A Covering Framework for Offline POMDPs Learning Using Belief Space Metric

Youheng Zhu
Northwestern University

Yiping Lu
Northwestern University

Abstract

In off-policy evaluation (OPE) for partially observable Markov decision processes (POMDPs), an agent must infer hidden states from past observations, which exacerbates both the curse of horizon and the curse of memory in existing OPE methods. This paper introduces a novel covering analysis framework that exploits the intrinsic metric structure of the belief space (distributions over latent states) to relax traditional coverage assumptions. By focusing on the policies with stability property, we derive error bounds that mitigate exponential blow-ups in horizon and memory length. Our unified analysis technique applies to a broad class of OPE algorithms, yielding concrete error bounds and coverage requirements expressed in terms of belief space metrics rather than raw history coverage. We illustrate the improved sample efficiency of this framework via case studies: the double sampling Bellman error minimization algorithm, and the memory-based future-dependent value functions (FDVF). In both cases, our coverage definition based on the belief-space metric yields tighter bounds.

1 INTRODUCTION

Off-policy evaluation (OPE) is a central problem in offline reinforcement learning, aiming to estimate the expected cumulative reward of a target policy π_e using data collected under a different behavior policy π_b. This setting arises naturally in real-world applications, where interactive data collection is often impractical or unsafe, and learning must rely solely on pre-collected offline trajectories. In this paper, we consider a more realistic yet challenging setting where only partial observations of the underlying latent states are available. This leads to non-Markovian observation dynamics that may depend on the entire history of action-observation pairs. Such scenarios are modeled by partially observable Markov decision processes (POMDPs), which offer greater expressiveness for real-world problems (Atrash et al., 2009; Lauri et al., 2022) but introduce significant complexity compared to fully observable MDPs.

Although for a POMDP, Markovian is restored when treating history trajectories as states, in which case the POMDP is reduced to an MDP problem, directly applying conventional MDP methods, such as Importance Sampling and Bellman residual minimization, inevitably leads to error bounds exponentially scaling with horizon H, a phenomenon termed the *curse of horizon*. For instance, in importance sampling, the sequential importance weights grow exponentially with the horizon, leading to an intractable variance in the estimation. To alleviate this issue, a method called the Future Dependent Value Function (FDVF) is proposed for memoryless policies but fail when memory-based policies are introduced, in which case the coverage scales exponentially with the memory length, facing the *curse of memory* (Zhang and Jiang, 2024).

To overcome the curses of horizon and memory in history-as-state MDPs, we reformulate the problem in the *belief space*, a central concept in POMDPs, defined as the space of probability distributions over latent states given the observed history of actions and observations. Each element in belief space—referred to as a belief state—serves as a proxy for historical trajectories. As an explicit computation of a belief state requires the latent dynamic to be transparent to the agent, it is most commonly used in POMDP planning literature. Utilizing the metric structure of belief spaces, planning methods like point-based value iteration (PBVI) achieve efficient solutions by sparsely covering belief subspaces (Shani et al., 2013; Lee et al., 2007; Zhang et al., 2014). Although belief-space structure has been

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

extensively studied in POMDP planning (Lee et al., 2007) and model learning (Zhang et al., 2012), its role in off-policy evaluation (OPE) remains largely under-explored. Notably, current offline learning approaches typically neglect this metric structure, treating history spaces explicitly, resulting in exponential dependence on the horizon length. This raises a critical question:

While the metric structure of belief space has proven effective for characterizing computational complexity in POMDP planning, can it similarly characterize sample complexity in offline learning? More specifically, can we exploit this belief space structure to circumvent exponential complexity in offline POMDP learning?

Our Contributions Motivated by this question, our work explores the idea of belief space metric structure, and studies the theoretical guarantees of some common model-free OPE algorithms using belief metric. The core idea of our framework is similar to that of state abstraction (Li et al., 2006), given that the complexity of belief space can be lowered significantly through an abstraction that contracts similar states. That is to say, if two history trajectories have similar belief states, they should be considered similar in the analysis. To do this, we restrict ourselves to a **subset of policies**, i.e. the policies with *stability*. This structural assumption on policy class is *rich enough* to contain all possible policies of our interest, and possesses nice properties for tighter coverage. The overall result of our analysis with comparison to existing results that suffer from the curse of horizon/memory is presented in Table 1 below. In general, our result mitigates the exponentiality of coverage especially under smoothness structure of belief space as shown in Example 1 and 2. To specify our contributions:

- We propose a framework of analysis that uses state abstraction induced by ε-covering to obtain a coverage on the abstract space, which adapts to a wide range of scenarios in the OPE problem. This framework easily generalizes to other algorithms or even other reinforcement learning tasks.
- We show in Table 1, Theorem 4 and 5 that our coverage obtained using belief space covering is no worse than the original coverage. We also show in Example 1 and 2 that our coverage resolves the curse or horizon/memory under specific smoothness property of the POMDP model.
- In Section 5.1, we complete a detailed analysis specifically for double sampling algorithm as an example of Bellman error minimization algorithms. In Section 5.2, we also present the pipeline on future dependent value function where the fast forgetting properties of POMDP and policies are adopted. We then show that FDVF admits a simpler analysis, free from any assumptions on the POMDP system itself. **This indicates that the "curse of memory" is much easier to handle than the "curse of horizon"**. Additionally, this provide an answer to the question left by (Zhang and Jiang, 2024), that with structural assumption on the policy, we can mitigate the "curse of memory".

2 RELATED WORKS

POMDP planning. In POMDP planning literature, the idea of point-based value iteration (PBVI) (Shani et al., 2013; Kurniawati et al., 2009; Poupart et al., 2011; Shani et al., 2008; Smith and Simmons, 2012; Spaan and Vlassis, 2005) is to compute on a finite subset of the entire belief space, aiming for an optimal policy. Notably, an important characteristic of PBVI is that its selection of belief subspace uses the metric structure in belief space, namely, every time the algorithm expands the belief subset, it searches for the furthest belief point w.r.t. the current belief subset that is one-step reachable, so that the reachable belief subset can be constructed as sparse as possible. Additionally, the connection between complexity and belief space metric was identified by (Lee et al., 2007; Zhang et al., 2014), which proved the existence of approximate algorithm with complexity polynomial to the covering number of reachable belief space.

Curse of Horizon and Curse of Memory in OPE. Numerous algorithms have addressed Off-Policy Evaluation (OPE) in fully observable MDPs, such as Importance Sampling (Precup et al., 2000; Jiang and Li, 2016; Jiang et al., 2019; Hu and Wager, 2023), Fitted Q-Iteration (FQE) (Ernst et al., 2005; Munos and Szepesvári, 2008; Le et al., 2019), Bellman residual minimization with double sampling (Baird III, 1995), min-max estimators (Antos et al., 2008; Chen and Jiang, 2019; Feng et al., 2019; Nachum et al., 2019; Uehara et al., 2021; Zanette and Wainwright, 2022), and marginalized importance sampling (Uehara et al., 2020). However, directly applying these approaches to Partially Observable MDPs (POMDPs) by treating each trajectory history as a distinct state encounters a fundamental challenge known as the *curse of horizon*: the error bounds become exponentially worse as the horizon grows, due to coverage assumptions expanding with the exponentially large history space. Alternatively, recent approaches such as the Future Dependent Value Function (FDVF) (Uehara et al., 2023; Zhang and Jiang, 2024) address this by shifting coverage requirements onto latent states, thus providing polynomial guarantees for memoryless policies. Nevertheless, this method is constrained by the *curse of memory*, as its complexity reverts to exponential when

extended to memory-based policies, due to the necessity of capturing dependencies between future observations and historical memory states, dramatically increasing coverage complexity.

3 PRELIMINARIES

Infinite-horizon Discounted POMDP: An infinite-horizon discounted POMDP can be specified as 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 kernel (i.e., the conditional distribution of the observation given the state), and $\mathbb{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel (i.e., the conditional distribution of the next state given the current state-action pair). We use $\Delta(\cdot)$ to represent probability distributions on the given space, and $|\cdot|$ for the cardinality of a set. For simplicity and without loss of generality, we assume discrete and finite spaces $\mathcal{S}, \mathcal{A}, \mathcal{O}$, of which the cardinality can be large.

The POMDP evolves as follows: starting from an initial latent state $s_1 \sim d_0(s)$, at each step h, the latent state s_h emits an observation o_h drawn from $\mathbb{O}(s_h)$, and the environment generates a reward r_h based on the current state-action pair (s_h, a_h). The state then transitions according to $s_{h+1} \sim \mathbb{T}(s_h, a_h)$. Crucially, in general POMDPs, the learner has no access to the latent state space $\mathcal{S}$; instead, only trajectories collected under an offline behavior policy are available.

We also consider the finite-horizon POMDP setting extensively discussed in Section 5.2. In the finite-horizon scenario, we set the discount factor $\gamma = 1$, and the agent interacts with the environment for a finite number of steps H.

Offline Data: 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 of $\mathcal{D}_1$ as shown below. In Section 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, as shown by $\mathcal{D}_2$.

$$\mathcal{D}_1 = \{(o_1^{[i]}, a_1^{[i]}, r_1^{[i]}, \cdots, o_{h_i}^{[i]}, a_{h_i}^{[i]}, r_{h_i}^{[i]}, o_{h_i+1}^{[i]})\}_{i=1}^n,$$

$$\mathcal{D}_2 = \{((o_1^{[i]}, a_1^{[i]}, r_1^{[i]}, \cdots, o_H^{[i]}, a_H^{[i]}, r_H^{[i]})\}_{i=1}^n$$

State Abstraction: For an MDP $\langle \mathcal{S}, \mathcal{A}, r, \gamma, P \rangle$ where $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ denotes the transition kernel, an abstraction ϕ is a mapping from $\mathcal{S}$ to an abstract state space $\mathcal{S}_\phi$, and the MDP is transformed into an abstract MDP $\langle \mathcal{S}_\phi, \mathcal{A}, r_\phi, \gamma, P_\phi \rangle$ where $r_\phi(\phi(s), a) := \mathbb{E}_{s' \sim p_{\phi(s)}}[r(s', a)]$ and $P_\phi(\phi(s_d)|\phi(s), a) := \mathbb{E}_{s' \sim p_{\phi(s)}}[\sum_{\phi(s'')=\phi(s_d)} P(s''|s', a)]$. Here $\{p_x\}_{x \in \mathcal{S}_\phi}$ is any family of distributions in which p_x being supported on $\phi^{-1}(x)$. For any function defined on the abstract system $f_{\text{bin}} : \mathcal{S}_\phi \to \mathbb{R}$, we define the lifted version of which as $[f_{\text{bin}}]_{\text{true}}(\cdot) := f_{\text{bin}}(\phi(\cdot))$. Similar for an abstract policy $\pi_\phi : \mathcal{S}_\phi \to \Delta(\mathcal{A})$, of which the lifted version $[\pi_\phi]_{\text{true}}(\cdot) := \pi_\phi(\phi(\cdot))$ In the following section, ϕ is often selected by ε, and is treated as equivalent. Conventionally, notations with super/subscripts ϕ is also used to specify functions defined on the abstract system, and whenever we say $f_\phi \in \mathcal{F}$ where $\mathcal{F}$ is a function class defined on the true system, we mean $\exists f \in \mathcal{F}, f(\phi(\cdot)) = f_\phi(\phi(\cdot))$.

Other Notations. For any random variable X_t in a POMDP, define its discounted occupancy under policy π by $d^\pi(x) := (1-\gamma)\sum_{k=0}^\infty \gamma^k \Pr_\pi(X_k = x)$. In particular, for state-action occupancy, $d^\pi(s, a) := (1-\gamma)\sum_{k=0}^\infty \gamma^k \Pr_\pi(S_k = s, A_k = a)$. We use d^D for the population distribution induced by the offline data-collection process, and $\mathbb{E}_{\mathcal{D}}[\cdot]$ for empirical expectation over the offline dataset. $J(\pi)$ represents the expected reward of a policy π, and $J_{\hat{Q}}(\pi)$ is the estimated reward of π using approximation function $\hat{Q}$.

4 UNIFIED ANALYSIS OVERVIEW

In this section, we briefly explain how the geometry of the belief state space can help characterize the sample complexity of off-policy evaluation for POMDPs, and what our result looks like in general. We also present the basics of belief space, abstraction on the belief space induced by an ε-cover, and the assumptions related to the belief metric.

Belief State Space and Smoothness Condition: Since one cannot observe the latent state directly, a prediction of the current state can be made using the information from the entire history of observations and actions. We denote the history at time step h to be $\tau_h = (o_1, a_1, o_2, a_2, \cdots, o_{h-1}, a_{h-1}) \in \mathcal{H}_h \subset \mathcal{H}$ and $\tau_h^+ := (\tau_h, o_h) \in \mathcal{H}_h^+ \subset \mathcal{H}^+$. Consequently 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 $\mathcal{B} = \{b : \exists h \in \mathbb{N}\ \exists \tau_h^+, \mathbf{b}(\tau_h^+) = b\}$. Consider a common case when such $\mathbf{b}$ is a bijection, then $\mathcal{B}$ becomes a perfect proxy for $\mathcal{H}$, of which the

Table 1: Comparison With Existing Coverage

Criteria	Existing Coverage With Curse of Horizon/Memory		Our Coverage using Belief Space Smoothness
Bellman Error Minimization (*e.g.* Double Sampling)			
Coverage Definition (Jiang and Xie, 2025)	$\left\|\frac{d^{\pi_e}(\tau_h,a)}{d^D(\tau_h,a)}\right\|_\infty$		$\left\|\frac{d_e^{\phi^{\pi_e}}(\phi(b),a)}{d^D(\phi(b),a)}\right\|_\infty$
Coverage Worst Case[1] Scale	$\lvert\mathcal{B}\rvert = \Theta((\lvert\mathcal{O}\rvert\lvert\mathcal{A}\rvert)^H)$	$>$	Covering$(\mathcal{B}, \Theta(n^{-1/2}))$[2]
Ability to handle $H \to \infty$	✘: Infinite		✓: Polynomial guarantee see example 1
Future Dependent Value Function			
Coverage Definition (Zhang and Jiang, 2024)	$\sup_{h,V}\sqrt{\frac{\mathbb{E}_{\pi_e}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_H)}V)(s_h,\tau_h)^2]}{\mathbb{E}_{\pi_b}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}}$		$\sup_{h,V}\sqrt{\frac{\mathbb{E}_{\pi_e^\phi}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h,\tau_{[h-T+1:h]})^2]}{\mathbb{E}_{\pi_b^\phi}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}}$
L_2 Belief Coverage (One-hot Belief) (Zhang and Jiang, 2024)	$\mathbb{E}_{\pi_b}\left[\left(\frac{d^{\pi_e}(s_h,\tau_h)}{d^{\pi_b}(s_h,\tau_h)}\right)^2\right]$	$>$ Theorem 4	$\mathbb{E}_{\pi_b^\phi}\left[\left(\frac{d_\phi^{\pi_e^\phi}(s_h,\tau_{[h-T+1:h]})}{d_\phi^{\pi_b^\phi}(s_h,\tau_{[h-T+1:h]})}\right)^2\right]$
L_∞ Belief Coverage (One-hot Belief) (Zhang and Jiang, 2024)	$\left\|\frac{d^{\pi_e}(s_h,\tau_h)}{d^{\pi_b}(s_h,\tau_h)}\right\|_\infty$	$>$ Theorem 5	$\left\|\frac{d_\phi^{\pi_e^\phi}(s_h,\tau_{[h-T+1:h]})}{d_\phi^{\pi_b^\phi}(s_h,\tau_{[h-T+1:h]})}\right\|_\infty$
L_∞ Worst Case[1] (One-hot Belief)	$\Theta((\lvert\mathcal{O}\rvert\lvert\mathcal{A}\rvert)^H)$	$>$	$\Theta((\lvert\mathcal{O}\rvert\lvert\mathcal{A}\rvert)^T)$
Ability to handle $H \to \infty$[3]	✘: Infinite		✓: Polynomial guarantee see example 2

cardinality grows exponentially with the horizon. In infinite horizon cases, $|\mathcal{B}| = \infty$, yet considering the compactness of a bounded subset of $\mathbb{R}^{|\mathcal{S}|}$, cluster points of $\mathcal{B}$ must exist. For simplicity, we assign distinct belief copies to histories that share the same belief state distribution, making the belief space metric a pseudo-metric. 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. Similarly for value function $\tilde{V}(\tau_h^+) = V(\mathbf{b}(\tau_h^+))$. Since $\tilde{V}, \tilde{\pi}, \tau_h^+ \in \mathcal{H}^+$ one-to-one correspond to $V, \pi, b \in \mathcal{B}$, we slightly abuse our notation and treat them as equivalent for the rest of the passage, i.e. whenever we mention $b \in \mathcal{B}$, we also mean the corresponding $\mathbf{b}^{-1}(b) \in \mathcal{H}^+$, especially when describing the algorithms, since they only see the data trajectories instead of actual beliefs.

Then we introduce the core idea of using belief space metric to lower the complexity of the potentially exponential belief space, that is through covering. By introducing an ε-cover as a abstraction of the original belief space, we can treat near belief states as one, making the space simpler. This is formalized below with a similar idea as an *ε-sufficient statistic* in (François-Lavet et al., 2019; Subramanian et al., 2022).

Abstraction Induced by Covering. Consider the belief space $\mathcal{B}$, for any $\varepsilon > 0$ and an ε-cover $\mathcal{C}_\varepsilon \subset \mathcal{B}$ (Defined in Appendix B). There exists an abstraction $\phi : \mathcal{B} \to \mathcal{C}_\varepsilon$ such that $\forall b \in \mathcal{B}$, $\|\phi(b) - b\|_1 \le \varepsilon$. Select any such ϕ, and a family of measure $\{p_x\}_{x\in\mathcal{C}_\varepsilon}$ mentioned in Section 3, then an abstract belief MDP is defined, we refer to which as the abstract system.

To obtain a meaningful result, it is important for us to limit our attention to a subset of all possible policies, i.e. those that presents stability. This is characterized by the two core structural assumptions on the policy of interest, primarily introduced in Lipschitz-MDP literature (Pirotta et al., 2015; Gelada et al., 2019):

Assumption 1 (Local Stability). $\forall b_1, b_2 \in \mathcal{B}$, $\|\pi(b_1) - \pi(b_2)\|_1 \le L_\pi \|b_1 - b_2\|_1$.

Assumption 2 (Value Stability). $\sup_{\substack{b_1,b_2\in\mathcal{B}\\ \varepsilon\ge 0,\phi_\varepsilon}} |V^{[\pi_{\phi_\varepsilon}]_{\text{true}}}(b_1) - V^{[\pi_{\phi_\varepsilon}]_{\text{true}}}(b_2)|/\|b_1 - b_2\|_1 \le L_V < \infty$.

Remark 1. *Assumption 1 is made by the intuition*

[1] "Worst-case coverage" refers to the worst case for the most exploratory data-collection distribution.

[2] Covering$(\mathcal{B}, \varepsilon)$ denotes the L_1 covering number of $\mathcal{B}$.

[3] For $H \to \infty$, we assume worst-case coverage grows as a subpolynomial power $\alpha_0 \le 1$ (not logarithmic, which would trivially remove the curse of horizon). In the FDVF case, specific forgetting rates may be required.

that a good belief state policy should treat two similar belief state similarly, and thus should itself have some local stability. Assumption 2 measures the stability of a policy's long-term return. As indicated by the following Theorem 1, it can also be viewed as a proxy for how closely a policy resembles the optimal policy.

Theorem 1 (Lemma 1 in (Lee et al., 2007)). *For any* $b_1, b_2 \in \mathcal{B}$, $|V^*(b_1) - V^*(b_2)| \leq \frac{R_{\max}}{1-\gamma}\|b_1 - b_2\|_1$.

Thus, V^* is Lipschitz in the belief metric with constant $\frac{R_{\max}}{1-\gamma}$. This provides intuition that "good" policies may admit moderate value-stability constants. For the optimal policy specifically, if one instead compares with the lifted optimal policy of the abstract system, Theorem 27 of (Subramanian et al., 2022) gives a corresponding abstraction error of order $O(R_{\max}/(1-\gamma)^2)$. Apart from the inherent stability of optimal value, people have studied this stability property not just in POMDPs, but also in continuous state MDPs. This line of work, e.g. (Pirotta et al., 2015; Asadi et al., 2018; Gelada et al., 2019), were explored in various contexts, but in such cases, the stability in value weren't as natural as in POMDPs, since unlike in belief spaces, the system dynamic in a continuous state MDP may not be smooth w.r.t. its intrinsic metric.

In general, the two assumptions hold with some finite constant L_π and L_V, but the worst case scaling of them could be exponentially large. However, policies with malignant stability are often bad and uninteresting, and it is efficient for us to only study the behavior of those good policies. With that said, our analysis applies to both cases, and the smaller the stability constants are, the more tractable our bound becomes. Either way, our bound will be no worse than the original.

4.1 Unified Analysis In a Nutshell

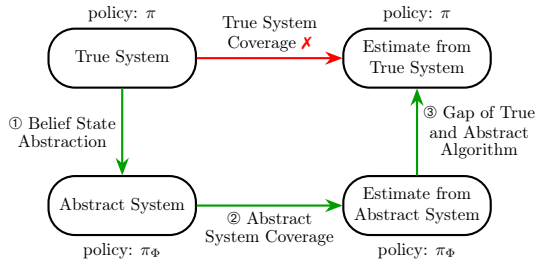

Figure 1: Pipeline of the analysis

Specifically as shown in Figure 1, in step 1, we descend the true belief space MDP system (resp. policy π) to an abstract system (resp. abstract policy π_ϕ). Using similar ideas of state abstraction, we control the abstraction error using the size of bins ε. In step 2, we execute the algorithm on the abstract system, with the coverage assumption for the abstract belief space, which can be much more tractable than the coverage of the true system due to the curse of horizon. We also provide Theorems 4, and 5 to show that abstract coverage is no worse than the original coverage. Eventually for step 3, we utilize the stability property of value function again to control the difference between the real and the virtually executed algorithm on the same offline data. Combining all the analysis above, we obtain an estimation error bound without incorporating the traditional coverage assumption.

A simple toy example illustrating why the history space can grow exponentially while the belief-space covering number remains small is deferred to Appendix E, this captures the intuition behind the red arrow and green arrow 2 in Figure 1.

In this paper, we construct the abstraction using an ε-cover $\mathcal{C}_\varepsilon$ of the belief space, with definition in Appendix B. We state the following helpful lemma for controlling abstraction error.

Lemma 1. $\forall a \in \mathcal{A}, b_1, b_2 \in \mathcal{B}$, $\mathbb{E}_{o\sim P(\cdot|b_1,a)}[\|b_1^{o,a} - b_2^{o,a}\|_1] \leq 2\|b_1 - b_2\|_1$. *Here* $b^{o,a}$ *denotes the updated next belief of* b *after taking action* a *and observing* o.

Remark 2. *Intuitively, after a pair of same action and observation* (a, o)*, the chances of two belief states sharing the same state becomes larger, resulting in the two next belief states become closer, i.e. a data processing inequality (DPI) should hold. However, such contraction property generally does not hold point wise as indicated in example 3, which also show that the Lipschitz value can go to infinity. The reason for that is that the belief update is a posterior instead of a Markov kernel, and a DPI only applies to the latter. However, the lemma shows that after taking expectation, the value is bounded by* 2*. The proof can be found in Appendix A.*

Proposition 1. *If for any* $\tau_h \in \mathcal{H}$, $\mathbf{b}(\tau_h)$ *is one-hot, then* 2 *in Lemma 1 can be replaced with* 1.

Theorem 2. *Under Assumption 1 and 2, for an abstraction* ϕ_ε *depending on* ε*, we have* $\|V^\pi_{\text{true}} - [V^{\pi_\phi}_{\text{bin}}]_{\text{true}}\|_\infty \leq L_\phi^{[1]}\varepsilon$*, where* $L_\phi^{[1]} := \frac{(L_\pi+1)R_{\max}+2L_V}{1-\gamma} + \frac{\gamma R_{\max}L_\pi + R_{\max}}{(1-\gamma)^2}$*. See Appendix B for the proof.*

Remark 3. *For finite horizon POMDP, simply replace* $(1-\gamma)^{-1}$ *with* H.

Therefore, our previous assumptions enable a principled reduction from the exponentially large belief MDP to an abstract belief MDP, with a tractable approximation error. The abstract state space has cardinality on the order of the covering number, potentially mitigating the curse of horizon, as illustrated in Examples 1 and 2. Formally speaking, we have the following meta-theorem, with the proof in Appendix A.

Theorem 3 (Meta-theorem). *Let* $\mathcal{M}$ *be a POMDP,* π

a policy, and an OPE procedure $\text{Alg} := \{\mathfrak{alg}, \mathfrak{est}\}$*, where* $\mathfrak{alg} : \mathcal{D} \to \mathcal{V}$ *learns an object* $\hat{Q}^\pi \in \mathcal{V}$ *from an offline dataset* $\mathcal{D}$ *of size* n*, and* $\mathfrak{est}$ *maps a learned object in* $\mathcal{V}$ *to an estimated value. We omit the dependence of* Alg *on* π *and on the offline data* $\mathcal{D}$ *when clear from context.*

Then for any $\varepsilon \geq 0$ *and an abstraction* $\phi : \mathcal{B} \to \mathcal{B}$ *such that* $\forall\, b_1, b_2 \in \mathcal{B}$*,* $\phi(b_1) = \phi(b_2) \Rightarrow \|b_1 - b_2\|_1 \leq \varepsilon$*, we denote the algorithm executed on the abstract system as* $\text{Alg}^\phi := \{\mathfrak{alg}^\phi, \mathfrak{est}^\phi\}$*. If assumption 1, 2 holds, and that there exists an* $L_\phi^{[2]}$ *such that* $|\mathfrak{est}(Q) - \mathfrak{est}^\phi(Q)| \leq L_\phi^{[2]}\varepsilon$ *for all* $Q \in \mathcal{V}$*, we also consider when* Alg *admits a finite sample estimation error on the abstract system of the form* $|\mathfrak{est}^\phi(\hat{Q}^\pi) - \mathfrak{est}^\phi(Q_\phi^\pi)| \leq C_\pi^\phi \cdot \sqrt{\|\mathcal{V}\|_\infty \cdot \left(\frac{1}{n}\log\frac{|\mathcal{V}|}{\delta}\right)^\alpha + L_\mathcal{E}\varepsilon}$*, w.p.* $> 1-\delta$*, where* C_π^ϕ *is a constant,* $\|\mathcal{V}\|_\infty$*,* $|V|$ *respectively denotes the boundedness and cardinality of the function class for function approximation. Then we have* $|\mathfrak{est}(\hat{Q}^\pi) - \mathfrak{est}(Q^\pi)| \leq L_\phi + C_\pi^\phi \cdot \sqrt{\|\mathcal{V}\|_\infty \cdot \left(\frac{1}{n}\log\frac{|\mathcal{V}|}{\delta}\right)^\alpha + L_\mathcal{E}\varepsilon}$*, w.p.* $> 1-\delta$*. Here,* $L_\phi := L_\phi^{[1]} + L_\phi^{[2]}$ *with* $L_\phi^{[1]}$ *defined in Theorem 2,* Q^π *and* Q_ϕ^π *represent the ground truth estimators.*

4.2 Why Coverage on Covering is Better?

In the following part, we showcase the general idea why our coverage is no worse than the original coverage by providing the two theorems as a complement to our Table 1. Since directly comparing the occupancy of π_e and the abstract occupancy of π_e^ϕ is difficult, so we turn to comparing the occupancy of $[\pi_e^\phi]_{\text{true}} := \tau_h \mapsto \pi_e^\phi(\phi(\tau_h))$, which generally have the same scaling as that of π_e. Proving the theorems (see Appendix E) uses an information-theoretic idea that the divergence between two probability measures becomes smaller on a coarser σ-algebra, using the variational representation of f-divergences.

Theorem 4. *Consider the* L_2 *belief coverage in the one-hot scenario. Then for any behavior policy* π_b *and truncation abstraction* ϕ_T*, there exists a* $d_\phi^D \in \Delta(\mathcal{S}^\phi \times \mathcal{H}_T)$*, such that for any* π_e*, along with its abstract policy* π_e^ϕ *and the corresponding lifted version* $[\pi_e^\phi]_{\text{true}}$*, we have* $\mathbb{E}_{d_\phi^D}\left[\left(\frac{d_\phi^{\pi_e^\phi}(s_h,\tau_{[h-T+1:h]})}{d_\phi^D(s_h,\tau_{[h-T+1:h]})}\right)^2\right] \leq \mathbb{E}_{\pi_b}\left[\left(\frac{d^{[\pi_e^\phi]_{\text{true}}}(s_h,\tau_h)}{d^{\pi_b}(s_h,\tau_h)}\right)^2\right]$.

Theorem 5. *Same result for the* L_∞ *belief coverage that* $\left\|\frac{d_\phi^{\pi_e^\phi}(s_h,\tau_{[h-T+1:h]})}{d_\phi^D(s_h,\tau_{[h-T+1:h]})}\right\|_\infty \leq \left\|\frac{d^{[\pi_e^\phi]_{\text{true}}}(s_h,\tau_h)}{d^{\pi_b}(s_h,\tau_h)}\right\|_\infty$.

Next, we provide illustrative examples to show the superiority our result under certain structures.

Example 1. *Consider a belief space with smoothness structure [Detailed Definition in Appendix E]. With coverage sublinear polynomial to the worst case, we have a finite sample guarantee of* $O(\frac{(C|\mathcal{S}|L_\mathcal{E} m R_{\max}^2)^{\frac{1}{4}}}{(1-\gamma)^{\frac{3}{2}}} \cdot (\frac{1}{n}\log\frac{|\mathcal{F}|}{\delta})^{\frac{1}{8}})$*, where* C*,* m *are constants related to the smoothness property.*

Example 2. *Consider a fast forgetting policy with forgetting speed* $T(\varepsilon) = O(\log\frac{1}{\varepsilon})$*, then with coverage sublinear polynomial to the worst case, we can obtain a finite sample guarantee of* $O(\frac{\max\{\|\mathcal{V}\|_\infty, \|\Theta\|_\infty\}}{(1-\gamma)^2} \cdot (\frac{1}{n}\log\frac{|\mathcal{V}||\Theta|}{\delta})^{\frac{1}{4}})$*. If we make a even stronger assumption than logarithmical scaling memory, i.e. strictly short-term memory, then the result goes back to what's discussed in (Uehara et al., 2023; Zhang and Jiang, 2024).*

5 EXAMPLES OF APPLICATION

In this section, we apply our unified analysis on two different types of OPE algorithms, namely, the double sampling Bellman error minimization algorithm and future-dependent value function, aiming for a more sample efficient guarantee.

5.1 Analysis on Bellman Error Minimization Algorithms

Double Sampling. Consider a Bellman error minimization algorithm using double sampling, each offline data contains two tuple (b, a, r, b'_A) and (b, a, r, b'_B) with the latter sampled independently after the system resets to belief b. The corresponding estimator can be written as $\hat{Q}^\pi = \arg\min_{f\in\mathcal{F}} \mathcal{E}(f, \pi)$ where $\mathcal{E}(f,\pi) = \mathbb{E}_\mathcal{D}[(f(b,a) - (r + \gamma f(b'_A, \pi)))(f(b,a) - (r + \gamma f(b'_B, \pi)))]$

Instead of assuming standard coverage on the true system, we adopt the following abstract covering assumption on the abstract system.

Assumption 3 (Abstract Policy Coverage)**.** $\|d^{\pi_\phi}/d^D\|_\infty \leq C_\pi(\phi) < \infty$

Remark 4. *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 3 lies in its potential to outperform coverage assumptions in the original space. Using an abstract belief space allows the exponentially large history space to be reduced to a space with size of* ε*-covering number.*

And also a standard realizability assumption.

Assumption 4 (Abstract Realizability). *$Q_\phi^{\pi_\phi} \in \mathcal{F}$, which according to our notation, is short for $\exists f \in \mathcal{F}, f(\phi(\cdot)) = Q_\phi^{\pi_\phi}(\phi(\cdot))$ since $Q_\phi^{\pi_\phi}$ is defined on the abstract system.*

Noticed that we previously assumed the stability of value function, whose equivalence to the Lipschitz 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 Lipschitz with regard to belief state.

Assumption 5 (Lipschitz function class). *$\forall f \in \mathcal{F}$, $\forall a \in \mathcal{A}$, $|f(b_1, a) - f(b_2, a)| \leq L_Q \|b_1 - b_2\|_1$.*

Then, we can provide the value of $L_\phi^{[2]}$ defined in Theorem 3 for this special case, and furthermore, the eventual guarantee for double sampling algorithm using the aforementioned assumptions and methods of analysis. Proofs in Appendix C.

Theorem 6. *If Assumption 5 holds, then for $L_\phi^{[2]}$ defined in Theorem 3, $L_\phi^{[2]} = \frac{R_{\max}}{1-\gamma} + L_Q$.*

Theorem 7. *If Assumptions 4, 5, 1, and 2 all hold, then we have:*

$$|J_{\hat{Q}^\pi}(\pi) - J(\pi)| \leq \inf_{\substack{\varepsilon \geq 0 \\ D(\varepsilon)}} \left(\frac{\sqrt{C_\pi(\varepsilon)}}{1-\gamma} \cdot \sqrt{\sqrt{\frac{32 R_{\max}^4}{n(1-\gamma)^4} \cdot \log \frac{2|\mathcal{F}|}{\delta}} + L_{\mathcal{E}} \varepsilon} + L_\phi \varepsilon \right)$$

where $L_{\mathcal{E}} = \frac{8R_{\max}}{1-\gamma} \cdot \big((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\big)$, L_ϕ is defined as in Theorem 3 and $D(\varepsilon)$ stands for such ε that satisfies realizability (Assumption 4).

Corollary 1 (Finite sample guarantee). *If Assumptions 5,1, and 2 all hold, then for all n satisfying $n \geq 8R_{\max}^4 (L_\phi / L_{\mathcal{E}})^4 \log(2|\mathcal{F}|/\delta)$, and the abstraction ϕ induced by ε-cover with $\varepsilon = \frac{1}{L_{\mathcal{E}}} \sqrt{\frac{32 R_{\max}^4}{n(1-\gamma)^4} \cdot \log \frac{2|\mathcal{F}|}{\delta}}$ satisfies Assumption 4, we have $|J_{\hat{Q}^\pi}(\pi) - J(\pi)| \leq \frac{2\sqrt{C_\pi^n}}{1-\gamma} \cdot \big(\frac{128 R_{\max}^4}{n(1-\gamma)^4} \cdot \log \frac{2|\mathcal{F}|}{\delta}\big)^{\frac{1}{4}}$, where $C_\pi^n := C_\pi\big(\frac{1}{L_{\mathcal{E}}} \sqrt{\frac{32 R_{\max}^4}{n(1-\gamma)^4} \cdot \log \frac{2|\mathcal{F}|}{\delta}}\big)$.*

Remark 5. *The guarantee obtained using our method relies on the coverage defined on the abstract system, which is more tractable than the original coverage as discussed in Remark 4 and Table 1. Moreover, with appropriate belief space smoothness condition (Example 1), our result provides a polynomial finite sample guarantee while the original bound goes to infinity.*

5.2 Future-Dependent Value Function.

FDVF was proposed targeting memoryless policies. Here we introduce the memory-based version of FDVF, which suffers from the "curse of memory" as discussed in (Zhang and Jiang, 2024). We first introduce the respective definition of future space $\mathcal{F}'$ as $f'_h := (o_h, a_h, o_{h+1}, a_{h+1}, \cdots, o_H, a_H) \in \mathcal{F}'_h \subset \mathcal{F}'$.

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. The future-dependent value function $V_{\mathcal{F}}$ is any such function that satisfies $\mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_h, \tau_h)|s_h, \tau_h] = V_{\mathcal{S}}^{\pi_e}(s_h, \tau_h)$ with the RHS being the value function of π_e, and is a zero point of the following two Bellman Residual Operators.

Definition 1 (Memory-Based Bellman Residual Operator). *We define $(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h, \tau_{[h-T+1:h]}) := \mathbb{E}_{\substack{a_{1:h}\sim\pi_e \\ a_{h+1:H}\sim\pi_b}}[r_h + V(f_{h+1})|s_h, \tau_{[h-T+1:h]}] - \mathbb{E}_{\substack{a_{1:H-1}\sim\pi_e \\ a_{h:H}\sim\pi_b}}[V(f_h)|s_h, \tau_{[h-T+1:h]}]$, and $(\mathcal{B}^{\mathcal{H}}V)(\tau_h) := \mathbb{E}_{\substack{a_{1:h}\sim\pi_e \\ a_{h+1:H}\sim\pi_b}}[r_h + V(f_{h+1})|\tau_h] - \mathbb{E}_{\substack{a_{1:h-1}\sim\pi_e \\ a_{h:H}\sim\pi_b}}[V(f_h)|\tau_h]$.*

Memory-Based Algorithm. For memory-based policies, we define $\mu(a_h, \tau_h^+) := \frac{\pi_e(a_h|\tau_h^+)}{\pi_b(a_h|\tau_h^+)}$, then the min-max algorithm is defined as follows:

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

FDVF Analysis Pipeline. The analysis of FDVF follows a structured framework that uses the previously introduced methodology, of which an illustrative Figure 2 can be found in Appendix D. All proofs can also be found in Appendix D.

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 window T.

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

To reuse the previous analysis, we also introduce an abstraction mapping $\phi : \mathcal{B} \to \mathcal{B}$ that operates directly on belief states. The mapping ϕ satisfies $\phi(\mathbf{b}(\tau_h^+)) = \mathbf{b}(\tilde{\phi}(\tau_h^+))$. Since this mapping ϕ depends on the time window length T, we denote it as ϕ_T. Notice that ϕ_T and $\tilde{\phi}_T$ are one-to-one, we treat them equivalently. Now we provide the fast-forgetting assumption of POMDP and the policy.

Assumption 6 (Fast-Forgetting 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 $\|b_1 - b_2\|_1 \leq \varepsilon$. The values of T satisfying this condition form a function of ε, denoted $T_0(\varepsilon)$.*

Assumption 7 (Fast-Forgetting Policy)**.** *For the abstraction mapping ϕ_T, it holds that for all $\varepsilon > 0$, there exists a $T \in \mathbb{N}^+$, 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 $\|\pi(\tau_h^{[1]+}) - \pi(\tau_h^{[2]+})\|_1 \leq L_\pi \varepsilon$. We denote the dependency of T on ε as $T_1(\varepsilon)$.*

Lemma 2 (Stability implies Fast-Forgetting)**.** *If Assumption 6 and Assumption 1 hold, then Assumption 7 holds automatically, with $T_1 = T_0$.*

Conditions: Controlling Differences between Real and Abstract Algorithm. Since our analysis is build on the requirement that the virtually executed algorithm and the actual algorithm bear little difference, we first propose some conditions to restrain ε from being too large.

Definition 2. *We define $\|\mathcal{V}\|_\infty := \max_{V \in \mathcal{V}} \|V\|_\infty$ (similar for Θ), $C_\mathcal{V} := \max\{\|\mathcal{V}\|_\infty + 1, \|\Theta\|_\infty\}$, $C_\mu := \max_h \max_{a_h, \tau_h^+} \mu(a_h, \tau_h^+)$, and $L_\mathcal{E} := 3\big(\frac{2H(C_\mu+1)L_\pi \|\mathcal{V}\|_\infty \|\Theta\|_\infty}{\min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)} + HC_\mu \|\mathcal{V}\|_\infty \|\Theta\|_\infty + \frac{3H^2 \max\{C_\mu \|\mathcal{V}\|_\infty \|\Theta\|_\infty, \frac{1}{2}\|\Theta\|_\infty^2\}}{\min\{\min_h \min_{o_h,\tau_h} P(o_h|\tau_h), \min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)/L_\pi\}}\big)$.*

Condition 1. *The ε is small enough that $L_\pi \varepsilon / \min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+) \leq \frac{1}{2}$.*

Condition 2. *The ε is small enough that $\frac{H\varepsilon}{\min\{\min_h \min_{o_h,\tau_h} P(o_h|\tau_h), \min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)/L_\pi\}} \leq 1$*

Condition 1 assumes non-zero entries for the behavior policy π_b, which is known and chosen by the learner. This assumption is also used in literature like (Zhang and Jiang, 2025), from which we adopt the same philosophy. In Condition 2, the probability $P(o_h|\tau_h)$ being uniformly above zero is a non-trivial assumption, but we'll show later in a simpler pipeline that this condition can actually be discarded.

We then state the following theorem about $L_\phi^{[2]}$.

Theorem 8. *If Assumptions 6, 7 and 9 hold, then for any $\varepsilon > 0$, with $T \geq \max\{T_0(\varepsilon), T_1(\varepsilon), T_2(\varepsilon)\}$, we have $L_\phi^{[2]} = \|\mathcal{V}\|_\infty$, with $L_\phi^{[2]}$ defined in Theorem 3.*

Theoretical Guarantee of FDVF. The following theorem showcases the guarantee for FDVF under our unified analysis, with the given condition that indicates our selection of ε should generally have the same scaling as finite-sample error term.

Condition 3. *For some uniform constant C, for the given ε, n, δ, $L_\mathcal{E}\varepsilon \leq \frac{eCHC_\mathcal{V}^2 C_\mu}{2n} \cdot \log \frac{4|\mathcal{V}||\Theta|}{\delta}$.*

Theorem 9 (Theoretical Guarantee of FDVF)**.** *Suppose the abstract realizability condition $V_\mathcal{F}^\phi \in \mathcal{V}$ and the Bellman completeness condition $\forall V \in \mathcal{V}, \mathcal{B}^\mathcal{H} V \in \Theta$ ($\mathcal{B}^\mathcal{H}$ here refers to the operator on the abstract system) hold, and Assumptions 2, 6, 7, and 9 are satisfied. For any $\varepsilon > 0$ satisfying condition 1, 2, 3, define $T = \max\{T_0(\varepsilon), T_1(\varepsilon), T_2(\varepsilon)\}$. Then, for some uniform constant c, with probability at least $1 - \delta$, we have:*

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

Corollary 2 (Boosted finite sample guarantee)**.** *For n large enough with necessary realizability and completeness condition, we have a finite sample guarantee:*

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

A Simpler Pipeline: Abstracting Only the Policy. Revisiting the above analysis and noticeably one step is actually unnecessary, namely, the abstraction from the original POMDP to the short-term memory POMDP. That's because the memory dependency of the policy is the real root of the "curse of memory." Notably, the introduction of Assumption 2 and 6 are all for the sake of bounding the abstraction error of the POMDP itself, and therefore can be eliminated for FDVF. **This shows a significant advantage of FDVF comparing to history-as-state MDP that the "curse of memory" is much easier to handle than "the curse of horizon", since for the latter, abstracting the POMDP itself is inevitable.** When we only abstract the policy, the previous condition 1 and 2 can be relaxed to condition 2′ for $H > 1$.

Condition 2′. *The ε is small enough that $HL_\pi \varepsilon / \min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+) \leq 1$.*

Theorem 10 (Tighter Theoretical Guarantee of FDVF)**.** *Suppose the abstract realizability condition $V_\mathcal{F}^\phi \in \mathcal{V}$ and the Bellman completeness condition $\forall V \in \mathcal{V}, \mathcal{B}^\mathcal{H} V \in \Theta$ ($\mathcal{B}^\mathcal{H}$ here refers to the operator on the abstract system) hold, and Assumptions 7 and 9 are satisfied. For any $\varepsilon > 0$ satisfying condition 2′, 3, define $T = \max\{T_1(\varepsilon), T_2(\varepsilon)\}$. Then, for some uniform*

constant c, c_1, c_2, with probability at least $1-\delta$, we have:

$$|J(\pi_e) - \mathbb{E}_{\pi_b}[\hat{V}(f_1)]| \leq L_\phi \varepsilon + \sqrt{H} \cdot \max_{h\in[H]} \sup_{V\in\mathcal{V}} \sqrt{\frac{\mathbb{E}_{\pi_e^\phi}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h,\tau_{[h-T+1:h]})^2]}{\mathbb{E}_{\pi_b^\phi}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}} \cdot \sqrt{\frac{cHC_{\mathcal{V}}'^2 C_\mu}{n} \log \frac{|\mathcal{V}||\Theta|}{\delta} + L_{\mathcal{E}}\varepsilon}$$

where $C_{\mathcal{V}}' := \max\{\|\mathcal{V}\|_\infty + 1, \|\Theta\|_\infty\}$, $L_\phi = R_{\max} H L_\pi + R_{\max} H^2 L_\pi + \|\mathcal{V}\|_\infty$ *and* $L_\varepsilon = 3 \cdot \Big(\frac{HL_\pi(c_1(C_\mu+1)\|\mathcal{V}\|_\infty\|\Theta\|_\infty + c_2 H \max\{C_\mu\|\mathcal{V}\|_\infty\|\Theta\|_\infty, \frac{1}{2}\|\Theta\|_\infty^2\})}{\min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)} + HC_\mu\|\mathcal{V}\|_\infty\|\Theta\|_\infty \Big)$

Remark 6. *The coverage in our result only takes in a history of window T instead of the entire horizon H, and Theorem 4, 5 proves that the L_2 and L_∞ belief coverage in the belief one-hot scenario are no worse than the original. Example 2 also shows a polynomial finite sample guarantee while the original bound does not exist, effectively mitigating the curse of memory. Despite that structural assumption on POMDP model is adopted for Theorem 9, this can be avoided by taking a simpler pipeline (i.e. Theorem 10) which gives us a even better result, indicating the advantage in tractability of memory to horizon.*

6 CONCLUSION

We developed a belief-space covering framework for offline POMDP off-policy evaluation that exploits the metric structure of beliefs to relax standard coverage assumptions. Intuitively, nearby beliefs induce similar dynamics and value estimates and can therefore be aggregated without sacrificing accuracy. Under stability assumptions, history-based coverage requirements reduce to covering numbers of the belief space, yielding abstract coverage that is generally no worse than the original and that can mitigate exponential dependence on horizon and memory when the reachable beliefs are suitably smooth. Instantiating this framework for double-sampling Bellman error minimization and for future-dependent value functions shows that belief-space abstractions can restore polynomial finite-sample guarantees in settings where history-as-state analyses become vacuous. More broadly, our results suggest that in POMDPs, the effective difficulty of OPE is governed less by raw trajectory length than by the complexity and stability of the induced belief dynamics. This perspective suggests future directions such as stability-regularized training, belief-aware representation learning, and extensions to other offline RL tasks and robust POMDP settings.

Acknowledgments

Youheng thanks Nan Jiang for useful discussions.

References

Antos, A., Szepesvári, C., and Munos, R. (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129.

Asadi, K., Misra, D., and Littman, M. (2018). Lipschitz continuity in model-based reinforcement learning. In *International conference on machine learning*, pages 264–273. PMLR.

Atrash, A., Kaplow, R., Villemure, J., West, R., Yamani, H., and Pineau, J. (2009). Development and validation of a robust speech interface for improved human-robot interaction. *International Journal of Social Robotics*, 1:345–356.

Baird III, L. C. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the twelfth international conference on machine learning*, pages 30–37.

Bennett, A., Kallus, N., Oprescu, M., Sun, W., and Wang, K. (2024). Efficient and sharp off-policy evaluation in robust markov decision processes. *Advances in Neural Information Processing Systems*, 37:112962–113000.

Bovy, E. M., Suilen, M., Junges, S., and Jansen, N. (2024). Imprecise probabilities meet partial observability: game semantics for robust pomdps. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6697–6706.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pages 1042–1051. PMLR.

Dutta, S., Caprio, M., Lin, V., Cleaveland, M., Jang, K. J., Ruchkin, I., Sokolsky, O., and Lee, I. (2025). Distributionally robust statistical verification with imprecise neural networks. In *Proceedings of the 28th ACM International Conference on Hybrid Systems: Computation and Control*, pages 1–12.

Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6.

Feng, Y., Li, L., and Liu, Q. (2019). A kernel loss for solving the bellman equation. *Advances in Neural Information Processing Systems*, 32.

François-Lavet, V., Rabusseau, G., Pineau, J., Ernst, D., and Fonteneau, R. (2019). On overfitting and asymptotic bias in batch reinforcement learning with partial observability. *Journal of Artificial Intelligence Research*, 65:1–30.

Gelada, C., Kumar, S., Buckman, J., Nachum, O., and Bellemare, M. G. (2019). Deepmdp: Learning continuous latent space models for representation

learning. In *International conference on machine learning*, pages 2170–2179. PMLR.

Hao, M., Su, P., Hu, L., Szabó, Z., Zhao, Q., and Shi, C. (2024). Off-policy evaluation with deeply-abstracted states. *arXiv preprint arXiv:2406.19531*.

Hu, Y. and Wager, S. (2023). Off-policy evaluation in partially observed markov decision processes under sequential ignorability. *The Annals of Statistics*, 51(4):1561–1585.

Jiang, B., Song, R., Li, J., and Zeng, D. (2019). Entropy learning for dynamic treatment regimes. *Statistica Sinica*, 29(4):1633.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pages 652–661. PMLR.

Jiang, N. and Xie, T. (2025). Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 40(4):570–596.

Kallus, N., Mao, X., Wang, K., and Zhou, Z. (2022). Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, pages 10598–10632. PMLR.

Kurniawati, H., Hsu, D., and Lee, W. S. (2009). SARSOP: Efficient Point-Based POMDP Planning by Approximating Optimally Reachable Belief Spaces. *Robotics: Science and Systems IV*.

Lauri, M., Hsu, D., and Pajarinen, J. (2022). Partially observable markov decision processes in robotics: A survey. *IEEE Transactions on Robotics*, 39(1):21–40.

Le, H., Voloshin, C., and Yue, Y. (2019). Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR.

Lee, W., Rong, N., and Hsu, D. (2007). What makes some pomdp problems easy to approximate? *Advances in neural information processing systems*, 20.

Li, L., Walsh, T. J., and Littman, M. L. (2006). Towards a unified theory of state abstraction for mdps. *AI&M*, 1(2):3.

Munos, R. and Szepesvári, C. (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5).

Nachum, O., Chow, Y., Dai, B., and Li, L. (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32.

Osogami, T. (2015). Robust partially observable markov decision process. In *International Conference on Machine Learning*, pages 106–115. PMLR.

Pirotta, M., Restelli, M., and Bascetta, L. (2015). Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100(2):255–283.

Poupart, P., Kim, K.-E., and Kim, D. (2011). Closing the gap: Improved bounds on optimal pomdp solutions. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 21, pages 194–201.

Precup, D., Sutton, R. S., and Singh, S. (2000). Eligibility traces for off-policy policy evaluation. In *ICML*, volume 2000, pages 759–766. Citeseer.

Shani, G., Brafman, R. I., and Shimony, S. E. (2008). Prioritizing point-based pomdp solvers. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(6):1592–1605.

Shani, G., Pineau, J., and Kaplow, R. (2013). A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*, 27:1–51.

Smith, T. and Simmons, R. (2012). Point-based pomdp algorithms: Improved analysis and implementation. *arXiv preprint arXiv:1207.1412*.

Spaan, M. T. and Vlassis, N. (2005). Perseus: Randomized point-based value iteration for pomdps. *Journal of artificial intelligence research*, 24:195–220.

Subramanian, J., Sinha, A., Seraj, R., and Mahajan, A. (2022). Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83.

Uehara, M., Huang, J., and Jiang, N. (2020). Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR.

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W., and Xie, T. (2021). Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*.

Uehara, M., Kiyohara, H., Bennett, A., Chernozhukov, V., Jiang, N., Kallus, N., Shi, C., and Sun, W. (2023). Future-dependent value-based off-policy evaluation in pomdps. *Advances in neural information processing systems*, 36:15991–16008.

Zanette, A. and Wainwright, M. J. (2022). Bellman residual orthogonalization for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:3137–3151.

Zhang, Y. and Jiang, N. (2024). On the curses of future and history in future-dependent value functions for off-policy evaluation. *arXiv preprint arXiv:2402.14703*.

Zhang, Y. and Jiang, N. (2025). Statistical tractability of off-policy evaluation of history-dependent policies in pomdps. *arXiv preprint arXiv:2503.01134*.

Zhang, Z., Hsu, D., and Lee, W. S. (2014). Covering number for efficient heuristic-based pomdp planning. In *International conference on machine learning*, pages 28–36. PMLR.

Zhang, Z., Littman, M., and Chen, X. (2012). 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.

7 Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A THE UNIFIED ANALYSIS

Proof of Meta-theorem 3

Proof. First, we have from Theorem 2,

$$\begin{aligned}|\mathfrak{est}(Q^\pi) - \mathfrak{est}^\phi(Q^\pi_\phi)| &= |\mathbb{E}_{b\sim d_0}[V^{\pi_\phi}_{\text{bin}}(\phi(b)) - V^\pi_{\text{true}}(b)]| \\ &\le \|[V^{\pi_\phi}_{\text{bin}}]_{\text{true}} - V^\pi_{\text{true}}\|_\infty \\ &\le L^{[1]}_\phi \varepsilon,\end{aligned}$$

where we use $b \sim d_0$ as a shorthand for $s_1 \sim d_0, o_1 \sim \mathbb{O}(s_1), b = \mathbf{b}(o_1)$. Then using triangle's inequality, we get

$$\begin{aligned}|\mathfrak{est}(\hat{Q}^\pi) - \mathfrak{est}^\phi(Q^\pi)| &\le |\mathfrak{est}(\hat{Q}^\pi) - \mathfrak{est}^\phi(\hat{Q}^\pi)| + |\mathfrak{est}^\phi(\hat{Q}^\pi) - \mathfrak{est}^\phi(Q^\pi_\phi)| + |\mathfrak{est}(Q^\pi) - \mathfrak{est}^\phi(Q^\pi_\phi)| \\ &\le L^{[1]}_\phi + L^{[2]}_\phi + C^\phi_\pi \cdot \sqrt{\|\mathcal{V}\|_\infty \cdot \left(\frac{1}{n}\log\frac{|\mathcal{V}|}{\delta}\right)^\alpha + L_\mathcal{E}\varepsilon}, \;\; w.p. > 1-\delta\end{aligned}$$

And that completes the proof. □

Lemma 3. *If for any ε that satisfies $D(\varepsilon)$, the following holds*

$$|J(\pi) - \hat{J}(\pi)| \le L_\phi\varepsilon + C^\phi_\pi \cdot \sqrt{\|\mathcal{V}\|_\infty \cdot \left(\frac{1}{n}\log\frac{|\mathcal{V}|}{\delta}\right)^\alpha + L_\mathcal{E}\varepsilon} \quad w.p. \ge 1-\delta. \tag{1}$$

Then

$$|J(\pi) - \hat{J}(\pi)| \le \inf_{\substack{\varepsilon\ge 0 \\ D(\varepsilon)}} \left(L_\phi\varepsilon + C^\phi_\pi \cdot \sqrt{\|\mathcal{V}\|_\infty \cdot \left(\frac{1}{n}\log\frac{|\mathcal{V}|}{\delta}\right)^\alpha + L_\mathcal{E}\varepsilon} \right) \quad w.p. \ge 1-\delta. \tag{2}$$

Proof. Let $\beta(\varepsilon) := L_\phi\varepsilon + C^\phi_\pi \cdot \sqrt{\|\mathcal{V}\|_\infty \cdot \left(\frac{1}{n}\log\frac{|\mathcal{V}|}{\delta}\right)^\alpha + L_\mathcal{E}\varepsilon}$, and $\beta^* := \inf_{\substack{\varepsilon\ge 0 \\ D(\varepsilon)}} \beta(\varepsilon)$. Then there exists a sequence of $\{\varepsilon_i\}_{i=1}^\infty$ satisfying $\varepsilon_i \ge 0$ and $D(\varepsilon_i)$, such that $\beta(\varepsilon_i) \downarrow \beta^*$. Then the family of events $\{E_i := \{\omega \in \Omega : |J(\pi) - \hat{J}(\pi)|(\omega) \le \beta(\varepsilon_i)\}\}_{i=1}^\infty$ is decreasing, with the limit being $E_\infty := \{\omega \in \Omega : |J(\pi) - \hat{J}(\pi)|(\omega) \le \beta^*\}$. It then suffice to prove the result by applying the monotone convergence theorem of measure, which shows that $\Pr(E_\infty) = \lim_{i\to\infty}\Pr(E_i) \ge 1-\delta$. □

Proof of Proposition 1

Proof. This is because for any $b_1, b_2 \in \mathcal{B}, a \in \mathcal{A}, o \in \mathcal{O}$, if $b_1 \ne b_2$, $b_1^{o,a}$ and $b_2^{o,a}$ would either be identical, thus $\frac{\|b_1^{o,a} - b_2^{o,a}\|_1}{\|b_1 - b_2\|_1} = 0$, or be different, thus $\frac{\|b_1^{o,a} - b_2^{o,a}\|_1}{\|b_1 - b_2\|_1} = 1$. □

Explanation of Example

Example 3. *Consider a latent MDP with one action a, two states s_1, s_2 and four observations o_1, o_2, o_3, o_4. The initial state is evenly distributed over s_1 and s_2, the emission probability of s_1 is $(0.5, 0, 0.5-\xi, \xi)$ and of s_2 is $(0, 0.5, \xi, 0.5-\xi)$. Then for two belief states $b_1 := \mathbf{b}(o_2) = (0,1)$ and $b_2 := \mathbf{b}(o_4) = (2\xi, 1-2\xi)$, simultaneously taking action a and observing o_3 makes the successive belief becomes $b_1^{o_3,a} = (0,1)$ and $b_2^{o_3,a} = (0.5, 0.5)$. This violates the contraction property as $\|b_1^{o_3,a} - b_2^{o_3,a}\|_1 \ge \frac{1}{4\xi}\|b_1 - b_2\|_1$ for any $\xi \le \frac{1}{4}$, also showing that the Lipschitz parameter can be arbitrarily large as $\xi \to 0$.*

Proof of Lemma 1.

Proof. Fix an action $a \in \mathcal{A}$. For a belief $b \in \mathcal{B}$, define the joint distribution over states and observations:

$$P_b(s, o|a) := b(s)\, P(o|s, a), \qquad P_b(o|a) = \sum_s P_b(s, o|a).$$

The posterior distribution over the current state is

$$\tilde{b}^{o,a}(s) := P(s|b,o,a) = \frac{P_b(s,o|a)}{P_b(o|a)} = P_b(s|o,a).$$

The next belief after observing o is

$$b^{o,a}(s') = \sum_s P(s'|s,a)\,\tilde{b}^{o,a}(s).$$

For each $o \in \mathcal{O}$, the total variation distance contracts under the state transition kernel:

$$\|b_1^{o,a} - b_2^{o,a}\|_1 = \|\tilde{b}_1^{o,a} P(\cdot|\cdot,a) - \tilde{b}_2^{o,a} P(\cdot|\cdot,a)\|_1 \le \|\tilde{b}_1^{o,a} - \tilde{b}_2^{o,a}\|_1.$$

Taking expectation with respect to $P(\cdot|b_1,a)$ gives

$$\begin{aligned}
\mathbb{E}_{o\sim P(\cdot|b_1,a)}[\|b_1^{o,a} - b_2^{o,a}\|_1] &\le \mathbb{E}_{o\sim P(\cdot|b_1,a)}[\|\tilde{b}_1^{o,a} - \tilde{b}_2^{o,a}\|_1] \\
&= \sum_o P_{b_1}(o|a) \sum_s \big|P(s|b_1,o,a) - P(s|b_2,o,a)\big| \\
&= \sum_{s,o} \big|P_{b_1}(s,o|a) - P_{b_1}(o|a) P_{b_2}(s|o,a)\big|.
\end{aligned}$$

Insert and subtract $P_{b_2}(s,o|a) = P_{b_2}(o|a)P_{b_2}(s|o,a)$, then apply the triangle inequality:

$$\le \sum_{s,o} \big|P_{b_1}(s,o|a) - P_{b_2}(s,o|a)\big| + \sum_{s,o} \big|P_{b_1}(o|a)P_{b_2}(s|o,a) - P_{b_2}(s,o|a)\big|.$$

All three mappings

$$k_1^a((o',s')|s) = \mathbb{I}(s=s')P(o'|s,a), \quad k_2^a((o',s')|o) = \mathbb{I}(o=o')P_{b_2}(s'|o,a), \quad k_3^a(o|s) = P(o|s,a)$$

where $\mathbb{I}(x=x')$ is the indicator function on $\mathcal{X}$, are Markov kernels. By the data processing inequality for total variation,

$$\sum_{s,o} \big|P_{b_1}(s,o|a) - P_{b_2}(s,o|a)\big| = \Big\| \sum_s k_1^a((\cdot,\cdot)|s) b_1(s) - \sum_s k_1^a((\cdot,\cdot)|s) b_2(s) \Big\|_1 \le \|b_1 - b_2\|_1,$$

$$\begin{aligned}
\sum_{s,o} \big|P_{b_1}(o|a)P_{b_2}(s|o,a) - P_{b_2}(s,o|a)\big| &= \Big\| \sum_o k_2^a((\cdot,\cdot)|o) P_{b_1}(o|a) - \sum_o k_2^a((\cdot,\cdot)|o) P_{b_2}(o|a) \Big\|_1 \\
&\le \|P_{b_1}(\cdot|a) - P_{b_2}(\cdot|a)\|_1 \\
&= \Big\| \sum_s b_1(s) P(\cdot|s,a) - \sum_s b_2(s) P(\cdot|s,a) \Big\|_1 \\
&= \Big\| \sum_s k_3^a(\cdot|s) b_1(s) - \sum_s k_3^a(\cdot|s) b_2(s) \Big\|_1 \\
&\le \|b_1 - b_2\|_1.
\end{aligned}$$

Combining the above inequalities yields

$$\mathbb{E}_{o\sim P(\cdot|b_1,a)}[\|b_1^{o,a} - b_2^{o,a}\|_1] \le 2\,\|b_1 - b_2\|_1.$$

By symmetry, the same bound holds when the expectation is taken with respect to $P(\cdot|b_2,a)$ instead of $P(\cdot|b_1,a)$. This completes the proof. □

B ABSTRACTION UNDER COVERING

The proof for bounding the belief abstraction error, i.e. Theorem 11 follows a similar idea from Theorem 9 and Proposition 48 in (Subramanian et al., 2022).

Definition 3. *A ε-cover $\mathcal{C}_\varepsilon$ is a subspace of the belief state space which satisfies:*

$$\mathcal{B} \subset \bigcup_{c \in \mathcal{C}_\varepsilon} \mathbf{B}(c, \varepsilon) \tag{3}$$

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.

Belief-MDP one-step update. Recall that in the main text we define the belief at time h as $b = \mathbf{b}(\tau_h^+) = \Pr(s_h \mid \tau_h^+)$, i.e. the posterior of the current latent state after observing the current observation o_h. Accordingly, throughout the appendix, whenever we write the next-belief update from b under action a, the observation involved is the *next* observation o_{h+1}.

For $b \in \Delta(\mathcal{S})$ and $a \in \mathcal{A}$, define the one-step predictive next-state distribution

$$\bar{b}^{\,a}(s') := \sum_{s \in \mathcal{S}} \mathbb{T}(s'|s, a) b(s). \tag{4}$$

Then the conditional distribution of the next observation is

$$P(o_+|b, a) := \sum_{s' \in \mathcal{S}} \mathbb{O}(o_+|s') \bar{b}^{\,a}(s') = \sum_{s' \in \mathcal{S}} \mathbb{O}(o_+|s') \sum_{s \in \mathcal{S}} \mathbb{T}(s'|s, a) b(s). \tag{5}$$

The updated next belief after taking action a and receiving next observation o_+ is

$$b^{a,o_+}(s') := \Pr(s_{h+1} = s' \mid b_h = b, a_h = a, o_{h+1} = o_+) = \frac{\mathbb{O}(o_+|s') \bar{b}^{\,a}(s')}{P(o_+|b, a)}. \tag{6}$$

Finally, the one-step transition kernel of the belief MDP is defined by

$$P(b'|b, a) := \sum_{o_+ \in \mathcal{O}:\, b^{a,o_+} = b'} P(o_+|b, a). \tag{7}$$

Equivalently,

$$P(b'|b, a) = \sum_{o_+ \in \mathcal{O}:\, b^{a,o_+} = b'} \sum_{s' \in \mathcal{S}} \mathbb{O}(o_+|s') \sum_{s \in \mathcal{S}} \mathbb{T}(s'|s, a) b(s). \tag{8}$$

When there is no ambiguity, we may still write $b^{o,a}$ in place of b^{a,o_+}.

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

$$|r(b_1, a) - r(b_2, a)| \le R_{\max} \|b_1 - b_2\|_1. \tag{9}$$

Proof. This is easily obtained from:

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

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

Lemma 5. *(Lemma 2 in (Zhang et al., 2012)) For any two belief points b_1, b_2 and all action a, we have $\sum_o |P(o|b_1,a) - P(o|b_2,a)| \leq \|b_1 - b_2\|_1$.*

Consequently, we put forward the following proposition.

Proposition 2. *For policy π satisfying Assumption 1, we have*

$$\sum_{o,a} |P(o|b_1,a)\pi(a|b_1) - P(o|b_1,a)\pi(a|b_2)| \leq L_\pi \|b_1 - b_2\|_1. \tag{10}$$

Proof. This is a direct application of the data processing inequality. Notice that $k^{b_1}((o',a')|a) = P(o'|b_1,a)\mathbb{I}(a=a')$ is a Markov kernel, then

$$\begin{aligned}\sum_{o,a} |P(o|b_1,a)\pi(a|b_1) - P(o|b_1,a)\pi(a|b_2)| &= \sum_{o',a'} \left| \sum_a k^{b_1}((o',a')|a)\pi(a|b_1) - \sum_a k^{b_1}((o',a')|a)\pi(a|b_2) \right| \\ &\leq \|\pi(\cdot|b_1) - \pi(\cdot|b_2)\|_1 \\ &\leq L_\pi \|b_1 - b_2\|_1,\end{aligned}$$

and the proof is done. □

The one-step error is easy to control, however, without model irrelevant state abstraction assumptions such as bisimulation, it is extremely difficult to control the accumulative error induced by infinite amount of steps. Fortunately, stability property of value function provides us with an alternative approach.

In the abstract MDP, the tuple $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$ is mapped by the abstraction ϕ to $\langle \mathcal{S}_\phi, \mathcal{A}, P_\phi, R_\phi, \gamma \rangle$, 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:

$$P_\phi(\phi(s')|\phi(s),a) = \mathbb{E}_{s \sim p_{\phi(s)}}[P_\phi(\phi(s')|s,a)]. \tag{11}$$

Notation clarification. Equation (11) applies to an abstract MDP built from any underlying MDP state space. To avoid confusion, we reserve $\mathbb{T}(s' \mid s,a)$ for the transition kernel of the latent (hidden-state) MDP. When the underlying MDP state space is the belief space $\mathcal{B}$, we instead write $P(b' \mid b,a)$ for the transition kernel of the belief MDP, and define the corresponding abstract transition kernel analogously as

$$P_\phi(\phi(b') \mid \phi(b),a) := \mathbb{E}_{\tilde{b} \sim p_{\phi(b)}} \left[\sum_{\phi(\bar{b}) = \phi(b')} P(\bar{b} \mid \tilde{b},a) \right].$$

Similarly, the abstract reward is defined by

$$R_\phi(\phi(b),a) := \mathbb{E}_{\tilde{b} \sim p_{\phi(b)}}[R(\tilde{b},a)].$$

Therefore, whenever the underlying states are beliefs, the transition of the abstract belief MDP is denoted by P_ϕ, not $\mathbb{T}$.

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 $V_{\text{true}}^{[\pi_\phi]_{\text{true}}}$ and the lifted value function $[V_{\text{bin}}^{\pi_\phi}]_{\text{true}}$ from 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 11. *If Assumption 2 holds, then the error between $V_{\text{true}}^{[\pi_\phi]_{\text{true}}}$ and the lifted abstract MDP's true value function $[V_{\text{bin}}^{\pi_\phi}]_{\text{true}}$ can be bounded as follows:*

$$\|V_{\text{true}}^{[\pi_\phi]_{\text{true}}} - [V_{\text{bin}}^{\pi_\phi}]_{\text{true}}\|_\infty \leq \frac{(R_{\max} + 2L_V)\varepsilon}{1-\gamma} + \frac{R_{\max}}{(1-\gamma)^2}\varepsilon \tag{12}$$

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 $b' \sim \text{bin}(\phi(b))$ 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, where the action a is determined by the policy π_ϕ. Once the update is complete, the resulting belief state is the sampled b'.

Also $b_{k+1} \sim b_k$ symbolizes the classical belief update.

With these notations 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

$$[V_{\text{bin}}^{\pi_\phi}]_{\text{true}}(b) = \mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots}} [r_\phi(\phi(b_1),a_1) + \gamma r_\phi(\phi(b_2),a_2) + \gamma^2 r_\phi(\phi(b_3),a_3) + \cdots] \tag{13}$$

Consider $V^{[k]}$ as

$$V^{[k]} = \mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots\\ b_k\sim\text{bin}(\phi(b_{k-1}))\\ b_{k+1}\sim b_k\\ \cdots}} [r_\phi(\phi(b_1),a_1) + \gamma r_\phi(\phi(b_2),a_2) + \gamma^2 r_\phi(\phi(b_3),a_3) + \cdots] \tag{14}$$

Then $V^{[0]}(b) = V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b)$. Next, for $\forall b$,

$$|V^{[k+1]}(b) - V^{[k]}(b)| \tag{15}$$

$$= \left| \mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots\\ b_k\sim\text{bin}(\phi(b_{k-1}))\\ b_{k+1}\sim b_k\\ b_{k+2}\sim b_{k+1}\\ \cdots}} [\gamma^k V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_{k+1})] - \mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots\\ b_k\sim\text{bin}(\phi(b_{k-1}))\\ b_{k+1}\sim\text{bin}(\phi(b_k))\\ b_{k+2}\sim b_{k+1}\\ \cdots}} [\gamma^k r_\phi(\phi(b_{k+1}),a) + \gamma^{k+1} V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_{k+2})] \right| \tag{16}$$

$$= \left| \mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots\\ b_k\sim\text{bin}(\phi(b_{k-1}))\\ b_{k+1}\sim b_k\\ b_{k+2}\sim b_{k+1}\\ \cdots}} [\gamma^k V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_{k+1})] - \mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots\\ b_k\sim\text{bin}(\phi(b_{k-1}))\\ b_{k+1}\sim\text{bin}(\phi(b_k))\\ b_{k+2}\sim b_{k+1}\\ \cdots}} [\gamma^k r_\phi(\phi(b_{k+1}),a) - \gamma^k r(b_{k+1},a) + \gamma^k V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_{k+1})] \right| \tag{17}$$

$$\leq \gamma^k R_{\max}\varepsilon + \gamma^k 2L_V\varepsilon + \gamma^k \frac{R_{\max}}{1-\gamma}\varepsilon \tag{18}$$

where the last inequality used the stability of value function (Assumption 2), Lemma 1 and Lemma 4, 5 since the next belief is sampled from the start of same bin and thus close enough. Specifically, it uses the fact that

$$\begin{aligned}
&\left|\mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_k,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_k^{o,a})] - \mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ b_t\sim p_{\phi(b_k)}\\ o\sim P(\cdot|b_t,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_t^{o,a})]\right| \\
&= \left|\mathbb{E}_{b_t\sim p_{\phi(b_k)}}\left[\mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_k,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_k^{o,a})] - \mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_t,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_t^{o,a})]\right]\right| \\
&\leq \left|\mathbb{E}_{b_t\sim p_{\phi(b_k)}}\left[\mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_k,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_k^{o,a})] - \mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_k,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_t^{o,a})]\right]\right| \\
&\qquad + \left|\mathbb{E}_{b_t\sim p_{\phi(b_k)}}\left[\mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_k,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_t^{o,a})] - \mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_t,a)}}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_t^{o,a})]\right]\right| \\
&\leq \left|\mathbb{E}_{b_t\sim p_{\phi(b_k)}}\left[\mathbb{E}_{\substack{a\sim\pi(\phi(b_k))\\ o\sim P(\cdot|b_k,a)}}[L_V\|b_k^{o,a} - b_t^{o,a}\|_1]\right]\right|
\end{aligned}$$

$$+\left|\mathbb{E}_{\substack{b_t\sim p_{\phi(b_k)}\\ a\sim\pi(\phi(b_k))}}\left[\langle P(\cdot|b_k,a)-P(\cdot|b_t,a), V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_t^{\cdot,a})\rangle\right]\right|$$
$$\leq 2L_V\varepsilon+\frac{R_{\max}}{1-\gamma}\varepsilon$$

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

$$|V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b)-[V_{\text{bin}}^{\pi_\phi}]_{\text{true}}(b)| \tag{19}$$
$$=\left|\sum_{k=0}^{\infty}\left(V^{[k+1]}(b)-V^{[k]}(b)\right)\right| \tag{20}$$
$$\leq \sum_{k=0}^{\infty}\left|\gamma^k R_{\max}\varepsilon+\gamma^k 2L_V\varepsilon+\gamma^k\frac{R_{\max}}{1-\gamma}\varepsilon\right| \tag{21}$$
$$\leq \frac{(R_{\max}+2L_V)\varepsilon}{1-\gamma}+\frac{R_{\max}}{(1-\gamma)^2}\varepsilon \tag{22}$$

□

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, which does not rely on any assumption on the POMDP model itself.

Theorem 12. *If Assumption 1 holds.*

$$\|V_{\text{true}}^{\pi}-V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty\leq\frac{R_{\max}L_\pi\varepsilon}{1-\gamma}+\frac{\gamma R_{\max}}{(1-\gamma)^2}L_\pi\varepsilon \tag{23}$$

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

$$\begin{aligned}\|V_{\text{true}}^{\pi}-V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty&=\|V_{\text{true}}^{\pi}-\mathcal{T}^{[\pi_\phi]_{\text{true}}}V_{\text{true}}^{\pi}+\mathcal{T}^{[\pi_\phi]_{\text{true}}}V_{\text{true}}^{\pi}-\mathcal{T}^{[\pi_\phi]_{\text{true}}}V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty\\&\leq\|V_{\text{true}}^{\pi}-\mathcal{T}^{[\pi_\phi]_{\text{true}}}V_{\text{true}}^{\pi}\|_\infty+\gamma\|V_{\text{true}}^{\pi}-V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty.\end{aligned} \tag{24}$$

Here for an MDP $\langle\mathcal{S},\mathcal{A},r,\gamma,P\rangle$, $(\mathcal{T}^\pi V)(s):=\mathbb{E}_{\substack{a'\sim\pi(\cdot|s)\\ s'\sim P(\cdot|s,a')}}[r(s,a')+\gamma V(s')]$ is the Bellman operator, which is a γ-Lipschitz compression operator w.r.t. the infinity norm. Consequently,

$$\|V_{\text{true}}^{\pi}-V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty\leq\frac{1}{1-\gamma}\|V_{\text{true}}^{\pi}-\mathcal{T}^{[\pi_\phi]_{\text{true}}}V_{\text{true}}^{\pi}\|_\infty. \tag{25}$$

For any b, we have

$$\begin{aligned}&|(V_{\text{true}}^{\pi}-\mathcal{T}^{[\pi_\phi]_{\text{true}}}V_{\text{true}}^{\pi})(b)|\\&=|(\mathcal{T}^{\pi}V_{\text{true}}^{\pi}-\mathcal{T}^{[\pi_\phi]_{\text{true}}}V_{\text{true}}^{\pi})(b)|\\&=\left|\mathbb{E}_{\substack{a\sim\pi(b)\\ b^{+1}\sim P(\cdot|b)}}\left[r+\gamma V_{\text{true}}^{\pi}(b^{+1})\right]-\mathbb{E}_{\substack{a\sim\pi_\phi(\phi(b))\\ b^{+1}\sim P(\cdot|b)}}\left[r+\gamma V_{\text{true}}^{\pi}(b^{+1})\right]\right|\end{aligned} \tag{26}$$

We first look at r,

$$\begin{aligned}|\mathbb{E}_{a\sim\pi(b)}[r]-\mathbb{E}_{a\sim\pi_\phi(\phi(b))}[r]|&=|\mathbb{E}_{a\sim\pi(b)}[r]-\mathbb{E}_{a\sim\pi(\phi(b))}[r]|\\&\leq R_{\max}L_\pi\varepsilon\end{aligned} \tag{27}$$

Then we look at V_{true}^{π},

$$\begin{aligned}&\left|\mathbb{E}_{\substack{a\sim\pi(b)\\ b^{+1}\sim P(\cdot|b)}}\left[\gamma V_{\text{true}}^{\pi}(b^{+1})\right]-\mathbb{E}_{\substack{a\sim\pi_\phi(\phi(b))\\ b^{+1}\sim P(\cdot|b)}}\left[\gamma V_{\text{true}}^{\pi}(b^{+1})\right]\right|\\&=\gamma\left|\sum_{o\in\mathcal{O}}\sum_{a\in\mathcal{A}}\left[\left(P(o|b,a)\pi(a|b)-P(o|b,a)\pi(a|\phi(b))\right)\cdot V^{\pi}(b^{o,a})\right]\right|\\&\leq\frac{\gamma R_{\max}}{1-\gamma}L_\pi\varepsilon\end{aligned} \tag{28}$$

where we used Proposition 2 for the final inequality. □

Proof of Theorem 2.

Proof. Combining Theorem 11, 12, we get

$$\begin{aligned}\|[V_{\text{bin}}^{\pi_\phi}]_{\text{true}} - V_{\text{true}}^{\pi}\|_\infty &\leq \|[V_{\text{bin}}^{\pi_\phi}]_{\text{true}} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty + \|V_{\text{true}}^{[\pi_\phi]_{\text{true}}} - V_{\text{true}}^{\pi}\|_\infty \\ &\leq \frac{(L_\pi+1)R_{\max}+2L_V}{1-\gamma}\varepsilon + \frac{\gamma R_{\max}L_\pi + R_{\max}}{(1-\gamma)^2}\varepsilon, \end{aligned} \tag{29}$$

thereby completing the proof. □

C DOUBLE SAMPLING ANALYSIS IN SECTION 5.1

Definition 4 (abstract algorithm)**.** *Consider the Bellman error minimization algorithm using double sampling, not only is it executed in the real world (simulator), but also virtually in the abstract system, using the same piece of offline data. The optimization target for the abstract algorithm can be written as*

$$\hat{Q}_\phi^{\pi_\phi} = \arg\min_{f\in\mathcal{F}} \mathcal{E}_\phi(f,\pi) \tag{30}$$

where

$$\mathcal{E}_\phi(f,\pi) = \mathbb{E}_{\mathcal{D}}[(f(\phi(b),a) - (r_\phi + \gamma f(\phi(b'_A),\pi_\phi)))(f(\phi(b),a) - (r_\phi + \gamma f(\phi(b'_B),\pi_\phi)))]. \tag{31}$$

Lemma 6 (MDP telescoping (Jiang and Xie, 2025))**.** *For an MDP $\langle \mathcal{S},\mathcal{A},r,\gamma,P\rangle$ and any function $Q:\mathcal{S}\times\mathcal{A}\to\mathbb{R}^+$, we have*

$$J_Q(\pi) - J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{d^\pi}[Q - \mathcal{T}^\pi Q] \tag{32}$$

where $(\mathcal{T}^\pi Q)(s,a) := r(s,a) + \gamma \mathbb{E}_{\substack{s'\sim P(\cdot|s,a)\\ a'\sim\pi(\cdot|s')}}[Q(s',a')]$ is the Bellman operator.

Lemma 7. *In the binned system, we have the following telescoping error*

$$|J_{\hat{Q}}(\pi_\phi) - J(\pi_\phi)| \leq \frac{\sqrt{C_\pi(\phi)}}{1-\gamma}\cdot\sqrt{\mathbb{E}_{d^D}[(\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q})^2]} \tag{33}$$

Proof. Recall the previously mentioned Lemma 6. Substituting it into the case of the abstract belief MDP gives:

$$J_{\hat{Q}}(\pi_\phi) - J(\pi_\phi) = \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_\phi}}[\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q}] \tag{34}$$

Therefore, we have:

$$\begin{aligned}|J_{\hat{Q}}(\pi_\phi) - J(\pi_\phi)| &= \left|\frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_\phi}}[\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q}]\right| \\ &\leq \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_\phi}}[|\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q}|] &(35)\\ &\leq \frac{1}{1-\gamma}\sqrt{\mathbb{E}_{d^{\pi_\phi}}[(\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q})^2]} &(36)\\ &\leq \frac{1}{1-\gamma}\sqrt{\mathbb{E}_{d^D}[\frac{d^\pi}{d^D}(\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q})^2]} &(37)\\ &\leq \frac{\sqrt{C_\pi(\phi)}}{1-\gamma}\cdot\sqrt{\mathbb{E}_{d^D}[(\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q})^2]} &(38)\end{aligned}$$

□

And we obviously have

$$\mathbb{E}_{d^{\mathcal{D}}}[\mathcal{E}_\phi(\hat{Q},\pi)] = \mathbb{E}_{d^{\mathcal{D}}}[(\hat{Q} - \mathcal{T}^{\pi_\phi}\hat{Q})^2] \tag{39}$$

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

$$|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^{\mathcal{D}}}[\mathcal{E}_\phi(f,\pi)]| \leq \sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}} \tag{40}$$

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 $R\max/(1-\gamma)$, 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):

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

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

$$\begin{aligned} &\Pr(|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^{\mathcal{D}}}[\mathcal{E}_\phi(f,\pi)]| \geq t) \leq 2\exp\left(-\frac{2t^2n(1-\gamma)^4}{16R_{\max}^4}\right) \\ \rightarrow\ &\Pr(|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^{\mathcal{D}}}[\mathcal{E}_\phi(f,\pi)]| > t) \leq 2\exp\left(-\frac{2t^2n(1-\gamma)^4}{16R_{\max}^4}\right) \end{aligned} \tag{41}$$

However, the goal of the proof is actually:

$$\Pr(\forall f \in \mathcal{F}, |\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^{\mathcal{D}}}[\mathcal{E}_\phi(f,\pi)]| \leq t) \tag{42}$$

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:

$$\begin{aligned} &\Pr(\forall f \in \mathcal{F}, |\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^{\mathcal{D}}}[\mathcal{E}_\phi(f,\pi)]| \leq t) \\ =\ &\Pr(A) = 1 - \Pr(\Omega\backslash A) = 1 - \Pr(\bigcup_{f\in\mathcal{F}} B_f) \\ \geq\ &1 - \sum_{f\in\mathcal{F}} \Pr(B_f) \geq 1 - 2|\mathcal{F}|\exp\left(-\frac{t^2n(1-\gamma)^4}{8R_{\max}^4}\right) \end{aligned} \tag{43}$$

The second-to-last step uses the subadditivity of probability (countable subadditivity), and the final step applies inequality (41).

Let $\delta = 2|\mathcal{F}|\exp\big(-t^2n(1-\gamma)^4/8R_{\max}^4\big)$, 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 (42) completes the proof. □

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

Remark 7. *For a finite function space where $|\mathcal{F}| < \infty$, the Bellman completeness Assumption 8 implies the realizability Assumption 4.*

Proposition 3. *Under the Bellman completeness Assumption 8, 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:*

$$|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^D}[\mathcal{E}_\phi(f,\pi)]| \lesssim \frac{R_{\max}^2}{n(1-\gamma)^2} \cdot \log \frac{|\mathcal{F}|}{\delta} \tag{44}$$

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 8 derived from Hoeffding's inequality, and cannot use the tighter concentration rate provided by Proposition 3.

Lemma 9. *If Assumption 5 holds, then*

$$|\mathcal{E}(f,\pi) - \mathcal{E}_\phi(f,\pi)| \leq \frac{4R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon \tag{45}$$

Proof.

$$\begin{aligned}
&|\mathcal{E}(f,\pi) - \mathcal{E}_\phi(f,\pi)| \\
=\ &|\mathbb{E}_{\mathcal{D}}[(f(b,a) - (r+\gamma f(b'_A,\pi)))(f(b,a) - (r+\gamma f(b'_B,\pi)))] - \\
&\quad \mathbb{E}_{\mathcal{D}}[(f(\phi(b),a) - (r_\phi + \gamma f(\phi(b'_A),\pi_\phi)))(f(\phi(b),a) - (r_\phi + \gamma f(\phi(b'_B),\pi_\phi)))]| \\
\leq\ &|\mathbb{E}_{\mathcal{D}}[\{(f(b,a) - f(\phi(b),a)) - (r(b,a) - r_\phi(\phi(b),a)) - \gamma(f(b'_A,\pi) - f(\phi(b'_A),\pi_\phi)\} \\
&\qquad \cdot (f(b,a) - (r+\gamma f(b'_B,\pi)))]| + \\
&|\mathbb{E}_{\mathcal{D}}[\{(f(b,a) - f(\phi(b),a)) - (r(b,a) - r_\phi(\phi(b),a)) - \gamma(f(b'_B,\pi) - f(\phi(b'_B),\pi_\phi)\} \\
&\qquad \cdot (f(b,a) - (r+\gamma f(b'_A,\pi)))]|.
\end{aligned} \tag{46}$$

Using the fact that

$$\begin{aligned}
&|f(b,\pi) - f(\phi(b),\pi_\phi)| \\
=\ &|\mathbb{E}_{\pi(a|b)}[f(b,a)] - \mathbb{E}_{\pi(a|\phi(b))}[f(\phi(b),a)]| \\
\leq\ &|\mathbb{E}_{\pi(a|b)}[f(b,a)] - \mathbb{E}_{\pi(a|\phi(b))}[f(b,a)]| + |\mathbb{E}_{\pi(a|\phi(b))}[f(b,a)] - \mathbb{E}_{\pi(a|\phi(b))}[f(\phi(b),a)]| \\
\leq\ &\frac{R_{\max}}{1-\gamma}\varepsilon + L_Q\varepsilon
\end{aligned} \tag{47}$$

we have

$$\begin{aligned}
&|\mathcal{E}(f,\pi) - \mathcal{E}_\phi(f,\pi)| \\
\leq\ &2 \cdot \frac{2R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon
\end{aligned} \tag{48}$$

□

Proposition 4. *If Assumption 4 , 5 holds, then*

$$|\mathbb{E}_{d^D}[\mathcal{E}_\phi(\hat{Q}^\pi,\pi)]| \leq \sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4} \cdot \log \frac{2|\mathcal{F}|}{\delta}} + \frac{8R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon \tag{49}$$

Proof. Using Lemma 8, we have

$$|\mathcal{E}_\phi(\hat{Q},\pi) - \mathbb{E}_{d^D}[\mathcal{E}_\phi(\hat{Q},\pi)]| \leq \sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4} \cdot \log \frac{2|\mathcal{F}|}{\delta}} \tag{50}$$

and

$$|\mathcal{E}_\phi(Q_\phi^{\pi_\phi}, \pi) - \mathbb{E}_{d^D}[\mathcal{E}_\phi(Q_\phi^{\pi_\phi}, \pi)]| \leq \sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4} \cdot \log \frac{2|\mathcal{F}|}{\delta}} \tag{51}$$

where $\mathbb{E}_{d^D}[\mathcal{E}_\phi(Q_\phi^{\pi_\phi}, \pi)] = 0$. Then using Lemma 9, we have with probability greater than $1-\delta$

$$|\mathcal{E}(\hat{Q}, \pi) - \mathbb{E}_{d^D}[\mathcal{E}_\phi(\hat{Q}, \pi)]| \leq \sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4} \cdot \log \frac{2|\mathcal{F}|}{\delta}} + \frac{4R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon \tag{52}$$

and

$$|\mathcal{E}(Q_\phi^{\pi_\phi}, \pi) - \mathbb{E}_{d^D}[\mathcal{E}_\phi(Q_\phi^{\pi_\phi}, \pi)]| \leq \sqrt{\frac{8R_{\max}^4}{n(1-\gamma)^4} \cdot \log \frac{2|\mathcal{F}|}{\delta}} + \frac{4R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon \tag{53}$$

Using the abstract realizability Assumption 4, we have $\mathcal{E}(\hat{Q}, \pi) = \min_{f \in \mathcal{F}} \mathcal{E}(f, \pi) \leq \mathcal{E}(Q_\phi^{\pi_\phi}, \pi)$, and consequently

$$|\mathbb{E}_{d^D}[\mathcal{E}_\phi(\hat{Q}^\pi, \pi)]| \leq \sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4} \cdot \log \frac{2|\mathcal{F}|}{\delta}} + \frac{8R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon. \tag{54}$$

□

And consequently,

Theorem 13. *If Assumption 4, 5 hold, then*

$$|J_{\hat{Q}^\pi}(\pi_\phi) - J(\pi_\phi)| \leq \frac{\sqrt{C_\pi(\phi)}}{1-\gamma} \cdot \sqrt{\sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4} \cdot \log \frac{2|\mathcal{F}|}{\delta}} + L_\mathcal{E}\varepsilon} \tag{55}$$

where

$$L_\mathcal{E} := \frac{8R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right) \tag{56}$$

Proof. The result follows directly from Proposition 4, Lemma 7, and (39). □

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

Theorem 14. *If Assumption 5 holds, then*

$$|J_{\hat{Q}^\pi}(\pi) - J_{\hat{Q}^\pi}(\pi_\phi)| \leq \frac{R_{\max}}{1-\gamma}\varepsilon + L_Q\varepsilon \tag{57}$$

Proof. We have

$$\begin{aligned}
&|J_{\hat{Q}^\pi}(\pi) - J_{\hat{Q}^\pi}(\pi_\phi)| \\
&= |\mathbb{E}_{b\sim d_0}[\hat{Q}^\pi(b, \pi)] - \mathbb{E}_{b\sim d_0}[\hat{Q}^\pi(\phi(b), \pi_\phi)]| \\
&= |\mathbb{E}_{b\sim d_0}[\hat{Q}^\pi(b, \pi) - \hat{Q}^\pi(\phi(b), \pi_\phi)]| \\
&\leq \frac{R_{\max}}{1-\gamma}\varepsilon + L_Q\varepsilon
\end{aligned} \tag{58}$$

□

Proof of Theorem 6.

Proof. Using Theorem 14 and the definition of $L_\phi^{[2]}$ in Theorem 3, we have $L_\phi^{[2]} = \frac{R_{\max}}{1-\gamma} + L_Q$ □

Proof of Theorem 7.

Proof. Combining Theorem 13, Theorem 6 and the Meta-theorem 3, we have $|J_{\hat{Q}^\pi}(\pi) - J(\pi)| \leq \frac{\sqrt{C_\pi(\varepsilon)}}{1-\gamma} \cdot \sqrt{\sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}} + L_{\mathcal{E}}\varepsilon} + L_\phi\varepsilon$, where $L_{\mathcal{E}} = \frac{8R_{\max}}{1-\gamma} \cdot \big((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\big)$. Note that when applying the Meta-theorem 3, we specify $\mathfrak{est}^\phi(\hat{Q}^\pi) = J_{\hat{Q}^\pi}(\pi_\phi)$ and $\mathfrak{est}^\phi(Q_\phi^\pi) = J(\pi_\phi)$.

Then, applying the result of Lemma 3 proves the result. □

Proof of Corollary 1.

Proof. Notice that the ε inside the square root always dominates the ε outside with ε small enough, therefore, we prove the corollary by substituting $L_{\mathcal{E}}\varepsilon$ with $\sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}}$, and then find the condition that the out side $L_\phi\varepsilon$ can be dominated by the term inside the square root. Such condition can be presented as $\big(\frac{32R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}\big)^{\frac{1}{4}} \leq \frac{L_{\mathcal{E}}}{L_\phi} \cdot \frac{\sqrt{2C_\pi^n}}{1-\gamma}$. Noticing that the coverage term is generally increasing, and is always bounded below by 1, we therefore provide a sufficient condition as $\big(\frac{32R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}\big)^{\frac{1}{4}} \leq \frac{L_{\mathcal{E}}}{L_\phi} \cdot \frac{\sqrt{2}}{1-\gamma}$. Solving it gives us the condition $n \geq 8R_{\max}^4(L_\phi/L_{\mathcal{E}})^4 \log(2|\mathcal{F}|/\delta)$, under which $|J_{\hat{Q}^\pi}(\pi) - J(\pi)| \leq \frac{2\sqrt{C_\pi^n}}{1-\gamma} \cdot \big(\frac{128R_{\max}^4}{n(1-\gamma)^4} \cdot \log\frac{2|\mathcal{F}|}{\delta}\big)^{\frac{1}{4}}$. □

D FUTURE-DEPENDENT VALUE FUNCTIONS

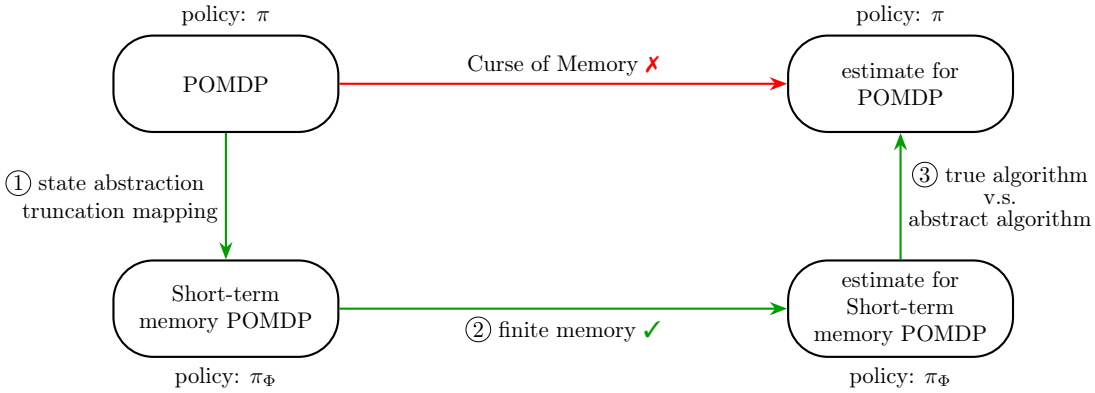

Figure 2: FDVF analysis pipeline

Definition 5. *MDPs $M_1 = \{\mathcal{S}_1, \mathcal{A}, P_1, R_1, H\}$ 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 $\varphi(M_1) := \{\varphi(\mathcal{S}_1), \mathcal{A}, P_1(\varphi(\cdot),\cdot), R_1(\varphi(\cdot),\cdot), H\} = M_2$*

Theorem 15. *For any POMDP $\mathcal{P}$ and $T \in \mathbb{N}^+$, there exists a short-term memory POMDP $\mathcal{P}_T$ with memory window T such that the belief MDP of $\mathcal{P}$, after abstraction by ϕ_T, is isomorphic to the belief MDP of $\mathcal{P}_T$.*

Proof. To prove the existence, it suffices to construct a short-term memory POMDP $\mathcal{P}_T$. Consider the belief MDP $\mathcal{M}$ of the original POMDP $\mathcal{P}$, and let $\mathcal{M}_T$ be the abstraction of $\mathcal{M}$ through ϕ_T. Let $\mathbf{b}$ denote the belief mapping in the original POMDP. Now, construct $\mathcal{P}_T = (\mathcal{S}', \mathcal{A}, \mathcal{O}, r', H, \mathbb{O}', \mathbb{T}')$ as follows:

Define $\mathcal{S}' := \bigcup_{i=0}^{T-1}(\mathcal{O}\times\mathcal{A})^i \times \mathcal{O}$, and the observation function as $\mathbb{O}'(o|s' = (o_1, a_1, \cdots, o_T)) := \mathbb{I}\{o = o_T\}$, which is a one-hot vector. The reward function is defined as $r'(s' = (o_1, a_1, \cdots, o_T), a) := R_\phi(\phi_T(\mathbf{b}(s')), a)$, and the transition probability as $\mathbb{T}(s_1'|s_0', a) := P_\phi(\phi_T(\mathbf{b}(s_1'))|\phi_T(\mathbf{b}(s_0')), a)$. Here P_ϕ and R_ϕ denote the transition and reward of the abstract belief MDP, defined analogously to (11) when the underlying state space is the belief space.

Next, we verify that the belief MDP $\mathcal{M}_T$ of POMDP $\mathcal{P}_T$ is indeed isomorphic to the abstraction of the belief MDP $\mathcal{M}$ of $\mathcal{P}$. Notice that for every historical sequence in $\mathcal{P}_T$, its state can be uniquely determined simply by taking the last T elements of the sequence. That is, the belief states in $\mathcal{P}_T$ are one-hot encoded. Thus, the

belief MDP of $\mathcal{P}_T$ is isomorphic to the hidden underlying MDP of $\mathcal{P}_T$. According to the definitions above, this underlying MDP is naturally isomorphic to $\mathcal{M}_T$.

This completes the construction and the proof. □

Remark 8. *As discussed above and in the main text, because 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 6), we can directly apply the conclusions from Theorem 2 for abstraction error control.*

Note that policy truncation is essential here. This is not only to directly reuse the conclusions from Theorem 2, but also due to the "curse of memory"—the memory of a policy can severely affect the quality of theoretical guarantees.

D.1 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 10. *If Assumption 7 hold, then for any $\varepsilon > 0$ that satisfies condition 1, with $T \geq T_1(\varepsilon)$, we have,*

$$|\mu(a_h, \tau_h^+) - \mu(a_h, \tau_{[h-T+1:h]}^+)| \leq \frac{2(C_\mu + 1)L_\pi \varepsilon}{\min_h \min_{a_h, \tau_h^+} \pi_b(a_h|\tau_h^+)} \tag{59}$$

Proof. Using Condition 1, we have

$$\begin{aligned}
|\mu(a_h, \tau_h^+) - \mu(a_h, \tau_{[h-T+1:h]}^+)| \leq & \frac{|\pi_e(a_h|\tau_h^+) - \pi_e(a_h|\tau_{[h-T+1:h]}^+)|}{\pi_b(a_h|\tau_h^+)} \\
& + \pi_e(a_h|\tau_{[h-T+1:h]}^+) \left| \frac{1}{\pi_b(a_h|\tau_h^+)} - \frac{1}{\pi_b(a_h|\tau_{[h-T+1:h]}^+)} \right| \qquad (60) \\
\leq & \frac{L_\pi \varepsilon}{\min_h \min_{a_h, \tau_h^+} \pi_b(a_h|\tau_h^+)} \left(1 + \frac{\pi_e(a_h|\tau_{[h-T+1:h]}^+)}{\pi_b(a_h|\tau_{[h-T+1:h]}^+)} \right) \qquad (61) \\
\leq & \frac{L_\pi \varepsilon}{\min_h \min_{a_h, \tau_h^+} \pi_b(a_h|\tau_h^+)} (2 + 2C_\mu) \qquad (62)
\end{aligned}$$

which proves the lemma. □

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 9 (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 $V \in \mathcal{V}$,*

$$|V(f_h, \tau_h) - V(f_{[h:h+T]}, \tau_{[h-T+1:h]})| \leq \|\mathcal{V}\|_\infty \varepsilon \tag{63}$$

The suitable values of T form a function of ε, denoted as $T_2(\varepsilon)$.

Note that the essential assumption here is that the "history" in the extended future is fast-forgetting. Since in the original literature of FDVF (Uehara et al., 2023), the future is by default truncated by a length M_F.

3. The difference in data-generating distribution between the real POMDP and the abstract short-term memory POMDP. This discrepancy arises from two sources, firstly that the behavior policy is truncated, and secondly, 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:

$$w^{\phi_T}(f_1) := \frac{\pi_b^{\phi_T}(a_1|\tau_1^+)}{\pi_b(a_1|\tau_1^+)} \cdot \frac{P^{\phi_T}(o_2|\tau_2)}{P(o_2|\tau_2)} \cdot \cdots \cdot \frac{\pi_b^{\phi_T}(a_H|\tau_H^+)}{\pi_b(a_H|\tau_H^+)} \tag{64}$$

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

Lemma 11. *If Assumptions 6 and 7 hold, then for any $\varepsilon > 0$, with $T \geq \max\{T_0(\varepsilon), T_1(\varepsilon)\}$, then for ε small enough, namely, when Condition 2 is satisfied, we have*

$$|w^{\phi_T}(f_1)| \leq 1 + \frac{3H\varepsilon}{\min\{\min_h \min_{o_h,\tau_h} P(o_h|\tau_h), \min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)/L_\pi\}} \tag{65}$$

Proof. We first show that using Condition 2, for any h, $\left|1 - \frac{\pi_b^{\phi_T}(a_h|\tau_h^+)}{\pi_b(a_h|\tau_h^+)}\right| \leq \frac{L_\pi \varepsilon}{\min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)}$.

Then using the fact that

$$P^{\phi_T}(o_h|\tau_h) = P^{\phi_T}(o_h|\mathbf{b}(\tau_{h-1}^+), a_{h-1}) = \mathbb{E}_{b\sim p_{\phi_T(\mathbf{b}(\tau_{h-1}^+))}}[P(o_h|b, a_{h-1})] \tag{66}$$

and that for all $b \in \phi_T^{-1}(\phi_T(\mathbf{b}(\tau_{h-1}^+)))$, $\|b - \mathbf{b}(\tau_{h-1}^+)\|_1 \leq \varepsilon$, which is described in Assumption 6. It then follows that

$$\begin{aligned}
|P(o_h|\tau_h) - P^{\phi_T}(o_h|\tau_h)| &= |P(o_h|\mathbf{b}(\tau_{h-1}^+), a_{h-1}) - \mathbb{E}_{b\sim p_{\phi_T(\mathbf{b}(\tau_{h-1}^+))}}[P(o_h|b, a_{h-1})]| \\
&= |\mathbb{E}_{b\sim p_{\phi_T(\mathbf{b}(\tau_{h-1}^+))}}[P(o_h|\mathbf{b}(\tau_{h-1}^+), a_{h-1}) - P(o_h|b, a_{h-1})]| \\
&\leq \mathbb{E}_{b\sim p_{\phi_T(\mathbf{b}(\tau_{h-1}^+))}}[|P(o_h|\mathbf{b}(\tau_{h-1}^+), a_{h-1}) - P(o_h|b, a_{h-1})|] \qquad (67) \\
&\leq \varepsilon \qquad (68)
\end{aligned}$$

where the last step uses Lemma 5. Now we have $\left|1 - \frac{P^{\phi_T}(o_h|\tau_h)}{P(o_h|\tau_h)}\right| \leq \frac{\varepsilon}{\min_h \min_{o_h,\tau_h} P(o_h|\tau_h)}$.

Therefore, $w^{\phi_T} \leq (1 + \frac{\varepsilon}{\min\{\min_h \min_{o_h,\tau_h} P(o_h|\tau_h), \min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)/L_\pi\}})^H$, combining Condition 2 we have $w^{\phi_T} \leq 1 + \frac{3H\varepsilon}{\min\{\min_h \min_{o_h,\tau_h} P(o_h|\tau_h), \min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)/L_\pi\}}$. □

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

Lemma 12. *Define*

$$\mathcal{E}_{\mathcal{V},\Theta}(V,\theta) := \sum_{h=1}^{H} \mathbb{E}_{\mathcal{D}}[\{\mu(a_h, \tau_h^+)(r_h + V(f_{h+1})) - V(f_h)\}\theta(\tau_h) - \frac{1}{2}\theta(\tau_h)^2] \tag{69}$$

$$\begin{aligned}
\mathcal{E}_{\mathcal{V},\Theta}^{\phi_T}(V,\theta) := \sum_{h=1}^{H} \mathbb{E}_{\mathcal{D}}[w^{\phi_T}(\{\mu(a_h, \tau_{h-T+1:h}^+)(r_h + V(\phi_T(f_{h+1}))) - V(\phi_T(f_h))\}\theta(\tau_h) \\
- \frac{1}{2}\theta(\tau_h)^2)]
\end{aligned} \tag{70}$$

where $\phi_T(f_h) = \phi_T(f_h', \tau_h) = (f_h', \phi_T(\tau_h))$. If Assumptions 6,7, and9 all hold, then for any $\varepsilon > 0$ satisfying condition 1, 2 and for any $V \in \mathcal{V}, \theta \in \Theta$, we have:

$$|\mathcal{E}_{\mathcal{V},\Theta}(V,\theta) - \mathcal{E}_{\mathcal{V},\Theta}^{\phi_T}(V,\theta)| \leq \frac{1}{3} L_{\mathcal{E}} \varepsilon \tag{71}$$

and $T = \max\{T_0(\varepsilon), T_1(\varepsilon), T_2(\varepsilon)\}$.

Proof. The proof of such lemma uses triangle's inequality, by summing up the error depicted in Lemma 10, Assumption 9, and Lemma 11, we get the result. □

Theorem 16. *Define*

$$\mathcal{E}_{\mathcal{V}}(V) := \max_{\theta\in\Theta} \mathcal{E}_{\mathcal{V},\Theta}(V,\theta), \quad \mathcal{E}_{\mathcal{V}}^{\phi_T}(V) := \max_{\theta\in\Theta} \mathcal{E}_{\mathcal{V},\Theta}^{\phi}(V,\theta) \tag{72}$$

If Assumptions 6,7, and9 all hold, then for any $\varepsilon > 0$ satisfying condition 1, 2 and for any $V \in \mathcal{V}$, we have:

$$|\mathcal{E}_{\mathcal{V}}(V) - \mathcal{E}_{\mathcal{V}}^{\phi_T}(V)| \leq \frac{2}{3} L_{\mathcal{E}} \varepsilon \tag{73}$$

and $T = \max\{T_0(\varepsilon), T_1(\varepsilon), T_2(\varepsilon)\}$.

Proof. This use the observation that if $\forall \theta, |\mathcal{E}_{\mathcal{V},\Theta}(V,\theta) - \mathcal{E}_{\mathcal{V},\Theta}^{\phi_T}(V,\theta)| \leq \frac{1}{3} L_{\mathcal{E}} \varepsilon$, then $|\sup_\theta \mathcal{E}_{\mathcal{V},\Theta}(V,\theta) - \sup_\theta \mathcal{E}_{\mathcal{V},\Theta}^{\phi_T}(V,\theta)| \leq 2 \cdot \frac{1}{3} L_{\mathcal{E}} \varepsilon$. □

Proof of Theorem 8.

Proof. To prove the theorem, we need to prove $L_\phi^{[2]} := \|\mathcal{V}\|_\infty$ satisfies $|\mathbb{E}_{\pi_b}[\hat{V}(f_1)] - \mathbb{E}_{\pi_b^\phi}[\hat{V}(f_1)]| \leq L_\phi^{[2]} \varepsilon$.

Recall that in our construction, the dynamic of the short-term memory POMDP ensures that the first T action-observation has the exact same dynamic as the true POMDP. Also notice that $\forall f_1^{[1]}, f_1^{[2]} \in \mathcal{F}_1$, if $f_1^{[1]}, f_1^{[2]}$ shares the first T pairs of action and observation, then $|\hat{V}(f_1^{[1]}) - \hat{V}(f_1^{[2]})| \leq \|\mathcal{V}\|_\infty \varepsilon$ as indicated by the property of the function class $\mathcal{V}$, Assumption 9. Then

$$\begin{aligned} |\mathbb{E}_{\pi_b}[\hat{V}(f_1)] - \mathbb{E}_{\pi_b^\phi}[\hat{V}(f_1)]| &\leq \mathbb{E}_{f_{1:T} \sim \pi_b}[|\mathbb{E}_{f_{T+1} \sim \pi_b}[\hat{V}(f_1)|f_{1:T}] - \mathbb{E}_{f_{T+1} \sim \pi_b^\phi}[\hat{V}(f_1)|f_{1:T}]|] \\ &\leq \|\mathcal{V}\|_\infty \varepsilon \end{aligned} \tag{74}$$

which proves the theorem. □

D.2 Theoretical Guarantee of FDVF

Proof of Theorem 9.

Proof. Let $\hat{V} := \arg\min_{V \in \mathcal{V}} \mathcal{E}_{\mathcal{V}}(V)$, and correspondingly $\hat{V}_\phi := \arg\min_{V \in \mathcal{V}} \mathcal{E}_{\mathcal{V}}^{\phi_T}(V)$. Our first goal is to show that with probability greater than $1 - \delta$,

$$\begin{aligned} |J^\phi(\pi_e^\phi) - \mathbb{E}_{\pi_b^\phi}[\hat{V}(f_1)]| \leq \sqrt{H} \cdot \max_{h \in [H]} \sup_{V \in \mathcal{V}} \sqrt{\frac{\mathbb{E}_{\pi_e^\phi}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)} V)(s_h, \tau_{[h-T+1:h]})^2]}{\mathbb{E}_{\pi_b^\phi}[(\mathcal{B}^{\mathcal{H}} V)(\tau_h)^2]}} \\ \cdot \sqrt{\frac{cHC_{\mathcal{V}}^2 C_\mu}{n} \log \frac{|\mathcal{V}||\Theta|}{\delta} + L_{\mathcal{E}} \varepsilon} \end{aligned} \tag{75}$$

To do this, we follow the proof provided by (Zhang and Jiang, 2024), define $X_{V,h}^\phi := \mu(a_h, \tau_h^+)(r_h + V(\phi_T(f_{h+1}))) - V(\phi_T(f_h))$, $X_{V,h} := \mu(a_h, \tau_h^+)(r_h + V(f_{h+1})) - V(f_h)$, then $\mathcal{E}_{\mathcal{V}}^{\phi_T}(V) = \frac{1}{2} \max_{\theta \in \Theta} \sum_{h=1}^H \mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X_{V,h}^\phi)^2 - (X_{V,h}^\phi - \theta(\tau_h))^2)]$, and such θ that achieves maximum is denoted as $\hat{\theta}_V$. Similarly, $\mathcal{E}_{\mathcal{V}}(V) = \frac{1}{2} \max_{\theta \in \Theta} \sum_{h=1}^H \mathbb{E}_{\mathcal{D}}[X_{V,h}^2 - (X_{V,h} - \theta(\tau_h))^2)$ and the θ that achieves maximum is represented by $\hat{\theta}_V^0$. According to a concentration argument using Bernstein's inequality as presented in first part (*Analysis of Inner Maximizer*) of the the proof of theorem 2 in (Zhang and Jiang, 2024), we arrive at an argument that indicates with probability greater than $1 - \delta/2$, for any $V \in \mathcal{V}, \theta \in \Theta$,

$$\begin{aligned} \left| \sum_{h=1}^H \mathbb{E}_{\mathcal{D}}[w^{\phi_T}(\hat{\theta}_V(\tau_h) - X_{V,h}^\phi)^2] - \sum_{h=1}^H \mathbb{E}_{\mathcal{D}}[w^{\phi_T}(X_{V,h}^\phi - (\mathcal{B}^{\mathcal{H}} V)(\tau_h))^2] \right| \\ \leq \frac{675 HC_{\mathcal{V}}^2 C_\mu \|w^{\phi_T}\|_\infty}{n} \cdot \log \frac{4|\mathcal{V}||\Theta|}{\delta} =: \eta \end{aligned} \tag{76}$$

Then we have

$$\begin{aligned}
&\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^{\phi}_{\hat{V},h})^2-(X^{\phi}_{\hat{V},h}-(\mathcal{B}^{\mathcal{H}}\hat{V})(\tau_h))^2)]\\
\leq&\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^{\phi}_{\hat{V},h})^2-(\hat{\theta}_{\hat{V}}(\tau_h)-X^{\phi}_{\hat{V},h})^2)]+\eta &(77)\\
\leq&\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[X^2_{\hat{V},h}-(\hat{\theta}_{\hat{V}}(\tau_h)-X_{\hat{V},h})^2]+\frac{1}{3}L_{\mathcal{E}}\varepsilon+\eta &(78)\\
\leq&\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[X^2_{\hat{V},h}-(\hat{\theta}^0_{\hat{V}}(\tau_h)-X_{\hat{V},h})^2]+\frac{1}{3}L_{\mathcal{E}}\varepsilon+\eta &(79)\\
\leq&\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^{\phi}_{\hat{V}_\phi,h})^2-(\hat{\theta}_{\hat{V}_\phi}(\tau_h)-X^{\phi}_{\hat{V}_\phi,h})^2)]+L_{\mathcal{E}}\varepsilon+\eta &(80)
\end{aligned}$$

Here, (77) uses (76). (78) uses Lemma 12 and the fact that $\mathcal{E}_{\mathcal{V},\Theta}(\hat{V},\hat{\theta}_{\hat{V}})=\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[X^2_{\hat{V},h}-(\hat{\theta}_{\hat{V}}(\tau_h)-X_{\hat{V},h})^2]$ and that $\mathcal{E}^{\phi}_{\mathcal{V},\Theta}(\hat{V},\hat{\theta}_{\hat{V}})=\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^{\phi}_{\hat{V},h})^2-(\hat{\theta}_{\hat{V}}(\tau_h)-X^{\phi}_{\hat{V},h})^2)]$. After that, (79) uses the fact that $\hat{\theta}^0_{\hat{V}}=\arg\max_{\theta\in\Theta}\mathcal{E}_{\mathcal{V},\Theta}(\hat{V},\theta)$, thus $\mathcal{E}_{\mathcal{V},\Theta}(\hat{V},\hat{\theta}^0_{\hat{V}})\geq\mathcal{E}_{\mathcal{V},\Theta}(\hat{V},\hat{\theta}_{\hat{V}})$. (80) uses Theorem 16 and the fact that $\min_{V\in\mathcal{V}}\mathcal{E}_{\mathcal{V}}(V)=\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[X^2_{\hat{V},h}-(\hat{\theta}^0_{\hat{V}}(\tau_h)-X_{\hat{V},h})^2]$ and $\min_{V\in\mathcal{V}}\mathcal{E}^{\phi}_{\mathcal{V}}(V)=\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^{\phi}_{\hat{V}_\phi,h})^2-(\hat{\theta}_{\hat{V}_\phi}(\tau_h)-X^{\phi}_{\hat{V}_\phi,h})^2)]$. Noticing that $\min_{V\in\mathcal{V}}\mathcal{E}_{\mathcal{V}}\leq\min_{V\in\mathcal{V}}\mathcal{E}^{\phi}_{\mathcal{V}}(V)+\sup_{V\in\mathcal{V}}|\mathcal{E}_{\mathcal{V}}(V)-\mathcal{E}^{\phi}_{\mathcal{V}}(V)|$ finish the derivation of (80).

After that, notice the abstract realizability assumption $\exists V^{\phi}_{\mathcal{F}}\in\mathcal{V}$, and that for any $V^{\phi}_{\mathcal{F}}$, we have

$$\begin{aligned}
&\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^{\phi}_{\hat{V}_\phi,h})^2-(\hat{\theta}_{\hat{V}_\phi}(\tau_h)-X^{\phi}_{\hat{V}_\phi,h})^2)]\\
\leq&\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^{\phi}_{V^{\phi}_{\mathcal{F}},h})^2-(\hat{\theta}_{V^{\phi}_{\mathcal{F}}}(\tau_h)-X^{\phi}_{V^{\phi}_{\mathcal{F}},h})^2)] &(81)\\
\leq&\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^{\phi}_{V^{\phi}_{\mathcal{F}},h})^2-(X^{\phi}_{V^{\phi}_{\mathcal{F}},h}-(\mathcal{B}^{\mathcal{H}}V^{\phi}_{\mathcal{F}})(\tau_h))^2)]+\eta=\eta &(82)
\end{aligned}$$

where the second last inequality uses the minimal property of $\hat{V}_\phi$, and the last inequality uses (76). The last equality is the result of the definition of $V^{\phi}_{\mathcal{F}}$ that it is the zero point of bellman residual operator $\mathcal{B}^{\mathcal{H}}$.

Combining (82) and (80), we have

$$\sum_{h=1}^{H}\mathbb{E}_{\mathcal{D}}[w^{\phi_T}((X^{\phi}_{\hat{V},h})^2-(X^{\phi}_{\hat{V},h}-(\mathcal{B}^{\mathcal{H}}\hat{V})(\tau_h))^2)]\leq L_{\mathcal{E}}\varepsilon+\frac{1350HC^2_{\mathcal{V}}C_{\mu}\|w^{\phi_T}\|_{\infty}}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta} \tag{83}$$

The next step is identical to the equation (17) in (Zhang and Jiang, 2024), which, with the help of Bernstein's inequality, gives us that with probability greater than $1-\delta/2$, for any $V\in\mathcal{V}$,

$$\begin{aligned}
&\left|\sum_{h=1}^{H}\{\mathbb{E}_{\mathcal{D}}-\mathbb{E}_{\pi^{\phi}_b}\}[w^{\phi_T}((X^{\phi}_{V,h})^2-(X^{\phi}_{V,h}-(\mathcal{B}^{\mathcal{H}}V)(\tau_h))^2)]\right|\\
\leq&\sqrt{\frac{58HC^2_{\mathcal{V}}C_{\mu}\|w^{\phi_T}\|_{\infty}}{n}\cdot\log\frac{4|\mathcal{V}|}{\delta}\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi^{\phi}_b}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}+\frac{27HC^2_{\mathcal{V}}C_{\mu}\|w^{\phi_T}\|_{\infty}}{n}\cdot\log\frac{4|\mathcal{V}|}{\delta}. &(84)
\end{aligned}$$

Therefore, combining (83) and (84) we get

$$\sum_{h=1}^{H}\mathbb{E}_{\pi_b^\phi}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2] \le L_{\mathcal{E}}\varepsilon + \frac{1377HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta} + \sqrt{\frac{58HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta}\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi_b^\phi}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}, \tag{85}$$

solving which gives us the final result that

$$\sum_{h=1}^{H}\mathbb{E}_{\pi_b^\phi}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2] \le \frac{1406HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta} + L_{\mathcal{E}}\varepsilon + \sqrt{\frac{80707H^2C_{\mathcal{V}}^4C_\mu^2\|w^{\phi_T}\|_\infty^2}{n^2}\cdot\left(\log\frac{4|\mathcal{V}||\Theta|}{\delta}\right)^2 + \frac{29HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta}\cdot 2L_{\mathcal{E}}\varepsilon} \tag{86}$$

$$\le \frac{(1406+\sqrt{80707+29C})HC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta} + L_{\mathcal{E}}\varepsilon \tag{87}$$

where the last step is subject to $2L_{\mathcal{E}}\varepsilon \le \frac{CHC_{\mathcal{V}}^2C_\mu\|w^{\phi_T}\|_\infty}{n}\cdot\log\frac{4|\mathcal{V}||\Theta|}{\delta}$. Notice that under condition 2, $\|w^{\phi_T}\|_\infty \le e$, so this requirement is covered by condition 3.

Then (75) is shown using the telescoping property of bellman residual operator and the fact that $\|w^{\phi_T}\|_\infty \le e$ as has been mentioned.

Now that we've obtained (75), it suffice to prove the theorem using the result from Theorem 8, which indicates that

$$|J(\pi_e) - J^\phi(\pi_e^\phi)| + |\mathbb{E}_{\pi_b}[\hat{V}(f_1)] - \mathbb{E}_{\pi_b^\phi}[\hat{V}(f_1)]| \le L_\phi\varepsilon. \tag{88}$$

And we prove the theorem by applying Meta-theorem 3. Note that when applying the Meta-theorem, we specify $\mathfrak{est}^\phi(\hat{V}) = \mathbb{E}_{\pi_b^\phi}[\hat{V}(f_1)]$ and $\mathfrak{est}^\phi(V_\phi^\pi) = J^\phi(\pi_e^\phi)$, the latter is the ground truth estimation on the abstract system. □

Corollary 3. *Under the conditions of the Theorem 9, with probability greater then $1-\delta$, we have:*

$$|J(\pi_e) - \mathbb{E}_{\pi_b}[\hat{V}(f_1)]| \le \inf_{\substack{\varepsilon\ge 0\\ D(\varepsilon)}}\left(L_\phi\varepsilon + \sqrt{H}\cdot\max_{h\in[H]}\sup_{V\in\mathcal{V}}\sqrt{\frac{\mathbb{E}_{\pi_e^\phi}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h,\tau_{[h-T+1:h]})^2]}{\mathbb{E}_{\pi_b^\phi}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}}\cdot\sqrt{\frac{cHC_{\mathcal{V}}^2C_\mu}{n}\log\frac{|\mathcal{V}||\Theta|}{\delta} + L_{\mathcal{E}}\varepsilon}\right) \tag{89}$$

where $D(\varepsilon)$ stands for such ε that satisfies abstract realizability, Bellman completeness and condition 1, 2 and 3.

Proof. Combining Lemma 3 and Theorem 9, we get the result. □

Proof of Corollary 2

Proof. This is obtained by choosing $L_{\mathcal{E}}\varepsilon = \frac{c'HC_{\mathcal{V}}^2C_\mu}{n}\log\frac{|\mathcal{V}||\Theta|}{\delta}$ for some constant c' in Theorem 9. □

D.3 A Simpler Pipeline: Abstracting Only the Policy

Proof of Theorem 10.

Proof. The new L_ϕ is obtained using Theorem 12 combining with Theorem 8, and the new $L_{\mathcal{E}}$ is analogous to the result of Lemma 12. The rest of the proof is exactly identical to that of Theorem 9. □

E WHY OUR COVERAGE IS BETTER

A Toy Example: History Space vs. Belief-Space Covering The following simple example illustrates the potential gap between the size of the history space and the covering complexity of the reachable belief space.

Consider a two-state POMDP with two non-informative observations, $|\mathcal{A}|$ actions, and a transition kernel that does not depend on the action:

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}.$$

Let the initial belief be $b_0 = (1, 0)$. At time t, the number of possible action histories is $|\mathcal{A}|^t$.

Because the observations are non-informative, the belief is completely determined by the time index t rather than by the full history. A direct calculation gives

$$b_t = \Big(0.5 + 0.5 \cdot 0.8^t,\ 0.5 - 0.5 \cdot 0.8^t\Big),$$

and therefore

$$\|b_t - (0.5, 0.5)\|_1 = 0.8^t.$$

Hence, up to horizon T, the reachable belief set

$$\mathcal{B}_{\le T} := \{b_t : 0 \le t \le T\}$$

lies on a one-dimensional contracting trajectory toward the stationary belief $(0.5, 0.5)$. Its ε-covering number satisfies

$$\text{Covering}(\varepsilon, \mathcal{B}_{\le T}) = 1 + \min\Big\{T, \Big\lfloor \log_{1.25} \frac{1}{\varepsilon} \Big\rfloor\Big\},$$

whereas the number of length-T histories is $|\mathcal{A}|^T$.

This example shows that, even though the history space grows exponentially with T, the reachable belief space can remain extremely simple: its covering complexity grows only logarithmically in $1/\varepsilon$. In the infinite-horizon case $T = \infty$, the number of histories is infinite, while

$$\text{Covering}(\varepsilon, \mathcal{B}) = 1 + \Big\lfloor \log_{1.25} \frac{1}{\varepsilon} \Big\rfloor.$$

In terms of Figure 1, the exponential dependence corresponds to the true-system coverage (red arrow), whereas the logarithmic dependence corresponds to the abstract-system coverage after passing to belief-space covering (green arrow 2).

Elaboration on example 1 In this example, we consider a belief space with a smoothness structure (Section 5.3 (Lee et al., 2007)) denoted as follow:

$\mathcal{B}$ is a bounded subset in a $|\mathcal{S}|$-dimensional vector space, assume that every belief $b \in \mathcal{B}$ can be represented by m basis vectors through linear combinations, and the magnitudes of both the basis elements and the linear coefficients are bounded above by a constant C. Then the covering number for our belief space scales as $O((C|\mathcal{S}|L_{\mathcal{E}}m)^m \cdot (\frac{32R_{\max}^4}{n(1-\gamma)^4} \cdot \log \frac{2|\mathcal{F}|}{\delta})^{\frac{m}{2}})$. We assume the coverage being sublinear polynomial w.r.t. its worst case (i.e. the covering number), specifically to the power of $\frac{1}{2m}$. Then we have a finite sample guarantee of $O(\frac{(C|\mathcal{S}|L_{\mathcal{E}}mR_{\max}^2)^{\frac{1}{4}}}{(1-\gamma)^{\frac{3}{2}}} \cdot (\frac{1}{n} \log \frac{|\mathcal{F}|}{\delta})^{\frac{1}{8}})$. Note that we only assume sublinear polynomial instead of logarithmic since the latter is too strong, and may directly resolve the exponentiality.

Proof of Theorem 4

Proof. We proof the theorem by constructing $d^D(\tau') = \sum_{\{\tau:\tilde{\phi}_T(\tau)=\tau'\}} d^{\pi_b}(\tau)$. Then noticing that $d_{\phi}^{\pi_e^{\phi}}(\tau') = \sum_{\{\tau:\tilde{\phi}_T(\tau)=\tau'\}} d^{[\pi_e^{\phi}]_{\text{true}}}(\tau)$ is automatically satisfied by how the abstraction ϕ_T is defined. Also, it's not difficult to notice that in the one-hot belief state scenario, $\frac{d^{[\pi_e^{\phi}]_{\text{true}}}(s_h,\tau_h)}{d^{\pi_b}(s_h,\tau_h)} = \frac{d^{[\pi_e^{\phi}]_{\text{true}}}(\tau_h)}{d^{\pi_b}(\tau_h)}$, and it's exactly the same for the short-term memory POMDP induced by $\tilde{\phi}_T$ as we constructed. Here, $\tilde{\phi}_T : \mathcal{H} \to \mathcal{H}_T \subset \mathcal{H}$.

Then, consider two σ-algebras $\mathcal{A} := \mathcal{P}(\mathcal{H})$ and $\mathcal{D} := \tilde{\phi}_T^{-1}(\mathcal{P}(\mathcal{H}_T))$, and it's obvious that $\mathcal{D} \subset \mathcal{A}$. Define the probability point measure P^{π_e} and P^{π_b} corresponding to the weight function $d^{[\pi_e^{\phi}]_{\text{true}}}$ and d^{π_b} on the σ-algebras $\mathcal{A}$, then the probability measure can also be restricted to the smaller σ-algebra $\mathcal{D}$. It is easy to notice that the two terms we try to compare coincides with the χ^2-divergence between P^{π_e} and P^{π_b}, where for the LHS we use the coarser σ-algebra $\mathcal{D}$, and use the finer σ-algebra $\mathcal{A}$ for the RHS.

Then we use the variational representation of χ^2-divergence to obtain our final result, by noticing that

$$\chi^2_{\mathcal{D}}(P^{\pi_e} \| P^{\pi_b}) = \sup_{g \in \mathcal{M}(\mathcal{D})} \mathbb{E}_{P^{\pi_e}}[g(\tau_h)] - \mathbb{E}_{P^{\pi_b}}[g(\tau_h)^2/4 + g(\tau_h)] \tag{90}$$

$$\chi^2_{\mathcal{A}}(P^{\pi_e} \| P^{\pi_b}) = \sup_{g \in \mathcal{M}(\mathcal{A})} \mathbb{E}_{P^{\pi_e}}[g(\tau_h)] - \mathbb{E}_{P^{\pi_b}}[g(\tau_h)^2/4 + g(\tau_h)] \tag{91}$$

Since $\mathcal{D} \subset \mathcal{A}$, any g that is $\mathcal{D}$ measurable is also $\mathcal{A}$ measurable, consequently

$$\chi^2_{\mathcal{D}}(P^{\pi_e} \| P^{\pi_b}) \leq \chi^2_{\mathcal{A}}(P^{\pi_e} \| P^{\pi_b}) \tag{92}$$

which proves the theorem. □

Proof of Theorem 5

Proof. Construct d^D exactly as in Theorem 4, then let $w^\star(\tau_h) = \frac{d^{[\pi_e^{\phi}]_{\text{true}}}(\tau_h)}{d^D(\tau_h)}$, and $\tau_h^\star$ is when achieves the maximum. Similarly, let $w^*(\tilde{\phi}_T(\tau)) = \frac{d^{\pi_e^{\phi}}(\tilde{\phi}_T(\tau))}{d^D(\tilde{\phi}_T(\tau))}$, and $\tilde{\phi}_T(\tau^*)$ is when achieves the maximum. It's obvious that $\forall \tau_h$ such that $\tilde{\phi}_T(\tau_h) = \tilde{\phi}_T(\tau^*)$, $w^\star(\tau_h) \leq w^\star(\tau_h^\star)$. Denote $\tau_h' := \arg\max_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} w^\star(\tau_h)$, then $w^*(\tilde{\phi}_T(\tau^*)) = \frac{\sum_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} d^{[\pi_e^{\phi}]_{\text{true}}}(\tau_h)}{\sum_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} d^D(\tau_h)}$. Notice that

$$\frac{\sum_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} d^{[\pi_e^{\phi}]_{\text{true}}}(\tau_h)}{\sum_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} d^D(\tau_h)} \leq \frac{\sum_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} d^D(\tau_h) \cdot d^{[\pi_e^{\phi}]_{\text{true}}}(\tau_h')/d^D(\tau_h')}{\sum_{\tilde{\phi}_T(\tau_h)=\tilde{\phi}_T(\tau^*)} d^D(\tau_h)} = w^\star(\tau_h')$$

Consequently, $w^*(\tilde{\phi}_T(\tau^*)) \leq w^\star(\tau_h') \leq w^\star(\tau_h^\star)$, which prove the theorem. □

F RELATION WITH DEEP ABSTRACTION

In this section, we discuss our relation with OPE methods that explicitly construct abstractions. Our method uses abstraction purely as a tool for analysis: we analyze existing algorithms without changing them, but simply reveal when and how these algorithms admit improved guarantees due to belief-space smoothness. On the contrary, some other methods actively construct an abstraction to simplify OPE. In such cases, algorithms running on an abstract system (thus simpler than the original system) may achieve smaller error guarantees. In this section, we briefly compare our idea and that of deep abstraction (Hao et al., 2024).

In (Hao et al., 2024), they designed a method that construct a deep abstraction in MDPs using the conventional abstraction techniques, by applying two different methods of abstraction recursively to obtain a deep abstraction. And provably, the variance of the abstracted system monotonously decreases as the abstraction goes deeper.

Comparison with our settings.

1. **Differences in type and strictness of abstraction:** The abstraction in this paper requires, at each step, an either forward-model-irrelevant condition or backward-model-irrelevant condition. As we know, bisimulation, whether or not in its approximate version, is a very strong condition to fulfill, and becomes especially restrictive in belief spaces with exponential cardinality and limited structure. Also, since it's using conventional abstraction skills, it doesn't require the metric structure of the state space. In contrast, our abstraction is based on an ε-net over the belief space, which leverages the metric geometry of the space and applies uniformly to a wide range of POMDPs regardless of structural assumptions.

2. **In solving the curse of horizon:** Indeed, (Hao et al., 2024) elegantly showed that the MSE monotonously decreases as the abstraction goes deeper. But to address the curse of memory/horizon via abstraction, one must analyze how coverage improves in the abstract space. Notably, directly applying their analysis to POMDPs reveals that Assumptions 2 and 4 implicitly hide an exponential constant within $O(1)$. This constant stems from the boundedness of the function class $\mathcal{W}$, which includes the MIS ratio $\hat{w}^\pi$ and is assumed finite under Assumption 2. While this is acceptable in MDPs where no curse of horizon/memory exists, in POMDPs, it is crucial to account for how abstraction influences this exponential term.

G FUTURE ALGORITHM DESIGN

While our paper focuses on the theoretical framework, the stability perspective of our analysis naturally inspires concrete algorithmic ideas for future work.

1. **Stability-regularized training:** Augment Bellman-error minimization or value-function fitting with an additional penalty term

$$\lambda \mathbb{E}_{\mathcal{D}^{\otimes 2}}[\mathbb{I}(\|\hat{b}_1 - \hat{b}_2\|_1 \leq \epsilon) \cdot |V(\hat{b}_1) - V(\hat{b}_2)|].$$

2. **Post-training stability selection:** Train multiple candidate policies, then select the one with the highest empirical stability measured over belief neighborhoods. Theoretically, this is equivalent to the above penalty approach as $\lambda \to 0$.

H LIMITATIONS

Despite our general result is provably no worse than the original coverage assumption, it is possible in some circumstances that the metric property of belief space cannot improve the coverage either. The simplest scenario to consider is when every history has a unique one-hot belief state, and the POMDP is merely equivalent to an MDP with exponentially large state space. In this case, the belief metric is a discrete metric, for $\forall b_1, b_2 \in \mathcal{B}, b_1 \neq b_2 \to \|b_1 - b_2\|_1 = 2$, and the covering number is exactly the cardinality of the space, which is exponential. This reveals the limitation of our analysis in cases when belief space is sparse, or when lack of some specific smoothness structure. However, information-theoretically, OPE problems for POMDPs always suffer from the curse of Horizon in the most general case as shown in (Zhang and Jiang, 2025), meaning that structural assumptions or specific properties of the system must be utilized to gain meaningful progress.

Another limitation is when sample size becomes too large comparing to the horizon H. Notice that in the finite sample argument provided by our result (e.g. Corollary 1), the abstract coverage depends on the approximation level ε, which is set to $O(n^{-1})$. If n becomes too large in this case, the $O(n^{-1})$-covering number will converge to the cardinality of the space $\mathcal{B}$ itself, which is exponential w.r.t. the horizon H. This also trivialize our analysis. Therefore, when considering finite horizon POMDPs, the horizon should be relatively large comparing to the sample size for our result to be valid.