A Note on Boosting POMDPs Guarantees: A Framework under Belief Space Smoothness

Youheng Zhu
CS Department
University of Illinois at Urbana Champaign
email: youheng@illinois.edu
supervised by: Nan Jiang

Abstract

1 Introduction

2 Preliminaries

In this section, we introduce the model setup and algorithms for off-line POMDPs.

2.1 Infinite-horizon Discounted POMDP

An infinite-horizon discounted POMDP is a 7-tuple: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, r, \gamma, \mathbb{O}, \mathbb{T} \rangle$ where $\gamma \in [0, 1)$ is the discount factor, $\mathcal{S}$ is the latent state space, $\mathcal{A}$ is the action space, $\mathcal{O}$ is the observation space, $r : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is the bounded reward function, $\mathbb{O} : \mathcal{S} \to \Delta(\mathcal{O})$ is the emission dynamic, and $\mathbb{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition dynamic. We use $\Delta(\cdot)$ to represent a probability distribution on a specific space, and $\mathbb{O}$ and $\mathbb{T}$ respectively indicates the probability of the observation given the current state, and that of the next-state given current state-action pair. We use $|\cdot|$ to denote the cardinality of a space. For simplicity yet without loosing the essence of the idea, we consider the spaces $\mathcal{S}, \mathcal{A}, \mathcal{O}$ to be discrete and finite.

The system dynamic can be uniquely demonstrated by the following procedure:

It is important to note that in a general POMDP, the latent state space $\mathcal{S}$ is unknown to the learner, and learners only get access to the trajectories sampled by a behavior policy, which is invariant under the off-line setting.

It is also convenient for us to introduce the finite-horizon setting for POMDPs, which is majorly talked about in chapter 6. In the finite horizon settings, the discounted factor γ is set to 1, and the agent stops at step H.

2.2 Belief State Space and Smoothness Condition

Since one cannot observe the latent state directly, an overall good prediction of the current state is by using the information from the entire history of observations and actions. In chapter 4 and 5, we denote the history at time step h to be

$$\tau_h = (o_1, a_1, o_2, a_2, \cdots, o_{h-1}, a_{h-1}, o_h) \in \mathcal{H}, \tag{1}$$

In chapter 6, we denote the history at time step h to be

$$\tau_h = (o_1, a_1, o_2, a_2, \cdots, o_{h-1}, a_{h-1}) \in \mathcal{H}, \tag{2}$$

and consequently one can predict the current state given the history data. The belief state $\mathbf{b}(\tau_h) = \Pr(s_h|\tau_h)$ is an element of $\Delta(\mathcal{S}) \subset \mathbb{R}^{|\mathcal{S}|}$ when $|\mathcal{S}| < \infty$. We use $\mathcal{B}$ to denote belief state space such that

$$\mathcal{B} = \{b : \exists h \in \mathbb{N}\ \exists \tau_h, \mathbf{b}(\tau_h) = b\} \tag{3}$$

Assumption 1. $\mathbf{b} : \mathcal{H} \to \mathcal{B}$ *is an injection (and thus a bijection).*

The assumption is especially natural when considering very large latent state space and therefore very high-dimensional belief state space. Consequently, $|\mathcal{B}| = \infty$ unless we truncate an infinitely long tail from the history that we consider and bares a reasonable truncation error. In a following example and section 5.2, we will discuss the case of $|\mathcal{B}| < \infty$, which gives us some worst-case properties.

With the assumption, we denote the policy of interest $\tilde{\pi}(\tau_h) = \pi(\mathbf{b}(\tau_h)) : \mathcal{H} \to \Delta(\mathcal{A})$, which is used to sample an action when given a history.

It is easy to see that a good belief state policy should treat two similar belief state similarly, and thus should have some smoothness condition with regard to the topology of the belief state space. We typically denote the type of policy we are interested in using the following assumption.

Assumption 2 (Lipchitz of Policy)**.** $\exists L_\pi, \|\pi(b_1) - \pi(b_2)\|_1 \leq L_\pi \|b_1 - b_2\|_1$.

2.3 Off-line Data

In chapter 4 and 5, the offline dataset $\mathcal{D}$ is collected using a behavior policy $\tilde{\pi}_b$. The process involves independently collecting n sample trajectories $(o_1, a_1, \cdots)$ from the POMDP. From each trajectory, a prefix of the first h elements is truncated to form a tuple $(o_1, a_1, r_1, o_2, a_2, r_2, \cdots, o_h, a_h, r_h, o_{h+1})$ where h is randomly selected. Finally, the dataset takes the form:

$$\mathcal{D} = \{(o_1^{[i]}, a_1^{[i]}, r_1^{[i]}, o_2^{[i]}, a_2^{[i]}, r_2^{[i]}, \cdots, o_{h_i}^{[i]}, a_{h_i}^{[i]}, r_{h_i}^{[i]}, o_{h_i+1}^{[i]})\}_{i=1}^n \tag{4}$$

In chapter 6, for the future-dependent value function (FDVF), the definition of offline data differs slightly. In the FDVF setting, we consider a finite-horizon POMDP of length H.

Again, a behavior policy π_b is used to interact with the environment and collect data. This time, the entire trajectory is treated as a single data point, i.e.,

$$\mathcal{D} = \{((o_1^{[i]}, a_1^{[i]}, r_1^{[i]}, o_2^{[i]}, a_2^{[i]}, r_2^{[i]}, \cdots, o_{H-1}^{[i]}, a_{H-1}^{[i]}, r_{H-1}^{[i]}, o_H^{[i]}, a_H^{[i]}, r_H^{[i]})\}_{i=1}^n \tag{5}$$

3 Overall Analysis In a Nutshell

By assuming the Lipchitz continuity of value function w.r.t. belief state, we arrived at a PAC sample complexity guarantee using belief space coverage assumption, the general proof follows the process as demonstrated below:

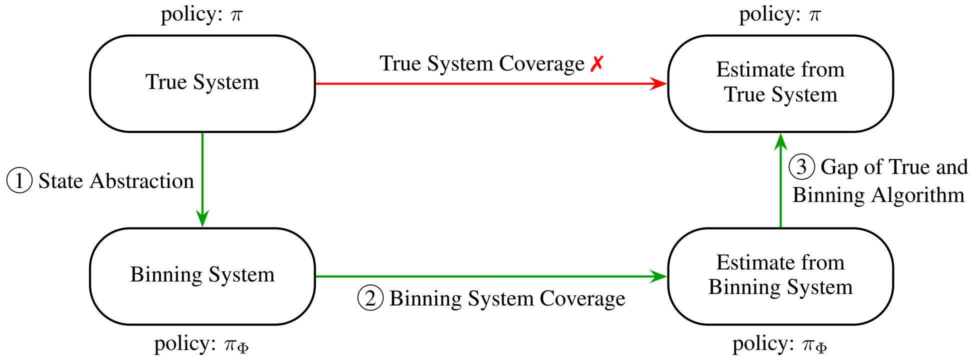

Figure 1: Pipeline of the analysis

Specifically in step 1, we descend the true belief space MDP system to a abstract binned system where the policy π is also descended to an abstract policy π_Φ. Using similar ideas of state abstraction, we bridge the abstraction gap using the size of bins ε. We also show that belief space bisimulation assumption can be replaced by Lipchitz value function using chaining techniques. In step 2, we employed the standard analysis for MDP, with the coverage assumption for the binned belief space, which can be much more tractable than the coverage of the true system due to "The Curse of Horizon". Eventually for step 3, we utilize the Lipchitz property of value function again to control the difference of the binned version and the true version of the same algorithm on the same off-line dataset. Combining all the analysis above, we put forward the final sample complexity guarantee.

4 Off-Policy Evaluation under Smooth Conditions

4.1 Abstraction under Covering

Definition 1. *A ε-cover $\mathcal{C}_\varepsilon$ is a subspace of the belief state space which satisfies:*

$$\mathcal{B} \subset \bigcup_{c \in \mathcal{C}_\varepsilon} \mathbf{B}(c, \varepsilon) \tag{6}$$

where $\mathbf{B}(c, \varepsilon)$ stands for an open ball centered at c with radius ε. The cardinality of $\mathcal{C}_\varepsilon$ is called ε-covering number. For every ε-cover $\mathcal{C}_\varepsilon$, there exist a partition of the belief state space, where each $c \in \mathcal{C}_\varepsilon$ acts as the representation element of the bin.

Building on this, we can attempt to characterize how certain important quantities behave when two belief states are sufficiently close. First, the following lemma provides a bound on the difference in expected rewards when the belief states are close.

Lemma 1. *For two belief states b_1 and b_2, $\forall a \in \mathcal{A}$, we have:*

$$|r(b_1, a) - r(b_2, a)| \leq R_{\max} \|b_1 - b_2\|_1. \tag{7}$$

For simplicity, we first look at the standard bisimulation setting.

Assumption 3 (Bisimulation)**.** *For two arbitrary belief points b_1 and b_2 in the bin represented by an element c in the ε-cover $\mathcal{C}_\varepsilon$, we have:*

$$\|\Phi P(b_1, a) - \Phi P(b_2, a)\|_1 \leq L_b \varepsilon.$$

This would put condition on the ε-cover $\mathcal{C}_\varepsilon$ and the partition it induces.

However, this kind of assumption is clearly unrealistic, and it is easy to find counterexamples. Therefore, directly replicating the method of state abstraction encounters obstacles. Before diving deeper into this discussion, we first prove the following lemma:

Lemma 2 (Contraction Property of the Belief Space)**.** *Let b_1 and b_2 be two belief states in the belief space $\mathcal{B}$. Use the following notation to represent the next belief state:*

$$b^{o,a} = \mathbf{b}(\mathbf{b}^{-1}(b) + a + o)$$

where $+$ denotes concatenation. In this study, we also use b^{+1} as a shorthand for $b^{o,a}$ when a specific (o, a) pair is not emphasized. Similarly, $b^{+2}, b^{+3}, \cdots$ follow the same notation rules.

The contraction lemma states that for all o, a,

$$\|b_1^{o,a} - b_2^{o,a}\|_1 \leq \eta \|b_1 - b_2\|_1$$

for some uniform $\eta \in (0, 1]$.

Lemma 3. *(Lemma 2 in [2]) For any two belief points b_1, b_2 satisfying $\|b_1 - b_2\|_1 \leq \varepsilon$, $|P(o|b_1, a) - P(o|b_2, a)| \leq \|b_1 - b_2\|_1 \leq \varepsilon$.*

Consequently, we put forward the following proposition.

Proposition 1. *For policy π satisfying Assumption 2, we have for $\forall o, a$*

$$|P(b_1^{o,a}|b_1) - P(b_2^{o,a}|b_2)| \leq (1 + L_\pi)\|b_1 - b_2\|_1. \tag{8}$$

Here the bound is somehow loose by a $|\mathcal{O}||\mathcal{A}|$ factor. A tighter result for Proposition 1 which will be useful later is

Proposition 2. *For any $o \in \mathcal{O}$,*

$$\left| \sum_{a \in \mathcal{A}} (P(b_1^{o,a}|b_1) - P(b_2^{o,a}|b_2)) \right| \leq (1 + L_\pi)\|b_1 - b_2\|_1. \tag{9}$$

The one-step error is easy to control, however, without bisimulation, it is extremely difficult to control the accumulative error induced by infinite amount of steps. With that said, we can try to obtain an upper bound for k-step transition error:

Proposition 3. $|P(b_1^{+k}|b_1) - P(b_2^{+k}|b_2)| \leq (1 + L_\pi)(1 + \eta)^k \|b_1 - b_2\|_1$

With Proposition 3, we can try to control the error propagated from the initial belief space difference:

$$\begin{aligned}
&|\mathbb{E}_{a,o,\cdots\sim b_1}\left[r(b_1^{+k}, a_k)\right] - \mathbb{E}_{a,o,\cdots\sim b_2}\left[r(b_1^{+k}, a_k)\right]| \\
&= |\langle (P(b_1^{[\cdot]}|b_1) - P(b_2^{[\cdot]}|b_2)), r(b_1^{[\cdot]}, a) \rangle| \qquad (10) \\
&\leq \|P(b_1^{+k}|b_1) - P(b_2^{+k}|b_2)\|_1 \cdot R_{\max}. \qquad (11)
\end{aligned}$$

Since we only get an L_∞ norm in Proposition 3, extending it to L_1 would need an extra $(|\mathcal{O}||\mathcal{A}|)^k$ expenses , making the error propagation explode drastically, not to mention that our horizon would go to infinity. Unless $\gamma < 1/(1 + \eta)|\mathcal{O}||\mathcal{A}|$, the error would be impossible to control in this analysis. This result also indicates that to control the propagation of error, additional assumptions are necessary—properties of the POMDP belief space alone are insufficient. As discussed previously, although double simulation can help, it still requires a new assumption to be effective. The most direct approach is to assume that error propagation is controllable. Another possible assumption is that the value function V satisfies a Lipschitz condition. The relationship between these two approaches will be discussed later.

Assumption 4 (Error Propagation Control)**.**

$$\exists L_H,\ \forall k,\ |\mathbb{E}_{a,o,\cdots\sim b_1}\left[r(b_1^{+k}, a_k)\right] - \mathbb{E}_{a,o,\cdots\sim b_2}\left[r(b_1^{+k}, a_k)\right]| \leq L_H R_{\max}\|b_1 - b_2\|_1.$$

Assumption 5 (Lipchitz of Value Function)**.**

$$\begin{aligned}
\exists L_V, |V^\pi(b_1) - V^\pi(b_2)| &\leq L_V\|b_1 - b_2\|_1 \\
|V^{\pi_\phi}(b_1) - V^{\pi_\phi}(b_2)| &\leq L_V\|b_1 - b_2\|_1
\end{aligned}$$

With these preparations, we can now begin to control the value function error between the abstract belief space and the true belief space. The following theorem shows that for any abstract policy π_ϕ satisfying the Lipschitz property, the error between its lifted value function in the abstract space and the value function of its lifted policy in the real space (in the ∞-norm sense) is controllable.

Theorem 1. *For a policy π satisfying the Lipschitz assumption 2, let π_ϕ be its abstracted version. Under Assumption 4, the following holds:*

$$\|[\tilde{V}_{\text{bin}}^{\pi_\phi}]_{\text{true}} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq \frac{L_H R_{\max}\varepsilon}{1 - \gamma} + \frac{R_{\max}\varepsilon}{1 - \gamma\eta} \tag{12}$$

Corollary 1. *Under Assumption 4 and Assumption 5, the following relation holds:*

For a belief space MDP, if it satisfies L_H-error propagation control, then its value function is L_V-Lipschitz continuous, where

$$L_V = \frac{R_{\max}}{1 - \gamma} L_H + \frac{R_{\max}}{1 - \gamma\eta}$$

Proof. This proof is almost identical to the proof of Theorem 1, and can be referenced there. □

Remark 1. *From the above corollary, it can be seen that the Lipschitz continuity assumption for value functions (Assumption 5) is actually weaker than the error propagation control assumption (Assumption 4), although the latter is more direct based on the prior analysis. In the remainder of this work, we mainly adopt the value function Lipschitz continuity assumption 5.*

Corollary 2. *For a policy π satisfying the Lipschitz assumption 2, and its abstracted version π_ϕ, under Assumption 5, we have the following bound:*

$$\|[\tilde{V}_{\text{bin}}^{\pi_\phi}]_{\text{true}} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq L_V\varepsilon \tag{13}$$

After obtaining Theorem 1, a natural follow-up question is whether the $\tilde{V}_{\text{bin}}^{\pi_\phi}$ in the theorem is the same as the true value function $V_{\text{bin}}^{\pi_\phi}$ of the abstract MDP. In fact, the two are not the same—$\tilde{V}_{\text{bin}}^{\pi_\phi}$ is at best an intermediate value rather than the actual value function within the abstract MDP.

In the abstract MDP, the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ is mapped by the abstraction ϕ to $(\mathcal{S}_\phi, \mathcal{A}, P_\phi, R_\phi, \gamma)$, which means that the transition dynamics P_ϕ in the abstract MDP are induced by the original MDP.

Specifically, the induced P_ϕ satisfies that there exists a family of probability measures $\{p_x\}_{x \in \mathcal{S}_\phi}$, where each p_x is defined on $\phi^{-1}(x)$, such that the transition probability from $\phi(s)$ to $\phi(s')$ under action a, namely, $P_\phi(\phi(s')|\phi(s), a)$ in the abstract MDP can be written as:

$$P_\phi(\phi(s')|\phi(s), a) = \mathbb{E}_{s \sim p_{\phi(s)}}[P_\phi(\phi(s')|s, a)] \tag{14}$$

Because this characterization of P_ϕ relies on the existence of such a family of probability measures without specifying their exact properties, any proof involving the value function $V_{\text{bin}}^{\pi_\phi}$ must treat the $\{p_x\}_{x \in \mathcal{S}_\phi}$ as arbitrary.

With this understanding, we now present the following theorem, which provides an upper bound on the error between $\tilde{V}_{\text{bin}}^{\pi_\phi}$ and the true value function $V_{\text{bin}}^{\pi_\phi}$ in the abstract MDP. Importantly, the proof of this theorem does not rely on the specific form of the measures $\{p_x\}_{x \in \mathcal{S}_\phi}$.

Theorem 2. *If Assumption 5 holds, then the error between $\tilde{V}_{\text{bin}}^{\pi_\phi}$ and the abstract MDP's true value function $V_{\text{bin}}^{\pi_\phi}$ can be bounded as follows:*

$$\|\tilde{V}_{\text{bin}}^{\pi_\phi} - V_{\text{bin}}^{\pi_\phi}\|_\infty \leq \frac{R_{\max}\varepsilon}{1-\gamma} + \frac{R_{\max} L_V \varepsilon}{(1-\gamma)^2} \tag{15}$$

Before ending this part, we'll need to fill the gap between the target policy and the abstracted policy to which the target policy descended. This is handled by the following theorem.

Theorem 3. *If Assumptions 1, 2, and 5 hold, then the following two inequalities hold simultaneously.*

$$\|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq \frac{R_{\max} L_\pi \varepsilon}{1-\gamma} + 2|\mathcal{O}||\mathcal{A}|\frac{\gamma R_{\max}}{(1-\gamma)^2}(1 + L_\pi)\varepsilon, \tag{16}$$

$$\|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq \frac{R_{\max} L_\pi \varepsilon}{1-\gamma} + |\mathcal{O}|\frac{\gamma R_{\max}}{(1-\gamma)^2}(1 + L_\pi)\varepsilon + |\mathcal{O}||\mathcal{A}|\frac{\gamma(1 + L_\pi)L_V\varepsilon^2}{1-\gamma}. \tag{17}$$

4.2 Algorithm on Belief Space MDP

Consider a Bellman error minimization algorithm using double sampling, whose optimization target can be written as

$$\hat{Q}^\pi = \arg\min_{f \in \mathcal{F}} \mathcal{E}(f, \pi) \tag{18}$$

where

$$\mathcal{E}(f, \pi) = \mathbb{E}_{\mathcal{D}}[(f(b, a) - (r + \gamma f(b'_A, \pi)))(f(b, a) - (r + \gamma f(b'_B, \pi)))]. \tag{19}$$

On the abstracted space, using the same data, the algorithm becomes

$$\hat{Q}_\phi^\pi = \arg\min_{f \in \mathcal{F}} \mathcal{E}_\phi(f, \pi) \tag{20}$$

where

$$\mathcal{E}_\phi(f, \pi) = \mathbb{E}_{\mathcal{D}}[(f(\phi(b), a) - (r_\phi + \gamma f(\phi(b'_A), \pi_\phi)))(f(\phi(b), a) - (r_\phi + \gamma f(\phi(b'_B), \pi_\phi)))]. \tag{21}$$

Despite the algorithm was computed in the true system, we consider a virtually executed algorithm, and adopts the following standard covering assumption.

Assumption 6 (Binned Policy Coverage)**.** $\|d^{\pi_\phi}/d^{\mathcal{D}}\|_\infty \leq C_\pi(\phi) < \infty$

Remark 2. *It is worth noting that the coverage $C_\pi(\phi)$ here depends on the specific abstraction mapping ϕ. Under the most exploratory data collection distribution d^D, the worst-case growth rate of $C_\pi(\phi)$ is approximately aligned with $|\mathcal{C}_\varepsilon|$, which denotes the ε-covering number.*

The benefit of the belief-policy coverage assumption 6 lies in its potential to outperform coverage assumptions in the original space. In the original (real) space, the value of the coverage assumption can grow exponentially, leading to the curse of history. In contrast, using an abstract belief space allows the exponentially large historical space to be reduced to a space whose size is just the ε-covering number. This makes better use of the smooth structural properties and relaxes the assumptions required.

Lemma 4. *In the binned system, we have the following telescoping error*

$$|J_{\hat{Q}_\phi^\pi}(\pi_\phi) - J(\pi_\phi)| \le \frac{\sqrt{C_\pi(\phi)}}{1-\gamma} \cdot \sqrt{\mathbb{E}_{d^{\mathcal{D}}}[(\hat{Q}_\phi^\pi - \mathcal{T}^{\pi_\phi}\hat{Q}_\phi^\pi)^2]} \tag{22}$$

And we obviously have

$$\mathbb{E}_{d^{\mathcal{D}}}[\mathcal{E}_\phi(\hat{Q}_\phi^\pi, \pi)] = \mathbb{E}_{d^{\mathcal{D}}}[(\hat{Q}_\phi^\pi - \mathcal{T}^{\pi_\phi}\hat{Q}_\phi^\pi)^2] \tag{23}$$

As the size of independent samples grows, the difference between the empirical estimate and the true expectation of the value above becomes closer, whose convergence speed can be characterized using concentration inequalities such as Hoeffding's or Bernstein's inequality. Using Hoeffding's inequality, we get the following lemma.

Lemma 5. *With probability at least $1-\delta$, for $\forall f \in \mathcal{F}$,*

$$|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^{\mathcal{D}}}[\mathcal{E}_\phi(f,\pi)]| \le \sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}} \tag{24}$$

In fact, Hoeffding's inequality only leverages the boundedness of the function. However, by introducing the Bellman completeness assumption below, we can also take the variance of the function into account and apply Bernstein's inequality to achieve a tighter convergence rate.

And the standard Bellman completeness assumption is as below:

Assumption 7 (Bellman Completeness)**.** $\forall f \in \mathcal{F}, \mathcal{T}^{\pi_\phi} f \in \mathcal{F}$.

Proposition 4. *Under the Bellman completeness assumption 7, we can obtain an upper bound with $O(1/n)$ convergence rate. Specifically, with probability at least $1-\delta$, for all $f \in \mathcal{F}$, the following holds:*

$$|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^{\mathcal{D}}}[\mathcal{E}_\phi(f,\pi)]| \lesssim \frac{R_{\max}^2}{n(1-\gamma)^2} \cdot \log\frac{|\mathcal{F}|}{\delta} \tag{25}$$

The detailed proof is left as an exercise for interested readers.

Of course, for the purpose of this discussion, the Bellman completeness assumption is not necessary—only the following realizability assumption is needed to achieve the goal. However, in this case, we can only characterize the concentration rate using Lemma 5 derived from Hoeffding's inequality, and cannot use the tighter concentration rate provided by Proposition 4.

Assumption 8 (Realizability)**.** $Q_\phi^\pi \in \mathcal{F}$.

Remark 3. *For a finite function space where $|\mathcal{F}| < \infty$, the Bellman completeness assumption 7 implies the realizability assumption 8.*

From the previous two lemmas, we have the following proposition.

Proposition 5. *If Assumption 8 holds, then with probability at least $1-\delta$, the following inequality holds:*

$$|\mathbb{E}_{d^{\mathcal{D}}}[\mathcal{E}_\phi(\hat{Q}_\phi^\pi, \pi)]| \le \sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}} \tag{26}$$

And consequently the following theorem.

Theorem 4. *Under Assumption 8,*

$$|J_{\hat{Q}_\phi^\pi}(\pi_\phi) - J(\pi_\phi)| \le \frac{\sqrt{C_\pi(\phi)}}{1-\gamma} \cdot \left(\frac{32R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}\right)^{\frac{1}{4}} \tag{27}$$

Proof. From Proposition 5 and Lemma 4, and noting Equation (23), the conclusion follows directly. □

This theorem shows that in the abstract belief MDP, the abstract algorithm can produce an approximately optimal function $\hat{Q}^{\pi}_{\phi}$. Compared to the true optimal Q^{π}_{ϕ}, the expected return difference between them can be controlled at the rate of $O(n^{-1/4})$. If the Bellman completeness assumption 7 is adopted instead of the realizability assumption 8, this rate can be improved to $O(n^{-1/2})$.

[**Youheng:** Potentially, the analysis can be extended to other algorithms such as double-robust or MIS, and the analysis will also be quite standard, so I'll keep it this way.]

4.3 Gap Between True Algorithm and Virtual Algorithm

Finally, since the agent operates in the real POMDP—i.e., the real belief MDP—rather than in the abstracted belief MDP, even when using the same sampling data, the two algorithms will inherently differ. The following will discuss how this difference can be controlled.

Noticed that we previously assumed the Lipchitz continuity of value function, whose equivalence to the Lipchitz continuity of Q-function at action a can be easily proven. We now assume the function class $\mathcal{F}$ we use to approximate Q-function is also Lipchitz with regard to belief state.

Assumption 9 (Lipchitz of Function Class)**.** $\exists L_Q$, $\forall f \in \mathcal{F}$, $\forall a \in \mathcal{A}$, $|f(b_1, a) - f(b_2, a)| \leq L_Q \|b_1 - b_2\|_1$.

Remark 4. *Note that the previous discussion assumed the function class satisfies the realizability assumption 8. Therefore, a necessary condition for Assumption 9 to coexist with it is that $L_Q \geq L_V$.*

With the assumption on the function class, we can therefore control the differences between $\mathcal{E}_\phi(f, \pi)$ and $\mathcal{E}(f, \pi)$ for the very same fixed $f \in \mathcal{F}$, which is stated in the lemma.

Lemma 6. *If assumption 9 holds, then*

$$|\mathcal{E}(f, \pi) - \mathcal{E}_\phi(f, \pi)| \leq \frac{4R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon \tag{28}$$

Then, we look at how $\hat{Q}^{\pi}$ and $\hat{Q}^{\pi}_{\phi}$ differs on the very same empirical bellman error $\mathcal{E}_\phi(\cdot, \pi)$.

Theorem 5. *If the Lipschitz assumption for the function class 9 holds, then we have:*

$$|\mathcal{E}_\phi(\hat{Q}^{\pi}, \pi) - \mathcal{E}_\phi(\hat{Q}^{\pi}_{\phi}, \pi)| \leq \frac{8R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon \tag{29}$$

To put things together, we have

Proposition 6. *If assumption 8 , 9 holds, then*

$$|\mathbb{E}_{d^{\mathcal{D}}}[\mathcal{E}_\phi(\hat{Q}^{\pi}, \pi)]| \leq \sqrt{\frac{2R_{\max}^2}{n(1-\gamma)^2} \cdot \log\frac{2|\mathcal{F}|}{\delta}} + \frac{8R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon \tag{30}$$

Proof. It is an direct result from proposition 5 and theorem 5. □

And consequently,

Theorem 6. *If assumption 8, 9 hold, then*

$$|J_{\hat{Q}^{\pi}}(\pi_\phi) - J(\pi_\phi)| \leq \frac{\sqrt{C_\pi(\phi)}}{1-\gamma} \cdot \sqrt{\sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}} + \frac{8R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon} \tag{31}$$

Proof. The result follows directly from proposition 6, lemma 4, and (23). Note that when applying Lemma 4 and Equation (23), $\hat{Q}^{\pi}_{\phi}$ should be replaced with $\hat{Q}^{\pi}$. □

The following series of theorems are all preparatory steps toward ultimately controlling the overall error.

Theorem 7. *If Assumptions 1,2, and5 hold, then we have:*

$$|J(\pi_\phi) - J(\pi)| \leq \frac{(L_\pi + 1)R_{\max}\varepsilon}{1-\gamma} + L_V\varepsilon + \frac{R_{\max}L_V\varepsilon}{(1-\gamma)^2} + 2|\mathcal{O}||\mathcal{A}|\frac{\gamma R_{\max}}{(1-\gamma)^2}(1+L_\pi)\varepsilon \tag{32}$$

Corollary 3. *If Assumptions 1,2, and5 hold, then the following inequality holds:*

$$\begin{aligned}|J(\pi_\phi) - J(\pi)| \leq{}& \frac{(L_\pi + 1)R_{\max}\varepsilon}{1-\gamma} + L_V\varepsilon + \frac{R_{\max}L_V\varepsilon}{(1-\gamma)^2} \\ &+ |\mathcal{O}|\frac{\gamma R_{\max}}{(1-\gamma)^2}(1+L_\pi)\varepsilon + |\mathcal{O}||\mathcal{A}|\frac{\gamma(1+L_\pi)L_V\varepsilon^2}{1-\gamma}\end{aligned} \tag{33}$$

Theorem 8. *If assumption 9 holds, then*

$$|J_{\hat{Q}^\pi}(\pi) - J_{\hat{Q}^\pi}(\pi_\phi)| \leq \frac{R_{\max}}{1-\gamma}\varepsilon + L_Q\varepsilon \tag{34}$$

Finally, by combining the above theorems, we arrive at the following conclusion:

Theorem 9. *If Assumptions 8,9,1,2, and5 all hold, then we have:*

$$\begin{aligned}|J_{\hat{Q}^\pi}(\pi) - J(\pi)| \leq{}& \frac{\sqrt{C_\pi(\varepsilon)}}{1-\gamma}\cdot\sqrt{\sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}} + \frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon} \\ &+ \frac{(L_\pi + 1)R_{\max}\varepsilon}{1-\gamma} + L_V\varepsilon + \frac{R_{\max}L_V\varepsilon}{(1-\gamma)^2} \\ &+ 2|\mathcal{O}||\mathcal{A}|\frac{\gamma R_{\max}}{(1-\gamma)^2}(1+L_\pi)\varepsilon + L_Q\varepsilon\end{aligned} \tag{35}$$

This completes the full theoretical analysis for the double sampling algorithm. A similar analysis can be performed for the doubly robust and marginal importance sampling methods. This section has already laid the groundwork for that—indeed, once their finite sample guarantees corresponding to (4) are given, the remaining steps in the analysis are nearly identical.

For a sample complexity guarantee, one can first decide with probability $1-\delta$, the entire error should be below ϵ, then one can set an appropriate ε so that the error ϵ can be balanced onto the terms with ε, and the terms with $1/n$. Take a simple example, suppose

$$\sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}} = 3\cdot\frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon. \tag{36}$$

Then one may solve the equation

$$\begin{aligned}\epsilon ={}& \frac{2\sqrt{C_\pi(\varepsilon)}}{1-\gamma}\cdot\sqrt{\frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon} + \frac{(L_\pi + 1)R_{\max}\varepsilon}{1-\gamma} \\ &+ L_V\varepsilon + \frac{R_{\max}L_V\varepsilon}{(1-\gamma)^2} + 2|\mathcal{O}||\mathcal{A}|\frac{\gamma R_{\max}}{(1-\gamma)^2}(1+L_\pi)\varepsilon + L_Q\varepsilon\end{aligned} \tag{37}$$

for the optimal $\varepsilon(\epsilon)$. After that, putting $\varepsilon(\epsilon)$ into (36) and one can solve the sample complexity n.

5 Quick-Forgetting Function Class and Finite-Horizon Guarantees

5.1 Quick-Forgetting Function Class: An Example

In this section, we consider an extended version of Lemma 2. The assumption states that there exists a constant $\eta' > 0$ such that:

$$\eta'\|b_1 - b_2\|_1 \leq \|b_1^{o,a} - b_2^{o,a}\|_1 \leq \eta\|b_1 - b_2\|_1$$

A possible example of a possible Lipchitz function class is the Quick-Forgetting function class $\mathcal{F}_q^{[m]}$, so that for $\forall f \in \mathcal{F}_q^{[m]}$,

$$f : \mathcal{H} \rightarrow [0, \frac{R_{\max}}{1-\gamma}] \tag{38}$$

$$f : \tau_h \mapsto V(\tau_{h-m:h}) \tag{39}$$

and

$$\forall \tau_h^{[1]}, \tau_h^{[2]},\ |f(\tau_h^{[1]}) - f(\tau_h^{[2]})| \leq L_F \cdot \eta_e^h \ \ s.t.\ \tau_{h-m:h}^{[1]} = \tau_{h-m:h}^{[2]} \tag{40}$$

for some $\eta_e \leq \eta'$.

In the sense that two distinct belief states will be close to each other after the same amount of history $\tau_{h-m:m}$ length m, by a factor of η^m, it is natural that the corresponding value function will be close enough under our assumption. Thus, by mapping two histories with the same m-step tail to the same value will be a good approximation.

With that said, we would also like to check out the Lipchitz guarantee that $\mathcal{F}_q^{[m]}$ provides. We first have the following lemma, which indicates the smoothness of $\mathcal{F}_q^{[m]}$ on the true belief states.

Lemma 7. *$\forall f \in \mathcal{F}_q^{[m]}$, we have*

$$\|f\|_{\text{lip}} = \max_{\substack{h_1,h_2 \in \mathcal{H} \\ h_1 \neq h_2}} \frac{|f(h_1) - f(h_2)|}{\|\mathbf{b}(h_1) - \mathbf{b}(h_2)\|_1} \leq F_m < \infty \tag{41}$$

for some uniform F_m.

5.2 Finite-Horizon Guarantees

5.2.1 Lipchitz Guarantees

For the finite horizon POMDP, any value function class is guaranteed to be Lipchitz under Assumption 1 with regard to some worst Lipchitz value.

This is because

$$\|f\|_{\text{lip}} = \max_{\substack{h_1,h_2 \in \mathcal{H} \\ h_1 \neq h_2}} \frac{|f(h_1) - f(h_2)|}{\|\mathbf{b}(h_1) - \mathbf{b}(h_2)\|_1} \tag{42}$$

is a finite number given that $\mathcal{B}$ is a finite state.

5.2.2 Covering Number Guarantees

For finite $\mathcal{B}$, the covering number is upper bounded by $|\mathcal{B}|$. However this could be exponential.

6 Future-Dependent Value Functions

In this section, we extend the proposed methodology and framework to future-dependent value functions (FDVFs). Due to space constraints and considerations of scope, we provide only a high-level theoretical analysis of FDVFs under this framework, without delving into the strengths or weaknesses of the assumptions involved. Interested researchers are encouraged to explore these issues further in future work. Moreover, some of the theorem proofs are omitted due to length limitations; readers are welcome to attempt the proofs themselves as exercises.

6.1 The FDVF Algorithm under Memory-Based Policies

When FDVF was first introduced, its primary purpose was to address the offline estimation problem under memoryless policies. The focus on memoryless policies stems from the so-called "curse of memory," as discussed in the concluding remarks of [1]. In short, the algorithm requires modifications to work under memory-based policies and suffers from the curse of memory in the absence of any assumptions. We begin by revising the definitions of FDVF and the Bellman residual operator introduced in Chapter 2, and we also introduce a new error control lemma.

Definition 2 (New Future-Dependent Value Function)**.** *Under memory-based conditions, define the future-dependent value function $V_{\mathcal{F}}$ as a function $V : \mathcal{F}'_h \times \mathcal{H}'_h \to \mathbb{R}$ that satisfies:*

$$\mathbb{E}_{\pi_b}[V(f_h, \tau_h)|s_h, \tau_h] = V_{\mathcal{S}}^{\pi_e}(s_h, \tau_h) \tag{43}$$

Remark 5. *Why do we need to modify the definition of FDVF here? This is because, in the original definition of FDVF for memoryless policies, s_h could serve as an information bottleneck. When extended to memory-based settings, we can choose to incorporate the historical sequence $\mathcal{H}$ in place of just s_h.*

However, this introduces a major complication—FDVF, which previously existed, may no longer exist.

In other words, the existence of FDVF is determined by the condition:

$$\mathcal{M}_{\mathcal{F},h} \times \mathcal{V}_{\mathcal{F},h} = V_{\mathcal{S},h}^{\pi_e}$$

In the original case where $|\mathcal{S}| \ll |\mathcal{F}_h|$, this underdetermined system has a solution, and it is not unique. But in the memory-based case, as the history grows, there will come a step h where $|\mathcal{S} \times \mathcal{H}_h| > |\mathcal{F}_h|$. Without a low-rank assumption, this linear system becomes unsolvable, and FDVF ceases to exist.

Without existence, the realizability assumptions required by double robustness and marginal importance sampling break down, leading to the failure of the algorithm.

The above definition effectively expands $\mathcal{F}$ so that each history τ_h has a unique sequence of (future-history) ordered pairs. It can be easily proven that under this construction, FDVF always exists.

In addition, one can define future-dependent value functions for policies with a time window T.

Definition 3 (New Future-Dependent Value Function (with Time Window))**.** *Under memory-based conditions, define the future-dependent value function $V_{\mathcal{F}}$ as a function $V : \mathcal{F}'_h \times \mathcal{H}'_{\max(h-H+1,0):h} \to \mathbb{R}$ that satisfies:*

$$\mathbb{E}_{\pi_b}[V(f_h, \tau_{[\max(h-H+1,0):h]})|s_h, \tau_{[\max(h-H+1,0):h]}] = V_{\mathcal{S}}^{\pi_e}(s_h, \tau_{[\max(h-H+1,0):h]}) \tag{44}$$

where $\tau_{[h-t+1:h]}$ denotes the last t elements truncated from the history trajectory.

Remark 6. *It can be seen that when the time window $T = 1$, this reduces to the standard FDVF case.*

From this point forward, for convenience, we will write (f_h, τ_h) simply as f_h. Similarly, we will treat $\mathcal{F}'$ as the original future space, and define $\mathcal{F} := \mathcal{F}' \times \mathcal{H}'$ as the new space of "(future-history) pairs." This is because τ_h can be considered a part of the extended future—or equivalently, the future is duplicated separately for each history sequence.

Definition 4 (New Bellman Residual Operator)**.**

$$\begin{aligned}(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h, \tau_{[h-T+1:h]}) := \mathbb{E}_{\substack{a_{1:h}\sim\pi_e \\ a_{h+1:H}\sim\pi_b}}[r_h + V(f_{h+1})|s_h, \tau_{[h-T+1:h]}] \\ - \mathbb{E}_{\substack{a_{1:H-1}\sim\pi_e \\ a_{h:H}\sim\pi_b}}[V(f_h)|s_h, \tau_{[h-T+1:h]}]\end{aligned} \tag{45}$$

$$(\mathcal{B}^{\mathcal{H}}V)(\tau_h) := \mathbb{E}_{\substack{a_{1:h}\sim\pi_e \\ a_{h+1:H}\sim\pi_b}}[r_h + V(f_{h+1})|\tau_h] - \mathbb{E}_{\substack{a_{1:h-1}\sim\pi_e \\ a_{h:H}\sim\pi_b}}[V(f_h)|\tau_h] \tag{46}$$

Lemma 8. *For any behavior policy π_b and target policy π_e—both of which may be memory-based—and any function $V : \mathcal{F} \to \mathbb{R}$, we have:*

$$J(\pi_e) - \mathbb{E}_{\pi_b}[V(f_1)] = \sum_{h=1}^{H} \mathbb{E}_{\tau_h\sim\pi_e}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)] \tag{47}$$

Proof. The proof of this lemma is almost identical to that of Lemma **??**. For the right-hand side of the equation, we can write:

$$\begin{aligned}&\sum_{h=1}^{H} \mathbb{E}_{\substack{a_{1:h-1}\sim\pi_e \\ a_{h:H}\sim\pi_b}}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)] \\ =& \sum_{h=1}^{H} \mathbb{E}_{\substack{a_{1:h-1}\sim\pi_e \\ a_{h:H}\sim\pi_b}}\left[\mathbb{E}_{\substack{a_{1:h}\sim\pi_e \\ a_{h+1:H}\sim\pi_b}}[r_h + V(f_{h+1})|\tau_h] - \mathbb{E}_{\substack{a_{1:h-1}\sim\pi_e \\ a_{h:H}\sim\pi_b}}[V(f_h)|\tau_h]\right]\end{aligned} \tag{48}$$

And we have:

$$\mathbb{E}_{\substack{a_{1:h-1}\sim\pi_e\\ a_{h:H}\sim\pi_b}}\left[\mathbb{E}_{\substack{a_{1:h}\sim\pi_e\\ a_{h+1:H}\sim\pi_b}}[V(f_{h+1})|\tau_h]\right] = \mathbb{E}_{\substack{a_{1:h}\sim\pi_e\\ a_{h+1:H}\sim\pi_b}}\left[\mathbb{E}_{\substack{a_{1:h}\sim\pi_e\\ a_{h+1:H}\sim\pi_b}}[V(f_{h+1})|\tau_h]\right] \tag{49}$$

$$= \mathbb{E}_{\substack{a_{1:h}\sim\pi_e\\ a_{h+1:H}\sim\pi_b}}[V(f_{h+1})] \tag{50}$$

$$= \mathbb{E}_{\substack{a_{1:h}\sim\pi_e\\ a_{h+1:H}\sim\pi_b}}\left[\mathbb{E}_{\substack{a_{1:h}\sim\pi_e\\ a_{h+1:H}\sim\pi_b}}[V(f_{h+1})|\tau_{h+1}]\right] \tag{51}$$

The last equality is clearly by the definition of conditional expectation, and the first equality holds because, under both $\mathbb{E}_{\substack{a_{1:h-1}\sim\pi_e\\ a_{h:H}\sim\pi_b}}[\cdot]$ and $\mathbb{E}_{\substack{a_{1:h}\sim\pi_e\\ a_{h+1:H}\sim\pi_b}}[\cdot]$, the random variable τ_h has the same pushed-forward distribution on $\mathbf{T}_h$.

From this, by the chain rule, all cross terms cancel out, thereby proving the lemma. □

Remark 7. *From Lemma 8 and the earlier Lemma* **??**, *one can observe that regardless of whether the Bellman residual operator is conditioned on s_h or τ_h, the proof actually relies on a single condition: the random variable used as the conditioning variable must have the same pushed-forward probability distribution under both $\mathbb{E}_{\substack{a_{1:h-1}\sim\pi_e\\ a_{h:H}\sim\pi_b}}[\cdot]$ and $\mathbb{E}_{\substack{a_{1:h}\sim\pi_e\\ a_{h+1:H}\sim\pi_b}}[\cdot]$. This means that if one uses the single-step observation o_h as the conditioning variable to define a new Bellman residual operator, the same type of error control lemma still holds.*

Lemma 9. *For any behavior policy π_b and evaluation policy π_e—both of which may have memory of length up to T—and any function $V : \mathcal{F} \to \mathbb{R}$, we have:*

$$J(\pi_e) - \mathbb{E}_{\pi_b}[V(f_1)] = \sum_{h=1}^{H} \mathbb{E}_{s_h,\tau_h\sim\pi_e}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h, \tau_{[h-T+1:h]})] \tag{52}$$

Remark 8. *Clearly, Lemma 9 is a direct extension of Lemma* **??**. *Furthermore, when the length of the time window T is set to* 0, *Lemma 9 immediately degenerates to the memoryless policy case covered by Lemma* **??**.

Memory-Based Doubly Robust Algorithm: For memory-based policies, the doubly robust algorithm is defined as follows:

$$\hat{V}_{\mathcal{F}} = \arg\min_{V\in\mathcal{V}} \max_{\theta\in\Theta} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{D}}[\{\mu(a_h, \tau_h^+)(r_h + V(f_{h+1})) - V(f_h)\}\theta(\tau_h) - \frac{1}{2}\theta(\tau_h)^2] \tag{53}$$

where $\tau_h^+ := (\tau_h, o_h)$, and $\mu(a_h, \tau_h^+) := \frac{\pi_e(a_h|\tau_h^+)}{\pi_b(a_h|\tau_h^+)}$. The correctness of the algorithm can be proven with the following theorem:

Theorem 10. *If $(\mathcal{B}^{\mathcal{H}}V)(\tau_h) \in \Theta$, then*

$$\max_{\theta\in\Theta} \sum_{h=1}^{H} \mathbb{E}_{\pi_b}[\{\mu(a_h, \tau_h^+)(r_h + V(f_{h+1})) - V(f_h)\}\theta(\tau_h) - \frac{1}{2}\theta(\tau_h)^2]$$

$$= \frac{1}{2}\sum_{h=1}^{H} \mathbb{E}_{\pi_b}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2] \tag{54}$$

This means that, as the number of samples tends to infinity, we can obtain an accurate estimate of $\mathbb{E}_{\pi_b}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]$. This is precisely the objective for which the estimator is designed.

Memory-Based Marginal Importance Sampling: For memory-based policies, the marginal importance sampling algorithm is given by:

$$\hat{V}_{\mathcal{F}} = \arg\min_{V\in\mathcal{V}} \max_{w\in\mathcal{W}} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{D}}[w(\tau_h) \cdot \{\mu(a_h, \tau_h^+)(r_h + V(f_{h+1})) - V(f_h)\}] \tag{55}$$

where $\mu(a_h, \tau_h^+) := \frac{\pi_e(a_h|\tau_h^+)}{\pi_b(a_h|\tau_h^+)}$. The correctness of the algorithm can be proven with the following theorem:

Theorem 11. *Suppose the realizability conditions hold, i.e., $V_{\mathcal{F}} \in \mathcal{V}, w^\star \in \mathcal{W}$, where $w^\star$ satisfies:*

$$\mathbb{E}_{\pi_b}[w^\star(\tau_h)(\mathcal{B}^{\mathcal{H}}V)(\tau_h)] = \mathbb{E}_{\pi_e}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)]$$

Then we have:

$$V_{\mathcal{F}} = \arg\min_{V \in \mathcal{V}} \max_{w \in \mathcal{W}} \sum_{h=1}^{H} \mathbb{E}_{\pi_b}[w(\tau_h) \cdot \{\mu(a_h, \tau_h^+)(r_h + V(f_{h+1})) - V(f_h)\}] \tag{56}$$

and

$$\max_{w \in \mathcal{W}} \sum_{h=1}^{H} \mathbb{E}_{\pi_b}[w(\tau_h) \cdot \{\mu(a_h, \tau_h^+)(r_h + V_{\mathcal{F}}(f_{h+1})) - V_{\mathcal{F}}(f_h)\}] = 0 \tag{57}$$

The proof is left as an exercise for interested readers.

6.2 FDVF Analysis Pipeline

As shown in Figure 2, the analysis of FDVF follows a structured framework that uses the previously introduced methodology: "abstraction, theoretical guarantees over the abstract space, and transition from abstract algorithm to the real algorithm." The detailed procedures will be introduced step-by-step in the following subsections.

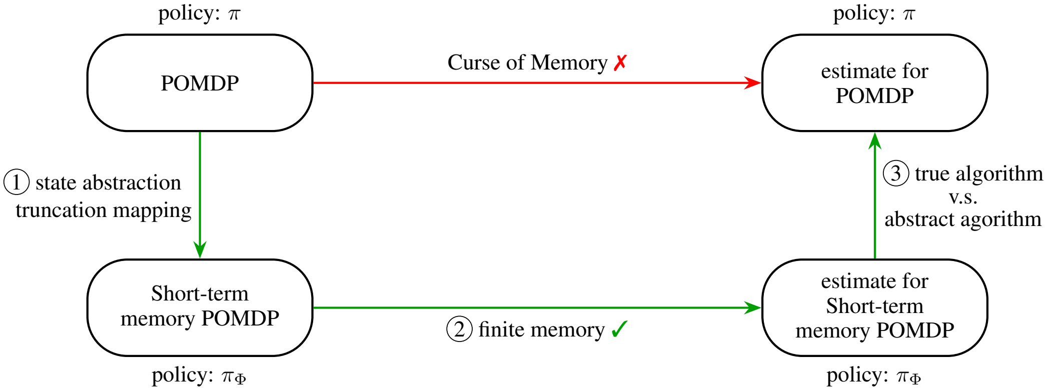

Figure 2: FDVF analysis pipeline

6.3 Abstraction Induced by Truncation Mapping

The first step in the approach is to introduce an abstraction mapping $\tilde{\phi} : \mathcal{H}^+ \to \mathcal{H}_T^+$, where T is the time window, and $\mathcal{H}_T^+ := \bigcup_{t=1}^{T} (\mathcal{O} \times \mathcal{A})^{t-1} \times \mathcal{O}$ denotes the set of history sequences constrained by the time window T. The specific definition of $\tilde{\phi}$ is as follows:

$$\tilde{\phi}(o_1, a_1, \ldots, o_h) := \begin{cases} (o_{h-T+1}, a_{h-T+1}, \ldots, o_h), & \text{if } h \geq T \\ \text{id}, & \text{if } h < T \end{cases} \tag{58}$$

To reuse the previous analysis, we also introduce an abstraction mapping $\phi : \mathcal{B} \to \mathcal{B}$ that operates directly on belief states, and it satisfies $\phi(\mathbf{b}(\tau_h^+)) = \mathbf{b}(\tilde{\phi}(\tau_h^+))$. Under appropriate assumptions, ϕ can be made a sufficiently good abstraction mapping. Since this mapping ϕ depends on the time window length T, we denote it as ϕ_T.

Assumption 10 (Fast-Forgotten POMDP)**.** *For the abstraction mapping ϕ_T defined above, the following holds: for all $\varepsilon > 0$, there exists $T \in \mathbb{N}^+$ such that for all $b_1, b_2 \in \mathcal{B}$, if $\phi_T(b_1) = \phi_T(b_2)$, then $\|b_1 - b_2\|_1 \leq \varepsilon$.*

With this abstraction mapping, we can induce an abstract belief MDP. Moreover, since the abstraction $\tilde{\phi}$ (or equivalently ϕ, as they correspond one-to-one) acts as a bijection for trajectory histories τ_h of length $h \leq T$, it effectively selects a representative from each partition for these trajectories.

Furthermore, observe that in this belief MDP, if for any two belief states $b, b' \in \mathcal{B}$ we have $\Pr(b'|b) > 0$, then the last $T-1$ components of the sequence $\mathbf{b}^{-1}(b')$ must exactly match the first $T-1$ components of the sequence $\mathbf{b}^{-1}(b)$. **This ensures the existence of a POMDP with the same $(\mathcal{O}, \mathcal{A})$ structure whose belief MDP is isomorphic to the belief MDP induced by ϕ.**

Remark 9. *Here, $M_1 = \mathcal{S}_1, \mathcal{A}, P_1, R_1, H$ and $M_2 = \mathcal{S}_2, \mathcal{A}, P_2, R_2, H$ are said to be isomorphic if there exists a bijection $\varphi : \mathcal{S}_1 \to \mathcal{S}_2$ such that*

$$\varphi(M_1) := (\varphi(\mathcal{S}_1), \mathcal{A}, P_1(\varphi(\cdot), \cdot), R_1(\varphi(\cdot), \cdot), H) = M_2$$

With this short-term memory POMDP (whose memory is limited to the time window T), we can now run the abstract algorithm on it. Naturally, the policy to be evaluated must also be abstracted (i.e., truncated), transforming from the original π into π_ϕ. To ensure that the truncation does not introduce excessive error, we must also invoke the Lipschitz continuity assumption on the policy from Assumption 2. Combined with the fast-forgetting POMDP assumption (Assumption 10), this implies that the policy itself possesses a kind of fast-forgetting property.

Lemma 10 (Fast-Forgotten Policy)**.** *If Assumption 10 and Assumption 2 hold, then for the previously defined $\tilde{\phi}_T$, it holds that for all $\varepsilon > 0$, there exists $T \in \mathbb{N}^+$ such that for all $\tau_h^{[1]+}, \tau_h^{[2]+} \in \mathcal{H}^+$ and all $\pi \in \pi_e, \pi_b$, if $\tilde{\phi}_T(\tau_h^{[1]+}) = \tilde{\phi}_T(\tau_h^{[2]+})$, then $\|\pi(\tau_h^{[1]+}) - \pi(\tau_h^{[2]+})\|_1 \leq \varepsilon$.*

As discussed above, since this short-term memory POMDP is induced by an abstraction mapping ϕ, and this abstraction mapping ϕ guarantees that all belief states mapped to the same representative are close to each other (Assumption 10), the arrow ① in Figure 2 can be directly used to apply the conclusions from Theorem 7 and its corollaries for error control.

Remark 10. *Note that policy truncation is necessary here. This is not only to directly reuse the conclusions from Theorem 7 and its corollaries, but also due to the "curse of memory"—the memory of a policy can severely affect the quality of theoretical guarantees.*

Finally, we abstract the content of this section into the following theorem.

Theorem 12. *If Assumption 10 and Assumption 2 hold, then for any $\varepsilon > 0$, there exists a time window $T \in \mathbb{N}^+$ such that*

$$|J(\pi_e) - J^{\phi_T}(\pi_e)| \leq L_\phi \varepsilon \tag{59}$$

where L_ϕ is a manually chosen constant, selected according to Theorem 7. Since we are considering finite-horizon POMDPs here, $1-\gamma$ should be replaced by H. In addition, the value of T that satisfies the condition can be regarded as a function of ε, denoted $T_0(\varepsilon)$.

6.4 Theoretical Guarantee of FDVF on Short-Term Memory POMDP

6.4.1 Theoretical Guarantee of the Abstract Algorithm

On the abstracted short-term memory POMDP (whose existence has been proven in the previous section), the doubly robust algorithm (135) can be executed. This algorithm clearly satisfies the following theoretical guarantee:

Theorem 13. *Assuming the realizability condition $V_{\mathcal{S}} \in \mathcal{V}$ and the Bellman completeness condition $\forall V \in \mathcal{V}, \mathcal{B}^{\mathcal{H}} V \in \Theta$, then for any $T \in \mathbb{N}^+$, with probability at least $1-\delta$, we have:*

$$\begin{aligned}|J^{\phi_T}(\pi_e) - \mathbb{E}_{\pi_b}[\hat{V}^{\phi_T}(f_1)]| \leq cH \max\{\max_{V\in\mathcal{V}} \|V\|_\infty + 1, \max_{\theta\in\Theta} \|\theta\|_\infty\} \cdot \|w^{\phi_T}\|_\infty \\ \cdot \max_{h\in[H]} \sup_{V\in\mathcal{V}} \sqrt{\frac{\mathbb{E}_{\pi_e}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)} V)(s_h, \tau_{[h-T+1:h]})^2]}{\mathbb{E}_{\pi_b}[(\mathcal{B}^{\mathcal{H}} V)(\tau_h)^2]}} \cdot \sqrt{\frac{C_\mu}{n} \log \frac{|\mathcal{V}||\Theta|}{\delta}}\end{aligned} \tag{60}$$

where $C_\mu := \max_h \max_{a_h, \tau_h^+} \mu(a_h, \tau_h^+)$, and w^{ϕ_T} stands for the importance weight that shifts from the true data generating distribution to the data generating distribution induced by π_b^ϕ on the short-term memory POMDP.

The characterization of $\|w^{\phi_T}\|_\infty$ can be found in lemma 12.

For marginal importance sampling, under suitable assumptions, a corresponding finite-sample error bound can also be written [1].

Since the policies considered here only have short-term memory T, this in a sense mitigates the "curse of memory." Thus, the analysis for arrow ② is completed.

6.4.2 Real Algorithm vs. Abstract Algorithm

The differences between the real algorithm and the abstract algorithm come from three aspects:

1. The discrepancy between $\mu(a_h, \tau_h^+) = \pi_e(a_h|\tau_h^+)/\pi_b(a_h|\tau_h^+)$ and the truncated version $\mu(a_h, \tau_{[h-T+1:h]}^+) = \pi_e(a_h|\tau_{[h-T+1:h]}^+)/\pi_b(a_h|\tau_{[h-T+1:h]}^+)$. This discrepancy can be controlled by the following lemma:

Lemma 11. *If Assumptions 2 and 10 hold, then for any $\varepsilon > 0$, there exists $T \in \mathbb{N}^+$ such that*

$$|\mu(a_h, \tau_h^+) - \mu(a_h, \tau_{[h-T+1:h]}^+)| \le \frac{c_1(C_\mu + 1)\varepsilon}{\min_h \min_{a_h, \tau_h^+} \pi_b(a_h|\tau_h^+)} \tag{61}$$

where c_1 is an absolute constant. The set of T values satisfying the condition can be denoted as a function of ε, written $T_1(\varepsilon)$.

2. The discrepancy between $V(f_h', \tau_h')$ and $V(f_h', \tau_{[h-T+1:h]}')$. This requires the function class to forget historical information quickly, as stated below:

Assumption 11 (Fast-Forgotten Function Class)**.** *Consider the function class used for estimation $\mathcal{V} : \mathcal{F} = (\mathcal{F}' \times \mathcal{H}) \to \mathbb{R}$. It satisfies that for all $\varepsilon > 0$, there exists $T \in \mathbb{N}^+$ such that for all $V \in \mathcal{V}$,*

$$|V(f_h, \tau_h) - V(f_h, \tau_{[h-T+1:h]})| \le \|\mathcal{V}\|_\infty \varepsilon \tag{62}$$

where $\|\mathcal{V}\|_\infty := \max_{V \in \mathcal{V}} \|V\|_\infty$. The suitable values of T form a function of ε, denoted as $T_2(\varepsilon)$.

Note that this assumption imposes no constraints on the "future" component—only the history is required to exhibit a fast-forgetting property.

3. The difference in data-generating distribution between the real POMDP and the abstract short-term memory POMDP. This discrepancy arises from two sources:

The behavior policy is truncated.

The transition probabilities of the POMDP differ slightly.

Let $w^\phi(f_1)$ denote the importance weight accounting for this distribution shift. Then we define:

$$w^{\phi_T}(f_1) := \frac{\pi_b^{\phi_T}(a_1|\tau_1^+)}{\pi_b(a_1|\tau_1^+)} \cdot \frac{P^{\phi_T}(o_2|\tau_2)}{P(o_2|\tau_2)} \cdot \dots \cdot \frac{\pi_b^{\phi_T}(a_H|\tau_H^+)}{\pi_b(a_H|\tau_H^+)} \tag{63}$$

Under the assumptions that both the POMDP and the policy are fast-forgetting, we have the following lemma:

Lemma 12. *If Assumptions 10 and 2 hold, then for any $\varepsilon > 0$, there exists $T \in \mathbb{N}^+$ such that:*

$$|w^{\phi_T}(f_1)| \le 1 + \frac{c_2 H \varepsilon}{\min\{\min_h \min_{o_h, \tau_h} P(o_h|\tau_h), \min_h \min_{a_h, \tau_h^+} \pi_b(a_h|\tau_h^+)\}} \tag{64}$$

The values of T satisfying this condition form a function of ε, denoted $T_3(\varepsilon)$.

To summarize, by considering all sources of error, we have the following theorem.

Theorem 14. *Define*

$$\mathcal{E}_{\mathcal{V},\Theta}(V) := \max_{\theta \in \Theta} \sum_{h=1}^H \mathbb{E}_\mathcal{D}[\{\mu(a_h, \tau_h^+)(r_h + V(f_{h+1}))V(f_h)\}\theta(\tau_h) - \frac{1}{2}\theta(\tau_h)^2] \tag{65}$$

$$\mathcal{E}_{\mathcal{V},\Theta}^{\phi_T}(V) := \max_{\theta \in \Theta} \sum_{h=1}^H \mathbb{E}_\mathcal{D}[\{\mu(a_h, \tau_{h-T+1:h}^+)(r_h + V(\phi_T(f_{h+1}))) - V(\phi_T(f_h))\}\theta(\tau_h) - \frac{1}{2}\theta(\tau_h)^2] \tag{66}$$

If Assumptions 10,2, and11 all hold, then for any $\varepsilon > 0$ and for any $V \in \mathcal{V}$, we have:

$$|\mathcal{E}_{\mathcal{V},\Theta}(V) - \mathcal{E}^{\phi_T}_{\mathcal{V},\Theta}(V)| \leq L_{\mathcal{E}}\varepsilon \tag{67}$$

where

$$L_{\mathcal{E}} := \left(\frac{c_1 H (C_\mu + 1)\|\mathcal{V}\|_\infty \|\Theta\|_\infty}{\min_h \min_{a_h, \tau_h^+} \pi_b(a_h|\tau_h^+)} + C_\mu \|\mathcal{V}\|_\infty \|\Theta\|_\infty + \right.$$
$$\left. \frac{c_2 H \max\{C_\mu \|\mathcal{V}\|_\infty \|\Theta\|_\infty, \frac{1}{2}\|\Theta\|_\infty^2\}}{\min\{\min_h \min_{o_h, \tau_h} P(o_h|\tau_h), \min_h \min_{a_h, \tau_h^+} \pi_b(a_h|\tau_h^+)\}} \right) \tag{68}$$

and $T = \max\{T_1(\varepsilon), T_2(\varepsilon), T_3(\varepsilon)\}$.

6.5 Theoretical Guarantee of FDVF

After completing the previous two subsections, we have essentially finished the error control of the three green arrows in Figure 2, thereby completing the overall analysis. The following theorem summarizes all the previous results.

Theorem 15 (Theoretical Guarantee of FDVF)**.** *Suppose the realizability condition $V_{\mathcal{S}} \in \mathcal{V}$ and the Bellman completeness condition $\forall V \in \mathcal{V}, \mathcal{B}^{\mathcal{H}} V \in \Theta$ hold, and Assumptions 10,2, and11 are satisfied. For any $\varepsilon > 0$, define $T = \max\{T_0(\varepsilon), T_1(\varepsilon), T_2(\varepsilon), T_3(\varepsilon)\}$.*

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

$$|J(\pi_e) - \mathbb{E}_{\pi_b}[\hat{V}(f_1)]| \leq L_\phi \varepsilon + cH \max\{\|\mathcal{V}\|_\infty + 1, \|\Theta\|_\infty\} \cdot \|w^{\phi_T}\|_\infty$$
$$\cdot \max_{h \in [H]} \sup_{V \in \mathcal{V}} \sqrt{\frac{\mathbb{E}_{\pi_e}[(\mathcal{B}^{(\mathcal{S}, \mathcal{H}_T)} V)(s_h, \tau_{[h-T+1:h]})^2]}{\mathbb{E}_{\pi_b}[(\mathcal{B}^{\mathcal{H}} V)(\tau_h)^2]}} \cdot \sqrt{\frac{C_\mu}{n} \log \frac{|\mathcal{V}||\Theta|}{\delta}} + L_{\mathcal{E}}\varepsilon \tag{69}$$

Corollary 4. *Under the conditions of the above theorem, with probability greater then $1 - \delta$, we have:*

$$|J(\pi_e) - \mathbb{E}_{\pi_b}[\hat{V}(f_1)]| \leq \inf_{\substack{\varepsilon \geq 0 \\ D(\varepsilon)}} \left(L_\phi \varepsilon + cH \max\{\|\mathcal{V}\|_\infty + 1, \|\Theta\|_\infty\} \cdot \|w^{\phi_{T(\varepsilon)}}\|_\infty \right.$$
$$\left. \cdot \max_{h \in [H]} \sup_{V \in \mathcal{V}} \sqrt{\frac{\mathbb{E}_{\pi_e}[(\mathcal{B}^{(\mathcal{S}, \mathcal{H}_{T(\varepsilon)})} V)(s_h, \tau_{[h-T(\varepsilon)+1:h]})^2]}{\mathbb{E}_{\pi_b}[(\mathcal{B}^{\mathcal{H}} V)(\tau_h)^2]}} \cdot \sqrt{\frac{C_\mu}{n} \log \frac{|\mathcal{V}||\Theta|}{\delta}} + L_{\mathcal{E}}\varepsilon \right) \tag{70}$$

where $D(\varepsilon)$ stands for such ε that satisfies realizability and Bellman completeness.

6.6 A Simpler Pipeline: Abstracting Only the Policy

Upon carefully revisiting the above analysis, it becomes apparent that one step in the full reuse of the framework is actually unnecessary—namely, the abstraction from the original POMDP to the short-term memory POMDP. In fact, what truly matters is the abstraction of the policy, since the memory dependency of the policy is the real root of the "curse of memory."

However, to abstract the policy, one inevitably relies on the Lipschitz continuity of the policy. And to avoid the curse of memory by adjusting the memory window, it becomes necessary to link the distance in the belief space with the fast-forgetting property—which, in turn, inevitably invokes assumptions about the POMDP itself. Therefore, under the current analytical framework, even if we avoid abstracting the POMDP directly, we cannot eliminate such assumptions entirely.

That said, a major advantage of abstracting only the policy is that, when controlling the error corresponding to arrow ③, modifying the POMDP itself leads to the appearance of terms like $\max_h \max_{o_h, \tau_h} P(o_h|\tau_h)$ in the denominator of the upper bound on the importance weights. These terms depend on the nature of the POMDP. Although this issue can be circumvented under high-probability assumptions, abstracting only the policy directly removes this term from the bound—which is a desirable outcome.

Ultimately, this simplified pipeline does not differ much from the full analysis presented earlier. In many cases, the difference in the results is only in polynomial factors. Interested readers are encouraged to conduct their own analysis.

References

[1] Yuheng Zhang and Nan Jiang. On the curses of future and history in future-dependent value functions for off-policy evaluation, 2024.

[2] Zongzhang Zhang, Michael Littman, and Xiaoping Chen. Covering number as a complexity measure for pomdp planning and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1853–1859, 2012.

A Supplementary Proofs

A.1 Proof of Lemma 1

Proof. This is easily obtained from:

$$
\begin{aligned}
|r(b_1, a) - r(b_2, a)| &= |\mathbb{E}_{s\sim b_1}[r(s, a)] - \mathbb{E}_{s\sim b_2}[r(s, a)]| \\
&= |\langle r(\cdot, a), b_1 - b_2\rangle| \\
&\leq R_{\max}\|b_1 - b_2\|_1.
\end{aligned}
$$

And it shows that when treating POMDPs as belief space MDPs, there's intrinsic smoothness within the dynamic. □

A.2 Proof of Lemma 2

Proof.

$$
\begin{aligned}
P(s'|\tau_h, o, a) &= \sum_s P(s'|s, \tau_h, o, a) \cdot P(s|\tau_h, o, a) \\
&= \sum_s P(s'|s, a) \cdot P(s|\tau_h, o, a)
\end{aligned}
$$

where

$$
\sum_{o,a} P(s|\tau_h, o, a) \cdot P(o, a|\tau_h) = P(s|\tau_h)
$$

Then

$$
\begin{aligned}
\|b_1^{o,a} - b_2^{o,a}\|_1 &= \sum_{s'} \left| \sum_s P(s'|s, a) \cdot (P(s|\tau_h, o, a) - P(s|\tau_h', o, a)) \right| \\
&\leq \sum_{s'} \sum_s P(s'|s, a) \cdot |(P(s|\tau_h, o, a) - P(s|\tau_h', o, a))| \\
&\leq \sum_s |(P(s|\tau_h, o, a) - P(s|\tau_h', o, a))| \\
&\leq \|b_1 - b_2\|_1
\end{aligned}
$$

Therefore,

$$
\eta := \sup_{\substack{b_1, b_2 \in \mathcal{B} \\ o\in\mathcal{O}, a\in\mathcal{A}}} \frac{\|b_1^{o,a} - b_2^{o,a}\|_1}{\|b_1 - b_2\|_1} \leq 1 \tag{71}
$$

□

A.3 Proof of Proposition 1

Proof. We decompose our target function as

$$
\begin{aligned}
&|P(b_1^{o,a}|b_1) - P(b_2^{o,a}|b_2)| \\
&= |P(o|b_1, a)\pi(a|b_1) - P(o|b_2, a)\pi(a|b_2)| \\
&= |P(o|b_1, a)\pi(a|b_1) - P(o|b_1, a)\pi(a|b_2) + P(o|b_1, a)\pi(a|b_2) - P(o|b_2, a)\pi(a|b_2)| \\
&\leq |P(o|b_1, a)(\pi(a|b_1) - \pi(a|b_2))| + |(P(o|b_1, a) - P(o|b_2, a))\pi(a|b_2)| \\
&\leq |P(o|b_1, a)| \cdot |\pi(a|b_1) - \pi(a|b_2)| + |P(o|b_1, a) - P(o|b_2, a)| \cdot |\pi(a|b_2)| \\
&\leq (1 + L_\pi)\|b_1 - b_2\|_1
\end{aligned} \tag{72}
$$

where we used Lemma 3 for the last inequality. □

A.4 Proof of Proposition 2

Proof. We follow the same idea in the proof of Proposition 1 and decompose the LHS as

$$
\begin{aligned}
&\Big|\sum_{a\in\mathcal{A}}(P(b_1^{o,a}|b_1) - P(b_2^{o,a}|b_2))\Big| \\
\leq\ &\Big|\sum_{a\in\mathcal{A}} P(o|b_1,a)(\pi(a|b_1) - \pi(a|b_2))\Big| + \Big|\sum_{a\in\mathcal{A}}(P(o|b_1,a) - P(o|b_2,a))\pi(a|b_2)\Big| \\
\leq\ &\max_{a\in\mathcal{A}} \big|P(o|b_1,a)\big| \cdot \Big|\sum_{a\in\mathcal{A}}(\pi(a|b_1) - \pi(a|b_2))\Big| + \\
&\qquad\qquad \max_{a\in\mathcal{A}} \big|P(o|b_1,a) - P(o|b_2,a)\big| \cdot \Big|\sum_{a\in\mathcal{A}} \pi(a|b_2)\Big| \\
\leq\ &(1+L_\pi)\|b_1 - b_2\|_1,
\end{aligned} \tag{73}
$$

which proves the result. □

A.5 Proof of Proposition 3

Proof. Using Proposition 1, we get

$$|P(b_1^{+1}|b_1) - P(b_2^{+1}|b_2)| \leq (1+L_\pi)\|b_1 - b_2\|_1. \tag{74}$$

Replacing b_1, b_2 with b_1^{+1}, b_2^{+2} and using Lemma 2, we have

$$|P(b_1^{+2}|b_1^{+1}) - P(b_2^{+2}|b_2^{+1})| \leq (1+L_\pi)\eta\|b_1 - b_2\|_1. \tag{75}$$

Therefore

$$
\begin{aligned}
&|P(b_1^{+2}|b_1) - P(b_2^{+2}|b_2)| \\
=\ &\Big|P(b_1^{+2}|b_1) - \sum_{o,a} P(b_2^{+2}|b_2^{o,a})P(b_1^{o,a}|b_1) + \sum_{o,a} P(b_2^{+2}|b_2^{o,a})P(b_1^{o,a}|b_1) - P(b_2^{+2}|b_2)\Big| && (76)\\
\leq\ &\Big|P(b_1^{+2}|b_1) - \sum_{o,a} P(b_2^{+2}|b_2^{o,a})P(b_1^{o,a}|b_1)\Big| + \Big|\sum_{o,a} P(b_2^{+2}|b_2^{o,a})P(b_1^{o,a}|b_1) - P(b_2^{+2}|b_2)\Big| && (77)\\
=\ &\Big|\sum_{o,a}\Big(P(b_1^{+2}|b_1^{o,a}) - P(b_2^{+2}|b_2^{o,a})\Big)P(b_1^{o,a}|b_1)\Big| + \Big|\sum_{o,a} P(b_2^{+2}|b_2^{o,a})\Big(P(b_1^{o,a}|b_1) - P(b_2^{o,a}|b_2)\Big)\Big| \\
\leq\ &\|P(b_1^{+2}|b_1^{[\cdot]}) - P(b_2^{+2}|b_2^{[\cdot]})\|_\infty \|P(b_1^{[\cdot]}|b_1)\|_1 + \|P(b_2^{+2}|b_2^{[\cdot]})\|_1 \|P(b_1^{[\cdot]}|b_1) - P(b_2^{[\cdot]}|b_2)\|_\infty && (78)\\
\leq\ &(1+L_\pi)\eta\|b_1 - b_2\|_1 \cdot 1 + 1 \cdot (1+L_\pi)\|b_1 - b_2\|_1 && (79)\\
\leq\ &(1+L_\pi)(1+\eta)\|b_1 - b_2\|_1 && (80)
\end{aligned}
$$

where in (79) we used the fact that $\|P(b_2^{+2}|b_2^{[\cdot]})\|_1 = P(b_2^{o_1,a_1,o_2,a_2}|b_2^{o_1,a_1}) \leq 1$. This is the consequence of Assumption 1 that every belief state has a unique history. It's worth mentioning that, the same as Proposition 1, every step there can be a $|\mathcal{O}||\mathcal{A}|$ factor loose.

After that, we recursively repeat the procedure above, and using mathematical induction, we get the result. □

A.6 Proof of Theorem 1

Proof. We first control $\mathbb{E}_{a,o,\cdots\sim b_1}\big[\sum_{k=0}^{\infty}\gamma^k r(b_1^{+k}, a_k)\big] - \mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^{\infty}\gamma^k r(b_2^{+k}, a_k)\big]$. After we do this, we can reduce the problem to the one we need using

$$
\begin{aligned}
&\Big|\sum_{b_1\in\mathrm{bin}(\phi(b))}\Big[p_{\phi(b)}(b_1)\mathbb{E}_{a,o,\cdots\sim b_1}\big[\sum_{k=0}^{\infty}\gamma^k r(b_1^{+k}, a_k)\big]\Big] - \mathbb{E}_{a,o,\cdots\sim b'}\big[\sum_{k=0}^{\infty}\gamma^k r(b'^{+k}, a_k)\big]\Big| \\
\leq\ &\Big|\mathbb{E}_{a,o,\cdots\sim b}\big[\sum_{k=0}^{\infty}\gamma^k r(b^{+k}, a_k)\big] - \mathbb{E}_{a,o,\cdots\sim b'}\big[\sum_{k=0}^{\infty}\gamma^k r(b'^{+k}, a_k)\big]\Big| && (81)
\end{aligned}
$$

which is already controlled.

To do this, we split the formula into two parts:

$$\begin{aligned}
&\Big|\mathbb{E}_{a,o,\cdots\sim b_1}\big[\sum_{k=0}^{\infty}\gamma^k r(b_1^{+k},a_k)\big]-\mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^{\infty}\gamma^k r(b_2^{+k},a_k)\big]\Big|\\
\leq&\Big|\mathbb{E}_{a,o,\cdots\sim b_1}\big[\sum_{k=0}^{\infty}\gamma^k r(b_1^{+k},a_k)\big]-\mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^{\infty}\gamma^k r(b_1^{+k},a_k)\big]\Big|+\\
&\Big|\mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^{\infty}\gamma^k r(b_1^{+k},a_k)\big]-\mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^{\infty}\gamma^k r(b_2^{+k},a_k)\big]\Big|. \qquad (82)
\end{aligned}$$

We first look at the first term.

$$\begin{aligned}
&\Big|\mathbb{E}_{a,o,\cdots\sim b_1}\big[\sum_{k=0}^{\infty}\gamma^k r(b_1^{+k},a_k)\big]-\mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^{\infty}\gamma^k r(b_1^{+k},a_k)\big]\Big|\\
=&\Big|\sum_{k=0}^{\infty}\Big(\mathbb{E}_{a,o,\cdots\sim b_1}\big[\gamma^k r(b_1^{+k},a_k)\big]-\mathbb{E}_{a,o,\cdots\sim b_2}\big[\gamma^k r(b_1^{+k},a_k)\big]\Big)\Big| \qquad (83)
\end{aligned}$$

which corresponds to the propagated error within each layer and summing them up. As discussed above, with Assumption 4, this term is dominated by $L_H R_{\max}\|b_1-b_2\|_1/(1-\gamma)$.

Next, we look at the second term which is not covered by Assumption 4.

$$\begin{aligned}
&\Big|\mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^{\infty}\gamma^k r(b_1^{+k},a_k)\big]-\mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^{\infty}\gamma^k r(b_2^{+k},a_k)\big]\Big|\\
=&\Big|\sum_{k=0}^{\infty}\Big(\mathbb{E}_{a,o,\cdots\sim b_2}\big[\gamma^k r(b_1^{+k},a_k)-\gamma^k r(b_2^{+k},a_k)\big]\Big)\Big|\\
\leq&\frac{R_{\max}\|b_1-b_2\|_1}{1-\gamma\eta} \qquad (84)
\end{aligned}$$

where we used Lemma 1 and Lemma 2. □

A.7 Proof of Theorem 2

Proof. We begin by clarifying and establishing the notation used in the proof. Fix an arbitrary family $\{p_x\}_{x\in\mathcal{S}_\phi}$, and let $b'\sim\mathrm{bin}(\phi(b))$ denote the expectation taken over the following sampling process:

1. Since $\phi(b)\in\mathcal{B}_\phi$ is the representative element of some partition of the belief space after binning, the set $\phi^{-1}(\phi(b))\subset\mathcal{B}$ is the corresponding element in the original belief space—i.e., the subset consisting of all belief states that are grouped into the same bin as b.

2. Sample a temporary belief state b_{temp} from $\phi^{-1}(\phi(b))$ according to the fixed distribution $p_{\phi(b)}$.

3. Starting from b_{temp}, perform the belief update procedure shown in (**??**), where the action a is determined by the policy π. Once the update is complete, the resulting belief state is the sampled b'.

With this notation established, we can proceed with the proof of the theorem. The main idea of the proof is to construct a chain rule argument. First, notice that

$$\tilde{V}^{\pi_\phi}_{\text{bin}}=\mathbb{E}_{b_1\sim\mathrm{bin}(\phi(b))}[V^{[\pi_\phi]_{\text{true}}}_{\text{true}}(b_1)] \qquad (85)$$

and

$$V^{\pi_\phi}_{\text{bin}}=\mathbb{E}_{\substack{b_1\sim\mathrm{bin}(\phi(b))\\ b_2\sim\mathrm{bin}(\phi(b_1))\\ \cdots}}\left[r_\phi(\phi(b_1),a_1)+\gamma r_\phi(\phi(b_2),a_2)+\gamma^2 r_\phi(\phi(b_3),a_3)+\cdots\right] \qquad (86)$$

Consider $V^{[k]}$ as

$$V^{[k]}=\mathbb{E}_{\substack{b_1\sim\mathrm{bin}(\phi(b))\\ b_2\sim\mathrm{bin}(\phi(b_1))\\ \cdots\\ b_k\sim\mathrm{bin}(\phi(b_{k-1}))\\ b_{k+1}\sim b_k\\ \cdots}}\left[r_\phi(\phi(b_1),a_1)+\gamma r_\phi(\phi(b_2),a_2)+\gamma^2 r_\phi(\phi(b_3),a_3)+\cdots\right] \qquad (87)$$

Then for $\forall b$,

$$
\begin{aligned}
& |V^{[k+1]}(b) - V^{[k]}(b)| && (88) \\
= & \Bigg| \mathbb{E}_{\substack{b_1 \sim \text{bin}(\phi(b)) \\ b_2 \sim \text{bin}(\phi(b_1)) \\ \cdots \\ b_k \sim \text{bin}(\phi(b_{k-1})) \\ b_{k+1} \sim b_k \\ b_{k+2} \sim b_{k+1} \\ \cdots}} [\gamma^k V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_{k+1})] - \\
& \qquad \mathbb{E}_{\substack{b_1 \sim \text{bin}(\phi(b)) \\ b_2 \sim \text{bin}(\phi(b_1)) \\ \cdots \\ b_k \sim \text{bin}(\phi(b_{k-1})) \\ b_{k+1} \sim \text{bin}(\phi(b_k)) \\ b_{k+2} \sim b_{k+1} \\ \cdots}} [\gamma^k r_\phi(\phi(b_{k+1}), a) + \gamma^{k+1} V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_{k+2})] \Bigg| && (89) \\
= & \Bigg| \mathbb{E}_{\substack{b_1 \sim \text{bin}(\phi(b)) \\ b_2 \sim \text{bin}(\phi(b_1)) \\ \cdots \\ b_k \sim \text{bin}(\phi(b_{k-1})) \\ b_{k+1} \sim b_k \\ b_{k+2} \sim b_{k+1} \\ \cdots}} [\gamma^k V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_{k+1})] - \\
& \quad \mathbb{E}_{\substack{b_1 \sim \text{bin}(\phi(b)) \\ b_2 \sim \text{bin}(\phi(b_1)) \\ \cdots \\ b_k \sim \text{bin}(\phi(b_{k-1})) \\ b_{k+1} \sim \text{bin}(\phi(b_k)) \\ b_{k+2} \sim b_{k+1} \\ \cdots}} [\gamma^k r_\phi(\phi(b_{k+1}), a) - \gamma^k r(b_{k+1}, a) + \gamma^k V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_{k+1})] \Bigg| && (90) \\
\leq & \gamma^k R_{\max} \varepsilon + \frac{\gamma^k}{1-\gamma} R_{\max} L_V \varepsilon && (91)
\end{aligned}
$$

where the last inequality used the Lipchitz of value function since the next belief is sampled from the same bin and thus close enough.

Finally, we do the chaining, and sums up all the $V^{[k+1]} - V^{[k]}$ to get for $\forall \phi(b)$,

$$
\begin{aligned}
& |\tilde{V}_{\text{bin}}^{\pi_\phi}(\phi(b)) - V_{\text{bin}}^{\pi_\phi}(\phi(b))| && (92) \\
= & \left| \sum_{k=1}^{\infty} \left(V^{[k+1]}(b) - V^{[k]}(b) \right) \right| && (93) \\
\leq & \sum_{k=1}^{\infty} \left| \gamma^k R_{\max} \varepsilon + \frac{\gamma^k}{1-\gamma} R_{\max} L_V \varepsilon \right| && (94) \\
\leq & \frac{R_{\max} \varepsilon}{1-\gamma} + \frac{R_{\max} L_V \varepsilon}{(1-\gamma)^2} && (95)
\end{aligned}
$$

□

A.8 Proof of Theorem 3

Proof. Using the fact that $V_{\text{true}}^{[\pi_\phi]_{\text{true}}} = \mathcal{T}^{[\pi_\phi]_{\text{true}}} V_{\text{true}}^{[\pi_\phi]_{\text{true}}}$,

$$
\begin{aligned}
\|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty &= \|V_{\text{true}}^{\pi} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V_{\text{true}}^{\pi} + \mathcal{T}^{[\pi_\phi]_{\text{true}}} V_{\text{true}}^{\pi} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \\
&\leq \|V_{\text{true}}^{\pi} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V_{\text{true}}^{\pi}\|_\infty + \gamma \|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty. \qquad (96)
\end{aligned}
$$

Consequently,

$$
\|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq \frac{1}{1-\gamma} \|V_{\text{true}}^{\pi} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V_{\text{true}}^{\pi}\|_\infty. \qquad (97)
$$

For any b, we have

$$\begin{aligned}
&|(V^{\pi}_{\text{true}} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V^{\pi}_{\text{true}})(b)| \\
&= |(\mathcal{T}^{\pi} V^{\pi}_{\text{true}} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V^{\pi}_{\text{true}})(b)| \\
&= \left| \mathbb{E}_{\substack{a\sim\pi(b) \\ b^{+1}\sim P(\cdot|b)}} \left[r + \gamma V^{\pi}_{\text{true}}(b^{+1}) \right] - \mathbb{E}_{\substack{a\sim\pi_\phi(\phi(b)) \\ b^{+1}\sim P(\cdot|b)}} \left[r + \gamma V^{\pi}_{\text{true}}(b^{+1}) \right] \right| \qquad (98)
\end{aligned}$$

We first look at r,

$$\begin{aligned}
|\mathbb{E}_{a\sim\pi(b)}[r] - \mathbb{E}_{a\sim\pi_\phi(\phi(b))}[r]| &= |\mathbb{E}_{a\sim\pi(b)}[r] - \mathbb{E}_{a\sim\pi(\phi(b))}[r]| \\
&\leq R_{\max} L_{\pi} \varepsilon \qquad (99)
\end{aligned}$$

Then we look at V^{π}_{true},

$$\begin{aligned}
&\left| \mathbb{E}_{\substack{a\sim\pi(b) \\ b^{+1}\sim P(\cdot|b)}} \left[\gamma V^{\pi}_{\text{true}}(b^{+1}) \right] - \mathbb{E}_{\substack{a\sim\pi_\phi(\phi(b)) \\ b^{+1}\sim P(\cdot|b)}} \left[\gamma V^{\pi}_{\text{true}}(b^{+1}) \right] \right| \\
&= \gamma \left| \sum_{o\in\mathcal{O}} \sum_{a\in\mathcal{A}} \left[\big(P(b^{o,a}|b) - P(\phi(b)^{o,a}|\phi(b)) \big) \cdot V^{\pi}(b^{o,a}) \right] \right| \\
&\leq \gamma \sum_{o\in\mathcal{O}} \left| \sum_{a\in\mathcal{A}} \left[\big(P(b^{o,a}|b) - P(\phi(b)^{o,a}|\phi(b)) \big) \cdot V^{\pi}(b^{o,a}) \right] \right| \\
&\leq 2|\mathcal{O}||\mathcal{A}| \frac{\gamma R_{\max}}{1-\gamma}(1+L_\pi)\varepsilon \wedge \left(|\mathcal{O}| \frac{\gamma R_{\max}}{1-\gamma}(1+L_\pi)\varepsilon + \gamma|\mathcal{O}||\mathcal{A}|(1+L_\pi)L_V \varepsilon^2 \right) \qquad (100)
\end{aligned}$$

where we used Proposition 1 or the tighter result Proposition 2. However, the latter would need Assumption 5 to hold. Note that we can't directly apply Assumption 4 here. □

A.9 Proof of Lemma 4

Proof. Recall the previously mentioned Lemma **??**. Substituting it into the case of the abstract belief MDP gives:

$$J_{\hat{Q}^{\pi}_{\phi}}(\pi_\phi) - J(\pi_\phi) = \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_\phi}}[\hat{Q}^{\pi}_{\phi} - \mathcal{T}^{\pi_\phi} \hat{Q}^{\pi}_{\phi}] \qquad (101)$$

Therefore, we have:

$$\begin{aligned}
|J_{\hat{Q}^{\pi}_{\phi}}(\pi_\phi) - J(\pi_\phi)| &= \left| \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_\phi}}[\hat{Q}^{\pi}_{\phi} - \mathcal{T}^{\pi_\phi} \hat{Q}^{\pi}_{\phi}] \right| \\
&\leq \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi_\phi}}[|\hat{Q}^{\pi}_{\phi} - \mathcal{T}^{\pi_\phi} \hat{Q}^{\pi}_{\phi}|] \qquad (102) \\
&\leq \frac{1}{1-\gamma} \sqrt{\mathbb{E}_{d^{\pi_\phi}}[(\hat{Q}^{\pi}_{\phi} - \mathcal{T}^{\pi_\phi} \hat{Q}^{\pi}_{\phi})^2]} \qquad (103) \\
&\leq \frac{1}{1-\gamma} \sqrt{\mathbb{E}_{d^{D}}[\frac{d^{\pi}}{d^{D}}(\hat{Q}^{\pi}_{\phi} - \mathcal{T}^{\pi_\phi} \hat{Q}^{\pi}_{\phi})^2]} \qquad (104) \\
&\leq \frac{\sqrt{C_\pi(\phi)}}{1-\gamma} \cdot \sqrt{\mathbb{E}_{d^{D}}[(\hat{Q}^{\pi}_{\phi} - \mathcal{T}^{\pi_\phi} \hat{Q}^{\pi}_{\phi})^2]} \qquad (105)
\end{aligned}$$

□

A.10 Proof of Lemma 5

Proof. For $\mathcal{E}_\phi(f, \pi)$, we first estimate an upper bound on its absolute value. Since $f \in \mathcal{F}$ is used to approximate a value function, its upper bound can be assumed to be no greater than $R\text{max}/(1-\gamma)$, i.e., the upper bound of the value function. Therefore, we can give a rough upper bound (possibly with a constant slack, which is acceptable since it's only a constant):

$$0 \leq \mathcal{E}_\phi(f, \pi) \leq \frac{4R^2_{\max}}{(1-\gamma)^2}$$

Thus, by Hoeffding's inequality, for any $f \in \mathcal{F}$, we have:

$$\begin{aligned} &\Pr(|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(f,\pi)]| \geq t) \leq 2\exp\left(-\frac{2t^2 n(1-\gamma)^4}{16R_{\max}^4}\right) \\ \to\ &\Pr(|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(f,\pi)]| > t) \leq 2\exp\left(-\frac{2t^2 n(1-\gamma)^4}{16R_{\max}^4}\right) \qquad (106) \end{aligned}$$

However, the goal of the proof is actually:

$$\Pr(\forall f \in \mathcal{F}, |\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(f,\pi)]| \leq t) \qquad (107)$$

For such problems, a common approach is to use the union bound. Let the probability space be $(\Omega, \Sigma, \Pr)$, and define the events:

$$\begin{aligned} A &:= \{\omega \in \Omega : \forall f \in \mathcal{F}, |\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(f,\pi)]|(\omega) \leq t\} \\ B_f &:= \{\omega \in \Omega : |\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(f,\pi)]|(\omega) > t\} \end{aligned}$$

Then:

$$\begin{aligned} &\Pr(\forall f \in \mathcal{F}, |\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(f,\pi)]| \leq t) \\ =\ &\Pr(A) \\ =\ &1 - \Pr(\Omega \backslash A) && (108) \\ =\ &1 - \Pr(\bigcup_{f\in\mathcal{F}} B_f) && (109) \\ \geq\ &1 - \sum_{f\in\mathcal{F}} \Pr(B_f) && (110) \\ \geq\ &1 - 2|\mathcal{F}|\exp\left(-\frac{t^2 n(1-\gamma)^4}{8R_{\max}^4}\right) && (111) \end{aligned}$$

The second-to-last step uses the subadditivity of probability (countable subadditivity), and the final step applies inequality (106).

Let $\delta = 2|\mathcal{F}|\exp\left(-t^2 n(1-\gamma)^4/8R_{\max}^4\right)$, then solving for t gives $t = \sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}}$

Substituting this into (107) completes the proof. □

A.11 Proof of Proposition 5

Proof. First, by the realizability assumption 8, we have the following inequality:

$$\mathcal{E}_\phi(Q_\phi^\pi, \pi) \geq \mathcal{E}_\phi(\hat{Q}_\phi^\pi, \pi) \geq 0 \qquad (112)$$

This is because $\hat{Q}_\phi^\pi = \arg\min_{f\in\mathcal{F}} \mathcal{E}_\phi(f,\pi)$. Also, from the fixed-point property of the Bellman operator, we have:

$$\mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(Q_\phi^\pi, \pi)] = 0 \qquad (113)$$

Now, applying Lemma 5, we obtain:

$$|\mathcal{E}_\phi(\hat{Q}_\phi^\pi,\pi) - \mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(\hat{Q}_\phi^\pi,\pi)]| \leq \sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}} \qquad (114)$$

and

$$|\mathcal{E}_\phi(Q_\phi^\pi,\pi) - \mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(Q_\phi^\pi,\pi)]| \leq \sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}} \qquad (115)$$

Combining (112),(113),(114), and (115), we get:

$$|\mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(\hat{Q}_\phi^\pi,\pi)]| \leq 2\sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}} \qquad (116)$$

□

A.12 Proof of Lemma 6

Proof.

$$\begin{aligned}
&|\mathcal{E}(f,\pi) - \mathcal{E}_\phi(f,\pi)| \\
&= |\mathbb{E}_\mathcal{D}[(f(b,a) - (r + \gamma f(b'_A,\pi)))(f(b,a) - (r + \gamma f(b'_B,\pi)))] - \\
&\quad \mathbb{E}_\mathcal{D}[(f(\phi(b),a) - (r_\phi + \gamma f(\phi(b'_A),\pi_\phi)))(f(\phi(b),a) - (r_\phi + \gamma f(\phi(b'_B),\pi_\phi)))]| \\
&\le |\mathbb{E}_\mathcal{D}[\{(f(b,a) - f(\phi(b),a)) - (r(b,a) - r_\phi(\phi(b),a)) - \gamma(f(b'_A,\pi) - f(\phi(b'_A),\pi_\phi)\} \\
&\qquad \cdot (f(b,a) - (r + \gamma f(b'_B,\pi)))]| + \\
&\quad |\mathbb{E}_\mathcal{D}[\{(f(b,a) - f(\phi(b),a)) - (r(b,a) - r_\phi(\phi(b),a)) - \gamma(f(b'_B,\pi) - f(\phi(b'_B),\pi_\phi)\} \\
&\qquad \cdot (f(b,a) - (r + \gamma f(b'_A,\pi)))]|.
\end{aligned} \tag{117}$$

Using the fact that

$$\begin{aligned}
&|f(b,\pi) - f(\phi(b),\pi_\phi)| \\
&= |\mathbb{E}_{\pi(a|b)}[f(b,a)] - \mathbb{E}_{\pi(a|\phi(b))}[f(\phi(b),a)]| \\
&\le |\mathbb{E}_{\pi(a|b)}[f(b,a)] - \mathbb{E}_{\pi(a|\phi(b))}[f(b,a)]| + |\mathbb{E}_{\pi(a|\phi(b))}[f(b,a)] - \mathbb{E}_{\pi(a|\phi(b))}[f(\phi(b),a)]| \\
&\le \frac{R_{\max}}{1-\gamma}\varepsilon + L_Q\varepsilon
\end{aligned} \tag{118}$$

we have

$$\begin{aligned}
&|\mathcal{E}(f,\pi) - \mathcal{E}_\phi(f,\pi)| \\
&\le 2 \cdot \frac{2R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon
\end{aligned} \tag{119}$$

□

A.13 Proof of Theorem 5

Proof.

$$\begin{aligned}
&\mathcal{E}_\phi(\hat{Q}^\pi,\pi) - \mathcal{E}_\phi(\hat{Q}^\pi_\phi,\pi) + \mathcal{E}(\hat{Q}^\pi_\phi,\pi) - \mathcal{E}(\hat{Q}^\pi,\pi) \\
&= \mathcal{E}_\phi(\hat{Q}^\pi,\pi) - \mathcal{E}(\hat{Q}^\pi,\pi) + \mathcal{E}(\hat{Q}^\pi_\phi,\pi) - \mathcal{E}_\phi(\hat{Q}^\pi_\phi,\pi) \\
&\le |\mathcal{E}_\phi(\hat{Q}^\pi,\pi) - \mathcal{E}(\hat{Q}^\pi,\pi)| + |\mathcal{E}(\hat{Q}^\pi_\phi,\pi) - \mathcal{E}_\phi(\hat{Q}^\pi_\phi,\pi)| \\
&\le \frac{8R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon
\end{aligned} \tag{120}$$

where we employ Lemma 6 for the last inequality.
Using the fact that

$$\mathcal{E}_\phi(\hat{Q}^\pi,\pi) - \mathcal{E}_\phi(\hat{Q}^\pi_\phi,\pi) \ge 0 \tag{121}$$

$$\mathcal{E}(\hat{Q}^\pi_\phi,\pi) - \mathcal{E}(\hat{Q}^\pi,\pi) \ge 0, \tag{122}$$

we have

$$|\mathcal{E}_\phi(\hat{Q}^\pi,\pi) - \mathcal{E}_\phi(\hat{Q}^\pi_\phi,\pi)| \le \frac{8R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon \tag{123}$$

□

A.14 Proof of Theorem 7

Proof.

$$\begin{align}
&|J(\pi_\phi) - J(\pi)| \nonumber\\
&= |\mathbb{E}_{b\sim d_0}[V_{\text{bin}}^{\pi_\phi}(\phi(b)) - V_{\text{true}}^{\pi}(b)]| \nonumber\\
&\leq \|[V_{\text{bin}}^{\pi_\phi}]_{\text{true}} - V_{\text{true}}^{\pi}\|_\infty \nonumber\\
&\leq \|[V_{\text{bin}}^{\pi_\phi}]_{\text{true}} - [\tilde{V}_{\text{bin}}^{\pi_\phi}]_{\text{true}}\|_\infty + \|[\tilde{V}_{\text{bin}}^{\pi_\phi}]_{\text{true}} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty + \|V_{\text{true}}^{[\pi_\phi]_{\text{true}}} - V_{\text{true}}^{\pi}\|_\infty \nonumber\\
&\leq \frac{(L_\pi+1)R_{\max}\varepsilon}{1-\gamma} + L_V\varepsilon + \frac{R_{\max}L_V\varepsilon}{(1-\gamma)^2} + 2|\mathcal{O}||\mathcal{A}|\frac{\gamma R_{\max}}{(1-\gamma)^2}(1+L_\pi)\varepsilon \tag{124}
\end{align}$$

□

A.15 Proof of Theorem 8

Proof.

$$\begin{align}
&|J_{\hat{Q}^\pi}(\pi) - J_{\hat{Q}^\pi}(\pi_\phi)| \nonumber\\
&= |\mathbb{E}_{b\sim d_0}[\hat{Q}^\pi(b,\pi)] - \mathbb{E}_{b\sim d_0}[\hat{Q}^\pi(\phi(b),\pi_\phi)]| \nonumber\\
&= |\mathbb{E}_{b\sim d_0}[\hat{Q}^\pi(b,\pi) - \hat{Q}^\pi(\phi(b),\pi_\phi)]| \nonumber\\
&\leq \frac{R_{\max}}{1-\gamma}\varepsilon + L_Q\varepsilon \tag{125}
\end{align}$$

□

A.16 Proof of Theorem 10

Proof. Let $X = \mu(a_h, \tau_h^+)(r_h + V(f_{h+1})) - V(f_h)$. Then:

$$\begin{align}
&\max_{\theta\in\Theta}\mathbb{E}_{\pi_b}[\{\mu(a_h,\tau_h^+)(r_h+V(f_{h+1})) - V(f_h)\}\theta(\tau_h) - \frac{1}{2}\theta(\tau_h)^2] \nonumber\\
&= \max_{\theta\in\Theta}\mathbb{E}_{\pi_b}[X(f_1)\theta(\tau_h) - \frac{1}{2}\theta(\tau_h)^2] \tag{126}\\
&= \frac{1}{2}\max_{\theta\in\Theta}\mathbb{E}_{\pi_b}[X(f_1)^2 - (\theta(\tau_h) - X(f_1))^2] \tag{127}\\
&= \frac{1}{2}\max_{\theta\in\Theta}\mathbb{E}_{\pi_b}[X(f_1)^2 - (\theta(\tau_h) - (\mathcal{B}^{\mathcal{H}}V)(\tau_h))^2 - ((\mathcal{B}^{\mathcal{H}}V)(\tau_h) - X(f_1))^2] \tag{128}\\
&= \frac{1}{2}\mathbb{E}_{\pi_b}[X(f_1)^2 - ((\mathcal{B}^{\mathcal{H}}V)(\tau_h) - X(f_1))^2] \tag{129}\\
&= \frac{1}{2}\mathbb{E}_{\pi_b}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2] \tag{130}
\end{align}$$

To justify (128), we first show the following identity:

$$\begin{align}
\mathbb{E}_{\pi_b}[Y(\tau_h)\cdot X(f_1)] &= \mathbb{E}_{\substack{a_{1:h-1,h+1:H}\sim\pi_b \\ a_h\sim\pi_e}}[Y(\tau_h)\cdot \mathbb{E}_{\substack{a_{1:h-1,h+1:H}\sim\pi_b \\ a_h\sim\pi_e}}[r_h + V(f_{h+1})|\tau_h]] \nonumber\\
&\qquad - \mathbb{E}_{\pi_b}[Y(\tau_h)\cdot\mathbb{E}_{\pi_b}[V(f_h)|\tau_h]] \tag{131}\\
&= \mathbb{E}_{\substack{a_{1:h-1,h+1:H}\sim\pi_b \\ a_h\sim\pi_e}}[Y(\tau_h)\cdot\{(\mathcal{B}^{\mathcal{H}}V)(\tau_h) + \mathbb{E}_{\pi_b}[V(f_h)|\tau_h]\}] \nonumber\\
&\qquad - \mathbb{E}_{\pi_b}[Y(\tau_h)\cdot\mathbb{E}_{\pi_b}[V(f_h)|\tau_h]] \tag{132}\\
&= \mathbb{E}_{\pi_b}[Y(\tau_h)\cdot\{(\mathcal{B}^{\mathcal{H}}V)(\tau_h) + \mathbb{E}_{\pi}[V(f_h)|\tau_h]\}] \nonumber\\
&\qquad - \mathbb{E}_{\pi_b}[Y(\tau_h)\cdot\mathbb{E}_{\pi_b}[V(f_h)|\tau_h]] \tag{133}\\
&= \mathbb{E}_{\pi_b}[Y(\tau_h)\cdot(\mathcal{B}^{\mathcal{H}}V)(\tau_h)] \tag{134}
\end{align}$$

Here, (131) uses the fact that $\mathbb{E}_{\substack{a_{1:h-1,h+1:H}\sim\pi_b \\ a_h\sim\pi_e}}[r_h + V(f_{h+1})|\tau_h] = \mathbb{E}_{\substack{a_{1:h}\sim\pi_e \\ a_{h+1:H}\sim\pi_e}}[r_h + V(f_{h+1})|\tau_h]$.

Now, to derive (128), we prove:

$$\begin{aligned}
&\mathbb{E}_{\pi_b}[(\theta(\tau_h) - (\mathcal{B}^{\mathcal{H}}V)(\tau_h))^2 + ((\mathcal{B}^{\mathcal{H}}V)(\tau_h) - X(f_1))^2] \\
=&\mathbb{E}_{\pi_b}[\theta(\tau_h)^2 - 2\theta(\tau_h)X(f_1) + X(f_1)^2] \\
&\quad + \mathbb{E}_{\pi_b}[2\theta(\tau_h)X(f_1) - 2\theta(\tau_h)(\mathcal{B}^{\mathcal{H}}V)(\tau_h) - 2(\mathcal{B}^{\mathcal{H}}V)(\tau_h)X(f_1) + 2(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2] \quad (135) \\
=&\mathbb{E}_{\pi_b}[(\theta(\tau_h) - X(f_1))^2] \quad (136)
\end{aligned}$$

The final equality uses (134).

For (130), it also follows directly using (134). This completes the proof. □

[**Youheng:** A lot of proofs for Chapter 6 todo.]

A.17 Proof of Existence of Short-Term Memory POMDP

Proof. For the original POMDP with observation-action pair $(\mathcal{O} - \mathcal{A})$, the goal is to construct □