markovian dynamics. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 5851–5858. IEEE, 2015.

Renato Paes Leme, Martin Pal, and Sergei Vassilvitskii. A field guide to personalized reserve prices. In *Proceedings of the 25th international conference on world wide web*, pages 1093–1102, 2016.

Christos Papadimitriou, George Pierrakos, Christos-Alexandros Psomas, and Aviad Rubinstein. On the complexity of dynamic mechanism design. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*, pages 1458–1475. SIAM, 2016.

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

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

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

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

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

Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.

Sheng Qiang and Mohsen Bayati. Dynamic pricing with demand covariates. *arXiv preprint arXiv:1604.07463*, 2016.

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

Alvin E Roth and Axel Ockenfels. Last-minute bidding and the rules for ending second-price auctions: Evidence from eBay and Amazon auctions on the internet. *American economic review*, 92(4):1093–1103, 2002.

Tim Roughgarden. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Mohit Sewak. *Deep reinforcement learning*. Springer, 2019.

Virag Shah, Ramesh Johari, and Jose Blanchet. Semi-parametric dynamic contextual pricing. *Advances in Neural Information Processing Systems*, 32, 2019.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.

Weiran Shen, Binghui Peng, Hanpeng Liu, Michael Zhang, Ruohan Qian, Yan Hong, Zhi Guo, Zongyao Ding, Pengjun Lu, and Pingzhong Tang. Reinforcement mechanism design: With applications to dynamic pricing in sponsored search auctions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2236–2243, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *nature*, 550(7676):354–359, 2017.

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

Steven Spielberg, Aditya Tulsyan, Nathan P Lawrence, Philip D Loewen, and R Bhushan Gopaluni. Toward self-driving processes: A deep reinforcement learning approach to control. *AIChE journal*, 65(10):e16689, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2021.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*, 2021.

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

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. Reward-free RL is no harder than reward-aware RL in linear Markov decision processes. In *International Conference on Machine Learning*, pages 22430–22456. PMLR, 2022.

Hanrui Wang, Kuan Wang, Jiacheng Yang, Linxiao Shen, Nan Sun, Hae-Seung Lee, and Song Han. GCN-RL circuit designer: Transferable transistor sizing with graph neural networks and reinforcement learning. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2020a.

Ruosong Wang, Simon S Du, Lin Yang, and Russ R Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 33:17816–17826, 2020b.

Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *Advances in Neural Information Processing Systems*, 34:13524–13536, 2021.

Yining Wang, Ruosong Wang, Simon Shaolei Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2020c.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on learning theory*, pages 4300–4354. PMLR, 2021.

Douglas J White. A survey of applications of Markov decision processes. *Journal of the operational research society*, 44(11):1073–1096, 1993.

Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I Jordan, and Haifeng Xu. Markov persuasion processes and reinforcement learning. In *ACM Conference on Economics and Computation*, 2022.

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

Xiaohong Wu, Yonggen Gu, Jie Tao, Guoqiang Li, Jingti Han, and Naixue Xiong. An effective data-driven cloud resource procurement scheme with personalized reserve prices. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(8):4693–4705, 2019.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

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

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

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

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34, 2021.

Mengxin Yu, Zhuoran Yang, and Jianqing Fan. Strategic decision-making in the presence of information asymmetry: Provably efficient RL with algorithmic instruments. *arXiv preprint arXiv:2208.11040*, 2022.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent Bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020a.

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

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

Weitong Zhang, Dongruo Zhou, and Quanquan Gu. Reward-free model-based reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34:1582–1593, 2021a.

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

Jun Zhao, Guang Qiu, Ziyu Guan, Wei Zhao, and Xiaofei He. Deep reinforcement learning for sponsored search real-time bidding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1021–1030, 2018.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C Parkes, and Richard Socher. The AI economist: Improving equality and productivity with AI-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C Parkes, and Richard Socher. The AI economist: Optimal economic policy design via two-level deep reinforcement learning. *arXiv preprint arXiv:2108.02755*, 2021.

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

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

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

Huozhi Zhou, Jinglin Chen, Lav R Varshney, and Ashish Jagmohan. Nonstationary reinforcement learning with linear function approximation. *Transactions on Machine Learning Research*, 2022.

You Zu, Krishnamurthy Iyer, and Haifeng Xu. Learning to persuade on the fly: Robustness against ignorance. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 927–928, 2021.