Reinforcement Learning from Partial Observation: Linear Function Approximation with Provable Sample Efficiency

Qi Cai [1] **Zhuoran Yang** [2] **Zhaoran Wang** [1]

Abstract

We study reinforcement learning for partially observed Markov decision processes (POMDPs) with infinite observation and state spaces, which remains less investigated theoretically. To this end, we make the first attempt at bridging partial observability and function approximation for a class of POMDPs with a linear structure. In detail, we propose a reinforcement learning algorithm (Optimistic Exploration via Adversarial Integral Equation or OP-TENET) that attains an ϵ-optimal policy within $O(1/\epsilon^2)$ episodes. In particular, the sample complexity scales polynomially in the intrinsic dimension of the linear structure and is independent of the size of the observation and state spaces. The sample efficiency of OP-TENET is enabled by a sequence of ingredients: (i) a Bellman operator with finite memory, which represents the value function in a recursive manner, (ii) the identification and estimation of such an operator via an adversarial integral equation, which features a smoothed discriminator tailored to the linear structure, and (iii) the exploration of the observation and state spaces via optimism, which is based on quantifying the uncertainty in the adversarial integral equation.

1. Introduction

Partial observability poses significant challenges for reinforcement learning, especially when the observation and state spaces are infinite. Given full observability, reinforcement learning is well studied empirically (Mnih et al., 2015; Silver et al., 2016; 2017) and theoretically (Auer et al., 2008; Osband et al., 2016; Azar et al., 2017; Jin et al., 2018; Yang & Wang, 2020; Jin et al., 2020b; Ayoub et al., 2020; Kakade et al., 2020; Du et al., 2021). In particular, for infinite state spaces, neural function approximators achieve remarkable successes empirically (Mnih et al., 2015; Berner et al., 2019; Arulkumaran et al., 2019), while linear function approximators become better understood theoretically (Yang & Wang, 2020; Jin et al., 2020b; Ayoub et al., 2020; Kakade et al., 2020; Du et al., 2021). In contrast, reinforcement learning in partially observed Markov decision processes (POMDPs) is less investigated theoretically despite its prevalence in practice (Cassandra et al., 1996; Hauskrecht & Fraser, 2000; Brown & Sandholm, 2018; Rafferty et al., 2011).

More specifically, partial observability poses both statistical and computational challenges. From a statistical perspective, it is challenging to predict future rewards, observations, or states due to a lack of the Markov property. In particular, predicting the future often involves inferring the distribution of the state (also known as the belief state) or its functionals as a summary of the history, which is already challenging even assuming the (observation) emission and (state) transition kernels are known (Vlassis et al., 2012; Golowich et al., 2022). Meanwhile, learning the emission and transition kernels faces various issues commonly encountered in causal inference (Zhang & Bareinboim, 2016). For example, they are generally nonidentifiable (Kallus et al., 2021). Even assuming they are identifiable, their estimation possibly requires a sample size that scales exponentially in the horizon and dimension (Jin et al., 2020a). Such statistical challenges are already prohibitive even for the evaluation of a policy (Nair & Jiang, 2021; Kallus et al., 2021; Bennett & Kallus, 2021), which forms the basis of policy optimization. From a computational perspective, it is known that policy optimization is generally intractable (Vlassis et al., 2012; Golowich et al., 2022). Moreover, infinite observation and state spaces amplify both statistical and computational challenges. On the other hand, most existing results are restricted to the tabular setting (Azizzadenesheli et al., 2016; Guo et al., 2016; Jin et al., 2020a; Xiong et al., 2021), where the observation and state spaces are finite.

In this paper, we study linear function approximation in POMDPs to address the statistical challenges amplified by infinite observation and state spaces. In particular, our con-

[1]Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, USA [2]Department of Statistics and Data Science, Yale University, New Haven, USA. Correspondence to: Qi Cai <qicai2022@u.northwestern.edu>, Zhuoran Yang <zhuoran.yang@yale.edu>, Zhaoran Wang <zhaoranwang@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

tribution is fourfold. First, we define a class of POMDPs with a linear structure and identify an ill conditioning measure for sample-efficient reinforcement learning. Such an ill conditioning measure corresponds to the undercompleteness in the tabular setting (Jin et al., 2020a). Second, we propose a reinforcement learning algorithm (OP-TENET), which applies to any POMDP admitting the aforementioned linear structure. Moreover, we use a minimax optimization formulation in OP-TENET such that the algorithm can be implemented in a computation-efficient manor even if the dataset is large. Third, we prove in theory that OP-TENET attains an ϵ-optimal policy within $O(1/\epsilon^2)$ episodes. In particular, the sample complexity scales polynomially in the intrinsic dimension of the linear structure and is independent of the size of the observation and state spaces. Fourth, our algorithm and analysis are based on new tools. In particular, the sample efficiency of OP-TENET is enabled by a sequence of ingredients: (i) a Bellman operator with finite memory, which represents the value function in a recursive manner, (ii) the identification and estimation of such an operator via an adversarial integral equation, which features a smoothed discriminator tailored to the linear structure, and (iii) the exploration of the observation and state spaces via optimism, which is based on quantifying the uncertainty in the adversarial integral equation.

1.1. Related Work

Our work is related to a line of recent work on the sample efficiency of reinforcement learning for POMDPs. In detail, Azizzadenesheli et al. (2016); Guo et al. (2016); Xiong et al. (2021) establish sample complexity guarantees for searching the optimal policy in POMDPs whose models are identifiable and can be estimated by spectral methods. However, Azizzadenesheli et al. (2016) and Guo et al. (2016) add extra assumptions such that efficient exploration of the POMDP can always be achieved by running arbitrary policies. In contrast, the upper bound confidence (UCB) method is used in Xiong et al. (2021) for adaptive exploration. However, they require strictly positive state transition and observation emission kernels to ensure fast convergence to the stationary distribution. The more related work is Jin et al. (2020a), which considers undercomplete POMDPs, in other words, the observations are more than the latent states. Their proposed algorithm can attain the optimal policy without estimating the exact model, but an observable component (Jaeger, 2000; Hsu et al., 2012), which is the same for our algorithm design, while only applies to tabular POMDPs.

In a broader context of reinforcement learning with partial observability, our work is related to several recent works on POMDPs with special structures. For example, Kwon et al. (2021) considers latent POMDPs, where each process has only one latent state, and the proposed algorithm efficiently infers the latent state using a short trajectory. Kozuno et al. (2021) considers POMDPs having tree-structured states with their positions in certain partitions being the observations. Compared with general POMDPs, these specially structures reduce the complexity of finding the optimal actions, and the corresponding algorithms use techniques closer to those for MDPs. Also, the aforementioned literature only consider tabular POMDPs.

In the contexture of reinforcement learning with function approximations, our work is related to a vast body of recent progress (Yang & Wang, 2020; Jin et al., 2020b; Cai et al., 2020; Du et al., 2021; Kakade et al., 2020; Agarwal et al., 2020; Zhou et al., 2021; Ayoub et al., 2020) on the sample efficiency of reinforcement learning for MDPs with linear function approximations. These works characterize the uncertainty in the regression for estimating either the model or value function of an MDP and use the uncertainty as a bonus on the rewards to encourage exploration. However, none of these approaches directly apply to POMDPs due to the latency of the states.

1.2. Notation

For any discrete or continuous set $\mathcal{X}$ and $p \in \mathbb{N}$, we denote by $L^p(\mathcal{X})$ the L^p space of functions over $\mathcal{X}$, and $\Delta(\mathcal{X})$ the set of probability density functions over $\mathcal{X}$ when $\mathcal{X}$ is continuous or probability mass functions when $\mathcal{X}$ is discrete. For any $d \in \mathbb{N}$, we denote by $[d]$ the set of integers from 1 to d. For a vector v and a matrix M, we denote by $[v]_i$ the i-th entry of v and $[M]_{i,j}$ the entry of M at the i-th row and j-th column. We denote by $\|\cdot\|_p$ the ℓ^p-norm of a vector or L^p-norm of a function. Also, for an operator M, we denote by $\|M\|_{p\mapsto q}$ the operator norm of M induced by the ℓ^p-norm or L^p-norm of the domain and ℓ^q-norm or L^q-norm of the range. We use the notation $\mathrm{linspan}(\cdot)$ and $\mathrm{conh}(\cdot)$ to represent the linear span and convex combination, respectively.

2. Background

2.1. POMDPs

We consider an episodic POMDP $(\mathcal{S}, \mathcal{A}, \mathcal{O}, H, \mathcal{T}, \mathcal{E}, \mu, r)$, where $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{O}$ are the state, action, and observation spaces, respectively, H is the length of each episode, $\mathcal{T}$ is the state transition kernel from a state-action pair to the next state, $\mathcal{E}$ is the observation emission kernel from a state to its observation, μ is the initial state distribution, and $r : \mathcal{O} \times \mathcal{A} \to [0, 1]$ is the reward function defined on the observation and action for each step. We assume that the action space $\mathcal{A}$ has a finite size $A \in \mathbb{N}$, but the state space $\mathcal{S}$ and observation space $\mathcal{O}$ can be infinite (with finite dimensions). Also, we consider the nonhomongeneous setting so that the state transition kernel and observation emission kernel can be different across each step. Hence, we use a

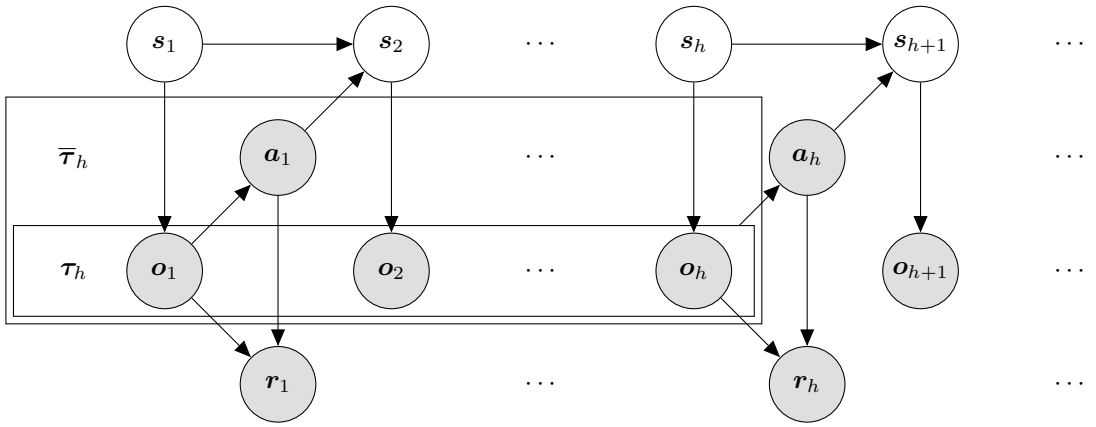

Figure 1: Directed acyclic graph of a POMDP. Here, we denote by $\tau_h = (o_1, \ldots, o_h)$ the observation history and $\overline{\tau}_h = (o_1, a_1, \ldots, o_{h-1}, a_{h-1}, o_h)$ the full history. See Section 2.1 for more details.

subscript $h \in \mathbb{N}$ to index the step. At the beginning of each episode, the agent receives the initial state $s_1 \sim \mu$. Then, the agent interacts with the environment as follows. At the h-th step, the agent receives the observation $o_h \sim \mathcal{E}_h(\cdot \,|\, s_h)$, takes an action a_h based on the observation history

$$\tau_h = (o_1, \ldots, o_h), \tag{2.1}$$

and receives the reward $r_h = r(o_h, a_h)$. Any mapping π from the observation history to the action is called a (deterministic) policy. We denote by Π the set of all such mappings. Note that the policy does not use the action history as an input. Such a restriction does not exclude the optimal policy, as the action history can be decoded from the observation history. Subsequently, the agent receives the next state s_{h+1} following $s_{h+1} \sim \mathcal{T}_h(\cdot \,|\, s_h, a_h)$. See Figure 1 for an illustration.

In a reinforcement learning problem, the environment is unknown, that is, the state transition kernel $\mathcal{T}$ and observation emission kernel $\mathcal{E}$ are unknown. We denote by $\{(\mathcal{T}^\theta, \mathcal{E}^\theta) : \theta \in \Theta\}$ the candidate class of $\mathcal{T}$ and $\mathcal{E}$, where θ is the parameter and Θ is the set of the parameter. We assume that the realizability condition holds, that is, there exists a parameter $\theta^* \in \Theta$ such that $\mathcal{T} = \mathcal{T}^{\theta^*}$ and $\mathcal{E} = \mathcal{E}^{\theta^*}$. Without loss of generality and for ease of presentation, we assume that μ, $\mathcal{T}_1$, $\mathcal{E}_1$, and $\mathcal{E}_2$ are known, which only account for the initialization. The goal is to find a policy that maximizes the expected total reward, that is,

$$\pi^* = \operatorname*{argmax}_{\pi \in \Pi} J(\theta^*, \pi), \tag{2.2}$$

$$\text{where } J(\theta, \pi) = \mathbb{E}_{\theta,\pi}\Big[\sum_{h=1}^{H} r_h\Big] \text{ for any } (\theta, \pi) \in \Theta \times \Pi.$$

Here, we write θ and π as the subscripts of the expectation to denote that the parameter of the state transition kernel and observation emission kernel have the parameter θ and the actions follow the policy π. In the sequel, we drop the subscript π if the expectation does not depends on it.

Additional Notation: Recall that we denote by Π the set of all policies. For notational simplicity, we denote by $\overline{\Pi}$ the set of mixing policies. A mixing policy selects a policy from Π randomly and executes such a policy throughout the episode. For any $h \in \mathbb{N}$, we denote by $\overline{\tau}_h$ the full history,

$$\overline{\tau}_h = (o_1, a_1, \ldots, o_{h-1}, a_{h-1}, o_h), \tag{2.3}$$

which includes the action history. We denote by Γ_h and $\overline{\Gamma}_h$ the sets of all histories τ_h and $\overline{\tau}_h$, respectively. Throughout the paper, we use bold letters for states, actions, and observations to emphasize that they are random variables in a POMDP, whose parameter and policy are specified in the context, while we use regular letters when they are deterministic values.

2.2. Linear Function Approximations

We specify the candidate class of the state transition kernel $\mathcal{T}$ and observation emission kernel $\mathcal{E}$. We define the following function classes of the conditional state distribution. In detail, we define

$$\begin{aligned}\mathcal{F}_{\mathrm{s}} = \{&p_\theta(s_h = \cdot \,|\, s_{h-1} = s, a_{h-1} = a) : \\ &(h, \theta, s, a) \in [H] \times \Theta \times \mathcal{S} \times \mathcal{A}\}, \\ \mathcal{F}'_{\mathrm{s}} = \{&p_{\theta,\pi}(s_h = \cdot \,|\, o_{h+1} = o, a_h = a) : \\ &(h, \theta, o, a, \pi) \in [H] \times \Theta \times \mathcal{O} \times \mathcal{A} \times \overline{\Pi}\}.\end{aligned}$$

Here, $p(\cdot)$ is the probability density function when the state space $\mathcal{S}$ is continuous and the probability mass function when $\mathcal{S}$ is discrete. The subscripts θ and π follow from (2.2). Note that conditioning on $a_h = a$ means that the agent takes the action a at the h-th step regardless of the

observation history, while the agent takes the other actions following the policy π as specified in the subscript. In the causal inference literature (Pearl, 2009), our notation corresponds to $\mathrm{do}(\boldsymbol{a}_h = a)$, which denotes the interventional distribution and differs from the observational distribution. Throughout this paper, we follow such a convention. Also, note that $\mathcal{F}_{\mathrm{s}}$ corresponds to the state distribution conditioning on the past, while $\mathcal{F}'_{\mathrm{s}}$ corresponds to that conditioning on the future. As a special case, we have $\mu \in \mathcal{F}_{\mathrm{s}}$ for $h = 1$ since $\boldsymbol{s}_0$ and $\boldsymbol{a}_0$ do not exist. We define the following function class of the conditional observation distribution,

$$\mathcal{F}_{\mathrm{o}} = \{p_{\theta,\pi}(\boldsymbol{o}_{h:h+2} = \cdot \,\big|\, \boldsymbol{a}_h = a, \boldsymbol{a}_{h+1} = a') : (h, \theta, a, a', \pi) \in [H] \times \Theta \times \mathcal{A}^2 \times \overline{\Pi}\}.$$

The following assumption restricts the above function classes to two low-dimensional subspaces.

Assumption 2.1 (Linear Function Approximations)**.** There exist $d_{\mathrm{s}}, d_{\mathrm{o}} \in \mathbb{N}$ and known distribution functions $\{\psi_i\}_{i=1}^{d_{\mathrm{s}}} \subset \Delta(\mathcal{S})$ and $\{\phi_i\}_{i=1}^{d_{\mathrm{o}}} \subset \Delta(\mathcal{O}^3)$ such that we have

- $\mathcal{F}_{\mathrm{s}}, \mathcal{F}'_{\mathrm{s}} \subset \mathrm{conh}(\{\psi_i\}_{i=1}^{d_{\mathrm{s}}})$,
- $\mathcal{F}_{\mathrm{o}} \subset \mathrm{conh}(\{\phi_i\}_{i=1}^{d_{\mathrm{o}}})$.

For ease of presentation, we denote $\{\psi_i\}_{i=1}^{d_{\mathrm{s}}}$ and $\{\phi_i\}_{i=1}^{d_{\mathrm{o}}}$ by ψ and ϕ, respectively, for the rest of the paper. Assumption 2.1 requires that $\mathcal{F}_{\mathrm{s}}$, $\mathcal{F}'_{\mathrm{s}}$, and $\mathcal{F}_{\mathrm{o}}$ are linearly represented by known bases ψ and ϕ. See, for example, Du et al. (2021) for the corresponding assumption in MDPs. Note that, when ψ and ϕ are the one-hot functions over $\mathcal{S}$ and $\mathcal{O}^3$, respectively, we recover the tabular setting (Jin et al., 2020a).

The following assumption ensures that the observation is informative for the state. For any $(h, \theta) \in [H] \times \Theta$, we define the observation operator $\mathbb{O}_h^\theta : L^1(\mathcal{S}) \to L^1(\mathcal{O})$ by

$$(\mathbb{O}_h^\theta f)(o) = \int_{\mathcal{S}} \mathcal{E}_h^\theta(o \,|\, s) \cdot f(s) \,\mathrm{d}s, \tag{2.4}$$

for any $f \in L^1(\mathcal{S})$ and $o \in \mathcal{O}$, which maps a state distribution to the observation distribution.

Assumption 2.2 (Invertible Observation Operators)**.** For any $(h, \theta) \in [H] \times \Theta$, there exist a known function $\mathcal{Z}_h^\theta : \mathcal{S} \times \mathcal{O} \to \mathbb{R}$ and the linear operator $\mathbb{Z}_h^\theta : L^1(\mathcal{O}) \to L^1(\mathcal{S})$ defined by

$$(\mathbb{Z}_h^\theta f)(s) = \int_{\mathcal{O}} \mathcal{Z}_h^\theta(s, o) \cdot f(o) \,\mathrm{d}o,$$

for any $f \in L^1(\mathcal{O})$ and $s \in \mathcal{S}$ such that we have

- $\mathbb{Z}_h^\theta \mathbb{O}_h^\theta f = f$ for any $f \in \mathrm{linspan}(\psi)$,
- $\|\mathbb{Z}_h^\theta\|_{1 \mapsto 1} \le \gamma$ for a constant $\gamma > 0$.

Assumption 2.2 requires that the observation operator $\mathbb{O}_h^\theta$ defined on $\mathrm{linspan}(\psi)$ is injective, which implies that it has a left inverse $\mathbb{Z}_h^\theta$. Note that the domain of $\mathbb{Z}_h^\theta$ naturally extends to $L^1(\mathcal{O})$. In other words, the observation distribution carries the full information of the state distribution. The (upper bound of the) operator norm γ is a measure of ill conditioning, which quantifies the fundamental difficulty, in terms of the information-theoretic limit, of reinforcement learning in the POMDP. See more discussion in Section B, where we prove that both Assumptions 2.1 and 2.2 hold if the state transition kernel and observation emission kernel admit certain a structure. Correspondingly, we provide a detailed form of the function $\mathcal{Z}_h^\theta$ in Section B. Also, we illustrate the connection to the tabular setting (Jin et al., 2020a) therein.

3. Algorithm

In this section, we first introduce the finite-memory Bellman operator in Section 3.1 and discuss its estimation in Section 3.2. Then, we present Algorithm 1, which performs optimistic exploration on top of operator estimation, in Section 3.3.

3.1. Finite-Memory Bellman Operator

To cast a POMDP as an MDP, it is necessary to aggregate the observation history and action history as the "state" in an MDP to retrieve the Markov property. In detail, for any $(h, \theta, \pi) \in [H] \times \Theta \times \Pi$, we define the full-memory Bellman operator $\mathbb{P}_h^{\theta,\pi} : L^\infty(\overline{\Gamma}_{h+1}) \to L^\infty(\overline{\Gamma}_h)$ by

$$\begin{aligned}
&(\mathbb{P}_h^{\theta,\pi} f)(\overline{\tau}_h) \\
&\quad = \mathbb{E}_{\theta,\pi}[f(\overline{\boldsymbol{\tau}}_{h+1}) \,|\, \overline{\boldsymbol{\tau}}_h = \overline{\tau}_h] \\
&\quad = \int_{\mathcal{O}} p_\theta\big(\boldsymbol{o}_{h+1} = o_{h+1} \,|\, \overline{\boldsymbol{\tau}}_h = \overline{\tau}_h, \boldsymbol{a}_h = \pi(\tau_h)\big) \\
&\qquad\qquad \cdot f\big(\overline{\tau}_h, \pi(\tau_h), o_{h+1}\big) \,\mathrm{d}o_{h+1},
\end{aligned} \tag{3.1}$$

for any $f \in L^\infty(\overline{\Gamma}_{h+1})$ and $\overline{\tau}_h \in \overline{\Gamma}_h$. Here, the second equality follows from $\overline{\boldsymbol{\tau}}_{h+1} = (\overline{\boldsymbol{\tau}}_h, \boldsymbol{a}_h, \boldsymbol{o}_{h+1})$ with $\boldsymbol{a}_h = \pi(\overline{\boldsymbol{\tau}}_h)$, which is defined in (2.3). In the sequel, the function f is set as the expected total reward conditioning on the $(h+1)$-step full history $\overline{\tau}_{h+1} \in \overline{\Gamma}_{h+1}$ and $\mathbb{P}_h^{\theta,\pi}$ maps it to the h-step counterpart, which resembles backward induction or dynamic programming in MDPs. We denote by $R : \overline{\Gamma}_{H+1} \to [0, H]$ the function that maps the $(H+1)$-step full history to the total reward, that is,

$$R(\overline{\tau}_{H+1}) = r(o_1, a_1) + \cdots + r(o_H, a_H), \tag{3.2}$$

for any $\overline{\tau}_{H+1} \in \overline{\Gamma}_{H+1}$. For any $h \in [H]$, the expected total reward satisfies

$$\mathbb{E}_{\theta,\pi}\Big[\sum_{i=1}^{H} \boldsymbol{r}_i \,\Big|\, \overline{\boldsymbol{\tau}}_h = \overline{\tau}_h\Big] = (\mathbb{P}_h^{\theta,\pi} \cdots \mathbb{P}_H^{\theta,\pi} R)(\overline{\tau}_h), \tag{3.3}$$

for any $\overline{\tau}_h \in \overline{\Gamma}_h$, where the equality follows from recursively applying (3.1) and the tower property of conditional expectation. A direct idea is to evaluate a policy π by estimating the parameter θ in (3.3) and optimize π in an iterative manner. However, estimating the operator $\mathbb{P}_h^{\theta,\pi}$ suffers from the curse of dimensionality since it requires estimating a distribution conditioning on the h-step full history $\overline{\boldsymbol{\tau}}_h$, which is high-dimensional.

From Full Memory to Finite Memory: We propose to bypass such an issue by exploiting the independence between the past observation and future observation conditioning on the current state. In detail, for any $(h, \theta, \pi) \in [H] \times \Theta \times \Pi$, we define the finite-memory Bellman operator $\mathbb{B}_h^{\theta,\pi} : L^\infty(\overline{\Gamma}_{h+1}) \to L^\infty(\overline{\Gamma}_h)$ by

$$(\mathbb{B}_h^{\theta,\pi} f)(\overline{\tau}_h) = \int_{\mathcal{O}^2} f(\overline{\tau}_h^\dagger, \pi(\tau_h^\dagger), \widetilde{o}_{h+1}) \cdot \mathcal{B}_{h,\pi(\tau_h^\dagger)}^\theta(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \, \mathrm{d}\widetilde{o}_h \, \mathrm{d}\widetilde{o}_{h+1}, \tag{3.4}$$

for any $f \in L^\infty(\overline{\Gamma}_{h+1})$ and $\overline{\tau}_h \in \overline{\Gamma}_h$. Here, the tail-mirrored full history $\overline{\tau}_h^\dagger$ and tail-mirrored observation history $\tau_h^\dagger$ are defined by

$$\overline{\tau}_h^\dagger = (\overline{\tau}_{h-1}, a_{h-1}, \widetilde{o}_h), \quad \tau_h^\dagger = (\tau_{h-1}, \widetilde{o}_h), \tag{3.5}$$

which switch the last observation o_h by $\widetilde{o}_h$ in the full history $\overline{\tau}_h$ and observation history τ_h, that is, $\overline{\tau}_h = (\overline{\tau}_{h-1}, a_{h-1}, o_h)$ and $\tau_h = (\tau_{h-1}, o_h)$. Also, the function $\mathcal{B}_{h,a}^\theta : \mathcal{O}^3 \to \mathbb{R}$ is defined by

$$\mathcal{B}_{h,a}^\theta(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) = \int_{\mathcal{S}} p_\theta(\widetilde{\boldsymbol{o}}_h = \widetilde{o}_h, \widetilde{\boldsymbol{o}}_{h+1} = \widetilde{o}_{h+1} \,|\, \widetilde{\boldsymbol{s}}_h = \widetilde{s}_h, \widetilde{\boldsymbol{a}}_h = a) \cdot \mathcal{Z}_h^\theta(\widetilde{s}_h, o_h) \, \mathrm{d}\widetilde{s}_h, \tag{3.6}$$

for any $o_h, \widetilde{o}_h, \widetilde{o}_{h+1} \in \mathcal{O}$ and $a \in \mathcal{A}$, where the function $\mathcal{Z}_h^\theta$ is defined in Assumption 2.2. Figure 2 illustrates the (random) variables in (3.4) and (3.6). In detail, $\widetilde{\boldsymbol{s}}_h$ is an independent replicate of $\boldsymbol{s}_h$, that is, they are independent and identically distributed conditioning on $\boldsymbol{s}_{h-1}$ and $\boldsymbol{a}_{h-1}$. Note that $\widetilde{\boldsymbol{s}}_h$ is constructed for ease of presentation, and does not exist in practice. Then, the action $\widetilde{\boldsymbol{a}}_h$, state $\widetilde{\boldsymbol{s}}_{h+1}$, and observations $\widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1}$ are similarly defined. In other words, their distribution conditioning on $\widetilde{\boldsymbol{s}}_h$ and $\overline{\boldsymbol{\tau}}_{h-1}$ mirrors the distribution of the action $\boldsymbol{a}_h$, state $\boldsymbol{s}_{h+1}$, and observations $\boldsymbol{o}_h, \boldsymbol{o}_{h+1}$ conditioning on $\boldsymbol{s}_h$ and $\overline{\boldsymbol{\tau}}_{h-1}$. When the state transition kernel and observation emission kernel have a specific parametrization, the function $\mathcal{Z}_h^\theta$ has a corresponding parametrization by Assumption 2.2, which allows us to parametrize the function $\mathcal{B}_{h,a}^\theta$ in (3.6). See Section B for an example where the state transition kernel and observation emission kernel admit a linear structure. Compared with the full-memory Bellman operator $\mathbb{P}_h^{\theta,\pi}$, the finite-memory Bellman operator $\mathbb{B}_h^{\theta,\pi}$ does not involve the distribution of $\boldsymbol{o}_{h+1}$ conditioning on $\overline{\boldsymbol{\tau}}_h$ and $\boldsymbol{a}_h$. Instead, it involves the distribution of $\widetilde{\boldsymbol{o}}_h$ and $\widetilde{\boldsymbol{o}}_{h+1}$ conditioning on $\widetilde{\boldsymbol{s}}_h$, where the distribution of $\widetilde{\boldsymbol{s}}_h$ is implied by the distribution of the single observation $\boldsymbol{o}_h$ via the function $\mathcal{Z}_h^\theta$. See the following paragraph for more discussion. Moreover, estimating $\mathbb{B}_h^{\theta,\pi}$ for each $h \in [H]$ only involves the distribution of $\boldsymbol{o}_{h-1}$, $\boldsymbol{o}_h$, and $\boldsymbol{o}_{h+1}$, which is low-dimensional. See Section 3.2 for more discussion.

How Finite Memory Works: For notational simplicity, we denote by σ_{h-1} the event

$$\overline{\boldsymbol{\tau}}_{h-1} = \overline{\tau}_{h-1}, \quad \boldsymbol{a}_{h-1} = a_{h-1} \tag{3.7}$$

for any $h \in [H+1]$. The following lemma implies that the finite-memory Bellman operator $\mathbb{B}_h^{\theta,\pi}$ is identical to the full-memory Bellman operator $\mathbb{P}_h^{\theta,\pi}$ in expectation conditioning on σ_{h-1}, which allows us to use $\mathbb{B}_h^{\theta,\pi}$ as a surrogate of $\mathbb{P}_{h,a}^\theta$.

Lemma 3.1 (Operators Equivalence in Expectation)**.** *For any $(h, \theta, \pi, \overline{\tau}_{h-1}, a_{h-1}) \in [H] \times \Theta \times \Pi \times \overline{\Gamma}_{h-1} \times \mathcal{A}$ and $f \in L^\infty(\overline{\Gamma}_{h+1})$, we have*

$$\mathbb{E}_\theta[(\mathbb{B}_h^{\theta,\pi} f)(\overline{\boldsymbol{\tau}}_h) - (\mathbb{P}_h^{\theta,\pi} f)(\overline{\boldsymbol{\tau}}_h) \,|\, \sigma_{h-1}] = 0.$$

Proof. See Section D.1 for a detailed proof. □

To see the intuition behind Lemma 3.1, note that by the definition of $\mathcal{Z}_h^\theta$ in Assumption 2.2, we have

$$\mathbb{E}_\theta[\mathcal{Z}_h^\theta(\widetilde{s}_h, \boldsymbol{o}_h) \,|\, \sigma_{h-1}] = p_\theta(\widetilde{\boldsymbol{s}}_h = \widetilde{s}_h \,|\, \sigma_{h-1}),$$

for any $(\widetilde{s}_h, \overline{\tau}_{h-1}, a_{h-1}) \in \mathcal{S} \times \overline{\Gamma}_{h-1} \times \mathcal{A}$. See Section D.1 for a derivation. In other words, $\mathcal{Z}_h^\theta$ serves as the bridge function in causal inference (Shi et al., 2020), which recovers the conditional distribution of $\widetilde{\boldsymbol{s}}_h$ from the conditional distribution of $\boldsymbol{o}_h$. Then, by taking the same conditional expectation on both sides of (3.6), we have

$$\begin{aligned} &\mathbb{E}_\theta[\mathcal{B}_{h,a}^\theta(\boldsymbol{o}_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \,|\, \sigma_{h-1}] \\ &= p_\theta(\widetilde{\boldsymbol{o}}_h = \widetilde{o}_h, \widetilde{\boldsymbol{o}}_{h+1} = \widetilde{o}_{h+1} \,|\, \sigma_{h-1}, \widetilde{\boldsymbol{a}}_h = a) \\ &= p_\theta(\boldsymbol{o}_h = \widetilde{o}_h, \boldsymbol{o}_{h+1} = \widetilde{o}_{h+1} \,|\, \sigma_{h-1}, \boldsymbol{a}_h = a), \end{aligned} \tag{3.8}$$

which is connected to the integral kernel $p_\theta(\boldsymbol{o}_{h+1} = o_{h+1} \,|\, \overline{\boldsymbol{\tau}}_h = \overline{\tau}_h, \boldsymbol{a}_h = a)$ on the right-hand side of (3.1) via the same conditional expectation

$$\begin{aligned} &\mathbb{E}_\theta[p_\theta(\boldsymbol{o}_{h+1} = o_{h+1} \,|\, \overline{\boldsymbol{\tau}}_h, \boldsymbol{a}_h = a) \,|\, \sigma_{h-1}] \\ &= \mathbb{E}_\theta[p_\theta(\boldsymbol{o}_{h+1} = o_{h+1} \,|\, \sigma_{h-1}, \boldsymbol{o}_h, \boldsymbol{a}_h = a) \,|\, \sigma_{h-1}] \\ &= p_\theta(\boldsymbol{o}_{h+1} = o_{h+1} \,|\, \sigma_{h-1}, \boldsymbol{a}_h = a) \\ &= \int_{\mathcal{O}} p_\theta(\boldsymbol{o}_h = o_h, \boldsymbol{o}_{h+1} = o_{h+1} \,|\, \sigma_{h-1}, \boldsymbol{a}_h = a) \, \mathrm{d}o_h. \end{aligned} \tag{3.9}$$

Here the second equality in (3.8) follows from the fact that $(\widetilde{\boldsymbol{a}}_h, \widetilde{\boldsymbol{o}}_h, \widetilde{a}_h)$ is an independent replicate of $(\boldsymbol{a}_h, \boldsymbol{o}_h, \boldsymbol{o}_{h+1})$,

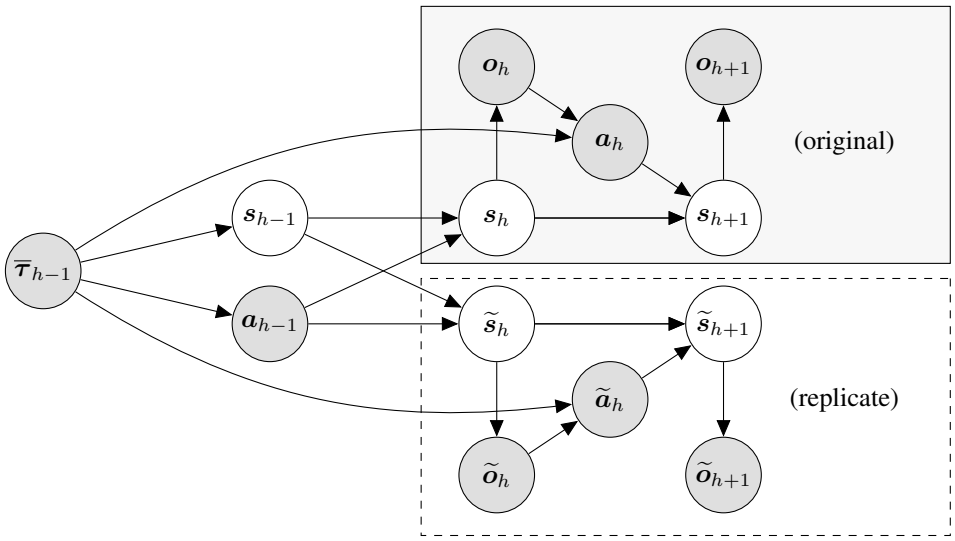

Figure 2: Illustration of the variables in the definition of $\mathbb{B}_h^{\theta,\pi}$ in (3.4) and (3.6). In detail, $\widetilde{s}_h$ is an independent replicate of s_h, that is, they are independent and identically distributed conditioning on s_{h-1} and a_{h-1}. Note that $\widetilde{s}_h$ is constructed for ease of presentation, and does not exist in practice. Then, the action $\widetilde{a}_h$, state $\widetilde{s}_{h+1}$, and observations $\widetilde{o}_h, \widetilde{o}_{h+1}$ are similarly defined. In other words, their distribution conditioning on $\widetilde{s}_h$ and $\overline{\tau}_{h-1}$ mirrors the distribution of the action a_h, state s_{h+1}, and observations o_h, o_{h+1} conditioning on s_h and $\overline{\tau}_{h-1}$. For notational simplicity, we define the tail-mirrored full history $\overline{\tau}_h^\dagger = (\overline{\tau}_{h-1}, a_{h-1}, \widetilde{o}_h)$ and tail-mirrored observation history $\tau_h^\dagger = (\tau_{h-1}, \widetilde{o}_h)$.

that is, they follow the same distribution conditioning on σ_{h-1}.

Backward Bellman Recursion: For any $(h, \theta, \pi) \in [H+1] \times \mathcal{A} \times \Pi$, we define the value function $V_h^{\theta,\pi} \in L^\infty(\overline{\Gamma}_h)$ by

$$V_h^{\theta,\pi}(\overline{\tau}_h) = (\mathbb{B}_h^{\theta,\pi} \cdots \mathbb{B}_H^{\theta,\pi} R)(\overline{\tau}_h), \tag{3.10}$$

for any $\overline{\tau}_h \in \overline{\Gamma}_h$, which gives the backward Bellman recursion

$$V_h^{\theta,\pi}(\overline{\tau}_h) = (\mathbb{B}_h^{\theta,\pi} V_{h+1}^{\theta,\pi})(\overline{\tau}_h), \tag{3.11}$$

for any $\overline{\tau}_h \in \overline{\Gamma}_h$. The following corollary is implied by Lemma 3.1, which relates the value function $V_h^{\theta,\pi}$ to the expected total reward. Note that $V_h^{\theta,\pi}$ does not correspond to the "reward-to-go" in the usual value function definition in MDPs since it involves all rewards across the H steps.

Corollary 3.2. *For any $(h, \theta, \pi) \in [H+1] \times \Theta \times \Pi$, we have*

$$\mathbb{E}_{\theta,\pi}\Big[V_h^{\theta,\pi}(\overline{\tau}_h) - \sum_{i=1}^{H} r_i \,\Big|\, \sigma_{h-1}\Big] = 0, \tag{3.12}$$

for any $(\overline{\tau}_{h-1}, a_{h-1}) \in \overline{\Gamma}_{h-1} \times \mathcal{A}$. For $h = 1$, we have $J(\theta, \pi) = \mathbb{E}[V_1^{\theta,\pi}(o_1)]$ since $\overline{\tau}_1 = o_1$ and $\sigma_0 = \varnothing$, which follows from the definition of $\overline{\tau}_h$ in (2.3).

Proof. See Section D.2 for a detailed proof. □

Corollary 3.2 allows us to evaluate a policy π by estimating $\{\mathbb{B}_h^{\theta^*,\pi}\}_{h=1}^H$ instead of $\{\mathbb{P}_h^{\theta^*,\pi}\}_{h=1}^H$. Meanwhile, $\{V_h^{\theta,\pi}\}_{h=1}^H$ play a critical role in analyzing the sample complexity.

3.2. Operator Estimation via Minimax Optimization

Although the finite-memory Bellman operator $\mathbb{B}_h^{\theta,\pi}$ defined in (3.4) does not involve the observation distribution conditioning on the history, that is, the distribution of o_{h+1} conditioning on $\overline{\tau}_h$ and o_h, it remains unclear how to estimate $\mathbb{B}_h^{\theta^*,\pi}$ in a sample-efficient manner. Note that, by the definition of $\mathbb{B}_h^{\theta^*,\pi}$, it suffices to estimate functions $\{\mathcal{B}_{h,a}^{\theta^*}\}_{a\in\mathcal{A}}$. To this end, we define the operator $\mathbb{F}_{h,a}^\theta : L^\infty(\mathcal{O}^3) \to L^\infty(\mathcal{O}^3)$ for any $(h, a, \theta) \in \{2, \ldots, H\} \times \mathcal{A} \times \Theta$ by

$$\begin{aligned}(\mathbb{F}_{h,a}^\theta f)(o_{h-1}, o_h, o_{h+1}) \\ = \int_{\mathcal{O}^2} f(o_{h-1}, \widetilde{o}_h, \widetilde{o}_{h+1}) \\ \cdot \mathcal{B}_{h,a}^\theta(o_h, \widetilde{o}_h, \widetilde{o}_{h+1})\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1},\end{aligned} \tag{3.13}$$

for any $f \in L^1(\mathcal{O}^3)$ and $o_{h-1}, o_h \in \mathcal{O}$. Note that $\mathbb{F}_{h,a}^\theta$ is a truncated version of $\mathbb{B}_h^{\theta,\pi}$, which drops a few variables that are redundant for operator estimation. The following lemma motivates the estimator of $\mathcal{B}_{h,a}^{\theta^*}$, which uses the definition of $\mathbb{F}_{h,a}^\theta$.

Lemma 3.3. *For any $(h, a, a', \pi) \in \{2, \ldots, H\} \times \mathcal{A}^2 \times \overline{\Pi}$,*

we have

$$\mathbb{E}_{X\sim\rho^{\pi}_{h,a,a'}}[(\mathbb{F}^{\theta^*}_{h,a'}f - f)(X)] = 0,$$

for any $f \in L^\infty(\mathcal{O}^3)$. *Here, the distribution* $\rho^{\pi}_{h,a,a'} \in \Delta(\mathcal{O}^3)$ *is defined by*

$$\begin{aligned}&\rho^{\pi}_{h,a,a'}(o_{h-1}, o_h, o_{h+1})\\ &\quad= p_{\theta^*,\pi}\big(\boldsymbol{o}_{h-1} = o_{h-1}, \boldsymbol{o}_h = o_h, \boldsymbol{o}_{h+1} = o_{h+1}\,|\\ &\qquad\qquad \boldsymbol{a}_{h-1} = a, \boldsymbol{a}_h = a'\big),\end{aligned}$$

for any $o_{h-1}, o_h, o_{h+1} \in \mathcal{O}$. *Also, we have*

$$\|\mathbb{F}^{\theta^*}_{h,a'}\|_{\infty\to\infty} \le \gamma.$$

Proof. See Section D.3 for a detailed proof. □

Minimax Optimization: For any $(h, a', \pi) \in \{2, \ldots, H\} \times \mathcal{A} \times \overline{\Pi}$, Lemma 3.3 allows us to estimate $\mathcal{B}^{\theta^*}_{h,a'}$ based on a dataset $\{\mathcal{D}_{h,a,a'}\}_{a\in\mathcal{A}}$, where the data points in $\mathcal{D}_{h,a,a'}$ are collected from the distribution $\rho^{\pi}_{h,a,a'}$. In other words, each episode involves three steps: (a) we execute the exploration policy π, which takes the actions $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{h-2}$, (b) we take the actions $\boldsymbol{a}_{h-1} = a$ and $\boldsymbol{a}_h = a'$ regardless of the observations, and (c) we add the observation tuple $(\boldsymbol{o}_{h-1}, \boldsymbol{o}_h, \boldsymbol{o}_{h+1})$ to $\mathcal{D}_{h,a,a'}$. Based on $\{\mathcal{D}_{h,a,a'}\}_{(h,a,a')\in\{2,\ldots,H\}\times\mathcal{A}^2}$, we estimate $\{\mathcal{B}^{\theta^*}_{h,a'}\}_{(h,a')\in\{2,\ldots,H\}\times\mathcal{A}}$ by solving the following minimax optimization problem,

$$\begin{aligned}\min_{\theta\in\Theta}\ \max_{f\in L^\infty(\mathcal{O}^3):\|f\|_\infty\le 1}\ \max_{(h,a,a')\in\{2,\ldots,H\}\times\mathcal{A}^2}&\\ \mathbb{E}_{X\sim\widehat{\mathcal{D}}_{h,a,a'}}[(\mathbb{S}\mathbb{F}^{\theta}_{h,a'}f - \mathbb{S}f)(X)].&\end{aligned} \tag{3.14}$$

Here, $\widehat{\mathcal{D}}_{h,a,a'}$ is the empirical distribution induced by the dataset $\mathcal{D}_{h,a,a'}$. Also, the projection operator $\mathbb{S} : L^1(\mathcal{O}^3) \to L^1(\mathcal{O}^3)$ satisfies that

$$\mathbb{E}_{X\sim p}[(\mathbb{S}f)(X)] = \int_{\mathcal{O}^3} f(x)\cdot p^\dagger(x)\,\mathrm{d}x, \tag{3.15}$$

for any $f \in L^\infty(\mathcal{O}^3)$ and $p \in \Delta(\mathcal{O}^3)$. Here, $p^\dagger \in L^1(\mathcal{O}^3)$ is the projection of p onto $\mathrm{linspan}(\{\phi_i\}_{i=1}^{d_\mathrm{o}})$. See the definition of $\mathbb{S}$ in the next paragraph. The minimax optimization problem in (3.14) is motivated by generative adversarial networks. To see the intuition behind (3.14), note that f serves as the discriminator and $F^{\theta}_{h,a'}$ serves as the generator. In detail, note that the function $\mathbb{F}^{\theta}_{h,a}f$ in (3.13) is constant with respect to the variable o_{h+1}. Thus, Lemma 3.3 implies that the true generator $F^{\theta^*}_{h,a'}$ recovers the distribution of $(\boldsymbol{o}_{h-1}, \boldsymbol{o}_h, \boldsymbol{o}_{h+1}) \sim d^{\pi}_{h,a,a'}$ (corresponding to the true parameter θ^*) from the marginal distribution of $(\boldsymbol{o}_{h-1}, \boldsymbol{o}_h)$. In this case, the true distribution and the (fake) distribution recovered by the generator can not be distinguished by any discriminator in $L^\infty(\mathcal{O}^3)$. When we train the generator and discriminator on a dataset, the discriminator class $L^\infty(\mathcal{O}^3)$ has a too large capacity. Therefore, we employ the projection operator $\mathbb{S}$ to enforce the finite-dimensional linear structure of $d^{\pi}_{h,a,a'}$, which reduces the capacity of the discriminator class. Such a projection operator guarantees the generalization power of the solution to (3.14).

Projection Operator via RKHS: In the following, we define the projection operator $\mathbb{S}$. To this end, we consider an RKHS $\mathcal{H}$ induced by a kernel function $\mathcal{K} : \mathcal{O}^3 \times \mathcal{O}^3 \to \mathbb{R}$. We define the corresponding RKHS embedding $\mathbb{K} : L^1(\mathcal{O}^3) \to \mathcal{H}$ by

$$(\mathbb{K}p)(x) = \int_{\mathcal{O}^3} \mathcal{K}(x, y)p(y)\,\mathrm{d}y, \tag{3.16}$$

for any $p \in L^1(\mathcal{O}^3)$ and $x \in \mathcal{O}^3$. Moreover, we define the matrix $G \in \mathbb{R}^{d_\mathrm{o}\times d_\mathrm{o}}$ by

$$\begin{aligned}[G]_{i,j} &= \langle \mathbb{K}\phi_i, \mathbb{K}\phi_j\rangle_{\mathcal{H}}\\ &= \mathbb{E}_{X\sim\phi_i, X'\sim\phi_j}[\mathcal{K}(X, X')],\end{aligned} \tag{3.17}$$

for any $i, j \in [d_\mathrm{o}]$. Recall that the distribution functions $\{\phi_i\}_{i=1}^{d_\mathrm{o}}$ are defined in Assumption 2.2. The following assumption specifies the regularity condition on $\mathcal{K}$ and $\{\phi_i\}_{i=1}^{d_\mathrm{o}}$.

Assumption 3.4. The kernel function $\mathcal{K}$ is bounded and continuous. In particular, we have $|\mathcal{K}(x, y)| \le 1$ for any $x, y \in \mathcal{O}^3$. Also, we have $\alpha = \lambda_{\min}(G) > 0$, where we denote by $\lambda_{\min}(\cdot)$ the minimum eigenvalue of a matrix and the matrix G is defined in (3.17).

Here, the continuity of $\mathcal{K}$ is defined with respect to the topology space $\mathcal{O}^3$. For example, $\mathcal{O}^3$ is (embedded as) a subset of some Euclidean space and the continuity of $\mathcal{K}$ is defined with respect to the corresponding Euclidean distance. The boundedness of $\mathcal{K}$ is satisfied by many kernel functions, for example, the radial basis function (RBF) kernel (Smola & Schölkopf, 1998). For the positive definiteness of the matrix G, note that, for any $v = (v_1, \ldots, v_{d_\mathrm{o}}) \in \mathbb{R}^{d_\mathrm{o}}$, we have

$$v^\top G v = \sum_{i,j=1}^{d_\mathrm{o}} v_i v_j \cdot \langle \mathbb{K}\phi_i, \mathbb{K}\phi_j\rangle_{\mathcal{H}} = \Big\|\sum_{i=1}^{d_\mathrm{o}} v_i \cdot \mathbb{K}\phi_i\Big\|_{\mathcal{H}}^2.$$

Therefore, to make G positive definite, it suffices to require $\mathbb{K}\phi_1, \ldots, \mathbb{K}\phi_{d_\mathrm{o}}$ to be linearly independent in $\mathcal{H}$. With the the kernel function $\mathcal{K}$ and matrix G defined above, we can verify that (3.15) holds for the operator $\mathbb{S}$ defined by

$$\begin{aligned}&(\mathbb{S}f)(x)\\ &\quad= \sum_{i,j\in[d_\mathrm{o}]} [G^{-1}]_{i,j}\cdot \mathbb{E}_{Y\sim\phi_i, Y'\sim\phi_j}\big[\mathcal{K}(x, Y)\cdot f(Y')\big],\end{aligned} \tag{3.18}$$

for any $f \in L^\infty(\mathcal{O}^3)$ and $x \in \mathcal{O}^3$. Here, the distance in the projection from p to $p^\dagger$ in (3.15) is defined by

$$d(p_1, p_2) = \|\mathbb{K}p_1 - \mathbb{K}p_2\|_{\mathcal{H}}, \tag{3.19}$$

for any $p_1, p_2 \in L^1(\mathcal{O}^3)$. See Section G.1 for a derivation.

3.3. Online Exploration via Optimistic Planning

We present the *Optimistic Exploration via Adversarial Integral Equation* (OP-TENET) algorithm, which incorporates operator estimation into optimistic planning to perform online exploration. In other words, we update the exploration policy in Section 3.2 in an iterative manner. We initialize OP-TENET with any policy $\pi_0 \in \Pi$ and a dataset

$$\{\mathcal{D}_{h,a,a'}\}_{(h,a,a')\in\{2,\ldots,H\}\times\mathcal{A}^2} = \varnothing, \tag{3.20}$$

which are updated subsequently in the K iterations. Each iteration consists of an exploration phase and a planning phase. In the following, we describe the k-th iteration for any $k \in [K]$.

Exploration Phase: Given the exploration policy π_{k-1}, we run an episode of the POMDP for each tuple $(h, a, a') \in \{2, \ldots, H\} \times \mathcal{A}^2$ following the data collecting scheme defined in Section 3.2 to add an observation tuple (o_{h-1}, o_h, o_{h+1}) into the dataset $\mathcal{D}_{h,a,a'}$. After the exploration phase of the k-th iteration, we have k observation tuples in the dataset $\mathcal{D}_{h,a,a'}$ for any (h, a, a'). Although the dataset is collected by the exploration policies $\pi_0, \ldots, \pi_{k-1}$ in the k iterations, we can regard it as a dataset collected by the mixing policy

$$\overline{\pi}_k = \mathrm{mixing}\{\pi_0, \ldots, \pi_{k-1}\}. \tag{3.21}$$

where each policy is sampled uniformly at random as defined in Section 2.1.

Planning Phase: We apply the operator estimation method defined in Section 3.2 to the updated dataset in (3.20) and construct a confidence set of the model parameter θ

$$\Theta_k = \Big\{\theta \in \Theta : L(\theta) \le \beta \cdot k^{-1/2}\Big\}, \tag{3.22}$$

for a constant $\beta > 0$, where $L(\theta)$ is defined as

$$L(\theta) = \max_{f\in L^\infty(\mathcal{O}^3):\|f\|_\infty\le 1} \max_{(h,a,a')\in\{2,\ldots,H\}\times\mathcal{A}^2} \mathbb{E}_{X\sim\widehat{\mathcal{D}}_{h,a,a'}}[(\mathbb{SF}^\theta_{h,a'}f - \mathbb{S}f)(X)]. \tag{3.23}$$

Given the confidence set defined in (3.22), we update the exploration policy by

$$\pi_k = \operatorname*{argmax}_{\pi\in\Pi} \max_{\theta\in\Theta_k} J(\theta, \pi), \tag{3.24}$$

which is the optimal policy with respect to the optimistic value estimator over parameters θ in the confidence set Θ_k. Recall that $J(\theta, \pi)$ is defined in (2.2). Note that we can perform the computation of (3.24) via a planning oracle for POMDPs (Golowich et al., 2022). In detail, we can reformulate (3.24) as

$$\theta_k = \operatorname*{argmax}_{\theta\in\Theta_k} J\big(\theta, \widehat{\pi}(\theta)\big), \quad \pi_k = \widehat{\pi}(\theta_k),$$

where the planning oracle $\widehat{\pi}(\cdot)$ outputs the optimal policy with respect to any parameter. The constraint $\theta \in \Theta_k$ can be further transformed as a part of the objective via the Lagrangian relaxation. Then, we can apply the stochastic gradient method to obtain θ_k in a computation-efficient manner. See Section C for more details. At the $(k+1)$-th iteration, we execute the exploration policy π_k to collect data, which serves as the next exploration phase. We present OP-TENET in Algorithm 1.

Algorithm 1 OP-TENET

Input: number of iterations K, confidence level β
Initialization: set π_0 as a deterministic policy
Initialization: update the dataset $\mathcal{D}_{h,a,a'} \leftarrow \varnothing$ for $(h, a, a') \in \{2, \ldots, H\} \times \mathcal{A}^2$
For $k = 1$ to K **do**
  For $(h, a, a') \in \{2, \ldots, H\} \times \mathcal{A}^2$ **do**
    Start a new episode
    Execute π_{k-1} to take the first $(h-2)$ actions
    Receive the observation o_{h-1}
    Take the action a, receive the observation o_h
    Take the action a', receive the observation o_{h+1}
    End the current episode
    Update the dataset

$$\mathcal{D}_{h,a,a'} \leftarrow \mathcal{D}_{h,a,a'} \cup \{(o_{h-1}, o_h, o_{h+1})\}$$

  Construct the confidence set Θ_k by (3.22)
  Update the policy

$$\pi_k \leftarrow \operatorname*{argmax}_{\pi\in\Pi} \max_{\theta\in\Theta_k} J(\theta, \pi)$$

Output: policy set $\{\pi_1, \ldots, \pi_K\}$

4. Theory

In this section, we analyze OP-TENET in Algorithm 1. In Section 4.1, we prove that the policies generated by Algorithm 1 converge to the optimal policy with a polynomial sample complexity. Due to space limit, we defer the proof sketch to Section A in the appendix, where we sketch the proof by three key lemmas.

4.1. Sample Efficiency

The following theorem characterizes the sample complexity of OP-TENET in Algorithm 1.

Theorem 4.1. *Under Assumptions 2.1, 2.2, and 3.4, for any $\delta > 0$, if we choose a confidence level β in Algorithm 1 to such that*

$$\beta \ge d_{\mathrm{o}}^{3/2}(\gamma + 1)/\alpha \cdot \sqrt{8\log(2KHA^2/\delta)}, \tag{4.1}$$

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

$$\frac{1}{K}\sum_{k=1}^{K}\big(J(\theta^*,\pi^*)-J(\theta^*,\pi_k)\big) \le \frac{4d_{\mathrm{s}}\gamma^2\beta H^2A^2\cdot\log K}{K^{1/2}}+\frac{4d_{\mathrm{s}}\gamma H^2}{K}. \tag{4.2}$$

Recall that d_{s} *and* d_{o} *are defined in Assumption 2.1,* γ *is defined in Assumption 2.2, and* α *is defined in Assumption 3.4.*

Note that the first term on the right-hand side of (4.2) is the leading term for a sufficiently large number of iterations K. Recall that the state distribution dimension d_{s} and observation distribution dimension d_{o} are defined in Assumption 2.1. Also, quantities γ and α are defined in Assumptions 2.2 and 3.4, respectively. By Theorem 4.1, if we run OP-TENET for K iterations and sample a policy from $\{\pi_1,\pi_K\}$ uniformly at random, the expected suboptimality of such a policy converges to zero with high probability at the rate of $K^{-1/2}$ up to logarithmic factors. Meanwhile, such a rate depends on H, $\mathcal{A}$, d_{s}, d_{o}, γ, and $1/\alpha$ polynomially. In other words, to obtain an ε-optimal policy for any suboptimality $\varepsilon>0$, it suffices to run

$$K=\mathrm{poly}(H,A,d_{\mathrm{s}},d_{\mathrm{o}},\gamma,1/\alpha)\cdot\widetilde{O}(1/\varepsilon^2) \tag{4.3}$$

iterations in OP-TENET to collect the data set. Note that the total number of episodes in K iterations is $(H-1)A^2K$. To our best knowledge, Theorem 4.1 is the first polynomial sample complexity upper bound for reinforcement learning in POMDPs that is independent of the number of states and observations. Moreover, the order of ε is optimal even in the MDP setting (Ayoub et al., 2020), which is a special case of POMDPs. In contrast to the sample complexity results in MDPs, a key difference of Theorem 4.1 is that it involves the (upper bound of the) operator norm γ of the bridge operator $\mathbb{Z}_h^\theta$, which is the left inverse of the observation operator $\mathbb{O}_h^\theta$. Recall that such a left inverse is defined with respect to the finite-dimensional subspace $\mathrm{linspan}(\{\phi_i\}_{i=1}^{d_{\mathrm{o}}})$ of $L^1(\mathcal{O}^3)$ in Assumption 2.2. The (upper bound of the) operator norm γ is a measure of ill conditioning, which quantifies the fundamental difficulty, in terms of the information-theoretic limit, of reinforcement learning in the POMDP. In the degenerate case where $\mathbb{O}_h^\theta$ is not invertible, Theorem 4.1 provides a trivial upper bound since we have $\gamma=\infty$. On the other hand, such a case contains examples that are fundamentally impossible to solve in a sample-efficient manner, which is implied by information theory (Jin et al., 2020a).

References

Agarwal, A., Henaff, M., Kakade, S., and Sun, W. PC-PG: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020.

Arulkumaran, K., Cully, A., and Togelius, J. Alphastar: An evolutionary computation perspective. In *Proceedings of the genetic and evolutionary computation conference companion*, pp. 314–315, 2019.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2008.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. Reinforcement learning of POMDPs using spectral methods. In *Conference on Learning Theory*, 2016.

Bennett, A. and Kallus, N. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *arXiv preprint arXiv:2110.15332*, 2021.

Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Brown, N. and Sandholm, T. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, 2020.

Cassandra, A. R., Kaelbling, L. P., and Kurien, J. A. Acting under uncertainty: Discrete bayesian models for mobile-robot navigation. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1996.

Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in RL. *arXiv preprint arXiv:2103.10897*, 2021.

Golowich, N., Moitra, A., and Rohatgi, D. Planning in observable POMDPs in quasipolynomial time. *arXiv preprint arXiv:2201.04735*, 2022.

Guo, Z. D., Doroudi, S., and Brunskill, E. A PAC RL algorithm for episodic POMDPs. In *Artificial Intelligence and Statistics*, 2016.

Hauskrecht, M. and Fraser, H. Planning treatment of ischemic heart disease with partially observable markov decision processes. *Artificial Intelligence in Medicine*, 18 (3):221–244, 2000.

Hsu, D., Kakade, S. M., and Zhang, T. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Jaeger, H. Observable operator models for discrete stochastic time series. *Neural computation*, 12(6):1371–1398, 2000.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, 2018.

Jin, C., Kakade, S. M., Krishnamurthy, A., and Liu, Q. Sample-efficient reinforcement learning of undercomplete POMDPs. *arXiv preprint arXiv:2006.12484*, 2020a.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2020b.

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. Information theoretic regret bounds for online nonlinear control. *arXiv preprint arXiv:2006.12466*, 2020.

Kallus, N., Mao, X., and Uehara, M. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

Kozuno, T., Ménard, P., Munos, R., and Valko, M. Model-free learning for two-player zero-sum partially observable Markov games with perfect recall. *arXiv preprint arXiv:2106.06279*, 2021.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. RL for latent MDPs: Regret guarantees and a lower bound. *arXiv preprint arXiv:2102.04939*, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Nair, Y. and Jiang, N. A spectral approach to off-policy evaluation for POMDPs. *arXiv preprint arXiv:2109.10502*, 2021.

Osband, I., Van Roy, B., and Wen, Z. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, 2016.

Pearl, J. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Pinelis, I. An approach to inequalities for the distributions of infinite-dimensional martingales. In *Probability in Banach Spaces*, pp. 128–134. Springer, 1992.

Pinelis, I. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, pp. 1679–1706, 1994.

Rafferty, A. N., Brunskill, E., Griffiths, T. L., and Shafto, P. Faster teaching by POMDP planning. In *International Conference on Artificial Intelligence in Education*, pp. 280–287. Springer, 2011.

Shi, X., Miao, W., and Tchetgen, E. T. A selective review of negative control methods in epidemiology. *Current epidemiology reports*, 7(4):190–202, 2020.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of Go without human knowledge. *nature*, 550(7676):354–359, 2017.

Smola, A. J. and Schölkopf, B. *Learning with kernels*, volume 4. Citeseer, 1998.

Vlassis, N., Littman, M. L., and Barber, D. On the computational complexity of stochastic controller optimization in POMDPs. *ACM Transactions on Computation Theory (TOCT)*, 4(4):1–8, 2012.

Xiong, Y., Chen, N., Gao, X., and Zhou, X. Sublinear regret for learning POMDPs. *arXiv preprint arXiv:2107.03635*, 2021.

Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, 2020.

Zhang, J. and Bareinboim, E. Markov decision processes with unobserved confounders: A causal approach. Technical report, Technical report, Technical Report R-23, Purdue AI Lab, 2016.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture Markov decision processes. In *Conference on Learning Theory*, 2021.

A. Proof Sketch

In this section, we sketch the proof of Theorem 4.1. In detail, we prove the theorem by three key lemmas. The following lemma provides a decomposition of the difference of expected total rewards in two POMDPs when the parameters are different and the policies are identical. Recall that the function J is defined in (2.2).

Lemma A.1 (Value Decomposition)**.** *Under Assumptions 2.1 and 2.2, we have*

$$\begin{aligned} &J(\theta, \pi) - J(\theta', \pi) \\ &\quad = \sum_{h=1}^{H} \mathbb{E}_{\theta', \pi}[(\mathbb{B}_h^{\theta, \pi} V_{h+1}^{\theta, \pi})(\overline{\boldsymbol{\tau}}_h) - (\mathbb{B}_h^{\theta', \pi} V_{h+1}^{\theta, \pi})(\overline{\boldsymbol{\tau}}_h)], \end{aligned}$$

for any $\theta, \theta' \in \Theta$ and $\pi \in \Pi$. Here, the function $V_{h+1}^{\theta, \pi}$ is defined in (3.10).

Proof. See Section F.1 for a detailed proof. □

For any $k \in [K]$, we denote by $\theta_k \in \Theta$ the model parameter that is selected in the planning phase of the k-th iteration OP-TENET (Algorithm 1), which is defined in (3.24), that is,

$$(\theta_k, \pi_k) = \operatorname*{argmax}_{(\theta, \pi) \in \Theta_k \times \Pi} J(\theta, \pi). \tag{A.1}$$

We define the state-dependent error $e_h^k : \mathcal{S} \to \mathbb{R}$ by

$$e_h^k(s_{h-1}) = \big|\mathbb{E}_{\theta^*, \pi_k}[(\mathbb{B}_h^{\theta_k, \pi_k} V_{h+1}^{\theta_k, \pi_k})(\overline{\boldsymbol{\tau}}_h) - (\mathbb{B}_h^{\theta^*, \pi_k} V_{h+1}^{\theta_k, \pi_k})(\overline{\boldsymbol{\tau}}_h) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}]\big|, \tag{A.2}$$

for any $(k, h, s_{h-1}) \in [K] \times [H] \times \mathcal{S}$. Conditioning on the event $\theta^* \in \Theta_k$, which is shown to occur with high probability in the following lemma, we have

$$J(\theta^*, \pi^*) - J(\theta^*, \pi_k) \le J(\theta_k, \pi_k) - J(\theta^*, \pi_k) \le \sum_{h=1}^{H} \mathbb{E}_{\theta^*, \pi_k}[e_h^k(\boldsymbol{s}_{h-1})], \tag{A.3}$$

which follows from Lemma A.1. Also, the following lemma characterizes the right-hand side of (A.3) when we replace the policy π_k by the mixing policy $\overline{\pi}_k$ defined in (3.21).

Lemma A.2 (Statistical Guarantee)**.** *Under Assumptions 2.1, 2.2, and 3.4, for any $\delta > 0$, by choosing the confidence level β in OP-TENET (Algorithm 1) such that it satisfies (4.1), with probability at least $1 - \delta$, we have*

- $\theta^* \in \Theta_k$,
- $\mathbb{E}_{\theta^*, \overline{\pi}_k}[e_h^k(\boldsymbol{s}_{h-1})] \le 2HA^2\gamma^2\beta \cdot k^{-1/2}$,

for any $(k, h) \in [K] \times [H]$.

Proof. See Section F.2 for a detailed proof. □

To characterize the right-hand side of (A.3), it remains to connect $\mathbb{E}_{\theta^*, \pi_k}[e_h^k(\boldsymbol{s}_{h-1})]$ with $\mathbb{E}_{\theta^*, \overline{\pi}_k}[e_h^k(\boldsymbol{s}_{h-1})]$, which involve different state distributions. The connection is established in the following lemma.

Lemma A.3 (Telescope of Error)**.** *Under Assumptions 2.1 and 2.2, for any $h \in [H]$, we have*

$$\sum_{k=1}^{K} \mathbb{E}_{\theta, \pi_k}[e_h^k(\boldsymbol{s}_{h-1})] \le 4\gamma d_{\mathrm{s}} H + 2d_{\mathrm{s}} \log K \cdot \max_{k \in [K]} \big(k \cdot \mathbb{E}_{\theta, \overline{\pi}_k}[e_h^k(\boldsymbol{s}_{h-1})]\big).$$

Proof. See Section F.3 for a detailed proof. □

Combining (A.3) with Lemmas A.2 and A.3, we obtain Theorem 4.1. See Section E for a detailed proof.

B. Examples: Linear Kernel POMDPs

In this section, we show examples of the candidate class of the state transition kernels and observation emission kernels, which satisfy our assumptions in the main paper. In particular, we consider the following definition of linear kernel POMDPs.

Definition B.1 (Linear Kernel POMDPs)**.** We say that

$$\mathcal{L} = \{(\mathcal{S}, \mathcal{A}, \mathcal{O}, H, \mathcal{T}^\theta, \mathcal{E}^\theta, \mu, r) : \theta \in \Theta\}$$

is a linear kernel POMDP set, if each state transition kernel $\mathcal{T}^\theta$ and observation emission kernel $\mathcal{E}^\theta$ in the set take the form,

$$\mathcal{T}_h^\theta(s' \,|\, s, a) = u(s')^\top M_{h,a}^\theta v(s), \quad \mathcal{E}_h^\theta(o \,|\, s) = q(o)^\top g_h^\theta(s), \tag{B.1}$$

for any $(h, \theta, s, s', a, o) \in [H] \times \Theta \times \mathcal{S}^2 \times \mathcal{A} \times \mathcal{O}$. Here, u, v, q, and g_h^θ are non-negative vector-valued functions with dimensions d_u, d_v, d_q, and d_q, respectively. The matrix $M_{h,a}^\theta \in \mathbb{R}^{d_u \times d_v}$ has non-negative entries. Moreover, we have $\mu \in \mathrm{conh}(\{[u(\cdot)]_i\}_{i=1}^{d_u})$ and

$$[u(\cdot)]_i \in \Delta(\mathcal{S}), \quad [q(\cdot)]_\ell \in \Delta(\mathcal{O}), \quad \text{for any } (i, j) \in [d_u] \times [d_q].$$

The following lemma shows that tabular POMDPs are linear kernel POMDPs.

Lemma B.2. *For any finite state space $\mathcal{S}$, finite observation space $\mathcal{O}$, action space $\mathcal{A}$, episode length H, initial distribution μ, and reward function r, we can define a linear kernel POMDP set $\{(\mathcal{S}, \mathcal{A}, \mathcal{O}, H, \mathcal{T}^\theta, \mathcal{E}^\theta, \mu, r) : \theta \in \Theta\}$ following Definition B.1, which consists of all possible POMDPs with the aforementioned elements.*

Proof. We define Θ as the set of all possible pair $(\widetilde{\mathcal{T}}, \widetilde{\mathcal{E}})$ such that $\widetilde{\mathcal{T}}$ is a state transition kernel and $\widetilde{\mathcal{E}}$ is an observation emission kernel with respect to the state space $\mathcal{S}$, action space $\mathcal{A}$, observation space $\mathcal{O}$, and episode length H. We let $d_u = d_v = |\mathcal{S}|$ and $d_q = |\mathcal{O}|$. For any $\theta = (\widetilde{\mathcal{T}}, \widetilde{\mathcal{E}}) \in \Theta$, we define

$$[M_{h,a}^\theta]_{s',s} = \widetilde{\mathcal{T}}_h(s' \,|\, s, a), \quad [q_h^\theta(\cdot)]_o = \widetilde{\mathcal{E}}_h(o \,|\, \cdot),$$

for any $(h, s, s', a, o) \in [H] \times \mathcal{S}^2 \times \mathcal{A} \times \mathcal{O}$. Also, we define

$$[u(\cdot)]_s = \mathbb{1}\{s = \cdot\}, \quad [v(\cdot)]_s = \mathbb{1}\{s = \cdot\}, \quad [q(\cdot)]_o = \mathbb{1}\{o = \cdot\},$$

for any $(s, o) \in \mathcal{S} \times \mathcal{O}$. Then, by noting that we have $\mathcal{T}^\theta = \widetilde{\mathcal{T}}$ and $\mathcal{E}^\theta = \widetilde{\mathcal{E}}$ following the definitions of $\mathcal{T}^\theta$ and $\mathcal{E}^\theta$ in Definition B.1, we conclude the proof of Lemma B.2. □

B.1. Verification of Assumption 2.1

Recall that in reinforcement learning for a POMDP, the state transition kernel and observation emission kernel are unknown elements of the POMDP. We say that the candidate class of the POMDP is a linear kernel POMDP set $\mathcal{L} = \{(\mathcal{S}, \mathcal{A}, \mathcal{O}, H, \mathcal{T}^\theta, \mathcal{E}^\theta, \mu, r) : \theta \in \Theta\}$ when the candidate class of the state transition kernel and observation emission kernel is $\{(\mathcal{T}^\theta, \mathcal{E}^\theta) : \theta \in \Theta\}$ and other elements of the POMDP are determined as $(\mathcal{S}, \mathcal{A}, \mathcal{O}, H, \mu, r)$. The following lemma shows that any linear kernel POMDP set satisfies the linear function approximation assumption (Assumption 2.1).

Lemma B.3. *When the candidate class of the POMDP is a linear kernel POMDP set $\mathcal{L}$ as defined in Definition B.1, we have that Assumption 2.1 holds with*

$$d_{\mathrm{s}} \le d_u(d_v + 1) \quad \textit{and} \quad d_{\mathrm{o}} \le d_q^3.$$

Recall that d_u, d_v, and d_q are the vector-valued function dimensions in the definition of $\mathcal{L}$, and $d_{\mathrm{s}}, d_{\mathrm{o}}$ are the number of basis distribution functions in Assumption 2.1.

Proof. We prove the lemma by constructing the basis distribution functions $\{\psi_i\}_{i=1}^{d_{\mathrm{s}}}$ and $\{\phi_i\}_{i=1}^{d_{\mathrm{o}}}$ satisfying Assumption 2.1.

State Distribution: Note that to make $\mathcal{F}_{\mathrm{s}} \in \mathrm{conh}(\{\psi_i\}_{i=1}^{d_{\mathrm{s}}})$, it suffices to let $\{[u]_i\}_{i=1}^{d_u}$ be a subset of $\{\psi_i\}_{i=1}^{d_{\mathrm{s}}}$. In the following, we construct the rest elements of $\{\psi_i\}_{i=1}^{d_{\mathrm{s}}}$ to make $\mathcal{F}'_{\mathrm{s}} \in \mathrm{conh}(\{\psi_i\}_{i=1}^{d_{\mathrm{s}}})$. For any $(h, \theta, \pi, s_h, a_h, o_{h+1}) \in [H] \times \Theta \times \mathcal{S} \times \mathcal{A} \times \mathcal{O}$, we have

$$\begin{aligned}
&p_{\theta,\pi}(\boldsymbol{s}_h = s_h, \boldsymbol{o}_{h+1} = o_{h+1} \,|\, \boldsymbol{a}_h = a_h) \\
&\quad = \int_{\mathcal{S}} \mathcal{E}^{\theta}_{h+1}(o_{h+1} \,|\, s_{h+1}) \cdot \mathcal{T}^{\theta}_{h}(s_{h+1} \,|\, s_h, a_h)\, \mathrm{d}s_{h+1} \cdot p_{\theta,\pi}(\boldsymbol{s}_h = s_h) \\
&\quad = \Big(\int_{\mathcal{S}} \mathcal{E}^{\theta}_{h+1}(o_{h+1} \,|\, s_{h+1}) \cdot u(s_{h+1})^{\top} M^{\theta}_{h,a_h}\, \mathrm{d}s_{h+1}\Big) v_h(s_h) \cdot p_{\theta,\pi}(\boldsymbol{s}_h = s_h),
\end{aligned} \tag{B.2}$$

where the second equality is by the form of $\mathcal{T}^{\theta}_h$ in Definition B.1. Similarly, we have

$$\begin{aligned}
p_{\theta,\pi}(\boldsymbol{s}_h = s_h) &= \mathbb{E}_{\theta,\pi}[\mathcal{T}^{\theta}_{h-1}(s_h \,|\, \boldsymbol{s}_{h-1}, \boldsymbol{a}_{h-1})] \\
&= u(s_h)^{\top} \mathbb{E}_{\theta,\pi}[M^{\theta}_{h,\boldsymbol{a}_{h-1}} v_{h-1}(\boldsymbol{s}_{h-1})].
\end{aligned} \tag{B.3}$$

For notational simplicity, we define the vectors

$$\zeta_1 = \Big(\int_{\mathcal{S}} \mathcal{E}^{\theta}_{h+1}(o_{h+1} \,|\, s_{h+1}) \cdot u(s_{h+1})^{\top} M^{\theta}_{h,a_h}\, \mathrm{d}s_{h+1}\Big)^{\top}, \tag{B.4}$$

$$\zeta_2 = \mathbb{E}_{\theta,\pi}[M^{\theta}_{h,\boldsymbol{a}_{h-1}} v_{h-1}(\boldsymbol{s}_{h-1})]. \tag{B.5}$$

Then, combining (B.2)-(B.5), we can write

$$\begin{aligned}
&p_{\theta,\pi}(\boldsymbol{s}_h = s_h \,|\, \boldsymbol{o}_{h+1} = o_{h+1}, \boldsymbol{a}_h = a_h) \\
&\quad = \frac{p_{\theta,\pi}(\boldsymbol{s}_h = s_h, \boldsymbol{o}_{h+1} = o_{h+1} \,|\, \boldsymbol{a}_h = a_h)}{p_{\theta,\pi}(\boldsymbol{o}_{h+1} = o_{h+1} \,|\, \boldsymbol{a}_h = a_h)} = \frac{\zeta_1^{\top} v_h(s_h) u(s_h)^{\top} \zeta_2}{p_{\theta,\pi}(\boldsymbol{o}_{h+1} = o_{h+1} \,|\, \boldsymbol{a}_h = a_h)}.
\end{aligned} \tag{B.6}$$

Note that we can rewrite (B.6) in a linear form,

$$p_{\theta,\pi}(\boldsymbol{s}_h = \cdot \,|\, \boldsymbol{o}_{h+1} = o_{h+1}, \boldsymbol{a}_h = a_h) = \Big\langle v(\cdot)u(\cdot)^{\top}, \frac{\zeta_2 \zeta_1^{\top}}{p_{\theta,\pi}(\boldsymbol{o}_{h+1} = o_{h+1} \,|\, \boldsymbol{a}_h = a_h)} \Big\rangle_{\mathrm{tr}},$$

where $\langle \cdot, \cdot \rangle_{\mathrm{tr}}$ represents the trace inner product of matrices. Therefore, we know that any function in $\mathcal{F}'_{\mathrm{s}}$ can be represented as a convex combination of the functions

$$\{[u(\cdot)]_i \cdot [v(\cdot)]_j\}_{i \in [d_u], j \in [d_v]}. \tag{B.7}$$

Then, by normalizing each function in (B.7) as a probability distribution function, we obtain $d_v d_u$ distribution functions, whose convex combination contains all elements of $\mathcal{F}'_{\mathrm{s}}$. Thus, by denoting the set of such distribution functions plus $\{u_i\}_{i=1}^{d_{\mathrm{u}}}$ by $\{\psi_i\}_{i=1}^{d_{\mathrm{s}}}$, we have

$$d_{\mathrm{s}} = d_u(d_v + 1), \quad \text{and} \quad \mathcal{F}'_{\mathrm{s}}, \mathcal{F}'_{\mathrm{s}} \subset \mathrm{conh}(\{\psi_i\}_{i=1}^{d_{\mathrm{s}}}).$$

Observation Distribution: In the following, we construct the basis distribution functions $\{\phi_i\}_{i=1}^{d_{\mathrm{o}}}$ such that $\mathcal{F}_{\mathrm{o}} \subset \mathrm{conh}(\{\phi_i\}_{i=1}^{d_{\mathrm{o}}})$. Note that, for any $(h, \pi, o_h, o_{h+1}, o_{h+2}, a_h, a_{h+1}) \in [H-1] \times \Theta \times \mathcal{O}^3 \times \mathcal{A}^2$, we have

$$\begin{aligned}
&p_{\theta,\pi}(\boldsymbol{o}_h = o_h, \boldsymbol{o}_{h+1} = o_{h+1}, \boldsymbol{o}_{h+2} = o_{h+2} \,|\, \boldsymbol{a}_h = a_h, \boldsymbol{a}_{h+1} = a_{h+1}) \\
&\quad = \int_{\mathcal{S}^3} p_{\theta,\pi}(\boldsymbol{s}_h = s_h) \cdot \mathcal{E}^{\theta}_h(o_h \,|\, s_h, a_h) \cdot \mathcal{T}^{\theta}_h(s_{h+1} \,|\, s_h, a_h) \cdot \mathcal{E}^{\theta}_{h+1}(o_{h+1} \,|\, s_{h+1}) \\
&\qquad\qquad \cdot \mathcal{T}^{\theta}_{h+1}(s_{h+2} \,|\, s_{h+1}, a_{h+1}) \cdot \mathcal{E}^{\theta}_h(o_h \,|\, s_h)\, \mathrm{d}s_h\, \mathrm{d}s_{h+1}\, \mathrm{d}s_{h+2} \\
&\quad = \sum_{i,j,\ell=1}^{d} \omega_{i,j,\ell} \cdot q_i(o_h) \cdot q_j(o_{h+1}) \cdot q_{\ell}(o_{h+2})
\end{aligned} \tag{B.8}$$

where $\{\omega_{i,j,\ell}\}_{i,j,\ell\in[d_q]}$ are defined by

$$\omega_{i,j,\ell} = \int_{\mathcal{S}^3} p_{\theta,\pi}(\boldsymbol{s}_h = s_h)\cdot \mathcal{T}_h^\theta(s_{h+1}\,|\,s_h,a_h)\cdot \mathcal{T}_{h+1}^\theta(s_{h+2}\,|\,s_{h+1},a_{h+1}) \\ \cdot [g_h^\theta(s_h)]_i\cdot [g_h^\theta(s_{h+1})]_j\cdot [g_h^\theta(s_{h+2})]_\ell\,\mathrm{d}s_h\,\mathrm{d}s_{h+1}\,\mathrm{d}s_{h+2}.$$

following the definition of $\mathcal{E}^\theta$ in Definition B.1. For any $i,j,\ell\in[d_q]$, we define the distribution function $\phi_{i,j,\ell}\in\Delta(\mathcal{O}^3)$ by

$$\phi_{i,j,\ell}(o_h,o_{h+1},o_{h+2}) = q_i(o_h)\cdot q_j(o_{h+1})\cdot q_\ell(o_{h+2})$$

for any $o_h,o_{h+1},o_{h+2}\in\mathcal{O}$. Then, by (B.8), we have $\mathcal{F}_{\mathrm{o}}\subset \mathrm{conh}(\{\phi_{i,j,\ell}\}_{i,j,\ell=1}^{d_q})$. Reorganizing the index, we can write $\{\phi_i\}_{i=1}^{d_{\mathrm{o}}} = \{\phi_{i,j,\ell}\}_{i,j,\ell=1}^{d_q}$ with $d_{\mathrm{o}} = d_q^3$.

Therefore, we conclude the proof of Lemma B.3. □

B.2. Verification of Assumption 2.2

For any $(h,i)\in[H]\times[d_{\mathrm{s}}]$, we define $\nu_{h,i}\in\Delta(\mathcal{O})$ by

$$\nu_{h,i}(o) = \int_{\mathcal{S}} \mathcal{E}_h^\theta(o_h\,|\,s_h)\cdot\psi_i(s_h)\,\mathrm{d}s_h \tag{B.9}$$

for any $o\in\mathcal{O}$. Recall that $\{\psi_i\}_{i=1}^{d_{\mathrm{s}}}$ are the basis distribution functions in Assumption 2.1, and we prove their existence when the candidate class of the POMDP is a linear kernel POMDP set in Section B.1. Let $\widetilde{\mathcal{K}}$ be a kernel function (different from the kernel function $\mathcal{K}$ in Section 3.2) defined on $\mathcal{O}\times\mathcal{O}$. For any $h\in[H]$, we define the matrix $\Lambda_h\in\mathbb{R}^{d_{\mathrm{s}}\times d_{\mathrm{s}}}$ by

$$[\Lambda_h]_{i,j} = \mathbb{E}_{\boldsymbol{o}\sim\nu_{h,i},\boldsymbol{o}'\sim\nu_{h,j}}[\widetilde{\mathcal{K}}(\boldsymbol{o},\boldsymbol{o}')],\quad \text{for any } i,j\in[d_{\mathrm{s}}]. \tag{B.10}$$

Similar to Assumption 3.4, the following assumption specifies the regularity condition on $\widetilde{\mathcal{K}}$ and $\{\nu_{h,i}\}_{h\in[H],i\in[d_{\mathrm{o}}]}$.

Assumption B.4. The kernel function $\widetilde{\mathcal{K}}$ is bounded. In particular, we have $|\widetilde{\mathcal{K}}(x,y)|\le 1$ for any $x,y\in\mathcal{O}$. Also, we have $\Lambda_h\succ 0$ for any $h\in[H]$.

Note that for any $h\in[H]$, similar to the discussion under Assumption 3.4, the positive definiteness of the matrix Λ_h requires the RKHS embedding of $\nu_{h,1},\ldots,\nu_{h,d_{\mathrm{s}}}$ to be linearly independent. Here, the RKHS and the corresponding RKHS embedding (operator) are defined with respect to the kernel function $\widetilde{\mathcal{K}}$. As a special case, when the state space $\mathcal{S}$ and observation space $\mathcal{O}$ are finite, we let

$$\widetilde{\mathcal{K}}(x,y) = \mathbb{1}\{x=y\},\quad \text{for any } x,y\in\mathcal{O},$$

and $\{\psi_i\}_{i=1}^{d_{\mathrm{s}}} = \{\mathbb{1}\{s=\cdot\}\}_{s\in\mathcal{S}}$ with $d_{\mathrm{s}} = |\mathcal{S}|$. Note that for a finite state space $\mathcal{S}$, the integral in (B.9) is defined with respect to the counting measure over $\mathcal{S}$. Then, Assumption B.4 is equivalent to requiring the vectors $\{\mathcal{E}_h(\cdot\,|\,s)\in\mathbb{R}^{|\mathcal{O}|}\}_{s\in\mathcal{S}}$ to be linearly independent for any $h\in[H]$, which recovers the undercompleteness assumption in (Jin et al., 2020a).

The following lemma shows that any linear kernel POMDP set satisfies the invertible observation operators assumption (Assumption 2.2), given the aforementioned linear independence condition.

Lemma B.5. *Suppose the candidate class of the POMDP is a linear kernel POMDP set $\mathcal{L}$ as defined in Definition B.1 and Assumption B.4 holds. Then, we have that Assumption 2.2 holds with*

$$\mathcal{Z}_h^\theta(s,o) = \sum_{i,j=1}^{d_{\mathrm{s}}}\psi_i(s)\cdot[(\Lambda_h)^{-1}]_{i,j}\cdot\mathbb{E}_{\boldsymbol{o}\sim\nu_{h,j}}[\widetilde{\mathcal{K}}(\boldsymbol{o},o)] \tag{B.11}$$

for any $(h,\theta,s,o)\in[H]\times\Theta\times\mathcal{S}\times\mathcal{O}$ and

$$\gamma = d\cdot\max_{h\in[H]}\|(\Lambda_h)^{-1}\|_{1\to 1} = d\cdot\max_{(h,j)\in[H]\times[d_{\mathrm{s}}]}\sum_{i=1}^{d_{\mathrm{s}}}\big|[(\Lambda_h)^{-1}]_{i,j}\big|.$$

Here, the matrix Λ_h is defined in (B.10).

Proof. By the definitions of the operators $\mathbb{Z}_h^\theta$ and $\mathbb{O}_h^\theta$, and function $\mathcal{Z}_h^\theta$ in Assumption 2.2, (2.4), and (B.11), respectively, for any $f \in L^1(\mathcal{S})$, we have

$$(\mathbb{Z}_h^\theta \mathbb{O}_h^\theta f)(s) = \sum_{i,j=1}^{d_s} \int_{\mathcal{S}\times\mathcal{O}^2} \psi_i(s) \cdot [(\Lambda_h)^{-1}]_{i,j} \cdot \nu_{h,j}(o') \cdot \widetilde{\mathcal{K}}(o', o) \cdot \mathcal{E}_h^\theta(o \,|\, s') \cdot f(s') \,\mathrm{d}s' \,\mathrm{d}o \,\mathrm{d}o'. \tag{B.12}$$

When $f \in \mathrm{linspan}(\{\psi_i\}_{i=1}^{d_s})$ with $f = \sum_{i=1}^{d_s} \psi_i \cdot c_i$, we have

$$\int_{\mathcal{S}} \mathcal{E}_h^\theta(o' \,|\, s') \cdot f(s') \,\mathrm{d}s' = \int_{\mathcal{S}} \sum_{\ell=1}^{d_s} \mathcal{E}_h^\theta(o' \,|\, s') \cdot \psi_\ell(s') \cdot c_\ell = \sum_{\ell=1}^{d_s} \nu_{h,\ell}(o') \cdot c_\ell, \tag{B.13}$$

where the last equality is by the definition of $\nu_{h,\ell}$ in (B.9). Plugging (B.13) into the right-hand side of (B.12), we obtain

$$\begin{aligned}
(\mathbb{Z}_h^\theta \mathbb{O}_h^\theta f)(s) &= \sum_{i,j=1}^{d_s} \int_{\mathcal{O}^2} \psi_i(s) \cdot [(\Lambda_h)^{-1}]_{i,j} \cdot \nu_{h,j}(o') \cdot \widetilde{\mathcal{K}}(o', o) \cdot \sum_{\ell=1}^{d_s} \nu_{h,\ell}(o') \cdot c_\ell \,\mathrm{d}o \,\mathrm{d}o' \\
&= \sum_{i,j,\ell=1}^{d_s} \psi_i(s) \cdot [(\Lambda_h)^{-1}]_{i,j} \cdot \Big(\int_{\mathcal{O}^2} \nu_{h,j}(o') \cdot \widetilde{\mathcal{K}}(o', o) \cdot \nu_{h,\ell}(o') \,\mathrm{d}o \,\mathrm{d}o' \Big) \cdot c_\ell \\
&= \sum_{i,j,\ell=1}^{d_s} \psi_i(s) \cdot [(\Lambda_h)^{-1}]_{i,j} \cdot [\Lambda_h]_{j,\ell} \cdot c_\ell.
\end{aligned}$$

Here, the last equality uses the definition of the matrix Λ_h in (B.10). By the definition of the inverse matrix, we have

$$\sum_{j=1}^{d_s} [(\Lambda_h)^{-1}]_{i,j} \cdot [\Lambda_h]_{j,\ell} = \mathbb{1}\{i = \ell\},$$

which implies

$$(\mathbb{Z}_h^\theta \mathbb{O}_h^\theta f)(s) = \sum_{i=1}^{d_s} \psi_i(s) \cdot c_i = f(s), \quad \text{for any } s \in \mathcal{S}.$$

In the following, we characterize the operator norm $\|\cdot\|_{1\to 1}$ of $\mathbb{Z}_h^\theta$. For any $(h, \theta, o) \in [H] \times \Theta \times \mathcal{O}$, we have

$$\begin{aligned}
\int_{\mathcal{S}} |\mathcal{Z}_h^\theta(s, o)| \,\mathrm{d}s &= \int_{\mathcal{S}} \Big| \sum_{i,j=1}^{d_s} \psi_i(s) \cdot [(\Lambda_h)^{-1}]_{i,j} \cdot \int_{\mathcal{O}} \nu_{h,j}(o') \cdot \widetilde{\mathcal{K}}(o', o) \,\mathrm{d}o' \Big| \,\mathrm{d}s \\
&\le \int_{\mathcal{S}} \sum_{i=1}^{d_s} \psi_i(s) \cdot \Big| \sum_{j=1}^{d_s} [(\Lambda_h)^{-1}]_{i,j} \cdot \int_{\mathcal{O}} \nu_{h,j}(o') \cdot \widetilde{\mathcal{K}}(o', o) \,\mathrm{d}o' \Big| \,\mathrm{d}s \\
&= \sum_{i=1}^{d_s} \Big| \sum_{j=1}^{d_s} [(\Lambda_h)^{-1}]_{i,j} \cdot \int_{\mathcal{O}} \nu_{h,j}(o') \cdot \widetilde{\mathcal{K}}(o', o) \,\mathrm{d}o' \Big|,
\end{aligned} \tag{B.14}$$

where the last equality is by the fact that ψ_i is a distribution function over $\mathcal{S}$ for any $i \in [d_s]$. The right-hand side of (B.14) is upper bounded by

$$\begin{aligned}
&\sum_{i=1}^{d_s} \Big| \sum_{j=1}^{d_s} [(\Lambda_h)^{-1}]_{i,j} \cdot \int_{\mathcal{O}} \nu_{h,j}(o') \cdot \widetilde{\mathcal{K}}(o', o) \,\mathrm{d}o' \Big| \\
&\quad \le \sum_{i=1}^{d_s} \sum_{j=1}^{d_s} \big| [(\Lambda_h)^{-1}]_{i,j} \big| \cdot \Big| \int_{\mathcal{O}} \nu_{h,j}(o') \cdot \widetilde{\mathcal{K}}(o', o) \,\mathrm{d}o' \Big| \\
&\quad \le \Big(\max_{j\in[d_s]} \sum_{i=1}^{d_s} \big| [(\Lambda_h)^{-1}]_{i,j} \big| \Big) \cdot \sum_{j=1}^{d_s} \Big| \int_{\mathcal{O}} \nu_{h,j}(o') \cdot \widetilde{\mathcal{K}}(o', o) \,\mathrm{d}o' \Big|.
\end{aligned} \tag{B.15}$$

For any $(h, j) \in [H] \times [d_{\mathrm{s}}]$, because the kernel function $\widetilde{\mathcal{K}}$ is uniformly bounded by 1, following Assumption B.4, and $\nu_{h,j}$ is a distribution function over $\mathcal{O}$, we have

$$\Big|\int_{\mathcal{O}} \nu_{h,j}(o') \cdot \widetilde{\mathcal{K}}(o', o)\,\mathrm{d}o'\Big| \le 1. \tag{B.16}$$

Combining (B.14)-(B.16), we have

$$\int_{\mathcal{S}} |\mathcal{Z}_h^\theta(s, o)|\,\mathrm{d}s \le \Big(\max_{j\in[d_{\mathrm{s}}]} \sum_{i=1}^{d_{\mathrm{s}}} \big|[(\Lambda_h)^{-1}]_{i,j}\big|\Big) \cdot d, \tag{B.17}$$

for any $(h, \theta, o) \in [H] \times \Theta \times \mathcal{O}$. Note that for any $f \in L^1(\mathcal{O})$, we have

$$\begin{aligned}
\|\mathbb{Z}_h^\theta f\|_1 &= \int_{\mathcal{S}} \Big|\int_{\mathcal{O}} \mathcal{Z}_h^\theta(s, o) \cdot f(o)\,\mathrm{d}o\Big|\,\mathrm{d}s \\
&\le \int_{\mathcal{S}\times\mathcal{O}} |\mathcal{Z}_h^\theta(s, o)| \cdot |f(o)|\,\mathrm{d}o\,\mathrm{d}s \\
&= \int_{\mathcal{O}} \Big(\int_{\mathcal{S}} |\mathcal{Z}_h^\theta(s, o)|\,\mathrm{d}s\Big) \cdot |f(o)|\,\mathrm{d}o,
\end{aligned}$$

combining which with (B.17), we obtain

$$\|\mathbb{Z}_h^\theta f\|_1 \le \Big(\max_{j\in[d_{\mathrm{s}}]} \sum_{i=1}^{d_{\mathrm{s}}} \big|[(\Lambda_h)^{-1}]_{i,j}\big|\Big) \cdot d \cdot \|f\|_1.$$

Therefore, we conclude the proof of Lemma B.5. □

C. Minimax Optimization in OP-TENET

In this section, we discuss the details on how to implement the computation of OP-TENET in practice. Recall that in the planning phase (introduced in Section 3.2) of each iteration of OP-TENET, we only consider the parameter θ such that

$$L(\theta) = \max_{f\in L^\infty(\mathcal{O}^3):\|f\|_\infty\le 1} \max_{(h,a,a')\in\{2,\ldots,H\}\times\mathcal{A}^2} \mathbb{E}_{X\sim\widehat{\mathcal{D}}_{h,a,a'}}[(\mathbb{S}\mathbb{F}_{h,a'}^\theta f - \mathbb{S}f)(X)]$$

is sufficiently small. Here, $\mathbb{S}$ and $\mathbb{F}_{h,a'}^\theta$ are operators defined in (3.18) and (3.13), respectively. Also, $\widehat{\mathcal{D}}_{h,a,a'}$ is the empirical distribution induced by $\mathcal{D}_{h,a,a'}$, which consists of k observation tuples with k being the iteration index. For ease of presentation, we assume that we have access to the following planning oracle.

Oracle C.1. We denote by $\widehat{\pi}$ a planning oracle for any given POMDP. In other word, the mapping $\widehat{\pi} : \Theta \to \Pi$ satisfies $\widehat{\pi}(\theta) \in \mathrm{argmax}_{\pi\in\Pi} J(\theta, \pi)$ for any $\theta \in \Theta$.

With the planning oracle defined above, we select the parameter θ_k by solving the following constrained optimization problem,

$$\min_{\theta\in\Theta} J\big(\theta, \widehat{\pi}(\theta)\big) \quad \text{s.t.} \quad L(\theta) \le \beta \cdot k^{-1/2}. \tag{C.1}$$

Then, we select the policy $\pi_k = \widehat{\pi}(\theta_k)$.

C.1. Lagrangian Relaxation

In the sequel, we handle the constraint in (C.1) via the Lagrangian relaxation. In detail, solving (C.1) is equivalent to solving the minimax optimization problem,

$$\min_{\theta\in\Theta} \max_{\lambda\ge 0} \; -J\big(\theta, \widehat{\pi}(\theta)\big) + \sum_{(h,a,a')\in\mathcal{I}} \lambda_{h,a,a'} \cdot \big(\widetilde{L}_{h,a,a'}(\theta) - \beta \cdot k^{-1/2}\big), \tag{C.2}$$

where $\lambda = (\lambda_{h,a,a'})_{(h,a,a')\in\mathcal{J}} \in \mathbb{R}^{(H-1)A^2}$ and $\widetilde{L}_{h,a,a'}(\theta)$ is defined by

$$\widetilde{L}_{h,a,a'}(\theta) = \max_{f\in L^\infty(\mathcal{O}^3):\|f\|_\infty\le 1} \mathbb{E}_{X\sim\widehat{\mathcal{D}}_{h,a,a'}}[(\mathbb{SF}^\theta_{h,a'}f - \mathbb{S}f)(X)]. \tag{C.3}$$

Here, for notational simplicity, we denote by $\mathcal{I}$ the set $\mathcal{I} = \{2,\ldots,H\}\times\mathcal{A}^2$. Note that for each $(h,a,a')\in\mathcal{I}$, we need to search a function within a ball in $L^\infty(\mathcal{O}^3)$. To this end, we propose to search functions within a large function approximator class, for example, a sufficiently large neural network. In detail, we denote by $f^w_{h,a,a'}$ the parametrization of the function approximator, where w is the parameter with a candidate set $\mathcal{W}$. For example, we can build a neural network whose input space is $\mathcal{O}^3$ and output space is $\mathbb{R}^{(H-1)A^2}$. Then, w represents the weights of all layers and

$$\big(f^w_{h,a,a'}(x)\big)_{(h,a,a')\in\mathcal{I}} \in \mathbb{R}^{(H-1)A^2}$$

is the output of the neural network corresponding any input $x\in\mathcal{O}^3$. Moreover, by properly choosing the activation function of the output layer, we are able to make $\|f^w_{h,a,a'}\|_\infty \le 1$ for any $w\in\mathcal{W}$ and $(h,a,a')\in\mathcal{I}$. Then, we approximately compute $\widetilde{L}_{h,a,a'}(\theta)$ in (C.3) by computing $\max_{w\in\mathcal{W}}\widehat{L}^w_{h,a,a'}(\theta)$, where

$$\widehat{L}^w_{h,a,a'}(\theta) = \mathbb{E}_{X\sim\widehat{\mathcal{D}}_{h,a,a'}}[(\mathbb{SF}^\theta_{h,a'}f^w_{h,a,a'} - \mathbb{S}f^w_{h,a,a'})(X)], \tag{C.4}$$

for any $w\in\mathcal{W}$. Combining (C.2)-(C.4), we approximately solve the constrained optimization problem in (C.1) by solving

$$\min_{\theta\in\Theta}\max_{\lambda\ge 0, w\in\mathcal{W}} \; -J\big(\theta,\widehat{\pi}(\theta)\big) + \sum_{(h,a,a')\in\mathcal{I}} \lambda_{h,a,a'}\cdot\big(\widehat{L}^w_{h,a,a'}(\theta) - \beta\cdot k^{-1/2}\big). \tag{C.5}$$

C.2. Stochastic Gradient Method

We denote by

$$\mathcal{L}(\theta,\lambda,w) = -J\big(\theta,\widehat{\pi}(\theta)\big) + \sum_{(h,a,a')\in\mathcal{I}} \lambda_{h,a,a'}\cdot\big(\widehat{L}^w_{h,a,a'}(\theta) - \beta\cdot k^{-1/2}\big)$$

the minimax objective in (C.5). In the sequel, we consider the stochastic gradient method for solving the minimax optimization problem in (C.5). In detail, suppose that we have unbiased stochastic gradient estimators g_θ, g_λ, and g_w such that

$$\begin{aligned}
\mathbb{E}[g_\theta(\theta,\lambda,w)] &= \nabla_\theta\mathcal{L}(\theta,\lambda,w),\\
\mathbb{E}[g_\lambda(\theta,\lambda,w)] &= \nabla_\lambda\mathcal{L}(\theta,\lambda,w),\\
\mathbb{E}[g_w(\theta,\lambda,w)] &= \nabla_w\mathcal{L}(\theta,\lambda,w),
\end{aligned}$$

for any $(\theta,\lambda,w)\in\Theta\times\mathbb{R}^{(H-1)A^2}\times\mathcal{W}$. Also, computing g_θ, g_λ, and g_w does not require access to the full data set $\mathcal{D} = \{\mathcal{D}_{h,a,a'}\}_{h\in\{2,\ldots,H\}\times\mathcal{A}^2}$ and thus has a low computation cost. In each iteration, starting from any (θ,λ,w) within the candidate set, we first update λ and w by running

$$\lambda \leftarrow \lambda + \eta_\lambda\cdot g_\lambda(\theta,\lambda,w), \quad w\leftarrow w + \eta_w\cdot g_w(\theta,\lambda,w)$$

for N_{dual} steps. Then, we update θ by running

$$\theta\leftarrow\theta - \eta_\theta\cdot g_\theta(\theta,\lambda,w).$$

for N_{primal} steps. Here, η_θ, η_λ, η_w are constant stepsizes. Note that after each update, if the updated parameter is not in the candidate set, we need to run an extra projection step, which replaces the updated parameter by its closest neighbor within the candidate set. Because we need to call the planning oracle $\widehat{\pi}$ after updating θ, which has a relatively high computation cost, it is better to set $N_{\mathrm{primal}} = 1$ and set N_{dual} as a large number.

In the sequel, we construct unbiased gradient estimators for the objective $\mathcal{L}(\theta,\lambda,w)$.

Construction of g_λ: To construct $g_\lambda(\theta, \lambda, x)$, note that we have

$$\partial_{\lambda_{h,a,a'}} \mathcal{L}(\theta, \lambda, w) = \widehat{L}^w_{h,a,a'}(\theta) - \beta \cdot k^{-1/2}$$

for any $(h, a, a') \in \{2, \ldots, H\} \times \mathcal{A}^2$. Let $\mathcal{B}$ be a batch of data sampled from $\mathcal{D}_{h,a,a'}$ uniformly at random. Then, by computing the batch average, we have

$$\mathbb{E}\Big[\frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}}(\mathbb{SF}^\theta_{h,a'} f^w_{h,a,a'} - \mathbb{S} f^w_{h,a,a'})(x) - \beta \cdot k^{-1/2}\Big] = \partial_{\lambda_{h,a,a'}} \mathcal{L}(\theta, \lambda, w). \tag{C.6}$$

For any $(i, j, x) \in [d_{\mathrm{o}}] \times [d_{\mathrm{o}}] \times \mathcal{O}^3$, let $Y_{i,j,x}$ and $Y'_{i,j,x}$ be independent random variables in $\mathcal{O}^3$ sampled from the distributions ϕ_i and ϕ_j, respectively. Then, by the definition of the operator $\mathbb{S}$ in (3.18), we have

$$(\mathbb{S} f^w_{h,a,a'})(x) = \sum_{i,j\in[d_{\mathrm{o}}]} \mathbb{E}\big[[G^{-1}]_{i,j} \cdot \mathcal{K}(x, Y_{i,j,x}) \cdot f^w_{h,a,a'}(Y'_{i,j,x})\big]. \tag{C.7}$$

Similarly, applying the definition of the operator $\mathbb{F}^\theta_{h,a'}$ in (3.13), we have

$$\begin{aligned}&(\mathbb{SF}^\theta_{h,a'} f^w_{h,a,a'})(x) \\ &= \sum_{i,j\in[d_{\mathrm{o}}]} \mathbb{E}\Big[[G^{-1}]_{i,j} \cdot \mathcal{K}(x, Y_{i,j,x}) \cdot \int_{\mathcal{O}^2} f^w_{h,a,a'}(\boldsymbol{o}_{h-1}, \widetilde{o}_h, \widetilde{o}_{h+1}) \cdot \mathcal{B}^\theta_{h,a}(\boldsymbol{o}_h, \widetilde{o}_h, \widetilde{o}_{h+1})\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1}\Big],\end{aligned} \tag{C.8}$$

where we denote $Y'_{i,j,x} = (\boldsymbol{o}_{h-1}, \boldsymbol{o}_h, \boldsymbol{o}_{h+1})$. Moreover, let ϕ_{ip} be a distribution supported on $\mathcal{O}^2$ and $\widetilde{Y}_{i,j,x} = (\widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1})$ be a random variable in $\mathcal{O}^2$ sampled from ϕ_{ip}. Then, following the idea of importance sampling, we can rewrite (C.8) as

$$\begin{aligned}&(\mathbb{SF}^\theta_{h,a'} f^w_{h,a,a'})(x) \\ &= \sum_{i,j\in[d_{\mathrm{o}}]} \mathbb{E}\Big[[G^{-1}]_{i,j} \cdot \mathcal{K}(x, Y_{i,j,x}) \cdot \frac{f^w_{h,a,a'}(\boldsymbol{o}_{h-1}, \widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1}) \cdot \mathcal{B}^\theta_{h,a}(\boldsymbol{o}_h, \widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1})}{\phi_{\mathrm{ip}}(\widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1})}\Big].\end{aligned} \tag{C.9}$$

Combining (C.6)-(C.9), we construct $g_\lambda(\theta, \lambda, x)$ by

$$[g_\lambda(\theta, \lambda, x)]_{h,a,a'} = \frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}}\big(\widehat{\mathbb{SF}^\theta_{h,a'} f^w_{h,a,a'}}(x) - \widehat{\mathbb{S} f^w_{h,a,a'}}(x)\big) - \beta \cdot k^{-1/2},$$

for any $(h, a, a') \in \{2, \ldots, H\} \times \mathcal{A}^2$, where

$$\begin{aligned}\widehat{\mathbb{SF}^\theta_{h,a'} f^w_{h,a,a'}}(x) &= \sum_{i,j\in[d_{\mathrm{o}}]} [G^{-1}]_{i,j} \cdot \mathcal{K}(x, Y_{i,j,x}) \cdot \frac{f^w_{h,a,a'}(\boldsymbol{o}_{h-1}, \widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1}) \cdot \mathcal{B}^\theta_{h,a}(\boldsymbol{o}_h, \widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1})}{\phi_{\mathrm{ip}}(\widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1})}, \\ \widehat{\mathbb{S} f^w_{h,a,a'}}(x) &= \sum_{i,j\in[d_{\mathrm{o}}]} [G^{-1}]_{i,j} \cdot \mathcal{K}(x, Y_{i,j,x}) \cdot f^w_{h,a,a'}(Y'_{i,j,x}).\end{aligned}$$

Construction of g_w: To construct $g_w(\theta, \lambda, x)$, note that we have

$$\begin{aligned}\nabla_w \mathcal{L}(\theta, \lambda, w) &= \sum_{(h,a,a')\in\mathcal{I}} \lambda_{h,a,a'} \cdot \nabla_w \widehat{L}^w_{h,a,a'}(\theta) \\ &= \sum_{(h,a,a')\in\mathcal{I}} \lambda_{h,a,a'} \cdot \mathbb{E}_{X\sim\widehat{\mathcal{D}}_{h,a,a'}}[(\nabla_w \mathbb{SF}^\theta_{h,a'} f^w_{h,a,a'} - \nabla_w \mathbb{S} f^w_{h,a,a'})(X)].\end{aligned}$$

Thus, following the similar argument as in (C.7)-(C.9), we construct $g_w(\theta, \lambda, w)$ as

$$g_w(\theta, \lambda, w) = \sum_{(h,a,a')\in\mathcal{I}} \lambda_{h,a,a'} \cdot \frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}}\big(\nabla_w \widehat{\mathbb{SF}^\theta_{h,a'} f^w_{h,a,a'}}(x) - \nabla_w \widehat{\mathbb{S} f^w_{h,a,a'}}(x)\big)$$

where

$$\nabla_w \widehat{\mathbb{SF}^{\theta}_{h,a'} f^w_{h,a,a'}}(x) = \sum_{i,j\in[d_o]} [G^{-1}]_{i,j} \cdot \mathcal{K}(x, Y_{i,j,x}) \cdot \frac{\nabla_w f^w_{h,a,a'}(\boldsymbol{o}_{h-1}, \widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1}) \cdot \mathcal{B}^{\theta}_{h,a}(\boldsymbol{o}_h, \widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1})}{\phi_{\mathrm{ip}}(\widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1})},$$
$$\nabla_w \widehat{\mathbb{SF}^{w}_{h,a,a'}}(x) = \sum_{i,j\in[d_o]} [G^{-1}]_{i,j} \cdot \mathcal{K}(x, Y_{i,j,x}) \cdot \nabla_w f^w_{h,a,a'}(Y'_{i,j,x}).$$

Here, $Y_{i,j,x}$, $\boldsymbol{o}_{h-1}$, $\boldsymbol{o}_h$, $\widetilde{\boldsymbol{o}}_h$, and $\widetilde{\boldsymbol{o}}_{h+1}$ are random variables defined the same as in the construction of g_λ.

Construction of g_θ: To construct $g_\theta(\theta, \lambda, x)$, note that we have

$$\begin{aligned}\nabla_w \mathcal{L}(\theta, \lambda, w) &= -\nabla_\theta J\big(\theta, \widehat{\pi}(\theta)\big) + \sum_{(h,a,a')\in\mathcal{I}} \lambda_{h,a,a'} \cdot \nabla_\theta \widehat{L}^w_{h,a,a'}(\theta) \\ &= -\nabla_\theta J\big(\theta, \widehat{\pi}(\theta)\big) + \sum_{(h,a,a')\in\mathcal{I}} \lambda_{h,a,a'} \cdot \mathbb{E}_{X\sim\widehat{\mathcal{D}}_{h,a,a'}}[(\nabla_\theta \mathbb{SF}^{\theta}_{h,a'} f^w_{h,a,a'})(X)]. \end{aligned} \tag{C.10}$$

Following the similar argument as in (C.8)-(C.9), we have the following unbiased estimator of the expectation in the second term on the right-hand side of (C.10),

$$\begin{aligned}&\frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}} \nabla_\theta \widehat{\mathbb{SF}^{\theta}_{h,a'} f^w_{h,a,a'}}(x) \\ &= \frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}}\sum_{i,j\in[d_o]} [G^{-1}]_{i,j} \cdot \mathcal{K}(x, Y_{i,j,x}) \cdot \frac{f^w_{h,a,a'}(\boldsymbol{o}_{h-1}, \widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1}) \cdot \nabla_\theta \mathcal{B}^{\theta}_{h,a}(\boldsymbol{o}_h, \widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1})}{\phi_{\mathrm{ip}}(\widetilde{\boldsymbol{o}}_h, \widetilde{\boldsymbol{o}}_{h+1})}. \end{aligned} \tag{C.11}$$

It remains to construct an unbiased estimator of the first term on the right-hand side of (C.10). Note that $\widehat{\pi}(\theta)$ is the maximizer of $J(\theta, \cdot)$. Thus, by the envelop theorem, it suffices to estimate

$$\big(\nabla_\theta J(\theta, \pi)\big)\big|_{\pi=\widehat{\pi}(\theta)}. \tag{C.12}$$

Note that for any $(\theta, \pi) \in \Theta \times \Pi$, we have

$$\begin{aligned}J(\theta, \pi) = \int_{\mathcal{S}^H\times\mathcal{O}^H} &R\big(o_1, \pi(\tau_1), \ldots, o_H, \pi(\tau_H)\big) \cdot \mu(s_1) \cdot \mathcal{E}_1(o_1 \,|\, s_1) \\ &\cdot \Big(\prod_{h=1}^{H-1} \mathcal{T}^{\theta}_h\big(s_{h+1} \,|\, s_h, \pi(\tau_h)\big) \cdot \mathcal{E}^{\theta}_{h+1}(o_{h+1} \,|\, s_{h+1})\Big)\, \mathrm{d}o_1 \cdots \mathrm{d}o_H\, \mathrm{d}s_1 \cdots \mathrm{d}s_H.\end{aligned}$$

Then, using the chain rule of the derivative, we have

$$\begin{aligned}\nabla_\theta J(\theta, \pi) = \sum_{i=1}^{H-1} \int_{\mathcal{S}^H\times\mathcal{O}^H} &R\big(o_1, \pi(\tau_1), \ldots, o_H, \pi(\tau_H)\big) \cdot \mu(s_1) \cdot \mathcal{E}_1(o_1 \,|\, s_1) \\ &\cdot \Big(\nabla_\theta \mathcal{T}^{\theta}_h\big(s_{h+1} \,|\, s_h, \pi(\tau_h)\big) \cdot \mathcal{E}^{\theta}_{h+1}(o_{h+1} \,|\, s_{h+1}) \\ &\qquad + \mathcal{T}^{\theta}_h\big(s_{h+1} \,|\, s_h, \pi(\tau_h)\big) \cdot \nabla_\theta \mathcal{E}^{\theta}_{h+1}(o_{h+1} \,|\, s_{h+1})\Big) \\ &\cdot \Big(\prod_{h=1, h\neq i}^{H-1} \mathcal{T}^{\theta}_h\big(s_{h+1} \,|\, s_h, \pi(\tau_h)\big) \cdot \mathcal{E}^{\theta}_{h+1}(o_{h+1} \,|\, s_{h+1})\Big)\, \mathrm{d}o_1 \cdots \mathrm{d}o_H\, \mathrm{d}s_1 \cdots \mathrm{d}s_H.\end{aligned} \tag{C.13}$$

Using the relation $\nabla \ln f = \nabla f / f$, we can rewrite the right-hand side of (C.13) to obtain

$$\begin{aligned}\nabla_\theta J(\theta, \pi) = \sum_{i=1}^{H-1} \int_{\mathcal{S}^H\times\mathcal{O}^H} &R\big(o_1, \pi(\tau_1), \ldots, o_H, \pi(\tau_H)\big) \cdot \mu(s_1) \cdot \mathcal{E}_1(o_1 \,|\, s_1) \\ &\cdot \Big(\nabla_\theta \ln \mathcal{T}^{\theta}_i\big(s_{i+1} \,|\, s_i, \pi(\tau_i)\big) + \nabla_\theta \ln \mathcal{E}^{\theta}_{i+1}(o_{i+1} \,|\, s_{i+1})\Big) \\ &\cdot \Big(\prod_{h=1}^{H-1} \mathcal{T}^{\theta}_h\big(s_{h+1} \,|\, s_h, \pi(\tau_h)\big) \cdot \mathcal{E}^{\theta}_{h+1}(o_{h+1} \,|\, s_{h+1})\Big)\, \mathrm{d}o_1 \cdots \mathrm{d}o_H\, \mathrm{d}s_1 \cdots \mathrm{d}s_H.\end{aligned}$$

Therefore, we have the following unbiased estimator of $\nabla_\theta J(\theta,\pi)$,

$$\widehat{\nabla_\theta J}(\theta,\pi) = R(\boldsymbol{o}_1,\boldsymbol{a}_1,\ldots,\boldsymbol{o}_H,\boldsymbol{a}_H)\cdot\sum_{i=1}^{H-1}\big(\nabla_\theta \ln \mathcal{T}_i^\theta(\boldsymbol{s}_{i+1}\,|\,\boldsymbol{s}_i,\boldsymbol{a}_i)+\nabla_\theta \ln \mathcal{E}_{i+1}^\theta(\boldsymbol{o}_{i+1}\,|\,\boldsymbol{s}_{i+1})\big),$$

where $(\boldsymbol{s}_1,\boldsymbol{o}_1,\boldsymbol{a}_1,\ldots,\boldsymbol{s}_H,\boldsymbol{o}_H,\boldsymbol{a}_H)$ is a trajectory of the POMDP with respect to the parameter θ and policy π. Note that for the given parameter θ, the trajectory can be obtained from a simulator rather than the real environment, which does not affect the sample complexity result in the main paper. Combining the above estimator with (C.10)-(C.12), we construct $g_\theta(\theta,\lambda,w)$ as

$$g_\theta(\theta,\lambda,w) = \widehat{\nabla_\theta J}\big(\theta,\widehat{\pi}(\theta)\big)+\sum_{(h,a,a')\in\mathcal{I}}\lambda_{h,a,a'}\cdot\frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}}\nabla_\theta\widehat{\mathbb{SF}^\theta_{h,a'}f^w_{h,a,a'}}(x),$$

where the second term is defined in (C.11).

D. Proofs for Section 3

In this section, we present the proofs for the results in Section 3.

D.1. Proof of Lemma 3.1

Proof. Following the notation in Lemma 3.1 and by the definition of $\mathbb{B}_h^{\theta,\pi}$ in (3.4), we have

$$\begin{aligned}&\mathbb{E}_\theta[(\mathbb{B}_h^{\theta,\pi}f)(\overline{\boldsymbol{\tau}}_h)\,|\,\sigma_{h-1}] \qquad\text{(D.1)}\\ &=\int_{\mathcal{S}\times\mathcal{O}^3} f\big(\overline{\tau}_h^\dagger,\pi(\tau_h^\dagger),\widetilde{o}_{h+1}\big)\cdot p_\theta\big(\boldsymbol{o}_h=\widetilde{o}_h,\boldsymbol{o}_{h+1}=\widetilde{o}_{h+1}\,|\,\boldsymbol{s}_h=\widetilde{s}_h,\boldsymbol{a}_h=\pi(\tau_h^\dagger)\big)\\ &\qquad\cdot\mathcal{Z}_h^\theta(\widetilde{s}_h,o_h)\cdot p_\theta(\boldsymbol{o}_h=o_h\,|\,\sigma_{h-1})\,\mathrm{d}o_h\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1}\,\mathrm{d}\widetilde{s}_h.\end{aligned}$$

Here, invoking Lemma G.1, we have

$$\int_{\mathcal{O}}\mathcal{Z}_h^\theta(\widetilde{s}_h,o_h)\cdot p_\theta(\boldsymbol{o}_h=o_h\,|\,\sigma_{h-1})\,\mathrm{d}o_h = p_\theta(\boldsymbol{s}_h=\widetilde{s}_h\,|\,\sigma_{h-1}).$$

Thus, we can rewrite (D.1) as

$$\begin{aligned}&\mathbb{E}_\theta[(\mathbb{B}_h^{\theta,\pi}f)(\overline{\boldsymbol{\tau}}_h)\,|\,\sigma_{h-1}] \qquad\text{(D.2)}\\ &=\int_{\mathcal{S}\times\mathcal{O}^2} f\big(\overline{\tau}_h^\dagger,\pi(\tau_h^\dagger),\widetilde{o}_{h+1}\big)\cdot p_\theta\big(\boldsymbol{o}_h=\widetilde{o}_h,\boldsymbol{o}_{h+1}=\widetilde{o}_{h+1}\,|\,\boldsymbol{s}_h=\widetilde{s}_h,\boldsymbol{a}_h=\pi(\overline{\tau}_h^\dagger)\big)\\ &\qquad\cdot p(\boldsymbol{s}_h=\widetilde{s}_h\,|\,\sigma_{h-1})\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1}\,\mathrm{d}\widetilde{s}_h\\ &=\int_{\mathcal{O}^2} f\big(\overline{\tau}_h^\dagger,\pi(\tau_h^\dagger),\widetilde{o}_{h+1}\big)\cdot p_{\theta,\pi}(\boldsymbol{o}_h=\widetilde{o}_h,\boldsymbol{o}_{h+1}=\widetilde{o}_{h+1}\,|\,\sigma_{h-1})\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1}.\end{aligned}$$

where the second equality uses the independence between $(\boldsymbol{o}_h,\boldsymbol{o}_{h+1})$ and $\boldsymbol{\tau}_{h-1}$ conditioning on $(\boldsymbol{s}_h,\boldsymbol{a}_h)$. Replacing the notations $\widetilde{o}_h$ and $\widetilde{o}_{h+1}$ of the integral variables on the right-hand side of (D.2) by o_h and o_{h+1}, respectively, we obtain

$$\begin{aligned}&\mathbb{E}_\theta[(\mathbb{B}_h^{\theta,\pi}f)(\overline{\boldsymbol{\tau}}_h)\,|\,\sigma_{h-1}] \qquad\text{(D.3)}\\ &=\int_{\mathcal{O}^2} f\big(\overline{\tau}_h,\pi(\tau_h),o_{h+1}\big)\cdot p_{\theta,\pi}(\boldsymbol{o}_h=o_h,\boldsymbol{o}_{h+1}=o_{h+1}\,|\,\sigma_{h-1})\,\mathrm{d}o_h\,\mathrm{d}o_{h+1}\\ &=\mathbb{E}_{\theta,\pi}[f(\overline{\boldsymbol{\tau}}_{h+1})\,|\,\sigma_{h-1}],\end{aligned}$$

where we denote $\overline{\tau}_h=(\overline{\tau}_{h-1},a_{h-1},o_h)$ and $\tau_h=(\tau_{h-1},o_h)$. On the other hand, by the tower property of the expectation and the definition of $\mathbb{P}_h^{\theta,\pi}$ in (3.1), we have

$$\mathbb{E}_{\theta,\pi}[f(\overline{\boldsymbol{\tau}}_{h+1})\,|\,\sigma_{h-1}] = \mathbb{E}_{\theta,\pi}\big[\mathbb{E}_{\theta,\pi}[f(\overline{\boldsymbol{\tau}}_{h+1})\,|\,\overline{\boldsymbol{\tau}}_h]\,\big|\,\sigma_{h-1}\big] = \mathbb{E}_{\theta,\pi}[(\mathbb{P}_h^{\theta,\pi}f)(\overline{\boldsymbol{\tau}}_h)\,|\,\sigma_{h-1}]. \qquad\text{(D.4)}$$

Combining (D.3) and (D.4), we conclude the proof of Lemma 3.1. □

D.2. Proof of Corollary 3.2

Proof. We prove the result in (3.12) by induction over $h \in [H+1]$. When $h = H+1$, by the definition of the value function in (3.10), we have

$$\mathbb{E}_{\theta,\pi}[V_{H+1}^{\theta,\pi}(\overline{\boldsymbol{\tau}}_{H+1}) \,|\, \sigma_H] = \mathbb{E}_{\theta,\pi}[R(\overline{\boldsymbol{\tau}}_{H+1}) \,|\, \sigma_H] = \mathbb{E}_{\theta,\pi}\Big[\sum_{i=1}^{H} \boldsymbol{r}_i \,\Big|\, \sigma_H\Big].$$

for any $(\overline{\tau}_H, a_H) \in \overline{\Gamma}_H \times \mathcal{A}$. Recall that the variables $\overline{\tau}_H$ and a_H appear in the event σ_H, which is defined in (3.7).

Assume that (3.12) holds when $h = j+1$ for some fixed $j \le H$. In other words, assume that we have

$$\mathbb{E}_{\theta,\pi}[V_{j+1}^{\theta,\pi}(\overline{\boldsymbol{\tau}}_{j+1}) \,|\, \sigma_j] = \mathbb{E}_{\theta,\pi}\Big[\sum_{i=1}^{H} \boldsymbol{r}_i \,\Big|\, \sigma_j\Big] \tag{D.5}$$

for any $(\overline{\tau}_j, a_j) \in \overline{\Gamma}_j \times \mathcal{A}$. Then, by the definition of the value function in (3.10) and invoking Lemma 3.1, we have

$$\begin{aligned}
\mathbb{E}_{\theta,\pi}[V_j^{\theta,\pi}(\overline{\boldsymbol{\tau}}_j) \,|\, \sigma_{j-1}] &= \mathbb{E}_{\theta,\pi}[(\mathbb{B}_j^{\theta,\pi} V_{j+1}^{\theta,\pi})(\overline{\boldsymbol{\tau}}_j) \,|\, \sigma_{j-1}] \\
&= \mathbb{E}_{\theta,\pi}[(\mathbb{P}_j^{\theta,\pi} V_{j+1}^{\theta,\pi})(\overline{\boldsymbol{\tau}}_j) \,|\, \sigma_{j-1}] \\
&= \mathbb{E}_{\theta,\pi}[V_{j+1}^{\theta,\pi}(\overline{\boldsymbol{\tau}}_{j+1}) \,|\, \sigma_{j-1}],
\end{aligned} \tag{D.6}$$

where the second equality uses the tower property of the conditional expectation. Combining (D.6) with the induction assumption in (D.5), we have that (3.12) holds when $h = j$. Thus, by induction we have that (3.12) holds for any $h \in [H+1]$.

Therefore, we conclude the proof of Corollary 3.2. □

D.3. Proof of Lemma 3.3

We prove a more general version of Lemma 3.3. In detail, we replace the true parameter θ^* in Lemma 3.3 by any parameter $\theta \in \Theta$.

Lemma D.1 (General Version of Lemma 3.3)**.** *For any $(h, \theta, a, a', \pi) \in \{2, \ldots, H\} \times \Theta \times \mathcal{A}^2 \times \overline{\Pi}$, we have*

$$\mathbb{E}_{X \sim \rho_{h,a,a'}^{\theta,\pi}}[(\mathbb{F}_{h,a'}^{\theta} f - f)(X)] = 0, \quad \textit{for any } f \in L^{\infty}(\mathcal{O}^3).$$

Here, the distribution $\rho_{h,a,a'}^{\theta,\pi} \in \Delta(\mathcal{O}^3)$ is defined by

$$\rho_{h,a,a'}^{\theta,\pi}(o_{h-1}, o_h, o_{h+1}) = p_{\theta,\pi}\big(\boldsymbol{o}_{h-1} = o_{h-1}, \boldsymbol{o}_h = o_h, \boldsymbol{o}_{h+1} = o_{h+1} \,|\, \boldsymbol{a}_{h-1} = a, \boldsymbol{a}_h = a'\big),$$

for any o_{h-1}, o_h, $o_{h+1} \in \mathcal{O}$. Also, we have $\|\mathbb{F}_{h,a'}^{\theta}\|_{\infty\to\infty} \le \gamma$.

Proof. By the definition of $\mathbb{F}_{h,a'}^{\theta}$ in (3.13), we have

$$\begin{aligned}
&\mathbb{E}_{X \sim \rho_{h,a,a'}^{\theta,\pi}}[(\mathbb{F}_{h,a'}^{\theta} f)(X)] \\
&= \int_{\mathcal{O}^3} (\mathbb{F}_{h,a'}^{\theta} f)(o_{h-1}, o_h, o_{h+1}) \cdot \rho_{h,a,a'}^{\theta,\pi}(o_{h-1}, o_h, o_{h+1}) \,\mathrm{d}o_{h-1} \,\mathrm{d}o_h \,\mathrm{d}o_{h+1} \\
&= \int_{\mathcal{O}^5} f(o_{h-1}, \widetilde{o}_h, \widetilde{o}_{h+1}) \cdot \mathcal{B}_{h,a'}^{\theta}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \\
&\qquad \cdot \rho_{h,a,a'}^{\theta,\pi}(o_{h-1}, o_h, o_{h+1}) \,\mathrm{d}o_{h-1} \,\mathrm{d}o_h \,\mathrm{d}o_{h+1} \,\mathrm{d}\widetilde{o}_h \,\mathrm{d}\widetilde{o}_{h+1}.
\end{aligned} \tag{D.7}$$

Here, by the definition of $\rho_{h,a,a'}^{\theta,\pi}$ in Lemma 3.3, we have

$$\int_{\mathcal{O}} \rho_{h,a,a'}^{\theta,\pi}(o_{h-1}, o_h, o_{h+1}) \,\mathrm{d}o_{h+1} = p_{\theta,\pi}(\boldsymbol{o}_{h-1} = o_{h-1}, \boldsymbol{o}_h = o_h \,|\, \boldsymbol{a}_{h-1} = a). \tag{D.8}$$

Combining (D.7) and (D.8), we obtain

$$\begin{aligned}&\mathbb{E}_{X\sim\rho^{\theta,\pi}_{h,a,a'}}[(\mathbb{F}^{\theta}_{h,a'}f)(X)] \\ &\quad=\int_{\mathcal{O}^4} f(o_{h-1},\widetilde{o}_h,\widetilde{o}_{h+1})\cdot\mathcal{B}^{\theta}_{h,a'}(o_h,\widetilde{o}_h,\widetilde{o}_{h+1}) \\ &\qquad\qquad\cdot p_{\theta,\pi}(\boldsymbol{o}_{h-1}=o_{h-1},\boldsymbol{o}_h=o_h\,|\,\boldsymbol{a}_{h-1}=a)\,\mathrm{d}o_{h-1}\,\mathrm{d}o_h\,\mathrm{d}o_{h+1}\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1},\end{aligned} \tag{D.9}$$

where $\mathcal{B}^{\theta}_{h,a'}(o_h,\widetilde{o}_h,\widetilde{o}_{h+1})$ takes the form

$$\mathcal{B}^{\theta}_{h,a'}(o_h,\widetilde{o}_h,\widetilde{o}_{h+1})=\int_{\mathcal{S}} p_\theta(\boldsymbol{o}_h=\widetilde{o}_h,\boldsymbol{o}_{h+1}=\widetilde{o}_{h+1}\,|\,\boldsymbol{s}_h=\widetilde{s}_h,\boldsymbol{a}_h=a')\cdot\mathcal{Z}^{\theta}_h(\widetilde{s}_h,o_h)\,\mathrm{d}\widetilde{s}_h$$

following the definition in (3.6). By the Markov property of the POMDP, we can write

$$\begin{aligned}&p_{\theta,\pi}(\boldsymbol{o}_{h-1}=o_{h-1},\boldsymbol{o}_h=o_h\,|\,\boldsymbol{a}_{h-1}=a) \\ &\quad=\int_{\mathcal{S}^2}\mathcal{E}^{\theta}_h(o_h\,|\,s_h)\cdot\mathcal{T}^{\theta}_h(s_h\,|\,s_{h-1},a)\cdot p_{\theta,\pi}(\boldsymbol{s}_{h-1}=s_{h-1},\boldsymbol{o}_{h-1}=o_{h-1})\,\mathrm{d}s_h\,\mathrm{d}s_{h-1}\end{aligned} \tag{D.10}$$

By Assumptions 2.1 and 2.2, we have

$$\int_{\mathcal{S}\times\mathcal{O}}\mathcal{Z}^{\theta}_h(\widetilde{s}_h,o_h)\cdot\mathcal{E}^{\theta}_h(o_h\,|\,s_h)\cdot\mathcal{T}^{\theta}_h(s_h\,|\,s_{h-1},a)\,\mathrm{d}o_h\,\mathrm{d}s_h=\mathcal{T}^{\theta}_h(\widetilde{s}_h\,|\,s_{h-1},a). \tag{D.11}$$

Combining (D.10) and (D.11), we obtain

$$\begin{aligned}&\int_{\mathcal{O}}\mathcal{Z}^{\theta}_h(\widetilde{s}_h,o_h)\cdot p_{\theta,\pi}(\boldsymbol{o}_{h-1}=o_{h-1},\boldsymbol{o}_h=o_h\,|\,\boldsymbol{a}_{h-1}=a)\,\mathrm{d}o_h\,\mathrm{d}s_h \\ &\quad=\int_{\mathcal{S}}\mathcal{T}^{\theta}_h(\widetilde{s}_h\,|\,s_{h-1},a)\cdot p_{\theta,\pi}(\boldsymbol{s}_{h-1}=s_{h-1},\boldsymbol{o}_{h-1}=o_{h-1})\,\mathrm{d}s_h\,\mathrm{d}s_{h-1} \\ &\quad=p_{\theta,\pi}(\boldsymbol{o}_{h-1}=o_{h-1},\boldsymbol{s}_h=\widetilde{s}_h\,|\,\boldsymbol{a}_{h-1}=a),\end{aligned}$$

which implies

$$\begin{aligned}&\int_{\mathcal{O}}\mathcal{B}^{\theta}_{h,a'}(o_h,\widetilde{o}_h,\widetilde{o}_{h+1})\cdot p_{\theta,\pi}(\boldsymbol{o}_{h-1}=o_{h-1},\boldsymbol{o}_h=o_h\,|\,\boldsymbol{a}_{h-1}=a)\,\mathrm{d}o_h \\ &\quad=p_{\theta,\pi}(\boldsymbol{o}_{h-1}=o_{h-1},\boldsymbol{o}_h=\widetilde{o}_h,\boldsymbol{o}_{h+1}=\widetilde{o}_{h+1}\,|\,\boldsymbol{a}_{h-1}=a,\boldsymbol{a}_h=a') \\ &\quad=\rho^{\theta,\pi}_{h,a,a'}(o_{h-1},\widetilde{o}_h,\widetilde{o}_{h+1}).\end{aligned} \tag{D.12}$$

Then, combining (D.9) and (D.12), we have

$$\begin{aligned}&\mathbb{E}_{X\sim\rho^{\theta,\pi}_{h,a,a'}}[(\mathbb{F}^{\theta}_{h,a'}f)(X)] \\ &\quad=\int_{\mathcal{O}^3} f(o_{h-1},\widetilde{o}_h,\widetilde{o}_{h+1})\cdot\rho^{\theta,\pi}_{h,a,a'}(o_{h-1},\widetilde{o}_h,\widetilde{o}_{h+1})\,\mathrm{d}o_{h-1}\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1}=\mathbb{E}_{X\sim\rho^{\theta,\pi}_{h,a,a'}}[f(X)].\end{aligned}$$

In the sequel, we prove $\|\mathbb{F}^{\theta}_{h,a'}\|_{\infty\to\infty}\le\gamma$. It suffices to prove

$$|(\mathbb{F}^{\theta}_{h,a'}f)(o_{h-1},o_h,o_{h+1})|=\Big|\int_{\mathcal{O}^2} f(o_{h-1},\widetilde{o}_h,\widetilde{o}_{h+1})\cdot\mathcal{B}^{\theta}_{h,a}(o_h,\widetilde{o}_h,\widetilde{o}_{h+1})\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1}\Big|\le\gamma, \tag{D.13}$$

for any $f\in L^\infty(\mathcal{O}^3)$ such that $\|f\|_\infty\le 1$ and $o_{h-1},o_h,o_{h+1}\in\mathcal{O}$. By the definition of the function $\mathcal{B}^{\theta}_{h,a}$ in (3.6), we have

$$\begin{aligned}&\Big|\int_{\mathcal{O}^2} f(o_{h-1},\widetilde{o}_h,\widetilde{o}_{h+1})\cdot\mathcal{B}^{\theta}_{h,a}(o_h,\widetilde{o}_h,\widetilde{o}_{h+1})\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1}\Big| \\ &\quad\le\int_{\mathcal{O}^2}|\mathcal{B}^{\theta}_{h,a}(o_h,\widetilde{o}_h,\widetilde{o}_{h+1})|\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1} \\ &\quad\le\int_{\mathcal{S}\times\mathcal{O}^2} p_\theta(\widetilde{\boldsymbol{o}}_h=\widetilde{o}_h,\widetilde{\boldsymbol{o}}_{h+1}=\widetilde{o}_{h+1}\,|\,\widetilde{\boldsymbol{s}}_h=\widetilde{s}_h,\widetilde{\boldsymbol{a}}_h=a)\cdot|\mathcal{Z}^{\theta}_h(\widetilde{s}_h,o_h)|\,\mathrm{d}\widetilde{s}_h\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1} \\ &\quad=\int_{\mathcal{S}}|\mathcal{Z}^{\theta}_h(\widetilde{s}_h,o_h)|\,\mathrm{d}\widetilde{s}_h,\end{aligned} \tag{D.14}$$

for any $o_{h-1}, o_h \in \mathcal{O}$. Note that by Assumption 2.2, we have

$$\int_{\mathcal{S}} |\mathcal{Z}_h^\theta(\widetilde{s}_h, o_h)| \, \mathrm{d}\widetilde{s}_h = \|\mathbb{Z}_h^\theta \delta_{o_h}\|_1 \le \gamma \cdot \|\delta_{o_h}\|_1 = \gamma, \tag{D.15}$$

for any $(h, \theta, o_h) \in [H] \times \Theta \times \mathcal{O}$. Here, δ_{o_h} is the Dirac delta function defined on $\mathcal{O}$, whose value is zero everywhere except at o_h, and whose integral over $\mathcal{O}$ is equal to one. Combining (D.14) and (D.15), we have that (D.13) holds.

Therefore, we conclude the proof of Lemma D.1. □

E. Proof of Theorem 4.1

Proof. For any $\delta > 0$, by the definition of (θ_k, π_k) in (A.1) and the first statement in Lemma A.2, with probability at least $1 - \delta$, it holds that

$$J(\theta^*, \pi^*) - J(\theta^*, \pi_k) \le J(\theta_k, \pi_k) - J(\theta^*, \pi_k) \tag{E.1}$$

for all $k \in [K]$. By further applying Lemma A.1 to the right-hand side of (E.1) and using the definition of the error function e_h^k in (A.2), we obtain

$$\begin{aligned} J(\theta^*, \pi^*) - J(\theta^*, \pi_k) &\le \sum_{h=1}^H \mathbb{E}_{\theta^*, \pi_k}[(\mathbb{B}_h^{\theta_k, \pi_k} V_{h+1}^{\theta_k, \pi_k})(\overline{\boldsymbol{\tau}}_h) - (\mathbb{B}_h^{\theta^*, \pi_k} V_{h+1}^{\theta_k, \pi_k})(\overline{\boldsymbol{\tau}}_h)] \\ &= \sum_{h=1}^H \mathbb{E}_{\theta^*, \pi_k}\big[\mathbb{E}_{\theta^*, \pi_k}[(\mathbb{B}_h^{\theta_k, \pi_k} V_{h+1}^{\theta_k, \pi_k})(\overline{\boldsymbol{\tau}}_h) - (\mathbb{B}_h^{\theta^*, \pi_k} V_{h+1}^{\theta_k, \pi_k})(\overline{\boldsymbol{\tau}}_h)] \,\big|\, \boldsymbol{s}_{h-1}\big] \\ &\le \sum_{h=1}^H \mathbb{E}_{\theta^*, \pi_k}[e_h^k(\boldsymbol{s}_{h-1})], \end{aligned} \tag{E.2}$$

where the equality uses the tower property of the expectation. Telescoping both sides of (E.2) for $k \in [K]$ and applying Lemma A.3, we obtain

$$\begin{aligned} &\sum_{k=1}^K J(\theta^*, \pi^*) - J(\theta^*, \pi_k) \\ &\quad \le \sum_{h=1}^H \sum_{k=1}^K \mathbb{E}_{\theta^*, \pi_k}[e_h^k(\boldsymbol{s}_{h-1})] \\ &\quad \le H d_{\mathrm{s}} \Big(4\gamma H + 2 \log K \cdot \max_{k \in [K]} \big(k \cdot \mathbb{E}_{\theta, \overline{\pi}_k}[e_h^k(\boldsymbol{s}_{h-1})]\big)\Big) \end{aligned} \tag{E.3}$$

By applying the second statement of Lemma A.2 to the right-hand side of (E.3), we further obtain

$$\begin{aligned} &\sum_{k=1}^K J(\theta^*, \pi^*) - J(\theta^*, \pi_k) \\ &\quad \le H d_{\mathrm{s}} \Big(4\gamma H + 2 \log K \cdot \max_{k \in [K]} \big(k \cdot 2HA^2\gamma^2\beta \cdot k^{-1/2}\big)\Big) \\ &\quad \le H d_{\mathrm{s}} \big(4\gamma H + 2 \log K \cdot 2HA^2\gamma^2\beta \cdot K^{1/2}\big), \end{aligned}$$

which concludes the proof of Theorem 4.1. □

F. Proofs for Section A

In this section, we present the proofs for the results in Section A.

F.1. Proof of Lemma A.1

Proof. By Corollary 3.2 and the definition of the value function in (3.10), we can write

$$J(\theta,\pi)=\mathbb{E}_{\theta',\pi}[V_1^{\theta,\pi}(\overline{\boldsymbol{\tau}}_1)],\quad J(\theta',\pi)=\mathbb{E}_{\theta',\pi}[V_{H+1}^{\theta,\pi}(\overline{\boldsymbol{\tau}}_{H+1})],$$

which implies

$$J(\theta,\pi)-J(\theta',\pi)=\sum_{h=1}^{H}\big(\mathbb{E}_{\theta',\pi}[V_h^{\theta,\pi}(\overline{\boldsymbol{\tau}}_h)]-\mathbb{E}_{\theta',\pi}[V_{h+1}^{\theta,\pi}(\overline{\boldsymbol{\tau}}_{h+1})]\big). \tag{F.1}$$

By the definition of $V_h^{\theta,\pi}$ in (3.10), we have

$$\mathbb{E}_{\theta',\pi}[V_h^{\theta,\pi}(\overline{\boldsymbol{\tau}}_h)]=\mathbb{E}_{\theta',\pi}[(\mathbb{B}_h^{\theta,\pi}V_{h+1}^{\theta,\pi})(\overline{\boldsymbol{\tau}}_h)]. \tag{F.2}$$

Also, by Lemma 3.1, we have

$$\mathbb{E}_{\theta',\pi}[V_{h+1}^{\theta,\pi}(\overline{\boldsymbol{\tau}}_{h+1})]=\mathbb{E}_{\theta',\pi}[(\mathbb{P}_h^{\theta',\pi}V_{h+1}^{\theta,\pi})(\overline{\boldsymbol{\tau}}_h)]=\mathbb{E}_{\theta',\pi}[(\mathbb{B}_h^{\theta',\pi}V_{h+1}^{\theta,\pi})(\overline{\boldsymbol{\tau}}_h)]. \tag{F.3}$$

Plugging (F.2) and (F.3) into the right-hand side of (F.1), we obtain

$$J(\theta,\pi)-J(\theta',\pi)=\sum_{h=1}^{H}\mathbb{E}_{\theta',\pi}[(\mathbb{B}_h^{\theta,\pi}V_{h+1}^{\theta,\pi})(\overline{\boldsymbol{\tau}}_h)-(\mathbb{B}_h^{\theta',\pi}V_{h+1}^{\theta,\pi})(\overline{\boldsymbol{\tau}}_h)],$$

which concludes the proof of Lemma A.1. □

F.2. Proof of Lemma A.2

Before proving Lemma A.2, we present several auxiliary lemmas for the proof of Lemma A.2. Recall that we define the projection operator $\mathbb{S}$ in (3.18). The following lemma verifies the projection property of $\mathbb{S}$ as mentioned in (3.15).

Lemma F.1. *For any $f\in L^\infty(\mathcal{O}^3)$ and $\rho\in\Delta(\mathcal{O}^3)$, we have*

$$\mathbb{E}_{X\sim\rho}[(\mathbb{S}f)(X)]=\int_{\mathcal{O}^3}f(x)\cdot\widehat{\rho}(x)\,\mathrm{d}x,$$

where $\widehat{\rho}$ is the projection of ρ onto $\mathrm{linspan}(\{\phi_i\}_{i=1}^{d_\mathrm{o}})$ with respect to the distance defined in (3.19) and takes the form

$$\widehat{\rho}(o_{h-1},o_h,o_{h+1})=\sum_{j\in[d_\mathrm{o}]}\phi_j(o_{h-1},o_h,o_{h+1})\cdot\sum_{i\in[d_\mathrm{o}]}[G^{-1}]_{i,j}\cdot\mathbb{E}_{X\sim\phi_i,Y\sim\rho}[\mathcal{K}(X,Y)], \tag{F.4}$$

for any $o_{h-1},o_h,o_{h+1}\in\mathcal{O}^3$.

Proof. See Section G.1 for a detailed proof. □

Recall that $\{\mathcal{D}_{h,a,a'}\}_{(h,a,a')\in\{2,\ldots,H\}\times\mathcal{A}^2}$ is the dataset in Algorithm 1, which is updated in each iteration. For any $k\in[K]$, we denote by $\mathcal{D}_{h,a,a'}^k$ the status of $\mathcal{D}_{h,a,a'}$ after the exploration phase of the k-th iteration of Algorithm 1. We denote by $\widehat{\mathcal{D}}_{h,a,a'}^k$ the empirical distribution induced by the dataset $\mathcal{D}_{h,a,a'}^k$. For any $(k,h,a,a')\in[K]\times\{2,\ldots,H\}\times\mathcal{A}^2$, as a special case of Lemma F.1 for $\rho=\widehat{\mathcal{D}}_{h,a,a'}^k$, we define the function $\widehat{\rho}_{h,a,a'}^k:\mathcal{O}^3\to\mathbb{R}$ by

$$\widehat{\rho}_{h,a,a'}^k(o_{h-1},o_h,o_{h+1})=\sum_{j\in[d_\mathrm{o}]}\phi_j(o_{h-1},o_h,o_{h+1})\cdot[\widehat{w}_{h,a,a'}^k]_j,$$

where the vector $\widehat{w}_{h,a,a'}^k\in\mathbb{R}^{d_\mathrm{o}}$ is defined by

$$[\widehat{w}_{h,a,a'}^k]_j=\sum_{i\in[d_\mathrm{o}]}[G^{-1}]_{i,j}\cdot\mathbb{E}_{X\sim\phi_i,Y\sim\widehat{\mathcal{D}}_{h,a,a'}^k}[\mathcal{K}(X,Y)], \tag{F.5}$$

for any $j\in[d_\mathrm{o}]$. Here, the matrix $G\in\mathbb{R}^{d_\mathrm{o}\times d_\mathrm{o}}$ is defined in (3.17). The following lemma shows that, with high probability, $\widehat{\rho}_{h,a,a'}^k$ converges to $\rho_{h,a,a'}^{\overline{\pi}_k}$ as k goes to infinity. The convergence is with respect to the L^1-norm in $\mathcal{O}^3$, which guarantees the generalization power of the solution to the minimax problem in (3.14).

Lemma F.2. *For any fixed $\delta > 0$, we define the event $\mathcal{G}$ as*

$$\|\widehat{\rho}^k_{h,a,a'} - \rho^{\overline{\pi}_k}_{h,a,a'}\|_1 \le d_{\mathrm{o}}^{3/2}/\alpha \cdot \sqrt{8\log(2KHA^2/\delta)} \cdot k^{-1/2},$$

for any $(k, h, a, a') \in [K] \times \{2, \ldots, H\} \times \mathcal{A}^2$. Then, it holds that $\mathcal{G}$ happens with probability at least $1 - \delta$.

Proof. See Section G.2 for a detailed proof. □

Moreover, for ease of presentation, we define the operator $\mathbb{V}^\theta_{h,a} : L^1(\mathcal{O}^3) \to L^1(\mathcal{O}^3)$ for any $(h, a, \theta) \in \{2, \ldots, H\} \times \mathcal{A} \times \Theta$ by

$$(\mathbb{V}^\theta_h f)(o_{h-1}, \widetilde{o}_h, \widetilde{o}_{h+1}) = \int_{\mathcal{O}^2} \mathcal{B}^\theta_{h,a}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \cdot f(o_{h-1}, o_h, o_{h+1}) \,\mathrm{d}o_h \,\mathrm{d}o_{h+1}, \tag{F.6}$$

for any $f \in L^1(\mathcal{O}^3)$ and $o_{h-1}, \widetilde{o}_h, \widetilde{o}_{h+1} \in \mathcal{O}$, which is the conjugate (i.e., transpose) of the operator $\mathbb{F}^\theta_{h,a}$ defined in (3.13). Recall that we define $\rho^{\theta,\pi}_{h,a,a'}$ in Lemma D.1, which is the general form of $\rho^\pi_{h,a,a'}$ for any $\theta \in \Theta$. The following lemma mirrors Lemma D.1.

Lemma F.3. *For any $(h, \theta, a, a', \pi) \in \{2, \ldots, H\} \times \Theta \times \mathcal{A}^2 \times \overline{\Pi}$, we have*

$$\|\mathbb{V}^\theta_{h,a'} \rho^{\theta,\pi}_{h,a,a'} - \rho^{\theta,\pi}_{h,a,a'}\|_1 = 0.$$

Also, we have $\|\mathbb{V}^\theta_{h,a'}\|_{1\to 1} \le \gamma$.

Proof. See Section G.3 for a detailed proof. □

Proof of Lemma A.2:

Proof. In the following, we condition on the event $\mathcal{G}$ defined in Lemma F.2, which happens with probability at least $1 - \delta$.

Proof of the first statement: Recall that we denote by $\mathcal{D}^k_{h,a,a'}$ the status of $\mathcal{D}_{h,a,a'}$ after the exploration phase of the k-th iteration of Algorithm 1. Correspondingly, we define the function $L^k : \Theta \to \mathbb{R}$ by

$$L^k(\theta) = \max_{f \in L^\infty(\mathcal{O}^3): \|f\|_\infty \le 1} \; \max_{(h,a,a') \in \{2,\ldots,H\} \times \mathcal{A}^2} \mathbb{E}_{X \sim \widehat{\mathcal{D}}^k_{h,a,a'}} [(\mathbb{S}\mathbb{F}^\theta_{h,a'} f - \mathbb{S} f)(X)],$$

for any $\theta \in \Theta$, which corresponds to the function L in (3.23) in the planning phase of the k-th iteration of Algorithm 1. To prove $\theta^* \in \Theta^k$, it suffices to prove that

$$L^k(\theta^*) \le \beta \cdot k^{-1/2}. \tag{F.7}$$

For notational simplicity, we write $\rho^k_{h,a,a'}$ in short for $\rho^{\overline{\pi}_k}_{h,a,a'}$. For any $f \in L^\infty(\mathcal{O}^3)$ such that $\|f\|_\infty \le 1$ and $(k, h, a, a') \in [K] \times \{2, \ldots, H\} \times \mathcal{A}^2$, we have

$$\begin{aligned}
&\mathbb{E}_{X \sim \widehat{\mathcal{D}}^k_{h,a,a'}} [(\mathbb{S}\mathbb{F}^{\theta^*}_{h,a'} f - \mathbb{S} f)(X)] \\
&\quad = \mathbb{E}_{X \sim \widehat{\mathcal{D}}^k_{h,a,a'}} [(\mathbb{S}\mathbb{F}^{\theta^*}_{h,a'} f - \mathbb{S} f)(X)] - \mathbb{E}_{X \sim \rho^k_{h,a,a'}} [(\mathbb{F}^{\theta^*}_{h,a'} f - f)(X)] \\
&\quad = \Big(\int_{\mathcal{O}^3} (\mathbb{F}^{\theta^*}_{h,a'} f)(x) \cdot \widehat{\rho}^k_{h,a,a'}(x) \,\mathrm{d}x - \mathbb{E}_{X \sim \rho^k_{h,a,a'}} [(\mathbb{F}^{\theta^*}_{h,a'} f)(X)] \Big) \\
&\qquad + \Big(\int_{\mathcal{O}^3} f(x) \cdot \widehat{\rho}^k_{h,a,a'}(x) \,\mathrm{d}x - \mathbb{E}_{X \sim \rho^k_{h,a,a'}} [f(X)] \Big),
\end{aligned} \tag{F.8}$$

where the first equality uses

$$\mathbb{E}_{X \sim \rho^k_{h,a,a'}} [(\mathbb{F}^{\theta^*}_{h,a'} f - f)(X)] = 0,$$

following Lemma 3.3 and the second equality is by Lemma F.1. Recall that $\|f\|_\infty \le 1$. By Holder's inequality and the definition of the event $\mathcal{G}$ in Lemma F.2, we have

$$\begin{aligned}&\int_{\mathcal{O}^3} f(x)\cdot \widehat{\rho}^k_{h,a,a'}(x)\,\mathrm{d}x - \mathbb{E}_{X\sim\rho^k_{h,a,a'}}[f(X)] \\ &\quad\le \|f\|_\infty\cdot\|\widehat{\rho}^k_{h,a,a'}-\rho^k_{h,a,a'}\|_1 \le d_{\mathrm{o}}^{3/2}/\alpha\cdot\sqrt{8\log(2KHA^2/\delta)}\cdot k^{-1/2}.\end{aligned} \tag{F.9}$$

Similarly, we have

$$\begin{aligned}&\int_{\mathcal{O}^3} (\mathbb{F}^{\theta^*}_{h,a'}f)(x)\cdot \widehat{\rho}^k_{h,a,a'}(x)\,\mathrm{d}x - \mathbb{E}_{X\sim\rho^k_{h,a,a'}}[(\mathbb{F}^{\theta^*}_{h,a'}f)(X)] \\ &\quad\le \|\mathbb{F}^{\theta^*}_{h,a'}f\|_\infty\cdot\|\widehat{\rho}^k_{h,a,a'}-\rho^k_{h,a,a'}\|_1 \\ &\quad\le d_{\mathrm{o}}^{3/2}\gamma/\alpha\cdot\sqrt{8\log(2KHA^2/\delta)}\cdot k^{-1/2},\end{aligned} \tag{F.10}$$

where the second inequality uses the fact $\|\mathbb{F}^{\theta^*}_{h,a'}\|_{\infty\to\infty}\le\gamma$ from Lemma 3.3. Then, by combining (F.8), (F.9), and (F.10) with the condition of β in (4.1), we have that the inequality in (F.7) holds for all $k\in[K]$.

Proof of the second statement: Invoking Lemma G.5, we have

$$\mathbb{E}_{\theta^*,\overline{\pi}_k}[e^k_h(\boldsymbol{s}_h)] \le \gamma^2 H\cdot\sum_{a,a'\in\mathcal{A}}\big\|\mathbb{V}^{\theta_k}_{h,a'}\rho^{\overline{\pi}_k}_{h,a,a'}-\rho^{\overline{\pi}_k}_{h,a,a'}\big\|_1, \tag{F.11}$$

for any $(k,h)\in[K]\times\{2,\ldots,H\}$. By the triangle inequality, we can write

$$\begin{aligned}&\big\|\mathbb{V}^{\theta_k}_{h,a'}\rho^{\overline{\pi}_k}_{h,a,a'}-\rho^{\overline{\pi}_k}_{h,a,a'}\big\|_1 \\ &\quad\le \big\|\mathbb{V}^{\theta_k}_{h,a'}\rho^{\overline{\pi}_k}_{h,a,a'}-\mathbb{V}^{\theta_k}_{h,a'}\rho^{k}_{h,a,a'}\big\|_1+\big\|\mathbb{V}^{\theta_k}_{h,a'}\rho^{k}_{h,a,a'}-\rho^{k}_{h,a,a'}\big\|_1+\big\|\rho^{k}_{h,a,a'}-\rho^{\overline{\pi}_k}_{h,a,a'}\big\|_1.\end{aligned}$$

By the definition of Θ_k in (3.22) and the fact $\theta_k\in\Theta_k$, we have

$$\big\|\mathbb{V}^{\theta_k}_{h,a'}\rho^{k}_{h,a,a'}-\rho^{k}_{h,a,a'}\big\|_1\le\beta\cdot k^{-1/2}$$

By the definition of the event $\mathcal{G}$ in Lemma F.2, we have

$$\big\|\rho^{k}_{h,a,a'}-\rho^{\overline{\pi}_k}_{h,a,a'}\big\|_1\le d_{\mathrm{o}}^{3/2}/\alpha\cdot\sqrt{8\log(2KHA^2/\delta)}\cdot k^{-1/2}.$$

Similarly, we have

$$\begin{aligned}\big\|\mathbb{V}^{\theta_k}_{h,a'}\rho^{\overline{\pi}_k}_{h,a,a'}-\mathbb{V}^{\theta_k}_{h,a'}\rho^{k}_{h,a,a'}\big\|_1 &\le \|\mathbb{V}^{\theta_k}_{h,a'}\|_{1\to1}\cdot\big\|\rho^{\overline{\pi}_k}_{h,a,a'}-\rho^{k}_{h,a,a'}\big\|_1 \\ &\le d_{\mathrm{o}}^{3/2}\gamma/\alpha\cdot\sqrt{8\log(2KHA^2/\delta)}\cdot k^{-1/2},\end{aligned} \tag{F.12}$$

where the second inequality uses the fact $\|\mathbb{V}^{\theta^*}_{h,a'}\|_{1\to1}\le\gamma$ from Lemma F.3. Combing (F.11)-(F.12) with the condition of β in (4.1), we obtain

$$\mathbb{E}_{\theta,\overline{\pi}_k}[e^k_h(\boldsymbol{s}_h)]\le 2\gamma^2\beta HA^2\cdot k^{-1/2},$$

for any $(k,h)\in[K]\times\{2,\ldots,H\}$.

Therefore, we conclude the proof of Lemma A.2. □

F.3. Proof of Lemma A.3

Proof. For any $h\in[H]$ and $\pi\in\overline{\Pi}$, we denote by μ^π_h the marginal distribution of $\boldsymbol{s}_h$ with respect to the policy π and the true parameter θ^*. By Assumption 2.1, we have

$$\mu^{\pi_k}_h\in\mathrm{conh}(\psi),\quad \mu^{\overline{\pi}_k}_h=\frac{1}{k}\sum_{i=0}^{k-1}\mu^{\pi_i}_h\in\mathrm{conh}(\psi),$$

for any $(k, h) \in \{0, \ldots, K\} \times [H]$. Here, π_0 is the initial policy and π_k with $k \in [K]$ is the trained policy in the k-th iteration of Algorithm 1. Thus, there exist vector $c_h^k, \overline{c}_h^k \in \Delta([d_{\mathrm{s}}]) \subset \mathbb{R}^{d_{\mathrm{s}}}$ such that

$$\mu_h^{\pi_k}(\cdot) = \psi(\cdot)^\top c_h^k, \quad \mu_h^{\overline{\pi}_k}(\cdot) = \psi(\cdot)^\top \overline{c}_h^k, \quad \overline{c}_h^k = (1/k) \cdot \sum_{i=0}^{k-1} c_h^i.$$

Also, we define the vector $b_h^k \in \mathbb{R}^{d_{\mathrm{s}}}$ by

$$[b_h^k]_i = \mathbb{E}_{\boldsymbol{s}_h \sim \psi_i}[e_{h+1}^k(\boldsymbol{s}_h)], \quad \text{for any } i \in [d_{\mathrm{s}}]. \tag{F.13}$$

Then, it holds that

$$\mathbb{E}_{\theta^*, \pi_k}[e_{h+1}^k(\boldsymbol{s}_h)] = (b_h^k)^\top c_h^k, \quad \mathbb{E}_{\theta^*, \overline{\pi}_k}[e_{h+1}^k(\boldsymbol{s}_h)] = (b_h^k)^\top \overline{c}_h^k. \tag{F.14}$$

For any $i \in [d_{\mathrm{s}}]$, we define $\underline{k}_i$ by

$$\underline{k}_i = \min\Big\{k \in [K] : \sum_{j=1}^{k} [c_h^j]_i \geq 1 \text{ or } k = K\Big\}. \tag{F.15}$$

Then, we can write

$$\begin{aligned} \sum_{k=1}^{K} \mathbb{E}_{\theta^*, \pi_k}[e_{h+1}^k(\boldsymbol{s}_h)] &= \sum_{i=1}^{d_{\mathrm{s}}} \sum_{k=1}^{K} [b_h^k]_i \cdot [c_h^k]_i \\ &= \sum_{i=1}^{d_{\mathrm{s}}} \Big(\sum_{k=1}^{\underline{k}_i} [b_h^k]_i \cdot [c_h^k]_i + \sum_{k=\underline{k}_i+1}^{K} [b_h^k]_i \cdot [c_h^k]_i \Big) \end{aligned} \tag{F.16}$$

The first summation term on the right-hand side of (F.16) can be upper bounded as

$$\sum_{i=1}^{d_{\mathrm{s}}} \sum_{k=1}^{\underline{k}_i} [b_h^k]_i \cdot [c_h^k]_i \leq 2\gamma H \cdot \sum_{i=1}^{d_{\mathrm{s}}} \sum_{k=1}^{\underline{k}_i} [c_h^k]_i \leq 4 d_{\mathrm{s}} \gamma H. \tag{F.17}$$

Here, the first inequality uses Lemma G.6, which provides an upper bound $2\gamma H$ for each $[b_h^k]_i$. Recall that $[b_h^k]_i$ is defined in (F.13). Also, the second inequality uses the fact

$$\sum_{k=1}^{\underline{k}_i} [c_h^k]_i = [c_h^{\underline{k}_i}]_i + \sum_{k=1}^{\underline{k}_i - 1} [c_h^k]_i \leq 1 + 1 \leq 2,$$

which is by the definition of $\underline{k}_i$ in (F.15) and the fact $[c_h^{\underline{k}_i}]_i \leq 1$ since $c_h^{\underline{k}_i} \in \Delta([d_{\mathrm{s}}])$. In the sequel, we characterize the second summation term on the right-hand side of (F.16). For any $i \in [d_{\mathrm{s}}]$ and $k \geq \underline{k}_i + 1$, we have

$$\begin{aligned} [b_h^k]_i \cdot [c_h^k]_i &\leq \mathbb{E}_{\theta^*, \overline{\pi}_k}[e_{h+1}^k(\boldsymbol{s}_h)] \cdot \frac{[b_h^k]_i \cdot [c_h^k]_i}{[b_h^k]_i \cdot [\overline{c}_h^k]_i} \\ &= \big(k \cdot \mathbb{E}_{\theta^*, \overline{\pi}_k}[e_{h+1}^k(\boldsymbol{s}_h)]\big) \cdot \frac{[c_h^k]_i}{\sum_{j=0}^{k-1} [c_h^j]_i} \\ &\leq \big(\max_{j \in [K]} \ell \cdot \mathbb{E}_{\theta^*, \overline{\pi}_\ell}[e_{h+1}^\ell(\boldsymbol{s}_h)]\big) \cdot \frac{[c_h^k]_i}{\sum_{j=1}^{k-1} [c_h^j]_i}, \end{aligned} \tag{F.18}$$

where the first inequality is by (F.14) and the second inequality uses the fact $[c_h^j]_0 \geq 0$. Note that for any (i, k) specified above, we have

$$\frac{[c_h^k]_i}{\sum_{j=0}^{k-1} [c_h^j]_i} \in [0, 1]$$

since it holds that $[c_h^k]_i \in [0,1]$ and $\sum_{j=0}^{k-1}[c_h^j]_i \geq 1$ by the definition of $\underline{k}_i$. Then, by applying the inequality $x \leq 2\log(1+x)$ for any $x \in [0,1]$, we have

$$\frac{[c_h^k]_i}{\sum_{j=1}^{k-1}[c_h^j]_i} \leq 2\log\Big(1+\frac{[c_h^k]_i}{\sum_{j=1}^{k-1}[c_h^j]_i}\Big) = 2\log\sum_{j=1}^{k}[c_h^j]_i - 2\log\sum_{j=1}^{k-1}[c_h^j]_i \tag{F.19}$$

Combining (F.18) and (F.19), we obtain

$$\begin{aligned}
\sum_{i=1}^{d_{\mathrm{s}}}\sum_{k=\underline{k}_i+1}^{K}[b_h^k]_i\cdot[c_h^k]_i &\leq \big(\max_{\ell\in[K]}\ell\cdot\mathbb{E}_{\theta^*,\overline{\pi}_\ell}[e_{h+1}^\ell(\boldsymbol{s}_h)]\big)\cdot\sum_{i=1}^{d_{\mathrm{s}}}\sum_{k=\underline{k}_i+1}^{K}\big(2\log\sum_{j=1}^{k}[c_h^j]_i - 2\log\sum_{j=1}^{k-1}[c_h^j]_i\big)\\
&= 2\big(\max_{\ell\in[K]}\ell\cdot\mathbb{E}_{\theta^*,\overline{\pi}_\ell}[e_{h+1}^\ell(\boldsymbol{s}_h)]\big)\cdot\sum_{i=1}^{d_{\mathrm{s}}}\big(\log\sum_{j=1}^{K}[c_h^j]_i - \log\sum_{j=1}^{\underline{k}_i}[c_h^j]_i\big)\\
&\leq 2\big(\max_{\ell\in[K]}\ell\cdot\mathbb{E}_{\theta^*,\overline{\pi}_\ell}[e_{h+1}^\ell(\boldsymbol{s}_h)]\big)\cdot d_{\mathrm{s}}\cdot\log K.
\end{aligned} \tag{F.20}$$

Plugging (F.17) and (F.20) into the right-hand side of (F.16), we obtain

$$\sum_{k=1}^{K}\mathbb{E}_{\theta^*,\pi_k}[e_{h+1}^k(\boldsymbol{s}_h)] \leq 4d_{\mathrm{s}}\gamma H + 2\big(\max_{\ell\in[K]}\ell\cdot\mathbb{E}_{\theta^*,\overline{\pi}_\ell}[e_{h+1}^\ell(\boldsymbol{s}_h)]\big)\cdot d_{\mathrm{s}}\cdot\log K,$$

which concludes the proof of Lemma A.3. □

G. Auxiliary Lemmas

In this section, we present (the proofs for) the auxiliary lemmas invoked in previous sections.

G.1. Proof of Lemma F.1

Proof. To see that $\widehat{\rho}$ defined in (F.4) is the projection, we consider the minimization problem

$$\min_{\rho'\in\mathrm{linspan}(\{\phi_i\}_{i=1}^{d_{\mathrm{o}}})}\|\mathbb{K}\rho'-\mathbb{K}\rho\|_{\mathcal{H}}^2 \tag{G.1}$$

for any $\rho \in \Delta(\mathcal{O}^3)$. Note that the objective can be written as

$$\begin{aligned}
&\|\mathbb{K}\rho'-\mathbb{K}\rho\|_{\mathcal{H}}^2\\
&\quad=\langle\mathbb{K}\rho',\mathbb{K}\rho'\rangle_{\mathcal{H}} - 2\cdot\langle\mathbb{K}\rho',\mathbb{K}\rho\rangle_{\mathcal{H}} + \langle\mathbb{K}\rho,\mathbb{K}\rho'\rangle_{\mathcal{H}}\\
&\quad=\mathbb{E}_{X\sim\rho',Y\sim\rho'}[\mathcal{K}(X,Y)] - 2\cdot\mathbb{E}_{X\sim\rho',Y\sim\rho}[\mathcal{K}(X,Y)] + \mathbb{E}_{X\sim\rho,Y\sim\rho}[\mathcal{K}(X,Y)].
\end{aligned} \tag{G.2}$$

Since we have $\rho' \in \mathrm{linspan}(\{\phi_i\}_{i=1}^{d_{\mathrm{o}}})$, there exists $w = (w_1,\ldots,w_{d_{\mathrm{o}}})^\top \in \mathbb{R}^{d_{\mathrm{o}}}$ such that

$$\rho'(x) = \sum_{j\in[d_{\mathrm{o}}]} w_j\cdot\phi_j(x), \quad \text{for any } x\in\mathcal{O}^3.$$

By the above form of ρ' and the definition of the matrix G in (3.17), we can further rewrite the right-hand side of (G.2) to obtain

$$\|\mathbb{K}\rho'-\mathbb{K}\rho\|_{\mathcal{H}}^2 = w^\top G w - 2\cdot\sum_{i=1}^{d_{\mathrm{o}}} w_i\cdot\mathbb{E}_{X\sim\phi_i,Y\sim\rho}[\mathcal{K}(X,Y)] + \mathbb{E}_{X\sim\rho,Y\sim\rho}[\mathcal{K}(X,Y)] \tag{G.3}$$

Plugging (G.3) into (G.1) and solving the obtained quadratic programming problem, we see that $\widehat{\rho}$ defined in (F.4) is the projection of ρ.

By the definition of $\mathbb{S}$ in (3.18), we have

$$\mathbb{E}_{X\sim\rho}[(\mathbb{S}f)(X)] = \sum_{i,j\in[d_o]} [G^{-1}]_{i,j}\cdot \mathbb{E}_{X\sim\rho,Y\sim\phi_i,Y'\sim\phi_j}\big[\mathcal{K}(X,Y)\cdot f(Y')\big], \tag{G.4}$$

where X, Y, Y' are independent variables in $\mathcal{O}^3$. By reorganizing terms in the summation, we can further write the right-hand side of (G.4) as

$$\begin{aligned}
&\sum_{i,j\in[d_o]} [G^{-1}]_{i,j}\cdot \mathbb{E}_{X\sim p,Y\sim\phi_i,Y'\sim\phi_j}\big[\mathcal{K}(X,Y)\cdot f(Y')\big] \\
&\quad= \sum_{j\in[d_o]} \mathbb{E}_{Y'\sim\phi_j}[f(Y')]\cdot \sum_{i\in[d_o]} [G^{-1}]_{i,j}\cdot \mathbb{E}_{X\sim p,Y\sim\phi_i}[\mathcal{K}(X,Y)] \\
&\quad= \int_{\mathcal{O}} f(x)\cdot \sum_{j\in[d_o]} \phi_j(x)\cdot \sum_{i\in[d_o]} [G^{-1}]_{i,j}\cdot \mathbb{E}_{X\sim p,Y\sim\phi_i}[\mathcal{K}(X,Y)]\,\mathrm{d}x,
\end{aligned}$$

combining which with the definition of $\widehat{\rho}$ in (F.4), we conclude the proof of Lemma F.1. □

G.2. Proof of Lemma F.2

Proof. For any $(k,h,a,a')\in[K]\times[2,\ldots,H]\times\mathcal{A}^2$, let $w^k_{h,a,a'}\in\mathbb{R}^{d_o}$ be the vector such that

$$\rho^{\overline{\pi}_k}_{h,a,a'}(o_{h-1},o_h,o_{h+1}) = \sum_{j\in[d_o]} \phi_i(o_{h-1},o_h,o_{h+1})\cdot [w^k_{h,a,a'}]_j, \tag{G.5}$$

for any $o_{h-1},o_h,o_{h+1}\in\mathcal{O}$, where $\{\phi_i\}_{i=1}^{d_o}$ are distribution functions defined in Assumption 2.1 and the existence of $w^k_{h,a,a'}$ is guaranteed by the assumption therein. Also, recall that we define $\widehat{w}^k_{h,a,a'}\in\mathbb{R}^{d_o}$ in (F.5) and we have

$$\widehat{\rho}^k_{h,a,a'}(o_{h-1},o_h,o_{h+1}) = \sum_{j\in[d_o]} \phi_j(o_{h-1},o_h,o_{h+1})\cdot [\widehat{w}^k_{h,a,a'}]_j. \tag{G.6}$$

For notational simplicity, we denote $\phi=(\phi_1,\ldots,\phi_{d_o})$ and write $\rho^k_{h,a,a'}$ in short for $\rho^{\overline{\pi}_k}_{h,a,a'}$. Then, we can rewrite (G.5) and (G.6) as

$$\rho^k_{h,a,a'}(\cdot) = \phi(\cdot)^\top w^k_{h,a,a'},\quad \widehat{\rho}^k_{h,a,a'}(\cdot) = \phi(\cdot)^\top \widehat{w}^k_{h,a,a'}.$$

Following the above definitions, we have

$$\begin{aligned}
&\|\widehat{\rho}^k_{h,a,a'} - \rho^k_{h,a,a'}\|_1 \\
&\quad= \int_{\mathcal{O}^3} |\phi(o_{h-1},o_h,o_{h+1})^\top(\widehat{w}^k_{h,a,a'} - w^k_{h,a,a'})|\,\mathrm{d}o_{h-1}\,\mathrm{d}o_h\,\mathrm{d}o_{h+1} \\
&\quad\le \int_{\mathcal{O}^3} \|\phi(o_{h-1},o_h,o_{h+1})\|_2\cdot\|\widehat{w}^k_{h,a,a'} - w^k_{h,a,a'}\|_2\,\mathrm{d}o_{h-1}\,\mathrm{d}o_h\,\mathrm{d}o_{h+1} \\
&\quad\le d_o\cdot\|\widehat{w}^k_{h,a,a'} - w^k_{h,a,a'}\|_2,
\end{aligned} \tag{G.7}$$

where the first inequality is by the Cauchy-Schwarz inequality and the last inequality uses

$$\begin{aligned}
&\int_{\mathcal{O}^3} \|\phi(o_{h-1},o_h,o_{h+1})\|_2\,\mathrm{d}o_{h-1}\,\mathrm{d}o_h\,\mathrm{d}o_{h+1} \\
&\quad\le \int_{\mathcal{O}^3} \|\phi(o_{h-1},o_h,o_{h+1})\|_1\,\mathrm{d}o_{h-1}\,\mathrm{d}o_h\,\mathrm{d}o_{h+1} \\
&\quad= \sum_{j=1}^{d_o}\int_{\mathcal{O}^3} \phi_j(o_{h-1},o_h,o_{h+1})\,\mathrm{d}o_{h-1}\,\mathrm{d}o_h\,\mathrm{d}o_{h+1} = d_o,
\end{aligned}$$

as $\{\phi_i\}_{i=1}^{d_o}$ are distribution functions over $\mathcal{O}^3$. Thus, to upper bound $\|\widehat{\rho}^k_{h,a,a'} - \rho^k_{h,a,a'}\|_1$, it suffices to upper bound $\|\widehat{w}^k_{h,a,a'} - w^k_{h,a,a'}\|_2$.

To this end, we define the vector $U^k_{h,a,a'} \in \mathbb{R}^{d_o}$ by

$$[U^k_{h,a,a'}]_i = \mathbb{E}_{X\sim\phi_i, Y\sim\widehat{\mathcal{D}}^k_{h,a,a'}}[\mathcal{K}(X,Y)], \tag{G.8}$$

where $\widehat{\mathcal{D}}^k_{h,a,a'}$ is the empirical distribution over $\mathcal{O}^3$ induced by the dataset $\mathcal{D}^k_{h,a,a'}$. Then, we can rewrite the definition of $\widehat{w}^k_{h,a,a'}$ in (F.5) as

$$\widehat{w}^k_{h,a,a'} = G^{-1}U^k_{h,a,a'},$$

where the matrix $G \in \mathbb{R}^{d_o\times d_o}$ is defined in (3.17). Then, we can write

$$\|\widehat{w}^k_{h,a,a'} - w^k_{h,a,a'}\|_2 = \|G^{-1}U^k_{h,a,a'} - G^{-1}Gw^k_{h,a,a'}\|_2 \le 1/\alpha \cdot \|U^k_{h,a,a'} - Gw^k_{h,a,a'}\|_2. \tag{G.9}$$

Moreover, by the definition of G in (3.17), we have

$$\begin{aligned}
[Gw^k_{h,a,a'}]_i &= \sum_{j\in[d_o]} [G]_{i,j}\cdot[w^k_{h,a,a'}]_j \\
&= \sum_{j\in[d_o]} \mathbb{E}_{X\sim\phi_i, Y\sim\phi_j}[\mathcal{K}(X,Y)]\cdot[w^k_{h,a,a'}]_j \\
&= \int_{\mathcal{O}^3\times\mathcal{O}^3} \phi_i(x)\cdot\mathcal{K}(x,y)\cdot\sum_{j\in[d_o]}\phi_j(y)\cdot[w^k_{h,a,a'}]_j\,\mathrm{d}x\,\mathrm{d}y,
\end{aligned} \tag{G.10}$$

By the definition of $w^k_{h,a,a'}$ in (G.5), we can write (G.10) as

$$[Gw^k_{h,a,a'}]_i = \int_{\mathcal{O}^3\times\mathcal{O}^3} \phi_i(x)\cdot\mathcal{K}(x,y)\cdot\rho^k_{h,a,a'}(y)\,\mathrm{d}x\,\mathrm{d}y = \mathbb{E}_{X\sim\phi_i, Y\sim\rho^k_{h,a,a'}}[\mathcal{K}(X,Y)] \tag{G.11}$$

Using the notation of the RKHS $\mathcal{H}$, we can further rewrite (G.8) and (G.11) as

$$[U^k_{h,a,a'}]_i = \langle\mathbb{K}\phi_i, \mathbb{K}\widehat{\mathcal{D}}^k_{h,a,a'}\rangle_{\mathcal{H}}, \quad [Gw^k_{h,a,a'}]_i = \langle\mathbb{K}\phi_i, \mathbb{K}\rho^k_{h,a,a'}\rangle_{\mathcal{H}}.$$

Recall that $\mathbb{K}$ is defined in (3.16). Therefore, using the Cauchy-Schwarz inequality for the inner product in $\mathcal{H}$, we have

$$\begin{aligned}
\|U^k_{h,a,a'} - Gw^k_{h,a,a'}\|_2^2 &= \sum_{i\in[d_o]} \big(\langle\mathbb{K}\phi_i, \mathbb{K}\widehat{\mathcal{D}}^k_{h,a,a'} - \mathbb{K}\rho^k_{h,a,a'}\rangle_{\mathcal{H}}\big)^2 \\
&\le \sum_{i\in[d_o]} \|\mathbb{K}\phi_i\|_{\mathcal{H}}^2\cdot\|\mathbb{K}\widehat{\mathcal{D}}^k_{h,a,a'} - \mathbb{K}\rho^k_{h,a,a'}\|_{\mathcal{H}}^2.
\end{aligned} \tag{G.12}$$

Since $\mathcal{K}$ is uniformly bounded by 1 as specified in Assumption 3.4, we have

$$\|\mathbb{K}\phi_i\|_{\mathcal{H}}^2 = \mathbb{E}_{X\sim\phi_i, Y\sim\phi_i}[\mathcal{K}(X,Y)] \le 1. \tag{G.13}$$

In the sequel, we characterize $\|\mathbb{K}\widehat{\mathcal{D}}^k_{h,a,a'} - \mathbb{K}\rho^k_{h,a,a'}\|_{\mathcal{H}}$ on the right-hand side of (G.12) for any fixed $(h,a,a') \in \{2,\ldots,H\}\times\mathcal{A}^2$. For notational simplicity, we denote by Y_j the data point that is added to $\mathcal{D}_{h,a,a'}$ in the j-th iteration of Algorithm 1. In other words, we have

$$\mathcal{D}^k_{h,a,a'} = \{Y_1,\ldots,Y_k\}, \quad \text{for any } k\in[K].$$

Then, the random function process $\{\mathcal{M}_j\}_{j\ge1}$ defined by

$$\mathcal{M}_j(\cdot) = (1/k)\cdot\Big(\sum_{i=1}^{\min\{j,k\}}\mathcal{K}(Y_i,\cdot) - \sum_{i=1}^{\min\{j,k\}}(\mathbb{K}\rho^{\pi_{i-1}}_{h,a,a'})(\cdot)\Big) \tag{G.14}$$

is a martingale in $\mathcal{H}$ adapted to the data filtration $\{\mathcal{U}_j\}_{j\geq 0}$ of Algorithm 1. In detail, for any $j \in [K]$, we have that $\mathcal{U}_j$ contains the information of all data collected in the first j iterations of Algorithm 1. Note that the data point Y_j follows from the distribution $\rho_{h,a,a'}^{\pi_{j-1}}$ conditioning on $\mathcal{U}_{j-1}$. Therefore, we have

$$\mathbb{E}[\mathcal{K}(Y_j, x) - (\mathbb{K}\rho_{h,a,a'}^{\pi_{j-1}})(x) \,|\, \mathcal{U}_{j-1}] = \mathbb{E}[\mathcal{K}(Y_j, x) \,|\, \mathcal{U}_{j-1}] - \mathbb{E}_{Y\sim\rho_{h,a,a'}^{\pi_{j-1}}}[\mathcal{K}(Y, x)] = 0,$$

for any fixed $x \in \mathcal{O}^3$, which implies that $\{\mathcal{M}_j\}_{j\geq 1}$ defined above is a martingale. Moreover, we have that the total quadratic variation of $\{\mathcal{M}_j\}_{j\geq 1}$ is upper bounded by

$$\sum_{i=1}^{k}(1/k^2)\cdot\|\mathcal{K}(Y_i,\cdot) - (\mathbb{K}\rho_{h,a,a'}^{\pi_i})(\cdot)\|_{\mathcal{H}}^2 \leq \sum_{i=1}^{k}(1/k^2)\cdot 4 = 4/k,$$

where the inequality uses the fact

$$\begin{aligned}&\|\mathcal{K}(Y_i,\cdot) - (\mathbb{K}\rho_{h,a,a'}^{\pi_i})(\cdot)\|_{\mathcal{H}}^2\\ &\quad= \mathcal{K}(Y_i,Y_i) + \mathbb{E}_{Y\sim\rho_{h,a,a'}^{\pi_i}, Y'\sim\rho_{h,a,a'}^{\pi_i}}[\mathcal{K}(Y,Y')] - 2\cdot\mathbb{E}_{Y\sim\rho_{h,a,a'}^{\pi_i}}[\mathcal{K}(Y_i,Y)] \leq 4\end{aligned}$$

following the same argument of (G.13). Then, invoking Lemma G.7 with

$$c^2 = 4/k \quad \text{and} \quad \varepsilon = \sqrt{8\log(2KHA^2/\delta)}\cdot k^{-1/2}$$

for any $\delta > 0$, with probability at least $1 - \delta/(KHA^2)$, it holds that

$$\|\mathcal{M}_k\|_{\mathcal{H}} = \|\mathbb{K}\widehat{\mathcal{D}}_{h,a,a'}^k - \mathbb{K}\rho_{h,a,a'}^k\|_{\mathcal{H}} \leq \sqrt{8\log(2KHA^2/\delta)}\cdot k^{-1/2}. \tag{G.15}$$

Here, the equality uses the definition of $\rho_{h,a,a'}^k = \rho_{h,a,a'}^{\overline{\pi}_k}$. Recall that $\overline{\pi}_k$ is the mixing policy that uniformly selects a policy from $\{\pi_0,\ldots,\pi_{k-1}\}$ at random, which implies

$$\rho_{h,a,a'}^k = \frac{1}{k}\sum_{i=0}^{k-1}\rho_{h,a,a'}^{\pi_i} \quad \text{and} \quad \mathbb{K}\rho_{h,a,a'}^k = \frac{1}{k}\sum_{i=0}^{k-1}\mathbb{K}\rho_{h,a,a'}^{\pi_i}.$$

Then, by further applying the union bound, we have that, with probability at least $1-\delta$, the inequality in (G.15) holds for any $(k,h,a,a') \in [K]\times\{2,\ldots,H\}\times\mathcal{A}^2$. Combining such an upper bound with (G.9), (G.12) and (G.13), we have

$$\|\widehat{w}_{h,a,a'}^k - w_{h,a,a'}^k\|_2 \leq d_{\mathrm{o}}^{1/2}/\alpha\cdot\sqrt{8\log(2KHA^2/\delta)}\cdot k^{-1/2}. \tag{G.16}$$

Combining (G.7) and (G.16), we know that, for any $\delta > 0$, we have

$$\|\widehat{\rho}_{h,a,a'}^k - \rho_{h,a,a'}^k\|_1 \leq d_{\mathrm{o}}^{3/2}/\alpha\cdot\sqrt{8\log(2KHA^2/\delta)}\cdot k^{-1/2},$$

for any $(k,h,a,a') \in [K]\times\{2,\ldots,H\}\times\mathcal{A}^2$, with probability at least $1-\delta$. Therefore, we conclude the proof of Lemma F.2. □

G.3. Proof of Lemma F.3

Proof. By the definition of $\mathbb{V}_{h,a}^{\theta}$ in (F.6), we have

$$\mathbb{E}_{X\sim p}[(\mathbb{F}_{h,a}^{\theta}f)(X)] = \int_{\mathcal{O}^3} f(x)\cdot(\mathbb{V}_{h,a}^{\theta}\rho)(x)\,\mathrm{d}x. \tag{G.17}$$

for any $f \in L^{\infty}(\mathcal{O}^3)$ and $\rho \in \Delta(\mathcal{O}^3)$. By combining (G.17) and Lemma D.1, we have

$$\int_{\mathcal{O}^3} f(x)\cdot(\mathbb{V}_{h,a}^{\theta}\rho_{h,a,a'}^{\theta,\pi} - \rho_{h,a,a'}^{\theta,\pi})(x)\,\mathrm{d}x,$$

for any $f \in L^\infty(\mathcal{O}^3)$, which implies

$$\|\mathbb{V}^\theta_{h,a}\rho^{\theta,\pi}_{h,a,a'} - \rho^{\theta,\pi}_{h,a,a'}\|_1 = 0.$$

For any $\rho \in L^1(\mathcal{O}^3)$ such that $\|\rho\|_1 = 1$, we have

$$\begin{aligned}
\|\mathbb{V}^\theta_{h,a}\rho\|_1 &= \int_{\mathcal{O}^3} \Big| \int_{\mathcal{O}^2} \mathcal{B}^\theta_{h,a}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \cdot \rho(o_{h-1}, o_h, o_{h+1}) \,\mathrm{d}o_h \,\mathrm{d}o_{h+1} \Big| \,\mathrm{d}o_{h-1} \,\mathrm{d}\widetilde{o}_h \,\mathrm{d}\widetilde{o}_{h+1} \\
&\le \int_{\mathcal{S}\times\mathcal{O}^5} p_\theta(\widetilde{\boldsymbol{o}}_h = \widetilde{o}_h, \widetilde{\boldsymbol{o}}_{h+1} = \widetilde{o}_{h+1} \,|\, \widetilde{\boldsymbol{s}}_h = \widetilde{s}_h, \widetilde{\boldsymbol{a}}_h = a) \cdot |\mathcal{Z}^\theta_h(\widetilde{s}_h, o_h)| \\
&\qquad\qquad \cdot |\rho(o_{h-1}, o_h, o_{h+1})| \,\mathrm{d}\widetilde{s}_h \,\mathrm{d}o_h \,\mathrm{d}o_{h+1} \,\mathrm{d}o_{h-1} \,\mathrm{d}\widetilde{o}_h \,\mathrm{d}\widetilde{o}_{h+1} \\
&= \int_{\mathcal{S}\times\mathcal{O}^3} |\mathcal{Z}^\theta_h(\widetilde{s}_h, o_h)| \cdot |\rho(o_{h-1}, o_h, o_{h+1})| \,\mathrm{d}\widetilde{s}_h \,\mathrm{d}o_h \,\mathrm{d}o_{h+1} \,\mathrm{d}o_{h-1},
\end{aligned} \tag{G.18}$$

where the inequality is by the definition of $\mathcal{B}^\theta_{h,a}$ in (3.6). Combining (G.18) with (D.15) from the proof of Lemma D.1, we have

$$\|\mathbb{V}^\theta_{h,a}\rho\|_1 \le \int_{\times\mathcal{O}^3} \gamma \cdot |\rho(o_{h-1}, o_h, o_{h+1})| \,\mathrm{d}o_h \,\mathrm{d}o_{h+1} \,\mathrm{d}o_{h-1} = \gamma,$$

which implies $\|\mathbb{V}^\theta_{h,a}\|_{1\to 1} \le \gamma$.

Therefore, we conclude the proof of Lemma F.3. □

G.4. Property of the Bridge Operator

Lemma G.1 (Bridge Property)**.** *Recall that we denote by σ_{h-1} the event*

$$\overline{\boldsymbol{\tau}}_{h-1} = \overline{\tau}_{h-1}, \quad \boldsymbol{a}_{h-1} = a_{h-1}.$$

For any $(h, \theta, \overline{\tau}_{h-1}, a_{h-1}) \in [H] \times \Theta \times \overline{\Gamma}_{h-1} \times \mathcal{A}$, we have

$$\mathbb{E}_\theta[\mathcal{Z}^\theta_h(\widetilde{s}_h, \boldsymbol{o}_h) \,|\, \sigma_{h-1}] = p_\theta(\widetilde{\boldsymbol{s}}_h = \widetilde{s}_h \,|\, \sigma_{h-1}).$$

Proof. By the tower property of the expectation and the Markov property of the POMDP, we have

$$\mathbb{E}_\theta[\mathcal{Z}^\theta_h(\widetilde{s}_h, \boldsymbol{o}_h) \,|\, \sigma_{h-1}] = \mathbb{E}_\theta\big[\mathbb{E}_\theta[\mathcal{Z}^\theta_h(\widetilde{s}_h, \boldsymbol{o}_h) \,|\, \boldsymbol{s}_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}] \,\big|\, \sigma_{h-1}\big]. \tag{G.19}$$

Note that, for any $s_{h-1} \in \mathcal{S}$, we have

$$\begin{aligned}
&\mathbb{E}_\theta[\mathcal{Z}^\theta_h(\widetilde{s}_h, \boldsymbol{o}_h) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}] \\
&= \int_{\mathcal{O}} \mathcal{Z}^\theta_h(\widetilde{s}_h, o_h) \cdot p_\theta(\boldsymbol{o}_h = o_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}) \,\mathrm{d}o_h \\
&= \int_{\mathcal{O}\times\mathcal{S}} \mathcal{Z}^\theta_h(\widetilde{s}_h, o_h) \cdot \mathcal{E}^\theta_h(o_h \,|\, s_h) \cdot p_\theta(\boldsymbol{s}_h = s_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}) \,\mathrm{d}s_h \,\mathrm{d}o_h.
\end{aligned} \tag{G.20}$$

We define the function $f : \mathcal{S} \to \mathbb{R}$ by

$$f(s_h) = p_\theta(\boldsymbol{s}_h = s_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}), \quad \text{for any } s_h \in \mathcal{S}.$$

Then, we have $f \in \mathcal{F}_{\mathrm{s}}$ and we can write the right-hand side of (G.20) as

$$(\mathbb{Z}^\theta_h \mathbb{O}^\theta_h f)(\widetilde{s}_h) = f(\widetilde{s}_h) = p_\theta(\boldsymbol{s}_h = \widetilde{s}_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1})$$

following Assumptions 2.1 and 2.2. In other words, we have

$$\begin{aligned}
&\mathbb{E}_\theta[\mathcal{Z}^\theta_h(\widetilde{s}_h, \boldsymbol{o}_h) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}] \\
&\quad = p_\theta(\boldsymbol{s}_h = \widetilde{s}_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}).
\end{aligned} \tag{G.21}$$

Combining (G.19) and (G.21) and using the Markov property of the POMDP, we obtain

$$\begin{aligned}
&\mathbb{E}_\theta[\mathcal{Z}_h^\theta(\widetilde{s}_h, \boldsymbol{o}_h) \,|\, \sigma_{h-1}] \\
&= \int_{\mathcal{S}} p_\theta(\boldsymbol{s}_h = \widetilde{s}_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}) \cdot p_\theta(\boldsymbol{s}_{h-1} = s_{h-1} \,|\, \sigma_{h-1}) \,\mathrm{d}s_{h-1} \\
&= \int_{\mathcal{S}} p_\theta(\boldsymbol{s}_h = \widetilde{s}_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \sigma_{h-1}) \cdot p_\theta(\boldsymbol{s}_{h-1} = s_{h-1} \,|\, \sigma_{h-1}) \,\mathrm{d}s_{h-1} \\
&= p_\theta(\boldsymbol{s}_h = \widetilde{s}_h \,|\, \sigma_{h-1}),
\end{aligned}$$

which concludes the proof of lemma G.1. □

The following lemma is a variant of Lemma 3.1, which adds the state information to the expectation condition.

Lemma G.2 (Variant of Lemma 3.1)**.** *For any $(h, \theta, \pi, s_{h-1}, \overline{\tau}_{h-1}, a_{h-1}) \in [H] \times \Theta \times \Pi \times \mathcal{S} \times \overline{\Gamma}_{h-1} \times \mathcal{A}$ and $f \in L^\infty(\overline{\Gamma}_{h+1})$, we have*

$$\mathbb{E}_\theta[(\mathbb{B}_h^{\theta,\pi} f)(\overline{\boldsymbol{\tau}}_h) - f(\overline{\boldsymbol{\tau}}_{h+1}) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \sigma_{h-1}] = 0.$$

Proof. The proof is very similar to the proof of Lemma 3.1. Following the notation in Lemma 3.1, by the definition of $\mathbb{B}_h^{\theta,\pi}$ in (3.4), we have

$$\begin{aligned}
&\mathbb{E}_\theta[(\mathbb{B}_h^{\theta,\pi} f)(\overline{\boldsymbol{\tau}}_h) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \sigma_{h-1}] \qquad \text{(G.22)} \\
&= \int_{\mathcal{S}\times\mathcal{O}^3} f\big(\overline{\tau}_h^\dagger, \pi(\tau_h^\dagger), \widetilde{o}_{h+1}\big) \cdot p_\theta\big(\boldsymbol{o}_h = \widetilde{o}_h, \boldsymbol{o}_{h+1} = \widetilde{o}_{h+1} \,|\, \boldsymbol{s}_h = \widetilde{s}_h, \boldsymbol{a}_h = \pi(\tau_h^\dagger)\big) \\
&\qquad\qquad \cdot \mathcal{Z}_h^\theta(\widetilde{s}_h, o_h) \cdot p_\theta(\boldsymbol{o}_h = o_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \sigma_{h-1}) \,\mathrm{d}o_h \,\mathrm{d}\widetilde{o}_h \,\mathrm{d}\widetilde{o}_{h+1} \,\mathrm{d}\widetilde{s}_h.
\end{aligned}$$

Here, by (G.20)-(G.21) in the proof of Lemma G.1, we have

$$\begin{aligned}
&\int_{\mathcal{O}} \mathcal{Z}_h^\theta(\widetilde{s}_h, o_h) \cdot p_\theta(\boldsymbol{o}_h = o_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \sigma_{h-1}) \,\mathrm{d}o_h \\
&= \int_{\mathcal{O}} \mathcal{Z}_h^\theta(\widetilde{s}_h, o_h) \cdot p_\theta(\boldsymbol{o}_h = o_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}) \,\mathrm{d}o_h \\
&= p_\theta(\boldsymbol{s}_h = \widetilde{s}_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}) \\
&= p_\theta(\boldsymbol{s}_h = \widetilde{s}_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \sigma_{h-1}).
\end{aligned}$$

Thus, we can rewrite (G.22) as

$$\begin{aligned}
&\mathbb{E}_\theta[(\mathbb{B}_h^{\theta,\pi} f)(\overline{\boldsymbol{\tau}}_h) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \sigma_{h-1}] \qquad \text{(G.23)} \\
&= \int_{\mathcal{S}\times\mathcal{O}^2} f\big(\overline{\tau}_h^\dagger, \pi(\tau_h^\dagger), \widetilde{o}_{h+1}\big) \cdot p_\theta\big(\boldsymbol{o}_h = \widetilde{o}_h, \boldsymbol{o}_{h+1} = \widetilde{o}_{h+1} \,|\, \boldsymbol{s}_h = \widetilde{s}_h, \boldsymbol{a}_h = \pi(\overline{\tau}_h^\dagger)\big) \\
&\qquad\qquad \cdot p(\boldsymbol{s}_h = \widetilde{s}_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \sigma_{h-1}) \,\mathrm{d}\widetilde{o}_h \,\mathrm{d}\widetilde{o}_{h+1} \,\mathrm{d}\widetilde{s}_h \\
&= \int_{\mathcal{O}^2} f\big(\overline{\tau}_h^\dagger, \pi(\tau_h^\dagger), \widetilde{o}_{h+1}\big) \cdot p_{\theta,\pi}(\boldsymbol{o}_h = \widetilde{o}_h, \boldsymbol{o}_{h+1} = \widetilde{o}_{h+1} \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \sigma_{h-1}) \,\mathrm{d}\widetilde{o}_h \,\mathrm{d}\widetilde{o}_{h+1}.
\end{aligned}$$

where the second equality uses the independence between $(\boldsymbol{o}_h, \boldsymbol{o}_{h+1})$ and $\boldsymbol{\tau}_{h-1}$ conditioning on $(\boldsymbol{s}_h, \boldsymbol{a}_h)$. Replacing the notations $\widetilde{o}_h$ and $\widetilde{o}_{h+1}$ of the integral variables on the right-hand side of (G.23) by o_h and o_{h+1}, respectively, we obtain

$$\begin{aligned}
&\mathbb{E}_\theta[(\mathbb{B}_h^{\theta,\pi} f)(\overline{\boldsymbol{\tau}}_h) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \sigma_{h-1}] \qquad \text{(G.24)} \\
&= \int_{\mathcal{O}^2} f\big(\overline{\tau}_h, \pi(\tau_h), o_{h+1}\big) \cdot p_{\theta,\pi}(\boldsymbol{o}_h = o_h, \boldsymbol{o}_{h+1} = o_{h+1} \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \sigma_{h-1}) \,\mathrm{d}o_h \,\mathrm{d}o_{h+1} \\
&= \mathbb{E}_{\theta,\pi}[f(\overline{\boldsymbol{\tau}}_{h+1}) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \sigma_{h-1}],
\end{aligned}$$

where we denote $\overline{\tau}_h = (\overline{\tau}_{h-1}, a_{h-1}, o_h)$ and $\tau_h = (\tau_{h-1}, o_h)$. Therefore, we conclude the proof of Lemma G.2. □

G.5. Properties of the Value Functions

Lemma G.3. *For any $(h, \pi, \theta, \overline{\tau}_h) \in [H] \times \Pi \times \Theta \times \overline{\Gamma}_h$, it holds that*

$$V_h^{\theta,\pi}(\overline{\tau}_h) = \int_{\mathcal{S}} \mathbb{E}_{\theta,\pi}\Big[\sum_{i=1}^{H} \boldsymbol{r}_i \,\Big|\, \boldsymbol{s}_h = s_h, \overline{\boldsymbol{\tau}}_{h-1} = \overline{\tau}_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}\Big] \cdot \mathcal{Z}_h^{\theta}(s_h, o_h)\,\mathrm{d}s_h. \tag{G.25}$$

Here, we denote $\overline{\tau}_h = (\overline{\tau}_{h-1}, a_{h-1}, o_h)$ following the definition in (2.3).

Proof. We prove the lemma by induction over $h \in [H]$. When $h = H$, by the definition of the value function in (3.10) and the definition of $\mathbb{B}_H^{\theta,\pi}$ in (3.4), we have

$$\begin{aligned} V_H^{\theta,\pi}(\overline{\tau}_H) &= (\mathbb{B}_H^{\theta,\pi} R)(\overline{\tau}_H) \\ &= \int_{\mathcal{O}^2} \Big(r\big(\widetilde{o}_H, \pi(\tau_H^\dagger)\big) + \sum_{i=1}^{H-1} r(o_h, a_h)\Big) \cdot \mathcal{B}^{\theta}_{h,\pi(\tau_H^\dagger)}(o_H, \widetilde{o}_H, \widetilde{o}_{H+1})\,\mathrm{d}\widetilde{o}_H\,\mathrm{d}\widetilde{o}_{H+1} \\ &= \int_{\mathcal{O}} \Big(r\big(\widetilde{o}_H, \pi(\tau_H^\dagger)\big) + \sum_{i=1}^{H-1} r(o_h, a_h)\Big) \cdot \Big(\int_{\mathcal{O}} \mathcal{B}^{\theta}_{h,\pi(\tau_H^\dagger)}(o_H, \widetilde{o}_H, \widetilde{o}_{H+1})\,\mathrm{d}\widetilde{o}_{H+1}\Big)\,\mathrm{d}\widetilde{o}_H. \end{aligned} \tag{G.26}$$

Recall that $\tau_H^\dagger$ is the tail-mirrored observation history defined in (3.5). Note that, by the definition of $\{\mathcal{B}^{\theta}_{H,a}\}_{a\in\mathcal{A}}$ in (3.6), we have

$$\begin{aligned} &\int_{\mathcal{O}} \mathcal{B}^{\theta}_{h,\pi(\tau_H^\dagger)}(o_H, \widetilde{o}_H, \widetilde{o}_{H+1})\,\mathrm{d}\widetilde{o}_{H+1} \\ &= \int_{\mathcal{S}\times\mathcal{O}} p_\theta(\widetilde{\boldsymbol{o}}_H = \widetilde{o}_H, \widetilde{\boldsymbol{o}}_{H+1} = \widetilde{o}_{H+1} \,|\, \widetilde{\boldsymbol{s}}_H = s_H, \widetilde{\boldsymbol{a}}_H = a) \cdot \mathcal{Z}_H^{\theta}(s_H, o_H)\,\mathrm{d}s_H\,\mathrm{d}\widetilde{o}_{H+1} \\ &= \int_{\mathcal{S}} p_\theta(\boldsymbol{o}_h = \widetilde{o}_H \,|\, \boldsymbol{s}_H = s_H, \boldsymbol{a}_H = a) \cdot \mathcal{Z}_H^{\theta}(s_H, o_H)\,\mathrm{d}s_H, \end{aligned} \tag{G.27}$$

where we use the fact that $\widetilde{\boldsymbol{s}}_H$, $\widetilde{\boldsymbol{a}}_H$, $\widetilde{\boldsymbol{o}}_H$, and $\widetilde{\boldsymbol{o}}_{H+1}$ in (3.6) have the same distribution of $\boldsymbol{s}_H$, $\boldsymbol{a}_H$, $\boldsymbol{o}_H$, and $\boldsymbol{o}_{H+1}$. Combining (G.26) and (G.27), we have that (G.25) holds for $h = H$.

Assume that (G.25) holds when $h = j+1$ for some fixed $j \le H-1$. Then, by the definition of the value function in (3.10), we have

$$V_j^{\theta,\pi}(\overline{\tau}_j) = (\mathbb{B}_j^{\theta,\pi} V_{j+1}^{\theta,\pi})(\overline{\tau}_j), \quad \text{for any } \overline{\tau}_j \in \overline{\Gamma}_j.$$

Applying the induction assumption and definition of $\mathbb{B}_j^{\theta,\pi}$ in (3.4), we obtain

$$\begin{aligned} (\mathbb{B}_j^{\theta,\pi} V_{j+1}^{\theta,\pi})(\overline{\tau}_j) = \int_{\mathcal{S}\times\mathcal{O}^2} &\mathbb{E}_{\theta,\pi}\Big[\sum_{i=1}^{H} \boldsymbol{r}_i \,\Big|\, \boldsymbol{s}_{j+1} = s_{j+1}, \overline{\boldsymbol{\tau}}_j = \overline{\tau}_j^\dagger, \boldsymbol{a}_j = \pi(\tau_j^\dagger)\Big] \cdot \mathcal{Z}_h^{\theta}(s_{j+1}, \widetilde{o}_{j+1}) \\ &\cdot \mathcal{B}^{\theta}_{j,\pi(\tau_j^\dagger)}(o_j, \widetilde{o}_j, \widetilde{o}_{j+1})\,\mathrm{d}s_{j+1}\,\mathrm{d}\widetilde{o}_j\,\mathrm{d}\widetilde{o}_{j+1}. \end{aligned} \tag{G.28}$$

Recall that $\tau_j^\dagger$ and $\overline{\tau}_j^\dagger$ are the tail-mirrored observation history and full history defined in (3.5), respectively. Note that, by the definition of $\{\mathcal{B}^{\theta}_{j,a}\}_{a\in\mathcal{A}}$ in (3.6), we have

$$\begin{aligned} &\mathcal{B}^{\theta}_{j,\pi(\tau_j^\dagger)}(o_j, \widetilde{o}_j, \widetilde{o}_{j+1}) \\ &= \int_{\mathcal{S}} p_\theta\big(\boldsymbol{o}_j = \widetilde{o}_j, \boldsymbol{o}_{j+1} = \widetilde{o}_{j+1} \,|\, \boldsymbol{s}_j = s_j, \boldsymbol{a}_j = \pi(\tau_j^\dagger)\big) \cdot \mathcal{Z}_j^{\theta}(s_j, o_j)\,\mathrm{d}s_j \\ &= \int_{\mathcal{S}} \mathcal{E}_h^{\theta}(\widetilde{o}_j \,|\, s_j) \cdot p_\theta\big(\boldsymbol{o}_{j+1} = \widetilde{o}_{j+1} \,|\, \boldsymbol{s}_j = s_j, \boldsymbol{a}_j = \pi(\tau_j^\dagger)\big) \cdot \mathcal{Z}_j^{\theta}(s_j, o_j)\,\mathrm{d}s_j \end{aligned} \tag{G.29}$$

Following the same argument in (G.20)-(G.21), we have

$$\begin{aligned}&\int_{\mathcal{O}} \mathcal{Z}_h^\theta(s_{j+1}, \widetilde{o}_{j+1}) \cdot p_\theta\big(\boldsymbol{o}_{j+1} = \widetilde{o}_{j+1} \,|\, \boldsymbol{s}_j = s_j, \boldsymbol{a}_j = \pi(\tau_j^\dagger)\big)\,\mathrm{d}\widetilde{o}_{j+1}\\ &\quad = p_\theta\big(\boldsymbol{s}_{j+1} = s_{j+1} \,|\, \boldsymbol{s}_j = s_j, \boldsymbol{a}_j = \pi(\tau_j^\dagger)\big),\end{aligned}$$

combining which with (G.29), we obtain

$$\begin{aligned}&\int_{\mathcal{O}} \mathcal{Z}_h^\theta(s_{j+1}, \widetilde{o}_{j+1}) \cdot \mathcal{B}^\theta_{j,\pi(\tau_j^\dagger)}(o_j, \widetilde{o}_j, \widetilde{o}_{j+1})\,\mathrm{d}\widetilde{o}_{j+1}\\ &\quad = \int_{\mathcal{S}} \mathcal{E}_h^\theta(\widetilde{o}_j \,|\, s_j) \cdot p_\theta\big(\boldsymbol{s}_{j+1} = s_{j+1} \,|\, \boldsymbol{s}_j = s_j, \boldsymbol{a}_j = \pi(\tau_j^\dagger)\big) \cdot \mathcal{Z}_j^\theta(s_j, o_j)\,\mathrm{d}s_j\\ &\quad = \int_{\mathcal{S}} p_\theta\big(\boldsymbol{o}_j = \widetilde{o}_j, \boldsymbol{s}_{j+1} = s_{j+1} \,|\, \boldsymbol{s}_j = s_j, \boldsymbol{a}_j = \pi(\tau_j^\dagger)\big) \cdot \mathcal{Z}_j^\theta(s_j, o_j)\,\mathrm{d}s_j. \qquad \text{(G.30)}\end{aligned}$$

Plugging (G.30) into the right-hand side of (G.28), we obtain

$$\begin{aligned}&(\mathbb{B}_j^{\theta,\pi} V_{j+1}^{\theta,\pi})(\overline{\tau}_j) \qquad \text{(G.31)}\\ &\quad = \int_{\mathcal{S}^2\times\mathcal{O}} \mathbb{E}_{\theta,\pi}\Big[\sum_{i=1}^H \boldsymbol{r}_i \,\Big|\, \boldsymbol{s}_{j+1} = s_{j+1}, \overline{\boldsymbol{\tau}}_j = \overline{\tau}_j^\dagger, \boldsymbol{a}_j = \pi(\tau_j^\dagger)\Big]\\ &\qquad\qquad \cdot p_\theta\big(\boldsymbol{o}_j = \widetilde{o}_j, \boldsymbol{s}_{j+1} = s_{j+1} \,|\, \boldsymbol{s}_j = s_j, \boldsymbol{a}_j = \pi(\tau_j^\dagger)\big) \cdot \mathcal{Z}_j^\theta(s_j, o_j)\,\mathrm{d}s_j\,\mathrm{d}s_{j+1}\,\mathrm{d}\widetilde{o}_j.\end{aligned}$$

Using the Markov property of the POMDP, we can simplify the right-hand side of (G.31) to obtain

$$(\mathbb{B}_j^{\theta,\pi} V_{j+1}^{\theta,\pi})(\overline{\tau}_j) = \int_{\mathcal{S}} \mathbb{E}_{\theta,\pi}\Big[\sum_{i=1}^H \boldsymbol{r}_i \,\Big|\, \boldsymbol{s}_j = s_j, \overline{\boldsymbol{\tau}}_{j-1} = \overline{\tau}_{j-1}, \boldsymbol{a}_{j-1} = a_{j-1}\Big] \cdot \mathcal{Z}_j^\theta(s_j, o_j)\,\mathrm{d}s_j,$$

which implies that (G.25) holds when $h = j$.

Therefore, we conclude the proof of Lemma G.3 by induction. □

Lemma G.4. *For any $(h, \theta, \overline{\tau}_h, \pi) \in [H] \times \Theta \times \overline{\Gamma}_h \times \Pi$, we have*

$$|V_h^{\theta,\pi}(\overline{\tau}_h)| \le \gamma H.$$

Recall that γ is defined in Assumption 2.2.

Proof. By Lemma G.3 and Assumption 2.2, we have

$$\begin{aligned}|V_h^{\theta,\pi}(\overline{\tau}_h)| &= \Big|\int_{\mathcal{S}} \mathbb{E}_{\theta,\pi}\Big[\sum_{i=1}^H \boldsymbol{r}_i \,\Big|\, \boldsymbol{s}_h = s_h, \overline{\boldsymbol{\tau}}_{h-1} = \overline{\tau}_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}\Big] \cdot \mathcal{Z}_h^\theta(s_h, o_h)\,\mathrm{d}s_h\Big|\\ &\le H \cdot \int_{\mathcal{S}} |\mathcal{Z}_h^\theta(s_h, o_h)|\,\mathrm{d}s_h = H \cdot \|\mathbb{Z}_h^\theta \delta_{o_h}\|_1 \le \gamma H \cdot \|\delta_{o_h}\|_1 = \gamma H,\end{aligned}$$

for any $(h, \theta, \overline{\tau}_h, \pi) \in [H] \times \Theta \times \overline{\Gamma}_h \times \Pi$. Here, δ_{o_h} is the Dirac delta function defined on $\mathcal{O}$, whose value is zero everywhere except at o_h, and whose integral over $\mathcal{O}$ is equal to one. Thus, we conclude the proof of Lemma G.4. □

G.6. Properties of the State-Dependent Error

Lemma G.5. *For any $(k, h) \in [K] \times \{2, \ldots, H\}$, we have*

$$\mathbb{E}_{\theta^*,\overline{\pi}_k}[e_h^k(\boldsymbol{s}_h)] \le \gamma^2 H \cdot \sum_{a,a' \in \mathcal{A}} \big\|\mathbb{V}_{h,a'}^{\theta_k} \rho_{h,a,a'}^{\overline{\pi}_k} - \rho_{h,a,a'}^{\overline{\pi}_k}\big\|_1.$$

Proof. By the definition of e_h^k in (A.2), we have

$$e_h^k(s_{h-1}) = \big|\mathbb{E}_{\theta^*,\pi_k}[(\mathbb{B}_h^{\theta_k,\pi_k} V_{h+1}^{\theta_k,\pi_k})(\overline{\boldsymbol{\tau}}_h) - (\mathbb{B}_h^{\theta^*,\pi_k} V_{h+1}^{\theta_k,\pi_k})(\overline{\boldsymbol{\tau}}_h) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}]\big|, \tag{G.32}$$

for any $s_{h-1} \in \mathcal{S}$. Note that, by the definition of $\mathbb{B}_h^{\theta,\pi}$ in (3.4), we have

$$\begin{aligned}
&\mathbb{E}_{\theta^*,\pi_k}[(\mathbb{B}_h^{\theta_k,\pi_k} V_{h+1}^{\theta_k,\pi_k})(\overline{\boldsymbol{\tau}}_h) - (\mathbb{B}_h^{\theta^*,\pi_k} V_{h+1}^{\theta_k,\pi_k})(\overline{\boldsymbol{\tau}}_h) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \overline{\boldsymbol{\tau}}_{h-1} = \overline{\tau}_{h-1}] \\
&\quad = \int_{\mathcal{O}^3} V_{h+1}^{\theta_k,\pi_k}\big(\overline{\tau}_h^\dagger, \pi_k(\tau_h^\dagger), \widetilde{o}_{h+1}\big) \cdot \Delta\mathcal{B}^{\theta_k,\theta^*}_{h,\pi_k(\tau_h^\dagger)}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \\
&\qquad\qquad \cdot p_{\theta^*}\big(\boldsymbol{o}_h = o_h \,\big|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = \pi_k(\tau_{h-1})\big)\,\mathrm{d}o_h\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1},
\end{aligned} \tag{G.33}$$

for any $(s_{h-1}, \overline{\tau}_{h-1}) \in \mathcal{S} \times \overline{\Gamma}_{h-1}$, where $\overline{\tau}_h^\dagger = (\overline{\tau}_{h-1}, a_{h-1}, \widetilde{o}_h)$, $\tau_h^\dagger = (\tau_{h-1}, \widetilde{o}_h)$, and

$$\Delta\mathcal{B}^{\theta_k,\theta^*}_{h,\pi_k(\tau_h^\dagger)}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) = \mathcal{B}^{\theta_k}_{h,\pi_k(\tau_h^\dagger)}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) - \mathcal{B}^{\theta^*}_{h,\pi_k(\tau_h^\dagger)}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}). \tag{G.34}$$

By replacing the actions $\pi_k(\tau_{h-1})$ and $\pi_k(\tau_h^\dagger)$ on the right-hand side of (G.33) by all possible action combinations, we have the inequality

$$\begin{aligned}
&\big|\mathbb{E}_{\theta^*,\pi_k}[(\mathbb{B}_h^{\theta_k,\pi_k} V_{h+1}^{\theta_k,\pi_k})(\overline{\boldsymbol{\tau}}_h) - (\mathbb{B}_h^{\theta^*,\pi_k} V_{h+1}^{\theta_k,\pi_k})(\overline{\boldsymbol{\tau}}_h) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \overline{\boldsymbol{\tau}}_{h-1} = \overline{\tau}_{h-1}]\big| \\
&\quad \le \sum_{a,a' \in \mathcal{A}} \Big| \int_{\mathcal{O}^3} V_{h+1}^{\theta_k,\pi_k}(\overline{\tau}_h^\dagger, a', \widetilde{o}_{h+1}) \cdot \Delta\mathcal{B}^{\theta_k,\theta^*}_{h,a'}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \\
&\qquad\qquad \cdot p_{\theta^*}\big(\boldsymbol{o}_h = o_h \,\big|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a\big)\,\mathrm{d}o_h\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1}\Big|.
\end{aligned} \tag{G.35}$$

Invoking Lemma G.4, we can further upper bound the left-hand side of (G.36) as

$$\begin{aligned}
&\big|\mathbb{E}_{\theta^*,\pi_k}[(\mathbb{B}_h^{\theta_k,\pi_k} V_{h+1}^{\theta_k,\pi_k})(\overline{\boldsymbol{\tau}}_h) - (\mathbb{B}_h^{\theta^*,\pi_k} V_{h+1}^{\theta_k,\pi_k})(\overline{\boldsymbol{\tau}}_h) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \overline{\boldsymbol{\tau}}_{h-1} = \overline{\tau}_{h-1}]\big| \\
&\quad \le \sum_{a,a' \in \mathcal{A}} \sup_{(\overline{\tau}_h^\dagger, \widetilde{o}_{h+1}) \in \overline{\Gamma}_h \times \mathcal{O}} |V_{h+1}^{\theta_k,\pi_k}(\overline{\tau}_h^\dagger, a', \widetilde{o}_{h+1})| \cdot \int_{\mathcal{O}^2} \Big| \int_{\mathcal{O}} \Delta\mathcal{B}^{\theta_k,\theta^*}_{h,a'}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \\
&\qquad\qquad \cdot p_{\theta^*}(\boldsymbol{o}_h = o_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a)\,\mathrm{d}o_h\Big|\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1} \\
&\quad \le \sum_{a,a' \in \mathcal{A}} \gamma H \cdot \int_{\mathcal{O}^2} \Big| \int_{\mathcal{O}} \Delta\mathcal{B}^{\theta_k,\theta^*}_{h,a'}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \\
&\qquad\qquad \cdot p_{\theta^*}(\boldsymbol{o}_h = o_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a)\,\mathrm{d}o_h\Big|\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1}.
\end{aligned} \tag{G.36}$$

Combining (G.32) and (G.36), and applying Jensen's inequality, we obtain

$$\begin{aligned}
e_h^k(s_{h-1}) \le \sum_{a,a' \in \mathcal{A}} \gamma H \cdot \int_{\mathcal{O}^2} \Big| \int_{\mathcal{O}} \Delta\mathcal{B}^{\theta_k,\theta^*}_{h,a'}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \\
\cdot p_{\theta^*}(\boldsymbol{o}_h = o_h \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \boldsymbol{a}_{h-1} = a)\,\mathrm{d}o_h\Big|\,\mathrm{d}\widetilde{o}_h\,\mathrm{d}\widetilde{o}_{h+1},
\end{aligned} \tag{G.37}$$

for any $s_{h-1} \in \mathcal{S}$.

In the sequel, we characterize the expectation of both sides of (G.37) with respect to the marginal distribution of $\boldsymbol{s}_{h-1}$. We define the function $f : \mathcal{S} \to \mathbb{R}$ by

$$f(s_{h-1}) = \int_{\mathcal{O}} \Delta\mathcal{B}^{\theta_k,\theta^*}_{h,a'}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \cdot p_{\theta^*,\overline{\pi}_k}(\boldsymbol{o}_h = o_h, \boldsymbol{s}_{h-1} = s_{h-1} \,|\, \boldsymbol{a}_{h-1} = a)\,\mathrm{d}o_h, \tag{G.38}$$

for any $s_{h-1} \in \mathcal{S}$. With f defined above and the inequality in (G.37), we have

$$\mathbb{E}_{\theta^*,\overline{\pi}_k}[e_H^k(\boldsymbol{s}_{h-1})] = \gamma H \cdot \sum_{a,a' \in \mathcal{A}} \int_{\mathcal{O}} \|f\|_1\,\mathrm{d}\widetilde{o}_h\widetilde{o}_{h+1}, \tag{G.39}$$

where we take the expectation of both sides of (G.37) with respect to the marginal distribution of $\boldsymbol{s}_{h-1}$, following the policy $\overline{\pi}_k$. Note that we can write the probability on the right-hand side of (G.38) as

$$\begin{aligned}&p_{\theta^*,\overline{\pi}_k}(\boldsymbol{o}_h = o_h, \boldsymbol{s}_{h-1} = s_{h-1} \,|\, \boldsymbol{a}_{h-1} = a)\\ &\quad = p_{\theta^*,\overline{\pi}_k}(\boldsymbol{s}_{h-1} = s_{h-1} \,|\, \boldsymbol{o}_h = o_h, \boldsymbol{a}_{h-1} = a) \cdot p_{\theta^*,\overline{\pi}_k}(\boldsymbol{o}_h = o_h \,|\, \boldsymbol{a}_{h-1} = a),\end{aligned}$$

which implies $f \in \mathcal{F}_{\mathrm{s}}' \subset \mathrm{linspan}(\{\psi_i\}_{i=1}^{d_{\mathrm{s}}})$ following Assumption 2.1. Then, by further applying Assumption 2.2, we obtain

$$\|f\|_1 = \|\mathbb{Z}_{h-1}^{\theta}\mathbb{O}_{h-1}^{\theta} f\|_1 \le \gamma \cdot \|\mathbb{O}_{h-1}^{\theta} f\|_1. \tag{G.40}$$

With f defined in (G.38) and $\mathbb{O}_{h-1}^{\theta}$ defined in (2.4), we can write

$$\begin{aligned}&\|\mathbb{O}_{h-1}^{\theta} f\|_1 \qquad\qquad (\text{G.41})\\ &= \int_{\mathcal{O}} \Big| \int_{\mathcal{O}\times\mathcal{S}} \Delta\mathcal{B}_{h,a'}^{\theta_k,\theta^*}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \cdot p_{\theta^*,\overline{\pi}_k}(\boldsymbol{o}_h = o_h, \boldsymbol{s}_{h-1} = s_{h-1} \,|\, \boldsymbol{a}_{h-1} = a)\\ &\qquad\qquad \cdot \mathcal{E}_{h-1}^{\theta}(o_{h-1} \,|\, s_{h-1}) \,\mathrm{d}o_h \,\mathrm{d}s_{h-1} \Big| \,\mathrm{d}o_{h-1}\\ &= \int_{\mathcal{O}} \Big| \int_{\mathcal{O}} \Delta\mathcal{B}_{h,a'}^{\theta_k,\theta^*}(o_h, \widetilde{o}_h, \widetilde{o}_{h+1}) \cdot p_{\theta^*,\overline{\pi}_k}(\boldsymbol{o}_{h-1} = o_{h-1}, \boldsymbol{o}_h = o_h \,|\, \boldsymbol{a}_{h-1} = a) \,\mathrm{d}o_h \Big| \,\mathrm{d}o_{h-1}\end{aligned}$$

Here, we can rewrite the probability on the right-hand side as

$$\begin{aligned}&p_{\theta^*,\overline{\pi}_k}(\boldsymbol{o}_{h-1} = o_{h-1}, \boldsymbol{o}_h = o_h \,|\, \boldsymbol{a}_{h-1} = a)\\ &\quad = \int_{\mathcal{O}} p_{\theta^*,\overline{\pi}_k}(\boldsymbol{o}_{h-1} = o_{h-1}, \boldsymbol{o}_h = o_h, \boldsymbol{o}_{h+1} = o_{h+1} \,|\, \boldsymbol{a}_{h-1} = a, \boldsymbol{a}_h = a') \,\mathrm{d}o_{h+1}\\ &\quad = \int_{\mathcal{O}} \rho_{h,a,a'}^{\overline{\pi}_k}(o_{h-1}, o_h, o_{h+1}) \,\mathrm{d}o_{h+1}.\end{aligned}$$

Recall that $\Delta\mathcal{B}_{h,a'}^{\theta_k,\theta^*}$ is defined in (G.34). Then, by applying the definition of $\mathbb{V}_{h,a'}^{\theta}$ in (F.6) for $\theta = \theta_k$ and $\theta = \theta^*$ to the right-hand side of (G.41) and integrating for $\widetilde{o}_h, \widetilde{o}_{h+1}$ over $\mathcal{O}^2$ for both sides, we have

$$\int_{\mathcal{O}^3} \|\mathbb{O}_{h-1}^{\theta} f\|_1 \,\mathrm{d}\widetilde{o}_h \,\mathrm{d}\widetilde{o}_{h+1} = \|\mathbb{V}_{h,a'}^{\theta_k}\rho_{h,a,a'}^{\overline{\pi}_k} - \mathbb{V}_{h,a'}^{\theta^*}\rho_{h,a,a'}^{\overline{\pi}_k}\|_1 = \|\mathbb{V}_{h,a'}^{\theta_k}\rho_{h,a,a'}^{\overline{\pi}_k} - \rho_{h,a,a'}^{\overline{\pi}_k}\|_1, \tag{G.42}$$

where the second equality is by Lemma F.3. Then, by combining (G.39), (G.40) and (G.42), we obtain

$$\begin{aligned}\mathbb{E}_{\theta^*,\overline{\pi}_k}[e_h^k(\boldsymbol{s}_{h-1})] &\le \gamma H \cdot \sum_{a,a'\in\mathcal{A}} \int_{\mathcal{O}^3} \|f\|_1 \,\mathrm{d}\widetilde{o}_h \,\mathrm{d}\widetilde{o}_{h+1}\\ &\le \gamma^2 H \cdot \sum_{a,a'\in\mathcal{A}} \int_{\mathcal{O}^3} \|\mathbb{O}_{h-1}^{\theta} f\|_1 \,\mathrm{d}\widetilde{o}_h \,\mathrm{d}\widetilde{o}_{h+1}\\ &= \gamma^2 H \cdot \sum_{a,a'\in\mathcal{A}} \|\mathbb{V}_h^{\theta_k}\rho_{h,a,a'}^{\overline{\pi}_k} - \rho_{h,a,a'}^{\overline{\pi}_k}\|_1,\end{aligned}$$

which concludes the proof of Lemma G.5. □

Lemma G.6. *For any $(k,h) \in [K] \times \{2,\ldots,H\}$ and $s_{h-1} \in \mathcal{S}$, we have*

$$e_h^k(s_{h-1}) \le 2\gamma H.$$

Proof. For any $(k, h, s_{h-1}, \overline{\tau}_{h-1}, a_{h-1}) \in [K] \times \{2,\ldots,H\} \times \mathcal{S} \times \overline{\Gamma}_{h-1} \times \mathcal{A}$, we have

$$\begin{aligned}&\big|\mathbb{E}_{\theta^*}[(\mathbb{B}_h^{\theta_k,\pi_k} V_{h+1}^{\theta_k,\pi_k})(\overline{\boldsymbol{\tau}}_h) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \overline{\boldsymbol{\tau}}_{h-1} = \overline{\tau}_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}]\big| \qquad (\text{G.43})\\ &\quad = \big|\mathbb{E}_{\theta^*}[V_h^{\theta_k,\pi_k}(\overline{\boldsymbol{\tau}}_h) \,|\, \boldsymbol{s}_{h-1} = s_{h-1}, \overline{\boldsymbol{\tau}}_{h-1} = \overline{\tau}_{h-1}, \boldsymbol{a}_{h-1} = a_{h-1}]\big| \le \gamma H,\end{aligned}$$

where the equality uses the definition of the value function in (3.10) and the inequality is by Lemma G.4. Similarly, by Lemma G.2, we have

$$\begin{aligned}&\big|\mathbb{E}_{\theta^*}[(\mathbb{B}_h^{\theta^*,\pi_k}V_{h+1}^{\theta_k,\pi_k})(\overline{\boldsymbol{\tau}}_h)\,|\,\boldsymbol{s}_{h-1}=s_{h-1},\overline{\boldsymbol{\tau}}_{h-1}=\overline{\tau}_{h-1},\boldsymbol{a}_{h-1}=a_{h-1}]\big| \\ &\quad=\big|\mathbb{E}_{\theta^*,\pi_k}[V_{h+1}^{\theta_k,\pi_k}(\overline{\boldsymbol{\tau}}_{h+1})\,|\,\boldsymbol{s}_{h-1}=s_{h-1},\overline{\boldsymbol{\tau}}_{h-1}=\overline{\tau}_{h-1},\boldsymbol{a}_{h-1}=a_{h-1}]\big|\le\gamma H.\end{aligned} \tag{G.44}$$

Combining (G.43) and (G.44), and using the triangle inequality, we have

$$\big|\mathbb{E}_{\theta^*}[(\mathbb{B}_h^{\theta_k,\pi_k}V_{h+1}^{\theta_k,\pi_k}-\mathbb{B}_h^{\theta^*,\pi_k}V_{h+1}^{\theta_k,\pi_k})(\overline{\boldsymbol{\tau}}_h)\,|\,\boldsymbol{s}_{h-1}=s_{h-1},\overline{\boldsymbol{\tau}}_{h-1}=\overline{\tau}_{h-1},\boldsymbol{a}_{h-1}=a_{h-1}]\big|\le 2\gamma H,$$

which, by Jensen's inequality, implies

$$e_h^k(s_{h-1})=\big|\mathbb{E}_{\theta^*,\pi_k}[(\mathbb{B}_h^{\theta_k,\pi_k}V_{h+1}^{\theta_k,\pi_k}-\mathbb{B}_h^{\theta^*,\pi_k}V_{h+1}^{\theta_k,\pi_k})(\overline{\boldsymbol{\tau}}_h)\,|\,\boldsymbol{s}_{h-1}=s_{h-1}]\big|\le 2\gamma H.$$

Therefore, we conclude the proof of Lemma G.6. □

G.7. Concentration Inequality

Lemma G.7. *Suppose that $\{\mathcal{M}_j\}_{j\ge1}$ is a martingale defined on a Hilbert space $\mathcal{H}$. For any $c>0$, if we have*

$$\sum_{j=1}^{\infty}\|\mathcal{M}_{j+1}-\mathcal{M}_j\|_{\mathcal{H}}^2\le c^2, \tag{G.45}$$

then, for any $\varepsilon>0$, it holds that

$$\mathcal{P}\Big(\sup_{j\ge1}\|\mathcal{M}_j\|_{\mathcal{H}}\ge\varepsilon\Big)\le 2\exp\Big\{-\frac{\varepsilon^2}{2c^2}\Big\}.$$

Proof. The lemma is a special case of Theorem 3.5 in (Pinelis, 1994) (see also, Theorem 3 in (Pinelis, 1992)), which is a more general result for martingales in Banach spaces. □