A Covering Framework for Offline POMDPs Learning using Belief Space Metric

Youheng Zhu
IEMS Department
Northwestern University
youhengzhu@u.northwestern.edu

Yiping Lu
IEMS Department
Northwestern University
yiping.lu@northwestern.edu

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) aims to estimate the return of a target policy using data collected by a different behavior policy. While well studied in fully observable Markov decision processes (MDPs), extending OPE to partially observable MDPs (POMDPs) stays challenging [28]. Partial observability induces non-Markovian dynamics over histories, so directly treating trajectories as states leads to *exponential dependence on the horizon*, or the *curse of horizon*. Recent work such as future-dependent value functions (FDVF) alleviates this issue for memoryless policies, but for memory-based policies, coverage scales exponentially with memory length, causing the *curse of memory* [27].

In the POMDP planning literature, however, it is well known that the *belief space*—the set of probability distributions over latent states—possesses a rich metric structure. Point-based value iteration (PBVI) and its variants [20, 11, 17, 19, 21, 22] exploit this geometry to sparsely cover reachable beliefs, leading to complexity polynomial in the covering number of belief space [20, 13, 29]. In contrast, most OPE methods in MDPs, such as importance sampling [18, 9, 8, 7], fitted Q-iteration [5, 15, 12], double sampling [3], min-max estimators [1, 4, 6, 16, 24, 26], and marginalized importance sampling [18, 9, 3, 24], largely ignore this structure. Applied naively to POMDPs, they suffer from the curse of horizon due to exponentially large history spaces. FDVF [25, 27] partially mitigates this by shifting coverage to latent states, but only for memoryless policies; with memory-based policies it reverts to the curse of memory. Therefore, we ask:

Can belief-space geometry similarly reduce the coverage thus complexity of OPE in POMDP settings?

We answer affirmatively by introducing a **covering framework** for OPE in POMDPs. Our key idea is to replace history-space coverage with ε-coverings in belief space, and by restricting our problem

Preprint. Under review.

to a subset of "good" policies, namely those that exhibit stability w.r.t. the beliefs, we yield error bounds that are provably no worse than existing ones and, under mild assumptions, avoid exponential blow-ups in horizon or memory length. To summarize our contribution:

- We propose a unified abstraction framework using belief-space coverings, applicable to a broad class of OPE algorithms.
- We prove that coverage in belief space is always at least as favorable as history-space coverage, and under structural assumptions yields polynomial guarantees.
- We demonstrate the framework on (i) double sampling Bellman error minimization and (ii) future-dependent value functions, obtaining tighter finite-sample bounds. Further discussions also indicates that the "curse of memory" is much easier to handle than the "curse of horizon", in the sense that the former requires no assumption on the POMDP itself.

Before we start, we would also like to guide the readers to the **definitions, terminologies and notations** of our core concepts, i.e. POMDPs, state abstraction [14], belief space, etc. in Appendix A.

Table 1: Two illustrative examples from our analysis. "Worst-case coverage" refers to the most exploratory data-collection distribution, where $\text{Covering}(\mathcal{B}, \varepsilon)$ denotes the L_1 covering number of belief space $\mathcal{B}$. For $H \to \infty$, we assume worst-case coverage grows as a subpolynomial power $\alpha_0 \leq 1$ (not logarithmic, which would trivially remove the curse of horizon). In the FDVF case, specific forgetting rates may be required.

Criteria	**Existing Coverage With Curse of Horizon/Memory**		**Our Coverage using Belief Space Smoothness**
Bellman Error Minimization (*e.g.* Double Sampling)			
Coverage Definition [10]	$\left\lVert \frac{d^{\pi_e}(\tau_h, a)}{d^D(\tau_h, a)} \right\rVert_\infty$		$\left\lVert \frac{d^{\pi_e^\phi}(\phi(b), a)}{d^D(\phi(b), a)} \right\rVert_\infty$
Coverage Worst Case Scale	$\lvert\mathcal{B}\rvert = \Theta((\lvert\mathcal{O}\rvert\lvert\mathcal{A}\rvert)^H)$	>	$\text{Covering}(\mathcal{B}, \Theta(n^{-1/2}))$
Ability to handle $H \to \infty$	✘: Infinite		✓: Polynomial guarantee see example 1
Future Dependent Value Function			
Coverage Definition [27]	$\sup_{h,V} \sqrt{\frac{\mathbb{E}_{\pi_e}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_H)}V)(s_h,\tau_h)^2]}{\mathbb{E}_{\pi_b}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}}$		$\sup_{h,V} \sqrt{\frac{\mathbb{E}_{\pi_e^\phi}[(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h,\tau_{[h-T+1:h]})^2]}{\mathbb{E}_{\pi_b^\phi}[(\mathcal{B}^{\mathcal{H}}V)(\tau_h)^2]}}$
L_2 Belief Coverage (One-hot Belief) [27]	$\mathbb{E}_{\pi_b}\left[\left(\frac{d^{\pi_e}(s_h, \tau_h)}{d^{\pi_b}(s_h, \tau_h)}\right)^2\right]$	> Theorem 5	$\mathbb{E}_{\pi_b^\phi}\left[\left(\frac{d_\phi^{\pi_e^\phi}(s_h, \tau_{[h-T+1:h]})}{d_\phi^{\pi_b^\phi}(s_h, \tau_{[h-T+1:h]})}\right)^2\right]$
L_∞ Belief Coverage (One-hot Belief) [27]	$\left\lVert \frac{d^{\pi_e}(s_h, \tau_h)}{d^{\pi_b}(s_h, \tau_h)} \right\rVert_\infty$	> Theorem 6	$\left\lVert \frac{d_\phi^{\pi_e^\phi}(s_h, \tau_{[h-T+1:h]})}{d_\phi^{\pi_b^\phi}(s_h, \tau_{[h-T+1:h]})} \right\rVert_\infty$
L_∞ Worst Case (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 $H \to \infty$	✘: Infinite		✓: Polynomial guarantee see example 2

2 Unified Analysis In a Nutshell

In this section, we briefly explain the core idea of how belief space metric can be used 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 are able to treat near belief states as one, making the space simpler. This process of abstraction is merely done conceptually, as an analytic tool, and does not explicitly appear in the algorithm. In fact, our framework is designed to be algorithm-agnostic, making it widely applicable.

To do this, 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 assumptions:

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

Assumption 2 (Value Stability)**.** $\sup_{\substack{b_1, b_2 \in \mathcal{B} \\ \varepsilon \geq 0, \phi_\varepsilon}} \frac{|V^{[\pi_{\phi_\varepsilon}]_{\mathrm{true}}}(b_1) - V^{[\pi_{\phi_\varepsilon}]_{\mathrm{true}}}(b_2)|}{\|b_1 - b_2\|_1} \leq L_V < \infty$

Remark 1. *Assumption 1 is made by the intuition 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 [13])**.** *For any $b_1, b_2 \in \mathcal{B}$, $|V^*(b_1) - V^*(b_2)| \leq \frac{R_{\max}}{1-\gamma} \|b_1 - b_2\|_1$. Here $V^*(b)$ is the optimal value function.*

2.1 Unified Analysis In a Nutshell

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 5, and 6 to show that abstract coverage is no worse than the original coverage. Eventually for step 3, we utilize the Lipchitz 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. In this paper, we construct the abstraction using a ε-cover $\mathcal{C}_\varepsilon$ of the belief space, with definition in Appendix A.

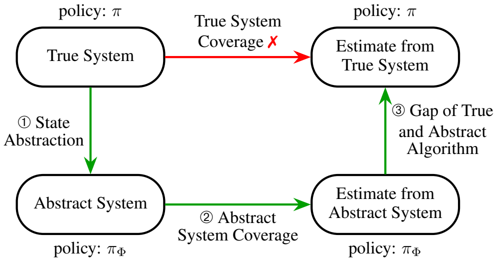

Figure 1: Pipeline of the analysis

3 Application of the Unified Analysis

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. The definitions for the two algorithms is presented in Appendix A, other notations may be introduced in either Appendix A or B. Some technical assumptions compensating or displaying alternative forms of the previous Assumption 1, 2 are illustrated in Appendix B as well.

Definition 1. *We define $L_\phi^{[1]} := \frac{(L_\pi+1)R_{\max}+2L_V}{1-\gamma} + \frac{\gamma R_{\max} L_\pi + R_{\max}}{(1-\gamma)^2}$ for the infinite horizon case, and $L_\phi^{[1]} := H(L_\pi + 1)R_{\max} + 2HL_V + \gamma H^2 R_{\max} L_\pi + H^2 R_{\max}$ for the finite case.*

3.1 Analysis on Bellman Error Minimization Algorithms

Theorem 2. *If Assumptions 4, 5, 1, and 2 all hold, with L_Q mentioned in Assumption 5, 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 $L_{\mathcal{E}} = \frac{8R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)$, $L_\phi = L_\phi^{[1]} + L_\phi^{[2]}$. $L_\phi^{[1]}$ is defined in Definition 1, $L_\phi^{[2]} = \frac{R_{\max}}{1-\gamma} + L_Q$ and $D(\varepsilon)$ stands for such ε that satisfies realizability (Assumption 4).

Corollary 1 (Finite sample guarantee)**.** *If Assumptions 5,1, and2 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 $\varepsilon = \frac{1}{L_{\mathcal{E}}}\sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}}$ satisfies Assumption 4, we have $|J_{\hat{Q}^\pi}(\pi) - J(\pi)| \leq \frac{2\sqrt{C_\pi^n}}{1-\gamma}\cdot\big(\frac{128R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}\big)^{\frac{1}{4}}$, where $C_\pi^n := C_\pi\big(\frac{1}{L_{\mathcal{E}}}\sqrt{\frac{32R_{\max}^4}{n(1-\gamma)^4}\cdot\log\frac{2|\mathcal{F}|}{\delta}}\big)$.*

3.2 Future-Dependent Value Function.

Theorem 3 (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 8 are satisfied. For any $\varepsilon > 0$ satisfying condition 1, 2, 3, define $T = \max\{T_0(\varepsilon), T_1(\varepsilon), T_2(\varepsilon)\}$, $L_\phi = L_\phi^{[1]} + \|\mathcal{V}\|_\infty$.*

Then, for some uniform constant c, with probability at least $1-\delta$, we have:

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

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

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

A Simpler Pipeline: Abstracting Only the Policy. Revisiting the above analysis, it becomes apparent that the abstraction from the original POMDP to the short-term memory POMDP is unnecessary. 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 all be eliminated in this case. **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 $H > 1$.

Theorem 4 (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 8 are satisfied. For any $\varepsilon > 0$ satisfying condition $2'$, 3, define $T = \max\{T_1(\varepsilon), T_2(\varepsilon)\}$, $L_\phi = L_\phi^{[1]} + \|\mathcal{V}\|_\infty$.*

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

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

where $C_{\mathcal{V}} := \max\{\|\mathcal{V}\|_\infty + 1, \|\Theta\|_\infty\}$, $L_\phi = R_{\max}HL_\pi + R_{\max}H^2L_\pi + \|\mathcal{V}\|_\infty$ and $L_\varepsilon = 3\big(\frac{HL_\pi(c_1(C_\mu+1)\|\mathcal{V}\|_\infty\|\Theta\|_\infty + c_2H\max\{C_\mu\|\mathcal{V}\|_\infty\|\Theta\|_\infty, \frac{1}{2}\|\Theta\|_\infty^2\})}{\min_h\min_{a_h,\tau_h^+}\pi_b(a_h|\tau_h^+)} + HC_\mu\|\mathcal{V}\|_\infty\|\Theta\|_\infty\big)$

Acknowledgments

YZ thanks Nan Jiang for useful discussions.

References

[1] András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

[2] Kavosh Asadi, Dipendra Misra, and Michael Littman. Lipschitz continuity in model-based reinforcement learning. In *International conference on machine learning*, pages 264–273. PMLR, 2018.

[3] Leemon Baird et al. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the twelfth international conference on machine learning*, pages 30–37, 1995.

[4] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pages 1042–1051. PMLR, 2019.

[5] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

[6] Yihao Feng, Lihong Li, and Qiang Liu. A kernel loss for solving the bellman equation. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[8] Binyan Jiang, Rui Song, Jialiang Li, and Donglin Zeng. Entropy learning for dynamic treatment regimes. *Statistica Sinica*, 29(4):1633, 2019.

[9] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pages 652–661. PMLR, 2016.

[10] Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 2024.

[11] Hanna Kurniawati, David Hsu, and Wee Sun Lee. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and systems*, volume 2008. Citeseer, 2008.

[12] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

[13] Wee Lee, Nan Rong, and David Hsu. What makes some pomdp problems easy to approximate? *Advances in neural information processing systems*, 20, 2007.

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

[15] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[16] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems*, 32, 2019.

[17] Pascal Poupart, Kee-Eung Kim, and Dongho Kim. 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, 2011.

[18] Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *ICML*, volume 2000, pages 759–766. Citeseer, 2000.

[19] Guy Shani, Ronen I Brafman, and Solomon Eyal Shimony. Prioritizing point-based pomdp solvers. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(6):1592–1605, 2008.

[20] Guy Shani, Joelle Pineau, and Robert Kaplow. A survey of point-based pomdp solvers. *Autonomous Agents and Multi-Agent Systems*, 27:1–51, 2013.

[21] Trey Smith and Reid Simmons. Point-based pomdp algorithms: Improved analysis and implementation. *arXiv preprint arXiv:1207.1412*, 2012.

[22] Matthijs TJ Spaan and Nikos Vlassis. Perseus: Randomized point-based value iteration for pomdps. *Journal of artificial intelligence research*, 24:195–220, 2005.

[23] Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022.

[24] Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

[25] Masatoshi Uehara, Haruka Kiyohara, Andrew Bennett, Victor Chernozhukov, Nan Jiang, Nathan Kallus, Chengchun Shi, and Wen Sun. Future-dependent value-based off-policy evaluation in pomdps. *Advances in neural information processing systems*, 36:15991–16008, 2023.

[26] Andrea Zanette and Martin J Wainwright. Bellman residual orthogonalization for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35:3137–3151, 2022.

[27] Yuheng Zhang and Nan Jiang. On the curses of future and history in future-dependent value functions for off-policy evaluation. *arXiv preprint arXiv:2402.14703*, 2024.

[28] Yuheng Zhang and Nan Jiang. Statistical tractability of off-policy evaluation of history-dependent policies in pomdps. *arXiv preprint arXiv:2503.01134*, 2025.

[29] Zongzhang Zhang, David Hsu, and Wee Sun Lee. Covering number for efficient heuristic-based pomdp planning. In *International conference on machine learning*, pages 28–36. PMLR, 2014.

A Definitions and Notations

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 $\mathcal{S}, \mathcal{A}, \mathcal{O}$.

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 Chapter 3.2. 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 chapter 6, for the future-dependent value function (FDVF), the definition of offline data differs slightly. In the FDVF setting, we consider a finite-horizon POMDP of length H. Again, a behavior policy π_b is used to interact with the environment and collect data. This time, the entire trajectory is treated as a single data point, 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 a MDP $(\mathcal{S}, \mathcal{A}, r, \gamma, P)$ 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 $(\mathcal{S}_\phi, \mathcal{A}, r_\phi, \gamma, P_\phi)$ 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: We denote state-action occupancy as $d^\pi(s, a) := (1 - \gamma) \sum_{k=1}^{\infty} \Pr_\pi(S_k = s, A_k = a)$. $J(\pi)$ represents the expected reward of a policy π, and $J_{\hat{Q}}(\pi)$ is the estimated reward of π using approximation function $\hat{Q}$. We use $\mathcal{F}$ to represent the function class we use for function approximation, in the bellman residual minimization setting it contains all candidate estimates of value functions, and in the FDVF setting it contains all candidate future dependent value functions.

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

$$\tau_h = (o_1, a_1, o_2, a_2, \cdots, o_{h-1}, a_{h-1}) \in \mathcal{H}_h \subset \mathcal{H}, \quad \tau_h^+ := (\tau_h, o_h) \in \mathcal{H}_h^+ \subset \mathcal{H}^+ \tag{3}$$

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 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.

A ε-Cover. For the definition of a ε-cover, we put:

Definition 2. *A ε-cover $\mathcal{C}_\varepsilon$ is a subspace of the belief state space which satisfies:*

$$\mathcal{B} \subset \bigcup_{c \in \mathcal{C}_\varepsilon} \mathbf{B}(c, \varepsilon) \tag{4}$$

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.

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

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

$$\hat{Q}^\pi = \operatorname*{arg\,min}_{f \in \mathcal{F}} \mathcal{E}(f, \pi), \quad \mathcal{E}(f, \pi) = \mathbb{E}_{\mathcal{D}}[(f(b, a) - (r + \gamma f(b'_A, \pi)))(f(b, a) - (r + \gamma f(b'_B, \pi)))]$$

Remark 2. *It is worth noting that the coverage $C_\pi(\phi)$ here depends on the specific abstraction mapping ϕ. Under the most exploratory data collection distribution d^D, the worst-case growth rate of $C_\pi(\phi)$ is approximately aligned with $|\mathcal{C}_\varepsilon|$, which denotes the ε-covering number. The benefit of the belief-policy coverage Assumption 3 lies in its potential to outperform coverage assumptions in the original space, while being generally no worse than that as well. Using an abstract belief space allows the exponentially large history space to be reduced to a space with size of ε-covering number.*

Future-Dependent Value Function: Setup and Definitions. 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 [27]. 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 3 (Memory-Based Bellman Residual Operator)**.**

$$\begin{aligned}(\mathcal{B}^{(\mathcal{S},\mathcal{H}_T)}V)(s_h, \tau_{[h-T+1:h]}) := \mathbb{E}_{\substack{a_{1:h}\sim\pi_e \\ a_{h+1:H}\sim\pi_b}}[r_h + V(f_{h+1}) | s_h, \tau_{[h-T+1:h]}] \\ - \mathbb{E}_{\substack{a_{1:H-1}\sim\pi_e \\ a_{h:H}\sim\pi_b}}[V(f_h) | s_h, \tau_{[h-T+1:h]}]\end{aligned} \tag{5}$$

$$(\mathcal{B}^{\mathcal{H}}V)(\tau_h) := \mathbb{E}_{\substack{a_{1:h}\sim\pi_e \\ a_{h+1:H}\sim\pi_b}}[r_h + V(f_{h+1}) | \tau_h] - \mathbb{E}_{\substack{a_{1:h-1}\sim\pi_e \\ a_{h:H}\sim\pi_b}}[V(f_h) | \tau_h] \tag{6}$$

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:

$$\hat{V}_{\mathcal{F}} = \operatorname*{arg\,min}_{V \in \mathcal{V}} \max_{\theta \in \Theta} \sum_{h=1}^{H} \mathbb{E}_{\mathcal{D}}[\{\mu(a_h, \tau_h^+)(r_h + V(f_{h+1})) - V(f_h)\}\theta(\tau_h) - \frac{1}{2}\theta(\tau_h)^2] \tag{7}$$

B Technical Assumptions and Conditions

In this section, we list some technical assumptions or variations of Assumption 1, 2 to help with the proof. **Note that some assumptions are not necessary for the tightest guarantee, but were listed here for the completeness of narrative.**

B.1 Double sampling example

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)**.** $\|d^{\pi_\phi}/d^D\|_\infty \le C_\pi(\phi) < \infty$

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

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

B.2 Future-dependent value function

Definition 4. *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 $L_{\mathcal{E}} := 3\big(\frac{2H(C_\mu+1)L_\pi\|\mathcal{V}\|_\infty\|\Theta\|_\infty}{\min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)} + HC_\mu\|\mathcal{V}\|_\infty\|\Theta\|_\infty + \frac{3H^2 \max\{C_\mu\|\mathcal{V}\|_\infty\|\Theta\|_\infty, \frac{1}{2}\|\Theta\|_\infty^2\}}{\min\{\min_h \min_{o_h,\tau_h} P(o_h|\tau_h), \min_h \min_{a_h,\tau_h^+} \pi_b(a_h|\tau_h^+)/L_\pi\}}\big)$.*

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 $\|b_1 - b_2\|_1 \le \varepsilon$. The values of T satisfying this condition form a function of ε, denoted $T_1(\varepsilon)$.*

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 $\|\pi(\tau_h^{[1]+}) - \pi(\tau_h^{[2]+})\|_1 \le L_\pi\varepsilon$. We denote the dependency of T on ε as $T_1(\varepsilon)$.*

Lemma 1 (Fast-Forgetting weaker than Lipchitz)**.** *If Assumption 6 and Assumption 1 hold, then Assumption 7 holds aotomatically, with $T_1 = T_0$.*

Assumption 8 (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 $V \in \mathcal{V}$,*

$$|V(f_h, \tau_h) - V(f_{[h:h+T]}, \tau_{[h-T+1:h]})| \le \|\mathcal{V}\|_\infty\varepsilon \tag{8}$$

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 [25], the future is by default truncated by a length M_F.

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.

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

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

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

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

C Why our Coverage is Better

In the following part, we showcase the general idea why our coverage is generally no worse than the original coverage by providing the three theorems. 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 C) 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-divergence.

Theorem 5. *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, we have*

$$\mathbb{E}_{d_\phi^D}\left[\left(\frac{d_\phi^{\pi_e^\phi}(s_h, \tau_{[h-T+1:h]})}{d_\phi^D(s_h, \tau_{[h-T+1:h]})}\right)^2\right] \leq \mathbb{E}_{\pi_b}\left[\left(\frac{d^{[\pi_e^\phi]_{\text{true}}}(s_h, \tau_h)}{d^{\pi_b}(s_h, \tau_h)}\right)^2\right] \sim \mathbb{E}_{\pi_b}\left[\left(\frac{d^{\pi_e}(s_h, \tau_h)}{d^{\pi_b}(s_h, \tau_h)}\right)^2\right] \tag{9}$$

Theorem 6. *Consider the L_∞ 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, we have*

$$\left\|\frac{d_\phi^{\pi_e^\phi}(s_h, \tau_{[h-T+1:h]})}{d_\phi^D(s_h, \tau_{[h-T+1:h]})}\right\|_\infty \leq \left\|\frac{d^{[\pi_e^\phi]_{\text{true}}}(s_h, \tau_h)}{d^{\pi_b}(s_h, \tau_h)}\right\|_\infty \sim \left\|\frac{d^{\pi_e}(s_h, \tau_h)}{d^{\pi_b}(s_h, \tau_h)}\right\|_\infty \tag{10}$$

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

Example 1. *In this example, we consider a belief space with a smoothness structure (Section 5.3 [13]) denoted as follow:*

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

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, namely, when the policy is strictly short-term memoried, then the result goes back to what's discussed in [25, 27].*

Proof of Theorem 5

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

$$\chi^2_{\mathcal{D}}(P^{\pi_e}\|P^{\pi_b}) = \sup_{g\in\mathcal{M}(\mathcal{D})} \mathbb{E}_{P^{\pi_e}}[g(\tau_h)] - \mathbb{E}_{P^{\pi_b}}[g(\tau_h)^2/4 + g(\tau_h)] \tag{11}$$

$$\chi^2_{\mathcal{A}}(P^{\pi_e}\|P^{\pi_b}) = \sup_{g\in\mathcal{M}(\mathcal{A})} \mathbb{E}_{P^{\pi_e}}[g(\tau_h)] - \mathbb{E}_{P^{\pi_b}}[g(\tau_h)^2/4 + g(\tau_h)] \tag{12}$$

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

$$\chi^2_{\mathcal{D}}(P^{\pi_e}\|P^{\pi_b}) \le \chi^2_{\mathcal{A}}(P^{\pi_e}\|P^{\pi_b}) \tag{13}$$

which proves the theorem. □

Proof of Theorem 6

Proof. Construct d^D exactly as in Theorem 5, 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 $\tilde{\phi}_T(\tau_h) = \tilde{\phi}_T(\tau^*)$, $w^\star(\tau_h) \le w^\star(\tau_h^\star)$. 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

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

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

D 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 a MDP with exponentially large state space. In this case, the belief metric is a discrete metric, for $\forall b_1, b_2 \in \mathcal{B}, b_1 \ne b_2 \to \|b_1 - b_2\|_1 = 2$, 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.

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.