Breaking the Order Barrier: Off-Policy Evaluation for Confounded POMDPs

Qi Kuang
School of Statistics and Data Science
Jiangxi University of Finance and Economics

Jiayi Wang
Department of Mathematical Sciences
University of Texas at Dallas

Fan Zhou *
School of Statistics and Data Science
MoE Key Laboratory of Interdisciplinary Research of Computation and Economics
Shanghai University of Finance and Economics

Zhengling Qi *
School of Business
George Washington University

Abstract

We consider off-policy evaluation (OPE) in Partially Observable Markov Decision Processes (POMDPs) with unobserved confounding. Recent advances have introduced bridge-function to circumvent unmeasured confounding and develop estimators for the policy value, yet the statistical error bounds of them related to the length of horizon T and the size of the state-action space $|\mathcal{O}||\mathcal{A}|$ remain largely unexplored. In this paper, we systematically investigate the finite-sample error bounds of OPE estimators in finite-horizon tabular confounded POMDPs. Specifically, we show that under certain rank conditions, the estimation error for policy value can achieve a rate of $\mathcal{O}(T^{1.5}/\sqrt{n})$, excluding the cardinality of the observation space $|\mathcal{O}|$ and the action space $|\mathcal{A}|$. With an additional mild condition on the concentrability coefficients in confounded POMDPs, the rate of estimation error can be improved to $\mathcal{O}(T/\sqrt{n})$. We also show that for a *fully history-dependent policy*, the estimation error scales as $\mathcal{O}\big(T/\sqrt{n}(|\mathcal{O}||\mathcal{A}|)^{\frac{T}{2}}\big)$, highlighting the exponential error dependence introduced by history-based proxies to infer hidden states. Furthermore, when the target policy is *memoryless policy*, the error bound improves to $\mathcal{O}\big(T/\sqrt{n}\sqrt{|\mathcal{O}||\mathcal{A}|}\big)$, which matches the optimal rate known for tabular MDPs. To the best of our knowledge, this is the first work to provide a comprehensive finite-sample analysis of OPE in confounded POMDPs.

1 Introduction

Partially Observable Markov Decision Processes (POMDPs) (Monahan, 1982) have become a practical framework for modeling decision-making under uncertainty across a wide range of applications (Albright, 1979; Monahan, 1982; Singh et al., 1994; Cassandra, 1998; Young et al., 2013; Bravo et al., 2019). In POMDPs, the agent must make decisions based on partial observations rather than full access to the underlying system state. Such partial observability often arises in real-world settings, making standard Markov Decision Processes (MDPs) inadequate for modeling the underlying data-generating processes. For example, in healthcare, clinical decision-making is frequently based on

*Corresponding authors: zhoufan@mail.shufe.edu.cn; qizhengling@email.gwu.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

partial information, while important latent factors like disease progression or genetic predispositions remain hidden or difficult to quantify.

A recent line of research has focused on off-policy evaluation (OPE) within the framework of confounded POMDPs (Tennenholtz et al., 2020; Nair and Jiang, 2021; Shi et al., 2022; Miao et al., 2022; Bennett and Kallus, 2024). Since confounded POMDPs inherently violate the Markov assumption and meanwhile encounter unmeasured confounders, existing approaches draw inspiration from proximal causal inference (Miao et al., 2018; Kallus et al., 2021; Cui et al., 2024), leveraging partial observations as a proxy to infer the hidden state, which enables the identification of the policy value. However, treating histories as proxy states poses fundamental hardness for POMDPs, as these histories can entail an exponentially large number of possibilities, thereby demanding a potentially exponential number of samples for accurate evaluation (Liu and Jin, 2022).

To avoid an explicit exponential dependence of the error on the cardinality of the observation space, recent work (Tennenholtz et al., 2020; Nair and Jiang, 2021) uses importance sampling (IS) but introduces an exponential-in-horizon quantity due to using the cumulative importance weights. To address the history-induced "curse of horizon"(Liu et al., 2018), recent work (Shi et al., 2022; Uehara et al., 2023; Zhang and Jiang, 2025a) employ function approximation under realizability assumptions and specific coverage assumptions. However, in tabular confounded POMDPs, despite the inevitable dependence on the observation and action sizes, the explicit dependence on the cardinality $|\mathcal{O}||\mathcal{A}|$ for history-dependent policies remains unspecified.

On the other hand, excluding the influence of cardinality, the optimal error rate achievable by OPE estimators in confounded POMDPs remains unclear. Recent work (Zhang and Jiang, 2025a,b) demonstrate an error rate of $\mathcal{O}(T^2/\sqrt{n})$, yet these pertain to *unconfounded* POMDPs, where the behavior policy is solely dependent on observed variables. In contrast, it has been shown that OPE estimators can achieve an error rate of $\mathcal{O}(T/\sqrt{n})$ in MDPs, which is known to be minimax optimal in both sample size and horizon length. For instance, Yin and Wang (2020) shows that marginal importance sampling (MIS) estimator can attain such error bounds for tabular setting, while Wang et al. (2024) further proves that nonparametric Fitted-Q Evaluation (FQE) estimators can achieve similar sharp dependence under the realizability assumption on the ratio functions. Given these insights, one naturally wonders if the successful analysis can be leveraged to better understand the horizon dependence for the convergence rate of the OPE estimator in the presence of unobserved confounding.

Specifically, drawing from the aforementioned literature, we seek to address three questions:

Q1: *How does the history-dependent policy explicitly influence the dependency of error bounds in confounded POMDPs?* **Q2:** *How does the error bound depend on the horizon T? Can the sharp linear rate, known for fully observable settings, be achieved in confounded POMDPs?* **Q3:** *What is the role of the concentrability coefficients (the ratio functions defined in Assumption 3) in improving the convergence rate of OPE estimators in confounded POMDPs?*

In this paper, we investigate the problem of off-policy evaluation under the confounded POMDPs, with a focus on addressing the three key questions. Compared to existing work on OPE in confounded POMDPs, our contributions are as follows: Firstly, we establish a two-step model-based approach to estimate the policy value, which relies on certain matrix invertibility conditions. We demonstrate the estimation error for fully history-dependent policy scale as $\mathcal{O}(T^{1.5}/\sqrt{n}(|\mathcal{O}||\mathcal{A}|)^{\frac{T}{2}})$, while the higher-order terms of error bound exhibit a stronger dependence on the size of observation-action space, scaling as $(|\mathcal{O}||\mathcal{A}|)^{\frac{3T}{2}}$. Secondly, excluding the influence of cardinality, as T grows, the first-order term of the error bound demonstrates a $T^{1.5}$ dependence, while the higher-order terms show a stronger T^3 dependence. Thirdly, by assuming the boundedness of certain concentrability coefficients in confounded POMDPs, without additionally estimating these probability ratio functions, we show that the first-order term of the error bound can be improved to $\mathcal{O}(T/\sqrt{n}(|\mathcal{O}||\mathcal{A}|)^{\frac{T}{2}})$. Lastly, when the target policy is reduced to a memoryless policy, the error bound improves to $\mathcal{O}(T/\sqrt{n}\sqrt{|\mathcal{O}||\mathcal{A}|})$, matching the sharpest rate of convergence known in the tabular MDPs setting.

2 Related Work

OPE in confounded POMDPs. A growing line of work has studied OPE in confounded POMDPs by leveraging proxy variables to identify policy value in the presence of unobserved confounders

(Zhang and Bareinboim, 2016; Shi et al., 2022; Lu et al., 2023; Hong et al., 2024b,a). While these methods have primarily focused on settings involving function approximation, theoretical analysis in tabular settings remains relatively limited. Besides, a line of research investigates the *unconfounded* setting (Uehara et al., 2023; Hu and Wager, 2023; Zhang and Jiang, 2025a,b). Among these, Zhang and Jiang (2025b) demonstrates that history-dependent policies can achieve polynomial sample complexity for OPE under certain coverage assumptions and further highlights the necessity of model-based approaches in this context. However, the unconfounded setting is less challenging than ours, as it fails to capture the complexities introduced by unobserved confounding. Crucially, none of the above works investigate whether optimal convergence rates can be achieved in confounded POMDPs.

OPE in MDPs. OPE in MDPs has been extensively studied, including importance sampling (IS) approaches (Precup, 2000) and their doubly robust variants (Dudík et al., 2011; Jiang and Li, 2016; Thomas and Brunskill, 2016), as well as marginalized importance sampling (MIS) methods (Xie et al., 2019; Kallus and Uehara, 2020; Yin and Wang, 2020). While IS-based estimators are broadly applicable, they suffer from high variance, leading to exponential dependence on the horizon length. Additionally, MIS-based estimators rely on the Markov property, which limits their direct application to confounded POMDPs. Yin and Wang (2020) demonstrate that MIS methods can achieve a rate of $\mathcal{O}(T/\sqrt{n})$. However, there has been no work investigating the potential for improving the dependence on horizon length for OPE in confounded POMDPs.

3 Preliminaries

POMDP Setup. We consider a finite-horizon episodic POMDP denoted by $\mathcal{M} := (\mathcal{S}, \mathcal{O}, \mathcal{A}, T, \nu_1, \{P_t\}_{t=1}^T, \{\mathcal{T}_t\}_{t=1}^T, \{r_t\}_{t=1}^T)$, where $\mathcal{S}$, $\mathcal{O}$ and $\mathcal{A}$ denote the state space, the observation space, and the action space respectively. In this paper, all $\mathcal{S}$, $\mathcal{O}$ and $\mathcal{A}$ are finite. The integer T is the total length of the horizon. We use $\nu_1 \in \Delta(\mathcal{S})$ to denote the distribution of the initial state, where $\Delta(\Omega)$ is a class of all probability distributions over the space Ω. Denote $\{P_t\}_{t=1}^T$ to be the collection of state transition kernels over $\mathcal{S} \times \mathcal{A}$ to $\mathcal{S}$, and $\{\mathcal{T}_t\}_{t=1}^T$ to be the collection of observation emission kernels over $\mathcal{S}$ to $\mathcal{O}$. We use $\{r_t\}_{t=1}^T$ denote the collection of reward functions, i.e., $r_t : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ at each time step t. Finally, we let O_0 denote the prior observation, which provides prior information about the initial state S_1.

In a standard POMDP, at each time step t, given the current (hidden) state S_t, an observation $O_t \sim \mathcal{T}_t(\cdot \mid S_t)$ is observed. The agent then selects an action A_t according to a certain policy, receives a reward R_t with $\mathbb{E}[R_t \mid S_t = s_t, A_t = a_t] = r_t(s_t, a_t)$ for every (s_t, a_t), and the environment transitions to the next state S_{t+1} according to $P_t(\cdot \mid S_t, A_t)$.

Off-policy Evaluation (OPE) under Unmeasured Confounding. This paper aims to estimate the value of a potentially history-dependent policy for POMDPs using offline data. Define the observed history by $H_t := (O_1, A_1, ..., O_t, A_t) \in \mathcal{H}_t$, where $\mathcal{H}_t := \prod_{j=1}^t (\mathcal{O} \times \mathcal{A})$ is the space of observable history up to time t. The *target policy* to be evaluated is denoted by $\{\pi_t\}_{t=1}^T$, where $\pi_t : \mathcal{O} \times \mathcal{H}_{t-1} \to \Delta(\mathcal{A})$ is *history-dependent*, illustrated by the green arrows in Figure 1. Given the target policy $\pi := \{\pi_t\}_{t=1}^T$, the policy value is defined as

$$\mathcal{V}(\pi) := \mathbb{E}_{S_1 \sim \nu_1}\Big[\mathbb{E}^\pi\Big[\sum_{t=1}^T R_t \mid S_1\Big]\Big],$$

where $\mathbb{E}^\pi$ is taken with respect to the distribution induced by the policy π.

In the offline setting, an agent cannot interact with the environment but only has access to a pre-collected dataset generated by some *behavior policy* $\{\pi_t^b\}_{t=1}^T$. We assume that the behavior policy depends on the unobserved state S_t, i.e., $\pi_t^b : \mathcal{S} \to \Delta(\mathcal{A})$ for each t. Unlike previous work (Uehara et al., 2023; Zhang and Jiang, 2025a,b), which restricts the behavior policy to be *memoryless* and dependent only on the current observation (i.e., $a_t \sim \pi_t^b(\cdot|o_t)$), our approach allows the behavior policy to depend on the latent state, making it more aligned with real-world scenarios and thus introducing additional complexity in the analysis. We use $\mathcal{P}$ to denote the offline data distribution under the behavior policy and summarize the data as $\mathcal{D} := \{o_0^i, (o_t^i, a_t^i, r_t^i)_{t=1}^T\}_{i=1}^n$, which consists of n i.i.d. samples drawn from $\mathcal{P}$. The full data generating process in so-called confounded POMDPs is depicted in Figure 1.

Notations. Throughout this paper, we assume that $\mathbb{E}$ is taken with respect to the offline distribution. We use uppercase letters such as $(S_t, O_t, A_t, R_t, H_t)$ to denote random variables and lowercase letters such as $(s_t, o_t, a_t, r_t, h_t)$ to denote their realizations, unless stated otherwise. We use the notation $X \perp\!\!\!\perp Y \mid Z$ when X and Y are conditionally independent given Z under the offline distribution. For random variables X, Y and Z taking values on $\{x_1, \ldots, x_m\}$, $\{y_1, \ldots, y_n\}$, and $\{z_1, \ldots, z_q\}$, respectively, $\mathbb{P}(X)$ denotes a n length vector with entries $\mathbb{P}_i(X) := p(X = x_i)$ and $\mathbb{P}(X \mid Y)$ denotes a $n \times m$ matrix with entries $\mathbb{P}_{i,j}(X \mid Y) := p(X = x_j \mid Y = y_i)$. Similarly, $\mathbb{P}(X = x \mid Y)$ denotes the n length vector with entries $\mathbb{P}_i(X \mid Y) := p(X = x \mid Y = y_i)$, and $\mathbb{P}(X, Z = z \mid Y)$ denotes a $n \times m$ matrix with entries $\mathbb{P}_{i,j}(X, Z = z \mid Y) := p(X = x_j, Z = z \mid Y = y_i)$. For brevity, we sometimes abbreviate these as $\mathbb{P}(x \mid Y)$ and $\mathbb{P}(X, z \mid Y)$. For any two sequences $\{a_n\}_{n=1}^{\infty}$, $\{b_n\}_{n=1}^{\infty}$, $a_n \lesssim b_n$ denotes $a_n \leq C b_n$ for some $N, C > 0$ and every $n > N$. For a matrix A, we denote its Moore-Penrose inverse by $A^{\dagger}$ and its smallest singular value by $\sigma_{min}(A)$.

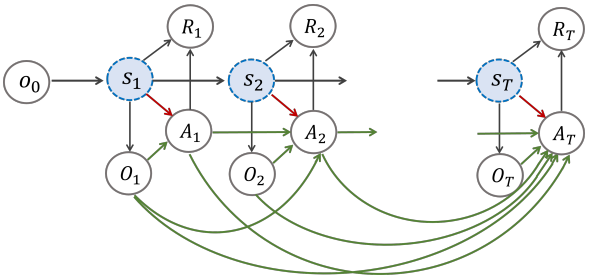

Figure 1: The directed acyclic graph of the data-generating process in confounded POMDPs, where states S_t are not observed. Red arrows indicate the generation of actions via the behavior policy, while green arrows indicate the generation through a target history-dependent policy. Under the offline distribution $\mathcal{P}$, we have the conditional independency $O_0 \perp\!\!\!\perp (O_t, O_{t+1}, R_t) \mid S_t, A_t, H_{t-1}$ for any $t \in [T]$.

4 Methods

In this section, we introduce the proposed off-policy evaluation method for confounded POMDPs. Before considering how to estimate $\mathcal{V}(\pi)$, we first consider the problem of identification.

4.1 Identification

In general, identification is not possible for OPE in the presence of unobserved confounders. The fundamental challenge arises from two sources: (i) **partial observability**, which breaks the conditional independence (Markov property) essential for classical OPE estimators; and (ii) **unmeasured confounding**, which renders direct estimation of the policy value $\mathcal{V}(\pi)$ intractable. Crucially, failure to account for such confounding leads to biased estimates of the policy value (Shi et al., 2022), as the hidden state simultaneously influences both the action and future rewards and transitions.

To address these issues, a natural strategy is to use the observed history to infer information about the hidden state (Shi et al., 2022; Hong et al., 2024a,b). This approach is motivated by the insight that the entire history contains rich information that can help reconstruct a proxy for the unobserved state. To enable the observed history for identifying the policy value, we impose rank conditions.

Assumption 1 (Invertibility). *For each $t \in [T]$ and $a_t \in \mathcal{A}$, assume $\mathbb{P}(O_t \mid A_t = a_t, S_t) \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{O}|}$ has full row rank and $\mathbb{P}(O_t \mid A_t = a_t, O_0, H_{t-1}) \in \mathbb{R}^{|\mathcal{O}||\mathcal{H}_{t-1}| \times |\mathcal{O}|}$ has full column rank.*

A necessary condition for the rank condition is that $|\mathcal{O}| > |\mathcal{S}|$. Invertibility of $\mathbb{P}(O_t \mid A_t = a_t, S_t)$ guaranties that sufficient information about the hidden states is encoded in the observations across time steps. The invertibility of $\mathbb{P}(O_t \mid A_t = a_t, O_0, H_{t-1})$ further ensures the recovery of information about the hidden states S_t from the observable history $\{O_0, H_{t-1}\}$. Notably, leveraging the entire history as a proxy relaxes the standard rank condition that requires $\mathbb{P}(O_t \mid A_t = a_t, O_{t-1})$, using one-step observation as a proxy, to be invertible (Tennenholtz et al., 2020; Shi et al., 2022). This is particularly beneficial in real-world applications where the observation space is often larger than the hidden state space.

Theorem 1. *Under Assumptions 1, $\mathcal{V}(\pi)$ is identified as*

$$\begin{aligned}\mathcal{V}(\pi) = &\sum_{t=1}^{T}\sum_{r_t}\sum_{h_t} r_t\Big(\prod_{i=1}^{t}\pi_i(a_i \mid o_i, h_{i-1})\Big)\mathbb{P}(O_0)\mathbb{P}(O_1|O_0)\mathbb{P}^{\dagger}(O_1 \mid A_1 = a_1, O_0)\\ &\Big(\prod_{i=1}^{t-1}\mathbb{P}(O_{i+1}, O_i = o_i \mid A_i = a_i, O_0, H_{i-1})\mathbb{P}^{\dagger}(O_{i+1} \mid A_{i+1} = a_{i+1}, O_0, H_i)\Big)\\ &\mathbb{P}(R_t = r_t, O_t = o_t \mid A_t = a_t, O_0, H_{t-1}).\end{aligned}$$

Theorem 1 states that the policy value can be expressed entirely in terms of observable variables. This expression is derived by decomposing the marginal distribution $p^{\pi}(r_t)$ using the transition dynamics, reward, and target policy, based on the identity $\mathcal{V}(\pi) = \mathbb{E}^{\pi}[\sum_{t=1}^{T} R_t] = \sum_{t=1}^{T}\sum_{r_t} r_t p^{\pi}(r_t)$. Similar results were initially introduced by Tennenholtz et al. (2020). Specifically, the invertibility of the action-conditioned probability matrices $\mathbb{P}(O_{t+1} \mid A_t = a_t, O_0, H_t)$ allows us to algebraically reconstruct the distribution of observations and rewards under the target policy using offline data collected by the behavior policy. Intuitively, conditioning on the action blocks the confounding path through the unobserved states, isolating the confounding effect of the action. Consequently, these action-conditioned probabilities, as proxy functions, can effectively correct the confounding influence, thus enabling us to identify the policy value. In addition, we extend the identification results beyond the tabular setting (see Theorem 5 in the Appendix), showing that the completeness condition (Assumption 5 in the Appendix), which is a generalization of the rank conditions in Assumption 1, is sufficient for the identification theorem to hold in more general settings.

4.2 Estimation via Value Functions

According to Theorem 1, we can directly perform OPE by estimating those conditional matrices relying solely on the offline data. However, the presence of multiple matrix inverses in the product is computationally expensive and may lead to instability in the estimation. To address this, we are motivated by the structure of the identification in Theorem 1 and propose to solve a Bellman-type equation to estimate the policy value as a more stable and computationally efficient approach.

Bellman-like recursions in confounded POMDPs. Under the invertibility assumptions stated in Assumption 1, the proxy $\{O_0, H_{t-1}\}$ is sufficient to construct a Bellman-like recursion based entirely on observable variables (see Appendix B.1 for further details). Specifically, we can formulate a system of linear integral equations (1), whose solution defines a sequence of proxy value functions $\{b^{\pi}_{V,t} : \mathcal{A}\times\mathcal{O}\times\mathcal{H}_{t-1}\to\mathbb{R}\}_{t=1}^{T}$.

$$\begin{aligned}&\mathbb{E}\Big[b^{\pi}_{V,t}(A_t, O_t, H_{t-1}) \,\Big|\, O_0, A_t, H_{t-1}\Big]\\ &= \mathbb{E}\Big[R_t\pi_t(A_t \mid O_t, H_{t-1}) + \sum_{a'\in\mathcal{A}} b^{\pi}_{V,t+1}(a', O_{t+1}, H_t)\pi_t(A_t \mid O_t, H_{t-1}) \,\Big|\, O_0, A_t, H_{t-1}\Big],\end{aligned} \tag{1}$$

where $b^{\pi}_{V,T+1} \equiv 0$. Note that the invertibility of $\mathbb{P}(O_t \mid A_t = a_t, O_0, H_{t-1})$ ensures that the value functions $\{b^{\pi}_{V,t}\}_{t=1}^{T}$ are uniquely defined solutions to this system. In the context of confounded POMDPs, these proxy functions play an analogous role to the value functions in MDPs, enabling the estimation of policy value through these functions. This naturally suggests a two-step procedure for OPE in confounded POMDPs: (i) estimate the conditional probabilities, and (ii) compute the value function from (1) using the estimated probabilities.

To illustrate the model-based estimation in the tabular setting, we first define two key conditional probability matrices for any given $a_t \in \mathcal{A}$, r_t, and $o_{t+1} \in \mathcal{O}$,

$$\begin{aligned}\mathbf{P}_{a_t} :=& \mathbb{P}(O_t \mid O_0, A_t = a_t, H_{t-1}) \in \mathbb{R}^{|\mathcal{O}||\mathcal{H}_{t-1}|\times|\mathcal{O}|},\\ \mathbf{P}_{a_t, r_t, o_{t+1}} :=& \mathbb{P}(O_{t+1} = o_{t+1}, R_t = r_t, O_t \mid O_0, A_t = a_t, H_{t-1}) \in \mathbb{R}^{|\mathcal{O}||\mathcal{H}_{t-1}|\times|\mathcal{O}|}.\end{aligned}$$

Then, we can rewrite the value recursion from equation (1) in matrix form as:

$$\mathbf{P}_{a_t}\mathbf{B}_t = \sum_{r_t, o_{t+1}} \mathbf{P}_{a_t, r_t, o_{t+1}}\big[r_t \cdot \pi_t(a_t|\mathbf{o}_t, h_{t-1}) + \sum_{a'\in\mathcal{A}} b^{\pi}_{V,t+1}(a', o_{t+1}, h_{t-1}, a_t, \mathbf{o}_t)\odot\pi_t\big], \tag{2}$$

where $\mathbf{B}_t := b^{\pi}_{V,t}(a_t, \mathbf{o}_t, h_{t-1}) \in \mathbb{R}^{|\mathcal{O}|}$, $\mathbf{P}_{a_t}, \mathbf{P}_{a_t,r_t,o_{t+1}} \in \mathbb{R}^{|\mathcal{O}||\mathcal{H}_{t-1}|\times|\mathcal{O}|}$ are conditional probability matrices defined earlier, $\odot$ denotes the element-wise product, and $r_t\pi_t(a_t|\mathbf{o}_t, h_{t-1}) + \pi_t(a_t|\mathbf{o}_t, h_{t-1}) \odot \sum_{a'} b^{\pi}_{V,t+1}(a', o_{t+1}, h_{t-1}, a_t, \mathbf{o}_t) \in \mathbb{R}^{|\mathcal{O}|}$ is a $|\mathcal{O}|$ length vector. To ensure a unique solution of $\mathbf{B}_t$, the matrix $\mathbf{P}_{a_t}$ must be full column rank, i.e. $\mathrm{rank}(\mathbf{P}_{a_t}) = |\mathcal{O}|$. This is a mild and typically reasonable assumption in practice, since $|\mathcal{O}| \ll |\mathcal{O}||\mathcal{H}_{t-1}|$ as t increases. Then, the solution to the linear system (2) is given by

$$\mathbf{B}_t = \mathbf{P}^{\dagger}_{a_t} \sum_{r_t,o_{t+1}} \mathbf{P}_{a_t,r_t,o_{t+1}} \big[r_t \cdot \pi_t(a_t|\mathbf{o}_t, h_{t-1}) + \sum_{a'\in\mathcal{A}} b^{\pi}_{V,t+1}(a', o_{t+1}, h_{t-1}, a_t, \mathbf{o}_t) \odot \pi_t\big],$$

where $\mathbf{P}^{\dagger}_{a_t} = \big(\mathbf{P}^{\top}_{a_t}\mathbf{P}_{a_t}\big)^{-1}\mathbf{P}^{\top}_{a_t} \in \mathbb{R}^{|\mathcal{O}|\times|\mathcal{O}||\mathcal{H}_{t-1}|}$ is the Moore–Penrose inverse of $\mathbf{P}_{a_t}$. Given $b^{\pi}_{V,T+1} \equiv 0$, the value functions $\{b^{\pi}_{V,t}\}_{t=1}^{T}$ can be solved iteratively, starting from $t = T$ and proceeding backward. Specifically, for any $a_t \in \mathcal{A}, o_t \in \mathcal{O}, h_{t-1} \in \mathcal{H}_t$, the update rule gives

$$b^{\pi}_{V,t}(a_t, o_t, h_{t-1}) = \pi_t(a_t|o_t, h_{t-1})\psi_t^{\top}\mathbf{P}^{\dagger}_{a_t} \sum_{r_t,o'} \mathbf{P}_{a_t,r_t,o_{t+1}}\psi_t\big(r_t + \sum_{a'\in\mathcal{A}} b^{\pi}_{V,t+1}(a', o', h_t)\big), \quad (3)$$

where $\psi_t(o_t) \in \mathbb{R}^{|\mathcal{O}|}$ is a one-hot encoding vector for the observation o_t.

Two-stage estimation. To enable the estimation of value functions, we estimate the conditional probability matrices $\widehat{\mathbf{P}}_{a_t}, \widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}}$ from the dataset $\mathcal{D}$ at time step t. Each matrix is constructed entry-wise as follows:

$$\begin{aligned}
\widehat{p}(o_t \mid o_0, a_t, h_{t-1}) &= \frac{\sum_{i=1}^{n} \mathbb{1}\big\{(o_t^i, h_{t-1}^i, o_0^i, a_t^i) = (o_t, h_{t-1}, o_0, a_t)\big\}}{n_{o_0,h_{t-1},a_t}}, \\
\widehat{p}(o_{t+1}, o_t, r_t \mid o_0, a_t, h_{t-1}) &= \frac{\sum_{i=1}^{n} \mathbb{1}\big\{(o_{t+1}^i, o_t^i, r_t^i, o_0^i, h_{t-1}^i, a_t^i) = (o_{t+1}, o_t, r_t, o_0, h_{t-1}, a_t)\big\}}{n_{o_0,h_{t-1},a_t}},
\end{aligned} \quad (4)$$

where $n_{o_0,h_{t-1},a_t} = \sum_{i=1}^{n} \mathbb{1}\{(o_0^i, h_{t-1}^i, a_t^i) = (o_0, h_{t-1}, a_t)\}$ denotes the number of the triplet (o_0, h_{t-1}, a_t) being visited among n independent episodes.

Then, the sequence of estimated value functions $\{\widehat{b}_{V,t}\}_{t=1}^{T}$ can be solved iteratively by

$$\widehat{b}_{V,t}(a_t, o_t, h_{t-1}) = \pi_t(a_t|o_t, h_{t-1})\psi_t^{\top}\widehat{\mathbf{P}}^{\dagger}_{a_t} \sum_{r_t,o'} \widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}}\psi_t\big(r_t + \sum_{a'\in\mathcal{A}} \widehat{b}_{V,t+1}(a', o', h_t)\big), \quad (5)$$

where $\widehat{b}_{V,T+1} = 0$, $\widehat{\mathbf{P}}^{\dagger}_{a_t} = \big(\widehat{\mathbf{P}}^{\top}_{a_t}\widehat{\mathbf{P}}_{a_t}\big)^{-1}\widehat{\mathbf{P}}^{\top}_{a_t} \in \mathbb{R}^{|\mathcal{O}|\times|\mathcal{O}||\mathcal{H}_{t-1}|}$. By Lemma 12, these empirical estimators are unbiased, i.e. $\mathbb{E}[\widehat{p}(o_t \mid o_0, a_t, h_{t-1})] = p(o_t \mid o_0, a_t, h_{t-1})$, which implies that $\mathbb{E}[\widehat{\mathbf{P}}_{a_t}] = \mathbf{P}_{a_t}$. Consequently, the recursion in (5) ultimately yields asymptotically unbiased estimates of the true value functions. Moreover, the $\widehat{\mathbf{P}}_{a_t}$ must be invertible for the recursion to proceed, which implicitly requires that each tuple (o_0, h_{t-1}, a_t) must be observed sufficiently many times. To ensure numerical stability and avoid division by zero, we assume each (o_0, h_{t-1}, a_t) in the data is sufficiently sampled. If the triple (o_0, h_{t-1}, a_t) is not collected in the offline data, we set $\widehat{p}(\cdot|o_0, a_t, h_{t-1}) = 0$.

Policy value estimation. After computing the value functions via (5) for T iterations, we plug the estimated value function $\widehat{b}_{V,1}$ into $\mathcal{V}(\pi)$ to obtain the empirical estimator as

$$\widehat{\mathcal{V}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\Big[\sum_{a\in\mathcal{A}} \widehat{b}_{V,1}(a, o_1^i)\Big].$$

Algorithm 1 summarizes the proposed algorithm for OPE in confounded POMDPs.

5 Theoretical Results

In this section, we study the theoretical properties of our method under certain technical assumptions. Our primary goal is to establish a finite-sample error upper bound for $\mathcal{V}(\pi) - \widehat{\mathcal{V}}(\pi)$. Specifically, this upper bound will depend on several factors, including the sample size n, the horizon length T, and the size of the observation space $|\mathcal{O}|$ and action space $|\mathcal{A}|$. To begin with, we impose the following key assumptions that are used in the theoretical analysis.

Algorithm 1 Tabular Off-Policy Evaluation for Confounded POMDPs

Input: Dataset $\mathcal{D}$, the target policy $\{\pi_t\}_{t=1}^T$, and initialize $\widehat{b}_{V,T+1} = 0$.
for $t = T, \ldots, 1$ **do**
  Estimation of conditional probability: obtain $\widehat{\mathbf{P}}_{a_t}$ and $\widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}}$ by (4)
  Estimation of value functions: obtain $\widehat{b}_{V,t}$ by (5)
end for
Output: obtain estimated policy value $\widehat{\mathcal{V}}(\pi)$ by $\widehat{b}_{V,1}$.

Assumption 2. *The following conditions hold.*

(a) (Coverage) For each $t \in [T]$, $\mathbb{E}\Big[\prod_{t'=1}^{t}\Big(\frac{\pi_{t'}(A_{t'}|O_{t'},H_{t'-1})}{\pi^b_{t'}(A_{t'}|S_{t'})}\Big)^2\Big] \le C_{\pi^b} < \infty$*;*

(b) (Invertibility) For each $t \in [T]$, $a_t \in \mathcal{A}$, $\mathrm{rank}(\mathbf{P}_{a_t}) = \mathrm{rank}(\widehat{\mathbf{P}}_{a_t}) = |\mathcal{O}|$*, and the smallest singular value* $\sigma_{min}(\mathbf{P}_{a_t})$ *satisfies* $\sigma_{min}(\mathbf{P}_{a_t}) \ge \frac{C_P^{-1}}{\sqrt{|\mathcal{O}||\mathcal{H}_{t-1}|}}$*, where* $0 < C_P < \infty$ *is a constant;*

(c)(Sufficient visitation of observations) For each $t \in [T]$*, the samples used to construct each entry of* $\widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}}$ *satisfy* $n_{o_0,h_{t-1},a_t} \ge np_t^{\pi_b}(o_0,h_{t-1},a_t)(1-\theta_{t,ij})$*, where* $0 < \theta_{t,ij} < 1$ *with* $\sum_{i,j}\theta_{t,ij} = 1$*, and we denote* $\theta^* := \min_{t,i,j}\theta_{t,ij}$.

Assumption $2(a)$ imposes a bounded second moment condition on the cumulative importance ratio (concentrability coefficients), which is milder compared to directly bounding the importance weight $\pi_t(a_t|o_t,h_{t-1})/\pi_t^b(a_t|s_t)$, a common assumption in the OPE literature. The invertibility condition (b) requires $\mathbf{P}_{a_t}$ and $\widehat{\mathbf{P}}_{a_t}$ to be well-conditioned matrices, which is necessary for the uniqueness of the solution to the iteration equations (3) and (5). Furthermore, we require the lower bound of the smallest singular value of $\mathbf{P}_{a_t}$ to decay in proportion to $1/\sqrt{|\mathcal{O}||\mathcal{H}_{t-1}|}$. This is informed by random matrix theory, where for an $M \times N$ random matrix (M fixed, $N \to \infty$), the smallest singular value decays at a rate of $\mathcal{O}(1/\sqrt{N})$ (Rudelson and Vershynin, 2009). Assumption $2(c)$ requires a sufficient number of samples for each triple (o_0, h_{t-1}, a_t), ensuring consistent estimation of the conditional probability matrices. Specifically, this requires the sample size $n \ge \frac{\mathrm{polylog}(|\mathcal{O}|^T,|\mathcal{A}|^T,T)}{\min_{t,o_0,h_{t-1},a_t} p_t^{\pi^b}(o_0,h_{t-1},a_t)}$, and further details can be found in Appendix C. Assumption 2(c) is introduced to simplify the proof and can be relaxed using a truncation argument, similar to the approach in Yin and Wang (2020).

We now present the main theorem that provides the upper bound for $|\mathcal{V}(\pi) - \widehat{\mathcal{V}}(\pi)|$.

Theorem 2. *Under Assumptions 1 and 2. Then, with probability at least* $1-\delta$*, it holds that*

$$\begin{aligned}\big|\mathcal{V}(\pi) - \widehat{\mathcal{V}}(\pi)\big| \lesssim & \frac{T^{1.5}}{\sqrt{n}}(1-\theta^*)^{-\frac{1}{2}}C_{\pi^b}^{\frac{1}{2}}C_P|\mathcal{O}|^{\frac{T}{2}}|\mathcal{A}|^{\frac{T}{2}} \\ & + \frac{T^{1.5}}{n}(1-\theta^*)^{-1}C_{\pi^b}^{\frac{1}{2}}C_P^2|\mathcal{O}|^{\frac{3T}{2}}|\mathcal{A}|^{\frac{3T}{2}}\sqrt{\log(T^2|\mathcal{O}|^T|\mathcal{A}|^T/\delta)} \\ & + \frac{T^3}{n^{\frac{3}{2}}}(1-\theta^*)^{-\frac{3}{2}}C_{\pi^b}^{\frac{1}{2}}C_P^3|\mathcal{O}|^{\frac{5T}{2}}|\mathcal{A}|^{\frac{5T}{2}}\log(T^2|\mathcal{O}|^T|\mathcal{A}|^T/\delta).\end{aligned} \tag{6}$$

In Theorem 2, the first term in (6) scales as $\mathcal{O}(T^{1.5}/\sqrt{n})$, which we refer to as the first-order term. The remaining terms, which are higher-order terms, exhibit a stronger dependence on T but converge more quickly due to the faster rates of the sample size n. These results respond to **Q2**. For clarity, we omit the detailed form of the higher-order terms here, and the full specifications are discussed in Appendix A. Compared to the order of $\mathcal{O}(T^2/\sqrt{n})$ obtained in both unconfounded POMDPs (Uehara et al., 2023; Zhang and Jiang, 2025a,b) and confounded POMDPs (Bennett and Kallus, 2024), we achieve a sharper dependence on the horizon by leveraging the fact that the variance of the first-order term can be decomposed into a sum of T individual expectations of the conditional variance. Moreover, the first-order term in the upper bound (6) exhibits an exponential dependence of order $(|\mathcal{O}||\mathcal{A}|)^{\frac{T}{2}}$ on the observation and action space, while the higher-order term shows a strong dependence of order $(|\mathcal{O}||\mathcal{A}|)^{\frac{\beta T}{2}}$ with $\beta \ge 3$. These results respond to **Q1**. The increased complexity arises from the fully history-dependent policy, leading to challenges in evaluating policies as the effective policy domain expands over time with the growth of the history space $|\mathcal{H}_t|$. This complexity highlights the inherent statistical challenges of evaluating fully history-dependent policies in confounded POMDPs.

Corollary 1. *Under the conditions in Theorem 2, for the memoryless policy dependent on the current observation, i.e.* $\pi : \mathcal{O} \to \Delta(\mathcal{A})$*, with high probability, it holds that*

$$\big|\mathcal{V}(\pi) - \widehat{\mathcal{V}}(\pi)\big| = \mathcal{O}\Big(\frac{T^{1.5}}{\sqrt{n}}(1-\theta^*)^{-\frac{1}{2}} C_{\pi^b}^{\frac{1}{2}} C_P |\mathcal{O}|^{\frac{1}{2}} |\mathcal{A}|^{\frac{1}{2}}\Big). \tag{7}$$

As shown in Corollary 1, when the target policy reduces to a memoryless policy, the exponential dependence on the observation and action spaces is no longer present. The upper bound in (7) instead exhibits a polynomial dependence on the horizon T, as well as on the cardinalities of the observation-action space. We now turn to a refined analysis of the upper bound in (6). Before proceeding, we assume the existence of the following sequence of ratio functions.

Assumption 3. *We assume the existence of real-valued functions* $\{b_{W,t}^{\pi} : \mathcal{A} \times \mathcal{O} \times \mathcal{H}_{t-1} \to \mathbb{R}\}_{t=1}^{T}$ *that satisfy the following conditional moment restrictions,*

$$\mathbb{E}\Big[b_{W,t}^{\pi}(A_t, H_{t-1}, O_0) \;\Big|\; S_t, A_t, H_{t-1}\Big] = \frac{\omega_t(S_t, H_{t-1})}{\pi_t^b(A_t \mid S_t)}, \tag{8}$$

where $\omega_t(S_t, H_{t-1}) = p_t^{\pi}(S_t, H_{t-1}) / p_t^{\pi^b}(S_t, H_{t-1})$*. We denote* $\rho_t(o_t, s_t, a_t, h_{t-1}) := \mathbb{E}\big[b_{W,t}^{\pi}(a_t, h_{t-1}, O_0)\pi_t(a_t \mid o_t, h_{t-1}) | o_t, s_t, a_t, h_{t-1}\big]$*. We further assume there exists a constant* $0 < C_W < \infty$ *such that*

$$\sup_{t, o_t, s_t, a_t, h_{t-1}} \rho_t(o_t, s_t, a_t, h_{t-1}) \le C_W. \tag{9}$$

These types of weight functions $\{b_{W,t}^{\pi}\}_{t=1}^{T}$ are also utilized in OPE for confounded POMDPs (Shi et al., 2022; Bennett and Kallus, 2024), which allows for the adjustment of the observed data distribution to account for the influence of unobserved confounders. The ratio ρ_t plays the role of a concentrability coefficient in confounded POMDPs. Specifically, the following identity holds:

$$\rho_t(o_t, s_t, a_t, h_{t-1}) = \frac{p_t^{\pi}(o_t, s_t, a_t, h_{t-1})}{p_t^{\pi^b}(o_t, s_t, a_t, h_{t-1})} = \frac{p_t^{\pi}(s_t, h_{t-1})\pi_t(a_t \mid o_t, h_{t-1})p(o_t \mid s_t)}{p_t^{\pi^b}(s_t, h_{t-1})\pi_t^b(a_t \mid s_t)p(o_t \mid s_t)}.$$

In the MDPs, ρ_t reduces to the ratio of the state-action marginal density, which quantifies the mismatch between the marginal state-action distributions of the target and behavior policies. Analogous to the fully observable case, we impose the bounded ratio in (9), which ensures that the offline data distribution $\mathcal{P}$ can calibrate the distribution induced by the target policy π.

Lemma 1. *For the target policy* $\pi : \mathcal{O} \times \mathcal{H}_{t-1} \to \Delta(\mathcal{A})$*, we have*

$$\sum_{t=1}^{T} \mathbb{E}^{\pi}\Big[\mathit{Var}\Big[R_t + \sum_{a \in \mathcal{A}} b_{V,t+1}^{\pi}(a, O_{t+1}, H_t) \;\Big|\; O_t, S_t, A_t, H_{t-1}\Big]\Big] \le \mathit{Var}^{\pi}\Big[\sum_{t=1}^{T} R_t\Big],$$

where Var^{π} *is taken with respect to the distribution induced by* π*.*

Lemma 1 establishes a variance decomposition bound that is analogous to Lemma 3.4 in Yin and Wang (2020), which focuses on the OPE in MDPs. Notably, it shows that the sum of the conditional variances across time steps is upper bounded by the variance of the cumulative reward, which is on the order of T^2, since $\mathrm{Var}^{\pi}[\sum_{t=1}^{T} R_t] \lesssim T^2$. This result can be derived by iteratively applying the law of total variance. The complete derivation is provided in Appendix C.1.2.

Theorem 3. *Under Assumptions 1, 2, and 3, we have the following results:*

(a) *for history-dependent policy, i.e.* $\pi_t : \mathcal{O} \times \mathcal{H}_{t-1} \to \Delta(\mathcal{A})$*, with high probability it holds that*

$$\big|\mathcal{V}(\pi) - \widehat{\mathcal{V}}(\pi)\big| = \mathcal{O}\Big(\frac{T}{\sqrt{n}}(1-\theta^*)^{-\frac{1}{2}} C_W^{\frac{1}{2}} C_P |\mathcal{O}|^{\frac{T}{2}} |\mathcal{A}|^{\frac{T}{2}}\Big); \tag{10}$$

(b) *for memoryless policy, i.e.* $\pi_t : \mathcal{O} \to \Delta(\mathcal{A})$*, with high probability it holds that*

$$\big|\mathcal{V}(\pi) - \widehat{\mathcal{V}}(\pi)\big| = \mathcal{O}\Big(\frac{T}{\sqrt{n}}(1-\theta^*)^{-\frac{1}{2}} C_W^{\frac{1}{2}} C_P |\mathcal{O}|^{\frac{1}{2}} |\mathcal{A}|^{\frac{1}{2}}\Big). \tag{11}$$

Based on Lemma 1, we derive Theorem 3. We omit the higher-order terms here, as they remain identical to those in Theorem 2. Notably, the constant of the coverage condition C_{π^b} is replaced by the constant of the bounded ratio C_W. Compared to the results in Theorem 2, the upper bound in (10) demonstrates an improvement in the dependence on the horizon length T in the first-order term, reducing it from $T^{1.5}$ to T, which is the sharpest known dependence of the horizon for tabular POMDPs. The bounded ratio function in Assumption 3 here provides an enhanced method for deriving the bound on the first-order term, thereby achieving faster convergence. These results answer the question **Q3**. In the memoryless case, the upper bound in (11) matches the optimal order of $\mathcal{O}(T/\sqrt{n}\sqrt{|\mathcal{O}||\mathcal{A}|})$ established for tabular MDPs (Yin and Wang, 2020). This finding highlights that a linear sample complexity in T is sufficient for evaluating the memoryless policy in confounded POMDPs under these conditions. In addition, it is crucial to emphasize that our approach does not require the estimation of weight functions $\{b^{\pi}_{W,t}\}_{t=1}^{T}$, focusing solely on the estimation of value functions $\{b^{\pi}_{V,t}\}_{t=1}^{T}$. This distinguishes our approach from prior work such as Shi et al. (2022), which necessitates the estimation of both components.

6 Simulation Results

We conduct a simulation study to examine the behavior of the error $|\mathcal{V}(\pi) - \widehat{\mathcal{V}}(\pi)|$ with respect to sample size n and horizon T. The primary objective is to provide empirical validation for our theoretical results. To this end, we use a relatively simple simulation setup that ensures clarity in demonstration. Specifically, we evaluate our approach in a simulated POMDP environment characterized by a one-dimensional discrete state/observation space, a discrete reward space, and binary actions. Concretely, we set $\mathcal{A} = \{0, 1\}, \mathcal{S} = \mathcal{O} = \{0, 1, 2\}$ for all t. The initial observation is given by $O_0 \sim \mathrm{Unif}(\{0, 1, 2\})$ and $S_t \sim \mathrm{Unif}(\{0, 1, 2\})$, and the transition dynamic is given by $O_t \sim P_t(\cdot|S_t)$, where $P_t(O_t|S_t) = \mathbb{1}\{O_t = S_t\}(1 - 3\epsilon/2) + \epsilon/2$. The immediate reward is set to be $R_t = 2/\{1 + \exp(-2S_tA_t - 3) - 1\}$. We collected offline data using a time-homogeneous behavioral policy $\pi^b_t(1|S_t) = 1/\{1 + \exp(-0.6S_t + 1)\} = 1 - \pi^b_t(0|S_t)$. For experimental details, we set $\epsilon = 0.2$, initialize $\widehat{b}_{V,T+1} = 0$, estimate conditional probability matrices as described in (4), and iteratively compute the value function $\widehat{b}_{V,t}$ over T steps according to (5).

We evaluate two target policies. (1) For the memoryless target policy $\pi_t(1|O_t) = 1/\{1 + \exp(-0.8O_t + 1)\} = 1 - \pi_t(0|O_t)$, the conditional probability $\mathbf{P}_{a_t}$ and $\mathbf{P}_{a_t,r_t,o_{t+1}}$ are conditioned on O_0. We evaluate its value using sample sizes $n = 200, 400, \ldots, 1000$, and horizon lengths $T = 20, 60, 100, 140$. The results, shown in Figure 2(a), reveal a nearly linear relationship between $|\mathcal{V}(\pi) - \widehat{\mathcal{V}}(\pi)|$ and T, which aligns with our theoretical results as shown in Theorem 3. (2) For the fully history-dependent target policy setting, to simplify computation, we fix the action space to a single action, $\mathcal{A} = \{1\}$. In this case, $\pi_t(1|O_t, H_{t-1}) = 1$ and the historical information H_{t-1} is only used to estimate the conditional probability matrices. We evaluate the policy value using sample sizes $n = 1000, 4000, 7000, 10000$, and horizon lengths $T = 2, 4, 6$. Figure 2(b) presents the logarithm of $|\mathcal{V}(\pi) - \widehat{\mathcal{V}}(\pi)|$ versus T. For $n = 1000$ setting, we observe noticeable fluctuations due to the increased size of the conditional probability matrices as T grows, which requires more samples to estimate each entry accurately. Nonetheless, across different n, we observe an approximately linear relationship between the logarithmic error and the horizon T. These experimental results are consistent with the theoretical findings presented in Theorem 3.

7 Conclusion

In this paper, we study the problem of OPE in confounded POMDPs, where both partial observability and unobserved confounding pose significant challenges to policy evaluation. To address these challenges, we propose a model-based framework that leverages observable history as a proxy for hidden states. Under suitable invertibility conditions, we establish identification results for history-dependent policies. Our theoretical analysis demonstrates that the proposed method achieves a convergence rate of $\mathcal{O}(\frac{T}{\sqrt{n}}|\mathcal{O}|^{\frac{T}{2}}|\mathcal{A}|^{\frac{T}{2}})$. An important direction for future research is to investigate the minimax-optimal rate for off-policy evaluation in confounded POMDPs and extend the framework to continuous state and action spaces.

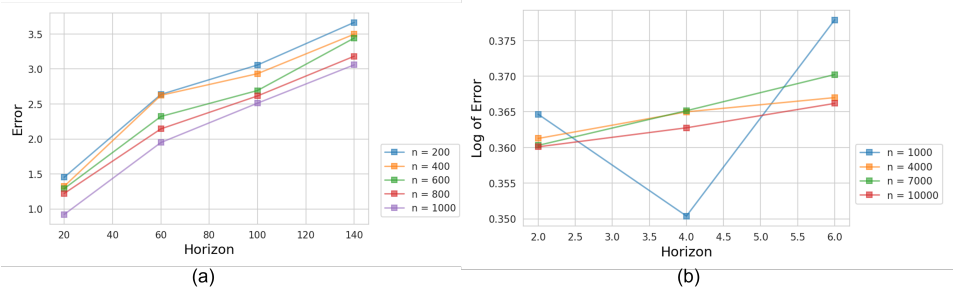

Figure 2: (a) Results for $|\mathcal{V}(\pi) - \widehat{\mathcal{V}}(\pi)|$ when the target policy is memoryless. (b) Results for $\log(|\mathcal{V}(\pi) - \widehat{\mathcal{V}}(\pi)|)$ when the target policy is fully history-dependent.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and constructive suggestions, which have significantly improved the quality of this paper. Qi Kuang acknowledges funding from the Early-Career Young Scientists and Technologists Project of Jiangxi Province (No. 20252BEJ730126) and the National Natural Science Foundation of China (No. 12571286). Fan Zhou acknowledges support from the Shanghai Research Center for Data Science and Decision Technology.

References

Albright, S. C. (1979). Structural results for partially observable markov decision processes. *Operations Research*, 27(5):1041–1053.

Bennett, A. and Kallus, N. (2024). Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *Operations Research*, 72(3):1071–1086.

Bravo, R. Z. B., Leiras, A., and Cyrino Oliveira, F. L. (2019). The use of uavs in humanitarian relief: an application of pomdp-based methodology for finding victims. *Production and Operations Management*, 28(2):421–440.

Cassandra, A. R. (1998). A survey of pomdp applications. In *Working notes of AAAI 1998 fall symposium on planning with partially observable Markov decision processes*, volume 1724.

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. (2024). Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*.

Hong, M., Qi, Z., and Xu, Y. (2024a). Model-based reinforcement learning for confounded pomdps. In *Forty-first International Conference on Machine Learning*.

Hong, M., Qi, Z., and Xu, Y. (2024b). A policy gradient method for confounded POMDPs. In *The International Conference on Learning Representations*.

Hu, Y. and Shiu, J.-L. (2018). Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Econometric Theory*, 34(3):659–693.

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, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 652–661, New York, New York, USA. PMLR.

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

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63.

Kress, R., Maz'ya, V., and Kozlov, V. (1989). *Linear integral equations*, volume 82. Springer.

Liu, Q. and Jin, C. (2022). Partially observable rl: Benign structures and simple generic algorithms. *https://qinghual2020.github.io*.

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31.

Lu, M., Min, Y., Wang, Z., and Yang, Z. (2023). Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. In *The International Conference on Learning Representations*.

Miao, R., Qi, Z., and Zhang, X. (2022). Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models. *arXiv preprint arXiv:2209.10064*.

Miao, W., Shi, X., and Tchetgen, E. T. (2018). A confounding bridge approach for double negative control inference on causal effects. *arXiv preprint arXiv:1808.04945*.

Monahan, G. E. (1982). State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms. *Management science*, 28(1):1–16.

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

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Precup, D. (2000). Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80.

Rudelson, M. and Vershynin, R. (2009). Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 62(12):1707–1739.

Shi, C., Uehara, M., Huang, J., and Jiang, N. (2022). A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes. In *International Conference on Machine Learning*, pages 20057–20094. PMLR.

Singh, S. P., Jaakkola, T., and Jordan, M. I. (1994). Learning without state-estimation in partially observable markovian decision processes. In *Machine Learning Proceedings 1994*, pages 284–292. Elsevier.

Stewart, G. W. and Sun, J.-g. (1990). *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic.

Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X., and Miao, W. (2020). An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*.

Tennenholtz, G., Shalit, U., and Mannor, S. (2020). Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10276–10283.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International conference on machine learning*, pages 2139–2148. PMLR.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434.

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.

Wang, J., Qi, Z., and Wong, R. K. (2024). A fine-grained analysis of fitted q-evaluation: Beyond parametric models. In *International Conference on Machine Learning*.

Xie, T., Ma, Y., and Wang, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in neural information processing systems*, 32.

Yin, M. and Wang, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR.

Young, S., Gasic, M., Thomson, B., and Williams, J. D. (2013). Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

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

Zhang, Y. and Jiang, N. (2025a). On the curses of future and history in future-dependent value functions for off-policy evaluation. *Advances in Neural Information Processing Systems*, 37:124756–124790.

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

The Appendix is organized as follows:

Section A presents the detailed statements for Theorem 2. Section B presents the detailed proof of Theorem 1. Section C presents the detailed proof of Theorem 2. Section D outlines the technical lemmas essential for the proofs. The code to implement the simulation is available at `https://github.com/kuangqi927/Confoundedpomdp`.

A Detailed Statements of Theorem 2

We begin by presenting the detailed statements for Theorem 2. We omit the full proof for Theorem 3, which differs from Theorem 2 only in the first-order term.

Theorem 4. *Under Assumptions 1 and 2, suppose the sample size n is sufficiently large, then, with probability at least $1-\delta$, it holds that*

$$\begin{aligned}\left|\mathcal{V}(\pi)-\widehat{\mathcal{V}}(\pi)\right| \lesssim & \frac{T^{1.5}}{\sqrt{n}}(1-\theta^*)^{-\frac{1}{2}} C_{\pi^b}^{\frac{1}{2}} C_P |\mathcal{O}|^{\frac{T}{2}} |\mathcal{A}|^{\frac{T}{2}} \\ & + \frac{T^{1.5}}{n}(1-\theta^*)^{-1} C_{\pi^b}^{\frac{1}{2}} C_P^2 |\mathcal{O}|^{\frac{3T}{2}} |\mathcal{A}|^{\frac{3T-1}{2}} \sqrt{\log(T^2|\mathcal{O}|^T|\mathcal{A}|^T/\delta)} \\ & + \frac{T^3}{n}(1-\theta^*)^{-1} C_{\pi^b}^{\frac{1}{2}} C_P^2 |\mathcal{O}|^{\frac{5T}{2}} |\mathcal{A}|^{\frac{5T}{2}} \sqrt{\log(T^2|\mathcal{O}|^T|\mathcal{A}|^T/\delta)} \\ & + \frac{T^3}{n^{\frac{3}{2}}}(1-\theta^*)^{-\frac{3}{2}} C_{\pi^b}^{\frac{1}{2}} C_P^3 |\mathcal{O}|^{\frac{5T}{2}} |\mathcal{A}|^{\frac{5T}{2}} \log(T^2|\mathcal{O}|^T|\mathcal{A}|^T/\delta) \\ & + \frac{T}{\sqrt{n}} \log(T^2|\mathcal{A}|/\delta).\end{aligned} \tag{12}$$

The first four terms are related to sub-optimality terms. The proof of the first-order term is provided in Appendix C.1.1, and the proof of the higher-order terms is given in Appendix C.1.5. The proof of the last term related to the empirical process is presented in Appendix C.2.

B Proof of Theorem 1

For convenience, we omit the uppercase letters and abbreviate $\mathbb{P}(X, Z=z \mid Y)$ as $\mathbb{P}(X, z \mid Y)$.

Proof. By the definition of policy value, we have

$$\begin{aligned}\mathcal{V}(\pi) =& \sum_{t=1}^{T}\sum_{r_t} r_t p^{\pi}(r_t) \\ =& \sum_{t=1}^{T}\sum_{r_t} r_t \sum_{o_0,s_1,a_1,o_1,\ldots,s_t,a_t,o_t} p(r_t \mid o_0,s_1,a_1,o_1,\ldots,s_t,a_t,o_t)p(o_0,s_1,a_1,o_1,\ldots,s_t,a_t,o_t) \\ =& \sum_{t=1}^{T}\sum_{r_t} r_t \sum_{o_0,s_1,a_1,o_1,\ldots,s_t,a_t,o_t} p(r_t \mid s_t,a_t)p(o_0,s_1,a_1,o_1,\ldots,s_t,a_t,o_t).\end{aligned}$$

We now recursively decompose the joint probability $p(o_0,s_1,a_1,o_1,\ldots,s_t,a_t,o_t)$ as follows,

$$\begin{aligned}& p(o_0,s_1,a_1,o_1,\ldots,s_t,a_t,o_t) \\ &= p(a_t \mid o_0,s_1,a_1,o_1,\ldots,s_t,o_t)p(o_0,s_1,a_1,o_1,\ldots,s_t,o_t) \\ &= \pi_t(a_t \mid o_t,h_{t-1})p(o_t \mid o_0,s_1,a_1,o_1,\ldots,s_t)p(o_0,s_1,a_1,o_1,\ldots,s_t) \\ &= \pi_t(a_t \mid o_t,h_{t-1})p(o_t \mid s_t)p(o_0,s_1,a_1,o_1,\ldots,s_t) \\ &= \pi_t(a_t \mid o_t,h_{t-1})p(o_t \mid s_t)p(s_t \mid o_0,s_1,a_1,o_1,\ldots,o_{t-1})p(o_0,s_1,a_1,o_1,\ldots,o_{t-1}) \\ &= \pi_t(a_t \mid o_t,h_{t-1})p(o_t \mid s_t)p(s_t \mid s_{t-1},a_{t-1})p(o_0,s_1,a_1,o_1,\ldots,o_{t-1}).\end{aligned}$$

Hence, by induction, we obtain that

$$p(o_0,s_1,a_1,o_1,\ldots,s_t,a_t,o_t) = \left(\prod_{i=1}^{t}\pi_i(a_i \mid o_i,h_{i-1})p(o_i \mid s_i)\right)\left(\prod_{i=1}^{t-1}p(s_{i+1} \mid s_i,a_i)\right)p(s_1|o_0)p(o_0).$$

Note that $O_t \perp\!\!\!\perp A_t, A_{t-1} \mid S_t$, we can rewrite as

$$\begin{aligned}
&p(r_t \mid s_t, a_t)p(o_0, s_1, a_1, o_1, \ldots, s_t, a_t, o_t) \\
&= p(r_t, o_t \mid s_t, a_t)\left(\prod_{i=1}^{t}\pi_i(a_i \mid o_i, h_{i-1})\right)\left(\prod_{i=1}^{t-1}p(s_{i+1}, o_i \mid s_i, a_i)\right)p(s_1|o_0)p(o_0).
\end{aligned}$$

We now rewrite the policy value in the vector form as

$$\begin{aligned}
\mathcal{V}(\pi) = \sum_{t=1}^{T}\sum_{r_t} r_t \sum_{a_1, o_1, \ldots, a_t, o_t}&\left(\prod_{i=1}^{t}\pi_i(a_i \mid o_i, h_{i-1})\right) \\
&\mathbb{P}(O_0)\mathbb{P}(S_1|O_0)\left(\prod_{i=1}^{t-1}\mathbb{P}(S_{i+1}, o_i \mid a_i, S_i)\right)\mathbb{P}(r_t, o_t \mid a_t, S_t).
\end{aligned} \tag{13}$$

Thus, the summation uses only observable variables.

We now aim to eliminate the hidden states in (13). We invoke Lemma 2 of Tchetgen et al. (2020), and the difference lies in that we use the full history $\{O_0, H_{t-1}\}$ as a proxy to infer hidden state S_t instead of one-step observation O_{t-1}. This yields the following key causal structure.

$$\begin{aligned}
&\mathbb{P}(S_t, o_{t-1} \mid a_{t-1}, S_{t-1})\mathbb{P}(S_{t+1}, o_t \mid a_t, S_t) \\
=&\mathbb{P}(O_{t-1} \mid a_{t-1}, S_{t-1})\mathbb{P}^\dagger(O_{t-1} \mid a_{t-1}, O_0, H_{t-2})\mathbb{P}(O_t, o_{t-1} \mid a_{t-1}, O_0, H_{t-2}) \\
&\mathbb{P}^\dagger(O_t \mid a_t, O_0, H_{t-1})\mathbb{P}(S_{t+1}, o_t \mid a_t, O_0, H_{t-1}),
\end{aligned} \tag{14}$$

and

$$\begin{aligned}
&\mathbb{P}(S_t, o_{t-1} \mid a_{t-1}, S_{t-1})\mathbb{P}(r_t, o_t \mid a_t, S_t) \\
=&\mathbb{P}(O_{t-1} \mid a_{t-1}, S_{t-1})\mathbb{P}^\dagger(O_{t-1} \mid a_{t-1}, O_0, H_{t-2})\mathbb{P}(O_{t-1} \mid a_{t-1}, O_0, H_{t-2}) \\
&\mathbb{P}^\dagger(O_t \mid a_t, O_0, H_{t-1})\mathbb{P}(r_t, o_t \mid a_t, O_0, H_{t-1}),
\end{aligned} \tag{15}$$

and

$$\mathbb{P}(S_t, o_{t-1} \mid a_{t-1}, O_0, H_{t-2})\mathbb{P}(O_t \mid a_t, S_t) = \mathbb{P}(O_t, o_{t-1} \mid a_{t-1}, O_0, H_{t-2}). \tag{16}$$

Combining these three equations (14), (15), and (16), we eliminate the latent states from (13) and yield

$$\begin{aligned}
&\mathbb{P}(O_0)\mathbb{P}(S_1|O_0)\left(\prod_{i=1}^{t-1}\mathbb{P}(S_{i+1}, o_i \mid a_i, S_i)\right)\mathbb{P}(r_t, o_t \mid a_t, S_t) \\
=&\mathbb{P}(O_0)\mathbb{P}(O_1|O_0) \\
&\mathbb{P}^\dagger(O_1 \mid a_1, O_0)\mathbb{P}(O_2, o_1 \mid a_1, O_0)\mathbb{P}^\dagger(O_2 \mid a_2, O_0, H_1)\mathbb{P}(O_3, o_2 \mid a_2, O_0, H_1) \\
&\mathbb{P}^\dagger(O_3 \mid a_3, O_0, H_2)\mathbb{P}(O_4, o_3 \mid a_3, O_0, H_2)\mathbb{P}^\dagger(O_4 \mid a_4, O_0, H_3)\mathbb{P}(O_5, o_4 \mid a_4, O_0, H_3) \\
&\cdots \\
&\mathbb{P}^\dagger(O_{t-1} \mid a_{t-1}, O_0, H_{t-2})\mathbb{P}(O_{t-1} \mid a_{t-1}, O_0, H_{t-2})\mathbb{P}^\dagger(O_t \mid a_t, O_0, H_{t-1})\mathbb{P}(r_t, o_t \mid a_t, O_0, H_{t-1}) \\
=&\mathbb{P}(O_0)\mathbb{P}(O_1|O_0)\mathbb{P}^\dagger(O_1 \mid a_1, O_0)\left(\prod_{i=1}^{t-1}\mathbb{P}(O_{i+1}, o_i \mid a_i, O_0, H_{i-1})\mathbb{P}^\dagger(O_{i+1} \mid a_{i+1}, O_0, H_i)\right) \\
&\mathbb{P}(r_t, o_t \mid a_t, O_0, H_{t-1}).
\end{aligned}$$

Here, we require $\mathbb{P}(O_t \mid a_t, S_t)$ to be invertible.

Putting all together gives the conclusion, which expresses the policy value entirely in terms of observable variables. □

B.1 Identification via value function

In this section, we present a complete proof of the identification results beyond the tabular setting.

Theorem 5 (Identification)**.** *Under Assumptions 4 and 5, the policy value for π can be identified as*

$$\mathcal{V}(\pi) = \mathbb{E}\Big[\sum_{a\in\mathcal{A}} b^{\pi}_{V,1}(a, O_1)\Big]. \tag{17}$$

Proof.

$$
\begin{aligned}
\mathcal{V}(\pi) =& \mathbb{E}^{\pi}\Big[\sum_{t=1}^{T} R_t\Big] \text{ (by definition of policy value)} \\
=& \mathbb{E}_{S_1\sim\nu_1}\sum_{t=1}^{T}\mathbb{E}^{\pi}\Big[R_t \;\Big|\; S_1\Big] \text{ (by the law of total expectation)} \\
=& \mathbb{E}_{S_1\sim\nu_1}\sum_{t=1}^{T}\mathbb{E}^{\pi}\Big[R_t \;\Big|\; S_1, H_0\Big] \text{ (by } H_0=\emptyset) \\
=& \mathbb{E}_{S_1\sim\nu_1}\left[\mathbb{E}\left[\sum_a b^{\pi}_{V,1}(a, O_1, H_0) \;\Big|\; S_1, H_0\right]\right] \text{ (by Lemma 2)} \\
=& \mathbb{E}_{S_1\sim\nu_1}\left[\mathbb{E}\left[\sum_a b^{\pi}_{V,1}(a, O_1) \;\Big|\; S_1\right]\right] \text{ (by } H_0=\emptyset) \\
=& \mathbb{E}\left[\sum_a b^{\pi}_{V,1}(a, O_1)\right].
\end{aligned}
$$

Consequently, we complete the proof. □

Assumption 1 ⇒ Assumption 4, 5. In tabular settings, the rank conditions are sufficient conditions for the completeness condition to hold. $\mathbb{P}(O_t \mid a_t, S_t)$ being invertible means different states map to distinguishable observation distributions. Similarly, invertibility of $\mathbb{P}(O_t \mid a_t, O_0, H_{t-1})$ ensures the history can span the full observation space. These together imply that the observable history $\{O_0, H_{t-1}\}$ contains enough information to separate functions of the hidden state S_t, which is exactly what completeness requires. Besides, the invertibility of $\mathbb{P}(O_t \mid a_t, O_0, H_{t-1})$ ensures the existence of value functions in Assumption 4.

Assumption 4. *There exist real-valued functions* $\{b^{\pi}_{V,t} : \mathcal{A}\times\mathcal{O}\times\mathcal{H}_{t-1}\to\mathbb{R}\}_{t=1}^{T}$ *that satisfy the following conditional moment restrictions:*

$$
\begin{aligned}
&\mathbb{E}\Big[b^{\pi}_{V,t}\left(A_t, O_t, H_{t-1}\right) \;\Big|\; O_0, A_t, H_{t-1}\Big] \\
&= \mathbb{E}\Big[R_t\pi_t(A_t \mid O_t, H_{t-1}) + \sum_{a'} b^{\pi}_{V,t+1}\left(a', O_{t+1}, H_t\right)\pi_t(A_t \mid O_t, H_{t-1}) \;\Big|\; O_0, A_t, H_{t-1}\Big].
\end{aligned}
$$

Beyond the tabular setting, the existence of these value functions is justified by some mild regularity conditions utilizing tools from singular value decomposition in functional analysis (Kress et al., 1989).

Assumption 5 (Completeness)**.** *For any measurable function* $g_t : \mathcal{S}\times\mathcal{A}\times\mathcal{H}_{t-1}\to\mathbb{R}$ *and any* $1\le t\le T$*,*

$$
\mathbb{E}\big[g_t(S_t, A_t, H_{t-1}) \mid A_t, H_{t-1}, O_0\big] = 0
$$

almost surely if and only if $g_t(S_t, A_t, H_{t-1}) = 0$ *almost surely.*

The completeness assumption is widely used in statistics and econometrics. For example, it plays a crucial role in instrumental variable estimation, where the identification of structural functions often depends on the completeness (Newey and Powell, 2003; Hu and Shiu, 2018).

Under Assumptions 4 and 5, we obtain another sequence of conditional moment restrictions (18) that are projected onto (S_t, A_t, H_{t-1}).

$$
\begin{aligned}
\mathbb{E}\Big[b^{\pi}_{V,t}(A_t, H_{t-1}, O_t) \;\Big|\; A_t, H_{t-1}, S_t\Big] = \mathbb{E}\Big[&R_t\pi_t(A_t \mid O_t, H_{t-1}) \\
&+ \sum_{a'\in\mathcal{A}} b^{\pi}_{V,t+1}(a', H_t, O_{t+1})\pi_t(A_t \mid O_t, H_{t-1}) \;\Big|\; A_t, H_{t-1}, S_t\Big].
\end{aligned} \tag{18}
$$

Lemma 2. *Under Assumptions 4 and 5, for all* $t = 1, ..., T$*, it holds that*

$$
\mathbb{E}\Big[\sum_a b^{\pi}_{V,t}(a, O_t, H_{t-1}) \mid S_t, H_{t-1}\Big] = \sum_{j=t}^{T}\mathbb{E}^{\pi}\Big[R_t \mid S_t, H_{t-1}\Big]. \tag{19}
$$

B.2 Proof of Lemma 2

Proof. We prove it by induction. At the step $t = T$, we have

$$
\begin{aligned}
&\mathbb{E}^{\pi}\left[R_T \;\middle|\; S_T, H_{T-1}\right] \\
=&\mathbb{E}\left[\mathbb{E}^{\pi}\left[\mathbb{E}\left[R_T \;\middle|\; S_T, H_{T-1}, O_T, A_T\right] \;\middle|\; S_T, H_{T-1}, O_T\right] \;\middle|\; S_T, H_{T-1}\right] \\
&\text{(by law of total expectation)} \\
=&\mathbb{E}^{\pi}\left[\mathbb{E}\left[\sum_a \mathbb{E}\left[R_T \;\middle|\; S_T, H_{T-1}, O_T, A_T = a\right]\pi_T(a \mid O_T, H_{T-1}) \;\middle|\; S_T, H_{T-1}, O_T\right] \;\middle|\; S_T, H_{T-1}\right] \\
=&\mathbb{E}\left[\sum_a \mathbb{E}\left[R_T \;\middle|\; S_T, H_{T-1}, O_T, A_T = a\right]\pi_T(a \mid O_T, H_{T-1}) \;\middle|\; S_T, H_{T-1}\right] \\
=&\mathbb{E}\left[\sum_a \mathbb{E}\left[R_T \;\middle|\; S_T, H_{T-1}, A_T = a\right]\pi_T(a \mid O_T, H_{T-1}) \;\middle|\; S_T, H_{T-1}\right] \\
&\text{(by } R_T \perp\!\!\!\perp O_T \mid S_T, A_T, H_{T-1}\text{)} \\
=&\sum_a \mathbb{E}\left[R_T \;\middle|\; S_T, H_{T-1}, A_T = a\right]\mathbb{E}\left[\pi_T(a \mid O_T, H_{T-1}) \;\middle|\; S_T, H_{T-1}\right] \\
=&\sum_a \mathbb{E}\left[R_T \;\middle|\; S_T, H_{T-1}, A_T = a\right]\mathbb{E}\left[\pi_T(a \mid O_T, H_{T-1}) \;\middle|\; S_T, H_{T-1}, A_T = a\right] \\
&\text{(by } O_T \perp\!\!\!\perp A_T \mid S_T, H_{T-1}\text{)} \\
=&\sum_a \mathbb{E}\left[R_T\pi_T(a \mid O_T, H_{T-1}) \;\middle|\; S_T, H_{T-1}, A_T = a\right] \quad \text{(by } O_T \perp\!\!\!\perp R_T \mid S_T, A_T, H_{T-1}\text{)} \\
=&\sum_a \mathbb{E}\left[b^{\pi}_{V,T}(O_T, H_{T-1}, a) \;\middle|\; S_T, H_{T-1}, A_T = a\right] \quad \text{(by Equation 18)} \\
=&\sum_a \mathbb{E}\left[b^{\pi}_{V,T}(O_T, H_{T-1}, a) \;\middle|\; S_T, H_{T-1}\right] \quad \text{(by } O_T \perp\!\!\!\perp A_T \mid S_T, H_{T-1}\text{)} \\
=&\mathbb{E}\left[\sum_a b^{\pi}_{V,T}(a, O_T, H_{T-1}) \;\middle|\; S_T, H_{T-1}\right]
\end{aligned}
$$

According to the above derivation, we have shown $\mathbb{E}\big[\sum_a b^{\pi}_{V,j}(a, O_j, H_{j-1}) \mid S_j, H_{j-1}\big] = \mathbb{E}^{\pi}\big[\sum_{t=j}^{T} R_t \mid S_j, H_{j-1}\big]$ when $j = T$. We proceed with the derivation by induction. Assume that $\mathbb{E}\big[\sum_a b^{\pi}_{V,j}(a, O_j, H_{j-1}) \mid S_j, H_{j-1}\big] = \mathbb{E}^{\pi}\big[\sum_{t=j}^{T} R_t \mid S_j, H_{j-1}\big]$ holds for $j = k+1$, we will show that it also holds for $j = k$.

For $j = k$, we first notice that

$$
\mathbb{E}^{\pi}\left[\sum_{t=k}^{T} R_t \;\middle|\; S_k, H_{k-1}\right] = \mathbb{E}^{\pi}\left[R_k \;\middle|\; S_k, H_{k-1}\right] + \mathbb{E}^{\pi}\left[\sum_{t=k+1}^{T} R_t \;\middle|\; S_k, H_{k-1}\right].
$$

Next, we analyze these two terms separately. Analyzing the first term is the same as $\mathbb{E}^{\pi}\left[R_T \mid S_T, H_{T-1}\right]$ by replacing T with k.

$$
\begin{aligned}
&\mathbb{E}^{\pi}\left[R_k \;\middle|\; S_k, H_{k-1}\right] \\
=&\mathbb{E}\left[\mathbb{E}^{\pi}\left[\mathbb{E}\left[R_k \;\middle|\; S_k, H_{k-1}, O_k, A_k\right] \;\middle|\; S_k, H_{k-1}, O_k\right] \;\middle|\; S_k, H_{k-1}\right] \\
&\quad \text{(by law of total expectation)} \\
=&\mathbb{E}^{\pi}\left[\mathbb{E}\left[\sum_a \mathbb{E}\left[R_k \;\middle|\; S_k, H_{k-1}, O_k, A_k = a\right]\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}, O_k\right] \;\middle|\; S_k, H_{k-1}\right] \\
=&\mathbb{E}\left[\sum_a \mathbb{E}\left[R_k \;\middle|\; S_k, H_{k-1}, O_k, A_k = a\right]\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}\right] \\
=&\mathbb{E}\left[\sum_a \mathbb{E}\left[R_k \;\middle|\; S_k, H_{k-1}, A_k = a\right]\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}\right] \\
&\quad \text{(by } R_k \perp\!\!\!\perp O_k \mid S_k, A_k, H_{k-1}\text{)} \\
=&\sum_a \mathbb{E}\left[R_k \;\middle|\; S_k, H_{k-1}, A_k = a\right]\mathbb{E}\left[\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}\right] \\
=&\sum_a \mathbb{E}\left[R_k \;\middle|\; S_k, H_{k-1}, A_k = a\right]\mathbb{E}\left[\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}, A_k = a\right] \\
&\quad \text{(by } O_k \perp\!\!\!\perp A_k \;\middle|\; S_k, H_{k-1}\text{)} \\
=&\sum_a \mathbb{E}\left[R_k\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}, A_k = a\right] \\
&\quad \text{(by } O_k \perp\!\!\!\perp R_k \;\middle|\; S_k, A_k, H_{k-1}\text{)} \\
=&\sum_a \mathbb{E}\left[R_k\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}, A_k = a\right]
\end{aligned}
\tag{20}
$$

For the second term, we have

$$
\begin{aligned}
&\mathbb{E}^{\pi}\left[\sum_{t=k+1}^{T} R_t \;\middle|\; S_k, H_{k-1}\right] \\
=&\mathbb{E}^{\pi}\left[\mathbb{E}^{\pi}\left[\sum_{t=k+1}^{T} R_t \;\middle|\; S_{k+1}, H_k, S_k\right] \;\middle|\; S_k, H_{k-1}\right] \text{ (by law of total expectation)} \\
=&\mathbb{E}^{\pi}\left[\mathbb{E}^{\pi}\left[\sum_{t=k+1}^{T} R_t \;\middle|\; S_{k+1}, H_k\right] \;\middle|\; S_k, H_{k-1}\right] \text{ (by } \{R_t\}_{t=k+1}^{T} \perp\!\!\!\perp S_k \mid S_{k+1}, H_k) \\
=&\mathbb{E}^{\pi}\left[\mathbb{E}\left[\sum_{a'} b^{\pi}_{V,k+1}(a', O_{k+1}, H_k) \;\middle|\; S_{k+1}, H_k\right] \;\middle|\; S_k, H_{k-1}\right] \\
=&\mathbb{E}^{\pi}\left[\mathbb{E}\left[\sum_{a'} b^{\pi}_{V,k+1}(a', O_{k+1}, H_k) \;\middle|\; S_{k+1}, H_k, S_k\right] \;\middle|\; S_k, H_{k-1}\right] (O_{k+1} \perp\!\!\!\perp S_k \mid S_{k+1}, H_k) \\
=&\mathbb{E}^{\pi}\left[\sum_{a'} b^{\pi}_{V,k+1}(a', O_{k+1}, H_k) \;\middle|\; S_k, H_{k-1}\right] \text{ (by law of total expectation)} \\
=&\mathbb{E}^{\pi}\left[\mathbb{E}\left[\sum_{a'} b^{\pi}_{V,k+1}(a', O_{k+1}, H_k) \;\middle|\; S_k, H_{k-1}, O_k, A_k\right] \;\middle|\; S_k, H_{k-1}\right] \text{ (by law of total expectation)} \\
=&\mathbb{E}\left[\sum_{a} \mathbb{E}\left[\sum_{a'} b^{\pi}_{V,k+1}(a', O_{k+1}, H_k) \;\middle|\; S_k, H_{k-1}, O_k, A_k = a\right] \pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}\right] \\
=&\mathbb{E}\left[\sum_{a} \mathbb{E}\left[\sum_{a'} b^{\pi}_{V,k+1}(a', O_{k+1}, H_k)\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}, O_k, A_k = a\right] \;\middle|\; S_k, H_{k-1}\right] \\
=&\sum_{a} \mathbb{E}\left[\mathbb{E}\left[\sum_{a'} b^{\pi}_{V,k+1}(a', O_{k+1}, H_k)\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}, O_k, A_k = a\right] \;\middle|\; S_k, H_{k-1}\right] \\
=&\sum_{a} \mathbb{E}\left[\mathbb{E}\left[\sum_{a'} b^{\pi}_{V,k+1}(a', O_{k+1}, H_k)\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}, O_k, A_k = a\right] \;\middle|\; S_k, H_{k-1}, A_k = a\right] \\
&\text{(by } O_k \perp\!\!\!\perp A_k \mid S_k) \\
=&\sum_{a} \mathbb{E}\left[\sum_{a'} b^{\pi}_{V,k+1}(a', O_{k+1}, H_k)\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}, A_k = a\right] \text{ (by law of total expectation)}
\end{aligned}
\tag{21}
$$

Combining Equations (20) and (21), we have

$$
\begin{aligned}
&\mathbb{E}^{\pi}\left[\sum_{t=k}^{T} R_t \;\middle|\; S_k, H_{k-1}\right] \\
=&\mathbb{E}^{\pi}\left[R_k \;\middle|\; S_k, H_{k-1}\right] + \mathbb{E}^{\pi}\left[\sum_{t=k+1}^{T} R_t \;\middle|\; S_k, H_{k-1}\right] \\
=&\sum_{a} \mathbb{E}\left[R_k \pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}, A_k = a\right] \\
+&\sum_{a} \mathbb{E}\left[\sum_{a'} b^{\pi}_{V,k+1}(a', O_{k+1}, H_k)\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}, A_k = a\right] \\
&\text{(by Equations (20) and (21))} \\
=&\sum_{a} \mathbb{E}\left[R_k \pi_k(a \mid O_k, H_{k-1}) + \sum_{a'} b^{\pi}_{V,k+1}(a', O_{k+1}, H_k)\pi_k(a \mid O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}, A_k = a\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_a \mathbb{E}\left[b^\pi_{V,k}(a, O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}, A_k = a \right] \quad \text{(by Equation 18)} \\
&= \sum_a \mathbb{E}\left[b^\pi_{V,k}(a, O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1} \right] \quad \text{(by } O_k \perp\!\!\!\perp A_k \;\middle|\; S_k, H_{k-1}) \\
&= \mathbb{E}\left[\sum_a b^\pi_{V,k}(a, O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1} \right]
\end{aligned}
$$

Therefore, $\mathbb{E}^\pi\left[\sum_{t=k}^T R_t \;\middle|\; S_k, H_{k-1}\right] = \mathbb{E}\left[b^\pi_{V,k}(O_k, H_{k-1}) \;\middle|\; S_k, H_{k-1}\right]$ also holds for $j = k$, if it holds for $j = k+1$. By the induction argument, the proof is completed. □

B.3 Identification via weight functions

We provide an alternative identification formula with weight functions $\{b^\pi_{W,t}\}_{t=1}^T$.

Theorem 6 (Identification with weight functions)**.** *Under Assumptions 4 and 5, the policy value can be identified by weight functions as*

$$
\mathcal{V}(\pi) = \mathbb{E}\Big[\sum_{t=1}^T R_t \pi_t(A_t \mid O_t, H_{t-1}) b^\pi_{W,t}(A_t, H_{t-1}, O_0) \Big].
$$

Proof.

$$
\begin{aligned}
\mathcal{V}(\pi) &= \sum_{t=1}^T \mathbb{E}^\pi\left[R_t(S_t, A_t)\right] \\
&= \sum_{t=1}^T \mathbb{E}^\pi\left[\mathbb{E}^\pi\left[R_t(S_t, A_t) \;\middle|\; O_t, S_t, H_{t-1}\right]\right] \\
&= \sum_{t=1}^T \mathbb{E}^\pi\left[\sum_a R_t(S_t, a)\pi_t(a \mid O_t, H_{t-1})\right] \\
&= \sum_{t=1}^T \mathbb{E}^\pi\left[\mathbb{E}^\pi\left[\sum_a R_t(S_t, a)\pi_t(a|O_t, H_{t-1}) \;\middle|\; S_t, H_{t-1}\right]\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[\omega_t(S_t, H_{t-1}) \cdot \mathbb{E}^\pi\left[\sum_a R_t(S_t, a)\pi_t(a \mid O_t, H_{t-1}) \;\middle|\; S_t, H_{t-1}\right]\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[\omega_t(S_t, H_{t-1}) \sum_{a\in\mathcal{A}} R_t(S_t, a)\pi_t(a|O_t, H_{t-1})\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[\omega_t(S_t, H_{t-1}) \sum_{a\in\mathcal{A}} R_t(S_t, a)\pi^b_t(a|S_t)\frac{\pi_t(a|O_t, H_{t-1})}{\pi^b_t(a|S_t)}\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\omega_t(S_t, H_{t-1})R_t(S_t, A_t)\frac{\pi_t(A_t|O_t, H_{t-1})}{\pi^b_t(A_t \mid S_t)} \;\middle|\; S_t, H_{t-1}, O_t\right]\right] \quad \text{(by } A_t \sim \pi^b_t(\cdot|S_t)) \\
&= \sum_{t=1}^T \mathbb{E}\left[R_t(S_t, A_t)\pi_t(A_t \mid O_t, H_{t-1})\frac{\omega_t(S_t, H_{t-1})}{\pi^b_t(A_t \mid S_t)}\right] \\
&= \sum_{t=1}^T \mathbb{E}\left[R_t(S_t, A_t)\pi_t(A_t \mid O_t, H_{t-1})\mathbb{E}\left[b^\pi_{W,t}(A_t, H_{t-1}, O_0) \;\middle|\; S_t, H_{t-1}, A_t\right]\right] \quad \text{(by Equation 18)} \\
&= \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[R_t(S_t, A_t)\pi_t(A_t \mid O_t, H_{t-1})b^\pi_{W,t}(A_t, H_{t-1}, O_0) \;\middle|\; S_t, H_{t-1}, A_t\right]\right]
\end{aligned}
$$

$$
\begin{aligned}
&(O_0 \perp\!\!\!\perp (O_t, R_t) \mid S_t, A_t, H_{t-1}) \\
&= \sum_{t=1}^{T} \mathbb{E}\left[R_t(S_t, A_t)\pi_t(A_t|O_t, H_{t-1})b^{\pi}_{W,t}(A_t, H_{t-1}, O_0)\right].
\end{aligned}
$$

□

C Proof of Theorem 2 and 3

In this section, we provide the proof for Theorems 2 and 3 in the main text. We only show the proof for the fully history-dependent policy, as the memoryless case can be derived straightforwardly by excluding the history H_{t-1} from the conditional probabilities $\mathbf{P}_{a_t}$ and $\mathbf{P}_{a_t,r_t,o_{t+1}}$. Before proceeding with the analysis, we first recall and introduce some notations.

Additional Notations. For $t = 1, \ldots, T$, let $\mathcal{D}_t$ represent the collection of historical data up to time step t, i.e., $\mathcal{D}_t = \{o_0^i, (o_{t'}^i, a_{t'}^i, r_{t'}^i)_{t'=1}^t\}_{i=1}^n$ with $\mathcal{D}_0 = \{o_0^i\}_{i=1}^n$. For simplicity, we omit the superscript π^b when referring to the expectation, variance, or probability under the distribution induced by π^b. To distinguish between different sources of randomness, we use $\mathcal{E}$ to denote expectations over random variables (capital letters) and $\mathbb{E}$ to represent expectations over offline data, when both $\mathcal{E}$ and $\mathbb{E}$ appear simultaneously. We use $\mathbf{I}$ to represent the identity matrix, with its dimension being clear from the context. If a non-negative random variable X satisfies $P(X \leq c\Xi(n,T)) \to 0$ as $c \to 0$ for any n, T, we write $X = \mathcal{O}_P(\Xi(n,T))$ with high probability.

Note that we assume a sufficiently large n in Assumption 2(c), where we require a sufficient number of samples for each triple (o_0, h_{t-1}, a_t), ensuring consistent estimation of the conditional probability matrices. Specifically, we require $n_{o_0,h_{t-1},a_t} \geq np_t^{\pi^b}(o_0, h_{t-1}, a_t)(1-\theta_{t,ij})$, where n_{o_0,h_{t-1},a_t} represents the count of the triple (o_0, h_{t-1}, a_t) in the data, and $p_t^{\pi^b}(o_0, h_{t-1}, a_t)$ is the probability density under the behavior policy π^b. Define the event $E := \{\exists\, t, o_0, h_{t-1}, a_t \text{ s.t. } n_{o_0,h_{t-1},a_t} \geq np_t^{\pi^b}(o_0, h_{t-1}, a_t)(1-\theta_{t,ij})\}$. Then, combining the multiplicative Chernoff bound and a union bound over each t, o_0, h_{t-1}, and a_t, we have

$$
\begin{aligned}
\mathbb{P}[E^c] &\leq \sum_t \sum_{o_0} \sum_{h_{t-1}} \sum_{a_t} \mathbb{P}\left[n_{o_0,h_{t-1},a_t} < np_t^{\pi^b}(o_0, h_{t-1}, a_t)(1-\theta_{t,ij})\right] \\
&\leq T|\mathcal{O}|^T|\mathcal{A}|^T e^{-\frac{\theta^{*2} n \min_{t,o_0,h_{t-1},a_t} p_t^{\pi^b}(o_0,h_{t-1},a_t)}{2}}.
\end{aligned}
$$

To ensure the number of samples is sufficiently large, the $\mathbb{P}[E^c]$ should be sufficiently small. This requires the sample size satisfying $n \geq \frac{\text{polylog}(|\mathcal{O}|^T, |\mathcal{A}|^T, T)}{\min_{t,o_0,h_{t-1},a_t} p_t^{\pi^b}(o_0,h_{t-1},a_t)}$.

Then, we consider the following decomposition of the error

$$
\begin{aligned}
\mathcal{V}(\pi) - \widehat{\mathcal{V}}(\pi) &= \mathbb{E}\Big[\sum_a b^{\pi}_{V,1}(a, O_1)\Big] - \widehat{\mathbb{E}}\Big[\sum_a \widehat{b}_{V,1}(a, O_1)\Big] \\
&= \mathbb{E}\Big[\sum_a b^{\pi}_{V,1}(a, O_1)\Big] - \mathbb{E}\Big[\sum_a \widehat{b}_{V,1}(a, O_1)\Big] \\
&\quad + \mathbb{E}\Big[\sum_a \widehat{b}_{V,1}(a, O_1)\Big] - \widehat{\mathbb{E}}\Big[\sum_a \widehat{b}_{V,1}(a, O_1)\Big].
\end{aligned}
$$

We begin by analyzing the sub-optimality term $\mathbb{E}[\sum_a b^{\pi}_{V,1}(a, O_1)] - \mathbb{E}[\sum_a \widehat{b}_{V,1}(a, O_1)]$. The second term $\mathbb{E}[\sum_a \widehat{b}_{V,1}(a, O_1)] - \widehat{\mathbb{E}}[\sum_a \widehat{b}_{V,1}(a, O_1)]$ can be upper bounded by the uniform law of large numbers according to the empirical processes.

C.1 Bounding $\mathbb{E}[\sum_a b^{\pi}_{V,1}(a, O_1)] - \mathbb{E}[\sum_a \widehat{b}_{V,1}(a, O_1)]$

We first decompose the sub-optimality term into three parts. Note that

$$
\mathbb{E}\left[\sum_a b^{\pi}_{V,1}(a, O_1) - \sum_a \widehat{b}_{V,1}(a, O_1)\right]
$$

$$
\begin{aligned}
=&\mathcal{E}\Bigg[\sum_{a_1} b^{\pi}_{V,1}(a_1, O_1) - \sum_{a_1}\pi_1(a_1|O_1)\psi_1^\top(O_1)\widehat{\mathbf{P}}^\dagger_{a_1}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1(O_1)\Big(r_1 + \sum_{a_2}\widehat{b}_{V,2}\big(a_2,o_2,(O_1,a_1)\big)\Big)\Bigg]\\
=&\mathcal{E}\Bigg[\sum_{a_1} b^{\pi}_{V,1}(a_1, O_1) - \sum_{a_1}\pi_1(a_1|O_1)\psi_1^\top(O_1)\widehat{\mathbf{P}}^\dagger_{a_1}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1(O_1)\Big(r_1 + \sum_{a_2} b^{\pi}_{V,2}\big(a_2,o_2,(O_1,a_1)\big)\Big)\\
&+\sum_{a_1}\pi_1(a_1|O_1)\psi_1^\top(O_1)\widehat{\mathbf{P}}^\dagger_{a_1}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1(O_1)\Big(\sum_{a_2} b^{\pi}_{V,2}\big(a_2,o_2,(O_1,a_1)\big) - \sum_{a_2}\widehat{b}_{V,2}\big(a_2,o_2,(O_1,a_1)\big)\Big)\Bigg]\\
=&\mathcal{E}\Bigg[\sum_{a_1} b^{\pi}_{V,1}(a_1, O_1) - \sum_{a_1}\pi_1(a_1|O_1)\psi_1^\top(O_1)\widehat{\mathbf{P}}^\dagger_{a_1}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1(O_1)\Big(r_1 + \sum_{a_2} b^{\pi}_{V,2}\big(a_2,o_2,(O_1,a_1)\big)\Big)\\
&+\sum_{a_1}\pi_1(a_1|O_1)\psi_1^\top(O_1)\widehat{\mathbf{P}}^\dagger_{a_1}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1(O_1)\Big(\sum_{a_2} b^{\pi}_{V,2}\big(a_2,o_2,(O_1,a_1)\big) - \sum_{a_2}\pi_2\big(a_2|o_2,(O_1,a_1)\big)\\
&\psi_2^\top(o_2)\widehat{\mathbf{P}}^\dagger_{a_2}\sum_{r_2,o_3}\widehat{\mathbf{P}}_{a_2,r_2,o_3}\psi_2(o_2)\Big(r_2 + \sum_{a_3}\widehat{b}_{V,3}\big(a_3,o_3,(O_1,a_1,o_2,a_2)\big)\Big)\Big)\Bigg]\\
=&\cdots\\
=&\mathcal{E}\big(E_1\big) + \mathcal{E}\big(E_2\big) + \mathcal{E}\big(E_3\big).
\end{aligned}
$$

We replace the data dependent terms $\widehat{\mathbf{P}}^\dagger_{a_t}$ and $\widehat{\mathbf{P}}^\dagger_{a_t,r_t,o_{t+1}}$ with their population counterparts $\mathbf{P}^\dagger_{a_t}$ and $\mathbf{P}^\dagger_{a_t,r_t,o_{t+1}}$, respectively. This error decomposition is then the sum of $\mathcal{E}(E_1)$, $\mathcal{E}(E_2)$ and $\mathcal{E}(E_3)$, where

$$
\begin{aligned}
E_1 =& \sum_{a_1}\pi_1\psi_1^\top\mathbf{P}^\dagger_{a_1}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1\Big[\sum_{a_1} b^{\pi}_{V,1}(a_1, O_1) - \Big(r_1 + \sum_{a_2} b^{\pi}_{V,2}\big(a_2,o_2,(O_1,a_1)\big)\Big)\Big]\\
&+\sum_{a_1}\pi_1\psi_1^\top\mathbf{P}^\dagger_{a_1}\sum_{r_1,o_2}\mathbf{P}_{a_1,r_1,o_2}\psi_1\sum_{a_2}\pi_2\psi_2^\top\mathbf{P}^\dagger_{a_2}\sum_{r_2,o_3}\widehat{\mathbf{P}}_{a_2,r_2,o_3}\psi_2\\
&\Big[\sum_{a_2} b^{\pi}_{V,2}\big(a_2,o_2,(O_1,a_1)\big) - \Big(r_2 + \sum_{a_2} b^{\pi}_{V,3}\big(a_3,o_3,(O_1,a_1,o_2,a_2)\big)\Big)\Big]\\
&+\cdots\\
&+\sum_{a_1}\pi_1\psi_1^\top\mathbf{P}^\dagger_{a_1}\sum_{r_1,o_2}\mathbf{P}_{a_1,r_1,o_2}\psi_1\sum_{a_2}\pi_2\psi_2^\top\mathbf{P}^\dagger_{a_2}\sum_{r_2,o_3}\mathbf{P}_{a_2,r_2,o_3}\psi_2\cdots\sum_{a_T}\pi_T\psi_T^\top\mathbf{P}^\dagger_{a_T}\sum_{r_T}\widehat{\mathbf{P}}_{a_T,r_T}\psi_T\\
&\Big[\sum_{a_T} b^{\pi}_{V,T}\big(a_T,o_T,(O_1,a_1,o_2,\cdots,a_{T-1})\big) - r_T\Big],\\
E_2 =&\Big(\sum_{a_1}\pi_1\psi_1^\top\widehat{\mathbf{P}}^\dagger_{a_1}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1 - \sum_{a_1}\pi_1\psi_1^\top\mathbf{P}^\dagger_{a_1}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1\Big)\\
&\Big[\sum_{a_1} b^{\pi}_{V,1}(a_1, O_1) - \Big(r_1 + \sum_{a_2} b^{\pi}_{V,2}\big(a_2,o_2,(O_1,a_1)\big)\Big)\Big],\\
&+\Big(\sum_{a_1}\pi_1\psi_1^\top\widehat{\mathbf{P}}^\dagger_{a_1}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1\sum_{a_2}\pi_2\psi_2^\top\widehat{\mathbf{P}}^\dagger_{a_2}\sum_{r_2,o_3}\widehat{\mathbf{P}}_{a_2,r_2,o_3}\psi_2\\
&-\sum_{a_1}\pi_1\psi_1^\top\mathbf{P}^\dagger_{a_1}\sum_{r_1,o_2}\mathbf{P}_{a_1,r_1,o_2}\psi_1\sum_{a_2}\pi_2\psi_2^\top\mathbf{P}^\dagger_{a_2}\sum_{r_2,o_3}\widehat{\mathbf{P}}_{a_2,r_2,o_3}\psi_2\Big)\\
&\Big[\sum_{a_2} b^{\pi}_{V,2}\big(a_2,o_2,(O_1,a_1)\big) - \Big(r_2 + \sum_{a_2} b^{\pi}_{V,3}\big(a_3,o_3,(O_1,a_1,o_2,a_2)\big)\Big)\Big]\\
&\cdots\\
&+\Big(\sum_{a_1}\pi_1\psi_1^\top\widehat{\mathbf{P}}^\dagger_{a_1}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1\sum_{a_2}\pi_2\psi_2^\top\widehat{\mathbf{P}}^\dagger_{a_2}\sum_{r_2,o_3}\widehat{\mathbf{P}}_{a_2,r_2,o_3}\psi_2\cdots\sum_{a_T}\pi_T\psi_T^\top\widehat{\mathbf{P}}^\dagger_{a_T}\sum_{r_T}\widehat{\mathbf{P}}_{a_T,r_T}\psi_T
\end{aligned}
$$

$$
\begin{aligned}
&- \sum_{a_1} \pi_1 \psi_1^\top \mathbf{P}_{a_1}^\dagger \sum_{r_1,o_2} \mathbf{P}_{a_1,r_1,o_2} \psi_1 \sum_{a_2} \pi_2 \psi_2^\top \mathbf{P}_{a_2}^\dagger \sum_{r_2,o_3} \mathbf{P}_{a_2,r_2,o_3} \psi_2 \cdots \sum_{a_T} \pi_T \psi_T^\top \mathbf{P}_{a_T}^\dagger \sum_{r_T} \widehat{\mathbf{P}}_{a_T,r_T} \psi_T \Bigg) \\
&\Bigg[\sum_{a_T} b_{V,T}^\pi \big(a_T, o_T, (O_1, a_1, o_2, \cdots, a_{T-1})\big) - r_T \Bigg], \\
E_3 = &\sum_{a_1} b_{V,1}^\pi(a_1, O_1) - \sum_{a_1} \pi_1 \psi_1^\top \widehat{\mathbf{P}}_{a_1}^\dagger \sum_{r_1,o_2} \widehat{\mathbf{P}}_{a_1,r_1,o_2} \psi_1 \sum_{a_1} b_{V,1}^\pi(a_1, O_1) \\
&+ \sum_{a_1} \pi_1 \psi_1^\top \widehat{\mathbf{P}}_{a_1}^\dagger \sum_{r_1,o_2} \widehat{\mathbf{P}}_{a_1,r_1,o_2} \psi_1 \Bigg[\sum_{a_2} b_{V,2}^\pi - \sum_{a_2} \pi_2 \psi_2^\top \widehat{\mathbf{P}}_{a_2}^\dagger \sum_{r_2,o_3} \widehat{\mathbf{P}}_{a_2,r_2,o_3} \psi_2 \sum_{a_2} b_{V,2}^\pi \Bigg] \\
&+ \cdots \\
&+ \sum_{a_1} \pi_1 \psi_1^\top \widehat{\mathbf{P}}_{a_1}^\dagger \sum_{r_1,o_2} \widehat{\mathbf{P}}_{a_1,r_1,o_2} \psi_1 \cdots \sum_{a_{T-1}} \pi_T \psi_{T-1}^\top \widehat{\mathbf{P}}_{a_{T-1}}^\dagger \sum_{r_{T-1},o_T} \widehat{\mathbf{P}}_{a_{T-1},r_{T-1},o_T} \psi_{T-1} \\
&\Bigg[\sum_{a_T} b_{V,T}^\pi - \sum_{a_T} \pi_T \psi_T^\top \widehat{\mathbf{P}}_{a_T}^\dagger \sum_{r_T} \widehat{\mathbf{P}}_{a_T,r_T} \psi_T \sum_{a_T} b_{V,T}^\pi \Bigg].
\end{aligned}
$$

Note that these terms ultimately decompose into a sum over all possible observable history trajectories. Simplifying these terms, we have

$$
\begin{aligned}
E_1 = &\sum_{t=1}^T \Bigg(\sum_{a_1} \pi_1 \psi_1^\top \mathbf{P}_{a_1}^\dagger \cdots \sum_{r_t,o_{t+1}} \widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}} \psi_t \Bigg) \Bigg[\sum_{a_t} b_{V,t}^\pi - \Big(r_t + \sum_{a_{t+1}} b_{V,t+1}^\pi \Big) \Bigg] \\
E_2 = &\sum_{t=1}^T \Bigg(\sum_{a_1} \pi_1 \psi_1^\top \widehat{\mathbf{P}}_{a_1}^\dagger \cdots \sum_{r_t,o_{t+1}} \widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}} \psi_t - \sum_{a_1} \pi_1 \psi_1^\top \mathbf{P}_{a_1}^\dagger \cdots \sum_{r_t,o_{t+1}} \widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}} \psi_t \Bigg) \\
&\Bigg[\sum_{a_t} b_{V,t}^\pi - \Big(r_t + \sum_{a_{t+1}} b_{V,t+1}^\pi \Big) \Bigg] \\
E_3 = &\sum_{t=1}^T \Bigg(\sum_{a_1} \pi_1 \psi_1^\top \widehat{\mathbf{P}}_{a_1}^\dagger \cdots \sum_{r_{t-1},o_t} \widehat{\mathbf{P}}_{a_{t-1},r_{t-1},o_t} \psi_{t-1} \Bigg[\sum_{a_t} b_{V,t}^\pi - \sum_{a_t} \pi_t \psi_t^\top \widehat{\mathbf{P}}_{a_t}^\dagger \sum_{r_t,o_{t+1}} \widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}} \psi_t \sum_{a_t} b_{V,t}^\pi \Bigg] \Bigg).
\end{aligned}
$$

It is straightforward to show $E_3 = 0$ by noticing that

$$
\begin{aligned}
&\sum_{a_t} \pi_t(a_t|o_t, h_{t-1}) \psi_t^\top(o_t) \widehat{\mathbf{P}}_{a_t}^\dagger \sum_{r_t,o_{t+1}} \widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}} \psi_t(o_t) \sum_{a_t} b_{V,t}^\pi(a_t, o_t, h_{t-1}) \\
&= \sum_{a_t} \pi_t(a_t|o_t, h_{t-1}) \psi_t^\top(o_t) \widehat{\mathbf{P}}_{a_t}^\dagger \widehat{\mathbf{P}}_{a_t} \psi_t(o_t) \sum_{a_t} b_{V,t}^\pi(a_t, o_t, h_{t-1}) \\
&= \psi_t^\top(o_t) \psi_t(o_t) \sum_{a_t} b_{V,t}^\pi(a_t, o_t, h_{t-1}) \\
&= \sum_{a_t} b_{V,t}^\pi(a_t, o_t, h_{t-1}).
\end{aligned}
$$

Throughout the proof we use the trick that

$$
\begin{aligned}
&\sum_{a_t} b_{V,t}^\pi(a_t, o_t, h_{t-1}) - \sum_{a_t} \pi_t(a_t|o_t, h_{t-1}) \psi_t^\top \widehat{\mathbf{P}}_{a_t}^\dagger \sum_{r_t,o_{t+1}} \widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}} \psi_t \big(r_t + \sum_{a_{t+1}} b_{V,t+1}^\pi(a_{t+1}, o_{t+1}, h_t) \big) \\
= &\sum_{a_t} \pi_t(a_t|o_t, h_{t-1}) \psi_t^\top \widehat{\mathbf{P}}_{a_t}^\dagger \sum_{r_t,o_{t+1}} \widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}} \psi_t \Big[\sum_{a_t} b_{V,t}^\pi(a_t, o_t, h_{t-1}) - \big(r_t + \sum_{a_{t+1}} b_{V,t+1}^\pi(a_{t+1}, o_{t+1}, h_t) \big) \Big].
\end{aligned}
$$

Thus, it suffices to bound $\mathcal{E}(E_1)$ and $\mathcal{E}(E_2)$ separately.

C.1.1 Bounding $\mathcal{E}(E_1)$

Note that $\mathcal{E}(E_1) := \sum_{t=1}^T \mathcal{E}(I_t)$ as the sum of T terms where

$$
\mathcal{E}(I_1) = \mathcal{E} \Bigg\{ \sum_{a_1} \frac{\pi_1(a_1|O_1)}{\pi_1^b(a_1|S_1)} \pi_1^b(a_1|S_1) \psi_1^\top \mathbf{P}_{a_1}^\dagger \sum_{r_1,o_2} \widehat{\mathbf{P}}_{a_1,r_1,o_2} \psi_1 \Big[\sum_{a_1} b_{V,1}^\pi(a_1, O_1) - \Big(r_1 + \sum_{a_2} b_{V,2}^\pi \big(a_2, o_2, (O_1, a_1)\big) \Big) \Big] \Bigg\},
$$

$$
\begin{aligned}
&= \mathcal{E}\Bigg\{ \frac{\pi_1(A_1|O_1)}{\pi_1^b(A_1|S_1)} \psi_1^\top \mathbf{P}_{A_1}^\dagger \sum_{r_1,o_2} \widehat{\mathbf{P}}_{A_1,r_1,o_2} \psi_1 \Big[\sum_{a_1} b_{V,1}^\pi(a_1, O_1) - \Big(r_1 + \sum_{a_2} b_{V,2}^\pi(a_2, o_2, H_1) \Big) \Big] \Bigg\}, \\
\mathcal{E}(I_2) &= \mathcal{E}\Bigg\{ \sum_{a_1} \frac{\pi_1(a_1|O_1)}{\pi_1^b(a_1|S_1)} \pi_1^b(a_1|S_1) \psi_1^\top \mathbf{P}_{a_1}^\dagger \sum_{r_1,o_2} \mathbf{P}_{a_1,r_1,o_2} \psi_1 \Big(\sum_{a_2} \frac{\pi_2(a_2|o_2, O_1, a_1)}{\pi_2^b(a_2|S_2)} \pi_2^b(a_2|S_2) \\
&\qquad \psi_2^\top \mathbf{P}_{a_2}^\dagger \sum_{r_2,o_3} \widehat{\mathbf{P}}_{a_2,r_2,o_3} \psi_2 \Big[\sum_{a_2} b_{V,2}^\pi(a_2, o_2, (O_1, a_1)) - \Big(r_2 + \sum_{a_2} b_{V,3}^\pi(a_3, o_3, (O_1, a_1, o_2, a_2)) \Big) \Big] \Big) \Bigg\} \\
&= \mathcal{E}\Bigg\{ \frac{\pi_1}{\pi_1^b} \psi_1^\top \mathbf{P}_{A_1}^\dagger \mathbf{P}_{A_1} \psi_1 \Big(\frac{\pi_2}{\pi_2^b} \psi_2^\top \mathbf{P}_{A_2}^\dagger \sum_{r_2,o_3} \widehat{\mathbf{P}}_{A_2,r_2,o_3} \psi_2 \Big[\sum_{a_2} b_{V,2}^\pi(a_2, O_2, H_1) - \Big(r_2 + \sum_{a_2} b_{V,3}^\pi(a_3, o_3, H_2) \Big) \Big] \Big) \Bigg\} \\
&= \mathcal{E}\Bigg\{ \frac{\pi_1(A_1|O_1)\pi_2(A_2|O_2, H_1)}{\pi_1^b(A_1|S_1)\pi_2^b(A_2|S_2)} \psi_2^\top \mathbf{P}_{A_2}^\dagger \sum_{r_2,o_3} \widehat{\mathbf{P}}_{A_2,r_2,o_3} \psi_2 \Big[\sum_{a_2} b_{V,2}^\pi(a_2, O_2, H_1) - \Big(r_2 + \sum_{a_3} b_{V,3}^\pi(a_3, o_3, H_2) \Big) \Big] \Bigg\}, \\
&\cdots \\
\mathcal{E}(I_T) &= \mathcal{E}\Bigg\{ \Big(\prod_{t=1}^T \frac{\pi_t}{\pi_t^b} \Big) \psi_T^\top \mathbf{P}_{A_T}^\dagger \sum_{r_T} \widehat{\mathbf{P}}_{A_T,r_T} \psi_T \; r_T \Bigg\}.
\end{aligned}
$$

Here, we simplify the expression by rewriting the summation over a_t, o_t as the expectation of A_t, O_t, and use the fact that $A_t \perp\!\!\!\perp O_t \mid S_t$.

First, we need to verify that the expectation $\mathbb{E}[\mathcal{E}(E_1)] = 0$.

For $\mathcal{E}(I_1)$, it suffices to show $\mathbb{E}[\widehat{\mathbf{P}}_{a_1,r_1,o_2}] = \mathbf{P}_{a_1,r_1,o_2}$. Note that for each entry of $\widehat{\mathbf{P}}_{a_1,r_1,o_2}$, the empirical transition probability is an unbiased estimator of its population transition probability (Lemma 12).

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{1}\big\{(o_2^i, o_1^i, r_1^i, o_0^i, a_1^i) = (o_2, o_1, r_1, o_0, a_1)\big\}}{\sum_{i=1}^n \mathbb{1}\big\{(o_0^i, a_1^i) = (o_0, a_1)\big\}} \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{1}\big\{(o_2^i, o_1^i, r_1^i, o_0^i, a_1^i) = (o_2, o_1, r_1, o_0, a_1)\big\}}{\sum_{i=1}^n \mathbb{1}\big\{(o_0^i, a_1^i) = (o_0, a_1)\big\}} \;\Big|\; \{o_0^{(i)}, a_1^{(i)}\}_{i=1}^n \right] \right] \\
&= \mathbb{E}\left[\frac{n_{o_0,a_1} p(o_2, o_1, r_1 \mid o_0, a_1)}{n_{o_0,a_1}} \;\Big|\; \{o_0^{(i)}, a_1^{(i)}\}_{i=1}^n \right] \\
&= p(o_2, o_1, r_1 \mid o_0, a_1).
\end{aligned}
$$

Thus, $\mathbb{E}[\widehat{\mathbf{P}}_{a_1,r_1,o_2}] = \mathbf{P}_{a_1,r_1,o_2}$, which yields

$$
\begin{aligned}
&\mathbb{E}\big[\mathcal{E}(I_1)\big] \\
&= \mathbb{E}\Bigg\{ \mathcal{E}\Bigg(\frac{\pi_1}{\pi_1^b} \psi_1^\top \mathbf{P}_{A_1}^\dagger \sum_{r_1,o_2} \widehat{\mathbf{P}}_{A_1,r_1,o_2} \psi_1 \Big[\sum_{a_1} b_{V,1}^\pi(a_1, O_1) - \Big(r_1 + \sum_{a_2} b_{V,2}^\pi(a_2, o_2, H_1) \Big) \Big] \Bigg) \Bigg\} \\
&= \mathbb{E}\Bigg\{ \frac{\pi_1}{\pi_1^b} \psi_1^\top \mathbf{P}_{A_1}^\dagger \sum_{r_1,o_2} \mathbf{P}_{A_1,r_1,o_2} \psi_1 \Big[\sum_{a_1} b_{V,1}^\pi(a_1, O_1) - \Big(r_1 + \sum_{a_2} b_{V,2}^\pi(a_2, o_2, H_1) \Big) \Big] \Bigg\} \\
&= \mathbb{E}\Bigg\{ \sum_{a_1} \pi_1(a_1|O_1) \psi_1^\top \mathbf{P}_{a_1}^\dagger \sum_{r_1,o_2} \mathbf{P}_{a_1,r_1,o_2} \psi_1 \Big[\sum_{a_1} b_{V,1}^\pi(a_1, O_1) - \Big(r_1 + \sum_{a_2} b_{V,2}^\pi(a_2, o_2, a_1, O_1) \Big) \Big] \Bigg\} \\
&= \mathbb{E}\Bigg\{ \sum_{a_1} b_{V,1}^\pi(a_1, O_1) - \sum_{a_1} b_{V,1}^\pi(a_1, O_1) \Bigg\} = 0.
\end{aligned}
$$

For $\mathcal{E}(I_t), t > 1$, we can show the similar results that $\mathbb{E}[\widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}}] = \mathbf{P}_{a_t,r_t,o_{t+1}}$. Thus, $\mathbb{E}[\mathcal{E}(E_1)] = 0$.

Then, it suffices to derive the bound for the variance of $\mathcal{E}(E_1)$. By applying the law of total variance (Lemma 3), we have

$$
\mathrm{Var}\Big[\mathcal{E}(E_1)\Big]
$$

$$
\begin{aligned}
=&\text{Var}\Bigg\{\sum_{t=1}^{T}\mathcal{E}\Bigg(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^{b}}\psi_t^\top\mathbf{P}_{A_t}^\dagger\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t}b_{V,t}^{\pi}-\big(r_t+\sum_{a_{t+1}}b_{V,t+1}^{\pi}\big)\Big]\Bigg)\Bigg\}\\
=&\sum_{t=1}^{T}\mathbb{E}\Bigg\{\text{Var}\Bigg[\mathcal{E}\Bigg(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^{b}}\psi_t^\top\mathbf{P}_{A_t}^\dagger\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t}b_{V,t}^{\pi}-\big(r_t+\sum_{a_{t+1}}b_{V,t+1}^{\pi}\big)\Big]\Bigg)\,\Big|\,\mathcal{D}_t\Bigg]\Bigg\}.
\end{aligned}
$$

Next we consider bounding $\text{Var}\big[\mathcal{E}(E_1)\big]$ under different conditions.

(1) Bounding $\text{Var}\big[\mathcal{E}(E_1)\big]$ without considering weight function Note that

$$
\begin{aligned}
&\text{Var}\Big[\mathcal{E}(E_1)\Big]\\
=&\sum_{t=1}^{T}\mathbb{E}\Bigg\{\text{Var}\Bigg[\mathcal{E}\Bigg(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^{b}}\psi_t^\top\mathbf{P}_{A_t}^\dagger\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t}b_{V,t}^{\pi}-\big(r_t+\sum_{a_{t+1}}b_{V,t+1}^{\pi}\big)\Big]\Bigg)\,\Big|\,\mathcal{D}_t\Bigg]\Bigg\}\\
\lesssim&T^2\sum_{t=1}^{T}\mathbb{E}\Bigg\{\text{Var}\Bigg[\mathcal{E}\Bigg(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^{b}}\psi_t^\top\mathbf{P}_{A_t}^\dagger\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Bigg)\,\Big|\,\mathcal{D}_t\Bigg]\Bigg\}\\
=&T^2\sum_{t=1}^{T}\mathbb{E}\Bigg\{\mathcal{E}\Bigg(\prod_{t'=1}^{t}\Big(\frac{\pi_{t'}}{\pi_{t'}^{b}}\Big)^2\text{Var}\Bigg[\psi_t^\top\mathbf{P}_{A_t}^\dagger\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\,\Big|\,\mathcal{D}_t\Bigg]\Bigg)\Bigg\}\\
\overset{(i)}{\leq}&T^2\sum_{t=1}^{T}\mathbb{E}\Bigg\{\mathcal{E}\Bigg(\prod_{t'=1}^{t}\Big(\frac{\pi_{t'}}{\pi_{t'}^{b}}\Big)^2\psi_t^\top\mathbf{P}_{A_t}^\dagger\sum_{r_t,o_{t+1}}\text{Cov}\Bigg[\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\,\Big|\,\mathcal{D}_t\Bigg](\mathbf{P}_{A_t}^\dagger)^\top\psi_t\Bigg)\Bigg\}\\
=&T^2\sum_{t=1}^{T}\mathbb{E}\Bigg\{\mathcal{E}\Bigg(\prod_{t'=1}^{t}\Big(\frac{\pi_{t'}}{\pi_{t'}^{b}}\Big)^2\psi_t^\top\mathbf{P}_{A_t}^\dagger\sum_{r_t,o_{t+1}}\text{diag}(\mathbf{1}\{E_{o_0,h_{t-1}}\})\text{Cov}\Bigg[\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\,\Big|\,\mathcal{D}_t\Bigg](\mathbf{P}_{A_t}^\dagger)^\top\psi_t\Bigg)\Bigg\}\\
\overset{(ii)}{\leq}&T^2\sum_{t=1}^{T}\mathcal{E}\Bigg(\prod_{t'=1}^{t}\Big(\frac{\pi_{t'}}{\pi_{t'}^{b}}\Big)^2\Big|\psi_t^\top\mathbf{P}_{A_t}^\dagger\mathbb{E}\Big[\text{diag}(\mathbf{1}\{E_{o_0,h_{t-1}}\}/n_{o_0,A_t,h_{t-1}})\text{diag}(\mathbf{P}_{A_t}\psi_t)\Big](\mathbf{P}_{A_t}^\dagger)^\top\psi_t\Big|\Bigg)\Bigg\}\\
\leq&\frac{T^2}{n}(1-\theta)^{-1}\sum_{t=1}^{T}\mathbb{E}\Bigg\{\prod_{t'=1}^{t}\Big(\frac{\pi_{t'}}{\pi_{t'}^{b}}\Big)^2\Big|\psi_t^\top\mathbf{P}_{A_t}^\dagger\text{diag}(\mathbf{P}_{A_t}\psi_t)(\mathbf{P}_{A_t}^\dagger)^\top\psi_t\Big|\Bigg\}\\
\leq&\frac{T^2}{n}(1-\theta^*)^{-1}C_{\pi^b}\sum_{t=1}^{T}|\mathcal{A}|\big\|\psi_t\big\|_2^2\,\big\|\mathbf{P}_{a_t}^\dagger\big\|_2^2\,\big\|\text{diag}\big(\mathbf{P}_{a_t}\psi_t\big)\big\|_2\\
\overset{(iii)}{\leq}&\frac{T^2}{n}(1-\theta^*)^{-1}C_{\pi^b}C_P^2\sum_{t=1}^{T}|\mathcal{A}||\mathcal{O}||\mathcal{H}_{t-1}|\\
=&\frac{T^2}{n}(1-\theta^*)^{-1}C_P^2C_{\pi^b}\sum_{t=1}^{T}|\mathcal{O}|^t\,|\mathcal{A}|^t\\
\leq&\frac{T^3}{n}(1-\theta^*)^{-1}C_P^2C_{\pi^b}|\mathcal{O}|^T\,|\mathcal{A}|^T.
\end{aligned}
$$

The inequality (i) follows from the independence of each $\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}$, since $\{r_t^{(i)},o_{t+1}^{(i)}\}_{i=1}^n$ partitions the n episodes into disjoint sets according to all combinations of r_t and o_{t+1}. We define $E_{o_0,h_{t-1}}=\{n_{o_0,A_t,h_{t-1}}>np_t^{\pi^b}(1-\theta_{t,ij})\}$, and $\mathbf{1}\{E_{o_0,h_{t-1}}\}$ is a vector of length $|\mathcal{O}||\mathcal{H}_{t-1}|$, where each element corresponds to $E_{o_0,h_{t-1}}$. The inequality (ii) utilizes the fact that for any $a_t\in\mathcal{A}$

$$
\begin{aligned}
&\text{Cov}\Big[\widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}}\psi_t\mid\mathcal{D}_t\Big]\\
\leq&\,\text{diag}(\mathbf{P}_{a_t,r_t,o_{t+1}}\psi_t/n_{o_0,a_t,h_{t-1}})-\mathbf{P}_{a_t,r_t,o_{t+1}}\psi_t\psi_t^\top\mathbf{P}_{a_t,r_t,o_{t+1}}^\top/n_{o_0,a_t,h_{t-1}}\\
\preceq&\,\text{diag}(\mathbf{P}_{a_t,r_t,o_{t+1}}\psi_t/n_{o_0,a_t,h_{t-1}}).
\end{aligned}
\tag{22}
$$

The inequality (iii) follows from that

$$\left\|\text{diag}\left(\mathbf{P}_{a_t}\psi_t\right)\right\|_2 \leq 1, \text{ and } \left\|\mathbf{P}_{a_t}^{\dagger}\right\|_2 \leq \frac{1}{\sigma_{min}(\mathbf{P}_{a_t})} \leq C_P\sqrt{|\mathcal{O}||\mathcal{H}_{t-1}|}.$$

Therefore, we obtain the upper bound of $\mathcal{E}(E_1)$

$$\mathcal{E}(E_1) = \mathcal{O}_P\left(\frac{T^{1.5}}{\sqrt{n}}(1-\theta^*)^{-1/2}C_P C_{\pi^b}^{\frac{1}{2}}|\mathcal{O}|^{\frac{T}{2}}|\mathcal{A}|^{\frac{T}{2}}\right)$$

(2) Bounding $\mathcal{E}(E_1)$ with the weight function

$$\begin{aligned}
&\text{Var}\Big[\mathcal{E}(E_1)\Big] \\
=&\sum_{t=1}^{T}\mathbb{E}\Bigg\{\text{Var}\Bigg[\mathcal{E}\Bigg(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\psi_t^{\top}\mathbf{P}_{A_t}^{\dagger}\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t}b_{V,t}^{\pi} - \big(r_t + \sum_{a_{t+1}}b_{V,t+1}^{\pi}\big)\Big]\Bigg)\Bigg|\mathcal{D}_t\Bigg]\Bigg\} \\
\leq&\frac{C_W C_P^2(1-\theta^*)^{-1}}{n}\sum_{t=1}^{T}|\mathcal{O}||\mathcal{H}_{t-1}||\mathcal{A}| \\
&\mathbb{E}^{\pi}\Bigg\{\text{Var}\Bigg[\sum_{a_t}b_{V,t}^{\pi}(a_t,O_t,H_{t-1}) - \Big(R_t + \sum_{a_{t+1}}b_{V,t+1}^{\pi}(a_{t+1},O_{t+1},H_t)\Big)\Bigg|O_t,S_t,A_t,H_{t-1}\Bigg]\Bigg\} \\
&\text{(by Equation (24))} \\
=&\frac{C_W C_P^2(1-\theta^*)^{-1}}{n}|\mathcal{O}|^T|\mathcal{A}|^T\sum_{t=1}^{T}\mathbb{E}^{\pi}\Bigg\{\text{Var}\Bigg[R_t + \sum_{a}b_{V,t+1}^{\pi}(a,O_{t+1},H_t)\Bigg|O_t,S_t,A_t,H_{t-1}\Bigg]\Bigg\} \\
\leq&\frac{C_W C_P^2(1-\theta^*)^{-1}}{n}|\mathcal{O}|^T|\mathcal{A}|^T\,\text{Var}^{\pi}\Big[\sum_{t=1}^{T}R_t\Big] \quad \text{(by Lemma 1)} \\
\leq&\frac{T^2}{n}C_W C_P^2(1-\theta^*)^{-1}|\mathcal{O}|^T|\mathcal{A}|^T.
\end{aligned}$$

For the first inequality, we examine the cases for $t = 1, 2$ separately in the following. The result for $t > 2$ follows by a similar argument. Thus, incorporating the ratio function, we obtain

$$\mathcal{E}(E_1) = \mathcal{O}_P\left(\frac{T}{\sqrt{n}}C_W^{\frac{1}{2}}C_P(1-\theta^*)^{-\frac{1}{2}}|\mathcal{O}|^{\frac{T}{2}}|\mathcal{A}|^{\frac{T}{2}}\right)$$

- For $t = 1$, we have

$$\begin{aligned}
&\mathbb{E}\Bigg\{\text{Var}\Bigg[\mathcal{E}\Bigg(\frac{\pi_1(A_1|O_1)}{\pi_1^b(A_1|S_1)}\psi_1^{\top}\mathbf{P}_{A_1}^{\dagger}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{A_1,r_1,o_2}\psi_1\Big[\sum_{a_1}b_{V,1}^{\pi} - \Big(r_1 + \sum_{a_2}b_{V,2}^{\pi}\Big)\Big]\Bigg)\Bigg|\mathcal{D}_1\Bigg]\Bigg\} \\
=&\mathbb{E}\Bigg\{\text{Var}\Bigg[\mathcal{E}\Bigg(\mathcal{E}\Big(b_{W,1}^{\pi}(A_1,O_0)\Big|A_1,S_1\Big)\pi_1(A_1|O_1) \\
&\psi_1^{\top}\mathbf{P}_{A_1}^{\dagger}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{A_1,r_1,o_2}\psi_1\Big[\sum_{a_1}b_{V,1}^{\pi}(a_1,O_1) - \Big(r_1 + \sum_{a_2}b_{V,2}^{\pi}(a_2,o_2,H_1)\Big)\Big]\Bigg)\Bigg|\mathcal{D}_1\Bigg]\Bigg\} \\
=&\mathbb{E}\Bigg\{\text{Var}\Bigg[\mathcal{E}\Bigg(b_{W,1}^{\pi}(A_1,O_0)\pi_1(A_1|O_1) \\
&\psi_1^{\top}\mathbf{P}_{A_1}^{\dagger}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{A_1,r_1,o_2}\psi_1\Big[\sum_{a_1}b_{V,1}^{\pi}(a_1,O_1) - \Big(r_1 + \sum_{a_2}b_{V,2}^{\pi}(a_2,o_2,H_1)\Big)\Big]\Bigg)\Bigg|\mathcal{D}_1\Bigg]\Bigg\}
\end{aligned}$$

$$
\begin{aligned}
&\text{(by } O_0 \perp\!\!\!\perp O_1 \mid S_1, A_1) \\
\leq &\mathbb{E}\Bigg\{\mathcal{E}\Bigg(\Big(b^{\pi}_{W,1}(A_1, O_0)\pi_1(A_1|O_1)\Big)^2 \\
&\psi_1^\top \mathbf{P}^\dagger_{A_1} \sum_{r_1,o_2} \mathrm{Cov}\Big[\widehat{\mathbf{P}}_{A_1,r_1,o_2}\psi_1 \;\Big|\; \mathcal{D}_1\Big]\Big[\sum_{a_1} b^{\pi}_{V,1}(a_1, O_1) - \Big(r_1 + \sum_{a_2} b^{\pi}_{V,2}(a_2, o_2, H_1)\Big)\Big]^2 (\mathbf{P}^\dagger_{A_1})^\top \psi_1\Bigg)\Bigg\} \\
\overset{(i)}{\leq} &\mathcal{E}\Bigg\{\Big(b^{\pi}_{W,1}(A_1, O_0)\pi_1(A_1|O_1)\Big)^2 \\
&\psi_1^\top \mathbf{P}^\dagger_{A_1} \mathbb{E}\Big[\mathrm{diag}(\frac{\mathbf{1}\{E_{o_0,h_{t-1}}\}}{n_{o_0,A_t,h_{t-1}}})\mathrm{diag}\big(\mathbf{P}_{A_1}\psi_1\big)\Big](\mathbf{P}^\dagger_{aA_1})^\top \psi_1 \Big[\sum_{a_1} b^{\pi}_{V,1}(a_1, O_1) - \Big(R_1 + \sum_{a_2} b^{\pi}_{V,2}(a_2, O_2, H_1)\Big)\Big]^2\Bigg\} \\
\leq &\frac{(1-\theta^*)^{-1}}{n}|\mathcal{A}|\,\mathbb{E}\Bigg\{\Big(b^{\pi}_{W,1}(A_1, O_0)\pi_1(A_1|O_1)\Big)^2 \\
&\psi_1^\top \mathbf{P}^\dagger_{a_1} \mathrm{diag}\big(\mathbf{P}_{a_1}\psi_1\big)(\mathbf{P}^\dagger_{a_1})^\top \psi_1 \Big[\sum_{a_1} b^{\pi}_{V,1}(a_1, O_1) - \Big(R_1 + \sum_{a_2} b^{\pi}_{V,2}(a_2, O_2, H_1)\Big)\Big]^2\Bigg\} \\
\overset{(ii)}{\leq} &\frac{C_P^2(1-\theta^*)^{-1}}{n}|\mathcal{O}||\mathcal{A}| \\
&\mathbb{E}\Bigg\{b^{\pi}_{W,1}(A_1, O_0)\pi_1(A_1|O_1)\Big[\sum_{a_1} b^{\pi}_{V,1}(a_1, O_1) - \Big(R_1 + \sum_{a_2} b^{\pi}_{V,2}(a_2, O_2, H_1)\Big)\Big]\Bigg\}^2 \\
= &\frac{C_P^2(1-\theta^*)^{-1}}{n}|\mathcal{O}||\mathcal{A}| \\
&\mathbb{E}\Bigg\{\mathrm{Var}\Bigg[b^{\pi}_{W,1}(A_1, O_0)\pi_1(A_1|O_1)\Big[\sum_{a_1} b^{\pi}_{V,1}(a_1, O_1) - \Big(R_1 + \sum_{a_2} b^{\pi}_{V,2}(a_2, O_2, H_1)\Big)\Big] \;\Big|\; O_1, S_1, A_1\Bigg]\Bigg\}
\end{aligned}
$$

The inequality (i) holds by applying the same argument as in the previous part that bounding $\mathrm{Var}[\mathcal{E}(E_1)]$ without using the weight function. The inequality (ii) holds since that, for any $t \in [T]$,

$$
\begin{aligned}
\Big|\psi_t^\top \mathbf{P}^\dagger_{a_t}\, \mathrm{diag}\big(\mathbf{P}_{a_t}\psi_t\big)\, (\mathbf{P}^\dagger_{a_t})^\top \psi_t\Big| &\leq \big\|\psi_t\big\|_2^2 \big\|\mathbf{P}^\dagger_{a_t}\big\|_2^2 \big\|\mathrm{diag}\big(\mathbf{P}_{a_t}\psi_t\big)\big\|_2 \\
&\leq C_P^2|\mathcal{O}||\mathcal{H}_{t-1}|.
\end{aligned} \tag{23}
$$

The last equality follows from the zero mean property that

$$
\begin{aligned}
&\mathbb{E}\Bigg\{b^{\pi}_{W,t}(A_t, O_0, H_{t-1})\pi_t(A_t|O_t, H_{t-1})\Big[\sum_{a_t} b^{\pi}_{V,1}(a_t, O_t, H_{t-1}) - \Big(R_t + \sum_{a_{t+1}} b^{\pi}_{V,t+1}(a_{t+1}, O_{t+1}, H_t)\Big)\Big]\Bigg\} \\
= &\mathbb{E}\Bigg\{b^{\pi}_{W,t}(A_t, O_0, H_{t-1})\mathbb{E}\Bigg(\pi_t(A_t|O_t, H_{t-1}) \\
&\Big[\sum_{a_t} b^{\pi}_{V,1}(a_t, O_t, H_{t-1}) - \Big(R_t + \sum_{a_{t+1}} b^{\pi}_{V,t+1}(a_{t+1}, O_{t+1}, H_t)\Big)\Big] \;\Big|\; O_0, A_t, H_{t-1}\Bigg)\Bigg\} \\
= &\mathbb{E}\Bigg\{b^{\pi}_{W,t}(A_t, O_0, H_{t-1})\mathbb{E}\Bigg(\pi_t(A_t|O_t, H_{t-1})\sum_{a_t} b^{\pi}_{V,1}(a_t, O_t, H_{t-1}) - b^{\pi}_{V,1}(A_t, O_t, H_{t-1}) \;\Big|\; O_0, A_t, H_{t-1}\Bigg)\Bigg\} \\
&\text{(by Assumption 4)} \\
= &\mathbb{E}\Bigg\{b^{\pi}_{W,t}(A_t, O_0, H_{t-1})\Bigg(\pi_t(A_t|O_t, H_{t-1})\sum_{a_t} b^{\pi}_{V,1}(a_t, O_t, H_{t-1}) - b^{\pi}_{V,1}(A_t, O_t, H_{t-1})\Bigg)\Bigg\} \\
= &\mathbb{E}\Bigg\{\frac{\omega_t(S_t, H_{t-1})}{\pi_t^b(A_t|S_t)}\mathbb{E}\Bigg\{\sum_{a_t} b^{\pi}_{V,1}(a_t, O_t, H_{t-1})\pi_t(A_t|O_t, H_{t-1}) - b^{\pi}_{V,1}(A_t, O_t, H_{t-1}) \;\Big|\; S_t, A_t, O_t, H_{t-1}\Bigg\}\Bigg\}
\end{aligned}
$$

(by Assumption 3)

$$
\begin{aligned}
=&\mathbb{E}\Bigg\{ \sum_a \pi_t^b(a|S_t)\frac{\omega_t(S_t,H_{t-1})}{\pi_t^b(a|S_t)}\mathbb{E}\Bigg\{ \sum_{a_t} b^\pi_{V,1}(a_t,O_t,H_{t-1})\pi_t(a|O_t,H_{t-1}) - b^\pi_{V,1}(a,O_t,H_{t-1}) \\
&\quad \Big|\, S_t, A_t=a, O_t, H_{t-1}\Bigg\}\Bigg\} \\
=&\mathbb{E}\Bigg\{\omega_t(S_t,H_{t-1})\mathbb{E}\Bigg\{ \sum_{a_t} b^\pi_{V,1}(a_t,O_t,H_{t-1}) - \sum_a b^\pi_{V,1}(a,O_t,H_{t-1}) \,\Big|\, S_t, A_t=a, O_t, H_{t-1}\Bigg\}\Bigg\} \\
=&0.
\end{aligned}
$$

For the last equation, we further have

$$
\begin{aligned}
&\mathbb{E}\Bigg\{ \mathrm{Var}\Bigg[b^\pi_{W,1}(A_1,O_0)\pi_1(A_1|O_1)\Big[\sum_{a_1} b^\pi_{V,1}(a_1,O_1) - \Big(R_1 + \sum_{a_2} b^\pi_{V,2}(a_2,O_2,H_1)\Big)\Big] \,\Big|\, O_1,S_1,A_1\Bigg]\Bigg\} \\
=&\mathbb{E}\Bigg\{ \mathrm{Var}\Bigg[\mathbb{E}\Big(b^\pi_{W,1}(A_1,O_0)\pi_1(A_1|O_1) \,\Big|\, O_1,S_1,A_1\Big) \\
&\quad \Big[\sum_{a_1} b^\pi_{V,1}(a_1,O_1) - \Big(R_1 + \sum_{a_2} b^\pi_{V,2}(a_2,O_2,H_1)\Big)\Big] \,\Big|\, O_1,S_1,A_1\Bigg]\Bigg\} \quad (\text{by } O_0 \perp\!\!\!\perp O_1 \mid S_1, A_1) \\
\overset{(i)}{\leq}& C_W\, \mathbb{E}\Bigg\{ \mathbb{E}\Big(b^\pi_{W,1}(A_1,O_0)\pi_1(A_1|O_1) \,\Big|\, O_1,S_1,A_1\Big) \\
&\quad \mathrm{Var}\Bigg[\sum_{a_1} b^\pi_{V,1}(a_1,O_1) - \Big(R_1 + \sum_{a_2} b^\pi_{V,2}(a_2,O_2,H_1)\Big) \,\Big|\, O_1,S_1,A_1\Bigg]\Bigg\} \quad (\text{by Assumption 3 }) \\
=&C_W\, \mathbb{E}\Bigg\{ \frac{p_1^\pi(S_1)\pi_1(A_1|O_1)p(O_1|S_1)}{p_1^{\pi^b}(S_1)\pi_1^b(A_1|S_1)p(O_1|S_1)} \mathrm{Var}\Bigg[\sum_{a_1} b^\pi_{V,1}(a_1,O_1) - \Big(R_1 + \sum_{a_2} b^\pi_{V,2}(a_2,O_2,H_1)\Big) \,\Big|\, O_1,S_1,A_1\Bigg]\Bigg\} \\
&\quad \Bigg(\text{by } \mathbb{E}\to\mathbb{E}^\pi: \quad \frac{p_t^\pi(a_t,o_t,s_t,h_{t-1})}{p_t^{\pi^b}(a_t,o_t,s_t,h_{t-1})} = \frac{p_t^\pi(s_t,h_{t-1})\pi_t(a_t|o_t,h_{t-1})p(o_t|s_t)}{p_t^{\pi^b}(s_t,h_{t-1})\pi_t^b(a_t|s_t)p(o_t|s_t)}\Bigg) \\
=&C_W\, \mathbb{E}^\pi\Bigg\{ \mathrm{Var}\Bigg[\sum_{a_1} b^\pi_{V,1}(a_1,O_1) - \Big(R_1 + \sum_{a_2} b^\pi_{V,2}(a_2,O_2,H_1)\Big) \,\Big|\, O_1,S_1,A_1\Bigg]\Bigg\}
\end{aligned}
$$

where the inequality (i) follows from the Assumption 3, which states the ratio function satisfies

$$
\sup_{t,o_t,s_t,a_t,h_{t-1}} \mathbb{E}\Big[b^\pi_{W,t}(a_t,h_{t-1},O_0)\pi_t(a_t \mid o_t,h_{t-1}) \,\Big|\, o_t,s_t,a_t,h_{t-1}\Big] \leq C_W.
$$

Thus, we have

$$
\begin{aligned}
&\mathbb{E}\Bigg\{ \mathrm{Var}\Bigg[\mathcal{E}\Bigg(\frac{\pi_1(A_1|O_1)}{\pi_1^b(A_1|S_1)} \psi_1^\top \mathbf{P}_{A_1}^\dagger \sum_{r_1,o_2} \widehat{\mathbf{P}}_{A_1,r_1,o_2}\psi_1 \Big[\sum_{a_1} b^\pi_{V,1}(a_1,O_1) - \Big(r_1 + \sum_{a_2} b^\pi_{V,2}(a_2,o_2,H_1)\Big)\Big]\Bigg) \,\Big|\, \mathcal{D}_1\Bigg]\Bigg\} \\
\leq& \frac{C_W C_P^2 (1-\theta^*)^{-1}}{n} |\mathcal{O}||\mathcal{A}|\ \mathbb{E}^\pi\Bigg\{ \mathrm{Var}\Bigg[\sum_{a_1} b^\pi_{V,1}(a_1,O_1) - \Big(R_1 + \sum_{a_2} b^\pi_{V,2}(a_2,O_2,H_1)\Big) \,\Big|\, O_1,S_1,A_1\Bigg]\Bigg\}.
\end{aligned}
$$

- For $t=2$, by applying the similar arguments, we have

$$
\mathbb{E}\Bigg\{ \mathrm{Var}\Bigg[\mathcal{E}\Bigg(\frac{\pi_1}{\pi_1^b}\frac{\pi_2}{\pi_2^b} \psi_2^\top \mathbf{P}_{A_2}^\dagger \sum_{r_2,o_3} \widehat{\mathbf{P}}_{A_2,r_2,o_3}\psi_2 \Big[\sum_{a_2} b^\pi_{V,2}(a_2,O_2,H_1) - \Big(r_2 + \sum_{a_3} b^\pi_{V,2}(a_3,o_3,H_2)\Big)\Big]\Bigg) \,\Big|\, \mathcal{D}_2\Bigg]\Bigg\}
$$

$$
\begin{aligned}
=&\mathbb{E}\Bigg\{\text{Var}\Bigg[\mathcal{E}\Bigg(\frac{p_1^\pi(S_1)\pi_1(A_1|O_1)}{p_1^{\pi^b}(S_1)\pi_1^b(A_1|S_1)}\sum_{a_2}\pi_2(a_2|O_2,H_1)\psi_2^\top\mathbf{P}_{a_2}^\dagger\sum_{r_2,o_3}\widehat{\mathbf{P}}_{a_2,r_2,o_3}\psi_2 \\
&\Bigg[\sum_{a_2}b_{V,2}^\pi(a_2,O_2,H_1)-\Big(r_2+\sum_{a_3}b_{V,2}^\pi(a_3,o_3,(O_2,a_2,H_1))\Big)\Bigg]\Bigg)\;\Big|\;\mathcal{D}_2\Bigg]\Bigg\} \\
=&\mathbb{E}\Bigg\{\text{Var}\Bigg[\mathcal{E}\Bigg(\frac{p_2^\pi(S_2,H_1)\pi_2(A_2|O_2,H_1)}{p_2^{\pi^b}(S_2,H_1)\pi_2^b(A_2|S_2)}\psi_2^\top\mathbf{P}_{A_2}^\dagger\sum_{r_2,o_3}\widehat{\mathbf{P}}_{A_2,r_2,o_3}\psi_2 \\
&\Bigg[\sum_{a_2}b_{V,2}^\pi(a_2,o_2,H_1)-\Big(r_2+\sum_{a_3}b_{V,2}^\pi(a_3,o_3,(O_2,A_2,H_1))\Big)\Bigg]\Bigg)\;\Big|\;\mathcal{D}_2\Bigg]\Bigg\}\qquad\text{(by Lemma 4)} \\
=&\mathbb{E}\Bigg\{\text{Var}\Bigg[\mathcal{E}\Bigg(b_{W,2}^\pi(A_2,O_0,H_1)\pi_2(A_2|O_2,H_1)\psi_2^\top\mathbf{P}_{A_2}^\dagger\sum_{r_2,o_3}\widehat{\mathbf{P}}_{A_2,r_2,o_3}\psi_2 \\
&\Bigg[\sum_{a_2}b_{V,2}^\pi(a_2,o_2,H_1)-\Big(r_2+\sum_{a_3}b_{V,2}^\pi(a_3,o_3,H_2)\Big)\Bigg]\Bigg)\;\Big|\;\mathcal{D}_2\Bigg]\Bigg\} \\
\overset{(i)}{\leq}&\frac{C_P^2(1-\theta^*)^{-1}}{n}|\mathcal{O}||\mathcal{H}_1||\mathcal{A}| \\
&\text{Var}\Bigg\{b_{W,2}^\pi(A_2,O_0,H_1)\pi_2(A_2|O_2,H_1)\Big[\sum_{a_2}b_{V,2}^\pi(a_2,o_2,H_1)-\Big(R_2+\sum_{a_3}b_{V,2}^\pi(a_3,O_3,H_2)\Big)\Big]\Bigg\} \\
=&\frac{C_P^2(1-\theta^*)^{-1}}{n}|\mathcal{O}|^2|\mathcal{A}|^2\,\mathbb{E}\Bigg\{\text{Var}\Bigg[\mathbb{E}\Big[b_{W,2}^\pi(A_2,O_0,H_1)\pi_2(A_2|O_2,H_1)\;\Big|\;O_2,S_2,A_2,H_1\Big] \\
&\Big[\sum_{a_2}b_{V,2}^\pi(a_2,O_2,H_1)-\Big(R_2+\sum_{a_3}b_{V,2}^\pi(a_3,O_3,H_2)\Big)\Big]\;\Big|\;O_2,S_2,A_2,H_1\Bigg]\Bigg\} \\
\overset{(ii)}{\leq}&\frac{C_WC_P^2(1-\theta^*)^{-1}}{n}|\mathcal{O}|^2|\mathcal{A}|^2\,\mathbb{E}\Bigg\{\mathbb{E}\Big[b_{W,2}^\pi(A_2,O_0,H_1)\pi_2(A_2|O_2,H_1)\;\Big|\;O_2,S_2,A_2,H_1\Big] \\
&\text{Var}\Bigg[\sum_{a_2}b_{V,2}^\pi(a_2,O_2,H_1)-\Big(r_2+\sum_{a_3}b_{V,2}^\pi(a_3,o_3,H_2)\Big)\;\Big|\;O_2,S_2,A_2,H_1\Bigg]\Bigg\} \\
=&\frac{C_WC_P^2(1-\theta^*)^{-1}}{n}|\mathcal{O}|^2|\mathcal{A}|^2\,\mathbb{E}\Bigg\{\frac{p_2^\pi(S_2,H_1)\pi_2(A_2\mid O_2,H_1)}{p_2^{\pi^b}(S_2,H_1)\pi_2^b(A_2\mid S_2)} \\
&\text{Var}\Bigg[\sum_{a_2}b_{V,2}^\pi(a_2,O_2,H_1)-\Big(R_2+\sum_{a_3}b_{V,2}^\pi(a_3,O_3,H_2)\Big)\;\Big|\;O_2,S_2,A_2,H_1\Bigg]\Bigg\} \\
=&\frac{C_WC_P^2(1-\theta^*)^{-1}}{n}|\mathcal{O}|^2|\mathcal{A}|^2\,\mathbb{E}^\pi\Bigg\{\text{Var}\Bigg[\sum_{a_2}b_{V,2}^\pi(a_2,O_2,H_1)-\Big(R_2+\sum_{a_3}b_{V,2}^\pi(a_3,O_3,H_2)\Big)\;\Big|\;O_2,S_2,A_2,H_1\Bigg]\Bigg\}
\end{aligned}
$$

The inequality (i) is derived using the same arguments as in the proof for the case $t = 1$ and equation (23). The inequality (ii) is a consequence of Assumption 3, as utilized in the proof for the case $t = 1$.

Therefore, for any $t \in [T]$, by the same reasoning, we have

$$\begin{aligned}
&\mathbb{E}\Bigg\{ \mathrm{Var}\Bigg[\mathcal{E}\Bigg(\prod_{t'=1}^{t} \frac{\pi_{t'}}{\pi^b_{t'}} \psi_t^\top \mathbf{P}^\dagger_{A_t} \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{A_t, r_t, o_{t+1}} \psi_t \Big[\sum_{a_t} b^\pi_{V,t} - \big(r_t + \sum_{a_{t+1}} b^\pi_{V,t+1} \big) \Big] \Bigg) \;\Big|\; \mathcal{D}_t \Bigg] \Bigg\} \\
\leq & \frac{C_W C_P^2 (1-\theta^*)^{-1}}{n} |\mathcal{O}|^2 |\mathcal{A}|^2 \\
& \mathbb{E}^\pi \Bigg\{ \mathrm{Var}\Bigg[\sum_{a_t} b^\pi_{V,t}(a_t, O_t, H_{t-1}) - \Big(R_t + \sum_{a_{t+1}} b^\pi_{V,t+1}(a_{t+1}, O_{t+1}, H_t) \Big) \;\Big|\; O_t, S_t, A_t, H_{t-1} \Bigg] \Bigg\}
\end{aligned} \tag{24}$$

C.1.2 Proof of Lemma 1

Proof. let's suppress the target policy π for simplicity. Denote $\widetilde{\mathcal{D}}_t = \{O_{1:t}, S_{1:t}, A_{1:t}, R_{1:t-1}\}$. First, note that

$$\begin{aligned}
&\mathbb{E}^\pi \Bigg[R_t + \sum_{a'} b^\pi_{V,t+1}(a', O_{t+1}, H_t) \;\Big|\; O_t, S_t, A_t, H_{t-1} \Bigg] \\
=& \sum_a \mathbb{E}\Bigg[R_t + \sum_{a'} b^\pi_{V,t+1}(a', O_{t+1}, H_t) \;\Big|\; O_t, S_t, A_t = a, H_{t-1} \Bigg] \pi_t(a \mid O_t, H_{t-1}) \\
=& \sum_a \mathbb{E}\Bigg[\Big(R_t + \sum_{a'} b^\pi_{V,t+1}(a', O_{t+1}, H_t) \Big) \pi_t(a \mid O_t, H_{t-1}) \;\Big|\; O_t, S_t, A_t = a, H_{t-1} \Bigg] \\
=& \mathbb{E}\Bigg[\mathbb{E}\Bigg[\sum_a \mathbb{E}\Bigg[\Big(R_t + \sum_{a'} b^\pi_{V,t+1}(a', O_{t+1}, H_t) \Big) \pi_t(a \mid O_t, H_{t-1}) \Big| O_t, S_t, A_t = a, H_{t-1} \Bigg] \Big| S_t, A_t = a, H_{t-1} \Bigg] \Bigg] \\
=& \mathbb{E}\Bigg[\sum_a \mathbb{E}\Bigg[\Big(R_t + \sum_{a'} b^\pi_{V,t+1}(a', O_{t+1}, H_t) \Big) \pi_t(a \mid O_t, H_{t-1}) \;\Big|\; S_t, A_t = a, H_{t-1} \Bigg] \Bigg] \\
& \text{(by the law of total expectation)} \\
=& \mathbb{E}\Bigg[\sum_a \mathbb{E}\Bigg[b^\pi_{V,t}(a, O_t, H_{t-1}) \;\Big|\; S_t, A_t = a, H_{t-1} \Bigg] \Bigg] \qquad \text{(by Equation (18))} \\
=& \sum_a b^\pi_{V,t}(a, O_t, H_{t-1})
\end{aligned} \tag{25}$$

Then, by iteratively applying the law of total variance, we have

$$\begin{aligned}
&\mathrm{Var}\Bigg[\sum_{t=1}^{h} R_t + \sum_a b^\pi_{V,h+1}(a, O_{h+1}, H_h) \Bigg] \\
=& \mathbb{E}\Bigg[\mathrm{Var}\Bigg[\sum_{t=1}^{h} R_t + \sum_a b^\pi_{V,h+1}(a, O_{h+1}, H_h) \;\Big|\; \widetilde{\mathcal{D}}_h \Bigg] \Bigg] + \mathrm{Var}\Bigg[\mathbb{E}\Bigg[\sum_{t=1}^{h} R_t + \sum_a b^\pi_{V,h+1}(a, O_{h+1}, H_h) \;\Big|\; \widetilde{\mathcal{D}}_h \Bigg] \Bigg] \\
=& \mathbb{E}\Bigg[\mathrm{Var}\Bigg[R_h + \sum_a b^\pi_{V,h+1}(a, O_{h+1}, H_h) \;\Big|\; O_h, S_h, A_h, H_{h-1} \Bigg] \Bigg] \\
& + \mathrm{Var}\Bigg[\sum_{t=1}^{h-1} R_t + \mathbb{E}\Bigg[R_h + \sum_a b^\pi_{V,h+1}(a, O_{h+1}, H_h) \;\Big|\; O_h, S_h, A_h, H_{h-1} \Bigg] \Bigg] \\
=& \mathbb{E}\Bigg[\mathrm{Var}\Bigg[R_h + \sum_a b^\pi_{V,h+1}(a, O_{h+1}, H_h) \;\Big|\; S_h, A_h, H_{h-1} \Bigg] \Bigg] \\
& + \mathrm{Var}\Bigg[\sum_{t=1}^{h-1} R_t + \sum_a b^\pi_{V,h}(a, O_h, H_{h-1}) \Bigg] \qquad \text{(by Equation (25))}
\end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}\left[\mathrm{Var}\left[R_h+\sum_a b^\pi_{V,h+1}(a,O_{h+1},H_h)\;\middle|\;O_h,S_h,A_h,H_{h-1}\right]\right]\\
&+\mathbb{E}\left[\mathrm{Var}\left[\sum_{t=1}^{h-1}R_t+\sum_a b^\pi_{V,h}(a,O_h,H_{h-1})\;\middle|\;\widetilde{\mathcal{D}}_{h-1}\right]\right]\\
&+\mathrm{Var}\left[\mathbb{E}\left[\sum_{t=1}^{h-1}R_t+\sum_a b^\pi_{V,h}(a,O_h,H_{h-1})\;\middle|\;\widetilde{\mathcal{D}}_{h-1}\right]\right]\\
=&\mathbb{E}\left[\mathrm{Var}\left[R_h+\sum_a b^\pi_{V,h+1}(a,O_{h+1},H_h)\;\middle|\;O_h,S_h,A_h,H_{h-1}\right]\right]\\
&+\mathbb{E}\left[\mathrm{Var}\left[R_{h-1}+\sum_a b^\pi_{V,h}(a,O_h,H_{h-1})\;\middle|\;O_{h-1},S_{h-1},A_{h-1},H_{h-2}\right]\right]\\
&+\mathrm{Var}\left[\sum_{t=1}^{h-2}R_t+\mathbb{E}\left[R_{h-1}+\sum_a b^\pi_{V,h}(a,O_h,H_{h-1})\;\middle|\;O_{h-1},S_{h-1},A_{h-1},H_{h-2}\right]\right]\\
=&\mathbb{E}\left[\mathrm{Var}\left[R_h+\sum_a b^\pi_{V,h+1}(a,O_{h+1},H_h)\;\middle|\;O_h,S_h,A_h,H_{h-1}\right]\right]\\
&+\mathbb{E}\left[\mathrm{Var}\left[R_{h-1}+\sum_a b^\pi_{V,h}(a,O_h,H_{h-1})\;\middle|\;O_{h-1},S_{h-1},A_{h-1},H_{h-2}\right]\right]\\
&+\mathrm{Var}\left[\sum_{t=1}^{h-2}R_t+\sum_a b^\pi_{V,h-1}(a,O_{h-1},H_{h-2})\right]\qquad\text{(by Equation (25))}\\
=&\cdots\\
\geq&\sum_{t=1}^{h}\mathbb{E}^\pi\Big[\mathrm{Var}\Big[R_t+\sum_a b^\pi_{V,t+1}(a,O_{t+1},H_t)\;\Big|\;O_t,S_t,A_t,H_{t-1}\Big]\Big],
\end{aligned}
$$

Then, letting $h=T$ and noting that $b^\pi_{V,T+1}=0$, we get the desired result. □

C.1.3 Proof of Lemma 3

Lemma 3. *The variance of $\mathcal{E}(E_1)$ can be decomposed as follows,*

$$
\begin{aligned}
&\mathit{Var}\left\{\sum_{t=1}^{T}\mathcal{E}\left(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi^b_{t'}}\psi_t^\top\mathbf{P}^\dagger_{a_t}\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t}b^\pi_{V,t}-\big(r_t+\sum_{a_{t+1}}b^\pi_{V,t+1}\big)\Big]\right)\right\}\\
=&\sum_{t=1}^{T}\mathbb{E}\left\{\mathit{Var}\left[\mathcal{E}\left(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi^b_{t'}}\psi_t^\top\mathbf{P}^\dagger_{a_t}\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t}b^\pi_{V,t}-\big(r_t+\sum_{a_{t+1}}b^\pi_{V,t+1}\big)\Big]\right)\;\middle|\;\mathcal{D}_t\right]\right\}.
\end{aligned}
$$

Proof. By iteratively applying the law of total variance, we have

$$
\begin{aligned}
&\mathrm{Var}\left\{\sum_{t=1}^{T}\mathcal{E}\left(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi^b_{t'}}\psi_t^\top\mathbf{P}^\dagger_{A_t}\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t}b^\pi_{V,t}-\big(r_t+\sum_{a_{t+1}}b^\pi_{V,t+1}\big)\Big]\right)\right\}\\
=&\mathbb{E}\left\{\mathrm{Var}\left[\sum_{t=1}^{T}\mathcal{E}\left(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi^b_{t'}}\psi_t^\top\mathbf{P}^\dagger_{A_t}\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t}b^\pi_{V,t}-\big(r_t+\sum_{a_{t+1}}b^\pi_{V,t+1}\big)\Big]\right)\;\middle|\;\mathcal{D}_T\right]\right\}\\
&+\mathrm{Var}\left\{\mathbb{E}\left[\sum_{t=1}^{T}\mathcal{E}\left(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi^b_{t'}}\psi_t^\top\mathbf{P}^\dagger_{A_t}\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t}b^\pi_{V,t}-\big(r_t+\sum_{a_{t+1}}b^\pi_{V,t+1}\big)\Big]\right)\;\middle|\;\mathcal{D}_T\right]\right\}\\
=&\mathbb{E}\left\{\mathrm{Var}\left[\sum_{t=1}^{T}\mathcal{E}\left(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi^b_{t'}}\psi_t^\top\mathbf{P}^\dagger_{A_t}\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t}b^\pi_{V,t}-\big(r_t+\sum_{a_{t+1}}b^\pi_{V,t+1}\big)\Big]\right)\;\middle|\;\mathcal{D}_T\right]\right\}
\end{aligned}
$$

$$
\begin{aligned}
&+\mathrm{Var}\left\{\sum_{t=1}^{T-1}\mathcal{E}\left(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\psi_t^\top \mathbf{P}_{A_t}^\dagger \sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t} b_{V,t}^\pi - \big(r_t+\sum_{a_{t+1}} b_{V,t+1}^\pi\big)\Big]\right)\right\}\\
=&\mathbb{E}\left\{\mathrm{Var}\left[\sum_{t=1}^{T}\mathcal{E}\left(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\psi_t^\top \mathbf{P}_{A_t}^\dagger \sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t} b_{V,t}^\pi - \big(r_t+\sum_{a_{t+1}} b_{V,t+1}^\pi\big)\Big]\right)\Big|\,\mathcal{D}_T\right]\right\}\\
&+\mathbb{E}\left\{\mathrm{Var}\left[\sum_{t=1}^{T-1}\mathcal{E}\left(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\psi_t^\top \mathbf{P}_{A_t}^\dagger \sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t} b_{V,t}^\pi - \big(r_t+\sum_{a_{t+1}} b_{V,t+1}^\pi\big)\Big]\right)\Big|\,\mathcal{D}_{T-1}\right]\right\}\\
&+\mathrm{Var}\left\{\mathbb{E}\left[\sum_{t=1}^{T-1}\mathcal{E}\left(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\psi_t^\top \mathbf{P}_{A_t}^\dagger \sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t} b_{V,t}^\pi - \big(r_t+\sum_{a_{t+1}} b_{V,t+1}^\pi\big)\Big]\right)\Big|\,\mathcal{D}_{T-1}\right]\right\}\\
=&\cdots\\
=&\sum_{t=1}^{T}\mathbb{E}\left\{\mathrm{Var}\left[\mathcal{E}\left(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\psi_t^\top \mathbf{P}_{A_t}^\dagger \sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t} b_{V,t}^\pi - \big(r_t+\sum_{a_{t+1}} b_{V,t+1}^\pi\big)\Big]\right)\Big|\,\mathcal{D}_t\right]\right\}.
\end{aligned}
$$

□

C.1.4 Proof of Lemma 4

Lemma 4. *For any function $f_t : \mathcal{A}\times\mathcal{O}\times\mathcal{H}_{t-1}\to\mathbb{R}^d$, the following holds:*

$$
\begin{aligned}
&\mathbb{E}\left[\frac{p_t^\pi(S_t,H_{t-1})}{p_t^{\pi^b}(S_t,H_{t-1})\pi_t^b(A_t\mid S_t)}f_t(A_t,O_t,H_{t-1})\right]\\
=&\mathbb{E}\left[\frac{p_{t-1}^\pi(S_{t-1},H_{t-2})\pi_{t-1}(A_{t-1}\mid O_{t-1},H_{t-2})}{p_{t-1}^{\pi^b}(S_{t-1},H_{t-2})\pi_{t-1}^b(A_{t-1}\mid S_{t-1})}\sum_a f_t(a,O_t,H_{t-1})\right].
\end{aligned}
$$

Proof. For the left-hand side of the equation, we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{p_t^\pi\left(S_t,H_{t-1}\right)}{p_t^{\pi^b}\left(S_t,H_{t-1}\right)\pi_t^b\left(A_t\mid S_t\right)}f_t\left(A_t,O_t,H_{t-1}\right)\right]\\
=&\int\left[\frac{p_t^\pi\left(s_t,h_{t-1}\right)}{p_t^{\pi^b}\left(s_t,h_{t-1}\right)\pi_t^b\left(a_t\mid s_t\right)}f_t\left(a_t,o_t,h_{t-1}\right)\right]p_t^{\pi^b}\left(a_t,o_t,s_t,h_{t-1}\right)da_t do_t ds_t dh_{t-1}\\
=&\int\left[\frac{p_t^\pi\left(s_t,h_{t-1}\right)}{p_t^{\pi^b}\left(s_t,h_{t-1}\right)\pi_t^b\left(a_t\mid s_t\right)}f_t\left(a_t,o_t,h_{t-1}\right)\right]\pi_t^b\left(a_t\mid s_t\right)p\left(o_t\mid s_t\right)p_t^{\pi^b}\left(s_t,h_{t-1}\right)da_t do_t ds_t dh_{t-1}\\
&\quad\text{by } O_t \perp\!\!\!\perp A_t\mid S_t\\
=&\int_{o_t,s_t,h_{t-1}}\sum_a\frac{p_t^\pi\left(s_t,h_{t-1}\right)}{p_t^{\pi^b}\left(s_t,h_{t-1}\right)\pi_t^b\left(a\mid s_t\right)}f_t\left(a,o_t,h_{t-1}\right)\pi_t^b\left(a\mid s_t\right)p\left(o_t\mid s_t\right)p_t^{\pi^b}\left(s_t,h_{t-1}\right)do_t ds_t dh_{t-1}\\
=&\int_{o_t,s_t,h_{t-1}}\sum_a f_t\left(a,o_t,h_{t-1}\right)p\left(o_t\mid s_t\right)p_t^\pi\left(s_t,h_{t-1}\right)do_t ds_t dh_{t-1}.
\end{aligned}
$$

The third equality holds by taking expectation with respect to A_t.

Now we analyze the right-hand side term.

$$
\begin{aligned}
&\mathbb{E}\left[\frac{p_{t-1}^{\pi}\left(S_{t-1}, H_{t-2}\right)\pi_{t-1}\left(A_{t-1}\mid O_{t-1}, H_{t-2}\right)}{p_{t-1}^{\pi^b}\left(S_{t-1}, H_{t-2}\right)\pi_{t-1}^{b}\left(A_{t-1}\mid S_{t-1}\right)}\sum_{a} f_t\left(a, O_t, H_{t-1}\right)\right]\\
=&\int\left[\frac{p_{t-1}^{\pi}\left(s_{t-1}, h_{t-2}\right)\pi_{t-1}\left(a_{t-1}\mid o_{t-1}, h_{t-2}\right)}{p_{t-1}^{\pi^b}\left(s_{t-1}, h_{t-2}\right)\pi_{t-1}^{b}\left(a_{t-1}\mid s_{t-1}\right)}\sum_{a} f_t\left(a, o_t, h_{t-1}\right)\right]\\
&p^{\pi^b}\left(o_t, a_{t-1}, o_{t-1}, s_{t-1}, h_{t-2}\right) do_t da_{t-1} do_{t-1} ds_{t-1} dh_{t-2}\\
=&\int\left[\frac{p_{t-1}^{\pi}\left(s_{t-1}, h_{t-2}\right)\pi_{t-1}\left(a_{t-1}\mid o_{t-1}, h_{t-2}\right)}{p_{t-1}^{\pi^b}\left(s_{t-1}, h_{t-2}\right)\pi_{t-1}^{b}\left(a_{t-1}\mid s_{t-1}\right)}\sum_{a} f_t\left(a, o_t, h_{t-1}\right)\right]\\
&\int p^{\pi^b}\left(o_t, s_t, a_{t-1}, o_{t-1}, s_{t-1}, h_{t-2}\right) ds_t do_t da_{t-1} do_{t-1} ds_{t-1} dh_{t-2} \quad \text{(marginalization over } S_t\text{)}\\
=&\int\frac{p_{t-1}^{\pi}\left(s_{t-1}, h_{t-2}\right)\pi_{t-1}\left(a_{t-1}\mid o_{t-1}, h_{t-2}\right)}{p_{t-1}^{\pi^b}\left(s_{t-1}, h_{t-2}\right)\pi_{t-1}^{b}\left(a_{t-1}\mid s_{t-1}\right)}\sum_{a} f_t\left(a, o_t, h_{t-1}\right)\\
&p^{\pi^b}\left(o_t, s_t, a_{t-1}, o_{t-1}, s_{t-1}, h_{t-2}\right) ds_t do_t da_{t-1} do_{t-1} ds_{t-1} dh_{t-2}\\
=&\int p_{t-1}^{\pi}\left(s_{t-1}, h_{t-2}\right)\pi_{t-1}\left(a_{t-1}\mid o_{t-1}, h_{t-2}\right)\sum_{a} f_t\left(a, o_t, h_{t-1}\right)\\
&p\left(o_t\mid s_t\right) p\left(s_t\mid a_{t-1}, s_{t-1}\right) p\left(o_{t-1}\mid s_{t-1}\right) do_t ds_t da_{t-1} do_{t-1} ds_{t-1} dh_{t-2}\\
=&\int_{o_{o_t}, s_t, h_{t-1}}\sum_{a} f_t\left(a, o_t, h_{t-1}\right) p\left(o_t\mid s_t\right) p_t^{\pi}\left(s_t, h_{t-1}\right) do_t ds_t dh_{t-1}.
\end{aligned}
$$

The last equality is by noticing that

$$p_t^{\pi}(s_t, h_{t-1}) = p(s_t \mid s_{t-1}, a_{t-1})\pi_t(a_{t-1} \mid o_{t-1}, h_{t-2})p(o_{t-1} \mid s_{t-1})p_{t-1}^{\pi}(s_{t-1}, h_{t-2}).$$

Then the proof is completed by comparing these two terms. □

C.1.5 Bounding $\mathcal{E}(E_2)$

Similar as before, we treat $\mathcal{E}(E_2) = \sum_{t=1}^{T}\mathcal{E}(I_t)$ as the sum of T terms where

$$
\begin{aligned}
\mathcal{E}(I_1) =&\mathcal{E}\Bigg\{\Bigg(\sum_{a_1}\pi_1(a_1|O_1)\psi_1^{\top}(O_1)\widehat{\mathbf{P}}_{a_1}^{\dagger}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1(O_1) - \sum_{a_1}\pi_1(a_1|O_1)\psi_1^{\top}(O_1)\mathbf{P}_{a_1}^{\dagger}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1(O_1)\Bigg)\\
&\Bigg[\sum_{a_1} b_{V,1}^{\pi}(a_1, O_1) - \Big(r_1 + \sum_{a_2} b_{V,2}^{\pi}\big(a_2, o_2, (O_1, a_1)\big)\Big)\Bigg]\Bigg\},\\
\mathcal{E}(I_2) =&\mathcal{E}\Bigg\{\Bigg(\sum_{a_1}\pi_1(a_1|O_1)\psi_1^{\top}(O_1)\widehat{\mathbf{P}}_{a_1}^{\dagger}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1(O_1)\cdot\sum_{a_2}\pi_2\big(a_2|o_2,(O_1,a_1)\big)\psi_2^{\top}\widehat{\mathbf{P}}_{a_2}^{\dagger}\sum_{r_2,o_3}\widehat{\mathbf{P}}_{a_2,r_2,o_3}\psi_2\\
&-\sum_{a_1}\pi_1(a_1|O_1)\psi_1^{\top}(O_1)\mathbf{P}_{a_1}^{\dagger}\sum_{r_1,o_2}\mathbf{P}_{a_1,r_1,o_2}\psi_1(O_1)\cdot\sum_{a_2}\pi_2\big(a_2|o_2,(O_1,a_1)\big)\psi_2^{\top}\mathbf{P}_{a_2}^{\dagger}\sum_{r_2,o_3}\widehat{\mathbf{P}}_{a_2,r_2,o_3}\psi_2\Bigg)\\
&\Bigg[\sum_{a_2} b_{V,2}^{\pi}\big(a_2, o_2, (O_1, a_1)\big) - \Big(r_2 + \sum_{a_3} b_{V,3}^{\pi}\big(a_3, o_3, (O_1, a_1, o_2, a_2)\big)\Big)\Bigg]\Bigg\},\\
&\cdots\\
\mathcal{E}(I_T) =&\mathcal{E}\Bigg\{\Bigg(\sum_{a_1}\pi_1\psi_1^{\top}\widehat{\mathbf{P}}_{a_1}^{\dagger}\sum_{r_1,o_2}\widehat{\mathbf{P}}_{a_1,r_1,o_2}\psi_1\sum_{a_2}\pi_2\psi_2^{\top}\widehat{\mathbf{P}}_{a_2}^{\dagger}\sum_{r_2,o_3}\widehat{\mathbf{P}}_{a_2,r_2,o_3}\psi_2\cdots\sum_{a_T}\pi_T\psi_T^{\top}\mathbf{P}_{a_T}^{\dagger}\sum_{r_T}\widehat{\mathbf{P}}_{a_T,r_T}\psi_T\\
&-\sum_{a_1}\pi_1\psi_1^{\top}\mathbf{P}_{a_1}^{\dagger}\sum_{r_1,o_2}\mathbf{P}_{a_1,r_1,o_2}\psi_1\sum_{a_2}\pi_2\psi_2^{\top}\mathbf{P}_{a_2}^{\dagger}\sum_{r_2,o_3}\mathbf{P}_{a_2,r_2,o_3}\psi_2\cdots\sum_{a_T}\pi_T\psi_T^{\top}\mathbf{P}_{a_T}^{\dagger}\sum_{r_T}\widehat{\mathbf{P}}_{a_T,r_T}\psi_T\Bigg)\\
&\Bigg[\sum_{a_T} b_{V,T}^{\pi}\big(a_T, o_T, (O_1, a_1, \cdots, a_{T-1})\big) - r_T\Bigg]\Bigg\}.
\end{aligned}
$$

Next, we aim to derive the bound of $\mathrm{Var}[\mathcal{E}(E_2)]$. We first split $\mathcal{E}(E_2)$ into two parts:

$$
\begin{aligned}
&\mathcal{E}(E_2)\\
=&\mathcal{E}\Bigg\{\sum_{t=1}^{T}\Big(\sum_{a_1}\pi_1\psi_1^\top\widehat{\mathbf{P}}_{a_1}^\dagger\cdots\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}}\psi_t-\sum_{a_1}\pi_1\psi_1^\top\mathbf{P}_{a_1}^\dagger\cdots\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{a_t,r_t,o_{t+1}}\psi_t\Big)\\
&\quad\Big[\sum_{a_t}b_{V,t}^\pi(a_t,o_t,(O_1,A_1,\ldots,o_{t-1},a_{t-1}))-\Big(r_t+\sum_{a_{t+1}}b_{V,t+1}^\pi(a_{t+1},o_{t+1},(O_1,A_1,\ldots,o_t,a_t))\Big)\Big]\Bigg\}\\
=&\mathcal{E}\Bigg\{\sum_{t=1}^{T}\Big(\Big(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\Big)\psi_1^\top\widehat{\mathbf{P}}_{A_1}^\dagger\cdots\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t-\Big(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\Big)\psi_1^\top\mathbf{P}_{A_1}^\dagger\cdots\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big)\\
&\quad\Big[\sum_{a_t}b_{V,t}^\pi(a_t,o_t,(O_1,A_1,\ldots,o_{t-1},A_{t-1}))-\Big(r_t+\sum_{a_{t+1}}b_{V,t+1}^\pi(a_{t+1},o_{t+1},(O_1,A_1,\ldots,o_t,A_t))\Big)\Big]\Bigg\}\\
=&\mathcal{E}\Bigg\{\sum_{t=1}^{T}\Big(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\Big)\psi_1^\top\mathbf{P}_{A_1}^\dagger\cdots\sum_{r_{t-1},o_t}\mathbf{P}_{A_{t-1},r_{t-1},o_t}\psi_{t-1}\Big(\psi_t^\top\widehat{\mathbf{P}}_{A_t}^\dagger-\psi_t^\top\mathbf{P}_{A_t}^\dagger\Big)\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\\
&\quad\Big[\sum_{a_t}b_{V,t}^\pi(a_t,o_t,(O_1,A_1,\ldots,o_{t-1},A_{t-1}))-\Big(r_t+\sum_{a_{t+1}}b_{V,t+1}^\pi(a_{t+1},o_{t+1},(O_1,A_1,\ldots,o_t,A_t))\Big)\Big]\Bigg\}\\
&+\mathcal{E}\Bigg\{\sum_{t=1}^{T}\Big(\Big(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\Big)\psi_1^\top\widehat{\mathbf{P}}_{A_1}^\dagger\cdots\sum_{r_{t-1},o_t}\widehat{\mathbf{P}}_{A_{t-1},r_{t-1},o_t}\psi_{t-1}-\Big(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\Big)\psi_1^\top\mathbf{P}_{A_1}^\dagger\cdots\sum_{r_{t-1},o_t}\mathbf{P}_{A_{t-1},r_{t-1},o_t}\psi_{t-1}\Big)\\
&\quad\psi_t^\top\widehat{\mathbf{P}}_{A_t}^\dagger\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\Big[\sum_{a_t}b_{V,t}^\pi(a_t,o_t,(O_1,A_1,\ldots,o_{t-1},A_{t-1}))-\Big(r_t+\sum_{a_{t+1}}b_{V,t+1}^\pi(a_{t+1},o_{t+1},(O_1,A_1,\ldots,o_t,A_t))\Big)\Big]\Bigg\}\\
=&E_{2,1}+E_{2,2}
\end{aligned}
$$

(1) Bounding $\mathcal{E}(E_{2,1})$ We first verify that $\mathbb{E}[\mathcal{E}(E_{21})]=0$. For each $t\in[T]$, it holds that

$$
\begin{aligned}
&\mathbb{E}\Bigg\{\mathbb{E}\Bigg\{\mathcal{E}\Bigg\{\sum_{t=1}^{T}\Big(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\Big)\psi_1^\top\mathbf{P}_{A_1}^\dagger\cdots\sum_{r_{t-1},o_t}\mathbf{P}_{A_{t-1},r_{t-1},o_t}\psi_{t-1}\Big(\psi_t^\top\widehat{\mathbf{P}}_{A_t}^\dagger-\psi_t^\top\mathbf{P}_{A_t}^\dagger\Big)\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t\\
&\quad\Big[\sum_{a_t}b_{V,t}^\pi(a_t,o_t,(O_1,A_1,\ldots,o_{t-1},A_{t-1}))-\Big(r_t+\sum_{a_{t+1}}b_{V,t+1}^\pi(a_{t+1},o_{t+1},(O_1,A_1,\ldots,o_t,A_t))\Big)\Big]\Bigg\}\Big|\mathcal{D}_t\Bigg\}\Bigg\}\\
=&\mathbb{E}\Bigg\{\mathcal{E}\Bigg\{\sum_{t=1}^{T}\Big(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\Big)\psi_1^\top\mathbf{P}_{A_1}^\dagger\cdots\sum_{r_{t-1},o_t}\mathbf{P}_{A_{t-1},r_{t-1},o_t}\psi_{t-1}\Big(\psi_t^\top\widehat{\mathbf{P}}_{A_t}^\dagger-\psi_t^\top\mathbf{P}_{A_t}^\dagger\Big)\sum_{r_t,o_{t+1}}\mathbf{P}_{A_t,r_t,o_{t+1}}\psi_t\\
&\quad\Big[\sum_{a_t}b_{V,t}^\pi(a_t,o_t,(O_1,A_1,\ldots,o_{t-1},A_{t-1}))-\Big(r_t+\sum_{a_{t+1}}b_{V,t+1}^\pi(a_{t+1},o_{t+1},(O_1,A_1,\ldots,o_t,A_t))\Big)\Big]\Bigg\}\Bigg\}\\
=&\mathbb{E}\Bigg\{\mathcal{E}\Bigg\{\sum_{t=1}^{T}\Big(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi_{t'}^b}\Big)\psi_1^\top\mathbf{P}_{A_1}^\dagger\cdots\sum_{r_{t-1},o_t}\mathbf{P}_{A_{t-1},r_{t-1},o_t}\psi_{t-1}\Big(\psi_t^\top\mathbf{P}_{A_t}^\dagger-\psi_t^\top\mathbf{P}_{A_t}^\dagger\Big)\sum_{r_t,o_{t+1}}\mathbf{P}_{A_t,r_t,o_{t+1}}\psi_t\\
&\quad\Big[\sum_{a_t}b_{V,t}^\pi(a_t,o_t,(O_1,A_1,\ldots,o_{t-1},A_{t-1}))-\Big(r_t+\sum_{a_{t+1}}b_{V,t+1}^\pi(a_{t+1},o_{t+1},(O_1,A_1,\ldots,o_t,A_t))\Big)\Big]\Bigg\}\Bigg\}\\
=&0.
\end{aligned}
$$

With probability at least $1-T|\mathcal{A}|\delta$, we have the upper bound such that

$$
\begin{aligned}
&\mathrm{Var}\Big[\mathcal{E}(E_{2,1})\Big] \\
\overset{(i)}{=}&\sum_{t=1}^{T}\mathbb{E}\Bigg\{\mathrm{Var}\Bigg[\mathcal{E}\Bigg(\Big(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi^b_{t'}}\Big)\psi_1^\top\mathbf{P}^\dagger_{A_1}\cdots\sum_{r_{t-1},o_t}\mathbf{P}_{A_{t-1},r_{t-1},o_t}\psi_{t-1}\Big(\psi_t^\top\widehat{\mathbf{P}}^\dagger_{A_t}-\psi_t^\top\mathbf{P}^\dagger_{A_t}\Big)\sum_{r_t,o_{t+1}}\widehat{\mathbf{P}}_{A_t,r_t,o_{t+1}}\psi_t \\
&\Big[\sum_{a_t}b^\pi_{V,t}(a_t,o_t,(O_1,A_1,\ldots,o_{t-1},A_{t-1}))-\Big(r_t+\sum_{a_{t+1}}b^\pi_{V,t+1}(a_{t+1},o_{t+1},(O_1,A_1,\ldots,o_t,A_t))\Big)\Big]\Bigg)\,\Big|\,\mathcal{D}_t\Bigg]\Bigg\} \\
\le&\frac{T^2}{n}(1-\theta^*)^{-1}|\mathcal{A}|\sum_{t=1}^{T} \\
&\mathbb{E}\Bigg\{\mathcal{E}\Bigg(\Big(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi^b_{t'}}\Big)\psi_1^\top\mathbf{P}^\dagger_{A_1}\cdots\sum_{r_{t-1},o_t}\mathbf{P}_{A_{t-1},r_{t-1},o_t}\psi_{t-1}\Big(\psi_t^\top\widehat{\mathbf{P}}^\dagger_{A_t}-\psi_t^\top\mathbf{P}^\dagger_{A_t}\Big)\sum_{r_t,o_{t+1}}\mathrm{diag}\Big(\mathbf{P}_{A_t,r_t,o_{t+1}}\psi_t\Big)\Big)^2\Bigg\} \\
=&\frac{T^2}{n}(1-\theta^*)^{-1}|\mathcal{A}|\sum_{t=1}^{T}\mathbb{E}\Bigg\{\mathcal{E}\Bigg(\Big(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi^b_{t'}}\Big)\psi_1^\top\mathbf{P}^\dagger_{A_1}\cdots\mathbf{P}_{A_{t-1}}\psi_{t-1}\psi_t^\top\Big(\widehat{\mathbf{P}}^\dagger_{A_t}-\mathbf{P}^\dagger_{A_t}\Big)\mathrm{diag}\Big(\mathbf{P}_{A_t}\psi_t\Big)\Big)^2\Bigg\} \\
=&\frac{T^2}{n}(1-\theta^*)^{-1}|\mathcal{A}|\sum_{t=1}^{T}\mathbb{E}\Bigg\{\mathcal{E}\Bigg(\Big(\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi^b_{t'}}\Big)\psi_t^\top\Big(\widehat{\mathbf{P}}^\dagger_{A_t}-\mathbf{P}^\dagger_{A_t}\Big)\mathrm{diag}\Big(\mathbf{P}_{A_t}\psi_t\Big)\Big)^2\Bigg\} \\
\le&\frac{T^2}{n}(1-\theta^*)^{-1}|\mathcal{A}|\sum_{t=1}^{T}\mathbb{E}\Big[\prod_{t'=1}^{t}\frac{\pi_{t'}}{\pi^b_{t'}}\Big]^2\max_{a_t\in\mathcal{A}}\|\psi_t\|_2^2\|\mathbf{P}^\dagger_{a_t}-\widehat{\mathbf{P}}^\dagger_{a_t}\|_2^2\|\mathrm{diag}(\mathbf{P}_{a_t}\psi_t)\|_2^2 \\
\overset{(ii)}{\le}&\frac{T^3}{n}(1-\theta^*)^{-1}|\mathcal{A}|C_{\pi^b}\sum_{t=1}^{T}\max_{a_t\in\mathcal{A}}\|\widehat{\mathbf{P}}^\dagger_{a_t}\|_2^2\|\mathbf{P}_{a_t}-\widehat{\mathbf{P}}_{a_t}\|_2^2\|\mathbf{P}^\dagger_{a_t}\|_2^2 \qquad \text{(by Assumption 2(a))} \\
\overset{(iii)}{\lesssim}&\frac{T^2}{n}(1-\theta^*)^{-1}C_{\pi^b}C_P^4|\mathcal{O}|^2|\mathcal{H}_{T-1}|^2\sum_{t=1}^{T}\max_{a_t\in\mathcal{A}}\|\mathbf{P}_{a_t}-\widehat{\mathbf{P}}_{a_t}\|_2^2 \qquad \text{(by Equation (26))} \\
\le&\frac{T^2}{n}(1-\theta^*)^{-1}C_{\pi^b}C_P^4|\mathcal{A}|^2|\mathcal{O}|^2|\mathcal{H}_{T-1}|^2\cdot|\mathcal{O}||\mathcal{H}_{T-1}|\sum_{t=1}^{T}\frac{\log(|\mathcal{O}||\mathcal{H}_{t-1}|/\delta)}{n(1-\theta^*)} \quad \text{(by Lemma 5)} \\
\lesssim&\frac{T^3}{n^2}(1-\theta^*)^{-2}C_{\pi^b}C_P^4|\mathcal{O}|^{3T}|\mathcal{A}|^{3T-1}\log(|\mathcal{O}|^T|\mathcal{A}|^{T-1}/\delta)
\end{aligned}
$$

The equality (i) holds by recursively applying the law of total variance (same as Lemma 3) and $\mathbb{E}\big[\mathcal{E}(I_t)\mid\mathcal{D}_t\big]=0$. The inequality (ii) uses the fact that $A^{-1}-B^{-1}=A^{-1}(B-A)B^{-1}$, $\|\psi_t\|_2=1$, and $\big\|\mathrm{diag}\big(\mathbf{P}_{A_t}\psi_t\big)\big\|_2\le 1$. For inequality (iii), by Lemma 11, for any $a_t\in\mathcal{A}$, we have

$$\sigma_{min}(\widehat{\mathbf{P}}_{a_t})\ge\sigma_{min}(\mathbf{P}_{a_t})-\|\widehat{\mathbf{P}}_{a_t}-\mathbf{P}_{a_t}\|_2.$$

Thus, for any $a_t\in\mathcal{A}$, we have

$$
\begin{aligned}
\|\widehat{\mathbf{P}}^\dagger_{a_t}\|_2=&\frac{1}{\sigma_{min}(\widehat{\mathbf{P}}_{a_t})} \\
\le&\frac{1}{\sigma_{min}(\mathbf{P}_{a_t})-\|\widehat{\mathbf{P}}_{a_t}-\mathbf{P}_{a_t}\|_2} \qquad (26)\\
\lesssim&\frac{1}{\sigma_{min}(\mathbf{P}_{a_t})} \\
\le&C_P\sqrt{|\mathcal{O}||\mathcal{H}_{t-1}|} \qquad \text{(by Assumption 2).}
\end{aligned}
$$

Here, we assume a sufficiently large sample size n, ensuring $\|\widehat{\mathbf{P}}_{a_t}-\mathbf{P}_{a_t}\|_2$ converge to 0.

Therefore, with probability at least $1-T|\mathcal{A}|\delta$, we obtain

$$\mathcal{E}(E_{2,1})\lesssim\frac{T^{1.5}}{n}(1-\theta^*)^{-1}C_{\pi^b}^{\frac{1}{2}}C_P^2|\mathcal{O}|^{\frac{3T}{2}}|\mathcal{A}|^{\frac{3T-1}{2}}\sqrt{\log(|\mathcal{O}|^T|\mathcal{A}|^{T-1}/\delta)} \qquad (27)$$

(2) Bounding $\mathcal{E}(E_{2,2})$ Next, we derive the bound for $\mathcal{E}(E_{2,2})$.

For ease of expression, we denote

$$\begin{aligned}\mathcal{P}_t :=& \psi_t^\top \mathbf{P}_{A_t}^\dagger \sum_{r_t, o_{t+1}} \mathbf{P}_{A_t, r_t, o_{t+1}} \psi_t, \\ \widehat{\mathcal{P}}_t :=& \psi_t^\top \widehat{\mathbf{P}}_{A_t}^\dagger \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{A_t, r_t, o_{t+1}} \psi_t.\end{aligned}$$

Notice that, for each $t \in [T]$, with probability at least $1 - |\mathcal{A}|\delta$, $\left|\mathcal{P}_t - \widehat{\mathcal{P}}_t\right|$ can be bounded by

$$\begin{aligned}\left|\mathcal{P}_t - \widehat{\mathcal{P}}_t\right| =& \left|\psi_t^\top \mathbf{P}_{A_t}^\dagger \mathbf{P}_{A_t} \psi_t - \psi_t^\top \widehat{\mathbf{P}}_{A_t}^\dagger \widehat{\mathbf{P}}_{A_t} \psi_t\right| \\ =& \left|\psi_t^\top \mathbf{P}_{A_t}^\dagger \mathbf{P}_{A_t} \psi_t - \psi_t^\top \widehat{\mathbf{P}}_{A_t}^\dagger \mathbf{P}_{A_t} \psi_t + \psi_t^\top \widehat{\mathbf{P}}_{A_t}^\dagger \mathbf{P}_{A_t} \psi_t - \psi_t^\top \widehat{\mathbf{P}}_{A_t}^\dagger \widehat{\mathbf{P}}_{A_t} \psi_t\right| \\ =& \left|\psi_t^\top \left(\mathbf{P}_{A_t}^\dagger - \widehat{\mathbf{P}}_{A_t}^\dagger\right) \mathbf{P}_{A_t} \psi_t + \psi_t^\top \widehat{\mathbf{P}}_{A_t}^\dagger \left(\mathbf{P}_{A_t} - \widehat{\mathbf{P}}_{A_t}\right) \psi_t\right| \\ \leq& \|\psi_t\|_2^2 \|\mathbf{P}_{A_t}^\dagger - \widehat{\mathbf{P}}_{A_t}^\dagger\|_2 \|\mathbf{P}_{A_t}\|_2 + \|\psi_t\|_2^2 \|\widehat{\mathbf{P}}_{A_t}^\dagger\|_2 \|\mathbf{P}_{A_t} - \widehat{\mathbf{P}}_{A_t}\|_2 \\ \leq& \|\widehat{\mathbf{P}}_{A_t}^\dagger\|_2 \|\widehat{\mathbf{P}}_{A_t} - \mathbf{P}_{A_t}\|_2 + \|\widehat{\mathbf{P}}_{A_t}^\dagger\|_2 \|\mathbf{P}_{A_t} - \widehat{\mathbf{P}}_{A_t}\|_2 \\ \leq& 2 \max_{a_t \in \mathcal{A}} \|\widehat{\mathbf{P}}_{a_t}^\dagger\|_2 \|\widehat{\mathbf{P}}_{a_t} - \mathbf{P}_{a_t}\|_2 \\ \lesssim& C_P |\mathcal{A}| \sqrt{|\mathcal{O}||\mathcal{H}_{t-1}|} \frac{(1-\theta^*)^{-\frac{1}{2}}}{\sqrt{n}} \sqrt{|\mathcal{O}||\mathcal{H}_{t-1}|} \sqrt{\log(|\mathcal{O}||\mathcal{H}_{t-1}|/\delta)} \\ & \text{(by Equation (26) and Lemma 5)} \\ =& \frac{(1-\theta^*)^{-\frac{1}{2}}}{\sqrt{n}} C_P |\mathcal{O}|^t |\mathcal{A}|^t \sqrt{\log(|\mathcal{O}|^t |\mathcal{A}|^{t-1}/\delta)} \\ :=& \xi_{t,n,\delta}\end{aligned} \tag{28}$$

Notice that, for each $t \in [T]$, with probability at least $1 - |\mathcal{A}|\delta$, we have

$$
\begin{aligned}
&\left|\psi_t^\top \widehat{\mathbf{P}}_{A_t}^\dagger \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{A_t, r_t, o_{t+1}} \psi_t \Big[\sum_{a_t} b_{V,t}^\pi - \Big(r_t + \sum_{a_{t+1}} b_{V,t+1}^\pi \Big) \Big] \right| \\
=&\left|\psi_t^\top \widehat{\mathbf{P}}_{A_t}^\dagger \mathbf{P}_{A_t} \mathbf{P}_{A_t}^\dagger \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{A_t, r_t, o_{t+1}} \psi_t \Big[\sum_{a_t} b_{V,t}^\pi - \Big(r_t + \sum_{a_{t+1}} b_{V,t+1}^\pi \Big) \Big] \right| \\
\leq &\max_{a_t \in \mathcal{A}} \left|\psi_t^\top \widehat{\mathbf{P}}_{a_t}^\dagger \mathbf{P}_{a_t} \mathbf{P}_{a_t}^\dagger \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{a_t, r_t, o_{t+1}} \psi_t \Big[\sum_{a_t} b_{V,t}^\pi - \Big(r_t + \sum_{a_{t+1}} b_{V,t+1}^\pi \Big) \Big] \right| \\
\leq &\max_{a_t \in \mathcal{A}} \left\|\psi_t \widehat{\mathbf{P}}_{a_t}^\dagger \mathbf{P}_{a_t}\right\|_2 \left\| \mathbf{P}_{a_t}^\dagger \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{a_t, r_t, o_{t+1}} \psi_t \Big[\sum_{a_t} b_{V,t}^\pi - \Big(r_t + \sum_{a_{t+1}} b_{V,t+1}^\pi \Big) \Big] \right\|_2 \\
\overset{(i)}{\leq} &\max_{a_t \in \mathcal{A}} \left\|\widehat{\mathbf{P}}_{a_t}^\dagger \mathbf{P}_{a_t}\right\|_2 \frac{T}{\sqrt{n}} (1-\theta^*)^{-\frac{1}{2}} C_P \sqrt{|\mathcal{O}||\mathcal{H}_{t-1}||\mathcal{A}|} \\
\overset{(ii)}{\leq} &\max_{a_t \in \mathcal{A}} \left\|\mathbf{I} + \widehat{\mathbf{P}}_{a_t}^\dagger \Big(\mathbf{P}_{a_t} - \widehat{\mathbf{P}}_{a_t} \Big) \right\|_2 \frac{T}{\sqrt{n}} (1-\theta^*)^{-\frac{1}{2}} C_P \sqrt{|\mathcal{O}||\mathcal{H}_{t-1}||\mathcal{A}|} \\
\leq &\max_{a_t \in \mathcal{A}} \Big(\|\mathbf{I}\|_2 + \|\widehat{\mathbf{P}}_{a_t}^\dagger\|_2 \|\mathbf{P}_{a_t} - \widehat{\mathbf{P}}_{a_t}\|_2 \Big) \frac{T}{\sqrt{n}} (1-\theta^*)^{-\frac{1}{2}} C_P \sqrt{|\mathcal{O}||\mathcal{H}_{t-1}||\mathcal{A}|} \\
\leq &\Big(1 + C_P \frac{(1-\theta^*)^{-\frac{1}{2}}}{\sqrt{n}} |\mathcal{O}||\mathcal{H}_{t-1}||\mathcal{A}| \sqrt{\log(|\mathcal{O}||\mathcal{H}_{t-1}|/\delta)} \Big) \frac{T}{\sqrt{n}} (1-\theta^*)^{-\frac{1}{2}} C_P \sqrt{|\mathcal{O}||\mathcal{H}_{t-1}||\mathcal{A}|} \\
&\quad \text{(by Equation (26) and Lemma 5)} \\
=&\Big(1 + \frac{(1-\theta^*)^{-\frac{1}{2}}}{\sqrt{n}} C_P |\mathcal{O}|^t |\mathcal{A}|^t \sqrt{\log(|\mathcal{O}|^t |\mathcal{A}|^{t-1}/\delta)} \Big) \frac{T}{\sqrt{n}} (1-\theta^*)^{-\frac{1}{2}} C_P |\mathcal{O}|^{\frac{t}{2}} |\mathcal{A}|^{\frac{t}{2}} \\
=&T\Big(1 + \xi_{t,n,\delta} \Big) \zeta_{t,n}.
\end{aligned}
\tag{29}
$$

Here, we denote

$$
\zeta_{t,n} := \frac{(1-\theta^*)^{-\frac{1}{2}}}{\sqrt{n}} C_P |\mathcal{O}|^{\frac{t}{2}} |\mathcal{A}|^{\frac{t}{2}}.
$$

For (i), by using the similar argument in the proof of bounding $\mathcal{E}(E_1)$, we have

$$
\max_{a_t \in \mathcal{A}} \left\| \mathbf{P}_{a_t}^\dagger \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{a_t, r_t, o_{t+1}} \psi_t \Big[\sum_{a_t} b_{V,t}^\pi - \Big(r_t + \sum_{a_{t+1}} b_{V,t+1}^\pi \Big) \Big] \right\|_2 \leq \mathcal{O}_P\Big(\frac{T}{\sqrt{n}} (1-\theta^*)^{-\frac{1}{2}} C_P \sqrt{|\mathcal{O}||\mathcal{H}_{t-1}||\mathcal{A}|} \Big).
$$

The inequality (ii) uses the fact that $A^{-1}B = I + A^{-1}(B - A)$.

Then, for $\mathcal{E}(E_{2,2})$, we have

$$
\begin{aligned}
&\mathcal{E}(E_{2,2}) \\
=&\mathcal{E}\Bigg(\sum_{t=1}^T \Bigg(\Big(\prod_{t'=1}^{t-1} \frac{\pi_{t'}}{\pi_{t'}^b} \Big) \psi_1^\top \widehat{\mathbf{P}}_{A_1}^\dagger \cdots \sum_{r_{t-1}, o_t} \widehat{\mathbf{P}}_{A_{t-1}, r_{t-1}, o_t} \psi_{t-1} - \Big(\prod_{t'=1}^{t-1} \frac{\pi_{t'}}{\pi_{t'}^b} \Big) \psi_1^\top \mathbf{P}_{A_1}^\dagger \cdots \sum_{r_{t-1}, o_t} \mathbf{P}_{A_{t-1}, r_{t-1}, o_t} \psi_{t-1} \Bigg) \\
&\quad \frac{\pi_t}{\pi_t^b} \psi_t^\top \widehat{\mathbf{P}}_{A_t}^\dagger \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{A_t, r_t, o_{t+1}} \psi_t \\
&\quad \Big[\sum_{a_t} b_{V,t}^\pi(a_t, o_t, (O_1, A_1, \ldots, o_{t-1}, A_{t-1})) - \Big(r_t + \sum_{a_{t+1}} b_{V,t+1}^\pi(a_{t+1}, o_{t+1}, (O_1, A_1, \ldots, o_t, A_t)) \Big) \Big] \Bigg) \\
=&\mathcal{E}\Bigg(\sum_{t=1}^T \Bigg(\prod_{t'=1}^{t-1} \frac{\pi_{t'}}{\pi_{t'}^b} \widehat{\mathcal{P}}_{t'} - \prod_{t'=1}^{t-1} \frac{\pi_{t'}}{\pi_{t'}^b} \mathcal{P}_{t'} \Bigg) \frac{\pi_t}{\pi_t^b} \psi_t^\top \widehat{\mathbf{P}}_{A_t}^\dagger \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{A_t, r_t, o_{t+1}} \psi_t \\
&\quad \Big[\sum_{a_t} b_{V,t}^\pi(a_t, o_t, (O_1, A_1, \ldots, o_{t-1}, A_{t-1})) - \Big(r_t + \sum_{a_{t+1}} b_{V,t+1}^\pi(a_{t+1}, o_{t+1}, (O_1, A_1, \ldots, o_t, A_t)) \Big) \Big] \Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathcal{E}\Bigg(\sum_{t=1}^{T} \Big(\prod_{t'=1}^{t} \frac{\pi_{t'}}{\pi^b_{t'}} \Big) \Big(\prod_{t'=1}^{t-1} \widehat{\mathcal{P}}_{t'} - \prod_{t'=1}^{t-1} \mathcal{P}_{t'} \Big) \psi_t^\top \widehat{\mathbf{P}}^\dagger_{A_t} \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{A_t, r_t, o_{t+1}} \psi_t \\
&\quad \Big[\sum_{a_t} b^\pi_{V,t}(a_t, o_t, (O_1, A_1, \ldots, o_{t-1}, A_{t-1})) - \Big(r_t + \sum_{a_{t+1}} b^\pi_{V,t+1}(a_{t+1}, o_{t+1}, (O_1, A_1, \ldots, o_t, A_t)) \Big) \Big] \Bigg) \\
&\leq C_{\pi^b}^{\frac{1}{2}} \sum_{t=1}^{T} \Bigg\{ \mathcal{E}\Bigg(\Big(\prod_{t'=1}^{t-1} \Big(\widehat{\mathcal{P}}_{t'} - \mathcal{P}_{t'} + \mathcal{P}_{t'} \Big) - \prod_{t'=1}^{t-1} \mathcal{P}_{t'} \Big) \psi_t^\top \widehat{\mathbf{P}}^\dagger_{A_t} \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{A_t, r_t, o_{t+1}} \psi_t \\
&\quad \Big[\sum_{a_t} b^\pi_{V,t}(a_t, o_t, (O_1, A_1, \ldots, o_{t-1}, A_{t-1})) - \Big(r_t + \sum_{a_{t+1}} b^\pi_{V,t+1}(a_{t+1}, o_{t+1}, (O_1, A_1, \ldots, o_t, A_t)) \Big) \Big] \Bigg)^2 \Bigg\}^{\frac{1}{2}} \\
&\qquad \text{(by Assumption 2(a) and Cauchy-Schwarz inequality)} \\
&\leq C_{\pi^b}^{\frac{1}{2}} \sum_{t=1}^{T} \Bigg\{ \mathcal{E}\Bigg(\Big| \prod_{t'=1}^{t-1} \Big(\widehat{\mathcal{P}}_{t'} - \mathcal{P}_{t'} + \mathcal{P}_{t'} \Big) - \prod_{t'=1}^{t-1} \mathcal{P}_{t'} \Big| \cdot \Big| \psi_t^\top \widehat{\mathbf{P}}^\dagger_{A_t} \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{A_t, r_t, o_{t+1}} \psi_t \\
&\quad \Big[\sum_{a_t} b^\pi_{V,t}(a_t, o_t, (O_1, A_1, \ldots, o_{t-1}, A_{t-1})) - \Big(r_t + \sum_{a_{t+1}} b^\pi_{V,t+1}(a_{t+1}, o_{t+1}, (O_1, A_1, \ldots, o_t, A_t)) \Big) \Big] \Big| \Bigg)^2 \Bigg\}^{\frac{1}{2}} \\
&= C_{\pi^b}^{\frac{1}{2}} \sum_{t=1}^{T} \Bigg\{ \mathcal{E}\Bigg(\Big| \sum_{(\delta_1, \ldots, \delta_{t-1}) \in \{0,1\}^{t-1} / \{1\}^{t-1}} \Big(\mathcal{P}_1 \Big)^{\delta_1} \Big(\widehat{\mathcal{P}}_1 - \mathcal{P}_1 \Big)^{1-\delta_1} \cdots \Big(\mathcal{P}_{t-1} \Big)^{\delta_{t-1}} \Big(\widehat{\mathcal{P}}_{t-1} - \mathcal{P}_{t-1} \Big)^{1-\delta_{t-1}} \Big| \\
&\quad \cdot \Big| \psi_t^\top \widehat{\mathbf{P}}^\dagger_{A_t} \sum_{r_t, o_{t+1}} \widehat{\mathbf{P}}_{A_t, r_t, o_{t+1}} \psi_t \Big[\sum_{a_t} b^\pi_{V,t} - \Big(r_t + \sum_{a_{t+1}} b^\pi_{V,t+1} \Big) \Big] \Big| \Bigg)^2 \Bigg\}^{\frac{1}{2}} \\
&\leq C_{\pi^b}^{\frac{1}{2}} \sum_{t=1}^{T} \sum_{(\delta_1, \ldots, \delta_{t-1}) \in \{0,1\}^{t-1} / \{1\}^{t-1}} \Big| \mathcal{P}_1 \Big|^{\delta_1} \Big| \widehat{\mathcal{P}}_1 - \mathcal{P}_1 \Big|^{1-\delta_1} \cdots \Big| \mathcal{P}_{t-1} \Big|^{\delta_{t-1}} \Big| \widehat{\mathcal{P}}_{t-1} - \mathcal{P}_{t-1} \Big|^{1-\delta_{t-1}} \cdot T \Big(1 + \xi_{t,n,\delta} \Big) \zeta_{t,n} \\
&\qquad \text{(by Equation (29))} \\
&\leq C_{\pi^b}^{\frac{1}{2}} \sum_{t=1}^{T} \Bigg\{ \Big(\big| \mathcal{P}_1 \big|^{\delta_1} + \big| \widehat{\mathcal{P}}_1 - \mathcal{P}_1 \big|^{1-\delta_1} \Big) \cdots \Big(\big| \mathcal{P}_{t-1} \big|^{\delta_{t-1}} + \big| \widehat{\mathcal{P}}_{t-1} - \mathcal{P}_{t-1} \big|^{1-\delta_{t-1}} \Big) - \big| \mathcal{P}_1 \big|^{\delta_1} \cdots \big| \mathcal{P}_{t-1} \big|^{\delta_{t-1}} \Bigg\} \\
&\quad \cdot T \Big(1 + \xi_{t,n,\delta} \Big) \zeta_{t,n} \\
&\leq C_{\pi^b}^{\frac{1}{2}} \sum_{t=1}^{T} \Bigg[\Big(\big| 1 + \big| \widehat{\mathcal{P}}_1 - \mathcal{P}_1 \big| \Big) \cdots \Big(\big| 1 + \big| \widehat{\mathcal{P}}_{t-1} - \mathcal{P}_{t-1} \big| \Big) - 1 \Bigg] \cdot T \Big(1 + \xi_{t,n,\delta} \Big) \zeta_{t,n} \\
&\leq T C_{\pi^b}^{\frac{1}{2}} \sum_{t=1}^{T} \Bigg[\prod_{t'=1}^{t-1} \Big(1 + \xi_{t',n} \Big) - 1 \Bigg] \Big(1 + \xi_{t,n,\delta} \Big) \zeta_{t,n} \qquad \text{(by Equation (28))} \\
&\leq T C_{\pi^b}^{\frac{1}{2}} \sum_{t=1}^{T} \Bigg[\Big(1 + \xi_{t,n,\delta} \Big)^t - 1 \Bigg] \Big(1 + \xi_{t,n,\delta} \Big) \zeta_{t,n} \\
&\overset{(i)}{\lesssim} T C_{\pi^b}^{\frac{1}{2}} \sum_{t=1}^{T} t \, \xi_{t,n,\delta} \Big(1 + \xi_{t,n,\delta} \Big) \zeta_{t,n} \\
&\leq T^3 C_{\pi^b}^{\frac{1}{2}} \Big(1 + \xi_{T,n,\delta} \Big) \xi_{T,n,\delta} \zeta_{T,n}.
\end{aligned}
$$

The inequality (i) holds since $\left(1+\xi_{t,n,\delta}\right)^t \lesssim t\xi_{t,n,\delta}$ with $\xi_{t,n,\delta} < 1$, which is ensured by assuming a sufficiently large sample size n. Therefore, with the probability at least $1 - |\mathcal{A}|T^2\delta$, we obtain

$$\begin{aligned}\mathcal{E}(E_{2,2}) \lesssim & T^3 C_{\pi^b}^{\frac{1}{2}} \Big(1+\xi_{T,n,\delta}\Big)\xi_{T,n,\delta}\zeta_{T,n} \\ = & \frac{T^3}{n}(1-\theta^*)^{-1} C_{\pi^b}^{\frac{1}{2}} C_P^2 |\mathcal{O}|^{\frac{3T}{2}} |\mathcal{A}|^{\frac{3T}{2}} \sqrt{\log(|\mathcal{O}|^T|\mathcal{A}|^{T-1}/\delta)} \\ & + \frac{T^3}{n^{\frac{3}{2}}}(1-\theta^*)^{-\frac{3}{2}} C_{\pi^b}^{\frac{1}{2}} C_P^3 |\mathcal{O}|^{\frac{5T}{2}} |\mathcal{A}|^{\frac{5T}{2}} \log(|\mathcal{O}|^T|\mathcal{A}|^{T-1}/\delta)\end{aligned} \tag{30}$$

Combining with (27) and (30), with the probability of at least $1 - |\mathcal{A}|(T^2+T)\delta$, the upper bound of $\mathcal{E}(E_2)$ is

$$\begin{aligned}\mathcal{E}(E_2) \lesssim & \frac{T^{1.5}}{n}(1-\theta^*)^{-1} C_{\pi^b}^{\frac{1}{2}} C_P^2 |\mathcal{O}|^{\frac{3T}{2}} |\mathcal{A}|^{\frac{3T-1}{2}} \sqrt{\log(|\mathcal{O}|^T|\mathcal{A}|^{T-1}/\delta)} \\ & + \frac{T^3}{n}(1-\theta^*)^{-1} C_{\pi^b}^{\frac{1}{2}} C_P^2 |\mathcal{O}|^{\frac{3T}{2}} |\mathcal{A}|^{\frac{3T}{2}} \sqrt{\log(|\mathcal{O}|^T|\mathcal{A}|^{T-1}/\delta)} \\ & + \frac{T^3}{n^{\frac{3}{2}}}(1-\theta^*)^{-\frac{3}{2}} C_{\pi^b}^{\frac{1}{2}} C_P^3 |\mathcal{O}|^{\frac{5T}{2}} |\mathcal{A}|^{\frac{5T}{2}} \log(|\mathcal{O}|^T|\mathcal{A}|^{T-1}/\delta).\end{aligned}$$

C.1.6 Proof of Lemma 5

Lemma 5. *Let P be the $d_1 \times d_2$ matrix whose rows each sum to one. Let $\widehat{P}$ denote its entry-wise estimated counterpart, where each entry $\widehat{p}_{ij} \in (0,1)$ is an unbiased estimator of $p_{ij} \in (0,1)$ based on n_{ij} independent samples satisfying $\sum_{j=1}^{d_2} n_{ij} = N$ for each $i = 1, \ldots, d_1$. With probability at least $1-\delta$, the spectral norm bound holds that*

$$\left\|\widehat{P} - P\right\|_2 \lesssim \sqrt{\frac{d_1 \log(d_1/\delta)}{N}}.$$

Proof. Without loss of generality, we assume $d_1 > d_2$.

We begin by expressing $\widehat{P} - P$ as a sum of independent random matrices

$$\widehat{P} - P = \frac{1}{N}\sum_{k=1}^{N} E_k, \quad E_k = \begin{pmatrix} (E_k^{(1)})^\top \\ \vdots \\ (E_k^{(d_1)})^\top \end{pmatrix}, \quad E_k^{(i)} = X_k^{(i)} - P_i.$$

Here, $\widehat{P}_i = \frac{1}{N}\sum_{k=1}^{N} X_k^{(i)}$ is an empirical average of N independent multinomial trials, P_i denotes the i-th row of P, and $X_k^{(i)} \in \mathbb{R}^{d_2}$ is a one-hot vector drawn independently from a multinomial distribution with mean P_i.

Then, we apply the matrix Bernstein inequality to the sum $\sum_{k=1}^{N} E_k/N$. Note that $\mathbb{E}[E_k] = 0$, $\|E_k\| \le 2\sqrt{d_1}$, thus it is easy to check that $\sigma^2 = \frac{d_1}{N}$. By applying Lemma 6, for any $t > 0$, we have

$$\mathbb{P}\Big(\left\|\widehat{P} - P\right\|_2 \ge t\Big) = \mathbb{P}\Big(\Big\|\sum_{k=1}^{N} E_k/N\Big\|_2 \ge t\Big) \le (d_1+d_2)\exp\left(\frac{-t^2/2}{\frac{d_1}{N} + \frac{2\sqrt{d_1}}{N}t/3}\right).$$

Setting $t = C\sqrt{\frac{d_1\log(d_1/\delta)}{N}}$ for a constant, with probability at least $1-\delta$, the spectral norm of the deviation

$$\left\|\widehat{P} - P\right\|_2 \lesssim \sqrt{\frac{d_1 \log(d_1/\delta)}{N}}.$$

□

C.2 Bounding $\mathbb{E}[\sum_a \widehat{b}_{V,1}(a, O_1)] - \widehat{\mathbb{E}}[\sum_a \widehat{b}_{V,1}(a, O_1)]$

We first introduce the concept of Rademacher complexity, which is used to measure the size of function classes. Given any real-valued function class $\mathcal{G}$ defined over a random variable X and any radius $\delta > 0$, the population Rademacher complexity is defined as

$$\mathcal{R}(\mathcal{G}) = \mathbb{E}_{\epsilon,X}\left[\sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(X_i)\right|\right],$$

where $\{X_i\}_{i=1}^n$ are i.i.d. copies of X and $\{\epsilon_i\}_{i=1}^n$ are i.i.d. Rademacher random variables taking values in $\{-1, +1\}$ with equal probability. The empirical Rademacher complexity is given by

$$\mathcal{R}_n(\mathcal{G}) = \mathbb{E}_{\epsilon}\left[\sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i g(X_i)\right|\right].$$

If the function class $\mathcal{G}$ is bounded by T, by applying Lemma 10, the empirical Rademacher complexity can be bounded by

$$\mathcal{R}_n(\mathcal{G}) \le T\sqrt{\frac{2\log|\mathcal{G}|}{n}}. \tag{31}$$

Note that we can separate $\mathbb{E}[\sum_a \widehat{b}_{V,1}(a, O_1)] - \widehat{\mathbb{E}}[\sum_a \widehat{b}_{V,1}(a, O_1)]$ into two parts,

$$\begin{aligned}
&\left|\mathbb{E}\Big[\sum_a \widehat{b}_{V,1}(a, O_1)\Big] - \widehat{\mathbb{E}}\Big[\sum_a \widehat{b}_{V,1}(a, O_1)\Big]\right| \\
=&\left|\mathbb{E}\Big[\sum_a \widehat{b}_{V,1}(a, O_1) - \sum_a b^{\pi}_{V,1}(a, O_1)\Big] - \widehat{\mathbb{E}}\Big[\sum_a \widehat{b}_{V,1}(a, O_1) - \sum_a b^{\pi}_{V,1}(a, O_1)\Big]\right. \\
&\left. + \mathbb{E}\Big[\sum_a b^{\pi}_{V,1}(a, O_1)\Big] - \widehat{\mathbb{E}}\Big[\sum_a b^{\pi}_{V,1}(a, O_1)\Big]\right| \\
\le&\left|\mathbb{E}\Big[\sum_a \widehat{b}_{V,1}(a, O_1) - \sum_a b^{\pi}_{V,1}(a, O_1)\Big] - \widehat{\mathbb{E}}\Big[\sum_a \widehat{b}_{V,1}(a, O_1) - \sum_a b^{\pi}_{V,1}(a, O_1)\Big]\right| \\
&+ \left|\mathbb{E}\Big[\sum_a b^{\pi}_{V,1}(a, O_1)\Big] - \widehat{\mathbb{E}}\Big[\sum_a b^{\pi}_{V,1}(a, O_1)\Big]\right| \\
=&(a) + (b).
\end{aligned}$$

For (a), note that $\sum_a \widehat{b}_{V,1} - \sum_a b^{\pi}_{V,1} \in [-T, T]$. By applying Lemma 8 , with probability at least $1 - \delta$, we have

$$\begin{aligned}
(a) =&\left|\mathbb{E}\Big[\sum_a \widehat{b}_{V,1}(a, O_1) - \sum_a b^{\pi}_{V,1}(a, O_1)\Big] - \widehat{\mathbb{E}}\Big[\sum_a \widehat{b}_{V,1}(a, O_1) - \sum_a b^{\pi}_{V,1}(a, O_1)\Big]\right| \\
\le& \sup_{b\in\mathcal{B}_1}\left(\mathbb{E}\Big[\sum_a b(a, O_1) - \sum_a b^{\pi}_{V,1}(a, O_1)\Big] - \widehat{\mathbb{E}}\Big[\sum_a b(a, O_1) - \sum_a b^{\pi}_{V,1}(a, O_1)\Big]\right) \\
\le&\mathcal{R}_n(\mathcal{B}_1) + T\sqrt{\frac{\log(2/\delta)}{2n}} \\
\le&2c_0 T\sqrt{\frac{\log(c_1/\delta)}{n}} \qquad \text{(by Equation (31))}
\end{aligned} \tag{32}$$

where c_0 and c_1 are constants independent of n, T.

For (b), note that $\sum_a b^{\pi}_{V,1} \in [0, T]$. By using Hoeffding's inequality, with probability at least $1 - \delta$, we have

$$(b) = \left|\mathbb{E}\Big[\sum_a b^{\pi}_{V,1}(a, O_1)\Big] - \widehat{\mathbb{E}}\Big[\sum_a b^{\pi}_{V,1}(a, O_1)\Big]\right| \le c_0' T\sqrt{\frac{\log(c_1'/\delta)}{n}}, \tag{33}$$

where c_0 and c_1 are constants independent of n, T.

Combining (32) and (33), with probability at least $1-2\delta$, it holds that

$$\left|\mathbb{E}\Big[\sum_a \widehat{b}_{V,1}(a,O_1)\Big] - \widehat{\mathbb{E}}\Big[\sum_a \widehat{b}_{V,1}(a,O_1)\Big]\right| \lesssim T\sqrt{\frac{\log 1/\delta}{n}}.$$

D Additional Lemmas

Lemma 6 (Matrix Bernstein inequality, Tropp (2012))**.** *Suppose $X_1,\dots,X_n$ are mean-zero, $d_1\times d_2$ random matrices such that $\|X_i\|\le C$ almost surely for all $i\in\{1,\dots,n\}$. Then for any $t\ge 0$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| \ge t\right) \le (d_1+d_2)\exp\left\{\frac{-t^2/2}{\sigma^2+Ct/3}\right\}$$

where $\sigma^2 = \max\left\{\|\sum_{i=1}^n \mathbb{E}[X_i^\top X_i]\|, \|\sum_{i=1}^n \mathbb{E}[X_i X_i^\top]\|\right\}$.

Lemma 7 (McDiarmid's inequality)**.** *Suppose $X_1,\cdots,X_n\in\mathcal{X}$ are independent random variables and the function $f:\mathcal{X}^n\to\mathbb{R}$ is a mapping. For all $i\in\{1,\cdots,n\}$, and for all $x_1,\cdots,x_n,x_i'\in\mathcal{X}$, the function f satisfies*

$$|f(x_1,\cdots,x_{i-1},x_i,x_{i+1},\cdots,x_n) - f(x_1,\cdots,x_{i-1},x_i',x_{i+1},\cdots,x_n)| \le c_i,$$

then, we have

$$\mathbb{P}\left(\Big|f(X_1,\cdots,X_n) - \mathbb{E}f(X_1,\cdots,X_n)\Big| \ge t\right) \le 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Lemma 8. *Suppose $\mathcal{F}$ is a finite function class with $\|f\|_\infty\le M$, with probability at least $1-\delta$, we have*

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right| \le 2\mathcal{R}_n(\mathcal{F}) + M\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Proof. Using the symmetrization technique and McDiarmid's inequality (Lemma 7), for any $\delta>0$, we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right| > 2\mathcal{R}_n(\mathcal{F}) + \epsilon\right) \le 2\exp(-\frac{2n\epsilon^2}{M^2}).$$

Setting $2\exp\left(-\frac{2n\epsilon^2}{M^2}\right)=\delta$ and solving ϵ completes this probability inequality. □

Lemma 9 (Hoeffding's inequality)**.** *Suppose $X_1,X_2,\dots,X_n$ are independent random variables where $a\le X_i\le b$ almost surely. For any $t>0$, we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[\frac{1}{n}\sum_{i=1}^n X_i]\right| > t\right) \le 2\cdot\exp\left\{-\frac{2nt^2}{(b-a)^2}\right\}.$$

Lemma 10 (Massart's finite class lemma)**.** *Let $\mathcal{G}$ be a finite set of functions. Suppose that all functions in $\mathcal{G}$ are bounded, i.e., $\sup_{g\in\mathcal{G}}\|g\|_\infty\le B$. Then, the empirical Rademacher complexity is bounded by*

$$\mathcal{R}_n(\mathcal{G}) \le B\sqrt{\frac{2\log|\mathcal{G}|}{n}}.$$

Lemma 11 (Theorem 4.11 in Stewart and Sun (1990))**.** *Let $A,B\in\mathbb{R}^{m\times n}$ be the same-dimensional matrix with singular value*

$$\sigma_1(A)\ge\sigma_2(A)\ge\sigma_r(A),$$
$$\sigma_1(B)\ge\sigma_2(B)\ge\sigma_r(B).$$

Then for any unitarily invariant norm $\|\cdot\|$,

$$diag\big(\sigma_i(A)-\sigma_i(B)\big) \le \|A-B\|.$$

Lemma 12 (Unbiased empirical transition probability)**.** *The empirical transition probability estimator* $\widehat{P}(s'|s,a) = \frac{N(s,a,s')}{N(s,a)}$ *is unbiased, that* $\mathbb{E}[\widehat{P}(s'|s,a)] = P(s'|s,a)$*, where* $N(s,a,s') = \sum_{i=1}^{N} \mathbb{1}\{s_i = s, a_i = a, s_{i+1} = s'\}$ *and* $N(s,a) = \sum_{i=1}^{N} \mathbb{1}\{s_i = s, a_i = a\}$.

Proof. Given $N(s,a)$, $N(s,a,s')$ is a binomial variable such that

$$N(s,a,s')|N(s,a) \sim \text{Binomial}\big(N(s,a), P(s'|s,a)\big).$$

Thus, the expectation of this binomial variable is

$$\mathbb{E}\left[\frac{N(s,a,s')}{N(s,a)}\middle| N(s,a)\right] = P(s'|s,a).$$

By applying the law of total expectation, we have

$$\mathbb{E}\left[\frac{N(s,a,s')}{N(s,a)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{N(s,a,s')}{N(s,a)}\middle| N(s,a)\right]\right] = P(s'|s,a).$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We explicitly introduce our contributions in both the abstract and introduction. The introduction raises existing unexplored questions that we aim to address in this work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in the Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide complete proof for each theoretical result, and the assumptions and detailed proof are provided in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a detailed description of experimental details in the simulation section and explain the results to substantiate our claims.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Our work primarily focuses on theoretical analysis. Therefore, we only conduct simulation study.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide detailed simulation settings in the main text, including the structure of confounded POMDPs, the behavior policy used to collect data, and the two target policies to be evaluated.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our simulation studies justify our theoretical results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All simulations were performed on an Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential of confounded POMDPs to accurately model real-world decision-making tasks, such as precise medical applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release our code if this manuscript is acceptable.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: This research does not involve LLMs as any important, original, or non-standard components

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.