
Pessimism meets VCG: Learning Dynamic Mechanism Design via Offline Reinforcement Learning

Boxiang Lyu¹ Zhaoran Wang² Mladen Kolar¹ Zhuoran Yang³

Abstract

Dynamic mechanism design has garnered significant attention from both computer scientists and economists in recent years. By allowing agents to interact with the seller over multiple rounds, where agents' reward functions may change with time and are state-dependent, the framework is able to model a rich class of real-world problems. In these works, the interaction between agents and sellers is often assumed to follow a Markov Decision Process (MDP). We focus on the setting where the reward and transition functions of such an MDP are not known a priori, and we are attempting to recover the optimal mechanism using an a priori collected data set. In the setting where the function approximation is employed to handle large state spaces, with only mild assumptions on the expressiveness of the function class, we are able to design a dynamic mechanism using offline reinforcement learning algorithms. Moreover, learned mechanisms approximately have three key desiderata: efficiency, individual rationality, and truthfulness. Our algorithm is based on the pessimism principle and only requires a mild assumption on the coverage of the offline data set. To the best of our knowledge, our work provides the first offline RL algorithm for dynamic mechanism design without assuming uniform coverage.

1. Introduction

Mechanism design studies how best to allocate goods among rational agents (Maskin, 2008; Myerson, 2008; Roughgarden, 2010). Dynamic mechanism design focuses on ana-

lyzing optimal allocation rules in a changing environment, where demands for goods, the amount of available goods, and their valuations can vary over time (Bergemann & 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 & Parkes, 2003). Existing approaches in the literature require knowledge of the problem, such as the evaluation of goods by agents (Bergemann & Välimäki, 2010; Pavan et al., 2014), the transition dynamics of the system (Doepke & Townsend, 2006), or the policy that maximizes social welfare (Parkes & Singh, 2003; Parkes et al., 2004). Unfortunately, such knowledge is often not available in practice.

A practical approach we take in this paper 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 paper, 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 & Välimäki, 2019), we use offline RL with a general function approximation (Xie et al., 2021) to approximate the dynamic VCG mechanism.

Related Works. Parkes & Singh (2003) and Parkes et al. (2004) studied dynamic mechanism design from an MDP perspective. The proposed mechanisms can implement so-

¹Booth School of Business, University of Chicago, Chicago, IL, USA ²Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, USA ³Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA. Correspondence to: Boxiang Lyu <blyu@chicagobooth.edu>.

cial welfare-maximizing policies in a truth-revealing Bayes-Nash equilibrium both exactly and approximately. [Bapna & Weber \(2005\)](#) studied the dynamic auction setting from a multi-arm bandit perspective. Using the notion of marginal contribution, [Bergemann & 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 & Segal \(2013\)](#) extended both the VCG and AGV mechanisms ([d’Aspremont & 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 [Cavallaro \(2009\)](#), [Bergemann & Pavan \(2015\)](#), and [Bergemann & Välimäki \(2019\)](#) for recent surveys on dynamic mechanism design. Our paper builds on the mechanism in [Parkes \(2007\)](#) and [Bergemann & 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. \(2020\)](#) provided an algorithm that recovers the VCG mechanism in a stationary multi-arm bandit setting. [Cen & Shah \(2021\)](#), [Dai & Jordan \(2021\)](#), [Jagadeesan et al. \(2021\)](#), and [Liu et al. \(2021\)](#) 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.

Our paper 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 & Wang, 2021](#); [Uehara & 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 & Wang \(2021\)](#) proposed an instance-optimal method for tabular MDPs. [Uehara & 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 suboptimality. 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. \(2021\)](#).

Concurrent with our work, [Lyu et al. \(2022\)](#) studies the learning of a dynamic VCG mechanism in the online RL setting, where the mechanism is recovered through multiple rounds of interaction with the environment. Our work features several significant differences as we focus on general function approximation, whereas [Lyu et al. \(2022\)](#) only considers linear function approximation. We also focus on the offline RL setting, where the mechanism designer is not allowed to interact with the environment.

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 2.3 and 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 ([Carbaljal & Ely, 2013](#); [Bergemann & 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 $[z] = \{1, 2, \dots, z\}$. 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 $n \geq n_0$. We use $\tilde{\mathcal{O}}(\cdot)$ to denote $\mathcal{O}(\cdot)$ ignoring log factors. Unless stated otherwise, we use $\|\cdot\|$ to denote the ℓ_2 -norm

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 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \{r_{i,h}\}_{i=0,h=1}^{n,H})$, where \mathcal{S} is the state space, \mathcal{A} is the seller’s action space, H is the length of each episode, and $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ is the transition kernel, where $\mathbb{P}_h(s'|s, a)$ 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 $r_{i,h} : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ denote the reward function of an agent i at step h and $r_{0,h} : \mathcal{S} \times \mathcal{A} \rightarrow [-R_{\max}, -n + R_{\max}]$ 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}_π 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 $V_h^\pi(\cdot; r) : \mathcal{S} \rightarrow \mathbb{R}$ 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) $Q_h^\pi(\cdot, \cdot; r) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ 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 $g : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, 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

$$\begin{aligned} (\mathbb{T}_{h,r}^\pi g)(x, a) &= r_h(x, a) \\ &+ \mathbb{E}_\pi [g(s_{h+1}, \pi_{h+1}) | s_h = x, a_h = a], \end{aligned} \quad (2.1)$$

where \mathbb{E}_π is taken over the randomness in the transition kernel \mathbb{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

¹In mechanism design literature the reward function is often called “value function.” We use the term “reward function” throughout the paper to avoid confusion with state- and action-value functions.

reward as $\tilde{r}_{i,h}(s_h, a_h) \in [0, 1]$, which may be different from the true reward. The seller receives a reward $r_{0,h}(s, a)$ and transitions to some state $s' \sim \mathbb{P}_h(\cdot|s, a)$. 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, $\tilde{r}_{i,h}$, 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 $\tilde{r}_{i,h} : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ 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 $\tilde{r}_{i,h}$. 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 $\tilde{R}, \tilde{R}_{-i}$ 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 $\tilde{\mathcal{R}}$ be that for the reported reward functions. When all agents are truthful, $\tilde{\mathcal{R}} = \mathcal{R}$. We also let

$$\begin{aligned} Q_h^*(\cdot, \cdot; r) &= \max_{\pi \in \Pi} Q_h^\pi(\cdot, \cdot; r), \quad V_h^*(\cdot; r) = \max_{\pi \in \Pi} V_h^\pi(\cdot; r), \\ \pi_r^* &= \arg \max_{\pi \in \Pi} V_1^\pi(s_0; r), \quad \forall r \in \mathcal{R} \cup \tilde{\mathcal{R}}. \end{aligned}$$

As a shorthand notation, let $\pi^* = \pi_R^*$, $\pi_{-i}^* = \pi_{R_{-i}}^*$, $\tilde{\pi}^* = \pi_{\tilde{R}}^*$, and $\tilde{\pi}_{-i}^* = \pi_{\tilde{R}_{-i}}^*$. Following Kandasamy et al. (2020), 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

$$U_i^\pi(p_i) = \mathbb{E}_\pi \left[\sum_{h=1}^H r_{i,h}(s_h, a_h) \right] - p_i = V_1^\pi(s_0; r_i) - p_i.$$

The seller’s utility is similarly defined as

$$\begin{aligned} U_0^\pi(\{p_i\}_{i=1}^n) &= \mathbb{E}_\pi \left[\sum_{h=1}^H r_{0,h}(s_h, a_h) \right] + \sum_{i=1}^n p_i \\ &= V_1^\pi(s_0; r_0) + \sum_{i=1}^n p_i. \end{aligned}$$

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 & Välimäki (2010).

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 2.1 (Dynamic VCG Mechanism). When agents interact according to the aforementioned MDP, assuming the transition kernel \mathbb{P} and the reported reward functions $\{\tilde{r}_i\}_{i=0}^n$ are known, the VCG mechanism selects $\tilde{\pi}^*$, the social welfare maximizing policy based on the reported rewards, and charges the agent i price $p_i : \mathcal{S} \rightarrow \mathbb{R}$, given by $p_i = V_1^*(s_0; \tilde{R}_{-i}) - V_1^{\tilde{\pi}^*}(s_0; \tilde{R}_{-i})$. More generally, when the mechanism chooses to implement some arbitrary policy π , the VCG price for the agent i is given by

$$p_i = V_1^*(s_0; \tilde{R}_{-i}) - V_1^\pi(s_0; \tilde{R}_{-i}). \quad (2.2)$$

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 & 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 2.2. With \mathbb{P} and the reported rewards $\{\tilde{r}_i\}_{i=0}^n$ known, choosing $\tilde{\pi}^*$ and charging p_i for all $i \in [n]$ according to (2.2) ensures that the mechanism satisfies truthfulness, individual rationality, and efficiency simultaneously.

Proof. See Appendix B for a detailed proof. \square

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

$$\text{SubOpt}(\pi; s_0) = V_1^*(s_0; R) - V_1^\pi(s_0; R). \quad (2.3)$$

For any $i \in [n]$, let $p_i^*(s_0) = V_1^*(s_0; R_{-i}) - V_1^{\tilde{\pi}^*}(s_0; R_{-i})$ 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

$$\begin{aligned} \text{SubOpt}_i(\pi, \{p_i\}_{i=1}^n; s_0) &= U_i^{\tilde{\pi}^*}(p_i^*) - U_i^\pi(p_i) \\ &= V_1^{\tilde{\pi}^*}(s_0; r_i) - p_i^*(s_0) - V_1^\pi(s_0; r_i) + p_i, \end{aligned} \quad (2.4)$$

and the seller's suboptimality is

$$\begin{aligned} \text{SubOpt}_0(\pi, \{p_i\}_{i=1}^n; s_0) &= U_0^{\tilde{\pi}^*}(\{p_i^*\}_{i=1}^n) - U_0^\pi(\{p_i\}_{i=1}^n) \\ &= V_1^{\tilde{\pi}^*}(s_0; r_0) + \sum_{i=1}^n p_i^* - V_1^\pi(s_0; r_0) - \sum_{i=1}^n p_i. \end{aligned} \quad (2.5)$$

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, $\mathcal{D} = \{(x_h^\tau, a_h^\tau, \{\tilde{r}_{i,h}^\tau\}_{i=1}^n, x_{h+1}^\tau)\}_{h,\tau=1}^{H,K}$. 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 & Sun, 2021; Zanette et al., 2021).

Consider some general function class $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \dots \times \mathcal{F}_H$. For each $h \in [H]$, we use some arbitrary yet bounded function class $\mathcal{F}_h \subseteq \mathcal{S} \times \mathcal{A} \rightarrow [-(H-h+1)R_{\max}, (H-h+1)R_{\max}]$ to approximate $Q_h^\pi(\cdot, \cdot; r)$ for arbitrary π and $r \in \tilde{\mathcal{R}}$. 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 2.3 (Approximate Realizability). For any $r \in \tilde{\mathcal{R}}$ and $\pi \in \{\mathcal{S} \rightarrow \Delta(\mathcal{A})\}^H$, there exists some $f_r^\pi \in \mathcal{F}$ such that for all $h \in [H]$,

$$\sup_{\pi' \in \{\mathcal{S} \rightarrow \Delta(\mathcal{A})\}^H} \mathbb{E}_{\pi'_h} [\|f_{h,r}^{\pi'}(\cdot, \cdot; r) - Q_h^\pi(\cdot, \cdot; r)\|^2] \leq \epsilon_{\mathcal{F}}.$$

Intuitively, Assumption 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 2.4 (Approximate Completeness). For any $h \in [H]$, $r \in \tilde{\mathcal{R}}$, and $\pi \in \{\mathcal{S} \rightarrow \Delta(\mathcal{A})\}^H$, we have

$$\sup_{f \in \mathcal{F}_{h+1}} \inf_{f' \in \mathcal{F}_h} \mathbb{E}_{\mu_h} [\|f' - \mathbb{T}_{h,r}^\pi f\|^2] \leq \epsilon_{\mathcal{F}, \mathcal{F}}.$$

Assumption 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 & 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 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 $\pi \in \{\mathcal{S} \rightarrow \Delta(\mathcal{A})\}^H$. In particular,

$$C^\pi(\nu) = \max_{f^1, f^2 \in \mathcal{F}} \max_{h \in [H]} \max_{r \in \mathcal{R}} \frac{\mathbb{E}_{\nu_h} [\|f_h^1 - \mathbb{T}_{h,r}^\pi f_{h+1}^2\|^2]}{\mathbb{E}_{\mu_h} [\|f_h^1 - \mathbb{T}_{h,r}^\pi f_{h+1}^2\|^2]}.$$

The coefficient controls how well the Bellman estimation error shifts from one distribution to another for any Bellman transition operator \mathbb{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 $\tilde{\pi}, \tilde{p}_i$ 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 $U_i^{\tilde{\pi}}(\tilde{p}_i) - U_i^\pi(p_i) = \mathcal{O}(K^{-\alpha})$ 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.

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 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 $\tilde{\pi}$ such that the social welfare suboptimality $\text{SubOpt}(\tilde{\pi}; s_0)$ is small.
2. For all $i \in [n]$, estimate the VCG price p_i , defined in (2.2), as $\hat{p}_i = G_{-i}^{(1)}(s_0) - G_{-i}^{(2)}(s_0)$, where $G_{-i}^{(1)}(s_0)$ estimates $V_1^*(s_0; \tilde{R}_{-i})$ and $G_{-i}^{(2)}(s_0)$ estimates $V_1^{\tilde{\pi}}(s_0; \tilde{R}_{-i})$.

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 & Sun, 2021).

A greater challenge lies in implementing Step 2 and showing that the price estimates, $\{\hat{p}_i\}_{i=1}^n$, 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, $\tilde{\pi}$. The construction of $G_{-i}^{(1)}(s_0)$ is more challenging, involving two separate steps: (1) learning a fictitious policy that approximately maximizes $V_1^\pi(s_0; \tilde{R}_{-i})$ 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 & Sun, 2021; Xie et al., 2021). For example, when $G_{-i}^{(1)}(s_0)$ is a pessimistic estimate of $V_1^*(s_0; \tilde{R}_{-i})$, the price estimate \hat{p}_i is a ‘‘lower bound,’’ at least in the first term, of the actual price p_i derived in (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 & Välimäki, 2019), whereas a company wishing to maximize their profit would prefer having the agents ‘‘pay for’’ uncertainty (Friedman & Parkes, 2003).

To allow for such flexibility, we introduce hyperparameters $\zeta_1, \zeta_2 \in \{\text{PES}, \text{OPT}\}$, 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 suboptimality, we focus on the two extreme cases, $(\zeta_1, \zeta_2) = (\text{PES}, \text{OPT})$ and $(\zeta_1, \zeta_2) = (\text{OPT}, \text{PES})$, 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.

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 $f_h(s, a) - \mathbb{T}_{h,r}^\pi f_{h+1}(s, a)$ 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 $r \in \tilde{\mathcal{R}}$, we define $L_{h,r}(f_h, f'_{h+1}, \pi; \mathcal{D})$ as

$$\begin{aligned} L_{h,r}(f_h, f'_{h+1}, \pi; \mathcal{D}) &= \frac{1}{K} \sum_{\tau=1}^K (f_h(s_h^\tau, a_h^\tau) - r_h(s_h^\tau, a_h^\tau) \\ &\quad - f'_{h+1}(s_{h+1}^\tau, \pi_{h+1}))^2, \end{aligned}$$

where we slightly abuse the notation and let r_h^τ be the reported rewards $\tilde{r}_{i,h}^\tau$ summed over i according to the chosen reported reward function $r \in \tilde{\mathcal{R}}$. Recall that $\tilde{\mathcal{R}} = \{\tilde{R}_{-i}\}_{i=1}^n \cup \{\tilde{R}\}$ 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

$$\begin{aligned} B_{h,r}(f, \pi; \mathcal{D}) &= L_{h,r}(f_h, f_{h+1}, \pi; \mathcal{D}) \\ &\quad - \min_{g \in \mathcal{F}_h} L_{h,r}(g, f_{h+1}, \pi; \mathcal{D}). \end{aligned} \quad (3.1)$$

The goal of the policy evaluation algorithm is to solve the following regularized optimization problems:

$$\begin{aligned} \hat{Q}_r^\pi &= \arg \min_{f \in \mathcal{F}} -f_1(s_0, \pi) + \lambda \sum_{h=1}^H B_{h,r}(f, \pi; \mathcal{D}), \\ \check{Q}_r^\pi &= \arg \min_{f \in \mathcal{F}} f_1(s_0, \pi) + \lambda \sum_{h=1}^H B_{h,r}(f, \pi; \mathcal{D}), \end{aligned} \quad (3.2)$$

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 1.

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

Algorithm 1 Policy Evaluation

Require: Reported reward $r \in \tilde{\mathcal{R}}$, regularization coefficient λ , dataset $\mathcal{D} = \{(x_h^\tau, \omega_h^\tau, \{\tilde{r}_{i,h}^\tau\}_i^n)\}_{h,\tau=1}^{H,K}$, policy π .

- 1: For all h, τ , calculate r_h^τ as the sum of $\tilde{r}_{i,h}^\tau$ over i according to the reported reward function r .
 - 2: Obtain the optimistic and pessimistic estimates of Q_r^π using (3.2)
 - 3: Return action-value function estimates $\hat{Q}_r^\pi, \check{Q}_r^\pi$.
-

update rule can be written as

$$\hat{\pi}_{h,r}^{(t+1)}(a|s) \propto \hat{\pi}_{h,r}^{(t)}(a|s) \exp\left(\eta \hat{Q}_{h,r}^{(t)}(s, a)\right), \quad (3.3)$$

$$\check{\pi}_{h,r}^{(t+1)}(a|s) \propto \check{\pi}_{h,r}^{(t)}(a|s) \exp\left(\eta \check{Q}_{h,r}^{(t)}(s, a)\right), \quad (3.4)$$

where $\hat{Q}_{h,r}, \check{Q}_{h,r}$ are the action-value function estimates obtained from (3.2) (Bubeck, 2014; Cai et al., 2020; Xie et al., 2021).

For any set of T policies $\{\pi^{(t)}\}_{t=1}^T$, let $\text{Unif}(\{\pi^{(t)}\}_{t=1}^T)$ 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 $\text{Unif}(\{\hat{\pi}_r^{(t)}\}_{t=1}^T)$ and $\text{Unif}(\{\check{\pi}_r^{(t)}\}_{t=1}^T)$, 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 2.

Algorithm 2 Soft Policy Iteration for Episodic MDPs

Require: Reported reward $r \in \tilde{\mathcal{R}}$, regularization coefficient λ , dataset $\mathcal{D} = \{(x_h^\tau, \omega_h^\tau, \{\tilde{r}_{i,h}^\tau\}_i^n)\}_{h,\tau=1}^{H,K}$, number of iterations T , learning rate η .

- 1: Initialize optimistic and pessimistic policies, $\hat{\pi}_r^{(1)}$ and $\check{\pi}_r^{(1)}$, as the uniform policy.
 - 2: **for** $t = 1, \dots, T$ **do**
 - 3: Obtain the optimistic and pessimistic estimates of $Q_r^{\hat{\pi}_r^{(t)}}$ and $Q_r^{\check{\pi}_r^{(t)}}$ by Algorithm 1.
 - 4: Update policy estimates according to (3.3) and (3.4).
 - 5: **end for**
 - 6: Let $\hat{\pi}_r^{\text{out}} = \text{Unif}(\{\hat{\pi}_r^{(t)}\}_{t=1}^T)$, $\check{\pi}_r^{\text{out}} = \text{Unif}(\{\check{\pi}_r^{(t)}\}_{t=1}^T)$.
 - 7: Execute Algorithm 1 to construct optimistic action-value function \hat{Q}_r^{out} for $\hat{\pi}_r^{\text{out}}$ and pessimistic action-value function \check{Q}_r^{out} for $\check{\pi}_r^{\text{out}}$, respectively.
 - 8: **Return** $\{\hat{\pi}_r^{\text{out}}, \hat{Q}_r^{\text{out}}\}$ and $\{\check{\pi}_r^{\text{out}}, \check{Q}_r^{\text{out}}\}$.
-

We defer the pseudocode of our main algorithm to Appendix C in the form of Algorithm 3, as its construction is apparent given the two key subroutines above.

4. Main Results

We begin by formally defining the policy class induced by the policy improvement algorithm, Algorithm 2. It is a well-known result that policy iterates induced by mirror descent-style updates in (3.3) and (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

$$\begin{aligned} \Pi_{\text{SPI}} = & \Pi_{\text{It}} \\ & \cup \left\{ \pi : \pi = \text{Unif}(\{\pi^{(t)}\}_{t=1}^T, \{\pi^{(t)}\}_{t=1}^T \subset \Pi_{\text{It}}) \right\}. \end{aligned} \quad (4.1)$$

Before stating the main result, we introduce an additional notation. The statistical error Err^{stat} denotes

$$\begin{aligned} \text{Err}^{\text{stat}} = & \tilde{\mathcal{O}} \left(H(HR_{\max})^{5/3} K^{-1/3} \right) \\ & + \tilde{\mathcal{O}} \left(H \left((HR_{\max})^{1/3} \epsilon_{\mathcal{F}}^{1/3} + \sqrt{\epsilon_{\mathcal{F}} + \epsilon_{\mathcal{F}, \mathcal{F}}} \right) \right), \end{aligned}$$

while the optimization error Err^{opt} denotes

$$\text{Err}^{\text{opt}} = \tilde{\mathcal{O}} \left(H^2 R_{\max} \sqrt{1/T} \right).$$

To differentiate the policies learned under different truthfulness assumptions, let $\tilde{\pi} = \tilde{\pi}_R^{\text{out}}$ be the policy chosen by the algorithm when all agents are truthful, let $\tilde{\pi} = \tilde{\pi}_{r_i + \tilde{R}_{-i}}^{\text{out}}$ be the policy chosen when we only assume the agent i is truthful, and let $\tilde{\pi}_{\tilde{R}} = \tilde{\pi}_{\tilde{R}}^{\text{out}}$ be the policy chosen when no agent is truthful. Let $\tilde{\pi}^{(t)}, \tilde{\pi}_{\tilde{R}}^{(t)}, \tilde{\pi}_{r_i + \tilde{R}_{-i}}^{(t)}$ be the iterates of Algorithm 2 when learning these policies. Denote the prices charged by $\{\hat{p}_i\}_{i=1}^n, \{\tilde{p}_i\}_{i=1}^n$, and $\{\hat{p}_i, \tilde{R}_{-i}\}_{i=1}^n$, respectively.

We then summarize the performance of our learned mechanism with asymptotic bounds in Theorem 4.1. Theorem D.1 presented in Appendix D provides a more detailed result.

Theorem 4.1 (Informal). *With probability at least $1 - \delta$, with suitable choices of λ, δ , under Assumptions 2.3 and 2.4, the following claims hold simultaneously.*

1. *Algorithm 3 returns a mechanism that is asymptotically efficient. More specifically, assuming all agents report truthfully, we have*

$$\begin{aligned} \text{SubOpt}(\tilde{\pi}; s_0) \leq & \text{Err}^{\text{opt}} \\ & + \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi^*)} \right) \text{Err}^{\text{stat}}. \end{aligned}$$

2. *Assuming all agents report truthfully, when $(\zeta_1, \zeta_2) = (\text{PES}, \text{OPT})$, we have*

$$\begin{aligned} \text{SubOpt}_i(\tilde{\pi}, \{\hat{p}_i\}_{i=1}^n; s_0) \leq & \text{Err}^{\text{opt}} \\ & + \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi^*)} \right) \text{Err}^{\text{stat}}. \end{aligned}$$

When $(\zeta_1, \zeta_2) = (\text{OPT}, \text{PES})$, we have

$$\begin{aligned} \text{SubOpt}_i(\tilde{\pi}, \{\hat{p}_i\}_{i=1}^n; s_0) \leq & \text{Err}^{\text{opt}} + \text{Err}^{\text{stat}} \\ & \times \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi^*)} \right. \\ & \left. + \sqrt{C^{\tilde{\pi}_{-i}}(\tilde{\pi}_{-i})} + \sqrt{C^{\tilde{\pi}}(\tilde{\pi})} \right). \end{aligned}$$

3. *Assuming all agents report truthfully, when $(\zeta_1, \zeta_2) = (\text{PES}, \text{OPT})$, we have*

$$\begin{aligned} \text{SubOpt}_0(\tilde{\pi}, \{\hat{p}_i\}_{i=1}^n; s_0) \leq & n\text{Err}^{\text{opt}} + \text{Err}^{\text{stat}} \\ & \times \left(\sum_{i=1}^n \sqrt{C^{\tilde{\pi}_{-i}}(\tilde{\pi}_{-i})} + n\sqrt{C^{\tilde{\pi}}(\tilde{\pi})} \right. \\ & \left. + \sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}_{R_{-i}}^{(t)}}(\pi_{R_{-i}}^*)} \right). \end{aligned}$$

and, when $(\zeta_1, \zeta_2) = (\text{OPT}, \text{PES})$, we have

$$\begin{aligned} \text{SubOpt}_0(\tilde{\pi}, \{\hat{p}_i\}_{i=1}^n; s_0) \leq & n\text{Err}^{\text{opt}} + \text{Err}^{\text{stat}} \\ & \times \left(\sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T \sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\hat{\pi}_{R_{-i}}^{(t)})} \right. \\ & \left. + \sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T \sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\pi_{R_{-i}}^*)} \right). \end{aligned}$$

4. *Algorithm 3 returns a mechanism that is asymptotically individually rational. More specifically, even when other agents are untruthful, when $(\zeta_1, \zeta_2) = (\text{PES}, \text{OPT})$ and the agent i is truthful, their utility satisfies*

$$\begin{aligned} U_i^{\tilde{\pi}}(\tilde{p}_i) \geq & -\text{Err}^{\text{opt}} - \text{Err}^{\text{stat}} \\ & \times \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}_{R_{-i}}^{(t)}}(\tilde{\pi}_{R_{-i}}^*)} + \sqrt{C^{\tilde{\pi}_{R_{-i}}^{\text{out}}}(\tilde{\pi}_{R_{-i}}^{\text{out}})} \right. \\ & \left. + \frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi_{r_i + \tilde{R}_{-i}}^*)} \right). \end{aligned}$$

and when $(\zeta_1, \zeta_2) = (\text{OPT}, \text{PES})$ and the agent i is truthful, their utility satisfies

$$\begin{aligned} U_i^{\tilde{\pi}}(\tilde{p}_i) \geq & -\text{Err}^{\text{opt}} - \text{Err}^{\text{stat}} \\ & \times \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi_{r_i + \tilde{R}_{-i}}^*)} + \sqrt{C^{\tilde{\pi}_{R_{-i}}^{(t)}}(\tilde{\pi}_{R_{-i}}^*)} \right. \\ & \left. + \frac{1}{T} \sum_{t=1}^T \sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\hat{\pi}_{R_{-i}}^{(t)})} + \sqrt{C^{\tilde{\pi}}(\tilde{\pi})} \right). \end{aligned}$$

5. *Algorithm 3 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 $\zeta_2 = \text{OPT}$ the amount of utility gained by untruthful reporting is upper bounded as*

$$U_i^{\tilde{\pi}_{\tilde{R}}}(\hat{p}_i, \tilde{R}) - U_i^{\tilde{\pi}}(\tilde{p}_i) \leq \text{Err}^{\text{opt}} + \text{Err}^{\text{stat}} \times \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi_{r_i + \tilde{R}-i}^*)} + \sqrt{C^{\tilde{\pi}_{\tilde{R}}}(\tilde{\pi}_{\tilde{R}})} \right),$$

and when $\zeta_2 = \text{PES}$, the amount of utility gained by untruthful reporting is upper bounded as

$$U_i^{\tilde{\pi}_{\tilde{R}}}(\hat{p}_i, \tilde{R}) - U_i^{\tilde{\pi}}(\tilde{p}_i) \leq \text{Err}^{\text{opt}} + \text{Err}^{\text{stat}} \times \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi_{r_i + \tilde{R}-i}^*)} + \sqrt{C^{\tilde{\pi}}(\tilde{\pi})} \right).$$

Proof. See Appendix D for a detailed proof. \square

We make a few remarks about Theorem 4.1.

Dependence on the number of trajectories K . The only term that depends on the number of trajectories K is the statistical error Err^{stat} and it decays at the $\tilde{O}(K^{-1/3})$ 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 2.3 and 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 4.1 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 $\zeta_1 = \text{OPT}$. 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 F and the theorem's dependence on the covering number is made explicit in the non-asymptotic version, Theorem D.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., 2021) 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.

5. Proof Sketch

To prove the results in Theorem 4.1, we need to first analyze the concentration properties of the empirical Bellman error estimate, $B_{h,r}(f, \pi; \mathcal{D})$. As the function approximation class \mathcal{F} and the policy class Π often contains infinite elements, it is crucial that the tail bounds we obtain remain finite even when both $|\mathcal{F}|$ and $|\Pi|$ are infinite.

We begin by sketching out the concentration bounds for $B_{h,r}(f, \pi, \mathcal{D})$. Consider some arbitrary and fixed $h \in [H]$ and $r \in \tilde{\mathcal{R}}$. Let Z be the random vector $(s_h, a_h, r_h(s_h, a_h), s_{h+1})$, where $(s_h, a_h, s_{h+1}) \sim \mu_h$ and Z_j 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

$$g_{f,f'}^\pi(Z) = (f_h(s_h, a_h) - r_h - f'_{h+1}(s_{h+1}, \pi_{h+1}))^2 - (\mathbb{T}_{h,r}^\pi f'_{h+1}(s_h, a_h) - r_h - f'_{h+1}(s_{h+1}, \pi_{h+1}))^2, \quad (5.1)$$

and $g_{f,f'}^\pi(Z_j)$ its empirical counterpart evaluated on Z 's realization, Z_j . Recalling the definition of the Bellman transition operator $\mathbb{T}_{h,r}^\pi$, we can show that

$$\mathbb{E}_{Z \sim \mu_h} [g_{f,f'}^\pi(Z)] = \|f_h - \mathbb{T}_{h,r}^\pi f'_{h+1}\|_{2, \mu_h}^2.$$

The boundedness of functions in \mathcal{F} and reward functions $r \in \tilde{\mathcal{R}}$ ensure that

$$\text{Var}(g_{f,f'}^\pi(Z)) \leq 16H^2 R_{\max}^2 \|f_h - \mathbb{T}_{h,r}^\pi f'_{h+1}\|_{2, \mu_h}^2.$$

With both the expectation and variance bounded, we can derive a tail bound for the realizations $g_{f,f'}^\pi(Z_j)$, thereby

ensuring $\frac{1}{K} \sum_{j=1}^K g_{f,f'}^\pi(Z_j)$ is sufficiently close to $\|f_h - \mathbb{T}_{h,r}^\pi f_{h+1}'\|_{2,\mu_h}^2$ for a specific choice of $f, f' \in \mathcal{F}$ and $\pi \in \Pi$.

We then focus on the function $g_{f,f'}^\pi$ itself. Let $\mathcal{G}_{\mathcal{F},\Pi} = \{g_{f,f'}^\pi : f, f' \in \mathcal{F}, \pi \in \Pi\}$. Examining the definition of $g_{f,f'}^\pi(Z)$ in (5.1), we can directly control the covering number of $\mathcal{G}_{\mathcal{F},\Pi}$ using covering numbers of \mathcal{F}, Π , more formally introduced in Appendix C. Using a standard covering argument, we obtain a tail bound for $g_{f,f'}^\pi(Z)$ for all possible choices of $f, f' \in \mathcal{F}$ and $\pi \in \Pi$, even when both \mathcal{F} and Π are infinite, via the covering numbers of \mathcal{F} and Π .

Finally, we notice that $\frac{1}{K} \sum_{j=1}^K g_{f,f'}^\pi(Z_j)$ is close to $B_{h,r}(f, \pi; \mathcal{D})$ under Assumptions 2.3 and 2.4, linking the concentration behavior of $\frac{1}{K} \sum_{j=1}^K g_{f,f'}^\pi(Z_j)$ to the empirical losses $B_{h,r}(f, \pi; \mathcal{D})$ we observe.

5.1. Seller Suboptimality

We now sketch the proof for bounding the seller's optimality to provide some intuition on how to prove Theorem 4.1. Equation (D.4), given in the appendix, bounds $\text{SubOpt}_0(\tilde{\pi}, \{\hat{p}_i\}_{i=1}^n; s_0)$ as

$$\begin{aligned} & \text{SubOpt}_0(\tilde{\pi}, \{\hat{p}_i\}_{i=1}^n; s_0) \\ & \leq \sum_{i=1}^n \left(V_1^{\pi^*}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0) \right) \\ & \quad + \sum_{i=1}^n \left(G_{-i}^{(2)}(s_0) - V_1^{\tilde{\pi}}(s_0, R_{-i}) \right). \end{aligned}$$

The second term corresponds to the error bound of Algorithm 1. When $\zeta_2 = \text{OPT}$, the term exactly corresponds to the classic function evaluation error of the upper confidence bound methods. As such, it can be bounded using a combination of the distribution shift coefficient $C^{\tilde{\pi}}(\tilde{\pi})$ and the fact that $\hat{Q}_{R_{-i}}^{\tilde{\pi}}$ minimizes (3.2). When $\zeta_2 = \text{PES}$, we bound the term using the fact that the output of our policy evaluation algorithm is approximately pessimistic, similar to Lemma C.6 of Xie et al. (2021).

Next, we focus on the first term $G_{-i}^{(1)}(s_0) - V_1^{\pi^*}(s_0; R_{-i})$. When $\zeta_1 = \text{OPT}$, we use the following decomposition

$$\begin{aligned} & G_{-i}^{(1)}(s_0) - V_1^{\pi^*}(s_0; R_{-i}) \\ & = V_1^{\pi^*}(s_0; R_{-i}) - \frac{1}{T} \sum_{t=1}^T \hat{Q}_{1,R_{-i}}^{(t)}(s_0, \hat{\pi}_{1,R_{-i}}^{(t)}) \\ & \quad + \frac{1}{T} \sum_{t=1}^T \left(\hat{Q}_{1,R_{-i}}^{(t)}(s_0, \hat{\pi}_{1,R_{-i}}^{(t)}) - V_1^{\hat{\pi}_{R_{-i}}^{(t)}}(s_0; R_{-i}) \right) \\ & \quad + V_1^{\hat{\pi}_{R_{-i}}^{(t)}}(s_0; R_{-i}) - \hat{Q}_{1,R_{-i}}^{\text{out}}(s_0, \hat{\pi}_{1,-i}). \end{aligned}$$

The first term can be bounded using the properties of mirror descent (Bubeck, 2014). The latter two terms are function

evaluation errors, which we can bound in a similar way as $G_{-i}^{(2)}(s_0) - V_1^{\tilde{\pi}}(s_0, R_{-i})$. The first term can be similarly bounded when $\zeta_1 = \text{PES}$, completing the proof sketch.

6. Discussion

Our work provides the first algorithm that can provably learn the dynamic VCG mechanism with no prior knowledge, where the learned mechanism is asymptotically efficient, individually rational, and truthful. For future work, we aim to study the performance of our algorithm when the training set is corrupted with untruthful reports.

7. Acknowledgements

Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Simons Institute (Theory of Reinforcement Learning), Amazon, J.P. Morgan, and Two Sigma for their supports. Mladen Kolar acknowledges the William S. Fishman Faculty Research Fund at the University of Chicago Booth School of Business. Zhuoran Yang acknowledges Simons Institute (Theory of Reinforcement Learning).

References

- Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.
- Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.
- Athey, S. and Segal, I. An efficient dynamic mechanism. *Econometrica*, 81(6):2463–2485, 2013.
- Balcan, M.-F., Blum, A., Hartline, J. D., and Mansour, Y. Reducing mechanism design to algorithm design via machine learning. *Journal of Computer and System Sciences*, 74(8):1245–1270, 2008.
- Bapna, A. and Weber, T. A. Efficient dynamic allocation with uncertain valuations. *Available at SSRN 874770*, 2005.
- Bergemann, D. and Pavan, A. Introduction to symposium on dynamic contracts and mechanism design. *Journal of Economic Theory*, 159:679–701, 2015.
- Bergemann, D. and Välimäki, J. Efficient dynamic auctions. Technical report, Cowles Foundation for Research in Economics, Yale University, 2006.

- Bergemann, D. and Välimäki, J. The dynamic pivot mechanism. *Econometrica*, 78(2):771–789, 2010.
- Bergemann, D. and Välimäki, J. Dynamic mechanism design: An introduction. *Journal of Economic Literature*, 57(2):235–74, 2019.
- Bubeck, S. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.
- Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.
- Carbajal, J. C. and Ely, J. C. Mechanism design without revenue equivalence. *Journal of Economic Theory*, 148(1):104–133, 2013.
- Cavallo, R. Mechanism design for dynamic settings. *ACM SIGecom Exchanges*, 8(2):1–5, 2009.
- Cen, S. H. and Shah, D. Regret, stability, and fairness in matching markets with bandit learners. *arXiv preprint arXiv:2102.06246*, 2021.
- Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34, 2021.
- Clarke, E. H. Multipart pricing of public goods. *Public choice*, pp. 17–33, 1971.
- Dai, X. and Jordan, M. I. Learning strategies in decentralized matching markets under uncertain preferences. *Journal of Machine Learning Research*, 22(260):1–50, 2021.
- d’Aspremont, C. and Gérard-Varet, L.-A. Incentives and incomplete information. *Journal of Public economics*, 11(1):25–45, 1979.
- Doepke, M. and Townsend, R. M. Dynamic mechanism design with hidden income and hidden actions. *Journal of Economic Theory*, 126(1):235–285, 2006.
- Friedman, E. J. and Parkes, D. C. Pricing WiFi at Starbucks: issues in online mechanism design. In *Proceedings of the 4th ACM conference on Electronic commerce*, pp. 240–241, 2003.
- Gallien, J. Dynamic mechanism design for online commerce. *Operations Research*, 54(2):291–310, 2006.
- Gerding, E. H., Robu, V., Stein, S., Parkes, D. C., Rogers, A., and Jennings, N. R. Online mechanism design for electric vehicle charging. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pp. 811–818, 2011.
- Groves, T. Efficient collective choice when compensation is possible. *The Review of Economic Studies*, 46(2):227–241, 1979.
- Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.
- Hartline, J. D. Bayesian mechanism design. *Theoretical Computer Science*, 8(3):143–263, 2012.
- Jagadeesan, M., Wei, A., Wang, Y., Jordan, M., and Steinhardt, J. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34, 2021.
- Jin, C., Liu, Q., and Miryoosefi, S. Bellman Eluder dimension: New rich classes of RL problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*, 2021a.
- Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021b.
- Kakade, S. M., Lobel, I., and Nazerzadeh, H. Optimal dynamic mechanism design and the virtual-pivot mechanism. *Operations Research*, 61(4):837–854, 2013.
- Kandasamy, K., Gonzalez, J. E., Jordan, M. I., and Stoica, I. Mechanism design with bandit feedback. *arXiv preprint arXiv:2004.08924*, 2020.
- Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.
- Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative Q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.
- Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.
- Littman, M. L. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.
- Liu, L. T., Ruan, F., Mania, H., and Jordan, M. I. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22(211):1–34, 2021.
- Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

- Lyu, B., Meng, Q., Qiu, S., Wang, Z., Yang, Z., and Jordan, M. I. Learning dynamic mechanisms in unknown environments: A reinforcement learning approach. *arXiv preprint arXiv:2202.12797*, 2022.
- Maskin, E. S. Mechanism design: How to implement social goals. *American Economic Review*, 98(3):567–76, 2008.
- Myerson, R. B. Perspectives on mechanism design in economic theory. *American Economic Review*, 98(3):586–603, 2008.
- Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V. *Algorithmic Game Theory*. Cambridge University Press, 2007.
- Parkes, D. C. Online mechanisms. In Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. (eds.), *Algorithmic Game Theory*, pp. 411–439. Cambridge University Press, 2007.
- Parkes, D. C. and Singh, S. An MDP-based approach to online mechanism design. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pp. 791–798, 2003.
- Parkes, D. C., Singh, S., and Yanovsky, D. Approximately efficient online mechanism design. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, pp. 1049–1056, 2004.
- Pavan, A., Segal, I. R., and Toikka, J. Dynamic mechanism design: Incentive compatibility, profit maximization and information disclosure. *Profit Maximization and Information Disclosure (May 1, 2009)*, 2009.
- Pavan, A., Segal, I., and Toikka, J. Dynamic mechanism design: A Myersonian approach. *Econometrica*, 82(2): 601–653, 2014.
- Roughgarden, T. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.
- Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.
- Vickrey, W. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.
- Xie, T. and Jiang, N. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.
- Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.
- Yin, M. and Wang, Y.-X. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34, 2021.
- Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.
- Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., and Finn, C. Combo: Conservative offline model-based policy optimization. *Advances in Neural Information Processing Systems*, 34, 2021.
- Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. *arXiv preprint arXiv:2108.08812*, 2021.

A. Table of Notation

The following table summarizes the notation used in the paper.

Notation	Meaning
$r_{i,h}/\tilde{r}_{i,h}$	actual / reported reward function for agent i at step $h \in [H]$
$R_{-i,h}/(\tilde{R}_{-i,h})$	actual / reported sum of reward function across all participants sans agent i
$R_h/(\tilde{R}_h)$	actual / reported sum of reward functions across all participants
$\mathcal{R}/\tilde{\mathcal{R}}$	actual / reported reward functions of interest.
π_h	the policy taken by the seller at step $h \in [H]$
$\mathbb{T}_{h,r}^\pi$	policy specific Bellman transition operator
$C^\pi(\nu)$	Distribution shift coefficient (see Definition 2.5)
$C^{\pi_1}(\pi_2)$	Shorthand notation for $C^{\pi_1}(d_{\pi_2})$
$\hat{\pi}_{h,r}^{(t)}/(\check{\pi}_{h,r}^{(t)})$	optimistic / pessimistic policy estimate at the t -th iteration of Algorithm 2 with input $r \in \tilde{\mathcal{R}}$
$\hat{Q}_{h,r}^{(t)}/(\check{Q}_{h,r}^{(t)})$	optimistic / pessimistic action-value function estimate at the t -th iteration of Algorithm 2 with input $r \in \tilde{\mathcal{R}}$. Shorthand for $\hat{Q}_{h,r}^{\hat{\pi}_{h,r}^{(t)}}(\check{Q}_{h,r}^{\check{\pi}_{h,r}^{(t)}})$
$\hat{\pi}_{h,r}^{\text{out}}/(\check{\pi}_{h,r}^{\text{out}})$	optimistic / pessimistic policy output of Algorithm 2 with input $r \in \tilde{\mathcal{R}}$
$\hat{Q}_{h,r}^{\text{out}}/(\check{Q}_{h,r}^{\text{out}})$	optimistic / pessimistic action-value function estimate output of Algorithm 2 with input $r \in \tilde{\mathcal{R}}$. Shorthand for $\hat{Q}_{h,r}^{\hat{\pi}_{h,r}^{\text{out}}}(\check{Q}_{h,r}^{\check{\pi}_{h,r}^{\text{out}}})$

B. Proof of Mechanism Design Desiderata (Proposition 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 \mathbb{P} and $\{\tilde{r}_i\}_{i=0}^n$ given, the state-value functions $V_h^\pi(s_0, r)$ can be explicitly calculated for all $h \in [H], r \in \tilde{\mathcal{R}}$. We can then obtain exactly $\tilde{\pi}^*$ and directly calculate $p_i = V_1^*(s_0, \tilde{R}_{-i}) - V_1^{\tilde{\pi}^*}(s_0, \tilde{R}_{-i})$. Thus, the proposed mechanism is feasible when the rewards and transition kernel are known.

For convenience, let

$$\pi^{(1)} = \pi_{r_i + \tilde{R}_{-i}}^* = \arg \max_{\pi \in \Pi} V_1^\pi(s_0; r_i + \tilde{R}_{-i}) \quad \text{and} \quad \pi^{(2)} = \pi_{\tilde{R}}^* = \arg \max_{\pi \in \Pi} V_1^\pi(s_0; \tilde{R}),$$

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

$$p_i = V_1^*(s_0; \tilde{R}_{-i}) - V_1^{\pi^{(2)}}(s_0; \tilde{R}_{-i}).$$

Our goal is to then show that $V_1^{\pi^{(2)}}(s_0; \tilde{r}_i) \geq p_i$. That is, the value function of the *reported* reward is no less than the

price charged. Observe that

$$V_1^{\pi^{(2)}}(s_0; \tilde{r}_i) - \tilde{p}_i = V_1^{\pi^{(2)}}(s_0; \tilde{R}) - V_1^*(s_0; \tilde{R}_{-i}).$$

Let $\pi_{-i}^{(2)} = \arg \max_{\pi \in \Pi} V_1^\pi(s_0; \tilde{R}_{-i})$. Then we know that

$$V_1^{\pi^{(2)}}(s_0; \tilde{r}_i) - \tilde{p}_i \geq V_1^{\pi_{-i}^{(2)}}(s_0; \tilde{R}) - V_1^{\pi_{-i}^{(2)}}(s_0; \tilde{R}_{-i}) = V_1^{\pi_{-i}^{(2)}}(s_0; \tilde{r}_i) \geq 0.$$

3. Truthfulness: If $\tilde{r}_i = r_i$, that is, the agent i reports truthfully, they attain the following utility

$$U_i^{\pi^{(1)}}(p_i) = V_1^{\pi^{(1)}}(s_0; r_i) - V_1^*(s_0; \tilde{R}_{-i}) + V_1^{\pi^{(1)}}(s_0; \tilde{R}_{-i}) = V_1^{\pi^{(1)}}(s_0; r_i + \tilde{R}_{-i}) - V_1^*(s_0; \tilde{R}_{-i}).$$

When the agent reports some arbitrary \tilde{r}_i , the agent receives the following utility instead

$$U_i^{\pi^{(2)}}(p_i) = V_1^{\pi^{(2)}}(s_0; r_i) - V_1^*(s_0; \tilde{R}_{-i}) + V_1^{\pi^{(2)}}(s_0; \tilde{R}_{-i}) = V_1^{\pi^{(2)}}(s_0; r_i + \tilde{R}_{-i}) - V_1^*(s_0; \tilde{R}_{-i}).$$

Since $\pi^{(1)}$ maximizes $V_1^\pi(s_0; r_i + \tilde{R}_{-i})$, $u_i \geq \tilde{u}_i$ regardless of other agents' reported reward $\{\tilde{r}_j\}_{j \neq i}$ and the mechanism is truthful.

C. 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

$$\max_{h \in [H]} \sup_{s \in \mathcal{S}, a \in \mathcal{A}} |f_h^l(s, a) - f_h(s, a)| \leq \epsilon.$$

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

$$\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^2} \right)^{1/3}, \quad \eta = \sqrt{\frac{\log |\mathcal{A}|}{2H^2 R_{\max}^2 T}}, \quad (\text{C.1})$$

where

$$\epsilon_{\mathcal{S}} = \frac{5136}{K} H^4 R_{\max}^4 \log \left(56nH \cdot \mathcal{N}_\infty \left(\frac{19H^3 R_{\max}^3}{K}, \mathcal{F} \right) \cdot \mathcal{N}_{\infty,1} \left(\frac{19H^4 R_{\max}^4}{K}, \Pi_{\text{SPI}} \right) / \delta \right).$$

The pseudocode for our main algorithm can then be summarized as Algorithm 3.

D. Proof of Theorem 4.1

We re-state Theorem 4.1 in a finite sample form.

Theorem D.1 (Theorem 4.1 restated). *Suppose that λ, η are set according to (C.1) and Assumptions 2.3 and 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 $\tilde{\pi}$ is bounded as

$$\begin{aligned} \text{SubOpt}(\tilde{\pi}; s_0) &\leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3} (\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi^*)} \right) \left(2(HR_{\max})^{1/3} (\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

Algorithm 3 Offline VCG Learn

Require: Hyperparameters $\zeta_1, \zeta_2 \in \{\text{OPT}, \text{PES}\}$, regularization coefficient λ , number of iterations T , learning rate η .

- 1: Let $\tilde{\pi}_{\tilde{R}}^{\text{out}}$ be the pessimistic policy output of Algorithm 2 with $r = \tilde{R}$, T , and λ, η set according to (C.1).
- 2: **for** Agent $i = 1, 2, \dots, n$ **do**
- 3: Call Algorithm 2 with $r = \tilde{R}_{-i}$, T , and λ, η set according to (C.1).
- 4: If $\zeta_1 = \text{OPT}$, let $G_{-i}^{(1)}(s_0) = \hat{Q}_{1, \tilde{R}_{-i}}^{\text{out}}(s_0, \hat{\pi}_{1, \tilde{R}_{-i}}^{\text{out}})$. Otherwise let $G_{-i}^{(1)}(s_0) = \check{Q}_{1, \tilde{R}_{-i}}^{\text{out}}(s_0, \check{\pi}_{1, \tilde{R}_{-i}}^{\text{out}})$.
- 5: Call Algorithm 1 with $r = \tilde{R}_{-i}$, $\pi = \tilde{\pi}_{\tilde{R}}^{\text{out}}$, and λ set according to (C.1).
- 6: If $\zeta_2 = \text{OPT}$, let $G_{-i}^{(2)}(s_0) = \hat{Q}_{1, \tilde{R}_{-i}}^{\tilde{\pi}_{\tilde{R}}^{\text{out}}}(s_0, \tilde{\pi}_{1, \tilde{R}}^{\text{out}})$.
 Otherwise let $G_{-i}^{(2)}(s_0) = \check{Q}_{1, \tilde{R}_{-i}}^{\tilde{\pi}_{\tilde{R}}^{\text{out}}}(s_0, \tilde{\pi}_{1, \tilde{R}}^{\text{out}})$.
- 7: Set the estimated price $\hat{p}_i = G_{-i}^{(1)}(s_0) - G_{-i}^{(2)}(s_0)$.
- 8: **end for**
- 9: Return policy $\tilde{\pi}_{\tilde{R}}^{\text{out}}$ and estimated prices $\{\hat{p}_i\}_{i=1}^n$.

2. Assuming all agents report truthfully, when $(\zeta_1, \zeta_2) = (\text{PES}, \text{OPT})$, the agent i 's suboptimality, for all $i \in [n]$, satisfies

$$\begin{aligned} \text{SubOpt}_i(\tilde{\pi}, \{\hat{p}_i\}_{i=1}^n; s_0) &\leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} + 3\sqrt{\epsilon_{\mathcal{F}}} + 6(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi^*)} \right) \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F}, \mathcal{F}}} \right), \end{aligned}$$

and when $(\zeta_1, \zeta_2) = (\text{OPT}, \text{PES})$, the agent i 's suboptimality, for all $i \in [n]$, satisfies

$$\begin{aligned} \text{SubOpt}_i(\tilde{\pi}, \{\hat{p}_i\}_{i=1}^n; s_0) &\leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi^*)} + \sqrt{C^{\tilde{\pi}_{-i}}(\hat{\pi}_{-i})} + \sqrt{C^{\tilde{\pi}}(\tilde{\pi})} \right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F}, \mathcal{F}}} \right). \end{aligned}$$

3. Assuming all agents report truthfully, when $(\zeta_1, \zeta_2) = (\text{PES}, \text{OPT})$, the seller's suboptimality satisfies

$$\begin{aligned} \text{SubOpt}_0(\tilde{\pi}, \{\hat{p}_i\}_{i=1}^n; s_0) &\leq 2nH^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} + n\sqrt{\epsilon_{\mathcal{F}}} + 2n(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H \left(\sum_{i=1}^n \left(\sqrt{C^{\tilde{\pi}_{-i}}(\tilde{\pi}_{-i})} + \frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}_{R_{-i}}^{(t)}}(\pi_{-i}^*)} \right) + n\sqrt{C^{\tilde{\pi}}(\tilde{\pi})} \right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F}, \mathcal{F}}} \right), \end{aligned}$$

and when $(\zeta_1, \zeta_2) = (\text{OPT}, \text{PES})$, the seller's suboptimality satisfies

$$\begin{aligned} \text{SubOpt}_0(\tilde{\pi}, \{\hat{p}_i\}_{i=1}^n; s_0) &\leq 2nH^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} + 2n\sqrt{\epsilon_{\mathcal{F}}} + 4n(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H \left(\sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T \left(\sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\pi_{-i}^*)} + \sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\hat{\pi}_{R_{-i}}^{(t)})} \right) \right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F}, \mathcal{F}}} \right). \end{aligned}$$

4. (Asymptotic Individual Rationality) Even when other agents are untruthful, when $(\zeta_1, \zeta_2) = (\text{PES}, \text{OPT})$ and the agent i

is truthful, their utility is lower bounded by

$$\begin{aligned} U_i^{\tilde{\pi}}(\tilde{p}_i) &\geq -4H^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} - 3\sqrt{\epsilon_{\mathcal{F}}} - 6(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad - H \left(\frac{1}{T} \sum_{t=1}^T \left(\sqrt{C^{\tilde{\pi}^{(t)}}(\pi_{r_i+\tilde{R}_{-i}}^*)} + \sqrt{C^{\tilde{\pi}_{\tilde{R}_{-i}}^{(t)}}(\tilde{\pi}_{-i}^*)} \right) + \sqrt{C^{\tilde{\pi}_{\tilde{R}_{-i}}^{\text{out}}}(\tilde{\pi}_{\tilde{R}_{-i}}^{\text{out}})} \right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right), \end{aligned}$$

and when $(\zeta_1, \zeta_2) = (\text{OPT}, \text{PES})$, their utility is lower bounded by

$$\begin{aligned} U_i^{\tilde{\pi}}(\tilde{p}_i) &\geq -4H^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} - 2\sqrt{\epsilon_{\mathcal{F}}} - 4(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad - H \left(\frac{1}{T} \sum_{t=1}^T \left(\sqrt{C^{\tilde{\pi}^{(t)}}(\pi_{r_i+\tilde{R}_{-i}}^*)} + \sqrt{C^{\hat{\pi}_{\tilde{R}_{-i}}^{(t)}}(\tilde{\pi}_{-i}^*)} + \sqrt{C^{\hat{\pi}_{\tilde{R}_{-i}}^{(t)}}(\hat{\pi}_{\tilde{R}_{-i}}^{(t)})} \right) + \sqrt{C^{\tilde{\pi}}(\tilde{\pi})} \right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

5. (Asymptotic Truthfulness) Even when all the other agents are untruthful and irrespective of whether the agent i is truthful or not, when $\zeta_2 = \text{OPT}$, the amount of utility gained by untruthful reporting is upper bounded by

$$\begin{aligned} U_i^{\tilde{\pi}_{\tilde{R}}}(\hat{p}_{i,\tilde{R}}) - U_i^{\tilde{\pi}}(\tilde{p}_i) &\leq 2H^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + 2\sqrt{\epsilon_{\mathcal{F}}} + 4(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H \left(\frac{1}{T} \sum_{t=1}^T \left(\sqrt{C^{\tilde{\pi}^{(t)}}(\pi_{r_i+\tilde{R}_{-i}}^*)} + \sqrt{C^{\tilde{\pi}_{\tilde{R}}}(\tilde{\pi}_{\tilde{R}})} \right) \right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right), \end{aligned}$$

and when $\zeta_2 = \text{PES}$, the amount of utility gained by untruthful reporting is upper bounded by

$$\begin{aligned} U_i^{\tilde{\pi}_{\tilde{R}}}(\hat{p}_{i,\tilde{R}}) - U_i^{\tilde{\pi}}(\tilde{p}_i) &\leq 2H^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + 2\sqrt{\epsilon_{\mathcal{F}}} + 4(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H \left(\frac{1}{T} \sum_{t=1}^T \left(\sqrt{C^{\tilde{\pi}^{(t)}}(\pi_{r_i+\tilde{R}_{-i}}^*)} + \sqrt{C^{\tilde{\pi}}(\tilde{\pi})} \right) \right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

Proof of Theorem D.1. We will make use of the following concentration lemma.

Lemma D.2. For any fixed $h \in [H]$, $r \in \tilde{\mathcal{R}}$, and any policy class $\Pi \subset \{\mathcal{S} \rightarrow \Delta(\mathcal{A})\}^H$ we have

$$\begin{aligned} \Pr(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \\ &\quad \left| \mathbb{E}_{\mu_h} [\|f_h - \mathbb{T}_{h,r}^{\pi} f'_{h+1}\|^2] - L_{h,r}(f_h, f'_{h+1}, \pi; \mathcal{D}) + L_{h,r}(\mathbb{T}_{h,r}^{\pi} f'_{h+1}, f'_{h+1}, \pi; \mathcal{D}) \right| \\ &\quad \geq \epsilon (\alpha + \beta + \mathbb{E}_{\mu_h} [\|f_h - \mathbb{T}_{h,r}^{\pi} f'_{h+1}\|^2])) \\ &\leq 28 \left(\mathcal{N}_{\infty} \left(\frac{\epsilon\beta}{140HR_{\max}}, \mathcal{F} \right) \right)^2 \mathcal{N}_{\infty,1} \left(\frac{\epsilon\beta}{140H^2R_{\max}^2}, \Pi \right) \exp \left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{214(1+\epsilon)H^4R_{\max}^4} \right). \end{aligned}$$

for all $\alpha, \beta > 0$, $0 < \epsilon \leq 1/2$.

Proof. See Section F.1 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 D.3. For any policy class $\Pi \subset \{\mathcal{S} \rightarrow \Delta(\mathcal{A})\}^H$, let the “good event” $\mathcal{G}(\Pi)$ be defined as

$$\begin{aligned} \mathcal{G}(\Pi) = \{ \forall h \in [H], r \in \tilde{\mathcal{R}}, \pi \in \Pi, f, f' \in \mathcal{F} : \\ |\mathbb{E}_{\mu_h} [\|f_h - \mathbb{T}_{h,r}^{\pi} f'_{h+1}\|^2] - L_{h,r}(f_h, f'_{h+1}, \pi; \mathcal{D}) + L_{h,r}(\mathbb{T}_{h,r}^{\pi} f'_{h+1}, f'_{h+1}, \pi; \mathcal{D})| \\ \leq \epsilon_S + \frac{1}{2} \mathbb{E}_{\mu_h} [\|f_h - \mathbb{T}_{h,r}^{\pi} f'_{h+1}\|^2] \}, \end{aligned} \quad (\text{D.1})$$

where

$$\epsilon_S = \frac{5136}{K} H^4 R_{\max}^4 \log \left(56nH \cdot \mathcal{N}_{\infty} \left(\frac{19H^3 R_{\max}^3}{K}, \mathcal{F} \right) \cdot \mathcal{N}_{\infty,1} \left(\frac{19H^4 R_{\max}^4}{K}, \Pi \right) / \delta \right). \quad (\text{D.2})$$

Then $\mathcal{G}(\Pi)$ occurs with probability at least $1 - \delta$.

Proof. See Section F.2 for a detailed proof. \square

On the event $\mathcal{G}(\Pi)$, the best approximations of action-value functions, defined according to Assumption 2.3, have small empirical Bellman error estimates.

Corollary D.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 $Q_r^{\pi}(\cdot, \cdot; r)$ as defined in Assumption 2.3, $\pi \in \Pi$ and $r \in \tilde{\mathcal{R}}$. Then, for all $h \in [H]$, we have

$$B_{h,r}(f_r^{\pi,*}, \pi; \mathcal{D}) \leq 2\epsilon_S + 6\epsilon_{\mathcal{F}}.$$

Proof. See Section F.2 for a detailed proof. \square

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 D.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 $r \in \tilde{\mathcal{R}}$, $\pi \in \Pi$, $f \in \mathcal{F}$, if $B_{h,r}(f, \pi; \mathcal{D}) \leq \epsilon_0$, then

$$\mathbb{E}_{\mu_h} [\|f_h - \mathbb{T}_{h,r}^{\pi} f_{h+1}\|^2] \leq 2\epsilon_0 + 4\epsilon_S + 3\epsilon_{\mathcal{F},\mathcal{F}}.$$

Proof. See Section F.2 for a detailed proof. \square

We introduce the key properties of Algorithms 1 and 2 that we will use. The following lemma states that the outputs of Algorithm 1 are approximately optimistic and pessimistic.

Lemma D.6. For any $\pi = \{\pi_h\}_{h=1}^H \in \Pi_{\text{SPI}}$, reported reward $r \in \tilde{\mathcal{R}}$, and λ , conditioned on the event $\mathcal{G}(\Pi_{\text{SPI}})$, the following holds simultaneously for optimistic and pessimistic outputs of Algorithm 1:

1. $\check{Q}_{1,r}^{\pi}(s_0, \pi_1) + \lambda \sum_{h=1}^H B_{h,r}(\check{Q}_r^{\pi}, \pi; \mathcal{D}) \leq Q_1^{\pi}(s_0, \pi_1; r) + \sqrt{\epsilon_{\mathcal{F}}} + 2\lambda H \epsilon_S + 6\lambda H \epsilon_{\mathcal{F}};$
2. $\hat{Q}_{1,r}^{\pi}(s_0, \pi_1) - \lambda \sum_{h=1}^H B_{h,r}(\hat{Q}_r^{\pi}, \pi; \mathcal{D}) \geq Q_1^{\pi}(s_0, \pi_1; r) - \sqrt{\epsilon_{\mathcal{F}}} - 2\lambda H \epsilon_S - 6\lambda H \epsilon_{\mathcal{F}}.$

Proof. See Section E.1 for a detailed proof. \square

Additionally, the estimates given by Algorithm 1 are sufficiently good estimates of the ground truth action-value functions.

Lemma D.7. For any input $\pi = \{\pi_h\}_{h=1}^H \in \Pi_{\text{SPI}}$, reported reward $r \in \tilde{\mathcal{R}}$, when $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_S + 3\epsilon_{\mathcal{F}})^2} \right)^{1/3}$ and the event $\mathcal{G}(\Pi_{\text{SPI}})$ holds, the outputs of Algorithm 1 satisfy:

1. $Q_1^{\pi}(s_0, \pi_1; r) - \check{Q}_{1,r}^{\pi}(s_0, \pi_1) \leq H \sqrt{C^{\pi}(\pi)} (2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}});$

$$2. \hat{Q}_{1,r}^\pi(s_0, \pi_1) - Q_1^\pi(s_0, \pi_1; r) \leq H\sqrt{C^\pi(\pi)} \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).$$

Proof. See Section E.1 for a detailed proof. \square

Finally, we bound the difference between outputs of Algorithm 2 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 D.8. *For any comparator policy π (not necessarily in Π_{SPI}), any reported reward function $r \in \tilde{\mathcal{R}}$, with η set to $\sqrt{\frac{\log |\mathcal{A}|}{2H^2 R_{\max}^2 T}}$ and λ set to $\left(\frac{R_{\max}}{H^2(\epsilon_S + 3\epsilon_{\mathcal{F}})^2}\right)^{1/3}$ in Algorithm 2, the following claims hold conditioned on the event $\mathcal{G}(\Pi_{\text{SPI}})$:*

1. Let $\check{Q}_{1,r}^{(t)}$ and $\check{\pi}_r^{(t)}$ be the pessimistic value function estimate and policy estimate. Then

$$\begin{aligned} V_1^\pi(s_0; r) - \frac{1}{T} \sum_{t=1}^T \check{Q}_{1,r}^{(t)}(s_0, \check{\pi}_{1,r}^{(t)}) &\leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} \\ &+ H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\check{\pi}_r^{(t)}}(\pi)} \right) \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

2. Let $\hat{Q}_{1,r}^{(t)}$ and $\hat{\pi}_r^{(t)}$ be the optimistic value function estimate and policy estimate. Then

$$\begin{aligned} V_1^\pi(s_0; r) - \frac{1}{T} \sum_{t=1}^T \hat{Q}_{1,r}^{(t)}(s_0, \hat{\pi}_{1,r}^{(t)}) &\leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} \\ &+ H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\hat{\pi}_r^{(t)}}(\pi)} \right) \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

Proof. See Section E.2 for a detailed proof. \square

We then proceed with the proof as follows. We start by bounding the suboptimality of the output policy, defined according to equation (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 $\check{\pi} = \check{\pi}_R^{\text{out}}$ be the policy chosen by the algorithm when all agents are truthful, let $\tilde{\pi} = \check{\pi}_{r_i + \tilde{R}_{-i}}^{\text{out}}$ be the policy chosen when we only assume the agent i is truthful, and finally let $\tilde{\pi}_R = \check{\pi}_R^{\text{out}}$ be the policy chosen when none of the agents are truthful. Let the prices charged by the algorithm be $\{\hat{p}_i\}_{i=1}^n$, $\{\tilde{p}_i\}_{i=1}^n$, and $\{\hat{p}_{i,\tilde{R}}\}_{i=1}^n$, respectively.

Social Welfare Suboptimality Assuming all agents are truthful, we have $\tilde{r}_i = r_i$ for all i . Let π^* be the maximizer of $V_1^\pi(s_0; R)$ over π and let $\check{\pi}_R^{(t)}$ be the pessimistic policy iterate of Algorithm 2. We know that the social welfare suboptimality of $\check{\pi}$ is

$$\begin{aligned} \text{SubOpt}(\check{\pi}; s_0) &= V_1^{\pi^*}(s_0; R) - V_1^{\check{\pi}}(s_0; R) = V_1^{\pi^*}(s_0; R) - \frac{1}{T} \sum_{t=1}^T V_1^{\check{\pi}_R^{(t)}}(s_0; R) \\ &= \frac{1}{T} \sum_{t=1}^T \left(V_1^{\pi^*}(s_0; R) - Q_1^{\check{\pi}_R^{(t)}}(s_0, \check{\pi}_{1,R}^{(t)}; R) \right), \end{aligned}$$

as we recall that $\check{\pi}$ is the uniform mixture of policies $\{\check{\pi}_R^{(t)}\}_{t \in [T]}$. By Lemma D.6, we have

$$\text{SubOpt}(\check{\pi}; s_0) \leq \frac{1}{T} \sum_{t=1}^T \left(V_1^{\pi^*}(s_0; R) - \check{Q}_{1,R}^{(t)}(s_0, \check{\pi}_{1,R}^{(t)}; R) \right) + \sqrt{\epsilon_{\mathcal{F}}} + 2\lambda H \epsilon_S + 6\lambda H \epsilon_{\mathcal{F}}, \quad (\text{D.3})$$

where $\check{Q}_R^{(t)}$ is the pessimistic estimate of $Q(\cdot, \cdot; R)$ at the t -th iteration of Algorithm 2. When $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_S + 3\epsilon_{\mathcal{F}})^2}\right)^{1/3}$ and $\eta = \sqrt{\frac{\log |\mathcal{A}|}{2H^2 R_{\max}^2 T}}$, we apply Lemma D.8 to obtain

$$\begin{aligned} \text{SubOpt}(\check{\pi}; s_0) &\leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\check{\pi}_R^{(t)}}(\pi^*)} \right) \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

Individual Suboptimality Let π_{-i}^* be the maximizer of $V^\pi(s_0; R_{-i})$ over π . By Algorithm 3, the price \hat{p}_i is constructed as

$$\hat{p}_i = G_{-i}^{(1)}(s_0) - G_{-i}^{(2)}(s_0),$$

where $G_{-i}^{(1)}(s_0)$ is an estimate of $V^{\pi_{-i}^*}(s_0; R_{-i})$ obtained using Algorithm 2 and $G_{-i}^{(2)}(s_0)$ is an estimate of $V^{\check{\pi}}(s_0; R_{-i})$ for Algorithm 3's output policy, $\check{\pi}$. This observation will be extensively used in the remainder of the proof.

Assuming all agents are truthful, we have $\tilde{r}_i = r_i$ for all i . Recalling the construction of \hat{p}_i in Algorithm 3 line 3 and the definition of $\{p_i^*\}_{i=1}^n$ (see (2.2)), we have

$$\begin{aligned} \text{SubOpt}_i(\check{\pi}, \{\hat{p}_i\}_{i=1}^n; s_0) &= V_1^{\pi^*}(s_0; r_i) + V_1^{\pi^*}(s_0; R_{-i}) - V_1^{\pi_{-i}^*}(s_0; R_{-i}) - V_1^{\check{\pi}}(s_0; r_i) + G_{-i}^{(1)}(s_0) - G_{-i}^{(2)}(s_0) \\ &= V_1^{\pi^*}(s_0; R) - V_1^{\pi_{-i}^*}(s_0; R_{-i}) - V_1^{\check{\pi}}(s_0; r_i) + G_{-i}^{(1)}(s_0) - G_{-i}^{(2)}(s_0) \\ &\leq V_1^{\pi^*}(s_0; R) - V_1^{\check{\pi}}(s_0; R) + \left(G_{-i}^{(1)}(s_0) - V_1^{\pi_{-i}^*}(s_0; R_{-i}) \right) + \left(V_1^{\check{\pi}}(s_0; R_{-i}) - G_{-i}^{(2)}(s_0) \right) \\ &= \text{SubOpt}(\check{\pi}; s_0) + \left(G_{-i}^{(1)}(s_0) - V_1^{\pi_{-i}^*}(s_0; R_{-i}) \right) + \left(V_1^{\check{\pi}}(s_0; R_{-i}) - G_{-i}^{(2)}(s_0) \right). \end{aligned}$$

We have already bounded the first term and now focus on the two latter terms.

We begin by examining $G_{-i}^{(1)}(s_0) - V_1^{\pi_{-i}^*}(s_0; R_{-i})$.

- Suppose $\zeta_1 = \text{OPT}$. Since π_{-i}^* maximizes $V_1^{\pi_{-i}^*}(s_0; R_{-i})$ over π , we have

$$G_{-i}^{(1)}(s_0) - V_1^{\pi_{-i}^*}(s_0; R_{-i}) \leq G_{-i}^{(1)}(s_0) - V_1^{\hat{\pi}_{-i}}(s_0; R_{-i}).$$

Recall that $\hat{Q}_{R_{-i}}^{\text{out}}$ is the optimistic function estimate from the output of Algorithm 2, which is exactly the output of Algorithm 1 called on the policy returned by Algorithm 2, $\hat{\pi}_{-i}$. By Lemma D.7, we know that

$$\begin{aligned} G_{-i}^{(1)}(s_0) - V_1^{\hat{\pi}_{-i}}(s_0; R_{-i}) &\leq H \sqrt{C^{\hat{\pi}_{-i}}(\hat{\pi}_{-i})} \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

- Suppose $\zeta_1 = \text{PES}$. Since π_{-i}^* maximizes $V_1^{\pi_{-i}^*}(s_0; R_{-i})$ over π , we have

$$G_{-i}^{(1)}(s_0) - V_1^{\pi_{-i}^*}(s_0; R_{-i}) \leq G_{-i}^{(1)}(s_0) - V_1^{\check{\pi}_{-i}}(s_0; R_{-i}).$$

Recall that $G_{-i}^{(1)}(s_0) = \check{Q}_{1,R_{-i}}^{\text{out}}(s_0, \check{\pi}_{1,-i})$. When $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_S + 3\epsilon_{\mathcal{F}})^2}\right)^{1/3}$, by Lemma D.6 we have

$$G_{-i}^{(1)}(s_0) - V_1^{\pi_{-i}^*}(s_0; R_{-i}) \leq \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3}.$$

We perform a similar analysis for $V_1^{\tilde{\pi}}(s_0; R_{-i}) - G_{-i}^{(2)}(s_0)$ and when $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_S + 3\epsilon_{\mathcal{F}})^2}\right)^{1/3}$.

- When $\zeta_2 = \text{OPT}$, $V_1^{\tilde{\pi}}(s_0; R_{-i}) - G_{-i}^{(2)}(s_0) \leq \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3}$ by Lemma D.6.
- When $\zeta_2 = \text{PES}$, let $\tilde{Q}_{R_{-i}}^{\tilde{\pi}}$ be the pessimistic output of Algorithm 1 called on $\tilde{\pi}$. By Lemma D.7, we have

$$V_1^{\tilde{\pi}}(s_0; R_{-i}) - G_{-i}^{(2)}(s_0) \leq H\sqrt{C^{\tilde{\pi}}(\tilde{\pi})} \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right).$$

Seller Suboptimality We now turn our attention to the sellers' suboptimality. Assuming all agents are truthful, we have $\tilde{r}_i = r_i$ for all i . Recalling the definition of $\{p_i^*\}_{i=1}^n$ in (2.2), we have

$$\begin{aligned} \text{SubOpt}_0(\tilde{\pi}, \{\hat{p}_i\}_{i=1}^n; s_0) &= V_1^{\pi^*}(s_0; r_0) - V_1^{\tilde{\pi}}(s_0; r_0) + \sum_{i=1}^n \left(\max_{\pi' \in \Pi} V_1^{\pi'}(s_0; R_{-i}) - V_1^{\pi^*}(s_0; R_{-i}) \right) - \sum_{i=1}^n \hat{p}_i \\ &= \sum_{i=1}^n \max_{\pi' \in \Pi} V_1^{\pi'}(s_0; R_{-i}) - (n-1)V_1^{\pi^*}(s_0; R) - V_1^{\tilde{\pi}}(s_0; r_0) - \sum_{i=1}^n G_{-i}^{(1)}(s_0) + \sum_{i=1}^n G_{-i}^{(2)}(s_0) \\ &= \sum_{i=1}^n \left(\max_{\pi' \in \Pi} V_1^{\pi'}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0) \right) - (n-1)V_1^{\pi^*}(s_0; R) - V_1^{\tilde{\pi}}(s_0; r_0) + \sum_{i=1}^n G_{-i}^{(2)}(s_0) \quad (\text{D.4}) \\ &= \sum_{i=1}^n \left(V_1^{\pi^*}{}_{-i}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0) \right) + (n-1)(V_1^{\tilde{\pi}}(s_0; R) - V_1^{\pi^*}(s_0; R)) \\ &\quad + \sum_{i=1}^n \left(G_{-i}^{(2)}(s_0) - V_1^{\tilde{\pi}}(s_0, R_{-i}) \right) \\ &\leq \sum_{i=1}^n \left(V_1^{\pi^*}{}_{-i}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0) \right) + \sum_{i=1}^n \left(G_{-i}^{(2)}(s_0) - V_1^{\tilde{\pi}}(s_0, R_{-i}) \right), \end{aligned}$$

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

$$\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_S + 3\epsilon_{\mathcal{F}})^2}\right)^{1/3}, \quad \eta = \sqrt{\frac{\log |\mathcal{A}|}{2H^2 R_{\max}^2 T}}.$$

- When $\zeta_1 = \text{OPT}$, by Algorithm 3 line 3, we know that for any $i \in [n]$,

$$V_1^{\pi^*}{}_{-i}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0) = V_1^{\pi^*}{}_{-i}(s_0; R_{-i}) - \hat{Q}_{1,R_{-i}}^{\text{out}}(s_0, \hat{\pi}_{1,-i}).$$

By Lemma D.8, we know that

$$\begin{aligned} V_1^{\pi^*}{}_{-i}(s_0; R_{-i}) - \frac{1}{T} \sum_{t=1}^T \hat{Q}_{1,R_{-i}}^{(t)}(s_0, \hat{\pi}_{1,R_{-i}}^{(t)}) &\leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} \\ &\quad + H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\hat{\pi}_{R_{-i}}^{(t)})} \right) \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

By Lemma D.7 and recalling that $\hat{\pi}_{-i}$ is the uniform mixture of $\{\hat{\pi}_{R_{-i}}^{(t)}\}_{t \in [T]}$, we know that

$$\begin{aligned} &\frac{1}{T} \sum_{t=1}^T \hat{Q}_{1,R_{-i}}^{(t)}(s_0, \hat{\pi}_{1,R_{-i}}^{(t)}) - V_1^{\hat{\pi}_{-i}}(s_0; R_{-i}) \\ &= \frac{1}{T} \sum_{t=1}^T \left(\hat{Q}_{1,R_{-i}}^{(t)}(s_0, \hat{\pi}_{1,R_{-i}}^{(t)}) - V_1^{\hat{\pi}_{R_{-i}}^{(t)}}(s_0; R_{-i}) \right) \\ &\leq H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\hat{\pi}_{R_{-i}}^{(t)})} \right) \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

Lastly, by Lemma D.6, we also know that

$$V_1^{\hat{\pi}^{-i}}(s_0; R_{-i}) - \hat{Q}_{1,R_{-i}}^{\text{out}}(s_0, \hat{\pi}_{1,-i}) \leq \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3}.$$

Summing the three parts tells us that, for all $i \in [n]$, we have

$$\begin{aligned} & V_1^{\pi_{-i}^*}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0) \\ &= V_1^{\pi_{-i}^*}(s_0; R_{-i}) - \hat{Q}_{1,R_{-i}}^{\text{out}}(s_0, \hat{\pi}_{1,-i}) \\ &\leq 2H^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H\left(\frac{1}{T}\sum_{t=1}^T\left(\sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\pi_{-i}^*)} + \sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\hat{\pi}_{R_{-i}}^{(t)})}\right)\right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right) \end{aligned} \tag{D.5}$$

and

$$\begin{aligned} & \sum_{i=1}^n \left(V_1^{\pi_{-i}^*}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0) \right) \\ &\leq 2nH^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + n\sqrt{\epsilon_{\mathcal{F}}} + 2n(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H\left(\sum_{i=1}^n \frac{1}{T}\sum_{t=1}^T\left(\sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\pi_{-i}^*)} + \sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\hat{\pi}_{R_{-i}}^{(t)})}\right)\right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right). \end{aligned}$$

- When $\zeta_1 = \text{PES}$, by Algorithm 3 we know that for any $i \in [n]$,

$$V_1^{\pi_{-i}^*}(s_0; R_{-i}) - G_{-i}^{(1)}(s_0) = V_1^{\pi_{-i}^*}(s_0; R_{-i}) - \check{Q}_{1,R_{-i}}^{\text{out}}(s_0, \check{\pi}_{1,-i}).$$

By Lemma D.8, we know that

$$\begin{aligned} & V_1^{\pi_{-i}^*}(s_0; R_{-i}) - \frac{1}{T}\sum_{t=1}^T\check{Q}_{1,R_{-i}}^{(t)}(s_0, \check{\pi}_{1,R_{-i}}^{(t)}) \leq 2H^2R_{\max}\sqrt{\frac{2\log|\mathcal{A}|}{T}} \\ &\quad + H\left(\frac{1}{T}\sum_{t=1}^T\sqrt{C^{\check{\pi}_{R_{-i}}^{(t)}}(\pi_{-i}^*)}\right)\left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right). \end{aligned}$$

By Lemma D.6, we know that

$$\frac{1}{T}\sum_{t=1}^T\check{Q}_{1,R_{-i}}^{(t)}(s_0, \check{\pi}_{1,R_{-i}}^{(t)}) - V_1^{\check{\pi}_{-i}}(s_0; R_{-i}) \leq \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3}.$$

By Lemma D.7, we further know that

$$\begin{aligned} & V_1^{\check{\pi}_{-i}}(s_0; R_{-i}) - \check{Q}_{1,R_{-i}}^{\text{out}}(s_0, \check{\pi}_{1,-i}) \\ &\leq H\sqrt{C^{\check{\pi}_{-i}}(\check{\pi}_{-i})}\left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}\right). \end{aligned}$$

Summing the three parts together tells us that, for all $i \in [n]$ and any $C \geq 1$, we have

$$\begin{aligned}
 V_1^{\pi^*} (s_0; R_{-i}) - G_{-i}^{(1)} (s_0) &= V_1^{\pi^*} (s_0; R_{-i}) - \check{Q}_{1, R_{-i}}^{\text{out}} (s_0, \check{\pi}_{1, -i}) \\
 &\leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3} (\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} \\
 &\quad + H \left(\sqrt{C^{\check{\pi}_{-i}} (\check{\pi}_{-i})} + \frac{1}{T} \sum_{t=1}^T \sqrt{C^{\check{\pi}_{R_{-i}}^{(t)}} (\pi_{-i}^*)} \right) \\
 &\quad \times \left(2(HR_{\max})^{1/3} (\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F}, \mathcal{F}}} \right)
 \end{aligned} \tag{D.6}$$

and

$$\begin{aligned}
 \sum_{i=1}^n \left(V_1^{\pi^*} (s_0; R_{-i}) - G_{-i}^{(1)} (s_0) \right) \\
 \leq 2nH^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} + n\sqrt{\epsilon_{\mathcal{F}}} + 2n(HR_{\max})^{1/3} (\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} \\
 + H \left(\sum_{i=1}^n \sqrt{C^{\check{\pi}_{-i}} (\check{\pi}_{-i})} + \sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T \sqrt{C^{\check{\pi}_{R_{-i}}^{(t)}} (\pi_{-i}^*)} \right) \\
 \times \left(2(HR_{\max})^{1/3} (\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F}, \mathcal{F}}} \right).
 \end{aligned}$$

- When $\zeta_2 = \text{OPT}$, for all $i \in [n]$, let $\check{Q}_{R_{-i}}^{\check{\pi}}$ be the pessimistic estimate of $Q^{\check{\pi}}(\cdot, \cdot; R_{-i})$ returned by Algorithm 1. By Lemma D.7, we know

$$\begin{aligned}
 \sum_{i=1}^n \left(G_{-i}^{(2)} (s_0) - V_1^{\check{\pi}} (s_0, R_{-i}) \right) \\
 \leq nH \sqrt{C^{\check{\pi}} (\check{\pi})} \left(2(HR_{\max})^{1/3} (\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F}, \mathcal{F}}} \right).
 \end{aligned}$$

- When $\zeta_2 = \text{PES}$, $\sum_{i=1}^n \left(G_{-i}^{(2)} (s_0) - V_1^{\check{\pi}} (s_0, R_{-i}) \right) \leq n\sqrt{\epsilon_{\mathcal{F}}} + 2n(HR_{\max})^{1/3} (\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3}$ by Lemma D.6.

Plugging in the bound for SubOpt($\check{\pi}; s_0$) 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 $i' \in [n] \setminus i$. Recall that for any price p_i , the agents' expected utility under the chosen policy $\check{\pi}$ can be written as

$$\mathbb{E}_{d_{\check{\pi}}} [u_i] = V_1^{\check{\pi}} (s_0; r_i) - p_i.$$

According to Algorithm 3, we have

$$\begin{aligned}
 \mathbb{E}_{\check{\pi}} [u_i] &= V_1^{\check{\pi}} (s_0; r_i) - G_{-i}^{(1)} (s_0) + G_{-i}^{(2)} (s_0) \\
 &= V_1^{\check{\pi}} (s_0; r_i) + G_{-i}^{(2)} (s_0) - V^{\pi^*} (s_0; R_{-i}) + V^{\pi^*} (s_0; R_{-i}) - G_{-i}^{(1)} (s_0) \\
 &= (V^{\pi^*} (s_0; R) - V^{\pi^*} (s_0; R_{-i})) + V^{\check{\pi}} (s_0; r_i) + G_{-i}^{(2)} (s_0) - V^{\pi^*} (s_0; R) \\
 &\quad + V^{\pi^*} (s_0; R_{-i}) - G_{-i}^{(1)} (s_0) \\
 &\geq V^{\check{\pi}} (s_0; r_i) + G_{-i}^{(2)} (s_0) - V^{\pi^*} (s_0; R) + V^{\pi^*} (s_0; R_{-i}) - G_{-i}^{(1)} (s_0) \\
 &= G_{-i}^{(2)} (s_0) - V^{\check{\pi}} (s_0; R_{-i}) + V^{\check{\pi}} (s_0; R) - V^{\pi^*} (s_0; R) + V^{\pi^*} (s_0; R_{-i}) - G_{-i}^{(1)} (s_0),
 \end{aligned} \tag{D.7}$$

where the inequality comes from the fact that

$$(V^{\pi^*} (s_0; R) - V^{\pi^*} (s_0; R_{-i})) \geq (V^{\pi^*} (s_0; R) - V^{\pi^*} (s_0; R_{-i})) = V^{\pi^*} (s_0; r_i) \geq 0,$$

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 $G_{-i}^{(2)}(s_0) - V^{\tilde{\pi}}(s_0; R_{-i})$, respectively, when bounding the individual suboptimality for the agents. Also note that $V^{\tilde{\pi}}(s_0; R) - V^{\pi^*}(s_0; R) = -\text{SubOpt}(\tilde{\pi}; s_0)$ 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 $\zeta_1 = \text{OPT}$, by equation (D.5) we know that

$$\begin{aligned} G_{-i}^{(1)}(s_0) - V_1^{\pi_{-i}^*}(s_0; R_{-i}) &\geq -2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} - \sqrt{\epsilon_{\mathcal{F}}} - 2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad - H \left(\frac{1}{T} \sum_{t=1}^T \left(\sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\pi_{-i}^*)} + \sqrt{C^{\hat{\pi}_{R_{-i}}^{(t)}}(\hat{\pi}_{R_{-i}}^{(t)})} \right) \right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

- When $\zeta_1 = \text{PES}$, by equation (D.6) we know that

$$\begin{aligned} G_{-i}^{(1)}(s_0) - V_1^{\pi_{-i}^*}(s_0; R_{-i}) &\geq -2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} - \sqrt{\epsilon_{\mathcal{F}}} - 2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad - H \left(\sqrt{C^{\tilde{\pi}_{-i}}(\tilde{\pi}_{-i})} + \frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}_{R_{-i}}^{(t)}}(\pi_{-i}^*)} \right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

- When $\zeta_2 = \text{OPT}$, by Lemma D.6, we know that

$$G_{-i}^{(2)}(s_0) - V^{\tilde{\pi}}(s_0; R_{-i}) \geq -\sqrt{\epsilon_{\mathcal{F}}} - 2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3}.$$

- When $\zeta_2 = \text{PES}$, by Lemma D.7

$$G_{-i}^{(2)}(s_0) - V_1^{\tilde{\pi}}(s_0; R_{-i}) \geq -H \sqrt{C^{\tilde{\pi}}(\tilde{\pi})} \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right).$$

We now argue that our analysis holds even when the other agents are not truthful. Recall that $\tilde{\pi}$ is the output policy selected by Algorithm 3 when other agents report $\tilde{r}_{i'}$ and the agent i reports truthfully. Observe that here the decomposition in equation (D.7) can be written as

$$\begin{aligned} \mathbb{E}_{\tilde{\pi}}[u_i] &\geq \tilde{G}_{-i}^{(2)}(s_0) - V^{\tilde{\pi}}(s_0; \tilde{R}_{-i}) + V^{\tilde{\pi}}(s_0; r_i + \tilde{R}_{-i}) - V^{\pi_{r_i + \tilde{R}_{-i}}^*}(s_0; r_i + \tilde{R}_{-i}) \\ &\quad + V^{\pi_{-i}^*}(s_0; \tilde{R}_{-i}) - \tilde{G}_{-i}^{(1)}(s_0), \end{aligned}$$

where we recall that $\tilde{R}_{-i} = \sum_{i' \neq i} \tilde{r}_{i'}$, and $\pi_{r_i + \tilde{R}_{-i}}^*$ and π_{-i}^* maximize $V_1^{\pi}(s_0; r_i + \tilde{R}_{-i})$ and $V_1^{\pi}(s_0; \tilde{R}_{-i})$ over π , respectively. We also let $\tilde{G}_{-i}^{(1)}, \tilde{G}_{-i}^{(2)}$ be the estimates used in Algorithm 3 line 3 when other agents are reporting untruthfully.

Similar to the previous sections, we bound different terms under difference choices of ζ_1, ζ_2 .

- When $\zeta_1 = \text{OPT}$, similar to equation (D.5), we have

$$\begin{aligned} \tilde{G}_{-i}^{(1)}(s_0) - V_1^{\pi_{-i}^*}(s_0; \tilde{R}_{-i}) &\geq -2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} - \sqrt{\epsilon_{\mathcal{F}}} - 2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad - H \left(\frac{1}{T} \sum_{t=1}^T \left(\sqrt{C^{\hat{\pi}_{\tilde{R}_{-i}}^{(t)}}(\tilde{\pi}_{-i}^*)} + \sqrt{C^{\hat{\pi}_{\tilde{R}_{-i}}^{(t)}}(\hat{\pi}_{\tilde{R}_{-i}}^{(t)})} \right) \right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

- When $\zeta_1 = \text{PES}$, similar to equation (D.6), we have

$$\begin{aligned} \tilde{G}_{-i}^{(1)}(s_0) - V_1^{\tilde{\pi}^*_{-i}}(s_0; \tilde{R}_{-i}) &\geq -2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} - \sqrt{\epsilon_{\mathcal{F}}} - 2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad - H \left(\sqrt{C^{\tilde{\pi}^*_{-i, \text{out}}}_{\tilde{R}_{-i}}} + \frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}_{-i}}_{\tilde{R}_{-i}}} \right) \\ &\quad \times \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F}, \mathcal{F}}} \right). \end{aligned}$$

- When $\zeta_2 = \text{OPT}$, by Lemma D.6, we know

$$\tilde{G}_{-i}^{(2)}(s_0) - V^{\tilde{\pi}}(s_0; \tilde{R}_{-i}) \geq -\sqrt{\epsilon_{\mathcal{F}}} - 2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3}.$$

- When $\zeta_2 = \text{PES}$, by Lemma D.7

$$\tilde{G}_{-i}^{(2)}(s_0) - V_1^{\tilde{\pi}}(s_0; \tilde{R}_{-i}) \geq -H \sqrt{C^{\tilde{\pi}}(\tilde{\pi})} \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F}, \mathcal{F}}} \right),$$

where $\tilde{\pi}$ is the policy that the seller chooses when agent i reports truthfully and the other agents do not.

We finally focus on lower bounding $V^{\tilde{\pi}}(s_0; r_i + \tilde{R}_{-i}) - V^{\pi^*_{r_i + \tilde{R}_{-i}}}(s_0; r_i + \tilde{R}_{-i})$. Since $\tilde{\pi}$ is the uniform mixture of $\{\tilde{\pi}^{(t)}\}_{t \in [T]}$, we have

$$\begin{aligned} &V_1^{\pi^*_{r_i + \tilde{R}_{-i}}}(s_0; r_i + \tilde{R}_{-i}) - V_1^{\tilde{\pi}}(s_0; r_i + \tilde{R}_{-i}) \\ &= \frac{1}{T} \sum_{t=1}^T \left(V_1^{\pi^*_{r_i + \tilde{R}_{-i}}}(s_0; r_i + \tilde{R}_{-i}) - V_1^{\tilde{\pi}^{(t)}}(s_0; r_i + \tilde{R}_{-i}) \right) \\ &\leq \frac{1}{T} \sum_{t=1}^T \left(V_1^{\pi^*_{r_i + \tilde{R}_{-i}}}(s_0; r_i + \tilde{R}_{-i}) - \check{Q}_{1, r_i + \tilde{R}_{-i}}^{(t)}(s_0, \tilde{\pi}_1^{(t)}) \right) + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \end{aligned}$$

by Lemma D.6. By Lemma D.8, we know that

$$\begin{aligned} &\frac{1}{T} \sum_{t=1}^T \left(V_1^{\pi^*_{r_i + \tilde{R}_{-i}}}(s_0; r_i + \tilde{R}_{-i}) - \check{Q}_{1, r_i + \tilde{R}_{-i}}^{(t)}(s_0, \tilde{\pi}_1^{(t)}) \right) \leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} \\ &\quad + H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi^*_{r_i + \tilde{R}_{-i}})} \right) \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F}, \mathcal{F}}} \right). \end{aligned}$$

Therefore, we have

$$\begin{aligned} &V_1^{\pi^*_{r_i + \tilde{R}_{-i}}}(s_0; r_i + \tilde{R}_{-i}) - V_1^{\tilde{\pi}}(s_0; r_i + \tilde{R}_{-i}) \\ &\leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi^*_{r_i + \tilde{R}_{-i}})} \right) \left(2(HR_{\max})^{1/3}(\epsilon_{\mathcal{S}} + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_{\mathcal{S}} + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F}, \mathcal{F}}} \right). \end{aligned} \tag{D.8}$$

Flipping the signs yields the final bound.

Truthfulness Similar to above and let $\tilde{r}_{i'}$ be the potentially untruthful reward functions reported by other agents and let \tilde{r}_i be the untruthful reward function that the agent i may report. Furthermore, let $\tilde{R}_{-i} = \sum_{i' \neq i} \tilde{r}_{i'}$ and $\tilde{R} = \sum_{i=1}^n \tilde{r}_i$.

Let $\tilde{\pi}$ be the policy chosen by the seller when the agent i is truthful and other agents are possibly non-truthful and $\tilde{\pi}_{\tilde{R}}$ the policy chosen by Algorithm 3 when both the agent i and other agents are non-truthful. The agents' expected utilities for the two cases are

$$\begin{aligned}\mathbb{E}_{\tilde{\pi}}[u_i] &= V_1^{\tilde{\pi}}(s_0; r_i) + \tilde{G}_{-i}^{(2)}(s_0) - \tilde{G}_{-i}^{(1)}(s_0), \\ \mathbb{E}_{d_{\tilde{\pi}_{\tilde{R}}}}[u_i] &= V_1^{\tilde{\pi}_{\tilde{R}}}(s_0; r_i) + \tilde{G}_{-i}^{(2),'}(s_0) - \tilde{G}_{-i}^{(1),'}(s_0),\end{aligned}$$

where $\tilde{G}_{-i}^{(2)}(s_0)$ estimates $V^{\tilde{\pi}}(s_0; \tilde{R}_{-i})$ and $\tilde{G}_{-i}^{(2),'}(s_0)$ estimates $V^{\tilde{\pi}_{\tilde{R}}}(s_0; \tilde{R}_{-i})$.

Observe that both $\tilde{G}_{-i}^{(1)}(s_0)$ and $\tilde{G}_{-i}^{(1),'}(s_0)$ approximate $V_1^{\tilde{\pi}^*}(s_0; \tilde{R}_{-i})$ using the same algorithm, Algorithm 2. As the algorithm itself does not contain randomness and $\tilde{G}_{-i}^{(1)}(s_0)$ and $\tilde{G}_{-i}^{(1),'}(s_0)$ are constructed using the same parameters, the two terms must be equal. Then we have

$$\begin{aligned}\mathbb{E}_{\tilde{\pi}_{\tilde{R}}}[u_i] - \mathbb{E}_{\tilde{\pi}}[u_i] &= V_1^{\tilde{\pi}_{\tilde{R}}}(s_0; r_i) + \tilde{G}_{-i}^{(2),'}(s_0) - \left(V_1^{\tilde{\pi}}(s_0; r_i) + \tilde{G}_{-i}^{(2)}(s_0) \right) \\ &= V_1^{\tilde{\pi}_{\tilde{R}}}(s_0; r_i + \tilde{R}_{-i}) + \tilde{G}_{-i}^{(2),'}(s_0) - V_1^{\tilde{\pi}_{\tilde{R}}}(s_0; \tilde{R}_{-i}) - \left(V_1^{\tilde{\pi}}(s_0; r_i + \tilde{R}_{-i}) + \tilde{G}_{-i}^{(2)}(s_0) - V_1^{\tilde{\pi}}(s_0; \tilde{R}_{-i}) \right) \\ &= V_1^{\tilde{\pi}_{\tilde{R}}}(s_0; r_i + \tilde{R}_{-i}) - V_1^{\pi^*_{r_i + \tilde{R}_{-i}}}(s_0; r_i + \tilde{R}_{-i}) + \tilde{G}_{-i}^{(2),'}(s_0) - V_1^{\tilde{\pi}_{\tilde{R}}}(s_0; \tilde{R}_{-i}) \\ &\quad + V_1^{\pi^*_{r_i + \tilde{R}_{-i}}}(s_0; r_i + \tilde{R}_{-i}) - V_1^{\tilde{\pi}}(s_0; r_i + \tilde{R}_{-i}) + V_1^{\tilde{\pi}}(s_0; \tilde{R}_{-i}) - \tilde{G}_{-i}^{(2)}(s_0),\end{aligned}$$

where we recall that $\pi^*_{r_i + \tilde{R}_{-i}}$ is the maximizer of $V_1^\pi(s_0; r_i + \tilde{R}_{-i})$ over π (the social welfare maximizing policy when agent i reports truthfully). We then know that

$$V_1^{\tilde{\pi}_{\tilde{R}}}(s_0; r_i + \tilde{R}_{-i}) - V_1^{\pi^*_{r_i + \tilde{R}_{-i}}}(s_0; r_i + \tilde{R}_{-i}) \leq 0$$

and

$$\begin{aligned}\mathbb{E}_{\tilde{\pi}_{\tilde{R}}}[u_i] - \mathbb{E}_{\tilde{\pi}}[u_i] &\leq \left(\tilde{G}_{-i}^{(2),'}(s_0) - V_1^{\tilde{\pi}_{\tilde{R}}}(s_0; \tilde{R}_{-i}) \right) + \left(V_1^{\pi^*_{r_i + \tilde{R}_{-i}}}(s_0; r_i + \tilde{R}_{-i}) - V_1^{\tilde{\pi}}(s_0; r_i + \tilde{R}_{-i}) \right) \\ &\quad + \left(V_1^{\tilde{\pi}}(s_0; \tilde{R}_{-i}) - \tilde{G}_{-i}^{(2)}(s_0) \right).\end{aligned}$$

Let us focus on the middle term first. By (D.8), we have

$$\begin{aligned}V_1^{\pi^*_{r_i + \tilde{R}_{-i}}}(s_0; r_i + \tilde{R}_{-i}) - V_1^{\tilde{\pi}}(s_0; r_i + \tilde{R}_{-i}) &\leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} + \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} \\ &\quad + H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\tilde{\pi}^{(t)}}(\pi^*_{r_i + \tilde{R}_{-i}})} \right) \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right).\end{aligned}$$

We state the results conditioned on different values of ζ_2 as the bound no longer depends on ζ_1 .

- When $\zeta_2 = \text{OPT}$, by Lemma D.6, we have

$$V_1^{\tilde{\pi}}(s_0; \tilde{R}_{-i}) - \tilde{G}_{-i}^{(2)}(s_0) \leq \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3},$$

and by Lemma D.7,

$$\begin{aligned}\tilde{G}_{-i}^{(2),'}(s_0) - V_1^{\tilde{\pi}_{\tilde{R}}}(s_0; \tilde{R}_{-i}) &\leq H \sqrt{C^{\tilde{\pi}_{\tilde{R}}}(\tilde{\pi}_{\tilde{R}})} \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right).\end{aligned}$$

- When $\zeta_2 = \text{PES}$, by Lemma D.7,

$$V_1^{\tilde{\pi}}(s_0; \tilde{R}_{-i}) - \tilde{G}_{-i}^{(2)}(s_0) \leq H \sqrt{C^{\tilde{\pi}}(\tilde{\pi})} \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right),$$

and by Lemma D.6,

$$\tilde{G}_{-i}^{(2),\prime}(s_0) - V_1^{\tilde{\pi}}(s_0; \tilde{R}_{-i}) \leq \sqrt{\epsilon_{\mathcal{F}}} + 2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3}.$$

Combining the terms completes the proof. \square

E. Supporting Lemmas

In this section, we provide detailed proofs of supporting lemmas used in Section D.

E.1. Proofs for Algorithm 1

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 & Sun, 2021; Xie et al., 2021; Zanette et al., 2021). More precisely, let r be the input reward for Algorithm 1, π the input policy, and f^π the output. Let \mathcal{M}_{f^π} be the induced MDP. We formally state the result below.

Lemma E.1. *For any input policy π (not necessarily in Π_{SPI}) and input reward function r , Algorithm 1 returns a function f^π such that f^π is the Q -function of the policy π under the induced MDP \mathcal{M}_{f^π} , given by*

$$\mathcal{M}_{f^\pi} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r_{f^\pi}), \quad (\text{E.1})$$

where $r_{f^\pi, h} = r_h + f_h^\pi - \mathbb{T}_{h,r}^\pi f_{h+1}^\pi$. 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. \square

We immediately have the following corollary.

Corollary E.2. *Let f^π be any one of the two functions returned by Algorithm 1 for any input policy π (not necessarily in Π_{SPI}) and any input reward function r . Then, for all $h \in [H]$, we have*

$$|f_h^\pi(s, a) - Q_h^\pi(s, a; r)| \leq \sum_{h'=h}^H \mathbb{E}_{(S_{h'}, A_{h'}) \sim \pi | (s, a)} [|f_h^\pi - \mathbb{T}_{h,r}^\pi f_{h+1}^\pi|].$$

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 (E.1) and using Jensen's inequality concludes the proof. \square

We proceed to show that Algorithm 1 is approximately optimistic/pessimistic and bounding the estimation error of its outputs. We begin with the proof of Lemma D.6.

Proof of Lemma D.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 2.3. By Jensen's inequality, we have

$$|f_{1,r}^{\pi, *}(s_0, \pi_1) - Q_1^\pi(s_0, \pi_1; r)| \leq \mathbb{E}_{a \sim \pi_1(\cdot | s_0)} [|f_{1,r}^{\pi, *}(s_0, \pi_1) - Q_1^\pi(s_0, \pi_1; r)|] \leq \sqrt{\epsilon_{\mathcal{F}}}.$$

Additionally, using Lemma D.4 we know that, conditioned on the event $\mathcal{G}(\Pi_{\text{SPI}})$, for all $h \in [H]$ we have $B_{h,r}(f_r^{\pi, *}, \pi; \mathcal{D}) \leq 2\epsilon_S + 6\epsilon_{\mathcal{F}}$.

We then consider \check{Q}_r^π . By (3.2), we know that

$$\begin{aligned} \check{Q}_{1,r}^\pi(s_0, \pi) + \lambda \sum_{h=1}^H B_{h,r}(\check{Q}_r^\pi, \pi; \mathcal{D}) &\leq f_{1,r}^{\pi,*}(s_0, \pi) + \lambda \sum_{h=1}^H B_{h,r}(f_r^{\pi,*}, \pi; \mathcal{D}) \\ &\leq Q_1^\pi(s_0, \pi; r) + |f_{1,r}^{\pi,*}(s_0, \pi_1) - Q_1^\pi(s_0, \pi_1; r)| + 2\lambda H\epsilon_S + 6\lambda H\epsilon_{\mathcal{F}} \\ &\leq Q_1^\pi(s_0, \pi_1; r) + \sqrt{\epsilon_{\mathcal{F}}} + 2\lambda H\epsilon_S + 6\lambda H\epsilon_{\mathcal{F}}. \end{aligned}$$

Similarly for \hat{Q}_r^π , by (3.2), we have

$$\begin{aligned} \hat{Q}_{1,r}^\pi(s_0, \pi) - \lambda \sum_{h=1}^H B_{h,r}(\hat{Q}_r^\pi, \pi; \mathcal{D}) &\geq f_{1,r}^{\pi,*}(s_0, \pi) - \lambda \sum_{h=1}^H B_{h,r}(f_r^{\pi,*}, \pi; \mathcal{D}) \\ &\geq Q_1^\pi(s_0, \pi; r) - |f_{1,r}^{\pi,*}(s_0, \pi_1) - Q_1^\pi(s_0, \pi_1; r)| - 2\lambda H\epsilon_S - 6\lambda H\epsilon_{\mathcal{F}} \\ &\geq Q_1^\pi(s_0, \pi_1; r) - \sqrt{\epsilon_{\mathcal{F}}} - 2\lambda H\epsilon_S - 6\lambda H\epsilon_{\mathcal{F}}, \end{aligned}$$

thus completing the proof. \square

We prove that the action-value functions returned by Algorithm 1 are sufficiently good estimates.

Proof of Lemma D.7. By Corollary E.2, we have

$$\begin{aligned} \hat{Q}_{1,r}^\pi(s_0, \pi_1) - Q_1^\pi(s_0, \pi_1; r) &\leq \left| \sum_{h=1}^H \mathbb{E}_\pi \left[\hat{Q}_{h,r}^\pi - \mathbb{T}_{h,r}^\pi \hat{Q}_{h+1,r}^\pi \right] \right|, \\ Q_1^\pi(s_0, \pi_1; r) - \check{Q}_{1,r}^\pi(s_0, \pi_1) &\leq \left| \sum_{h=1}^H \mathbb{E}_\pi \left[\check{Q}_{h,r}^\pi - \mathbb{T}_{h,r}^\pi \check{Q}_{h+1,r}^\pi \right] \right|. \end{aligned}$$

Since the differences share similar forms, we can without loss of generality only consider \hat{Q}_r^π . Recall the definition of $C^\pi(\nu)$, given in Definition 2.5. We have

$$\begin{aligned} \left| \sum_{h=1}^H \mathbb{E}_\pi \left[\hat{Q}_{h,r}^\pi - \mathbb{T}_{h,r}^\pi \hat{Q}_{h+1,r}^\pi \right] \right| &\leq \sum_{h=1}^H \mathbb{E}_\pi \left[\left\| \hat{Q}_{h,r}^\pi - \mathbb{T}_{h,r}^\pi \hat{Q}_{h+1,r}^\pi \right\| \right] \\ &\leq \sqrt{C^\pi(\pi)} \sum_{h=1}^H \mathbb{E}_{\mu_h} \left[\left\| \hat{Q}_{h,r}^\pi - \mathbb{T}_{h,r}^\pi \hat{Q}_{h+1,r}^\pi \right\| \right], \end{aligned} \tag{E.2}$$

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 D.6, let $f_r^{\pi,*}$ be the best approximation of $Q^\pi(\cdot, \cdot; r)$ as defined in Assumption 2.3. Then

$$\lambda \sum_{h=1}^H B_{h,r}(\check{Q}_r^\pi, \pi; \mathcal{D}) \leq f_{1,r}^{\pi,*}(s_0, \pi_1) - \check{Q}_{1,r}^\pi(s_0, \pi_1) + 2\lambda H\epsilon_S + 6\lambda H\epsilon_{\mathcal{F}}.$$

Since $f_r^{\pi,*}, \check{Q}_{1,r}^\pi \in \mathcal{F}$, we have $f_r^{\pi,*}, \check{Q}_{1,r}^\pi \in [-HR_{\max}, HR_{\max}]$ and thus

$$\sum_{h=1}^H B_{h,r}(\check{Q}_r^\pi, \pi; \mathcal{D}) \leq \frac{2HR_{\max}}{\lambda} + 2H\epsilon_S + 6H\epsilon_{\mathcal{F}}.$$

By Corollary D.5, conditioned on $\mathcal{G}(\Pi_{\text{SPI}})$, we have

$$\begin{aligned} \sum_{h=1}^H \mathbb{E}_{\mu_h} \left[\left\| \hat{Q}_{h,r}^\pi - \mathbb{T}_{h,r}^\pi \hat{Q}_{h+1,r}^\pi \right\|^2 \right] &\leq 2 \sum_{h=1}^H B_{h,r}(\check{Q}_r^\pi, \pi; \mathcal{D}) + 4H\epsilon_S + 3H\epsilon_{\mathcal{F},\mathcal{F}} \\ &\leq \frac{4HR_{\max}}{\lambda} + 8H\epsilon_S + 12H\epsilon_{\mathcal{F}} + 3H\epsilon_{\mathcal{F},\mathcal{F}}. \end{aligned}$$

Plugging the bound back into (E.2) and applying Cauchy-Schwarz inequality gives us

$$\begin{aligned} \left| \sum_{h=1}^H \mathbb{E}_\pi \left[\check{Q}_{h,r}^\pi - \mathbb{T}_{h,r}^\pi \check{Q}_{h+1,r}^\pi \right] \right| &\leq \sqrt{H} \sqrt{C^\pi(\pi)} \sqrt{\frac{4HR_{\max}}{\lambda} + 8H\epsilon_S + 12H\epsilon_{\mathcal{F}} + 3H\epsilon_{\mathcal{F},\mathcal{F}}} \\ &= H \sqrt{C^\pi(\pi)} \sqrt{\frac{4R_{\max}}{\lambda} + 8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}}. \end{aligned}$$

Setting $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_S + 3\epsilon_{\mathcal{F}})} \right)^{1/3}$ and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \in \mathbb{R}_{\geq 0}$ completes the proof. \square

E.2. Proofs for Algorithm 2

We now turn to analyzing the policies selected in Algorithm 2. In particular, we focus on the mirror descent-style updates given in (3.3) and (3.4). We start by defining an abstract version of the procedure in Algorithm 2.

Definition E.3. 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 $\mathbb{E}_{a \in \mathcal{A}} [\log \pi_h(a|s)] = \sum_{a \in \mathcal{A}} \pi_h(a|s) \log \pi_h(a|s)$ for all π, h , and s . We continue with a standard analysis of the regret of actor-critic algorithms.

Lemma E.4. For any π (not necessarily in Π_{SPI}), for all $h \in [H]$ and $s \in \mathcal{S}$, setting $\eta = \sqrt{\frac{\log |\mathcal{A}|}{2H^2 R_{\max}^2 T}}$ in the procedure defined in E.3 ensures that

$$\sum_{t=1}^T \langle \pi_h(\cdot|s) - \pi_h^{(t)}(\cdot|s), f_h^{(t)}(s, \cdot) \rangle \leq 2HR_{\max} \sqrt{2T \log |\mathcal{A}|}.$$

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

$$\sum_{t=1}^T \langle \pi_h(\cdot|s) - \pi_h^{(t)}(\cdot|s), f_h^{(t)}(s, \cdot) \rangle \leq \sum_{t=1}^T \langle \pi_h^{(t+1)} - \pi_h^{(t)}(\cdot|s), f_h^{(t)}(s, \cdot) \rangle - \frac{1}{\eta} \mathbb{E}_{a \sim \pi_h^{(1)}} \left[\log \pi_h^{(1)}(a|s) \right],$$

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

$$\langle \pi_h(\cdot|s) - \pi_h^{(t)}(\cdot|s), f_h^{(t)}(s, \cdot) \rangle \leq \|f_h^{(t)}(s, \cdot)\|_\infty \sqrt{2\eta \langle \pi_h(\cdot|s) - \pi_h^{(t)}(\cdot|s), f_h^{(t)}(s, \cdot) \rangle}.$$

Recalling that all $f_h \in \mathcal{F}_h$ are bounded by HR_{\max} , we know that $\langle \pi_h(\cdot|s) - \pi_h^{(t)}(\cdot|s), f_h^{(t)}(s, \cdot) \rangle \leq 2\eta H^2 R_{\max}^2$. Following the proof in Section C.1 in Xie et al. (2021) completes our proof. \square

With the observations above, we proceed with proving Lemma D.8.

Proof of Lemma D.8. We analyze the pessimistic estimate and note that the analysis is similar for the other part. Let $\check{\pi}_r^{(t)}$ be the policy iterate of Algorithm 2 and $\check{Q}_r^{(t)}$ the corresponding value function estimate. We know that

$$\begin{aligned} V_1^\pi(s_0; r) - \frac{1}{T} \sum_{t=1}^T \check{Q}_{1,r}^{(t)}(s_0, \check{\pi}_{1,r}^{(t)}) &= \frac{1}{T} \sum_{t=1}^T \left(Q_1^\pi(s_0, \pi_1; r) - \check{Q}_{1,r}^{(t)}(s_0, \check{\pi}_{1,r}^{(t)}) \right) \\ &\leq \frac{1}{T} \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_\pi \left[\langle \check{Q}_{h,r}^{(t)}(s_h, \cdot), \pi_h(\cdot|s_h) - \check{\pi}_{h,r}^{(t)}(\cdot|s_h) \rangle \right] + \left| \frac{1}{T} \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_\pi \left[\check{Q}_{h,r}^{(t)} - \mathbb{T}_{h,r}^{\check{\pi}_r^{(t)}} \check{Q}_{h+1,r}^{(t)} \right] \right|, \end{aligned}$$

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 E.4, we know that when $\eta = \sqrt{\frac{\log |\mathcal{A}|}{2H^2 R_{\max}^2 T}}$, we have

$$\frac{1}{T} \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_\pi \left[\langle \check{Q}_{h,r}^{(t)}(s_h, \cdot), \pi_h(\cdot | s_h) - \check{\pi}_{h,r}^{(t)}(\cdot | s_h) \rangle \right] \leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}}.$$

For all $t \in [T]$, similar to the proof of Lemma D.7, when $\lambda = \left(\frac{R_{\max}}{H^2(\epsilon_S + 3\epsilon_{\mathcal{F}})^2} \right)^{1/3}$, we have

$$\left| \sum_{h=1}^H \mathbb{E}_\pi \left[\check{Q}_{h,r}^{(t)} - \mathbb{T}_{h,r}^{\check{\pi}_{h,r}^{(t)}} \check{Q}_{h+1,r}^{(t)} \right] \right| \leq H \sqrt{C^{\check{\pi}_r^{(t)}}(\pi)} \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right).$$

Notice that the distribution shift coefficient is changed from $C^\pi(\pi)$ to $C^{\check{\pi}_r^{(t)}}(\pi)$, as the policy specific Bellman operator \mathbb{T} is now induced by policy $\check{\pi}_r^{(t)}$ rather than π . Taking the average over t and applying the triangle inequality give us

$$\begin{aligned} & \left| \frac{1}{T} \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_\pi \left[\check{Q}_{h,r}^{(t)} - \mathbb{T}_{h,r}^{\check{\pi}_{h,r}^{(t)}} \check{Q}_{h+1,r}^{(t)} \right] \right| \\ & \leq H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\check{\pi}_r^{(t)}}(\pi)} \right) \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right). \end{aligned}$$

Combining the bounds, we have

$$\begin{aligned} V_1^\pi(s_0; r) - \frac{1}{T} \sum_{t=1}^T \check{Q}_{1,r}^{(t)}(s_0, \check{\pi}_{1,r}^{(t)}) & \leq 2H^2 R_{\max} \sqrt{\frac{2 \log |\mathcal{A}|}{T}} \\ & + H \left(\frac{1}{T} \sum_{t=1}^T \sqrt{C^{\check{\pi}_r^{(t)}}(\pi)} \right) \left(2(HR_{\max})^{1/3}(\epsilon_S + 3\epsilon_{\mathcal{F}})^{1/3} + \sqrt{8\epsilon_S + 12\epsilon_{\mathcal{F}} + 3\epsilon_{\mathcal{F},\mathcal{F}}} \right), \end{aligned}$$

which completes the proof. \square

F. Concentration Analysis

In this section, we prove the concentration lemmas used in Section D.

F.1. Proof of Lemma D.2

Proof of Lemma D.2. We start by including a minor adaptation of a useful result from Györfi et al. (2002) for completeness.

Theorem F.1 (Adaptation of Theorem 11.6 from Györfi et al. (2002)). *Let $B \geq 1$ and let \mathcal{G} be a class of functions $g: \mathbb{R}^d \rightarrow [0, B]$. Let Z_1, Z_2, \dots, Z_K be i.i.d. \mathbb{R}^d -valued random variables. Assume $\alpha > 0, 0 < \epsilon < 1$, and $K \geq 1$. Then*

$$\Pr \left(\sup_{g \in \mathcal{G}} \frac{\frac{1}{K} \sum_{j=1}^K g(Z_j) - \mathbb{E}[Z_j]}{\alpha + \frac{1}{K} \sum_{j=1}^K g(Z_j) + \mathbb{E}[Z_j]} > \epsilon \right) \leq 4\mathcal{N}_\infty \left(\frac{\alpha\epsilon}{5}, \mathcal{G} \right) \exp \left(-\frac{3\epsilon^2 \alpha K}{40B} \right).$$

Proof. By Theorem 11.6 from Györfi et al. (2002), we know that

$$\Pr \left(\sup_{g \in \mathcal{G}} \frac{\frac{1}{K} \sum_{j=1}^K g(Z_j) - \mathbb{E}[Z_j]}{\alpha + \frac{1}{K} \sum_{j=1}^K g(Z_j) + \mathbb{E}[Z_j]} > \epsilon \right) \leq 4\mathbb{E} \left[\mathcal{N}_1 \left(\frac{\alpha\epsilon}{5}, \mathcal{G}, \{Z_j\}_{j=1}^K \right) \right] \exp \left(-\frac{3\epsilon^2 \alpha K}{40B} \right),$$

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

$$\frac{1}{K} \sum_{j=1}^K |g(Z_j) - g^l(Z_j)| \leq \frac{\alpha\epsilon}{5}.$$

See Section 11.4 from Györfi et al. (2002) for a detailed proof of the statement above. We then show that for any $\{Z_j\}_{j=1}^K$, $\mathcal{N}_1\left(\frac{\alpha\epsilon}{5}, \mathcal{G}, \{Z_j\}_{j=1}^K\right) \leq \mathcal{N}_\infty\left(\frac{\alpha\epsilon}{5}, \mathcal{G}\right)$. Let $\{\tilde{g}^l\}_{l=1}^L$ 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

$$\frac{1}{K} \sum_{j=1}^K |g(Z_j) - \tilde{g}^l(Z_j)| \leq \frac{1}{K} \sum_{j=1}^K \frac{\alpha\epsilon}{5} = \frac{\alpha\epsilon}{5}.$$

Therefore $\{\tilde{g}^l\}_{l=1}^L$ satisfies the requirement above, concluding our proof. \square

Let $h \in [H], r \in \tilde{\mathcal{R}}$ be arbitrary and fixed and we begin by showing one side of the inequality holds, namely

$$\begin{aligned} & \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \mathbb{E}_{\mu_h} [\|f_h - \mathbb{T}_{h,r}^\pi f'_{h+1}\|^2] - L_{h,r}(f_h, f'_{h+1}, \pi; \mathcal{D}) + \right. \\ & \quad \left. L_{h,r}(\mathbb{T}_{h,r}^\pi f'_{h+1}, f'_{h+1}, \pi; \mathcal{D}) \geq \epsilon(\alpha + \beta + \mathbb{E}_{\mu_h} [\|f_h - \mathbb{T}_{h,r}^\pi f'_{h+1}\|^2])\right) \\ & \leq 14 \left(\mathcal{N}_\infty\left(\frac{\epsilon\beta}{140HR_{\max}}, \mathcal{F}\right) \right)^2 \mathcal{N}_{\infty,1}\left(\frac{\epsilon\beta}{140H^2R_{\max}^2}, \Pi\right) \exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{214(1+\epsilon)H^4R_{\max}^4}\right). \end{aligned}$$

for all $\alpha, \beta > 0, 0 < \epsilon \leq 1/2$.

Recall from (5.1) that

$$g_{f,f'}^\pi(Z) = (f_h(s_h, a_h) - r_h - f'_{h+1}(s_{h+1}, \pi_{h+1}))^2 - (\mathbb{T}_{h,r}^\pi f'_{h+1}(s_h, a_h) - r_h - f'_{h+1}(s_{h+1}, \pi_{h+1}))^2.$$

We begin by showing some basic properties of the random variable $g_{f,f'}^\pi(Z)$. Recall that by definition of the Bellman evaluation operator

$$\mathbb{T}_{h,r}^\pi f'_{h+1}(s_h, a_h) = \mathbb{E}_{\mathbb{P}} [r_h + f'_{h+1}(s_{h+1}, \pi_{h+1}) | s_h, a_h]. \quad (\text{F.1})$$

Since $\mathbb{T}_{h,r}^\pi f'_{h+1}(s_h, a_h) = \mathbb{E}_{\mu_h} [r_h + f'_{h+1}(s_{h+1}, \pi_{h+1}) | s_h, a_h]$, by law of total probability

$$\begin{aligned} \mathbb{E}_{Z \sim \mu_h} [g_{f,f'}^\pi(Z)] &= \mathbb{E}_{s_h, a_h \sim \mu_h} \left[\mathbb{E}_{s_{h+1} \sim \mu_h | s_h, a_h} [(f_h(s_h, a_h) - r_h - f'_{h+1}(s_{h+1}, \pi_{h+1}))^2 - \right. \\ & \quad \left. (\mathbb{T}_{h,r}^\pi f'_{h+1}(s_h, a_h) - r_h - f'_{h+1}(s_{h+1}, \pi_{h+1}))^2 | s_h, a_h] \right] \\ &= \mathbb{E}_{\mu_h} \left[\mathbb{E}_{s_{h+1} \sim \mu_h | s_h, a_h} [(f_h(s_h, a_h) + \mathbb{T}_{h,r}^\pi f'_{h+1}(s_h, a_h) - 2(r_h + f'_{h+1}(s_{h+1}, \pi_{h+1}))) \times \right. \\ & \quad \left. (f_h(s_h, a_h) - \mathbb{T}_{h,r}^\pi f'_{h+1}(s_h, a_h)) | s_h, a_h] \right] \\ &= \mathbb{E}_{\mu_h} [\|f_h(s_h, a_h) - \mathbb{T}_{h,r}^\pi f'_{h+1}(s_h, a_h)\|^2]. \end{aligned}$$

Additionally, recalling that $r_h \in [-R_{\max}, R_{\max}]$, $f'_{h+1} \in [-(H-h)R_{\max}, (H-h)R_{\max}]$, well as $f_h \in [-(H-h+1)R_{\max}, (H-h+1)R_{\max}]$, we know $g_{f,f'}^\pi(Z) \in [-16H^2R_{\max}^2, 16H^2R_{\max}^2]$. Lastly, notice that

$$\begin{aligned} \text{Var}(g_{f,f'}^\pi(Z)) &\leq \mathbb{E}[(g_{f,f'}^\pi(Z))^2] \\ &= \mathbb{E} \left[\mathbb{E}[(f_h(s_h, a_h) + \mathbb{T}_{h,r}^\pi f'_{h+1}(s_h, a_h) - 2(r_h + f'_{h+1}(s_{h+1}, \pi_{h+1})))^2 \times \right. \\ & \quad \left. (f_h(s_h, a_h) - \mathbb{T}_{h,r}^\pi f'_{h+1}(s_h, a_h))^2 | s_h, a_h] \right] \\ &\leq \mathbb{E}[16H^2R_{\max}^2 (f_h(s_h, a_h) - \mathbb{T}_{h,r}^\pi f'_{h+1}(s_h, a_h))^2] = 16H^2R_{\max}^2 \mathbb{E}[g_{f,f'}^\pi(Z)], \end{aligned} \quad (\text{F.2})$$

where for the last inequality we noticed that $f_h(s_h, a_h) + \mathbb{T}_{h,r}^\pi f'_{h+1}(s_h, a_h) - 2(r_h + f'_{h+1}(s_{h+1}, \pi_{h+1}))$ is bounded by $[-4HR_{\max}, 4HR_{\max}]$.

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 $0 < \epsilon \leq \frac{1}{2}$ 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

$$\mathbb{E}[g_{f'_n, f_n}^{\pi_n}(Z) | \{Z_j\}_{j=1}^K] - \frac{1}{K} \sum_{j=1}^K g_{f'_n, f_n}^{\pi_n}(Z_j) \geq \epsilon(\alpha + \beta + \mathbb{E}[g_{f'_n, f_n}^{\pi_n}(Z) | \{Z_j\}_{\tau=1}^K]),$$

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 $g_n = g_{f_n, f'_n}^{\pi_n}$. 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

$$\begin{aligned} & \Pr\left(\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K] - \frac{1}{K} \sum_{j=1}^K g_n(Z'_j) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2}\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K|\{Z_j\}_{j=1}^K]\right) \\ & \leq \frac{\text{Var}(g_n(Z)|\{Z_j\}_{j=1}^K)}{K\left(\frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2}\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K]\right)^2} \\ & \leq \frac{16H^2R_{\max}^2\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K]}{K\left(\frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2}\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K]\right)^2} \\ & \leq \frac{16H^2R_{\max}^2}{\epsilon^2(\alpha + \beta)K}, \end{aligned}$$

where the last inequality comes from the fact that $\frac{s_0}{(a+s_0)^2} \leq \frac{1}{4a}$ for all $s_0 \geq 0$ and $a > 0$. Thus, for all $K \geq \frac{128H^2R_{\max}^2}{\epsilon^2(\alpha + \beta)}$,

$$\Pr\left(\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K] - \frac{1}{K} \sum_{j=1}^K g_n(Z'_j) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2}\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K|\{Z_j\}_{j=1}^K]\right) \leq \frac{7}{8}.$$

We then know that

$$\begin{aligned} & \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^{\pi}(Z'_j) - \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^{\pi}(Z_j) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2}\mathbb{E}[g_{f_h, f'_{h+1}}^{\pi}(Z)]\right) \\ & \geq \Pr\left(\frac{1}{K} \sum_{j=1}^K g_n(Z'_j) - \frac{1}{K} \sum_{j=1}^K g_n(Z_j) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2}\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K]\right) \\ & \geq \Pr\left(\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K] - \frac{1}{K} \sum_{j=1}^K g_n(Z_j) \geq \epsilon(\alpha + \beta) + \epsilon\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K]\right) \\ & \quad \mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K] - \frac{1}{K} \sum_{j=1}^K g_n(Z'_j) \geq \epsilon(\alpha + \beta) + \epsilon\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K] \\ & = \mathbb{E}\left(\mathbb{1}\left\{\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K] - \frac{1}{K} \sum_{j=1}^K g_n(Z_j) \geq \epsilon(\alpha + \beta) + \epsilon\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K]\right\}\right) \\ & \quad \Pr\left(\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K] - \frac{1}{K} \sum_{j=1}^K g_n(Z'_j) \geq \epsilon(\alpha + \beta) + \epsilon\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K]\right) \\ & \geq \frac{7}{8} \Pr\left(\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K] - \frac{1}{K} \sum_{j=1}^K g_n(Z_j) \geq \epsilon(\alpha + \beta) + \epsilon\mathbb{E}[g_n(Z)|\{Z_j\}_{j=1}^K]\right) \\ & = \frac{7}{8} \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \mathbb{E}[g_{f_h, f'_{h+1}}^{\pi}(Z)] - \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^{\pi}(Z_j) \geq \epsilon(\alpha + \beta) + \epsilon\mathbb{E}[g_{f_h, f'_{h+1}}^{\pi}(Z)]\right). \end{aligned}$$

In other words, for $K \geq \frac{128H^2R_{\max}^2}{\epsilon^2(\alpha+\beta)}$,

$$\begin{aligned} & \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \mathbb{E}[g_{f_h, f'_{h+1}}^\pi(Z)] - \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z_j) \geq \epsilon(\alpha + \beta) + \epsilon \mathbb{E}[g_{f_h, f'_{h+1}}^\pi(Z)]\right) \\ & \leq \frac{8}{7} \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z'_j) \right. \\ & \quad \left. - \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z_j) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g_{f_h, f'_{h+1}}^\pi(Z)]\right). \quad (\text{F.3}) \end{aligned}$$

Replacement of Expectation by Empirical Mean of Ghost Sample We begin by noticing

$$\begin{aligned} & \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z'_i) - \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z_j) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g_{f_h, f'_{h+1}}^\pi(Z)]\right) \\ & \leq \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \right. \\ & \quad \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z'_i) - \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z_j) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g_{f_h, f'_{h+1}}^\pi(Z)], \\ & \quad \frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(Z'_i) - \mathbb{E}[(g_{f_h, f'_{h+1}}^\pi)^2(Z)] \leq \\ & \quad \left. \epsilon\left(\alpha + \beta + \frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(Z_j) + \mathbb{E}[(g_{f_h, f'_{h+1}}^\pi)^2(Z)]\right), \quad (\text{F.4}) \right. \\ & \quad \left. \frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(Z'_i) - \mathbb{E}[(g_{f_h, f'_{h+1}}^\pi)^2(Z)] \leq \right. \\ & \quad \left. \epsilon\left(\alpha + \beta + \frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(Z'_i) + \mathbb{E}[(g_{f_h, f'_{h+1}}^\pi)^2(Z)]\right)\right) \\ & + 2 \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{\frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(Z_j) - \mathbb{E}[(g_{f_h, f'_{h+1}}^\pi)^2(Z)]}{\left(\alpha + \beta + \frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(Z_j) + \mathbb{E}[(g_{f_h, f'_{h+1}}^\pi)^2(Z)]\right)}\right). \end{aligned}$$

Citing Theorem F.1, we may bound the second probability term on the right hand side as

$$\begin{aligned} & \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{\frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(Z_j) - \mathbb{E}[(g_{f_h, f'_{h+1}}^\pi)^2(Z)]}{\left(\alpha + \beta + \frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(Z_j) + \mathbb{E}[(g_{f_h, f'_{h+1}}^\pi)^2(Z)]\right)}\right) \\ & \leq 4N_\infty \left(\frac{(\alpha + \beta)\epsilon}{5}, \{g_{f_h, f'_{h+1}}^\pi : f, f' \in \mathcal{F}, \pi \in \Pi\}\right) \exp\left(-\frac{3\epsilon^2(\alpha + \beta)K}{40(16H^2R_{\max}^2)}\right). \end{aligned}$$

For the first probability term, notice that the second event in the conjunction implies

$$(1 + \epsilon) \mathbb{E}[(g_{f_h, f'_{h+1}}^\pi)^2(Z)] \geq (1 - \epsilon) \frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(Z_j) - \epsilon(\alpha + \beta),$$

which is equivalent to

$$\frac{1}{32H^2R_{\max}^2} \mathbb{E}[(g_{f_h, f'_{h+1}}^\pi)^2(Z)] \geq \frac{1 - \epsilon}{32H^2R_{\max}^2(1 + \epsilon)} \frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(Z_j) - \epsilon \frac{(\alpha + \beta)}{32H^2R_{\max}^2(1 + \epsilon)}.$$

A similar bound may be obtained for the term involving Z'_i . Noticing that by equation (F.2), we have $\mathbb{E}[g_{f_h, f'_{h+1}}^\pi(Z)] \geq \frac{1}{16H^2R_{\max}^2} \mathbb{E}[(g_{f_h, f'_{h+1}}^\pi)^2(Z)]$, and we know the first probability term in (F.4) can be bounded by

$$\begin{aligned} & \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z'_i) - \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z_j) \geq \frac{\epsilon}{2}(\alpha + \beta) + \right. \\ & \quad \left. \frac{\epsilon}{2} \left(\frac{1-\epsilon}{32H^2R_{\max}^2(1+\epsilon)} \frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(Z_j) - \frac{\epsilon(\alpha + \beta)}{32H^2R_{\max}^2} + \right. \right. \\ & \quad \quad \left. \left. \frac{1-\epsilon}{32H^2R_{\max}^2(1+\epsilon)} \frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(Z_j) - \frac{\epsilon(\alpha + \beta)}{32H^2R_{\max}^2} \right) \right) \\ & = \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z'_i) - \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z_j) \geq \frac{\epsilon}{2}(\alpha + \beta) - \right. \\ & \quad \left. \frac{\epsilon^2(\alpha + \beta)}{32H^2R_{\max}^2(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R_{\max}^2(1+\epsilon)} \left(\frac{1}{K} \sum_{j=1}^K ((g_{f_h, f'_{h+1}}^\pi)^2(Z'_j) + (g_{f_h, f'_{h+1}}^\pi)^2(Z_j)) \right) \right). \end{aligned}$$

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

$$\begin{aligned} & \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z'_j) - \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z_j) \geq \frac{\epsilon}{2}(\alpha + \beta) - \right. \\ & \quad \left. \frac{\epsilon^2(\alpha + \beta)}{32H^2R_{\max}^2(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R_{\max}^2(1+\epsilon)} \left(\frac{1}{K} \sum_{j=1}^K ((g_{f_h, f'_{h+1}}^\pi)^2(Z'_j) + (g_{f_h, f'_{h+1}}^\pi)^2(Z_j)) \right) \right) \\ & = \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K} \sum_{j=1}^K U_j (g_{f_h, f'_{h+1}}^\pi(Z'_j) - g_{f_h, f'_{h+1}}^\pi(Z_j)) \geq \frac{\epsilon}{2}(\alpha + \beta) - \right. \\ & \quad \left. \frac{\epsilon^2(\alpha + \beta)}{32H^2R_{\max}^2(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R_{\max}^2(1+\epsilon)} \left(\frac{1}{K} \sum_{j=1}^K ((g_{f_h, f'_{h+1}}^\pi)^2(Z'_j) + (g_{f_h, f'_{h+1}}^\pi)^2(Z_j)) \right) \right) \tag{F.5} \\ & \leq 2 \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K} \sum_{j=1}^K |U_j g_{f_h, f'_{h+1}}^\pi(Z_j)| \geq \frac{\epsilon}{4}(\alpha + \beta) - \right. \\ & \quad \left. \frac{\epsilon^2(\alpha + \beta)}{64H^2R_{\max}^2(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R_{\max}^2(1+\epsilon)} \frac{1}{K} \sum_{j=1}^K ((g_{f_h, f'_{h+1}}^\pi)^2(Z_j)) \right). \end{aligned}$$

Conditioning and Covering. We then condition the probability on $\{Z_j\}_{j=1}^K$. Fix some z_1, \dots, z_K and we consider instead

$$\begin{aligned} & \Pr\left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \left| \frac{1}{K} \sum_{j=1}^K U_j g_{f_h, f'_{h+1}}^\pi(z_j) \right| \geq \right. \\ & \quad \left. \frac{\epsilon(\alpha + \beta)}{4} - \frac{\epsilon^2(\alpha + \beta)}{64H^2R_{\max}^2(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R_{\max}^2(1+\epsilon)} \frac{1}{K} \sum_{j=1}^K (g_{f_h, f'_{h+1}}^\pi)^2(z_j) \right). \end{aligned}$$

Let $\delta > 0$ and let \mathcal{G}_δ be an ℓ_∞ δ -cover of $\mathcal{G}_{\mathcal{F}, \Pi} = \{g_{f_h, f'_{h+1}}^\pi : f, f' \in \mathcal{F}, \pi \in \Pi\}$. Fix some $(f, f', \pi) \in \mathcal{F} \times \mathcal{F} \times \Pi$ and there exists some $g \in \mathcal{G}_\delta$ such that $\sup_z |g(z) - g_{f_h, f'_{h+1}}^\pi(z)| < \delta$. We then know that

$$\left| \frac{1}{K} \sum_{j=1}^K U_j g_{f_h, f'_{h+1}}^\pi(z_j) \right| \leq \left| \frac{1}{K} \sum_{j=1}^K U_j g(z_j) \right| + \frac{1}{K} \sum_{j=1}^K |g_{f_h, f'_{h+1}}^\pi(z_j) - g(z_j)| \leq \left| \frac{1}{K} \sum_{j=1}^K U_j g(z_j) \right| + \delta$$

and

$$\begin{aligned}
 \frac{1}{K} \sum_{j=1}^K (g_{f_h, f_{h+1}}^\pi)^2(z_j) &= \frac{1}{K} \sum_{j=1}^K g^2(z_j) + \frac{1}{K} \sum_{j=1}^K ((g_{f_h, f_{h+1}}^\pi)^2(z_j) - g^2(z_j)) \\
 &= \frac{1}{K} \sum_{j=1}^K g^2(z_j) + \frac{1}{K} \sum_{j=1}^K (g_{f_h, f_{h+1}}^\pi(z_j) - g(z_j))(g_{f_h, f_{h+1}}^\pi(z_j) + g(z_j)) \\
 &\geq \frac{1}{K} \sum_{j=1}^K g^2(z_j) - 8H^2 R_{\max}^2 \frac{1}{K} \sum_{j=1}^K |g_{f_h, f_{h+1}}^\pi(z_j) - g(z_j)| \\
 &\geq \frac{1}{K} \sum_{j=1}^K g^2(z_j) - 8H^2 R_{\max}^2 \delta.
 \end{aligned}$$

Set $\delta = \frac{\beta\epsilon}{\text{enumerate5}}$. Notice that as $HR_{\max} \geq 1$, $0 < \epsilon \leq \frac{1}{2}$, we have

$$\frac{\epsilon\beta}{4} - \frac{\epsilon^2\beta}{64H^2R_{\max}^2(1+\epsilon)} - \delta - \delta \frac{\epsilon(1-\epsilon)}{8(1+\epsilon)} = \frac{\epsilon\beta}{2} - \frac{\epsilon^2\beta}{64H^2R_{\max}^2(1+\epsilon)} - \frac{\epsilon^2(1-\epsilon)\beta}{40(1+\epsilon)} \geq 0.$$

Therefore we have

$$\begin{aligned}
 \Pr \left\{ \exists f, f' \in \mathcal{F}, \pi \in \Pi : \left| \frac{1}{K} \sum_{j=1}^K U_j g_{f_h, f_{h+1}}^\pi(z_j) \right| \geq \right. \\
 \left. \frac{\epsilon(\alpha + \beta)}{4} - \frac{\epsilon^2(\alpha + \beta)}{64H^2R_{\max}^2(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R_{\max}^2(1+\epsilon)} \frac{1}{K} \sum_{j=1}^K (g_{f_h, f_{h+1}}^\pi)^2(z_j) \right\} \\
 \leq |\mathcal{G}_{\epsilon\beta/5}| \max_{g \in \mathcal{G}_{\epsilon\beta/5}} \Pr \left\{ \left| \frac{1}{K} \sum_{j=1}^K U_j g(z_j) \right| \geq \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{64H^2R_{\max}^2(1+\epsilon)} + \right. \\
 \left. \frac{\epsilon(1-\epsilon)}{64H^2R_{\max}^2(1+\epsilon)} \frac{1}{K} \sum_{j=1}^K g^2(z_j) \right\}. \tag{F.6}
 \end{aligned}$$

We then apply Bernstein's inequality to bound

$$\Pr \left\{ \left| \frac{1}{K} \sum_{j=1}^K U_j g(z_j) \right| \geq \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{64H^2R_{\max}^2(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R_{\max}^2(1+\epsilon)} \frac{1}{K} \sum_{j=1}^K g^2(z_j) \right\}$$

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,

$$\frac{1}{K} \sum_{j=1}^K \text{Var}(U_j g(z_j)) = \frac{1}{K} \sum_{j=1}^k g^2(z_j) \text{Var}(U_i) = \frac{1}{K} \sum_{j=1}^k g^2(z_j).$$

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 $|V_j| \leq 4H^2R_{\max}^2$. 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

$\sigma^2 = \frac{1}{K} \sum_{j=1}^K \text{Var}(U_j g(z_j)) = \frac{1}{K} \sum_{j=1}^k g^2(z_j)$. We then have for any $g \in \mathcal{G}_{\epsilon\beta/5}$

$$\begin{aligned} & \Pr \left\{ \left| \frac{1}{K} \sum_{j=1}^K U_j g(z_j) \right| \geq \frac{\epsilon\alpha}{4} - \frac{\epsilon^2\alpha}{64H^2R_{\max}^2(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R_{\max}^2(1+\epsilon)} \frac{1}{K} \sum_{j=1}^K g^2(z_j) \right\} \\ &= \Pr \left(\left| \frac{1}{K} \sum_{j=1}^k V_j \right| \geq A_1 + A_2\sigma^2 \right) \\ &\leq 2 \exp \left(-\frac{K(A_1 + A_2\sigma^2)^2}{2\sigma^2 + 2(A_1 + A_2\sigma^2) \frac{8H^2R^2}{3}} \right) \\ &= 2 \exp \left(-\frac{3KA_2}{16H^2R_{\max}^2} \frac{\left(\frac{A_1}{A_2} + \sigma^2\right)^2}{\frac{A_1}{A_2} + \left(1 + \frac{3}{8H^2R_{\max}^2A_2}\right)\sigma^2} \right) \\ &\leq 2 \exp \left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{140H^2R_{\max}^2(1+\epsilon)} \right), \end{aligned}$$

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 (F.5) and (F.6) gives us

$$\begin{aligned} & \Pr \left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z'_j) - \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z_j) \geq \frac{\epsilon}{2}(\alpha + \beta) - \right. \\ & \quad \left. \frac{\epsilon^2(\alpha + \beta)}{32H^2R_{\max}^2(1+\epsilon)} + \frac{\epsilon(1-\epsilon)}{64H^2R_{\max}^2(1+\epsilon)} \left(\frac{1}{K} \sum_{j=1}^K ((g_{f_h, f'_{h+1}}^\pi)^2(Z'_i) + (g_{f_h, f'_{h+1}}^\pi)^2(Z_j)) \right) \right) \\ & \leq 2\mathcal{N}_\infty \left(\frac{\epsilon\beta}{5}, \{g_{f_h, f'_{h+1}}^\pi : f, f' \in F, \pi \in \Pi\} \right) \exp \left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{140H^2R_{\max}^2(1+\epsilon)} \right). \end{aligned}$$

Recalling equations (F.4) and (F.5), we have

$$\begin{aligned} & \Pr \left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z'_i) - \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z_j) \geq \frac{\epsilon}{2}(\alpha + \beta) + \frac{\epsilon}{2} \mathbb{E}[g_{f_h, f'_{h+1}}^\pi(Z)] \right) \\ & \leq 4\mathcal{N}_\infty \left(\frac{\epsilon\beta}{5}, \{g_{f_h, f'_{h+1}}^\pi : f, f' \in F, \pi \in \Pi\} \right) \exp \left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{140H^2R_{\max}^2(1+\epsilon)} \right) \\ & \quad + 8\mathcal{N}_\infty \left(\frac{(\alpha + \beta)\epsilon}{5}, \{g_{f_h, f'_{h+1}}^\pi : f, f' \in \mathcal{F}, \pi \in \Pi\} \right) \exp \left(-\frac{3\epsilon^2(\alpha + \beta)K}{640H^2R_{\max}^2} \right). \end{aligned}$$

Plugging the result back into equation (F.3) and we finally know for $K \geq \frac{128H^2R_{\max}^2}{\epsilon^2(\alpha + \beta)}$,

$$\begin{aligned} & \Pr \left(\exists f, f' \in \mathcal{F}, \pi \in \Pi : \mathbb{E}[g_{f_h, f'_{h+1}}^\pi(Z)] - \frac{1}{K} \sum_{j=1}^K g_{f_h, f'_{h+1}}^\pi(Z_j) \geq \epsilon(\alpha + \beta) + \epsilon \mathbb{E}[g_{f_h, f'_{h+1}}^\pi(Z)] \right) \\ & \leq \frac{32}{7}\mathcal{N}_\infty \left(\frac{\epsilon\beta}{5}, \{g_{f_h, f'_{h+1}}^\pi : f, f' \in F, \pi \in \Pi\} \right) \exp \left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{140H^2R_{\max}^2(1+\epsilon)} \right) \\ & \quad + \frac{64}{7}\mathcal{N}_\infty \left(\frac{(\alpha + \beta)\epsilon}{5}, \{g_{f_h, f'_{h+1}}^\pi : f, f' \in \mathcal{F}, \pi \in \Pi\} \right) \exp \left(-\frac{3\epsilon^2(\alpha + \beta)K}{640H^2R_{\max}^2} \right) \\ & \leq 14\mathcal{N}_\infty \left(\frac{\epsilon\beta}{5}, \{g_{f_h, f'_{h+1}}^\pi : f, f' \in F, \pi \in \Pi\} \right) \exp \left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{214(1+\epsilon)H^4R_{\max}^4} \right). \end{aligned}$$

When $K < \frac{128H^2R_{\max}^2}{\epsilon^2(\alpha + \beta)}$, $\exp \left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{214(1+\epsilon)H^4R_{\max}^4} \right) \geq \exp \left(-\frac{128}{214} \right) \geq \frac{1}{14}$ and the claim trivially holds.

Bounding the Covering Number. Our final task is bounding $\mathcal{N}_\infty \left(\frac{\epsilon\beta}{5}, \{g_{f_h, f'_{h+1}}^\pi : f, f' \in \mathcal{F}, \pi \in \Pi\} \right)$ 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 $\frac{\epsilon\beta}{140H^2R_{\max}^2}$ -covering of Π with respect to $\|\cdot\|_{\infty,1}$. We then know that for any $f, f' \in \mathcal{F}, \pi \in \Pi$, there exists some $f^\dagger, f^\ddagger \in \mathcal{F}_0, \pi^\dagger \in \Pi_0$ such that

$$\begin{aligned} \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} |f_h(s,a) - f_h^\dagger(s,a)| &\leq \frac{\epsilon\beta}{140HR_{\max}}, \\ \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} |f'_{h+1}(s,a) - f_{h+1}^\ddagger(s,a)| &\leq \frac{\epsilon\beta}{140HR_{\max}}, \\ \sup_{s \in \mathcal{S}} \int_{a \in \mathcal{A}} |\pi_{h+1}(a|s) - \pi_{h+1}^\dagger(a|s)| &\leq \frac{\epsilon\beta}{140H^2R_{\max}^2}. \end{aligned}$$

Consider any arbitrary $z = (s, a, r, s') \sim \mu_h$. We know that

$$\begin{aligned} &\left| g_{f_h, f'_{h+1}}^{\pi_{h+1}}(z) - g_{f_h^\dagger, f_{h+1}^\ddagger}^{\pi_{h+1}^\dagger}(z) \right| \\ &= \left| (f_h(s,a) - r - f'_{h+1}(s', \pi_{h+1}))^2 - (\mathbb{T}_{h,r}^{\pi_{h+1}} f'_{h+1}(s,a) - r - f'_{h+1}(s', \pi_{h+1}))^2 - \right. \\ &\quad \left. (f_h^\dagger(s,a) - r - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger))^2 + (\mathbb{T}_{h,r}^{\pi_{h+1}^\dagger} f_{h+1}^\ddagger(s,a) - r - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger))^2 \right| \\ &\leq \left| (f_h(s,a) - r - f'_{h+1}(s', \pi_{h+1}))^2 - (f_h^\dagger(s,a) - r - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger))^2 \right| \\ &\quad + \left| (\mathbb{T}_{h,r}^{\pi_{h+1}} f'_{h+1}(s,a) - r - f'_{h+1}(s', \pi_{h+1}))^2 - (\mathbb{T}_{h,r}^{\pi_{h+1}^\dagger} f_{h+1}^\ddagger(s,a) - r - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger))^2 \right| \\ &\leq \left| f_h(s,a) + f_h^\dagger(s,a) - 2r - f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger) \right| \\ &\quad \times \left| f_h(s,a) - f_h^\dagger(s,a) + f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger) \right| \\ &\quad + \left| \mathbb{T}_{h,r}^{\pi_{h+1}} f'_{h+1}(s,a) + \mathbb{T}_{h,r}^{\pi_{h+1}^\dagger} f_{h+1}^\ddagger(s,a) - 2r - f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger) \right| \\ &\quad \times \left| \mathbb{T}_{h,r}^{\pi_{h+1}} f'_{h+1}(s,a) - \mathbb{T}_{h,r}^{\pi_{h+1}^\dagger} f_{h+1}^\ddagger(s,a) + f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger) \right| \\ &\leq 4HR_{\max} \left| f_h(s,a) - f_h^\dagger(s,a) + f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger) \right| \\ &\quad + 4HR_{\max} \left| \mathbb{T}_{h,r}^{\pi_{h+1}} f'_{h+1}(s,a) - \mathbb{T}_{h,r}^{\pi_{h+1}^\dagger} f_{h+1}^\ddagger(s,a) + f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\ddagger(s', \pi_{h+1}^\dagger) \right|, \end{aligned} \tag{F.7}$$

where for the last inequality we used the boundedness of functions in \mathcal{F}_h and \mathcal{F}_{h+1} . We then notice that

$$\begin{aligned}
 & \left| f_h(s, a) - f_h^\dagger(s, a) + f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\dagger(s', \pi_{h+1}^\dagger) \right| \\
 & \leq |f_h(s, a) - f_h^\dagger(s, a)| + |f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\dagger(s', \pi_{h+1}^\dagger)| \\
 & \leq \frac{\epsilon\beta}{140HR_{\max}} + |f'_{h+1}(s', \pi_{h+1}) - f'_{h+1}(s', \pi_{h+1}^\dagger)| + |f'_{h+1}(s', \pi_{h+1}^\dagger) - f_{h+1}^\dagger(s', \pi_{h+1}^\dagger)| \\
 & \leq \frac{\epsilon\beta}{140HR_{\max}} + \|\pi_{h+1} - \pi_{h+1}^\dagger\|_1 \|f'_{h+1}\|_\infty + |f'_{h+1}(s', \pi_{h+1}^\dagger) - f_{h+1}^\dagger(s', \pi_{h+1}^\dagger)| \\
 & \leq \frac{\epsilon\beta}{140HR_{\max}} + \frac{\epsilon\beta}{140H^2R_{\max}^2} HR_{\max} + |f'_{h+1}(s', \pi_{h+1}^\dagger) - f_{h+1}^\dagger(s', \pi_{h+1}^\dagger)| \\
 & \leq \frac{\epsilon\beta}{140HR_{\max}} + \frac{\epsilon\beta}{140HR_{\max}} + \mathbb{E}_{a' \sim \pi_{h+1}^\dagger(\cdot|s')} [|f'_{h+1}(s', a') - f_{h+1}^\dagger(s', a')|] \\
 & \leq \frac{3\epsilon\beta}{140HR_{\max}},
 \end{aligned}$$

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

$$\begin{aligned}
 & \left| \mathbb{T}_{h,r}^{\pi_{h+1}} f'_{h+1}(s, a) - \mathbb{T}_{h,r}^{\pi_{h+1}^\dagger} f_{h+1}^\dagger(s, a) + f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\dagger(s', \pi_{h+1}^\dagger) \right| \\
 & \leq \left| \mathbb{T}_{h,r}^{\pi_{h+1}} f'_{h+1}(s, a) - \mathbb{T}_{h,r}^{\pi_{h+1}^\dagger} f_{h+1}^\dagger(s, a) \right| + |f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\dagger(s', \pi_{h+1}^\dagger)| \\
 & \leq \left| \mathbb{T}_{h,r}^{\pi_{h+1}} f'_{h+1}(s, a) - \mathbb{T}_{h,r}^{\pi_{h+1}^\dagger} f_{h+1}^\dagger(s, a) \right| + \frac{2\epsilon\beta}{140HR_{\max}} \\
 & \leq \mathbb{E}_{s'' \sim \mathbb{P}_h(\cdot|s,a)} |f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\dagger(s', \pi_{h+1}^\dagger)| + \frac{2\epsilon\beta}{140HR_{\max}} \\
 & \leq \frac{4\epsilon\beta}{140HR_{\max}},
 \end{aligned}$$

where the second inequality uses the same reasoning as above to bound $|f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\dagger(s', \pi_{h+1}^\dagger)|$, the third Jensen's inequality, and the last inequality reuses the bound for $|f'_{h+1}(s', \pi_{h+1}) - f_{h+1}^\dagger(s', \pi_{h+1}^\dagger)|$ over arbitrary s' . Plugging these back into equation (F.7) shows

$$\left| g_{f_h, f_{h+1}}^{\pi_{h+1}}(z) - g_{f_h^\dagger, f_{h+1}^\dagger}^{\pi_{h+1}^\dagger}(z) \right| \leq \frac{7\epsilon\beta}{140HR_{\max}} \times 4HR_{\max} = \frac{\epsilon\beta}{5}.$$

Thus

$$\mathcal{N}_\infty \left(\frac{\epsilon\beta}{5}, \{g_{f_h, f_{h+1}}^\pi : f, f' \in F, \pi \in \Pi\} \right) \leq \left(\mathcal{N}_\infty \left(\frac{\epsilon\beta}{140HR_{\max}}, \mathcal{F} \right) \right)^2 \mathcal{N}_{\infty,1} \left(\frac{\epsilon\beta}{140H^2R_{\max}^2}, \Pi \right),$$

showing one side of the inequality holds.

To show the other side holds, simply replace $g_{f, f'}^\pi(Z)$ defined in equation 5.1 with its negative and repeat the analysis above. We then complete the proof by taking a union bound over both halves. \square

F.2. 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 D.3. Taking a union bound over all $h \in [H]$ and reported reward $r \in \tilde{\mathcal{R}}$ recalling that $|\tilde{\mathcal{R}}| \leq n + 1 \leq 2n$,

by Lemma D.2, we have

$$\begin{aligned} & \Pr\left(\exists h \in [H], r \in \tilde{\mathcal{R}}, f, f' \in \mathcal{F}, \pi \in \Pi : \right. \\ & \quad \left| \mathbb{E}_{\mu_h} [\|f_h - \mathbb{T}_{h,r}^\pi f'_{h+1}\|^2] - L_{h,r}(f_h, f'_{h+1}, \pi; \mathcal{D}) + L_{h,r}(\mathbb{T}_{h,r}^\pi f'_{h+1}, f'_{h+1}, \pi; \mathcal{D}) \right| \\ & \quad \geq \epsilon (\alpha + \beta + \mathbb{E}_{\mu_h} [\|f_h - \mathbb{T}_{h,r}^\pi f'_{h+1}\|^2]) \Big) \\ & \leq 56nH \left(\mathcal{N}_\infty \left(\frac{\epsilon\beta}{140HR_{\max}}, \mathcal{F} \right) \right)^2 \mathcal{N}_{\infty,1} \left(\frac{\epsilon\beta}{140H^2R_{\max}^2}, \Pi \right) \exp \left(-\frac{\epsilon^2(1-\epsilon)\alpha K}{214(1+\epsilon)H^4R_{\max}^4} \right). \end{aligned}$$

Letting $\alpha = \beta$ and $\epsilon = \frac{1}{2}$, setting the right hand side to δ , and solving for α gives us

$$\alpha \leq \frac{1}{K} \max \left\{ 5136H^4R_{\max}^4, 5136H^4R_{\max}^4 \log \frac{56nH\mathcal{N}_\infty \left(\frac{HR_{\max}}{K}, \mathcal{F} \right) \mathcal{N}_{\infty,1} \left(\frac{1}{K}, \Pi \right)}{\delta} \right\}.$$

As $\log 56 \geq 1$, $n, H \geq 1$, and $0 < \delta < 1$, the second term always dominates the first and we can simplify the inequality as

$$\alpha \leq \frac{5136H^4R_{\max}^4}{K} \log \frac{56nH\mathcal{N}_\infty \left(\frac{19H^3R_{\max}^3}{K}, \mathcal{F} \right) \mathcal{N}_{\infty,1} \left(\frac{19H^4R_{\max}^4}{K}, \Pi \right)}{\delta},$$

completing the proof. \square

Proof of Corollary D.4. For convenience, let $\hat{g}_{h,r}^\pi = \arg \min_{g \in \mathcal{F}_h} L_{h,r}(g, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D})$. We then know that

$$\begin{aligned} B_{h,r}(f_{h,r}^{\pi,*}, \pi; \mathcal{D}) &= L_{h,r}(f_{h,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) - L_{h,r}(\hat{g}_{h,r}^\pi, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) \\ &= L_{h,r}(f_{h,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) - L_{h,r}(\mathbb{T}_{h,r}^{\pi,*} f_{h+1,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) \\ &\quad - \left(L_{h,r}(\hat{g}_{h,r}^\pi, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) - L_{h,r}(\mathbb{T}_{h,r}^{\pi,*} f_{h+1,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) \right). \end{aligned}$$

By Lemma D.3, conditionally on the event $\mathcal{G}(\Pi)$ we have the following simultaneously:

$$\begin{aligned} L_{h,r}(f_{h,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) - L_{h,r}(\mathbb{T}_{h,r}^{\pi,*} f_{h+1,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) &\leq \epsilon_S + \frac{3}{2} \mathbb{E}_{\mu_h} \left[\|f_{h,r}^{\pi,*} - \mathbb{T}_{h,r}^{\pi,*} f_{h+1,r}^{\pi,*}\|^2 \right], \\ -L_{h,r}(\hat{g}_{h,r}^\pi, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) + L_{h,r}(\mathbb{T}_{h,r}^{\pi,*} f_{h+1,r}^{\pi,*}, f_{h+1,r}^{\pi,*}, \pi; \mathcal{D}) &\leq \epsilon_S, \end{aligned}$$

where the second inequality uses the fact that $\|\cdot\|^2$ is non-negative. Finally, noticing that

$$\begin{aligned} \mathbb{E}_{\mu_h} \left[\|f_{h,r}^{\pi,*} - \mathbb{T}_{h,r}^{\pi,*} f_{h+1,r}^{\pi,*}\|^2 \right] &\leq 2\mathbb{E}_{\mu_h} \left[\|f_{h,r}^{\pi,*} - Q_h^\pi(\cdot, \cdot; r)\|^2 \right] + 2\mathbb{E}_{\mu_h} \left[\|\mathbb{T}_{h,r}^{\pi,*} f_{h+1,r}^{\pi,*} - \mathbb{T}_{h,r}^{\pi,*} Q_h^\pi(\cdot, \cdot; r)\|^2 \right] \\ &\leq 2\epsilon_{\mathcal{F}} + 2\mathbb{E}_{\mu'_{h+1}} \left[\|f_{h+1,r}^{\pi,*} - Q_{h+1}^\pi(\cdot, \cdot; r)\|^2 \right] \\ &\leq 4\epsilon_{\mathcal{F}}, \end{aligned}$$

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 2.3. \square

Proof of Corollary D.5. Let $\hat{g}_{h,r}^\pi = \arg \min_{g \in \mathcal{F}_h} \mathbb{E}_{\mu_h} [\|g - \mathbb{T}_{h,r}^\pi f_{h+1,r}^\pi\|^2]$. Recalling the definition of $B_{h,r}$, we have

$$\begin{aligned} B_{h,r}(f_{h,r}^\pi, \pi; \mathcal{D}) &= L_{h,r}(f_{h,r}^\pi, f_{h+1,r}^\pi, \pi; \mathcal{D}) - \min_{g \in \mathcal{F}_h} L_{h,r}(g, f_{h+1,r}^\pi, \pi; \mathcal{D}) \\ &\geq L_{h,r}(f_{h,r}^\pi, f_{h+1,r}^\pi, \pi; \mathcal{D}) - L_{h,r}(\hat{g}_{h,r}^\pi, f_{h+1,r}^\pi, \pi; \mathcal{D}) \\ &= L_{h,r}(f_{h,r}^\pi, f_{h+1,r}^\pi, \pi; \mathcal{D}) - L_{h,r}(\mathbb{T}_{h,r}^\pi f_{h+1,r}^\pi, f_{h+1,r}^\pi, \pi; \mathcal{D}) \\ &\quad - \left(L_{h,r}(\hat{g}_{h,r}^\pi, f_{h+1,r}^\pi, \pi; \mathcal{D}) - L_{h,r}(\mathbb{T}_{h,r}^\pi f_{h+1,r}^\pi, f_{h+1,r}^\pi, \pi; \mathcal{D}) \right). \end{aligned}$$

By Lemma D.3, conditionally on the event $\mathcal{G}(\Pi)$ we have the following:

$$\begin{aligned} L_{h,r}(f_{h,r}^\pi, f_{h+1,r}^\pi, \pi; \mathcal{D}) - L_{h,r}(\mathbb{T}_{h,r}^\pi f_{h+1,r}^\pi, f_{h+1,r}^\pi, \pi; \mathcal{D}) &\geq -\epsilon_S + \frac{1}{2} \mathbb{E}_{\mu_h} [\|f_{h,r}^\pi - \mathbb{T}_{h,r}^\pi f_{h+1,r}^\pi\|^2], \\ -L_{h,r}(\widehat{g}_{h,r}^\pi, f_{h+1,r}^\pi, \pi; \mathcal{D}) + L_{h,r}(\mathbb{T}_{h,r}^\pi f_{h+1,r}^\pi, f_{h+1,r}^\pi, \pi; \mathcal{D}) &\geq -\epsilon_S - \frac{3}{2} \mathbb{E}_{\mu_h} [\|\widehat{g}_{h,r}^\pi - \mathbb{T}_{h,r}^\pi f_{h+1,r}^\pi\|^2]. \end{aligned}$$

Recalling that $B_{h,r}(f, \pi; \mathcal{D}) \leq \epsilon_0$, we have

$$\mathbb{E}_{\mu_H} [\|f_{h,r}^\pi - \mathbb{T}_{h,r}^\pi f_{h+1,r}^\pi\|^2] \leq 4\epsilon_S + 3\mathbb{E}_{\mu_h} [\|\widehat{g}_{h,r}^\pi - \mathbb{T}_{h,r}^\pi f_{h+1,r}^\pi\|^2] + 2\epsilon_0.$$

We conclude our proof by reminding ourselves of Assumption 2.4. □