Offline Reinforcement Learning: Role of State Aggregation and Trajectory Data

Zeyu Jia[1*] zyjia@mit.edu
Alexander Rakhlin[1] rakhlin@mit.edu
Ayush Sekhari[1] sekhari@mit.edu
Chen-Yu Wei[2] chenyu.wei@virginia.edu

[1]Massachusetts Institute of Technology, [2]University of Virginia

Abstract

We revisit the problem of offline reinforcement learning with value function realizability but without Bellman completeness. Previous work by Xie and Jiang (2021) and Foster et al. (2022) left open the question whether a bounded concentrability coefficient along with trajectory-based offline data admits a polynomial sample complexity. In this work, we provide a negative answer to this question for the task of offline policy evaluation. In addition to addressing this question, we provide a rather complete picture for offline policy evaluation with only value function realizability. Our primary findings are threefold: 1) The sample complexity of offline policy evaluation is governed by the concentrability coefficient in an aggregated Markov Transition Model jointly determined by the function class and the offline data distribution, rather than that in the original MDP. This unifies and generalizes the ideas of Xie and Jiang (2021) and Foster et al. (2022). 2) The concentrability coefficient in the aggregated Markov Transition Model may grow exponentially with the horizon length, even when the concentrability coefficient in the original MDP is small and the offline data is *admissible* (i.e., the data distribution equals the occupancy measure of some policy). 3) Under value function realizability, there is a generic reduction that can convert any hard instance with admissible data to a hard instance with trajectory data, implying that trajectory data offers no extra benefits than admissible data. These three pieces jointly resolve the open problem, though each of them could be of independent interest.

1 Introduction

In offline Reinforcement Learning (RL), the learner aims to either find the optimal policy (policy optimization) or evaluate a given policy (policy evaluation) based on pre-collected data (offline data), i.e. without any live interaction with the environment. This approach is relevant in situations where real-time engagement is either impractical or costly, such as in safety-critical applications like healthcare and autonomous driving. The main hurdle in offline RL is the discrepancy between the data's distribution and that of the target policy, which can significantly compromise our ability to evaluate policies in the environment. Consequently, much of the offline RL research focuses on addressing this data-policy mismatch.

To deal with large state spaces in RL, function approximation is used to generalize the learned knowledge across states. A key approach under this umbrella is *model-based offline RL*, which involves constructing explicit models for transition and reward functions. Theoretical research in this direction studies the sample complexity under *model realizability*—where we assume access to a model class that contains the underlying model (Uehara and Sun, 2021). In this setting, the sample complexity of offline RL depends on the concentrability coefficient and the model class's complexity, where the former quantifies the mismatch

[*]Authors are listed in alphabetical order by last name.

between the distribution induced by the offline data and the target policy (formally defined in Section 2.2.1) and the latter quantifies the diversity within the model class. Unfortunately, in many complex tasks, model-based offline RL is not applicable due to prohibitively large model complexity.

Alternatively, *value-based offline RL* simplifies the process by approximating only the target policy's value function, typically resulting in lower complexity. Despite being a simpler and more direct approach, achieving polynomial sample complexity under this approach has historically required additional assumptions beyond a bounded concentrability coefficient and value function realizability, such as Bellman completeness (Chen and Jiang, 2019), β-incompleteness with $\beta < 1$ (Zanette, 2023), density-ratio realizability (Xie and Jiang, 2020), pushforward concentrability (Xie and Jiang, 2021), etc. The feasibility of attaining polynomial sample complexity for offline RL with solely bounded concentrability coefficient and realizability remained open for a long time.

Foster et al. (2022) gave a negative answer to this open problem, showing that realizability and concentrability alone cannot ensure polynomial sample complexity in offline RL. Their lower bound construction, however, relies on the offline data being generated unnaturally—e.g. the offline data includes samples from trajectories that will never be visited by any policy or only includes samples from just the first two steps in an episode, with the later steps hidden from the learner. Thus, the following question remained open:

Is offline RL statistically efficient under value function realizability, bounded concentrability, and natural offline data distribution such as trajectories generated by a single behavior policy?

Our work answers this question in the negative for the problem of *offline policy evaluation* (OPE). Specifically, we show that: There exist MDP instances where even though the offline data consists of *trajectories* generated by a single behavior policy, the value function is realized by a small value function class, and the concentrability coefficient is polynomial in the horizon length, the worst-case sample complexity for offline policy evaluation scales exponentially with the horizon.

En route to establishing the above result, we provide a comprehensive characterization for offline policy evaluation with value function approximation, and show that in order to achieve sample efficiency, the concept of "concentrability" needs to be strengthened. In particular, we show that the sample complexity for offline policy evaluation is both upper and lower bounded via the "aggregated concealability coefficient", that denotes the distribution mismatch in an aggregated transition model obtained by clubbing together transitions from the states that have indistinguishable value functions under the given value function class (formal details in Section 3). On the lower bound side, our results generalize the lower bounds of Foster et al. (2022) since we can show that the aggregated concentrability coefficient is exponentially large in their construction. On the upper bound side, our works recovers the results of Xie and Jiang (2021) since we can show that aggregated concentrability coefficient is upper bounded by the pushforward concentrability coefficient, a notion of distribution mismatch that is used in their upper bound.

Having identified that the aggregated concentrability coefficient is the key quantity that governs the sample complexity of offline policy evaluation, we provide a simple example where the offline data distribution is admissible (i.e., the offline data distribution equals the occupancy measure of an offline policy), the standard concentrability coefficient is $O(H^3)$, but the aggregated concentrability is $2^{\Omega(H)}$, where H is the horizon length. Since aggregated concentrability governs the sample complexity of offline policy evaluation, this implies that realizability, concentrability, and admissibility are not sufficient for sample efficient OPE.

Finally, in order to obtain the lower bound for trajectory data, we provide a generic reduction that converts the aforementioned hard instance for admissible data into a hard instance for trajectory data, thus showing that in the worst case, OPE is no easier for trajectory data than for admissible data.

2 Preliminaries

2.1 Markov Decision Process

We consider the finite horizon setting. A **Markov Transition Model** (MTM), denoted by $M = \mathrm{MTM}(\mathcal{X}, \mathcal{A}, T, H, \rho)$, is parameterized by a state space $\mathcal{X}$, an action space $\mathcal{A} = \{\mathfrak{a}_1, \mathfrak{a}_2, \ldots\}$, a transition kernel $T : \mathcal{X} \times \mathcal{A} \mapsto \Delta(\mathcal{X})$, horizon length $H \in \mathbb{N}$, and initial distribution $\rho \in \Delta(\mathcal{X})$. We assume that the state space $\mathcal{X}$ is layered across time, i.e., $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots \cup \mathcal{X}_H$ with $\mathcal{X}_i \cap \mathcal{X}_j = \varnothing$ for any $i \neq j$. The initial distribution $\rho \in \Delta(\mathcal{X}_1)$ specifies the state distribution that every episode starts with. The transition kernel $T(x' \mid x, a)$ for $x \in \mathcal{X} \setminus \mathcal{X}_H$, $x' \in \mathcal{X}$ and $a \in \mathcal{A}$ specifies the probability of transitioning to state x' if the learner takes action a on state x. By the layering structure, $T(\cdot \mid x, a)$ is supported on $\mathcal{X}_h$ if $x \in \mathcal{X}_{h-1}$.

For a MTM M and policy $\pi : \mathcal{X} \to \triangle(\mathcal{A})$, we let $\mathbb{E}^{M,\pi}[\cdot]$ denote the expectation under the following process: $x_1 \sim \rho$; and for $h = 1, \ldots, H$, action $a_h \sim \pi(\cdot \mid x_h)$, and next state $x_{h+1} \sim T(\cdot \mid x_h, a_h)$. The state occupancy measure for a particular layer h is defined as $d_h^\pi(x; M) := \mathbb{E}^{M,\pi}[\mathbb{I}\{x_h = x\}]$, and the state-action occupancy measure is defined as $d_h^\pi(x, a; M) := d_h^\pi(x; M)\pi(a \mid x)$.

A **Markov Decision Process** (MDP), denoted by $M = \mathrm{MDP}(\mathcal{X}, \mathcal{A}, T, r, H, \rho)$, is a Markov Transition Model augmented with a reward function $r : \mathcal{X} \times \mathcal{A} \to \Delta([-1, 1])$. For $h \leq H$, the state value function of a policy $V_h^\pi(\cdot; M) : \mathcal{X}_h \mapsto \mathbb{R}$ is defined such that for any $x_h \in \mathcal{X}_h$, $V_h^\pi(x_h; M)$ is the total cumulative reward obtained by starting at state x_h at timestep h, and acting according to the policy π till the end of the episode, i.e. $V_h^\pi(x; M) = \mathbb{E}^{M,\pi}\big[\sum_{k=h}^H r(x_k, a_k) \mid x_h = x\big]$. We similarly define the state-action value function $Q_h^\pi : \mathcal{X}_h \times \mathcal{A} \mapsto \mathbb{R}$ such that for any $x_h \in \mathcal{X}_h$, $Q_h^\pi(x, a) = \mathbb{E}^{M,\pi}[\sum_{h'=h}^H r(x_{h'}, a_{h'}) \mid (x_h, a_h) = (x, a)]$. Whenever clear from the context, we omit the dependency on M and simply write $d_h^\pi(x), V_h^\pi(x)$ and $V^\pi(\rho)$, etc.

For the ease of notation, for any policy π, we define $T(x' \mid x, \pi) = \mathbb{E}_{a\sim\pi(\cdot|x)}[T(x' \mid x, a)]$ and $r(x, \pi) = \mathbb{E}_{a\sim\pi(\cdot|x)}[r(x, a)]$, and if π is deterministic (i.e., $\pi(\cdot \mid x)$ is always supported on a single action for every $x \in \mathcal{X}$), we use $\pi(x) \in \mathcal{A}$ to denote the action it chooses on the state x. Furthermore, whenever clear from the context, we overload the notation and use $r(x, a)$ to denote the expected value of reward distribution $r(x, a)$. We let Π denote the set of all non-stationary Markovian policies in the underlying MDP.

2.2 Offline RL Preliminaries

Throughout the paper, we consider the offline RL setting. In this setting, the learner is equipped with an offline data distribution[1] and can only gather data about the MDP by i.i.d. sampling from this offline distribution (however, the learner does not know the density function of μ). We consider three types of offline data models:

- **General Data:** The offline dataset is characterized by an offline distribution $\mu = (\mu_1, \ldots, \mu_{H-1})$ where $\mu_h \in \Delta(\mathcal{X}_h \times \mathcal{A})$ for $h \leq H - 1$. An offline sample comprises of $H - 1$ many tuples $(x_h, a_h, r_h, x'_{h+1})$, where for each $h \in [H - 1]$, $(x_h, a_h) \in \mathcal{X}_h \times \mathcal{A}$ is drawn from μ_h, $r_h \in [-1, 1]$ is drawn from $r(x_h, a_h)$, and $x'_{h+1} \in \mathcal{X}_{h+1}$ is drawn from $T(\cdot \mid x_h, a_h)$.

- **Admissible Data:** Similar to the General Data setting, but with the addition requirement that there exists an offline policy $\pi_b \in \Pi$ such that for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, $\mu_h(x, a) = d_h^{\pi_b}(x, a)$.

- **Trajectory Data:** Each offline sample is a complete trajectory $(x_1, a_1, r_1, x_2, a_2, r_2, \ldots, x_H)$ sampled in the underlying MDP using an offline policy π_b.

[1] Various works in the offline RL literature assume that instead of direct sampling access to the offline distribution μ, the learner is given an offline dataset D of n samples drawn from μ. These two settings are equivalent upto sampling.

The goal of the learner is to estimate the value of a given evaluation policy π_e by collecting samples from the offline data distribution. In particular, given access to the offline distribution μ, the learner would like to estimate $V^{\pi_e}(\rho)$ up to an accuracy of ε, i.e. return a $\widehat{V}$ such that

$$\left|\widehat{V} - V^{\pi_e}(\rho)\right| \le \varepsilon. \tag{1}$$

We are interested in quantifying the amount of data required to achieve (1) with high probability. It is not hard to see that learning with general or admissible data is more challenging than learning with trajectory data, because from a trajectory dataset one can generate a (x, a, r, x') dataset with $\mu_h(x, a) = d_h^{\pi_b}(x, a)$, but not vice versa.

Comparison to the Definition of Admissible Data in Foster et al. (2022). A result in Foster et al. (2022) also claims to provide an exponential lower bound for admissible data but in the discounted MDP setting. Their data distribution $\mu(x, a) = \frac{1}{2}(d_1^{\pi_b}(x, a) + d_2^{\pi_b}(x, a))$ is not considered as admissible in our definition above[2]. We note that their lower bound construction heavily relies on samples being drawn only from the first two steps, and no information from the third step or later should be revealed. Our admissible data, on the other hand, forces the information to be revealed on all steps. Our lower bound for the stronger notion of admissible data serves as an important step towards the lower bound for trajectory data (see Section 4.2).

2.2.1 Concentrability Coefficient

Throughout the paper, for simplicity, we consider a fixed and deterministic evaluation policy π_e that takes action $\mathfrak{a}_1$ on all the states, i.e. $\pi_e(x) = \mathfrak{a}_1$ for all $x \in \mathcal{X}$. The *concentrability coefficient* (Munos (2003); Chen and Jiang (2019)) of policy π_e in an MDP $M = (\mathcal{X}, \mathcal{A}, T, r, H, \rho)$ with respect to offline data distribution μ is defined as

$$\mathsf{C}(M, \mu, \pi_e) = \max_h \max_{x \in \mathcal{X}_h, a \in \mathcal{A}} \frac{d_h^{\pi_e}(x, a)}{\mu_h(x, a)}. \tag{2}$$

Whenever clear from the context, we skip π_e and use the notation $\mathsf{C}(M, \mu)$ to denote $\mathsf{C}(M, \mu, \pi_e)$.

2.2.2 Function Approximation

The learner is given a function set $\mathcal{F}$ that consists of functions of the form $\mathcal{X} \times \mathcal{A} \to [-1, 1]$. We make the following realizability assumption that the Q-function belongs to the function class $\mathcal{F}$.

Assumption 2.1 (Realizability). *We have $Q^{\pi_e} \in \mathcal{F}$.*

Under value function realizability and provided an offline data distribution μ, a common goal for offline policy evaluation is to design an algorithm that achieves (1) with probability at least $1 - \delta$ using only $\mathrm{poly}(\log|\mathcal{F}|, H, 1/\delta, 1/\varepsilon, \mathsf{C}(M, \mu))$ samples. In the following sections, we argue that this is impossible, unless we replace $\mathsf{C}(M, \mu)$ with another notion of concentrability (see Section 3 for details).

2.2.3 Offline Policy Evaluation Problem

An Offline Policy Evaluation (OPE) problem $\mathfrak{g}$ is given by a tuple $(M, \pi_e, \mu, \mathcal{F})$ where M denotes the underlying MDP, μ denotes the offline data distribution, π_e denotes the evaluation policy, and $\mathcal{F}$ denotes a state-action value function class. Given an OPE instance $\mathfrak{g}$ and a parameter $\varepsilon > 0$, the goal of the learner is to

[2] Since Foster et al. (2022) consider the discounted setting, data from the first two steps is enough to provide bounded concentrability coefficient. This is not true in our finite-horizon case.

estimate the value of the policy π_e in the MDP M upto precision ε in expectation, by only relying on samples drawn from the offline distribution μ.

We say that the OPE problem $\mathfrak{g}$ is *realizable* if $Q^{\pi_e}(\cdot; M) \in \mathcal{F}$. Furthermore, whenever μ is admissible and there exists a policy π_b such that $\mu_h = d_h^{\pi_b}$ for all $h \le H$, we often denote the OPE problem as $\mathfrak{g} = (M, \pi_e, \pi_b, \mathcal{F})$. Finally, in the case of trajectory data, we still use the notation $\mathfrak{g} = (M, \pi_e, \pi_b, \mathcal{F})$ to denote the OPE problem but explicitly clarify, whenever invoked, that the learner now has access to complete trajectories sampled using π_b.

3 State Aggregation in Offline RL

We start by considering the offline policy evaluation problem with general offline data, and introduce useful tools and notation for our main lower bounds for admissible and trajectory data, and our upper bound, in the following sections.

There is a rich literature on understanding the right structural assumptions for offline RL with general offline data. For a warm-up, when the underlying MDP is tabular, i.e. has a small number of states and actions, it is well-known that the concentrability coefficient governs the statistical complexity of offline policy evaluation. To give some intuition for this claim, and to set the foundation for what follows, let $x^\star$ denote the state that maximizes the right-hand side in the definition of the concentrability coefficient in (2), and for simplicity, suppose that $d^{\pi_e}(x^\star) \ge \varepsilon$. Now, consider two scenarios, the first where the MDP has a reward of $+1$ by taking any action in the state $x^\star$, and the second where the MDP has a reward of -1 by taking any action in $x^\star$; in both cases, we assume zero reward on all other states. Thus, to estimate the value of π_e up to precision ε, the learner needs to distinguish between the two scenarios, and the only way to do so is to observe a transition from $x^\star$ in the given offline dataset, which requires at least $\frac{1}{\mu(x^\star, \pi_e(x^\star))} \ge \mathsf{C}$ offline samples from μ, in expectation. To conclude, in tabular MDPs, the learner can explicitly keep track of different states in the MDP, and use the corresponding transition and reward behavior on these states to evaluate π_e, and thus the worst case scenarios for offline policy evaluation is when the offline data does not provide enough information about the parts of the MDP where π_e has high visitation probability, and thus concentrability coefficient governs the statistical complexity.

The offline policy evaluation problem unfortunately becomes more challenging when the MDP has a large state space and the learner has to rely on function approximation. For this regime, previous works by Xie and Jiang (2021) and Foster et al. (2022) hint that the difficulty of offline policy evaluation comes from the hardness of distinguishing states that have different transition behaviors but the same values. Recall that every piece of data in the offline dataset is of the form (x, a, r, x'). If x_1 and x_2 are two states appearing in the dataset such that $\mathcal{F}$ does not provide any information to distinguish them, i.e., $f(x_1, \cdot) = f(x_2, \cdot)$ for all $f \in \mathcal{F}$, then the learner has no guidance from $\mathcal{F}$ whether they are essentially the same state or not in terms of their rewards or dynamics behavior. There are also no clues from other parts of the dataset, since with high probability, every state only appears at most once in the dataset due to the large state space. Under such a challenging scenario, intuitively, the best the learner can do is to aggregate these two states together and treat them as the same item, to get the most out of the offline dataset and the given value function class. This algorithmic idea of "aggregation" is precisely what is used in the BVFT algorithm of Xie and Jiang (2021). In this following section, we formalize the argument that aggregating indistinguishable states is indeed *the best the learner can do* by showing a general lower bound in terms of aggregated concentrability coefficient. To establish our lower bound, in the next section, we formally define the notion of state aggregation and aggregated concentrability coefficient.

3.1 Aggregated Concentrability Coefficient

Over a given state space $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_H$, we can define a *state aggregation scheme* $\Phi = \Phi_1 \cup \cdots \cup \Phi_H$ as below. For any h, Φ_h defines a partition of $\mathcal{X}_h$ so that the following hold:

1) Every element $\phi \in \Phi_h$ is a subset of $\mathcal{X}_h$;

2) The subsets are disjoint, i.e., $\phi \cap \phi' = \varnothing$ for all $\phi, \phi' \in \Phi_h$;

3) The subsets cover $\mathcal{X}_h$, i.e., $\bigcup_{\phi \in \Phi_h} \phi = \mathcal{X}_h$.

An aggregated Markov Transition Model $\bar{M}$ is defined via a underlying Markov Transition Model $M = (\mathcal{X}, \mathcal{A}, T, H, \rho)$, state aggregation schemes Φ_h, and offline data distributions $\mu_h : \mathcal{X}_h \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ for $1 \leq h \leq H-1$. We write $\bar{M} = (M, \Phi, \mu)$. The aggregated transition dynamics for a policy π are defined by

$$\bar{T}(\phi' \mid \phi, \pi; \bar{M}) = \frac{\sum_{x \in \phi} \sum_{x' \in \phi'} \sum_{a \in \mathcal{A}} \pi(a \mid x) \mu_h(x, a) T(x' \mid x, a)}{\sum_{x \in \phi} \sum_{a \in \mathcal{A}} \pi(a \mid x) \mu_h(x, a)} \tag{3}$$

for $\phi \in \Phi_h$, $\phi' \in \Phi_{h+1}$. The aggregated occupancy measure for a policy π is defined as

$$\bar{d}_h^{\pi}(\phi; \bar{M}) := \mathbb{E}\left[\mathbb{I}\{\phi_h = \phi\} \;\middle|\; \phi_1 \sim \bar{\rho}(\cdot), \phi_{i+1} \sim \bar{T}(\cdot \mid \phi_i, \pi; \bar{M}), \forall 1 \leq i \leq h-1 \right],$$

where the initial distribution $\bar{\rho}$ is defined as $\bar{\rho}(\phi) := \sum_{x \in \phi} \rho(x)$.

Note that in general, it may not be meaningful to define aggregated transitions with respect to *actions*, i.e., $\bar{T}(\phi' \mid \phi, a; \bar{M})$. This is because states in the same aggregation may not even share the same action space. However, in the special case where states within the same aggregation share the same action space, the quantity $\bar{T}(\phi' \mid \phi, a; \bar{M})$ can be defined, which could be useful in simplifying the notation. We use this notation in our lower bound proof (Appendix C).

Definition 3.1 (Aggregated Concentrability Coefficient)**.** *For an aggregated* MDP $\bar{M} = (M, \Phi, \mu)$ *with underlying* MDP M*, aggregation scheme* Φ*, and offline distribution* μ*, we define the aggregated concentrability coefficient* $\bar{\mathsf{C}}_\varepsilon(M, \Phi, \mu)$ *as*

$$\bar{\mathsf{C}}_\varepsilon(M, \Phi, \mu) = \max_h \max_{\mathcal{I}} \left\{ \frac{\sum_{\phi \in \mathcal{I}} \bar{d}_h^{\pi_e}(\phi)}{\sum_{\phi \in \mathcal{I}} \sum_{x \in \phi} \mu_h(x, \pi_e(x))} \;\middle|\; \mathcal{I} \subseteq \Phi_h, \;\; \sum_{\phi \in \mathcal{I}} \bar{d}_h^{\pi_e}(\phi) \geq \varepsilon \right\}.$$

The aggregated concentrability coefficient is analogous to the standard concentrability coefficient defined in (2), but now under the aggregated transition model. The reason why the sum of aggregated occupancy measure is restricted to be at least ε above is because those ϕ with extremely small occupancy can be fully ignored during the policy evaluation process, while making no impact in the estimation error even if the above ratio is large.

3.2 A General Lower Bound in Terms of Aggregated Concentrability Coefficient

We now have all the necessary tools to state our first lower bound. The following theorem provides a general reduction that lifts any given instance of a Markov Transition Model, evaluation policy, offline data distribution, and aggregation scheme into a class of offline policy evaluation problems, and provides a statistical lower bound for offline policy evaluation for this class in terms of the aggregated concentrability coefficient.

Theorem 3.1. *Let $\varepsilon \in (0,1)$, M be a Markov Transition Model, Φ be an aggregation scheme over the states of M, π_e be a deterministic evaluation policy in M such that for any aggregation $\phi \in \Phi$ and states $x, x' \in \phi$ it holds that $\pi_e(x) = \pi_e(x')$, and μ be a general offline data distribution with standard concentrability coefficient $\mathsf{C}(M, \mu)$ and aggregated concentrability coefficient $\bar{\mathsf{C}}_\varepsilon(M, \Phi, \mu)$. Then, there exists a class $\mathfrak{G}$ of realizable* OPE *problems such that for every* OPE *problem $\mathfrak{g} = (M^{(\mathfrak{g})}, \pi_e{}^{(\mathfrak{g})}, \mu^{(\mathfrak{g})}, \mathcal{F}^{(\mathfrak{g})})$ in $\mathfrak{G}$,*

(a) The function class $\mathcal{F}^{(\mathfrak{g})}$ satisfies $Q^{\pi_e}(\cdot; M^{(\mathfrak{g})}) \in \mathcal{F}^{(\mathfrak{g})}$ (Assumption 2.1), and $|\mathcal{F}^{(\mathfrak{g})}| = 2$.

(b) Any pair of states x, x' that belong to the same aggregation $\phi \in \Phi$ satisfy $f(x, \cdot) = f(x', \cdot)$ for all $f \in \mathcal{F}^{(g)}$.

(c) The concentrability coefficient $\mathsf{C}(M^{(\mathfrak{g})}, \mu^{(\mathfrak{g})}) = \Theta(\mathsf{C}(M, \mu))$.

Furthermore, any offline policy evaluation algorithm that guarantees to estimate the value of $\pi_e{}^{(\mathfrak{g})}$ in the MDP *$M^{(\mathfrak{g})}$ up to precision ε, in expectation, for every* OPE *problem $\mathfrak{g} \in \mathfrak{G}$ must use*

$$\widetilde{\Omega}\left(\frac{\bar{\mathsf{C}}_\varepsilon(M, \Phi, \mu)}{\varepsilon}\right)$$

offline samples from $\mu^{(\mathfrak{g})}$ in some OPE *problem $\mathfrak{g} \in \mathfrak{G}$.*

The proof of Theorem 3.1 is deferred to Appendix C. In the proof, instead of directly using the given MTM M to construct the class $\mathfrak{G}$, we construct Block MDPs $\{M^{(\mathfrak{g})}\}_{\mathfrak{g} \in \mathfrak{G}}$ with latent state dynamics given by M (with three additional new latent states per layer). As shown in the appendix, this reduction ensures that the standard concentrability remains unchanged. Furthermore, we note that the function class $\mathcal{F}^{(\mathfrak{g})}$ and the evaluation policy $\pi_e{}^{(\mathfrak{g})}$ are the same for all instances $\mathfrak{g} \in \mathcal{G}$, and that the aggregated concentrability coefficient in $M^{(g)}$ is $\Theta(\bar{\mathsf{C}}_\varepsilon(M, \Phi, \mu))$ (see Proposition 2). We also have the following property.

Property 1. *In the construction in Theorem 3.1, if the offline distribution μ is admissible for the Markov Transition Model M, then for every* OPE *problem $\mathfrak{g} \in \mathfrak{G}$, the offline distribution $\mu^{(\mathfrak{g})}$ is also admissible for the corresponding* MDP *$M^{(\mathfrak{g})}$.*

3.3 Can Aggregated Concentrability be Larger than Standard Concentrability?

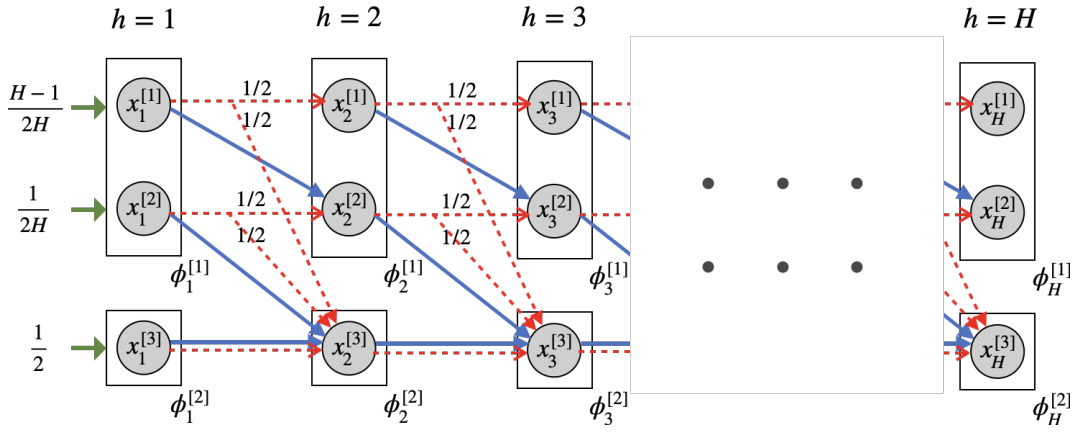


Figure 1: Markov Transition Model and aggregation scheme in Example 1. The blue arrows represent the transitions under action $\mathfrak{a}_1$, and the red arrows represent the transitions under $\mathfrak{a}_2$. The green arrows denote the initial distribution ρ.

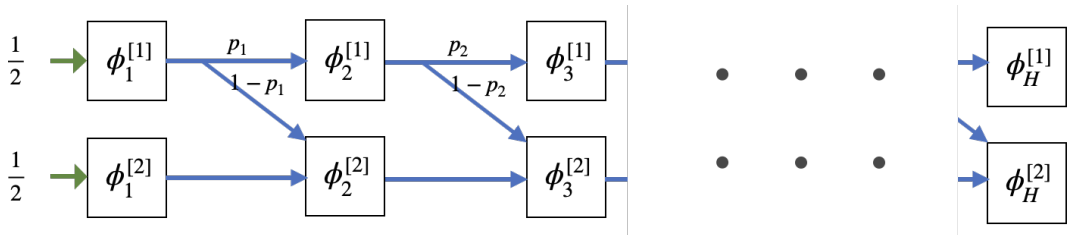


Figure 2: Dynamics for policy π_e in the aggregated MDP $\bar{M}$ in Example 1, where $p_h := \bar{T}\big(\phi_{h+1}^{[1]} \mid \phi_h^{[1]}, \pi_e; \bar{M}\big)$. As shown in Lemma 17, $p_h \geq \frac{H-1}{H+2}$ for all $1 \leq h \leq H-1$. The green arrows denote the initial distribution ρ in the aggregated MDP.

Theorem 3.1 indicates that the sample complexity of offline policy evaluation (with general data) grows with the aggregated concentrability coefficient $\bar{C}_\varepsilon(M, \Phi, \mu)$ instead of the standard concentrability coefficient $C(M, \mu)$. Given this lower bound, one may wonder how large can the aggregated concentrability be in comparison to the standard concentrability. In this section, we will demonstrate via an example that the gap could indeed be exponential.

Example 1. *Consider the example represented in Figure 1, where*

- ***Markov Transition Model*** M *consists of three states* $\{x_h^{[1]}, x_h^{[2]}, x_h^{[3]}\}$ *in each layer* $h \in [H]$*, and two actions* $\mathcal{A} = \{\mathfrak{a}_1, \mathfrak{a}_2\}$*. The initial distribution* ρ *(denoted by the solid green arrows) is defined such that* $\rho\big(x_1^{[1]}\big) = (H-1)/2H$*,* $\rho\big(x_1^{[2]}\big) = 1/2H$ *and* $\rho\big(x_1^{[3]}\big) = 1/2$*.*

  The transition dynamics are identical for every layer $h \in [H-1]$*, and is defined such that on taking action* $\mathfrak{a}_1$ *the agent transitions deterministically according to the solid blue arrows, and on taking action* $\mathfrak{a}_2$*, the agent transitions stochastically according to the dotted red arrows.*

- *For every* $h \in [H]$*, the* ***aggregation scheme*** $\Phi_h = \{\phi_h^{[1]}, \phi_h^{[2]}\}$ *with the aggregation* $\Phi_h^{[1]} = \{x_h^{[1]}, x_h^{[2]}\}$ *and* $\phi_h^{[2]} = \{x_h^{[3]}\}$*. The aggregated states are denoted by rectangular blocks.*

- *The* ***offline distribution*** μ *is the occupancy measure of an offline policy* π_b *defined such that* $\pi_b(x) = \frac{1}{H^2}\delta_{\mathfrak{a}_1}(\cdot) + \frac{H^2-1}{H^2}\delta_{\mathfrak{a}_2}(\cdot)$ *for all* $x \in \mathcal{X}$

- *The* ***evaluation policy*** π_e *such that* $\pi_e(x) = \delta(\mathfrak{a}_1)$ *for all* $x \in \mathcal{X}$*.*

Proposition 1. *For* $\varepsilon \leq 1/15$*, the Markov Transition Model* M*, aggregation scheme* Φ*, evaluation policy* π_e *and offline distribution* μ *given in Example 1, the standard concentrability coefficient* $C(M, \mu) = O(H^3)$*, whereas the aggregated concentrability coefficient* $\bar{C}_\varepsilon(M, \Phi, \mu) = \widetilde{\Omega}(2^H)$*.*

We now give a sketch for the proof of Proposition 1, with the full details deferred to Appendix D. We first calculate the upper bound on the standard concentrability coefficient. First, we argue that for $h \geq 3$, $d_h^{\pi_e}(x_h) = 1$ if $x_h = x_h^{[3]}$ and 0 otherwise. This can be easily observed from the transition of π_e (blue arrow) in Figure 1—following the blue arrow, the policy must stay in $x_h^{[3]}$ for $h \geq 3$. Next, we lower bound the state occupancy under π_b. We claim that

$$d_h^{\pi_b}\big(x_h^{[1]}\big) \geq \Omega\left(\frac{1}{2^h}\right), \quad d_h^{\pi_b}\big(x_h^{[2]}\big) \geq \Omega\left(\frac{1}{H2^h}\right), \quad \text{and} \quad d_h^{\pi_b}\big(x_h^{[3]}\big) \geq \frac{1}{2}. \tag{4}$$

The third inequality in (4) is easy to see since the occupancy on $x_h^{[3]}$ is non-decreasing w.r.t. h under any policy (Figure 1). To see the first two inequalities in (4), notice that since π_b chooses $\mathfrak{a}_2$ with probability $1 - \frac{1}{H^2}$, and $\mathfrak{a}_2$ carries $\frac{1}{2}$ of the weights from $x_h^{[1]}$ to $x_{h+1}^{[1]}$ (depicted by the red arrow in Figure 1), we have

$d_h^{\pi_b}(x_h^{[1]}) \geq \rho(x_1^{[1]})\left(\frac{1}{2}(1-\frac{1}{H^2})\right)^{h-1} = \Omega\left(\frac{1}{2^h}\right)$. Similarly, $d_h^{\pi_b}(x_h^{[2]}) \geq \rho(x_1^{[2]})\left(\frac{1}{2}(1-\frac{1}{H^2})\right)^{h-1} = \Omega\left(\frac{1}{H2^h}\right)$. With the calculation above and the fact that $\frac{\pi_e(a|x)}{\pi_b(a|x)} \leq H^2$, we conclude that $\mathsf{C} \leq H^2 \max_h \max_x \frac{d_h^{\pi_e}(x)}{d_h^{\pi_b}(x)} \leq O(H^3)$.

We now proceed to the lower bound on aggregated concentrability coefficient. From (3), we know that the aggregated dynamic for π_e, shown in Figure 2, is constructed by reweighting the transition in the original transition model using $\mu_h(\cdot) = d_h^{\pi_b}(\cdot)$. Since π_b takes action $\mathfrak{a}_2$ with large probability, and $\mathfrak{a}_2$ does not change the relative weight $\frac{d_h^{\pi_b}(x_h^{[1]})}{d_h^{\pi_b}(x_h^{[2]})}$ (see the red arrows in Figure 1), it can be shown that $\frac{d_h^{\pi_b}(x_h^{[1]})}{d_h^{\pi_b}(x_h^{[2]})} \geq \frac{\rho(x_1^{[1]})}{3\rho(x_1^{[2]})} = \frac{H-1}{3}$ for all h. This gives

$$p_h := \bar{T}(\phi_{h+1}^{[1]} \mid \phi_h^{[1]}, \pi_e) = \frac{d_h^{\pi_b}(x_h^{[1]}) \cdot 1 + d_h^{\pi_b}(x_h^{[2]}) \cdot (1/2)}{d_h^{\pi_b}(x_h^{[1]}) + d_h^{\pi_b}(x_h^{[2]})} \geq \frac{H-1}{H+2},$$

where the factors of 1 and $1/2$ are the probability of transitioning to $\phi_{h+1}^{[1]}$ from $x_h^{[1]}$ and $x_h^{[2]}$ following π_e. This further implies $\bar{d}_h^{\pi_e}(\phi_h^{[1]}) \geq \frac{1}{2}\left(\frac{H-1}{H+2}\right)^h$. Using a similar argument as for (4), we have $\bar{d}_h^{\pi_b}(\phi_h^{[1]}) \leq \frac{1}{2^h}$. These two bounds together imply that the aggregated concentrability is $2^{\Omega(H)}$.

4 Main Lower Bounds for Offline Policy Evaluation

4.1 Admissible Data

Example 1 provides an instance of a Markov Transition Model, aggregation scheme, evaluation policy and offline distribution for which the standard concentrability is $O(H^3)$ whereas the aggregated concentrated is $2^{\Omega(H)}$. Since the offline distribution μ in Example 1 is the occupancy measure d^{π_b} for the policy π_b, plugging Example 1 in Theorem 3.1 implies the following lower bound for offline policy evaluation with admissible offline data.

Theorem 4.1. *Let $\varepsilon \leq 1/15$, and horizon $H \geq 1$. Then, there exits a class $\mathfrak{G}_{\text{ADM}}$ of realizable* OPE *problems, such that for every* OPE *problem $\mathfrak{g} = (M^{(\mathfrak{g})}, {\pi_e}^{(\mathfrak{g})}, \mu^{(\mathfrak{g})}, \mathcal{F}^{(\mathfrak{g})}) \in \mathfrak{G}_{\text{ADM}}$, the concentrability coefficient of ${\pi_e}^{(\mathfrak{g})}$ w.r.t. $\mu^{(\mathfrak{g})}$ is $O(H^3)$, the offline distribution $\mu^{(\mathfrak{g})}$ is admissible for the* MDP *$M^{(\mathfrak{g})}$, and $|\mathcal{F}^{(\mathfrak{g})}| = 2$.*

Furthermore, any offline policy evaluation algorithm that guarantees to estimate the value of ${\pi_e}^{(\mathfrak{g})}$ in the MDP *$M^{(\mathfrak{g})}$ up to precision ε, in expectation, for every* OPE *problem $\mathfrak{g} \in \mathfrak{G}_{\text{ADM}}$ must use $2^{\Omega(H)}$ offline samples in some $\mathfrak{g} \in \mathfrak{G}_{\text{ADM}}$.*

The construction of the class $\mathfrak{G}_{\text{ADM}}$, and the proof of Theorem 4.1, are deferred to Appendix D. We remark that in all the OPE problem instances $\mathfrak{g} \in \mathfrak{G}_{\text{ADM}}$, the corresponding MDPs $M^{(\mathfrak{g})}$ share the same action space $\mathcal{A} = \{a_1, a_2\}$ (binary actions), state space $\mathcal{X}$ and horizon H, however, the transition dynamics, reward function and initial distribution could change with the instance. Furthermore, the policy ${\pi_e}^{(\mathfrak{g})}$ and the state-action value function class $\mathcal{F}^{(\mathfrak{g})}$ are also same across all instances $\mathfrak{g} \in \mathfrak{G}_{\text{ADM}}$.

Our lower bound in Theorem 4.1 considers admissible offline data distributions, where for any instance $\mathfrak{g} \in \mathfrak{G}_{\text{ADM}}$ and $h \leq H$, the offline distribution $\mu_h^{(\mathfrak{g})} = {d_h^{\pi_b}}^{(\mathfrak{g})}(\cdot; M^{(\mathfrak{g})})$, and the offline algorithm can draw samples of the form (x_h, a_h, r_h, x_{h+1}) from the process $(x_h, a_h) \sim \mu_h^{(\mathfrak{g})}$, $r_h \sim r^{(\mathfrak{g})}(x_h, a_h)$ and $x_{h+1} \sim T^{(\mathfrak{g})}(\cdot \mid x_h, a_h)$. Thus, Theorem 4.1 strengthens over the results of Foster et al. (2022), in which the offline data distribution is not equal to the occupancy measure of a single policy.

Having shown that bounded concentrability coefficient and realizability alone are not sufficient for statistically efficient offline policy evaluation, even if the offline distribution $\mu_h = d_h^{\pi_b}$ is admissible, we now ask what happens if the learner has access to complete offline trajectories $(x_1, a_1, r_1, s_2, \ldots, r_H, s_H)$ sampled using π_b. Unfortunately, for this scenario, the result of Theorem 4.1 no longer holds. This is because

the reduction in Theorem 3.1, which is a key tool in the proof of Theorem 4.1, does not prevent from leaking additional information when the learner has access to trajectories of length more than 2. In particular, by looking at the conditional distributions of x_3 after fixing actions a_1 and a_2 for the first two timesteps in that construction (which can be computed when given trajectory data that covers the first two timesteps), the learner can infer the value of π_e in the underlying MDP. In the next section, we develop additional tools to handle trajectory data.

4.2 Trajectory Data

In many real world applications, the offline dataset is collected by sampling trajectories of the form $(x_1, a_1, r_1, x_2, \ldots, x_H, a_H, r_H)$ and it remains to address whether access to the entire H-length trajectory instead of just the tuples (x, a, r, x') can allow the learner to circumvent the challenges introduced in previous subsections. In fact, Foster et al. (2022) left it as an open problem whether access to trajectory data can make offline RL statistically tractable. In this section, we answer this in the negative and show that in the worst case, access to trajectory data does not overcome the statistical inefficiencies of offline RL with just bounded concentrability coefficient and realizability.

Theorem 4.2. *Let $\varepsilon \le 1/15$, and horizon $H \ge 1$. Then, there exits a class $\mathfrak{G}_{\text{TRAJ}}$ of realizable* OPE *problems, such that for every* OPE *problem $\mathfrak{g} = (M^{(\mathfrak{g})}, {\pi_e}^{(\mathfrak{g})}, {\pi_b}^{(\mathfrak{g})}, \mathcal{F}^{(\mathfrak{g})}) \in \mathfrak{G}_{\text{TRAJ}}$, the learner has access to offline trajectories sampled using ${\pi_b}^{(\mathfrak{g})}$, the concentrability coefficient of ${\pi_e}^{(\mathfrak{g})}$ w.r.t. ${\pi_b}^{(\mathfrak{g})}$ is $O(H^3)$, and $|\mathcal{F}^{(\mathfrak{g})}| = 2$.*

Furthermore, any offline policy evaluation algorithm that estimates the value of ${\pi_e}^{(\mathfrak{g})}$ in the MDP *$M^{(\mathfrak{g})}$ up to precision $\varepsilon/(16H)$, in expectation, for every* OPE *problem $\mathfrak{g} \in \mathfrak{G}_{\text{TRAJ}}$ must use $2^{\Omega(H)}$ offline trajectories in some $\mathfrak{g} \in \mathfrak{G}_{\text{TRAJ}}$.*

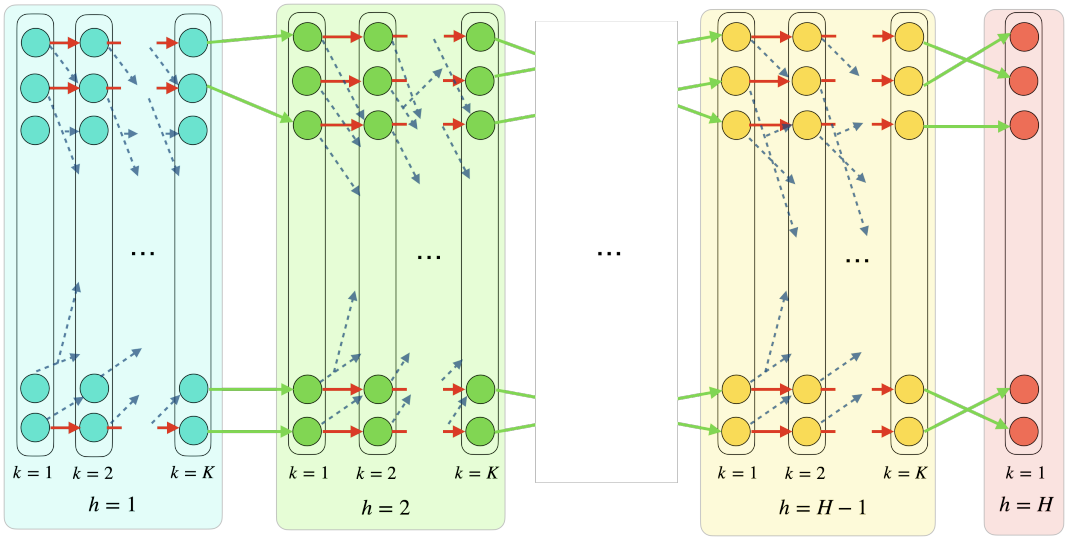


Figure 3: Lower bound construction for the proof of Theorem G.1. For each $h \in [H]$, the corresponding block denotes the K layers that are obtained using REPLICATOR by replicating the h-th layer in the given MDP M for K many times. The solid red arrows represent the transitions under the action $\mathfrak{a}_1$, the dotted blue arrows represent the transitions under the $\mathfrak{a}_2$ (under which we resample from the admissible distribution μ_h), the solid green arrows denote the transitions according to the original MDP M.

While the full proof is deferred to Appendix E, we present the main ideas and the key tools below. The primary reason why the lower bound from Theorem 4.1 does not hold under trajectory data is that access to trajectories spanning more than two timesteps in the underlying MDPs in that construction leaks additional

information, which can be exploited by the learner to evaluate π_e. In particular, given trajectory data, the learner can compute the marginal distribution over x_{h+2} given actions a_h and a_{h+1}, for $h \le H-2$, which can be used to identify the underlying instance in the class $\mathfrak{G}_{\text{ADM}}$ and thus compute π_e. Our key insight in the proof is to fix this problem of information leakage by introducing a general-purpose reduction from offline RL with admissible data to offline RL with trajectory data, which may be of independent interest. This reduction is obtained by using two new protocols called (a) the REPLICATOR protocol, and (b) the ADMISSIBLE-TO-TRAJECTORY protocol, which we describe below.

REPLICATOR: Given in Algorithm 1 in the appendix, the REPLICATOR protocol takes as input a realizable OPE problem $\mathfrak{g} = (M, \pi_b, \pi_e, \mathcal{F})$ where the MDP M has horizon H, and a parameter K, and converts it into another realizable OPE problem $\widetilde{\mathfrak{g}} = (\widetilde{M}, \widetilde{\pi}_b, \widetilde{\pi}_e, \mathcal{F})$ where the new MDP $\widetilde{M}$ has horizon $\widetilde{H} = (H-1)K+1$. We require that REPLICATOR satisfies the following property.

Property 2 (Informal; Formal version in Lemma 20)**.** *The realizable* OPE *problem* $\widetilde{\mathfrak{g}} \leftarrow \text{REPLICATOR}(\mathfrak{g}, K)$ *satisfies the following:*

(a) *Concentrability coefficient* $\sup_{\widetilde{h} \le \widetilde{H}} \left\| \frac{d_h^{\widetilde{\pi}_e}}{d_h^{\widetilde{\pi}_b}} \right\|_\infty \le 2 \sup_{h \le H} \left\| \frac{d_h^{\pi_e}}{d_h^{\pi_b}} \right\|_\infty$.

(b) *The value of the policy* $\widetilde{\pi}_e$ *in* $\widetilde{M}$ *is equal to the value of* π_e *in* M.

Our construction of REPLICATOR essentially replicates each layer in the given MDP M for K times (except for the last layer); see Figure 3 for illustration. In the following, we call these replicated layers as *sub-layers*. We first define the transition function. For the last sub-layer (i.e., for $k = K$) of each layer $h \le H-1$, the transition is exactly the same as that in the MDP from layer h to $h+1$ (denoted by the green arrows in Figure 3). For other sublayers with $k < K$, the transitions are designed such that: if the action $\mathfrak{a}_1$ is taken, then the state transitions to the same state in the next sub-layer (red arrows in Figure 3); and if the action $\mathfrak{a}_2$ is taken, the next state is sampled independently from the offline data distribution $\mu_h = d_h^{\pi_b}$ (blue arrows in Figure 3). Furthermore, the evaluation policy $\widetilde{\pi}_e$ in the new MDP is the same as π_e, and takes action $\mathfrak{a}_1$ on all states. The offline policy $\widetilde{\pi}_b$ is set as π_b for the last sub-layer (i.e. $k = K$), and is set as $\frac{1}{2}(\delta_{\mathfrak{a}_1} + \delta_{\mathfrak{a}_2})$ for the intermediate sub-layers with $k \le K-1$.

The rationale behind this design is that since $\widetilde{\pi}_b = \frac{1}{2}(\delta_{\mathfrak{a}_1} + \delta_{\mathfrak{a}_2})$, for each h, with probability $1 - 2^{-K+1}$ the offline policy will choose $\mathfrak{a}_2$ at least once in sub-layers $k = 1, \dots, K-1$. If $\mathfrak{a}_2$ is chosen at least once, then the state distribution at $k = K$ is equal to $\mu_h = d^{\pi_b}$ and independent from all previous layers $1, \dots, h-1$. As long as K is large enough, this happens with very high probability, which makes the offline data distribution at sub-layer $k = K$ resemble admissible data with distribution $\mu_h = d_h^{\pi_b}$, even when the data is actually a complete trajectory. It can be shown that this conversion preserves the concentrability coefficient up to a constant factor.

ADMISSIBLE-TO-TRAJECTORY: Given in Algorithm 2 in the appendix, this protocol takes as input K tuples of the form (x_h, a_h, r_h, x_{h+1}) sampled from an admissible offline distribution $\mu_h = d_h^{\pi_b}$, for every $h \in [H]$, and returns a trajectory $\widetilde{\tau}$ of length $\widetilde{H}$ in $\widetilde{M}$. We require that ADMISSIBLE-TO-TRAJECTORY satisfies the following property.

Property 3 (Informal; Formal version in Lemma 21)**.** *For a large class of offline policies* π_b*, the distribution of trajectory* $\widetilde{\tau}$*, constructed by* ADMISSIBLE-TO-TRAJECTORY *using offline data tuples from* d^{π_b}*, is close to the distribution of trajectories* $\widetilde{\tau}'$ *obtained using* $\widetilde{\pi}_b$ *in* $\widetilde{M}$.

The idea of ADMISSIBLE-TO-TRAJECTORY is straightforward: We already argue that REPLICATOR can simulate admissible data using trajectory data. Hence, with a reverse process, given an admissible dataset, we can create a new trajectory dataset in a new MDP that simulates the original admissible dataset.

With the above two protocols, the reduction of offline RL with admissible data to trajectory data is straightforward and is stated in Algorithm 3 in the appendix. At a high level, given a realizable OPE problem g with admissible offline distribution d^{π_b}, for some large enough K, we use the REPLICATOR protocol to create a realizable OPE problem $\widetilde{g}$ and use the ADMISSIBLE-TO-TRAJECTORY protocol to generate trajectory data corresponding to $\widetilde{\pi}_b$ in $\widetilde{M}$. Since, Property 2-(a) implies that the concentrability coefficient stays bounded and Property 2-(b) implies that the value to be evaluated remains unchanged, the above reduction provides a way to solve offline RL with admissible data by invoking an offline algorithm that requires trajectory data. Thus, if trajectory data with bounded concentrability coefficient is tractable, then so is admissible data by leveraging Algorithm 3, which contradicts Theorem 4.1. This implies that offline RL with trajectory data must also be statistically inefficient. The formal proof is deferred to Appendix E.

5 Upper Bound

Our lower bounds show that the worst-case sample complexity of offline policy evaluation grows with the aggregated concentrability. In this section, we complement our lower bounds with an upper bound of the form $\mathrm{poly}(\bar{\mathsf{C}}, H, \varepsilon^{-1}, \log|\mathcal{F}|)$. Taken together, the lower and upper bound suggest that aggregated concentrability, but not the standard concentrability, characterize the worst-case sample complexity of offline policy evaluation with value function approximation.

Theorem 5.1. *Let $\varepsilon > 0$, $\mathcal{F}$ be a state-action value function class that satisfies Assumption 2.1, π_{e} be an evaluation policy and μ be an offline (general) data distribution. Then, Algorithm 4 (a adaptation of the* BVFT *algorithm of Xie and Jiang (2021)) returns a $\widehat{V}$ such that $\left|\mathbb{E}_{x\sim\rho}[V^{\pi_{\mathrm{e}}}(x) - \widehat{V}]\right| = O(\varepsilon)$ after collecting*

$$n = \mathcal{O}\left(\frac{\bar{\mathsf{C}} \cdot H^6 \log(|\mathcal{F}|/\delta)}{\varepsilon^4}\right)$$

many (offline) samples from μ, where $\bar{\mathsf{C}} := \max_{f,f'\in\mathcal{F}} \bar{\mathsf{C}}_{\varepsilon^2/H^2}(M, \Phi(f,f'), \mu)$ and $\Phi(f,f')$ is the state aggregation scheme determined by $f, f' \in \mathcal{F}$ (see Definition F.2 for the precise definition).

The proof of Theorem 5.1 is deferred to Appendix F, wherein we also provide a generalization of this result that accounts for misspecification error in Assumption 2.1, and an upper bound for a slightly more challenging scenario where the learner only has access to the state value function class (instead of the state-action value function class). Note that the upper bound depends on the aggregations scheme $\Phi(f,f')$. The appearance of such an aggregation scheme in the upper bound is not surprising. In our lower bound in Theorem 3.1, while Φ is given as an input, the corresponding value function class $\mathcal{F}$ constructed for the class $\mathfrak{G}$ satisfies that $\Phi(f,f') = \Phi$ (see item-(b) in Theorem 3.1).

Algorithm 4 is an adaptation of the BVFT algorithm (Xie and Jiang, 2021) for offline policy evaluation. At a high-level, the key idea in the algorithm is to solve a minimax problem (with the objective determined by Bellman error) over pairs $(f,f') \in \mathcal{F}\times\mathcal{F}$, where for each pair, the algorithm creates a "tabular problem" by aggregating states with the same $(f(x), f'(x))$ value, and estimates the Bellman error for this tabular problem. Intuitively, this is probably the best the learner can do, since besides the value of $(f(x))_{f\in\mathcal{F}}$, the learner has no other ways to distinguish states in the large state space. Thus, due to aggregation, the upper bounds depends on aggregated concentrability coefficeint rather than the standard concentrability coefficient.

We remark that while Xie and Jiang (2021) do not present their upper bound in terms of the aggregated concentrability, this quantity already appears in their analysis (see Appendix C in Xie and Jiang (2021)). However, their final bound is represented with a stronger version of concentrability coefficient C_{pf} (*pushforward concentrability coefficient*, formally defined in Definition B.1). It is straightforward to show $\bar{\mathsf{C}} \le \mathsf{C}_{\mathrm{pf}}$ (Lemma 29). Our analysis basically follows theirs, but along the way does not relax $\bar{\mathsf{C}}$ to C_{pf}.

6 Conclusion

Our paper considers the problem of offline policy evaluation with value function approximation, where the function class does not satisfy Bellman completeness, and shows that its sample complexity is characterized by the aggregated concentrability coefficient—a notion of distribution mismatch in an aggregated MDP obtained by clubbing together transitions from the states that have indistinguishable value functions under the given value function class (formal details in Section 3). We provide an example of an MDP where the aggregated concentrability coefficient could be exponentially larger than the concentrability coefficient, using which we conclude that statistically efficient offline policy evaluation is not possible with bounded concentrability coefficient even if we assume access to trajectory data. This result thus highlights the necessity for further research into designing more effective strategies for dealing with the complexities inherent in offline reinforcement learning environments. We conclude with some discussion:

Could the Aggregated Concentrability be Smaller Than the Standard Concentrability? In Section 3.3, we demonstrated via an example that the aggregated concentrability can be exponentially larger than the standard concentrability. However, it is also quite easy to come up with situations where the aggregated concentrability is actually smaller than the standard one. For example, suppose the aggregations scheme $\Phi_h = \{\phi_h\}$ with $\phi_h = \mathcal{X}_h$, i.e. all states belong to a single aggregation. Here, the aggregated concentrability coefficient is exactly 1 since each layer has only one aggregation, whereas the standard concentrability coefficient could be arbitrary.

Gap Between Upper and Lower Bounds in terms of Dependence on ε The sample complexity in our upper bound (Theorem 5.1) scales with $\frac{1}{\varepsilon^4}$ instead of the more common $\frac{1}{\varepsilon^2}$. This is similar to Xie and Jiang (2021) and is because we divide the state space into $O(\frac{1}{\varepsilon^2})$ aggregations, each of which consists of states having the same value functions up to an accuracy ε. On the other hand, our lower bound has a $\Omega(\frac{1}{\varepsilon})$ dependence instead of the more common $\Omega(\frac{1}{\varepsilon^2})$. Improving the dependence on ε in either the upper or lower bounds is an interesting future research direction.

Connections to Other Notations of Concentrability. Various other notions of concentrability like pushforward concentrability C_{pf}, and action concentrability $\mathsf{C}_{\mathcal{A}}$ Definition B.1 are considered in the literature (Xie and Jiang, 2021). We show that $\bar{\mathsf{C}} \le \mathsf{C}_{\text{pf}}$ (Lemma 29) and $\bar{\mathsf{C}} \le (\mathsf{C}_{\mathcal{A}})^H$ (Lemma 30). Note that the sample complexity bound of $O\big((\mathsf{C}_{\mathcal{A}})^H\big)$ is also what we get by using *importance sampling* to perform offline policy evaluation.

Single Policy vs. All Policy Concentrability. The notions of *realizability* and *concentrability coefficient* adopted in our paper are only with respect to the given *evaluation policy*. This is also called *single policy concentrability* and is the standard assumption in offline policy evaluation literature. An alternative assumption that is used in the offline policy optimization literature is that of *all policy concentrability*, which states that the concentrability coefficient for all possible policies (in the MDP) w.r.t. the given offline data is bounded. While we restrict our discussions to the former, our construction in Example 1 has bounded concentrability coefficient for all policies (Appendix D). An interesting future research direction is to extend our lower bounds to the offline policy optimization setting.

Role of Realizable Value Function Class in Offline RL. In this paper, we considered the realizable setting (Assumption 2.1) where the learner has access to a value function class that contains Q^{π_e}, and showed that the statistical complexity of offline policy evaluation is governed by the aggregated concentrability coefficient for the aggregation scheme induced by the given function class. However, how important is this access to the

value function class? In particular, is statistically efficient offline RL feasible in the agnostic setting where the learner does not have any value function class? Unfortunately, as we show in Appendix G, agnostic offline policy evaluation is not statistically tractable in the worst case even when the learner is given trajectory offline data that has bounded pushforward concentrability coefficient (recall Lemma 29 that this implies bounded aggregated concentrability coefficient for any aggregation scheme). Hence, further structural assumptions on the underlying MDP or the policies are needed for tractable learning. Sekhari et al. (2021); Jia et al. (2024) explored some structural assumptions that enable agnostic learning in the online RL setting, and extending their work to the offline setting is an interesting future research direction.

How to Benefit from Trajectory Offline Data? Our work indicates that in the worst-case, trajectory offline data provides no additional statistical benefit over General or Admissible offline data in the standard offline RL setting with value function approximation and bounded concentrability coefficient. But not all MDPs are the worst-case. Can we expect some instance-dependent benefit from access to trajectory data in offline RL? Alternately, can we make further assumptions on the underlying MDP or the value function classes, that are benign enough to capture real-world scenarios, but allow the learner to better exploit trajectory data. Furthermore, it is also interesting to study whether we can get statistical or computational improvements under trajectory data when the Bellman Completeness property holds.

Acknowledgements

AS thanks Dylan Foster and Wen Sun for useful discussions.

References

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32, 2019.

Philip Amortila, Nan Jiang, and Tengyang Xie. A variant of the wang-foster-kakade lower bound for the discounted setting. *arXiv preprint arXiv:2011.01075*, 2020.

Philip Amortila, Dylan J Foster, Nan Jiang, Ayush Sekhari, and Tengyang Xie. Harnessing density ratios for online reinforcement learning. *arXiv preprint arXiv:2401.09681*, 2024.

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

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

Jinglin Chen and Nan Jiang. Offline reinforcement learning under value and density-ratio realizability: the power of gaps. In *Uncertainty in Artificial Intelligence*, pages 378–388. PMLR, 2022.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. In *Conference on Learning Theory*, pages 3489–3489. PMLR, 2022.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

Audrey Huang, Jinglin Chen, and Nan Jiang. Reinforcement learning in low-rank mdps with density features. *arXiv preprint arXiv:2302.02252*, 2023.

Jiawei Huang and Nan Jiang. On the convergence rate of off-policy policy optimization methods with density-ratio correction. In *International Conference on Artificial Intelligence and Statistics*, pages 2658–2705. PMLR, 2022.

Zeyu Jia, Gene Li, Alexander Rakhlin, Ayush Sekhari, and Nati Srebro. When is agnostic reinforcement learning statistically tractable? *Advances in Neural Information Processing Systems*, 36, 2024.

Nan Jiang. Batch value-function tournament. YouTube, 2021. Available at https://www.youtube.com/watch?v=IhOfTCY-oMg.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Michael Kearns, Yishay Mansour, and Andrew Ng. Approximate planning in large pomdps via reusable trajectories. *Advances in Neural Information Processing Systems*, 12, 1999.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Pac reinforcement learning with rich observations. *Advances in Neural Information Processing Systems*, 29, 2016.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

Zakaria Mhammedi, Dylan J Foster, and Alexander Rakhlin. Representation learning with multi-step inverse kinematics: An efficient and optimal approach to rich-observation rl. *arXiv preprint arXiv:2304.05889*, 2023.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pages 6961–6971. PMLR, 2020.

Rémi Munos. Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567. Citeseer, 2003.

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

Asuman E Ozdaglar, Sarath Pattathil, Jiawei Zhang, and Kaiqing Zhang. Revisiting the linear-programming framework for offline rl with general function approximation. In *International Conference on Machine Learning*, pages 26769–26791. PMLR, 2023.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

Paria Rashidinejad, Hanlin Zhu, Kunhe Yang, Stuart Russell, and Jiantao Jiao. Optimal conservative offline rl with general function approximation via augmented lagrangian. *arXiv preprint arXiv:2211.00716*, 2022.

Ayush Sekhari, Christoph Dann, Mehryar Mohri, Yishay Mansour, and Karthik Sridharan. Agnostic reinforcement learning with low-rank mdps and rich observations. *Advances in Neural Information Processing Systems*, 34:19033–19045, 2021.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

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

Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. In *International Conference on Machine Learning*, pages 12287–12297. PMLR, 2021.

Andrea Zanette. When is realizability sufficient for off-policy reinforcement learning? In *International Conference on Machine Learning*, pages 40637–40668. PMLR, 2023.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

Xuezhou Zhang, Yuda Song, Masatoshi Uehara, Mengdi Wang, Alekh Agarwal, and Wen Sun. Efficient reinforcement learning in block mdps: A model-free representation learning approach. In *International Conference on Machine Learning*, pages 26517–26547. PMLR, 2022.

Contents of Appendix

A Related Works

Offline RL is challenging due to lack of direct interaction with the environment. Existing theoretical works that provide polynomial sample complexity guarantees often rely on multiple assumptions to be satisfied simultaneously. Specifically, in the realm of value function approximation, three pivotal assumptions stand out: (value function) realizability, concentrability, and Bellman completeness (i.e. $\mathcal{T}_h f_{h+1} \subset \mathcal{F}_h$ for all $f_{h+1} \in \mathcal{F}_{h+1}$). The first two assumptions can be further categorized into single-policy concentrability (i.e., only the target policy has bounded concentrability) and all-policy concentrability (all policies in the MDP have bounded concentrability).

Bellman Completeness. If Bellman completeness holds, either all-policy realizability with single-policy concentrability (Xie et al., 2021) or single-policy realizability with all-policy concentrability (Chen and Jiang, 2019) can guarantee polynomial sample complexity for policy optimization. Furthermore, other classical algorithms like Fitted Q-Iteration (FQI) (Munos, 2003; Munos and Szepesvári, 2008; Antos et al., 2008) are proved to have finite sample guarantee in terms of concentrability. The Bellman completeness assumption, however, is deemed rather undesirable because it is non-monotone in the function class and thus may be severely violated when a rich function class is used. Several efforts have been made to remove this assumption, though all requiring new assumptions: Xie and Jiang (2021) showed that if a stronger version of concentrability, called *pushforward* concentrability, holds, then with only single-policy realizability, polynomial sample complexity can be achieved without Bellman completeness. Xie and Jiang (2020), Zhan et al. (2022), and Ozdaglar et al. (2023) introduced the notion of *density-ratio realizability* (different from value function realizability), and showed that this along with single-policy realizability and single-policy concentrability ensures polynomial sample complexity. Zanette (2023) relaxed Bellman completeness to the notion of β-*incompleteness* where Bellman completeness corresponds to $\beta = 0$. He proved that $\beta < 1$ along with realizability and concentrability admits polynomial sample complexity for policy evaluation.

The question of whether just realizability and concentrability alone are sufficient for sample efficient offline RL remained open until the work of Foster et al. (2022), who answered this in the negative. They gave two examples where polynomial samples is insufficient even with all-policy realizability and all-policy concentrability. However, their lower bounds heavily rely on the offline data distribution being non-admissible, leaving the admissible and the trajectory cases open (see definitions and comparison in Section 2.2).

Further works on offline RL include Liu et al. (2018); Uehara et al. (2020, 2021) that focus on policy evaluation, and Zhan et al. (2022); Huang and Jiang (2022); Chen and Jiang (2022); Rashidinejad et al. (2022); Ozdaglar et al. (2023) that focus on policy optimization.

Other Lower Bounds in Offline RL. There is another line of works showing exponential lower bound / impossibility results for offline policy evaluation with linear function approximation, but with concentrability replaced by other weaker notions of coverage (Wang et al., 2020; Amortila et al., 2020; Zanette, 2021), e.g. the linear coverability assumption that $\lambda_{\min}(\mathbb{E}_{(s,a)\sim\mu}\phi(s,a)\phi(s,a)^T) \geq 1/d$. However, their alternate assumptions do not imply concentrability; Furthermore, these prior works also do not consider trajectory data, as in our results. More positive results can be found in the literature of *model-based* approaches, for which we refer the reader to Uehara and Sun (2021) and the related works therin.

Online RL. In online RL, while value function realizability and Bellman completeness is still a common assumption, the bounded concentrability coefficient assumption can be replaced by some low rank structure on the Bellman error or its estimator (Jiang et al., 2017; Zanette et al., 2020; Du et al., 2021; Jin et al., 2021), which allow for efficient exploration. Recently, Xie et al. (2022) identified a new structural assumption called

coverability which resembles all-policy concentrability and ensures polynomial sample complexity when combined with Bellman completeness. There have been various works in online RL that attempt to relax the Bellman completeness assumption by instead assuming density ratio realizability Amortila et al. (2024), occupancy realizability Huang et al. (2023). Additionally, Krishnamurthy et al. (2016); Du et al. (2019); Misra et al. (2020); Zhang et al. (2022); Mhammedi et al. (2023), focus on the simpler setting of block MDPs (which is a special case of density ratio realizability). It is an interesting direction to further unify the common notions used in online and offline RL.

B Additional Definitions and Notation

In this section, we provide additional definitions and notations used in the appendix.

Definition B.1 (Pushforward Concentrability and Action Concentrability; Xie and Jiang (2021, Assumption 1))**.** *For a distribution $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$, if we further assume that*

(a) *There exists some $\mathsf{C}_{\mathcal{A}} > 0$ such that $\mu(a \mid x) \geq 1/\mathsf{C}_{\mathcal{A}}$ for any $x \in \mathcal{X}, a \in \mathcal{A}$,*

(b) *There exists some $\mathsf{C}_{\mathcal{X}} > 0$ such that the transition model satisfies $T(x' \mid x, a)/\mu(x') < \mathsf{C}_{\mathcal{X}}$, and the initial distribution ρ satisfies $\rho(x)/\mu(x) < \mathsf{C}_{\mathcal{X}}$ for any $x, x' \in \mathcal{X}, a \in \mathcal{A}$,*

then we say that the MDP*'s pushforward concentrability with respect to μ is $\mathsf{C}_{\mathrm{pf}} = \mathsf{C}_{\mathcal{X}}\mathsf{C}_{\mathcal{A}}$, and the* MDP*'s action concentrability with respect to μ is $\mathsf{C}_{\mathcal{A}}$.*

Aggregated Transitions with Actions. We further define the aggregated transitions with actions:

$$\bar{T}(\phi' \mid \phi, a; \bar{M}) := \frac{\sum_{x \in \phi} \sum_{x' \in \phi'} \mu(x, a) T(x' \mid x, a)}{\sum_{x \in \phi} \mu(x, a)} \tag{5}$$

Notice that when $\pi(x) = \delta_a(\cdot)$, i.e. π takes action a with probability 1 at all states, $\bar{T}(\phi' \mid \phi, a; \bar{M})$ in (5) agrees with $\bar{T}(\phi' \mid \phi, \pi; \bar{M})$ in (3).

Definition B.2 (Block MDP; Du et al. (2019); Misra et al. (2020))**.** *A block* MDP *is defined on top of a latent* MDP *$M = (\mathcal{Z}, \mathcal{A}, T, r, H, \rho)$, a rich observation state space $\mathcal{X}$ (partitioned into disjoint blocks $\mathcal{X}_z$ for each latent state z), a decoder function ξ and a conditional distribution $q(\cdot \mid z) \in \Delta(\mathcal{X}_z)$. The block* MDP *$\check{M} = (\mathcal{X}, \mathcal{A}, \check{T}, \check{r}, H, \check{\rho})$ with $\check{T}(x \mid x, a) = q(x' \mid \xi(x'))T(\xi(x') \mid \xi(x), a)$, $\check{r}(x, a) = r(\xi(x), a)$ and $\check{\rho}(x) = \rho(\xi(x))q(x \mid \xi(x))$.*

Definition B.3 (W-function of OPE problems)**.** *Given an* OPE *problem $(M, \mu, \pi_{\mathrm{e}}, \mathcal{F})$, the W-function: $W^{\pi_{\mathrm{e}}}(\cdot; \mu, M) : [H] \to \mathbb{R}$ is defined as*

$$W^{\pi_{\mathrm{e}}}(h; \mu, M) = \sum_{z \in \mathcal{Z}_h} \mu_h(z, \pi_{\mathrm{e}}(z)) Q_h^{\pi_{\mathrm{e}}}(z, \pi_{\mathrm{e}}(z); M). \tag{6}$$

Whenever clear from the context, the dependence on π_{e}, μ and M will be ignored.

Additional Notation. For $n \in \mathbb{N}$, we write $[n] = \{1, \ldots, n\}$. For a countable set $\mathcal{S}$, we write $\Delta(\mathcal{S})$ for the set of probability distributions on $\mathcal{S}$. For any function $u : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$ and distribution $\rho \in \Delta(\mathcal{X} \times \mathcal{A})$, we define the norms $\|u\|_{1,\rho} = \mathbb{E}_{(x,a)\sim\rho}[|u(x, a)|]$ and $\|u\|_{2,\rho} = \sqrt{\mathbb{E}_{(x,a)\sim\rho}[u^2(x, a)]}$. For a distribution $\mathbb{P} \in \Delta(\mathcal{X})$, we define the cross product of $\mathbb{P}^{\otimes n}$ to be a distribution over $\mathcal{X}^n$ such that $\mathbb{P}^{\otimes n}((x_1, \cdots, x_n)) = \prod_{i=1}^n \mathbb{P}(x_i)$, where $x_i \in \mathcal{X}$.

C Proof of Theorem 3.1

Suppose we are given the Markov Transition Model (MTM) $M = (\mathcal{Z}, \mathcal{A}, T, H, \rho)$, and a distribution μ over $\mathcal{Z} \times \mathcal{A}$. Φ is an aggregated scheme so that every $z \in \mathcal{Z}$ belongs to exact one of $\phi \in \Phi$, written as $z \in \phi$ (also all the latent states in ϕ should be at the same layer). We further define the aggregated function $\zeta : \mathcal{Z} \to \Phi$, where for any $z \in \phi$,

$$\zeta(z) := \phi \tag{7}$$

In the proof we will construct two class of offline policy evaluation (OPE) problems $\mathfrak{G}^{(1)}$ and $\mathfrak{G}^{(2)}$ from the given MDP M and distribution μ. And we will prove Theorem 3.1 by showing that there exists an OPE problem in $\mathfrak{G} = \mathfrak{G}^{(1)} \cup \mathfrak{G}^{(2)}$ that requires $\Omega(\bar{\mathsf{C}}/\varepsilon)$ number of samples for each layers to achieve accuracy $1/2$. The constructive proof is divided into three parts:

(i) Construct aggregated MDPs $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$ according to M and Φ such that the concentrability coefficients of $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$ are of order $\bar{\mathsf{C}}$ (Section C.2).

(ii) Construct two OPE problems $\mathfrak{g}^{(1)} = (M^{(1)}, \mu', \pi_e, \mathcal{F})$ and $\mathfrak{g}^{(2)} = (M^{(2)}, \mu', \pi_e, \mathcal{F})$ (Section C.3), where MDP $M^{(1)}$ and $M^{(2)}$ are obtained by adding three states u_h, v_h and w_h in each layers. Distribution μ' is obtained from μ after rearranging some probability to u_h, v_h and v_h. And we can show that the concentrability coefficients of $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$ can translate to the ratio between difference of value functions and difference of rewards between $M^{(1)}$ and $M^{(2)}$.

(iii) Construct two class of OPE problems $\mathfrak{G}^{(1)}$ and $\mathfrak{G}^{(2)}$ by lifting OPE problems $\mathfrak{g}^{(1)}$ and $\mathfrak{g}^{(2)}$ into rich observations (Section C.4).

C.1 Construction Sketch

In this subsection, we give a high-level sketch for the proof of Theorem 3.1. The full proof is detailed in the follow-up subsections.

Suppose we are given an arbitrary Markov Transition Model M with state space $\mathcal{Z}$, transition dynamics T and initial distribution ρ, any offline data distribution μ, and any state aggregation scheme Φ (see Figure 4(a)). Let $\mathcal{I}$ denote the set of aggregations that attains the maximum for $\bar{\mathsf{C}}_\varepsilon(M, \Phi, \mu)$ given in Definition 3.1. In Figure 4(a), $\mathcal{I}$ is represented with the bold rectangle (for simplicity, in Figure 4, $\mathcal{I}$ only includes a single aggregation that contains a single latent state $z^\star$, but in general $\mathcal{I}$ may include multiple aggregations each with multiple latent states). Based on M, we will construct two MDPs $M^{(1)}$ (with transitions $T^{(1)}$ and reward $r^{(1)}$) and $M^{(2)}$ (with transitions $T^{(2)}$ and reward $r^{(2)}$), and will argue that it is difficult for the learner to tell them apart when the MDPs are lifted to *block MDPs*.

1. *Modified* MDP *structure* M' (Section C.3): We construct an MDP structure M' with state space $\mathcal{Z}'$ that comprises of the state space $\mathcal{Z}$ (corresponding to M) along with three additional states u_h, v_h, w_h on each layer $h \in [H]$ (see Figure 4(b)). The transition dynamics T' in the M' is defined such that
   (a) Each of u_h, v_h, and w_h deterministically transitions to w_{h+1} under any action.
   (b) For any $z_1, z_2 \in \mathcal{Z}$ and $a \in \mathcal{A}$, $T'(z_2 \mid z_1, a) = (1 - 2/H)T(z_2 \mid z_1, a)$. In particular, the probability of each transition from $\mathcal{Z}_h$ to $\mathcal{Z}_{h+1}$ is decreased by a factor of $(1 - 2/H)$.
   (c) The remaining $2/H$ probability mass in T' is assigned to transitions from $\mathcal{Z}_h$ to u_{h+1} and v_{h+1}. These transitions are different for $M^{(1)}$ and $M^{(2)}$, and will be specified later.

Finally, we also define a new a modified aggregation scheme Φ' that comprises of Φ along with $3H$ more singleton aggregations, each consisting of u_h, v_h, w_h for $h \in [H]$.

2. *Reward functions* (Section C.2 and Section C.3): We create reward functions $r^{(1)}$ and $r^{(2)}$, for MDPs $M^{(1)}$ and $M^{(2)}$ respectively, such that non-zero rewards are only given to states in aggregation $\mathcal{I}$ and to $(u_h, v_h, w_h)_{h\in[H]}$ (see Figure 4(c)). In particular, we set

   - $r^{(1)}(z, \pi_e(z)) = \alpha$ and $r^{(2)}(z, \pi_e(z)) = -\alpha$ for any state $z \in \mathcal{I}$, for some properly chosen constant α.
   - $r^{(i)}(u_h, a) = 1$, $r^{(i)}(v_h, a) = -1$ and $r^{(i)}(w_h, a) = 0$ for any $h \in [H]$ and $a \in \mathcal{A}$.
   - $r^{(i)}(z, a) = 0$ for all other $z \in \mathcal{Z}$ and $a \in \mathcal{A}$.

3. *Value functions and missing transitions* (Section C.2 and Section C.3): We now proceed to the construction of state-action value functions $Q^{(1)}$ and $Q^{(2)}$ for the evaluation policy π_e, and the transition probabilities $T^{(1)}$ and $T^{(2)}$, for $M^{(1)}$ and $M^{(2)}$ respectively. These quantities are constructed so as to ensure that:

   (a) All states that belong to the same aggregation have the same value in both $M^{(1)}$ and $M^{(2)}$, and are thus indistinguishable via the value functions, i.e. for any aggregation $\phi \in \Phi'$, states $z_1, z_2 \in \phi$, and $a \in \mathcal{A}$,

   $$Q^{(1)}(z_1, a) = Q^{(1)}(z_2, a) \quad \text{and} \quad Q^{(2)}(z_1, a) = Q^{(2)}(z_2, a). \tag{8}$$

   (b) From any aggregation, the probability of transitioning to states u_h (or to states v_h) is same between $M^{(1)}$ and $M^{(2)}$, i.e. for any $\phi \in \Phi$ and $h \in [H]$,

   $$\begin{aligned} \forall a \in \mathcal{A}, \quad & \sum_{z\in\phi} \mu_h(z, a) T^{(1)}(u_{h+1} \mid z, a) = \sum_{z\in\phi} \mu_h(z, a) T^{(2)}(u_{h+1} \mid z, a), \\ \forall a \in \mathcal{A}, \quad & \sum_{z\in\phi} \mu_h(z, a) T^{(1)}(v_{h+1} \mid z, a) = \sum_{z\in\phi} \mu_h(z, a) T^{(2)}(v_{h+1} \mid z, a). \end{aligned} \tag{9}$$

   (c) For any $z_1, z_2 \in \mathcal{Z}$, we have $T^{(i)}(z_2 \mid z_1, a) = T'(z_2 \mid z_1, a)$ for all $a \in \mathcal{A}$.

Value functions and transitions that satisfy the above constraints are inductively constructed from time step $h = H$ to 1. Since each of the above constraints is a linear equation, the corresponding solutions can be obtained by solving a system of linear equations. At a high level, the reason why we added u_{h+1} and v_{h+1} — by splitting out some transition probabilities to u_{h+1} and v_{h+1}, and adjust their differences properly (notice that $Q^{(i)}(u_{h+1}, a) = 1$ and $Q^{(i)}(v_{h+1}, a) = -1$), we can calibrate the state-action values in ϕ, making them all equal.

Jointly solving (8), (9), and using the condition $T^{(i)}(u_{h+1} \mid z, a) + T^{(i)}(v_{h+1} \mid z, a) = \frac{2}{H}$, we can obtain the following solution:

$$Q^{(i)}(\phi, a) = \sum_{\phi' \in \Phi_{h+1}} \bar{T}(\phi' \mid \phi, a) V^{(i)}(\phi'), \tag{10}$$

where $Q^{(i)}(\phi, a)$ is the value of $Q^{(i)}(z, a)$ shared by all $z \in \phi$, Φ_{h+1} is the set of aggregations on layer $h+1$, and $\bar{T}(\phi' \mid \phi, a) = \frac{\sum_{z'\in\phi'} \sum_{z\in\phi} \mu_h(z,a) T'(z'|z,a)}{\sum_{z\in\phi} \mu_h(z,a)}$ is the aggregated transition. This is where the aggregated transition comes into the picture. With (10), the argument that the aggregated transition plays a role in the sample complexity is similar to the argument in the tabular case as outlined in Section 3. For formal proofs, see Lemma 4(a)-Lemma 4(c)

4. *Construction of offline distribution* μ' (Section C.3): For any $z \neq u_h, v_h, w_h$, we set $\mu'_h(z,a) = \frac{1}{2}\mu_h(z,a)$. Furthermore, for $z = u_h, v_h, w_h$, we define $\mu'_h(z,a) = \frac{1}{6}$ This construction ensures that both C and $\bar{\mathsf{C}}$ remain unchanged up to constant factors in the original M and in the modified $M^{(1)}$ and $M^{(2)}$. See Lemma 4(d) and Corollary 1 for formal proofs.

5. *Lifting to block MDPs* (Section C.4): We finally lift $M^{(1)}$ and $M^{(2)}$ to block MDPs where every state z serves as a *latent state* invisible to the learner. Instead of observing the latent state z, the learner only observes a *rich observation* from the set $\mathcal{X}$ corresponding to latent state z.

For the rest of this section, we provide a formal proof for Theorem 3.1.

C.2 Construction of Aggregated MDPs

We first construct two aggregated MDPs $\bar{M}^{(1)} = (\Phi, \mathcal{A}, \bar{T}, \bar{r}^{(1)}, H, \bar{\rho})$ and $\bar{M}^{(2)} = (\Phi, \mathcal{A}, \bar{T}, \bar{r}^{(2)}, H, \bar{\rho})$ of horizon H, whose state space is Φ and action space is $\mathcal{A}$. Furthermore, both of them have identical transition models and initial distributions given by:

- State Space Φ, and action space $\mathcal{A}$.
- **Transition model** $\bar{T}$ is defined as $\bar{T}(\phi \mid \phi, a) := \bar{T}(\phi \mid \phi, a; \bar{M})$, where $\bar{T}(\phi \mid \phi, a; \bar{M})$ is given in (3).
- **Initial distribution** $\bar{\rho}$ is defined as $\bar{\rho}(\phi) := \sum_{z \in \phi} \rho(z)$.

Suppose $h^* \in [H-1]$ and set $\mathcal{I} \subset \Phi_{h^*}$ attains the maximum in Definition 3.1. The reward function of $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$ are given by:

- **Reward function** $\bar{r}^{(1)}$ **for** $\bar{M}^{(1)}$**.** We set the reward to be 0 for all states $\phi \notin \mathcal{I}$. Furthermore, for $\phi \in \mathcal{I}$, we set the reward to be 0 for actions that would not have been chosen by π_{e}. On the remaining (ϕ, a) tuples, we set a non-zero reward given by:

$$\bar{r}^{(1)}(\phi, a) = \frac{\varepsilon}{2H \sum_{\phi \in \mathcal{I}} d^{\bar{\pi}_{\mathrm{e}}}_{h^*}(\phi; \bar{M})} \cdot \mathbb{I}\{\phi \in \mathcal{I}, a = \pi_{\mathrm{e}}(\phi)\}.$$

  The key intuition in the above choice of reward function is to ensure that only those states-action contribute to non-zero rewards for which $\phi \in \mathcal{I}$ and $a = \pi_{\mathrm{e}}(\phi)$; Hence, in order to receive a non-zero return, any agent in this MDP needs to first find states in $\mathcal{I}$ and then play action given by π_{e} on them. The denominator just consists of additional normalizing factors to ensure that the value is bounded by $\varepsilon/2H$.

- **Reward function** $\bar{r}^{(2)}$ **for** $\bar{M}^{(2)}$**.** The reward function is similar to $\bar{r}^{(1)}$, but with the negative sign. In particular, we define

$$\bar{r}^{(2)}(\phi, a) = -\frac{\varepsilon}{2H \sum_{\phi \in \mathcal{I}} d^{\bar{\pi}_{\mathrm{e}}}_{h^*}(\phi; \bar{M})} \cdot \mathbb{I}\{\phi \in \mathcal{I}, a = \pi_{\mathrm{e}}(\phi)\}.$$

Definition C.1 (Aggregated MDP)**.** *We define aggregated MDPs* $\bar{M}^{(1)} = (\Phi, \mathcal{A}, \bar{T}, \bar{r}^{(1)}, H, \bar{\rho})$ *and* $\bar{M}^{(2)} = (\Phi, \mathcal{A}, \bar{T}, \bar{r}^{(2)}, H, \bar{\rho})$ *as follows: transition model* $\bar{T}$ *is defined as* $\bar{T}(\phi \mid \phi, a) := \bar{T}(\phi \mid \phi, a; \bar{M})$*, where* $\bar{T}(\phi \mid \phi, a; \bar{M})$ *is given in* (3)*, reward functions are defined as*

$$\bar{r}^{(1)}(\phi, a) = \frac{\varepsilon}{2H \sum_{\phi \in \mathcal{I}} d^{\bar{\pi}_{\mathrm{e}}}_{h^*}(\phi; \bar{M})} \cdot \mathbb{I}\{\phi \in \mathcal{I}, a = \pi_{\mathrm{e}}(\phi)\},$$

and,

$$\bar{r}^{(2)}(\phi, a) = -\frac{\varepsilon}{2H \sum_{\phi \in \mathcal{I}} d^{\bar{\pi}_{\mathrm{e}}}_{h^*}(\phi; \bar{M})} \cdot \mathbb{I}\{\phi \in \mathcal{I}, a = \pi_{\mathrm{e}}(\phi)\}.$$

Lemma 1. *The value functions of $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$ satisfies that for any policy π and $\phi \in \Phi$,*

$$V^{\pi}(\phi; \bar{M}^{(1)}) = -V^{\pi}(\phi; \bar{M}^{(@)}).$$

Proof of Lemma 1. This lemma is easy to see after noticing that the transitions of $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$ are the same, while the reward functions of $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$ are of opposite signs. ■

Lemma 2. *The value functions of $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$ satisfies that*

$$V^{\bar{\pi}_e}(\bar{\rho}; \bar{M}^{(1)}) = \frac{\varepsilon}{2H}, \quad \text{and } V^{\bar{\pi}_e}(\bar{\rho}; \bar{M}^{(2)}) = -\frac{\varepsilon}{2H}.$$

Additionally, for each $\phi \in \Phi$,

$$0 \le V^{\bar{\pi}_e}(\phi; \bar{M}^{(1)}) \le \frac{1}{2H}, \qquad -\frac{1}{2H} \le V^{\bar{\pi}_e}(\phi; \bar{M}^{(2)}) \le 0. \tag{11}$$

Proof of Lemma 2. According to Lemma 1, we only need to prove results for $\bar{M}^{(1)}$. First of all, we can write the value functions as weighted averages of rewards with occupancy-measure-weights:

$$V^{\bar{\pi}_e}(\bar{\rho}; \bar{M}^{(1)}) = \sum_{\phi \in \mathcal{I}} r^{(1)}(\phi, \pi_e(\phi)) d_{h^*}^{\bar{\pi}_e}(\phi; \bar{M}) = \frac{\varepsilon}{2H}. \tag{12}$$

Next, We let $\mathcal{I}$ to be the set which attains the second maximum in (2) of the definition of $\mathsf{C}_{\varepsilon}(M, \mu)$. Then $\mathcal{I} \subset \Phi_h$ for some $h \in [H-1]$, and it also satisfies $\sum_{\phi \in \mathcal{I}} d_{h^*}^{\bar{\pi}_e}(\phi; \bar{M}) \ge \varepsilon$, which implies that

$$0 \le \bar{r}^{(1)}(\phi, a) \le \frac{\varepsilon}{2H \cdot \varepsilon} \le \frac{1}{2H}, \quad \forall \phi \in \Phi_h, a \in \mathcal{A} \qquad \text{and} \qquad \bar{r}^{(1)}(\phi, a) = 0, \quad \forall \phi \notin \Phi_h, a \in \mathcal{A}.$$

Hence for any $\phi \in \Phi$, the sum of rewards along any trajectory which starts from ϕ is always between 0 and $1/2H$ in $\bar{M}^{(1)}$. Therefore we get $0 \le V^{\bar{\pi}_e}(\phi; \bar{M}^{(1)}) \le 1/2H$ for any $\phi \in \Phi$. ■

C.3 Construction of Latent-State MDPs and OPE Problems

Based on $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$, we next construct two MDPs $M^{(1)}$ and $M^{(2)}$ which will be used as latent-state dynamics for rich observation MDPs that we construct in the next section. For $i \in \{1, 2\}$, we define

$$M^{(i)} = \mathrm{MDP}(\mathcal{Z}', \mathcal{A}, T^{(i)}, r^{(i)}, H, \rho),$$

where

- **State space $\mathcal{Z}'$** is defined such that $\mathcal{Z}' = \cup_{h=1}^{H} \mathcal{Z}'_h$ where, for each $h \in [H]$, in addition to the states in $\mathcal{Z}_h$, the set $\mathcal{Z}'_h$ contains three additional states $\{u_h, v_h, w_h\}$ for all $h \in [H]$. Formally, $\mathcal{Z}'_h := \mathcal{Z}_h \cup \{u_h, v_h, w_h\}$.

  The roles of u_h, v_h and w_h is to ensure that for every aggregated state $\phi \in \Phi$, each state $z \in \phi$ has the same value functione; How we achieve this will become clear later when we define the transition model $T^{(i)}$.

- **Initial distribution ρ** is the same as the initial distribution in M (the original MDP that was used in the construction of $\bar{M}^{(1)}$ and $\bar{M}^{(2)}$).

- **Reward function** $r^{(i)}$ is set as

$$r^{(i)}(z,a) = \begin{cases} \bar{r}^{(i)}(\zeta(z),a) & \text{if } z \in \mathcal{Z}_h \\ 1 & \text{if } z = u_h \\ -1 & \text{if } z = v_h \\ 0 & \text{if } z = w_h \end{cases}$$

for all $h \in [H-1]$, where $\zeta(z)$ is defined in (7). In particular, we use the same reward in $M^{(i)}$ as in $\bar{M}^{(i)}$ for the (old) states $z \in \mathcal{Z}$, and define new rewards for (newly added) states u_h, v_h and w_h. By definition, the reward $r^{(i)}(z',a) = r^{(i)}(z'',a)$ whenever z' and z'' belong the same aggregated state ϕ, for all $a \in \mathcal{A}$.

- **Transition model $T^{(i)}$.** The transitions are defined such that $T^{(i)}(z' \mid z,a)$ is proportional to $T(z' \mid z,a)$ for tuples $(z,a,z') \in \mathcal{Z}_h \times \mathcal{A} \times \mathcal{Z}_h$ that corresponds to transitions amongst (old) states that were also present in M, and the remaining probability mass is redirected it to new states $\{u_h, v_h, w_h\}_{h\in[H]}$. Formally, for $z \in \mathcal{Z}_h$ and action $a \in \mathcal{A}$, we set

$$T^{(i)}(z' \mid z,a) = \begin{cases} \left(1 - \frac{2}{H}\right) T(z' \mid z,a) & \text{if} \quad z' \in \mathcal{Z}_{h+1} \\ [\Delta T^{(i)}](z) + \frac{1}{H} & \text{if} \quad z' = u_{h+1} \\ -[\Delta T^{(i)}](z) + \frac{1}{H} & \text{if} \quad z' = v_{h+1} \\ 0 & \text{if} \quad z' = w_{h+1} \end{cases}, \tag{13}$$

where we defined

$$[\Delta T^{(i)}](z) := \frac{1}{2}\left(\sum_{\phi' \in \Phi_{h+1}} \bar{T}(\phi' \mid \zeta(z),a) V^{\bar{\pi}_e}(\phi'; \bar{M}^{(i)}) - \left(1 - \frac{2}{H}\right) \sum_{z' \in \mathcal{Z}_{h+1}} T(z' \mid z,a) V^{\bar{\pi}_e}(\zeta(z'); \bar{M}^{(i)}) \right). \tag{14}$$

Furthermore, for any $z \in \{u_h, v_w, w_h\}$ and $a \in \mathcal{A}$, $T^{(i)}$ transits to w_{h+1} with probability 1, i.e.

$$T^{(i)}(z' \mid z,a) = \begin{cases} 1 & \text{if} \quad z' = w_{h+1} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Intuitively, the above transitions imply that states $\{w_h\}_{h\in[H-1]}$ act as terminal states.

We further define distribution μ'_h over $\mathcal{Z}'_h$ of layer $h \in [H-1]$ as follows:

$$\mu'_h(z) = \begin{cases} \frac{\mu_h(z)}{2} & z \in \mathcal{Z}_h, \\ \frac{1}{6} & z \in \{u_h, v_h, w_h\}. \end{cases} \tag{16}$$

Additionally, to be formal, we define $\pi_b(z) = \pi_e(z) \equiv a_0$ (an arbitrary fixed state in $\mathcal{A}$) for any $z \in \{u_h, v_h, w_h\}$. And we define OPE problems $\mathfrak{g}^{(1)}$ and $\mathfrak{g}^{(2)}$ as

$$\mathfrak{g}^{(1)} = (M^{(1)}, \mu', \pi_e, \mathcal{F}), \quad \mathfrak{g}^{(2)} = (M^{(2)}, \mu', \pi_e, \mathcal{F}), \tag{17}$$

where $\mathcal{F} := \{Q^{(1)}, Q^{(2)}\}$ with $Q^{(1)}$ and $Q^{(2)}$ are Q-functions of $M^{(1)}$ and $M^{(2)}$ under policy π_e. Before we proceed, we note the following technical lemma.

Lemma 3. *For $i \in \{1,2\}$, $T^{(i)}$ constructed via (13) and (15) above is a valid transition model.*

Proof of Lemma 3. We first show that $T^{(i)}(z' \mid z,a) > 0$ for all $a \in \mathcal{A}$ and $z, z' \in \mathcal{Z}'_{h+1}$. This is trivial for $z' \in \mathcal{Z}_h$ or when $z' = w_{h+1}$. We next show that the same holds when $z' \in \{u_{h+1}, v_{h+1}\}$.

Note that due to Lemma 2, we have that $|V^{\bar{\pi}_e}(\phi; \bar{M}^{(i)})| \le \frac{1}{2H}$ for all $\phi \in \Phi$. Plugging this in (14), and using Triangle inequality, we get that

$$\left|[\Delta T^{(i)}](z)\right| \le \sup_{\phi \in \Phi} \left|V^{\bar{\pi}_e}(\phi; \bar{M}^{(i)})\right| \le \frac{1}{2H}.$$

Using this in (13), we immediately get that $T^{(i)}(z' \mid z,a) > 0$ for any $\varepsilon \le 1$. Furthermore, it is easy to check that for any $z \in \mathcal{Z}'_h$ and $a \in \mathcal{A}$, $\sum_{z' \in \mathcal{Z}'_{h+1}} T^{(i)}(z' \mid z,a) = 1$. Thus, $T^{(i)}$ is a valid transition model. ∎

We note the following technical lemma, which will be used in the rest of the analysis.

Lemma 4. *We have the following properties of value functions, Q-functions and occupancy measures of $M^{(1)}$ and $M^{(2)}$:*

(a) *For $i \in \{1,2\}$, and for any $\phi \in \Phi$, $z \in \phi$, and $a \in \mathcal{A}$,*

$$Q^{\pi_e}(z,a; M^{(i)}) = Q^{\bar{\pi}_e}(\zeta(z), a; \bar{M}^{(i)}).$$

(b) *Corresponding to the initial distribution ρ, the expected values satisfy*

$$V^{\pi_e}(\rho; M^{(1)}) - V^{\pi_e}(\rho; M^{(2)}) = \frac{\varepsilon}{H}.$$

(c) *For any $h \in [H-1]$, latent state $\phi \in \Phi_h$, action $a \in \mathcal{A}$ and latent state $z' \in \mathcal{Z}'$,*

$$\sum_{z \in \phi} \mu(z \mid z \in \phi, a) T^{(1)}(z' \mid z,a) = \sum_{z \in \phi} \mu(z \mid z \in \phi, a) T^{(2)}(z' \mid z,a).$$

(d) *For each $i \in \{1,2\}$, for any $h \in [H-1]$ and $z \in \mathcal{Z}_h$, and policy π,*

$$\frac{1}{16} d_h^{\pi}(z; M) \le d_h^{\pi}(z; M^{(i)}) \le d_h^{\pi}(z; M).$$

Proof of Lemma 4. Observe that, by construction, we have for $i \in \{1,2\}$,

$$Q^{(i)}(u_h, a) = 1, \qquad Q^{(i)}(v_h, a) = -1, \text{and} \qquad Q^{(i)}(w_h, a) = 0,$$

for any h and $a \in \mathcal{A}$.

Proof of (a): The proof follows via downward induction from $h = H$ to $h = 1$. First note that for any $z \in \mathcal{Z}_H$ and $a \in \mathcal{A}$, we have $Q^{\pi_e}(z,a; M^{(i)}) = 0 = Q^{\bar{\pi}_e}(\zeta(z), a; \bar{M}^{(i)})$; Thus the base case is satisfied. For the induction hypotesis, assume that the desired claim holds for $h+1$. Thus, for layer h, using Bellman equation for the policy π_e, we have that

$$\begin{aligned}
&Q^{\pi}(z,a; M^{(i)}) \\
&= r^{(i)}(z,a) + T^{(i)}(u_{h+1} \mid z,a) - T^{(i)}(v_{h+1} \mid z,a) + \sum_{z' \in \mathcal{Z}_{h+1}} T^{(i)}(z' \mid z,a) Q^{\pi_e}(z', \pi_e(z'); M^{(i)}) \\
&= \bar{r}^{(i)}(\zeta(z), a) + T^{(i)}(u_{h+1} \mid z,a) - T^{(i)}(v_{h+1} \mid z,a) + \sum_{z' \in \mathcal{Z}_{h+1}} T^{(i)}(z' \mid z,a) V^{\bar{\pi}_e}(\zeta(z'); \bar{M}^{(i)}) \\
&= \bar{r}^{(i)}(\zeta(z), a) + \sum_{\phi' \in \Phi_{h+1}} \bar{T}(\phi' \mid \zeta(z), a) V^{\bar{\pi}_e}(\phi'; \bar{M}^{(i)}) = Q^{\bar{\pi}_e}(\zeta(z), a; \bar{M}^{(i)}),
\end{aligned}$$

where in the last line, we used Bellman equation for policy π_e under the MDP $\bar{M}^{(1)}$. The above completes the induction step, thus showing that the claim holds for all $h \in [H]$.

Proof of (b): Using part-(a) above, we note that

$$V^{\pi_e}(z; M^{(1)}) = Q^{\pi_e}(z, \pi_e(z); M^{(1)}) = Q^{\bar{\pi}_e}(\zeta(z), \bar{\pi}_e(\zeta(z)); \bar{M}^{(1)}) = V^{\bar{\pi}_e}(\zeta(z); \bar{M}^{(1)}).$$

Similarly, we also get that $V^{\pi_e}(z; M^{(2)}) = V^{\bar{\pi}_e}(\zeta(z); \bar{M}^{(2)})$. The desired bound follows by noting that (12) implies

$$V^{\bar{\pi}_e}(\bar{\rho}; \bar{M}^{(1)}) - V^{\bar{\pi}_e}(\bar{\rho}; \bar{M}^{(2)}) = \frac{\varepsilon}{H},$$

which implies

$$V^{\pi_e}(\rho; M^{(1)}) - V^{\pi_e}(\rho; M^{(2)}) = \frac{\varepsilon}{H}.$$

Proof of (c): First note that whenever $z' \notin \{u_h, v_h\}_{h=1}^H$, we always have $T^{(1)}(z' \mid z, a) = T^{(2)}(z' \mid z, a)$ according to the definition of $T^{(1)}$ and $T^{(2)}$. We next show that the same holds for $z' = u_h$. We only need to verify $\sum_{z \in \phi} \mu(z, a)(T^{(1)}(u_h \mid z, a) - T^{(2)}(u_h \mid z, a)) = 0$.

$$\begin{aligned}
&\sum_{z \in \phi} \mu(z, a)(T^{(1)}(u_h \mid z, a) - T^{(2)}(u_h \mid z, a)) \\
&= \sum_{z \in \phi} \mu(z, a) \left(\sum_{\phi' \in \Phi_{h+1}} \bar{T}(\phi' \mid \zeta(z), a) \bar{V}^{(1)}(\phi') - \sum_{z' \in \mathcal{Z}_{h+1}} T^{(1)}(z' \mid z, a) \bar{V}^{(1)}(\zeta(z')) \right) \\
&= \sum_{z \in \phi} \mu(z, a) \sum_{\phi' \in \Phi_{h+1}} \bar{T}(\phi' \mid \phi, a) \bar{V}^{(1)}(\phi') - \sum_{\phi' \in \Phi_{h+1}} \sum_{z' \in \phi'} \bar{V}^{(1)}(\zeta(z')) \sum_{z \in \phi} \mu(z, a) T^{(1)}(z' \mid z, a) \\
&= \sum_{z \in \phi} \mu(z, a) \sum_{\phi' \in \Phi_{h+1}} \bar{T}(\phi' \mid \phi, a) \bar{V}^{(1)}(\phi') - \sum_{\phi' \in \Phi_{h+1}} \bar{V}^{(1)}(\phi') \sum_{z' \in \phi'} \sum_{z \in \phi} \mu(z, a) T^{(1)}(z' \mid z, a) \\
&= \sum_{z \in \phi} \mu(z, a) \sum_{\phi' \in \Phi_{h+1}} \bar{T}(\phi' \mid \phi, a) \bar{V}^{(1)}(\phi') - \sum_{\phi' \in \Phi_{h+1}} \bar{V}^{(1)}(\phi') \sum_{z \in \phi} \mu(z, a) \bar{T}(\phi' \mid \phi, a) = 0,
\end{aligned}$$

where the first equation uses the definition of $T^{(1)}$ and $T^{(2)}$ in (13) and Lemma 1, and the last equation follows from the definition of $\bar{T}$ in (3). Hence (c) is verified for $z' = u_h$. The proof for $z' = v_h$ follows similarly.

Proof of (d): We only prove the result for $M^{(1)}$; the proof for $M^{(2)}$ follows similarly. In fact, we show a slightly stronger result that for all $h \in [H]$ and $z \in \mathcal{Z}_h$,

$$\left(\frac{H-2}{H} \right)^{h-1} d_h^{\pi}(z; M) \le d_h^{\pi}(z, M^{(1)}) \le d_h^{\pi}(z, M). \tag{18}$$

The proof follows via induction over h. For the base case, note that for $h = 1$, by definition, we have $d_1^{\pi}(z; M) = \rho(z) = d_1^{\pi}(z; M^{(1)})$ for any $z \in \mathcal{Z}_1$, which implies (18).

For the induction step, suppose (18) holds for a certain $h \le H - 1$. For the upper bound, note that any $z \in \mathcal{Z}_{h+1}$ satisfies

$$d_{h+1}^{\pi}(z; M^{(1)}) = \sum_{z' \in \mathcal{Z}_h} d_h^{\pi}(z'; M^{(1)}) T^{(1)}(z \mid z', \pi(z')),$$

which combined with the upper bound in (18) implies

$$d_{h+1}^{\pi}(z; M^{(1)}) \le \sum_{z' \in \mathcal{Z}_h} d_h^{\pi}(z'; M) T(z \mid z', \pi(z')) = d_{h+1}^{\pi}(z, M).$$

For the lower bound, recall from the definition of $T^{(1)}$, which implies that

$$T^{(1)}(z \mid z', \pi(z')) = \frac{H-2}{H} T(z \mid z', \pi(z')).$$

Using the above with the lower bound in (18), we get that

$$\begin{aligned} d_{h+1}^{\pi}(z; M^{(1)}) &= \sum_{z' \in \mathcal{Z}'_h} d_h^{\pi}(z'; M^{(1)}) T^{(1)}(z \mid z', \pi(z')) \\ &\geq \left(\frac{H-2}{H}\right)^{h-1} \sum_{z' \in \mathcal{Z}_h} d_h^{\pi}(z'; M) \cdot \frac{H-2}{H} T(z \mid z', \pi(z')) = \left(\frac{H-2}{H}\right)^{h} d_h^{\pi}(z, M). \end{aligned}$$

The two bounds above imply the (18) also holds for $h+1$. This completes the induction step. ∎

This lemma has two direct corollaries:

Corollary 1. *The concentratbility coefficients $\mathsf{C}(M^{(1)}, \mu')$ of $M^{(1)}$ (or $\mathsf{C}(M^{(2)}, \mu')$ of $M^{(2)}$) satisfies that*

$$\mathsf{C}(M^{(1)}, \mu'), \quad \mathsf{C}(M^{(2)}, \mu') \leq 6\mathsf{C}(M, \mu, \pi_{\mathrm{e}}).$$

Proof of Corollary 1. This corollary directly follows from Definition 16 and Lemma 4 (d). ∎

Corollary 2. *For any two policies π and π' and any $i \in \{1, 2\}$, we have*

$$\sup_{h \in [H]} \sup_{z \in \mathcal{Z}'_h} \frac{d_h^{\pi}(z; M^{(i)})}{d_h^{\pi'}(z; M^{(i)})} \leq 48 \cdot \sup_{h \in [H]} \sup_{z \in \mathcal{Z}_h} \frac{d_h^{\pi}(z; M)}{d_h^{\pi'}(z; M)}$$

Proof of Corollary 2. First of all, for those $z \in \mathcal{Z}_h$ and $h \in [H-1]$, Lemma 4 (d) indicates that

$$\frac{d_h^{\pi}(z; M^{(i)})}{d_h^{\pi'}(z; M^{(i)})} \leq 16 \cdot \frac{d_h^{\pi}(z; M)}{d_h^{\pi'}(z; M)} \leq 48 \cdot \sup_{h \in [H]} \sup_{z \in \mathcal{Z}_h} \frac{d_h^{\pi}(z; M)}{d_h^{\pi'}(z; M)}.$$

Next, we verify cases where $z \in \{u_h, v_h, w_h\}_{h=1}^{H}$. Notice that the transition model of $M^{(i)}$ gives that

$$d_h^{\pi}(w_h; M^{(i)}) = d_{h-1}^{\pi}(u_h; M^{(i)}) + d_{h-1}^{\pi}(v_{h-1}; M^{(i)}),$$

we only need to verify that for any $z \in \{u_h, v_h\}_{h=1}^{H}$, we have

$$\frac{d_h^{\pi}(z; M^{(i)})}{d_h^{\pi'}(z; M^{(i)})} \leq 48 \cdot \sup_{h \in [H]} \sup_{z \in \mathcal{Z}_h} \frac{d_h^{\pi}(z; M)}{d_h^{\pi'}(z; M)}.$$

Without loss of generality, we only verify for $z = u_h$. We write

$$d_h^{\pi}(u_h; M^{(i)}) = \sum_{z \in \mathcal{Z}_{h-1}} T^{(i)}(u_h \mid z, \pi(z)) d_{h-1}^{\pi}(z).$$

According to the transition model of $M^{(i)}$, we have for any $z \in \mathcal{Z}_{h-1}$ and any $a \in \mathcal{A}$,

$$T^{(i)}(u_h \mid z, a) \in \left[\frac{1}{2H}, \frac{3}{2H}\right],$$

which indicates that for any policy π, π', we have

$$\sup_{z \in \mathcal{Z}_{h-1}} \frac{T^{(i)}(u_h \mid z, \pi(z))}{T^{(i)}(u_h \mid z, \pi'(z))} \in [1, 3].$$

Therefore, we have

$$\frac{d_h^{\pi}(u_h; M^{(i)})}{d_h^{\pi'}(u_h; M^{(i)})} \leq 3 \cdot \sup_{z \in \mathcal{Z}_{h-1}} \frac{d_{h-1}^{\pi}(z; M^{(i)})}{d_{h-1}^{\pi'}(z; M^{(i)})} \leq 48 \cdot \sup_{h \in [H]} \sup_{z \in \mathcal{Z}_h} \frac{d_h^{\pi}(z; M)}{d_h^{\pi'}(z; M)}.$$

∎

C.4 Construction of the Class $\mathfrak{G}$ of Offline Policy Evaluation Problems

In this section, we construct the class $\mathfrak{G}$ of OPE problems that are used in Theorem 3.1. The corresponding MDPs in $\mathfrak{G}$ are block MDPs based on $M^{(1)}$ and $M^{(2)}$ (constructed in the previous section), and certain decoder functions. The organization of this section is as follows:

- In Appendix C.4.1, we provide a general procedure to lift an OPE problem $(M, \mu, \pi_e, \mathcal{F})$ over state space $\mathcal{Z}$ into a block OPE problem $(\check{M}, \check{\mu}, \check{\pi}_e, \check{\mathcal{F}})$ with rich-observations in a set $\mathcal{X}$ and latent states in $\mathcal{Z}$, given a decoder function ψ.
- Then, in Appendix C.4.2, we first provide a class Ψ of decoder function and use the above procedure to construct the family $\mathfrak{G}$ of offline RL problems.

C.4.1 Lifting from OPE Problems to Block OPE Problems

In this section, we will discuss how to lift a normal OPE problem $(M, \mu, \pi_e, \mathcal{F})$ ($\mathcal{F}$ satisfies Q-realizability) into a block OPE problem $(\check{M}, \check{\mu}, \check{\pi}_e, \check{\mathcal{F}})$ where $\check{\mathcal{F}}$ satisfies (Q, W)-realizability.

Let $\mathcal{Z} = \bigcup_{h\in[H-1]} \mathcal{Z}_h$ be the state space of MDP M, and we fix $\mathcal{X} = \{\mathcal{X}(z)\}_{z\in\mathcal{Z}}$ to be a family of disjoint sets that denote rich-observations corresponding to latent states $z \in \mathcal{Z}$. Furthermore, let $\mathcal{X}_h = \{\mathcal{X}(z)\}_{z\in\mathcal{Z}_h}$. Then, for $M = (\mathcal{Z}, \mathcal{A}, T, r, H, \rho)$, we can define a block MDP

$$\check{M} = \text{MDP}(\mathcal{X}, \mathcal{A}, \check{T}, \check{R}, H, \check{\rho})$$

with latent state $\mathcal{Z}$ and rich observations in $\mathcal{X}$, where

- **State space**: $\mathcal{X}$ consists of rich-observations corresponding to latent states. We assume that the state space $\mathcal{X} = \bigcup_{h\in[H-1]} \mathcal{X}_h$ is layered.
- **Transition Model** depends only on the latent transition model T. In particular, for any $h \in [H-1]$, $x \in \mathcal{X}(z)$ corresponding to $z \in \mathcal{Z}_h$, and $x' \in \mathcal{X}(z')$ corresponding to $z' \in \mathcal{Z}_{h+1}$, we have

$$\check{T}(x' \mid x, a) := \frac{1}{|\mathcal{X}(z')|} \cdot T(z' \mid z, a).$$

- **Rewards**: For any $z \in \mathcal{Z}$ and $x \in \mathcal{X}$, and $a \in \mathcal{A}$, $\check{R}(x, a)$ is a $\{-1, 1\}$-valued random variable with expected value $r(z, a)$.
- **Initial distribution**: $\check{\rho} \in \Delta(\mathcal{X}_1)$ is defined such that for any $x \in \mathcal{X}_1$,

$$\check{\rho}(x) := \frac{1}{|\mathcal{X}(z)|} \cdot \rho(z) \tag{19}$$

  where z is such that $x \in \mathcal{X}(z)$.

In particular, corresponding to a latent state z, the observations are sampled from Uniform($\mathcal{X}(z)$). In order to construct respective offline RL problems on the MDP $\check{M}$, we also lift the offline data distribution μ to $\check{\mu}$, and offline policy π_e to $\check{\pi}_e$ as follows:

- **Offline distribution** For any $h \in [H-1]$, we define $\check{\mu}_h \in \Delta(\mathcal{X}_h \times \mathcal{A})$ such that for any $x \in \mathcal{X}_h$ and $a \in \mathcal{A}$,

$$\check{\mu}_h(x, a) := \frac{1}{|\mathcal{X}(z)|} \cdot \mu_h(z, a),$$

  where $z \in \mathcal{Z}_h$ is such that $x \in \mathcal{X}(z)$.

- **Evaluation policy** $\check{\pi}_{\mathrm{e}} : \mathcal{X} \mapsto \mathcal{A}$ is defined such that for any $x \in \mathcal{X}$, $\check{\pi}_{\mathrm{e}}(x) := \pi_{\mathrm{e}}(z)$ where z is such that $x \in \mathcal{X}(z)$.
- **Function class** $\check{\mathcal{F}}$ consists of tuples $(\check{f}, \check{W})$, where each $f \in \mathcal{F}$ generates a tuple $(\check{f}, \check{W})$ with $\check{f} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ defined as $\check{f}(x, a) = f(z, a)$ for any $x \in \mathcal{X}(z)$ and $\check{W} : [H-1] \to \mathbb{R}$ is defined as $\check{W}[H-1] = \sum_{x \in \mathcal{X}_h} \check{\mu}_h(x, \check{\pi}_{\mathrm{e}}(x)) \check{f}(x, \check{\pi}_{\mathrm{e}}(a))$.

The following lemma indicates that $\check{f}$ is the Q-function of block MDP $\check{M}$ as long as f is the Q-function of MDP M.

Lemma 5. *For any $h \in [H]$, $z \in \mathcal{Z}_h$, $x \in \mathcal{X}(z)$ and $a \in \mathcal{A}$, we have*

$$Q_h^{\check{\pi}_{\mathrm{e}}}(x, a; \check{M}) = Q_h^{\pi_{\mathrm{e}}}(z, a; M).$$

Proof of Lemma 5. We prove this equation by induction from $h = H$ to $h = 1$. When $h = H$, we have $Q_H^{\check{\pi}_{\mathrm{e}}}(x, a; \check{M}) = Q_H^{\pi_{\mathrm{e}}}(z, a; M)$ for any $z \in \mathcal{Z}_H$, $z \in \mathcal{X}(z)$ and $a \in \mathcal{A}$. Next, suppose $Q_{h+1}^{\check{\pi}_{\mathrm{e}}}(x, a; \check{M}) = Q_{h+1}^{\pi_{\mathrm{e}}}(z, a; M)$ holds for $z \in \mathcal{Z}_{h+1}$ and $x \in \mathcal{X}(z)$. According to the Bellman equation and definitions of $\check{T}, \check{R}$, for any $z \in \mathcal{Z}_h$ and $x \in \mathcal{X}(z)$, we have

$$\begin{aligned}
Q_h^{\check{\pi}_{\mathrm{e}}}(x, a; \check{M}) &= \mathbb{E}[\check{R}(x, a)] + \sum_{x' \in \mathcal{X}_{h+1}} \check{T}(x' \mid x, a) Q_{h+1}^{\check{\pi}_{\mathrm{e}}}(x', \check{\pi}_{\mathrm{e}}(x'); \check{M}) \\
&= r(z, a) + \sum_{z' \in \mathcal{Z}_{h+1}} Q_{h+1}^{\pi_{\mathrm{e}}}(z', \pi_{\mathrm{e}}(z')) \sum_{x' \in \mathcal{X}(z')} \check{T}(x' \mid x, a) \\
&= r(z, a) + \sum_{z' \in \mathcal{Z}_{h+1}} Q_{h+1}^{\pi_{\mathrm{e}}}(z', \pi_{\mathrm{e}}(z'); M) \sum_{z' \in \phi'} T(z' \mid z, a) = Q_h^{\pi_{\mathrm{e}}}(z, a; M),
\end{aligned}$$

where we use the induction hypothesis and the fact that $\check{\pi}_{\mathrm{e}}(x') = \pi_{\mathrm{e}}(z)$ for $x' \in \mathcal{X}(z')$ in the second equation. ■

Lemma 6. *For any $h \in [H-1]$,*

$$\sum_{x \in \mathcal{X}_h} \check{\mu}_h(x, \check{\pi}_{\mathrm{e}}(x)) Q_h^{\check{\pi}_{\mathrm{e}}}(x, \check{\pi}_{\mathrm{e}}(x); \check{M}) = \sum_{z \in \mathcal{Z}_h} \mu_h(z, \pi_{\mathrm{e}}(z)) Q_h^{\pi_{\mathrm{e}}}(z, \pi_{\mathrm{e}}(x); M).$$

Proof of Lemma 6. We first notice that $x \in \mathcal{X}(z)$, we have $\check{\pi}_{\mathrm{e}}(x) = \pi_{\mathrm{e}}(z)$. Hence according to Lemma 5 we only need to verify that, for any $z \in \mathcal{Z}$,

$$\sum_{x \in \mathcal{X}(z)} \check{\mu}_h(x, \check{\pi}_{\mathrm{e}}(x)) = \mu_h(z, \pi_{\mathrm{e}}(z)),$$

which is given by definition of $\check{\mu}_h$: $\check{\mu}_h(x, a) = \frac{\mu_h(z, a)}{|\mathcal{X}(z)|}$. ■

Lemma 7. *For any policy π over* MDP *M, let policy $\check{\pi}$ over $\check{M}$ to be $\check{\pi}(x) = \pi(z)$ for any $x \in \mathcal{X}(z)$. Then we have for any $h \in [H]$, $z \in \mathcal{Z}_h$ and $x \in \mathcal{X}(z)$,*

$$d_h^{\check{\pi}}(x; \check{M}) = \frac{d_h^{\pi}(z; M)}{|\mathcal{X}(z)|}. \tag{20}$$

Proof of Lemma 7. We will prove via induction on the layer of x. For $x \in \mathcal{X}_1$, (20) holds acording to the definition of initial distribution $\check{\rho}$. The induction from layer h to layer $h+1$ can be achieved by

$$d_{h+1}^{\check{\pi}}(x; \check{M}) = \sum_{x' \in \mathcal{X}_h} d_h^{\check{\pi}}(x'; \check{M}) \check{T}(x \mid x', \check{\pi}(x')) = \sum_{z' \in \mathcal{Z}_h} d_h^{\pi}(z'; M) \frac{T(z \mid z', \pi(z'))}{|\mathcal{X}(z)|} = \frac{d_h^{\pi}(z; M)}{|\mathcal{X}(z)|}$$

for any $x \in \mathcal{X}(z)$ and $z \in \mathcal{Z}_{h+1}$. ■

The above lemma has the following two corollaries:

Corollary 3. *The concentrability coefficient* $\mathsf{C}(\check{M}, \mu, \check{\pi}_{\mathrm{e}})$ *is the same as the concentrability coefficient* $\mathsf{C}(M, \mu, \pi_{\mathrm{e}})$.

Proof of Corollary 3. According to Lemma 7, we have for any $h \in [H-1]$, $z \in \mathcal{Z}_h$ and $x \in \mathcal{X}(z)$,

$$\frac{d_h^{\check{\pi}_{\mathrm{e}}}(x; \check{M})}{\check{\mu}(x)} = \frac{d_h^{\pi_{\mathrm{e}}}(z; M) / |\mathcal{X}(z)|}{\mu(z) / |\mathcal{X}(z)|} = \frac{d_h^{\pi_{\mathrm{e}}}(z; M)}{\mu(z)}.$$

Taking the supreme over all $h \in [H-1]$, $z \in \mathcal{Z}_h$ and $x \in \mathcal{X}(z)$, we get

$$\mathsf{C}(\check{M}, \mu, \check{\pi}_{\mathrm{e}}) = \sup_{h \in [H-1]} \sup_{x \in \mathcal{X}_h} \frac{d_h^{\check{\pi}_{\mathrm{e}}}(x; \check{M})}{\check{\mu}(x)} = \sup_{h \in [H-1]} \sup_{z \in \mathcal{Z}_h} \frac{d_h^{\pi_{\mathrm{e}}}(z; M)}{\mu(z)} = \mathsf{C}(M, \mu, \pi_{\mathrm{e}}).$$

■

Corollary 4. *For any two policies* π *and* π'*, let policy* $\check{\pi}$ *and* $\check{\pi}'$ *over* $\check{M}$ *to be* $\check{\pi}(x) = \pi(z)$ *and* $\check{\pi}'(x) = \pi'(z)$ *for any* $x \in \mathcal{X}(z)$*. Then we have*

$$\sup_{h \in [H]} \sup_{x \in \mathcal{X}_h} \frac{d_h^{\check{\pi}}(x; \check{M})}{d_h^{\check{\pi}'}(x; \check{M})} = \sup_{h \in [H]} \sup_{z \in \mathcal{Z}_h} \frac{d_h^{\pi}(z; M)}{d_h^{\pi'}(z; M)}.$$

Proof of Corollary 4. According to Lemma 7, we have for any $h \in [H]$, $z \in \mathcal{Z}_h$ and $x \in \mathcal{X}(z)$,

$$\frac{d_h^{\check{\pi}}(x; \check{M})}{d_h^{\check{\pi}'}(x; \check{M})} = \frac{d_h^{\pi}(z; M) / |\mathcal{X}(z)|}{d_h^{\pi'}(z; M) / |\mathcal{X}(z)|} = \frac{d_h^{\pi}(z; M)}{d_h^{\pi'}(z; M)}.$$

Taking the supreme over all $h \in [H]$, $z \in \mathcal{Z}_h$ and $x \in \mathcal{X}(z)$, we get

$$\sup_{h \in [H]} \sup_{x \in \mathcal{X}_h} \frac{d_h^{\check{\pi}}(x; \check{M})}{d_h^{\check{\pi}'}(x; \check{M})} = \sup_{h \in [H]} \sup_{z \in \mathcal{Z}_h} \frac{d_h^{\pi}(z; M)}{d_h^{\pi'}(z; M)}.$$

■

C.4.2 Construction of the family of offline RL problems

We will construct two OPE families $\mathfrak{G}^{(1)}$ and $\mathfrak{G}^{(2)}$ by lifting OPE problems $\mathfrak{g}^{(1)} = (M^{(1)}, \mu', \pi_{\mathrm{e}}, \mathcal{F})$ and $\mathfrak{g}^{(2)} = (M^{(2)}, \mu', \pi_{\mathrm{e}}, \mathcal{F})$ defined in (17) into Block OPE problems. Each Block OPE problem has the same observation space but a different emission distributions. Furthermore, each of these Block OPE problem has latent state space $\mathcal{Z}$ and is based on the same aggregation scheme Φ.

let $\{\mathcal{X}(\phi)\}_{\phi \in \mathcal{Z}}$ be a family of disjoint sets that denote rich-observations corresponding to aggregated states states $\phi \in \Phi$ such that

$$|\mathcal{X}(\phi)| \gtrsim \frac{|\phi|^3 H^8 \cdot \sup_{h \in H} |\Phi_h| \cdot \sup_{\phi \in \Phi} |\phi| \bar{\mathsf{C}}_{\varepsilon}(M, \Phi, \mu)^3}{\varepsilon^3}. \tag{21}$$

The observation space for all the Block-MDPs is given by $\mathcal{X} = \cup_{h=1}^{H} \mathcal{X}_h$, where

$$\mathcal{X}_h = \{u_h, v_h, w_h\} \cup \left(\cup_{\phi \in \Phi_h} \mathcal{X}(\phi)\right).$$

The Block-MDPs that we construct next will different in terms of which observations from $\mathcal{X}(\phi)$ will be assigned to latent states $z \in \phi$. To make this explicit, we rely on decoder functions that map $\psi : \mathcal{X} \mapsto \mathcal{Z}$. Without loss of generality, assume that all decoders that we will consider satisfy $\psi(u_h) = u_h$, $\psi(v_h) = v_h$ and $\psi(w_h) = w_h$ for all $h \in [H]$. Additionally, given a decoder ψ, we define the set $\mathcal{X}_\psi(z) = \{x \in \mathcal{X} \mid \psi(x) = z\}$. We finally define the set Ψ as the set of all possible decoders which ensure that for any $\phi \in \Phi$, each latent state $z \in \mathcal{Z}$ gets the same number of observations from $\mathcal{Z}$. In particular,

$$\Psi = \left\{ \psi : \mathcal{X} \mapsto \mathcal{Z} \mid \forall \phi \in \Phi, \forall z \in \phi : \ \mathcal{X}_\psi(z) \subseteq \mathcal{X} \text{ and } |\mathcal{X}_\psi(z)| = \frac{|\mathcal{X}(\phi)|}{|\phi|} \right\}.$$

Offline Policy Evaluation (OPE) Problem given ψ. Given a decoder ψ, and the above notation, we will lift OPE problem $\mathfrak{g}^{(1)} = (M^{(1)}, \mu', \pi_e, \mathcal{F})$ and $\mathfrak{g}^{(2)} = (M^{(2)}, \mu', \pi_e, \mathcal{F})$ into OPE problems $\mathfrak{g}^{(1)}_\psi = (\check{M}^{(1)}_\psi, \check{\mu}', \check{\pi}_e, \check{\mathcal{F}}_\psi)$ and $\mathfrak{g}^{(2)}_\psi = (\check{M}^{(2)}_\psi, \check{\mu}', \check{\pi}_e, \check{\mathcal{F}}_\psi)$ using the recipe in Appendix C.4.1, with $\check{\mathcal{F}}_\psi$ satisfies (Q, W)-realizability.

Family of OPE problems. We finally define the family $\mathfrak{G}$ for OPE problems as

$$\mathfrak{G} = \bigcup_{\psi \in \Psi} \{\mathfrak{g}^{(1)}_\psi, \mathfrak{g}^{(2)}_\psi\}.$$

We note the following useful technical lemma.

Lemma 8. *For any $\psi, \psi' \in \Psi$, we have $\check{\mathcal{F}}_\psi = \check{\mathcal{F}}_{\psi'}$.*

Proof of Lemma 8. We will only prove the results for $\check{M}^{(1)}_\psi$. To verify this, we only need to prove that for any $\psi, \psi' \in \Psi$, we have for any $1 \le h \le H$,

$$Q^{\check{\pi}_e}_h(\cdot; \check{M}^{(1)}_\psi) = Q^{\check{\pi}_e}_h(\cdot; \check{M}^{(1)}_{\psi'}) \quad \text{and} \quad W(h; \check{\mu}'_\psi, \check{M}^{(1)}_\psi)) = W(h; \check{\mu}'_\psi, \check{M}^{(1)}_{\psi'})).$$

The second equation directly follows from Lemma 6. In the next, we will verify the first equation. Lemma 5 gives that for any $\phi \in \Phi_h$ and $x \in \mathcal{X}(\phi)$ and $a \in \mathcal{A}$, we have

$$Q^{\check{\pi}_e}_h(x, a; \check{M}^{(1)}_\psi) = Q^{\pi_e}_h(\psi(x), a; M^{(1)}).$$

Next, Lemma 4 (a) indicates that

$$Q^{\pi_e}_h(\psi(x), a; M^{(1)}) = Q^{\bar{\pi}_e}_h(\zeta(\psi(x)), a; \bar{M}^{(1)}).$$

Notice that for every $x \in \mathcal{X}(\phi)$, we have $\zeta(\psi(x)) = \phi$ for any $\psi \in \Psi$. Hence for any $\psi \in \Psi$, we have $Q^{\check{\pi}_e}_h(x, a; \check{M}^{(1)}_\psi) = Q^{\bar{\pi}_e}_h(\phi, a; \bar{M}^{(1)})$ for every $\phi \in \Phi$ and $x \in \mathcal{X}(\phi)$, which is independent to ψ. ∎

Thus, moving forward, whenever clear from context, we will use the notation $\mathcal{F}$ to denote $\check{\mathcal{F}}_\psi$ for any $\psi \in \Psi$.

C.5 Proof of Theorem 3.1

After constructing the class $\mathfrak{G}$, we have finished the construction step. In the rest of the section we will prove the theorem by analyzing properties of OPE problems in $\mathfrak{G}$. In fact, we will prove the following stronger results than Theorem 3.1:

Theorem C.1. *Class $\mathfrak{G}$ satisfies the following properties:*

(a) Function class $\mathcal{F}$ with only two elements realizes the (Q, W)-function for all OPE *problems in $\mathfrak{G}$;*

(b) For any OPE *problem $(\check{M}, \mathcal{F}, \check{\mu}', \check{\pi}_e)$, we have $\mathsf{C}(\check{M}, \check{\mu}', \check{\pi}_e) \le 6\mathsf{C}(M, \mu, \pi_e)$;*

(c) Let $N = o\big(H\bar{\mathsf{C}}_\varepsilon(M,\Phi,\mu)/\varepsilon\big)$. For any algorithm which takes $\mathcal{D} = \cup_{h=1}^H \mathcal{D}_h$ as input and output the evaluation of value function $\widehat{V}$, there must exist some $\mathfrak{g} = (\check{M}, \mathcal{F}, \check{\mu}', \check{\pi}_e) \in \mathfrak{G}$ such that the algorithm fails to output ε/H-accurate evaluation with probability at least $1/2$ if the dataset $\mathcal{D}_h$ are collected according to $\check{\mu}$ and $\check{M}$ using N samples.

Proposition 2. *The aggregated concentrability coefficient of* OPE *problems $\mathfrak{g} = (\check{M}, \mathcal{F}, \check{\mu}', \check{\pi}_e) \in \mathfrak{G}$ is of the same order of the original one, i.e. $\bar{\mathsf{C}}_\varepsilon(\check{M}, \check{\Phi}, \check{\mu}') = \Theta(\bar{\mathsf{C}}_\varepsilon(M, \Phi, \mu))$ where the aggregation scheme $\check{\Phi}$ is defined over $\mathcal{X}$ such that $\check{\Phi}_h = \{\mathcal{X}(\phi) : \phi \in \Phi_h\} \cup \{\{u_h\}, \{v_h\}, \{w_h\}\}$.*

C.5.1 Technical Lemmas for Proof of Theorem C.1

In this subsection we provide several useful technical lemmas for proof of Theorem C.1.

To begin with, we denote

$$\check{\mu}_\psi(x) := \begin{cases} \mu(z)/|\mathcal{X}_\psi(z)| & \forall z \in \mathcal{Z}, x \in \mathcal{X}_\psi(z), \\ 0 & \forall x \in \{u_h, v_h, w_h\}_{h=1}^H \end{cases}.$$

Since the transition of u_h, v_h, w_h are already known, and $\check{\mu}_\psi(x) \ge \check{\mu}'_\psi(x)$ for all $x \in \mathcal{X}_\psi(z)$ and $z \in \mathcal{Z}$, in the following we only need to prove the results for OPE problems $\mathfrak{g} = (\check{M}, \mathcal{F}, \check{\mu}, \check{\pi}_e)$.

In the following, we use $\mathbb{P}_{h,n}(\cdot; \check{\mu}, \check{M})$ where $\check{M} = (\mathcal{X}, \mathcal{A}, H, \check{T}, \check{R}, \check{\rho})$ to denote the law of n tuples of (x, a, r, x') jointly, where each tuple is i.i.d. collecting as follows: first sample $(x, a) \sim \check{\mu}_h$, then sample $r \sim \check{R}(\cdot \mid x, a), x' \sim \check{T}(x' \mid x, a)$. Let

$$\mathbb{P}_{h,n}^{(1)} = \frac{1}{|\Psi|}\sum_{\psi\in\Psi} \mathbb{P}_{h,n}(\cdot; \check{\mu}_\psi, \check{M}_\psi^{(1)}) \quad \text{and} \quad \mathbb{P}_{h,n}^{(2)} = \frac{1}{|\Psi|}\sum_{\psi\in\Psi} \mathbb{P}_{h,n}(\cdot; \check{\mu}_\psi, \check{M}_\psi^{(2)}). \tag{22}$$

Furthermore, for any $\check{M}$, let $\mathbb{P}_n(\cdot; \check{\mu}, \check{M}) := \bigotimes_{h=1}^{H-1} \mathbb{P}_{h,n}(\cdot; \check{\mu}, \check{M})$, and using this notation, define

$$\mathbb{P}_n^{(1)} = \frac{1}{|\Psi|}\sum_{\psi\in\Psi} \mathbb{P}_n(\cdot; \check{\mu}_\psi, \check{M}_\psi^{(1)}) \quad \text{and} \quad \mathbb{P}_n^{(2)} = \frac{1}{|\Psi|}\sum_{\psi\in\Psi} \mathbb{P}_n(\cdot; \check{\mu}_\psi, \check{M}_\psi^{(2)}). \tag{23}$$

Additionally, since the state space $\mathcal{S} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_H$ is layered, we get that ψ can be separated across layers, and thus the above definitions imply that

$$\mathbb{P}_n^{(1)} = \bigotimes_{h=1}^{H-1} \mathbb{P}_{h,n}^{(1)} \quad \text{and} \quad \mathbb{P}_n^{(2)} = \bigotimes_{h=1}^{H-1} \mathbb{P}_{h,n}^{(2)}. \tag{24}$$

We have the following inequality for TV distance between product measures.

Lemma 9 (Polyanskiy and Wu (2014), I.33(b))**.** *For distributions $\mathbb{P}_1, \cdots, \mathbb{P}_H$ and $\mathbb{Q}_1, \cdots, \mathbb{Q}_H$, we have*

$$D_{\mathrm{TV}}\left(\bigotimes_{h=1}^{H-1} \mathbb{P}_h, \bigotimes_{h=1}^{H-1} \mathbb{Q}_h\right) \le \sum_{h=1}^{H-1} D_{\mathrm{TV}}(\mathbb{P}_h, \mathbb{Q}_h).$$

To prove this theorem, we first show the following lemma.

Lemma 10. *For any algorithm which takes $D_{h,n} = \{(x_{h,i}, a_{h,i}, r_{h,i}, x'_{h,i})\}_{i=1}^n$ where $h \in [H-1]$ as input and returns a value $\widehat{V}(D_{1:H-1,n})$ (where we use $D_{1:H-1,n}$ to denote $D_{1,n}, \cdots, D_{H-1,n}$), it must satisfy*

$$\sup_{\psi\in\Psi, i\in\{1,2\}} \mathbb{E}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(i)})}\left[\left|\widehat{V}(D_{1:H-1,n}) - V(\check{\rho};\check{M})\right|\right] \geq \frac{\varepsilon}{4H}\cdot\left(1 - \sum_{h=1}^{H-1} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \mathbb{P}_{h,n}^{(2)})\right).$$

Proof of Lemma 10. Lemma 8 gives that for any $\check{M}, \check{M}' \in \{\check{M}_\psi^{(1)} : \psi \in \Psi\}$, we have $Q_h^{\check{\pi}_e}(\cdot;\check{M}) = Q_h^{\check{\pi}_e}(\cdot;\check{M}')$, which implies that MDPs in $\{\check{M}_\psi^{(1)} : \psi \in \Psi\}$ share the same value function. Hence we have

$$V_h^{\check{\pi}_e}(\check{\rho};\check{M}) = V_h^{\check{\pi}_e}(\check{\rho};\check{M}'), \quad \forall M, M' \in \{\check{M}_\psi^{(1)} : \psi \in \Psi\}.$$

In the following, we denote the above quantity to be $V^{(1)}(\check{\rho})$. Similarly, we denote $V^{(2)}(\check{\rho})$ to be the counterpart for MDPs in $\{\check{M}_\psi^{(2)} : \psi \in \Psi\}$.

For any dataset $D_{1:H-1,n}$, we use $\delta(D_{1:H-1,n})$ to denote the following random variable:

$$\delta(D_{1:H-1,n}) = \mathbb{I}\left\{\widehat{V}(D_{1:H-1,n}) \leq \frac{V^{(1)}(\check{\rho}) + V^{(2)}(\check{\rho})}{2}\right\} \in \{0,1\},$$

Then for any $\psi \in \Psi$, we have

$$\begin{aligned}
&\mathbb{E}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(1)})}\left|\widehat{V}(D_{1:H-1,n}) - V(\check{\rho};\check{M})\right| \\
&\quad\geq \mathbb{E}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(1)})}\left[\delta(D_{1:H-1,n})\cdot\left|\widehat{V}(D_{1:H-1,n}) - V^{(1)}(\check{\rho})\right|\right] \\
&\quad\geq \mathbb{E}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(1)})}\left[\delta(D_{1:H-1,n})\cdot\left|\frac{V^{(1)}(\check{\rho}) + V^{(2)}(\check{\rho})}{2} - V^{(1)}(\check{\rho})\right|\right] \\
&\quad\geq \mathbb{P}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(1)})}(\delta(D_{1:H-1,n}) = 1)\cdot\frac{V^{(1)}(\check{\rho}) - V^{(2)}(\check{\rho})}{2} \\
&\quad= \frac{\varepsilon}{2H}\cdot\mathbb{P}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(1)})}(\delta(D_{1:H-1,n}) = 1),
\end{aligned}$$

where in the last equation we use (b) and Lemma 5. Similarly, for $\psi \in \Psi$, we have

$$\mathbb{E}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(2)})}\left|\widehat{V}(D_{1:H-1,n}) - V(\check{\rho};\check{M})\right| \geq \frac{\varepsilon}{2H}\cdot\mathbb{P}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(2)})}(\delta(D_{1:H-1,n}) = 0).$$

Therefore, we obtain that

$$\begin{aligned}
&\sup_{\psi\in\Psi, i\in\{1,2\}} \mathbb{E}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(i)})}\left[\left|\widehat{V}(D_{1:H-1,n}) - V(\check{\rho};\check{M})\right|\right] \\
&\geq \frac{1}{2|\Psi|}\sum_{\psi\in\Psi}\mathbb{E}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(1)})}\left[\left|\widehat{V}(D_{1:H-1,n}) - V(\check{\rho};\check{M})\right|\right] \\
&\quad + \frac{1}{2|\Psi|}\sum_{\psi\in\Psi}\mathbb{E}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(2)})}\left[\left|\widehat{V}(D_{1:H-1,n}) - V(\check{\rho};\check{M})\right|\right] \\
&\geq \frac{\varepsilon}{4H}\cdot\left(\frac{1}{|\Psi|}\sum_{\psi\in\Psi}\mathbb{P}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(1)})}(\delta(D_{1:H-1,n}) = 1) + \frac{1}{|\Psi|}\sum_{\psi\in\Psi}\mathbb{P}_{D_{1:H-1,n}\sim\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(2)})}(\delta(D_{,n}) = 0)\right) \\
&\overset{(i)}{\geq} \frac{\varepsilon}{4H}\cdot\left(1 - D_{\mathrm{TV}}\left(\frac{1}{|\Psi|}\sum_{\psi\in\Psi}\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(1)}), \quad \frac{1}{|\Psi|}\sum_{\psi\in\Psi}\mathbb{P}_n(\cdot;\check{\mu}_\psi,\check{M}_\psi^{(2)})\right)\right)
\end{aligned}$$

$$\overset{(ii)}{\geq} \frac{\varepsilon}{4H} \cdot \left(1 - D_{\mathrm{TV}}(\mathbb{P}_n^{(1)}, \mathbb{P}_n^{(2)})\right)$$
$$\overset{(ii)}{\geq} \frac{\varepsilon}{4H} \cdot \left(1 - \sum_{h=1}^{H-1} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \mathbb{P}_{h,n}^{(2)})\right),$$

where in the inequality (i) we use $\mathbb{P}(\mathcal{E}) + \mathbb{Q}(\mathcal{E}^c) \geq 1 - D_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$ for any event $\mathcal{E}$, in the inequality (ii) we use (23), and in the equation (iii) we use (24) and Lemma 9. ■

Hence we only need to upper bound the TV distance between $\mathbb{P}_{h,n}^{(1)}$ and $\mathbb{P}_{h,n}^{(2)}$, which is proved in the following lemma.

Lemma 11. *Suppose for every $\phi \in \Psi$, we have*

$$|\mathcal{X}(\phi)| \gtrsim \frac{|\phi|^3 H^8 \cdot \sup_{h \in H} |\Phi_h| \cdot \sup_{\phi \in \Phi} |\phi| \bar{\mathsf{C}}_\varepsilon(M, \Phi, \mu)^3}{\varepsilon^3}.$$

If $n \leq \frac{H}{8\varepsilon} \bar{\mathsf{C}}_\varepsilon(M, \Phi, \mu)$, then we have

$$\sum_{h=1}^{H-1} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \mathbb{P}_{h,n}^{(2)}) \leq \frac{1}{2}.$$

Proof of Lemma 11. At a high level, the proof contains three steps:

(i) We first define intermediate distributions $\mathbb{P}_{h,n}^{(0)}$ over tuples (x, a, r, x') and observe that via Triangle inequality that

$$\sum_{h=1}^{H-1} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \mathbb{P}_{h,n}^{(2)}) \leq \sum_{h=1}^{H-1} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \mathbb{P}_{h,n}^{(0)}) + \sum_{h=1}^{H-1} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(0)}, \mathbb{P}_{h,n}^{(2)}).$$

The final bound follows by showing that $\sum_{h=1}^{H-1} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(i)}, \mathbb{P}_{h,n}^{(0)}) \leq 1/4$ for all $i \in \{1, 2\}$.

(ii) Note that $\mathbb{P}_{h,n}^{(i)}$ is a distribution over tuples (x, a, r, x') where the instantaneous reward $r \sim \check{R}^{(i)}(r \mid x, a)$. We first simplify our objective a bit by converting $\mathbb{P}_{h,n}^{(i)}$ to $\widetilde{\mathbb{P}}_{h,n}^{(i)}$, where $\widetilde{\mathbb{P}}_{h,n}^{(i)}$ is a distribution over tuples (x, a, r, x') where $r \sim \check{R}^{(0)}(r \mid x, a)$ where $\check{R}^{(0)}$ is the reward function in $\mathbb{P}_{h,n}^{(0)}$. Another application of Triangle inequality implies that,

$$\sum_{h=1}^{H-1} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \mathbb{P}_{h,n}^{(0)}) \leq \sum_{h=1}^{H-1} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \widetilde{\mathbb{P}}_{h,n}^{(1)}) + \sum_{h=1}^{H-1} D_{\mathrm{TV}}(\widetilde{\mathbb{P}}_{h,n}^{(1)}, \mathbb{P}_{h,n}^{(0)}).$$

Bounding the first term above is straightforward.

(iii) We finally bound the term $D_{\mathrm{TV}}(\widetilde{\mathbb{P}}_{h,n}^{(1)}, \mathbb{P}_{h,n}^{(0)})$ for each $h \in [H-1]$ by delving further into the structure of the MDPs and the underlying data distribution in $\widetilde{\mathbb{P}}_n^{(1)}$. Most of the proof will be spend on bound.

Part-(i): Construction of $\mathbb{P}_n^{(0)}$. We first define additional notation. Let the distribution $\nu \in \Delta(\mathcal{X} \times \mathcal{A} \times \mathcal{X})$ such that for any $h \in [H-1]$, $\phi \in \Phi_h, \phi' \in \Phi_{h+1}$ and $x \in \mathcal{X}(\phi), x' \in \mathcal{X}(\phi')$, $a \in \mathcal{A}$, we have

$$\nu_h(x, a, x') = \sum_{z \in \phi} \sum_{z' \in \phi'} \frac{\mu(z, a) T^{(1)}(z' \mid z, a)}{|\mathcal{X}(\phi)||\mathcal{X}(\phi')|}, \tag{25}$$

Additionally, we define a reward distribution $\check{R}_h^{(0)} \in \Delta(\{-1,1\})$ such that

$$\check{R}_h^{(0)}(\cdot \mid x,a) = \begin{cases} \delta_1(\cdot) & x = u_{h+1}, \\ \delta_{-1}(\cdot) & x = v_{h+1}, \\ \frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot) & \text{otherwise}, \end{cases} \tag{26}$$

where we use $\delta_t(\cdot)$ denote the density of delta-distribution at t. Given ν and $\check{R}_h^{(0)}$ above, we define $\mathbb{P}_h^{(0)} \in \Delta(\mathcal{X}_h \times \mathcal{A} \times \{-1,1\} \times \mathcal{X}_{h+1})$ as

$$\mathbb{P}_h^{(0)}((x,a,r,x')) := \nu_h(x,a,x')\check{R}_h^{(0)}(r \mid x,a), \tag{27}$$

and set $\mathbb{P}_{h,n}^{(0)} = (\mathbb{P}_h^{(0)})^{\otimes n}$.

As a sanity check, note that

$$\begin{aligned} \sum_{(x,a,r,x') \in \mathcal{X}_h \times \mathcal{A} \times \{-1,1\} \times \mathcal{X}_{h+1}} \mathbb{P}_h^{(0)}((x,a,r,x')) &= \sum_{(x,a,x') \in \mathcal{X}_h \times \mathcal{A} \times \mathcal{X}_{h+1}} \nu_h(x,a,x') \\ &= \sum_{\phi \in \Phi_h} \sum_{\phi' \in \Phi'_{h+1} \cup \{u_{h+1}, v_{h+1}\}} \sum_{z \in \phi} \sum_{z' \in \phi'} \mu(z,a) T^{(1)}(z' \mid z,a) = 1, \end{aligned}$$

and thus $\mathbb{P}_h^{(0)}$ is a valid distribution; the above also implies that ν_h (defined above) is a valid distribution. Furthermore, while the above definition is based in $T^{(1)}$, we could have also defined $\mathbb{P}_h^{(0)}$ using $T^{(2)}$ and would have ended up with the same distribution since $\sum_{z \in \phi} \mu(z,a) T^{(1)}(z' \mid z,a) = \sum_{z \in \phi} \mu(z,a) T^{(2)}(z' \mid z,a)$ for any $\phi \in \Phi$, $z \in \mathcal{Z}'$ due to Lemma 4-(c).

Given $\mathbb{P}_{h,n}^{(0)}$, using Triangle inequality we have

$$\sum_{h \in [H-1]} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \mathbb{P}_{h,n}^{(2)}) \le \sum_{h \in [H-1]} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \mathbb{P}_{h,n}^{(0)}) + \sum_{h \in [H-1]} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(0)}, \mathbb{P}_{h,n}^{(2)}). \tag{28}$$

In the next part, we prove that $\sum_{h \in [H-1]} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \mathbb{P}_{h,n}^{(0)}) \le 1/4$. The proof for $\mathbb{P}_{h,n}^{(2)}$ follows similarly, and combining the two bound gives the desired statement.

Part-(ii): Construction of $\widetilde{\mathbb{P}}_n^{(1)}$ and bounding $D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \widetilde{\mathbb{P}}_{h,n}^{(1)})$. First recall that from (23), we can write

$$\mathbb{P}_{h,n}^{(1)}(\{(x_i,a_i,r_i,x_i')\}_{i=1}^n) = \frac{1}{|\Psi|} \sum_{\psi \in \Psi} \prod_{i=1}^n \check{\mu}_{h,\psi}(x_i,a_i) \check{R}_h^{(1)}(r_i \mid x_i,a_i) \check{T}_\psi^{(1)}(x_i' \mid x_i,a_i),$$

where $\check{R}_h^{(1)}$ is given by

$$\check{R}_h^{(1)}(\cdot \mid x,a) = \begin{cases} \delta_1(\cdot) & x = u_{h+1}, \\ \delta_{-1}(\cdot) & x = v_{h+1}, \\ \frac{\delta_1(\cdot)+\delta_{-1}(\cdot)}{2} + \frac{\delta_1(\cdot)-\delta_{-1}(\cdot)}{2} \cdot \frac{\varepsilon}{2H \sum_{\phi \in \mathcal{I}} d_{h^\star}^{\bar{\pi}_{\mathrm{e}}}(\phi;\bar{M})} & x \in \mathcal{X}(\phi) \text{ for } \phi \in \mathcal{I}, a = \check{\pi}_{\mathrm{e}}(x) \\ \frac{1}{2}\delta_1(\cdot) + \frac{1}{2}\delta_{-1}(\cdot) & \text{otherwise}. \end{cases} \tag{29}$$

We next define the distribution $\widetilde{\mathbb{P}}_{h,n}^{(1)}$ similar to $\mathbb{P}_{h,n}^{(1)}$, but where we use $\check{R}^{(0)}$ (given in (26)) instead of $\check{R}^{(1)}$ to remove the dependence on the rewards on i. In particular, we define

$$\tilde{\mathbb{P}}_{h,n}^{(1)}(\{(x_i,a_i,r_i,x_i')\}_{i=1}^n) = \frac{1}{|\Psi|} \sum_{\psi \in \Psi} \prod_{i=1}^n \check{\mu}_{h,\psi}(x_i,a_i) \check{R}_h^{(0)}(r_i \mid x_i,a_i) \check{T}_\psi^{(1)}(x_i' \mid x_i,a_i). \tag{30}$$

If we denote

$$\check{r}_h^{(1)}(x_i,a_i) := \mathbb{E}[\check{R}_h^{(1)}(\cdot \mid x,a)] \quad \text{and} \quad \check{r}_h^{(0)}(x_i,a_i) := \mathbb{E}[\check{R}_h^{(0)}(\cdot \mid x,a)].$$

Note that when $n \le \frac{H\bar{\mathsf{C}}_\varepsilon(M,\Phi,\mu)}{2\varepsilon}$, using the above definitions, we have

$$\begin{aligned}
&\sum_{h=1}^{H-1} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \tilde{\mathbb{P}}_{h,n}^{(1)}) \\
&\quad = \frac{1}{2} \sum_{h\in[H-1]} \sum_{\substack{\{(x_i,a_i,r_i,x_i')\}_{i=1}^n \\ \in(\mathcal{X}_h\times\mathcal{A}\times\{0,1\}\times\mathcal{X}_{h+1})^n}} \left|\mathbb{P}_{h,n}^{(1)}\big(\{(x_i,a_i,r_i,x_i')\}_{i=1}^n\big) - \tilde{\mathbb{P}}_{h,n}^{(1)}\big(\{(x_i,a_i,r_i,x_i')\}_{i=1}^n\big)\right| \\
&\quad \overset{(i)}{\le} \frac{1}{2} \sum_{h\in[H-1]} \sum_{\{(x_i,a_i)\}_{i=1}^n \in(\mathcal{X}_h\times\mathcal{A})^n} \frac{1}{|\Psi|} \sum_{\psi\in\Psi} \sum_{i=1}^n |\check{r}_h^{(1)}(x_i,a_i) - \check{r}_h^{(0)}(x_i,a_i)| \prod_{i=1}^n \check{\mu}_{h,\psi}(x_i,a_i) \\
&\quad \overset{(ii)}{=} \frac{1}{2|\Psi|} \sum_{\psi\in\Psi} \sum_{i=1}^n \sum_{h\in[H-1]} \sum_{(x_i,a_i)\in\mathcal{X}_h\times\mathcal{A}} \check{\mu}_{h,\psi}(x_i,a_i) |\check{r}_h^{(1)}(x_i,a_i) - \check{r}_h^{(0)}(x_i,a_i)| \\
&\quad \overset{(iii)}{\le} \frac{1}{2|\Psi|} \sum_{\psi\in\Psi} \sum_{i=1}^n \sum_{h\in[H-1]} \sum_{x_i\in\mathcal{X}_h} \check{\mu}_\psi(x_i,\check{\pi}_{\mathrm{e}}(x_i)) \frac{\varepsilon \mathbb{I}\{\zeta(\psi(x_i)) \in \mathcal{I})\}}{2H \sum_{\phi\in\mathcal{I}} d_h^{\bar{\pi}_{\mathrm{e}}}(\phi;\bar{M})} \\
&\quad = \frac{n\varepsilon}{4H \sum_{\phi\in\mathcal{I}} d_h^{\bar{\pi}_{\mathrm{e}}}(\phi;\bar{M})} \sum_{z:\zeta(z)\in\mathcal{I}} \mu(z,\pi_{\mathrm{e}}(z)),
\end{aligned}$$

where inequality (i) follows from Triangle Inequality, inequality (ii) follows by rearranging the terms and using the fact that $\sum_{x_i,a_i} \check{\mu}_{h,\psi}(x_i,a_i) = 1$. The inequality (iii) is from plugging in the forms of $\check{R}^{(1)}$ and $\check{R}^{(0)}$ from (26) and (29). Finally, the last line uses the fact that

$$\frac{1}{|\Psi|} \sum_{\psi\in\Psi} \sum_{x_i} \check{\mu}_\psi(x_i,\check{\pi}_{\mathrm{e}}(x_i)) \mathbb{I}\{\zeta(\psi(x_i)) \in \mathcal{I})\} = \sum_{z:\zeta(z)\in\mathcal{I}} \mu(z,\pi_{\mathrm{e}}(z)),$$

from the definition of $\check{\mu}$. Next, using the definition of $\bar{\mathsf{C}}_\varepsilon(M,\Phi,\mu)$ in Definition 3.1, and recalling that $\mathcal{I}$ is the maximizer aggregation in Definition 3.1, we get that

$$\bar{\mathsf{C}}_\varepsilon(M,\Phi,\mu) = \frac{\sum_{\phi\in\mathcal{I}} d_h^{\bar{\pi}_{\mathrm{e}}}(\phi;\bar{M})}{\sum_{z:\zeta(z)\in\mathcal{I}} \mu(z,\pi_{\mathrm{e}}(z))},$$

which implies that

$$\sum_{h=1}^{H-1} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \tilde{\mathbb{P}}_{h,n}^{(1)}) \le \frac{n\varepsilon}{4H\bar{\mathsf{C}}_\varepsilon(M,\Phi,\mu)} \le \frac{1}{8},$$

where the last inequality holds since $n \le \frac{H\bar{\mathsf{C}}_\varepsilon(M,\Phi,\mu)}{2\varepsilon}$.

Thus, using Triangle inequality,

$$\begin{aligned}
\sum_{h\in[H-1]} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \mathbb{P}_{h,n}^{(0)}) &\le \sum_{h\in[H-1]} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(1)}, \tilde{\mathbb{P}}_{h,n}^{(1)}) + \sum_{h\in[H-1]} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(0)}, \tilde{\mathbb{P}}_{h,n}^{(1)}) \\
&\le \frac{1}{8} + \sum_{h\in[H-1]} D_{\mathrm{TV}}(\mathbb{P}_{h,n}^{(0)}, \tilde{\mathbb{P}}_{h,n}^{(1)}).
\end{aligned} \tag{31}$$

Part-(iii): Bound on $D_{\mathrm{TV}}(\mathbb{P}^{(0)}_{h,n}, \tilde{\mathbb{P}}^{(1)}_{h,n})$. First note that, from Polyanskiy and Wu (2014, Proposition 7.13), we have

$$D_{\mathrm{TV}}(\mathbb{P}^{(0)}_{h,n}, \tilde{\mathbb{P}}^{(1)}_{h,n}) \le \frac{1}{2}\sqrt{D_{\chi^2}(\tilde{\mathbb{P}}^{(1)}_{h,n} \| \mathbb{P}^{(0)}_{h,n})}. \tag{32}$$

Using the form of χ^2-divergence, we note that

$$\begin{aligned}
&D_{\chi^2}(\tilde{\mathbb{P}}^{(1)}_{h,n} \| \mathbb{P}^{(0)}_{h,n}) \\
&\quad = \mathbb{E}_{\{(x_i,a_i,r_i,x_i')\}_{i=1}^n \sim \mathbb{P}^{(0)}_{h,n}} \left[\left(\frac{\tilde{\mathbb{P}}^{(1)}_{h,n}(\{(x_i,a_i,r_i,x_i')\}_{i=1}^n)}{\mathbb{P}^{(0)}_{h,n}(\{(x_i,a_i,r_i,x_i')\}_{i=1}^n)} \right)^2 \right] - 1 \\
&\quad = \mathbb{E}_{\{(x_i,a_i,r_i,x_i')\}_{i=1}^n \sim \mathbb{P}^{(0)}_{h,n}} \left[\left(\frac{\frac{1}{|\Psi|}\sum_{\psi\in\Psi}\prod_{i=1}^n \check{\mu}_{h,\psi}(x_i,a_i)\check{R}^{(0)}_h(r_i \mid x_i,a_i)\check{T}^{(1)}_{\psi}(x_i' \mid x_i,a_i)}{\prod_{i=1}^n \nu_h(x_i,a_i,x_i')\check{R}^{(0)}_h(r_i \mid x_i,a_i)} \right)^2 \right] - 1,
\end{aligned}$$

where the second line plugs in the definition of $\tilde{\mathbb{P}}^{(1)}_{h,n}$ in (30) and $\mathbb{P}^{(0)}_{h,n}$ in (27). We next note that the terms $\check{R}^{(0)}_h(r_i \mid x_i,a_i)$ will cancel out in the ratio, thus implying that the expression is independent of $\{r_i\}_{i=1}^n$. Furthermore, from the definition of $\mathbb{P}^{(0)}_{h,n}$ in (27), we note that sampling $\{(x_i,a_i,x_i')\}_{i=1}^n \sim \mathbb{P}^{(0)}_{h,n}$ is same as sampling $\{(x_i,a_i,x_i')\}_{i=1}^n \sim \nu_h^{\otimes n}$. Thus,

$$\begin{aligned}
&D_{\chi^2}(\tilde{\mathbb{P}}^{(1)}_{h,n} \| \mathbb{P}^{(0)}_{h,n}) \\
&\quad = \mathbb{E}_{\{(x_i,a_i,x_i')\}_{i=1}^n \sim \nu_h^{\otimes n}} \left[\left(\frac{\frac{1}{|\Psi|}\sum_{\psi\in\Psi}\prod_{i=1}^n \check{\mu}_{h,\psi}(x_i,a_i)\check{T}^{(1)}_{\psi}(x_i' \mid x_i,a_i)}{\prod_{i=1}^n \nu_h(x_i,a_i,x_i')} \right)^2 \right] - 1 \\
&\quad = \frac{1}{|\Psi|^2}\sum_{\psi_1,\psi_2\in\Psi} \mathbb{E}_{\{(x_i,a_i,x_i')\}_{i=1}^n \sim \nu_h^{\otimes n}} \left[\prod_{i=1}^n \frac{\check{\mu}_{h,\psi_1}(x_i,a_i)\check{\mu}_{h,\psi_2}(x_i,a_i)\check{T}^{(1)}_{\psi_1}(x_i' \mid x_i,a_i)\check{T}^{(1)}_{\psi_2}(x_i' \mid x_i,a_i)}{\nu_h(x_i,a_i,x_i')^2} \right] - 1 \\
&\quad = \frac{1}{|\Psi|^2}\sum_{\psi_1,\psi_2\in\Psi} \Bigg(\underbrace{\mathbb{E}_{(x,a,x')\sim\nu_h} \left[\frac{\check{\mu}_{h,\psi_1}(x,a)\check{\mu}_{h,\psi_2}(x,a)\check{T}^{(1)}_{\psi_1}(x' \mid x,a)\check{T}^{(1)}_{\psi_2}(x' \mid x,a)}{\nu_h(x,a,x')^2} \right]}_{:=\mathfrak{B}(\psi_1,\psi_2)} \Bigg)^n - 1,
\end{aligned} \tag{33}$$

where the second equality follows by expanding the square and rearranging the product. Finally, the last line exchange the expectation and the product by using the fact that tuples $\{(x_i,a_i,x_i')\}_{i=1}^n$ are i.i.d. sampled from ν_h. In the following, we will complete the proof by giving bounds on the terms $\mathfrak{B}(\psi_1,\psi_2)$ under various conditions on ψ_1 and ψ_2. However, we need additional notation before we proceed:

- Given any $h \in [H-1]$, $\phi \in \Phi_h$, $z, z' \in \phi$, and $\psi_1, \psi_2 \in \Psi$, we define

$$\theta_h(\phi, z_1, z_2; \psi_1, \psi_2) := \frac{|\mathcal{X}_{\psi_1}(z_1) \cap \mathcal{X}_{\psi_2}(z_2)|}{|\mathcal{X}(\phi)|} \tag{34}$$

  to denote the fraction of repeated observations between z and z' amongst all the observations that correspond to aggregation ϕ.

- Let $\xi = 1/(64H^2n) \in (0, 1/n)$. For any $h \in [H-1]$, $\phi \in \Phi_h$, and $z_1, z_2, \in \phi$, we define

$$\Gamma_h(\xi; \phi, z_1, z_2) = \left\{ (\psi_1,\psi_2) \in \Psi^2 \mid \theta(\phi, z_1, z_2; \psi_1, \psi_2) \le \frac{1+\xi}{|\phi|^2} \right\} \tag{35}$$

to denote the set of all pairs (ψ_1, ψ_2) for which the corresponding ratios $\theta(\phi, z_1, z_2; \psi_1, \psi_2)$ are small.

- Finally, we define the set

$$\Gamma(\xi) = \bigcap_{h\in[H-1],\phi\in\Phi_h\cup\{u_h,v_h\}} \left(\bigcap_{z_1,z_2,\in\phi} \Gamma(\xi;\phi,z_1,z_2) \right). \tag{36}$$

We now have all the necessary notation to proceed with the proof. We split the terms $\mathfrak{B}(\psi_1,\psi_2)$ appearing in (33) under two separate scenarios:

- Case 1: $\mathfrak{B}(\psi_1,\psi_2) \in \Gamma(\xi)$. Here, Lemma 12 (below) implies that

$$\mathfrak{B}(\psi_1,\psi_2) \le (1+\xi)^2.$$

- Case 2: $\mathfrak{B}(\psi_1,\psi_2) \notin \Gamma(\xi)$. Here, Lemma 13 (below) implies that

$$\mathfrak{B}(\psi_1,\psi_2) \le \sup_{\phi\in\Phi} |\phi|^4.$$

Combining the two above in (33), we get that

$$\begin{aligned}
D_{\chi^2}(\tilde{\mathbb{P}}^{(1)}_{h,n} \| \mathbb{P}^{(0)}_{h,n}) &\le \frac{1}{|\Psi|^2} \sum_{\psi_1,\psi_2\in\Psi} \left[\left(\mathbb{I}\{(\psi_1,\psi_2)\in\Gamma(\xi)\}(1+\xi)^{2n} + \mathbb{I}\{(\psi_1,\psi_2)\notin\Gamma(\xi)\} \sup_{\phi\in\Phi}|\phi|^{4n} \right) \right] - 1 \\
&\le \frac{1}{|\Psi|^2} \sum_{\psi_1,\psi_2\in\Psi} \left[\left(1 + 2n\xi + \mathbb{I}\{(\psi_1,\psi_2)\notin\Gamma(\xi)\} \sup_{\phi\in\Phi}|\phi|^{4n} \right) \right] - 1 \\
&= 2\xi n + \sup_{\phi\in\Phi}|\phi|^{4n} \cdot \frac{1}{|\Psi|^2} \sum_{\psi_1,\psi_2\in\Psi} \mathbb{I}\{(\psi_1,\psi_2)\notin\Gamma(\xi)\},
\end{aligned}$$

where the second line uses the fact that $(1+\xi)^n \le 1 + 2\xi n$ for any $\xi \le 1/n$. Next, we notice that for $\phi \in \cup_{h=1}^{H}\{u_h, v_h\}$ and $z_1, z_2 \in \phi$, since $|\phi| = 1$, we always have $(\psi_1,\psi_2) \in \Gamma(\xi;\phi,z_1,z_2)$. Therefore, we have

$$\Gamma(\xi) = \bigcap_{h\in[H-1],\phi\in\Phi_h} \left(\bigcap_{z_1,z_2,\in\phi} \Gamma(\xi;\phi,z_1,z_2) \right).$$

Lemma 14 gives that for any $h \in H, \phi \in \Phi_h$ and $z_1, z_2 \in \Phi$, we always have

$$\frac{1}{|\Psi|^2} \sum_{\psi_1,\psi_2\in\Psi} \mathbb{I}\{(\psi_1,\psi_2)\notin\Gamma_h(\xi;\phi,z_1,z_2)\} \le e^{-2\xi^2 \frac{|\mathcal{X}(\phi)|}{|\phi|^3}} \le \Gamma_h(\xi;\phi,z_1,z_2)\} \le e^{-2\xi^2 \inf_{\phi\in\Phi} \frac{|\mathcal{X}(\phi)|}{|\phi|^3}},$$

which indicates that

$$\begin{aligned}
\frac{1}{|\Psi|^2} \sum_{\psi_1,\psi_2\in\Psi} \mathbb{I}\{(\psi_1,\psi_2)\notin\Gamma(\xi)\} &\le \sum_{h\in[H-1],\phi\in\Phi_h} \sum_{z_1,z_2\in\phi} \frac{1}{|\Psi|^2} \sum_{\psi_1,\psi_2\in\Psi} \mathbb{I}\{(\psi_1,\psi_2)\notin\Gamma_h(\xi;\phi,z_1,z_2)\} \\
&\le H \cdot \sup_{h\in H}|\Phi_h| \cdot \sup_{\phi\in\Phi}|\phi|^2 \cdot e^{-2\xi^2 \inf_{\phi\in\Phi} \frac{|\mathcal{X}(\phi)|}{|\phi|^3}}
\end{aligned}$$

This suggests that

$$D_{\chi^2}(\tilde{\mathbb{P}}^{(1)}_{h,n} \| \mathbb{P}^{(0)}_{h,n}) \le 2\xi n + H \cdot \sup_{h\in H}|\Phi_h| \cdot \sup_{\phi\in\Phi}|\phi|^{4n+2} \cdot e^{-2\xi^2 \inf_{\phi\in\Phi} \frac{|\mathcal{X}(\phi)|}{|\phi|^3}}.$$

Finally, when

$$|\mathcal{X}(\phi)| \geq c_0 \cdot n^3 |\phi|^3 H^5 \cdot \sup_{h\in H} |\Phi_h| \cdot \sup_{\phi\in\Phi} |\phi| \qquad \forall \phi \in \Phi$$

for some sufficiently large constant c_0, with choice $\xi = 1/(64H^2 n)$ we will have

$$D_{\chi^2}(\tilde{\mathbb{P}}^{(1)}_{h,n} \| \mathbb{P}^{(0)}_{h,n}) \leq \frac{1}{16H^2}.$$

Hence when

$$|\mathcal{X}(\phi)| \gtrsim \frac{|\phi|^3 H^8 \cdot \sup_{h\in H} |\Phi_h| \cdot \sup_{\phi\in\Phi} |\phi| \bar{\mathsf{C}}_\varepsilon(M,\Phi,\mu)^3}{\varepsilon^3} \quad \text{and} \quad n \leq \frac{H}{8\varepsilon} \bar{\mathsf{C}}_\varepsilon(M,\Phi,\mu)$$

this together with (32) implies that

$$D_{\mathrm{TV}}(\mathbb{P}^{(0)}_{h,n}, \tilde{\mathbb{P}}^{(1)}_{h,n}) \leq \frac{1}{8H}.$$

According to (31), we have

$$\sum_{h\in[H-1]} D_{\mathrm{TV}}(\mathbb{P}^{(1)}_{h,n}, \mathbb{P}^{(0)}_{h,n}) \leq \frac{1}{8} + H \cdot \frac{1}{8H} \leq \frac{1}{4}.$$

Similarly, we can also prove that

$$\sum_{h\in[H-1]} D_{\mathrm{TV}}(\mathbb{P}^{(2)}_{h,n}, \mathbb{P}^{(0)}_{h,n}) \frac{1}{4}.$$

Plugging in these two in (28), we get the desired bound.

■

Lemma 12. *We have the following property for $\mathfrak{B}_h(\psi_1,\psi_2)$ defined in (33): For any $\psi_1,\psi_2 \in \Gamma(\xi)$, we have*

$$\mathfrak{B}_h(\psi_1,\psi_2) \leq (1+\xi)^2.$$

Proof of Lemma 12. From (33), recall that

$$\mathfrak{B}_h(\psi_1,\psi_2) = \mathbb{E}_{(x,a,x')\sim\nu_h}\left[\frac{\check{\mu}_{h,\psi_1}(x,a)\check{\mu}_{h,\psi_2}(x,a)\check{T}^{(1)}_{\psi_1}(x' \mid x,a)\check{T}^{(1)}_{\psi_2}(x' \mid x,a)}{\nu_h(x,a,x')^2}\right].$$

According to our construction of $\check{\mu}_{h,\psi}$ in Appendix C.4.1, for any $x \in \mathcal{X}(z)$, we have

$$\check{\mu}_{h,\psi}(x,a) = \frac{\mu_h(z,a)}{|\mathcal{X}(z)|} = \mu_h(z,a) \cdot \frac{|\phi|}{|\mathcal{X}(\phi)|}.$$

Additionally, according to our construction of $\check{T}^{(1)}_{\psi}(x' \mid x,a)$, for any $x \in \mathcal{X}(z)$ and $x' \in \mathcal{X}(z')$, we have

$$\check{T}^{(1)}_{\psi}(x' \mid x,a) = \frac{T^{(1)}(z' \mid z,a)}{|\mathcal{X}(\phi')|} = T^{(1)}(z' \mid z,a) \cdot \frac{|\phi'|}{|\mathcal{X}(\phi')|}.$$

This implies that for any $\phi \in \Phi_h, \phi' \in \Phi'_{h+1}$, and $x \in \mathcal{X}_{\psi_1}(z_1) \cap \mathcal{X}_{\psi_2}(z_2), x' \in \mathcal{X}_{\psi_1}(z'_1) \cap \mathcal{X}_{\psi_2}(z'_2)$, we have

$$\begin{aligned}
&\check{\mu}_{h,\psi_1}(x,a)\check{\mu}_{h,\psi_2}(x,a)\check{T}^{(1)}_{\psi_1}(x' \mid x,a)\check{T}^{(1)}_{\psi_2}(x' \mid x,a) \\
&\quad = \frac{|\phi|^2}{|\mathcal{X}(\phi)|^2} \cdot \frac{|\phi'|^2}{|\mathcal{X}(\phi')|^2} \cdot \mu_h(z_1,a)\mu_h(z_2,a)T^{(1)}(z_1 \mid z_1,a)T^{(1)}(z_2 \mid z_2,a). \qquad (37)
\end{aligned}$$

And the definition of θ in (34) gives that

$$|\mathcal{X}_{\psi_1}(z_1) \cap \mathcal{X}_{\psi_2}(z_2)| = \theta_h(\phi, z_1, z_2; \psi_1, \psi_2) \cdot |\mathcal{X}(\phi)|$$
$$|\mathcal{X}_{\psi_1}(z_1') \cap \mathcal{X}_{\psi_2}(z_2')| = \theta_{h+1}(\phi', z_1', z_2'; \psi_1, \psi_2) \cdot |\mathcal{X}(\phi)|$$

Hence we can write

$$\mathbb{E}_{(x,a,x')\sim\nu_h}\left[\frac{\check{\mu}_{h,\psi_1}(x,a)\check{\mu}_{h,\psi_2}(x,a)\check{T}^{(1)}_{\psi_1}(x' \mid x,a)\check{T}^{(1)}_{\psi_2}(x' \mid x,a)}{\nu_h(x,a,x')^2}\right]$$

$$= \sum_{x\in\mathcal{X}_h}\sum_{a\in\mathcal{A}}\sum_{x'\in\mathcal{X}_{h+1}\cup\{u_{h+1},v_{h+1}\}} \frac{\check{\mu}_{h,\psi_1}(x,a)\check{\mu}_{h,\psi_2}(x,a)\check{T}^{(1)}_{\psi_1}(x' \mid x,a)\check{T}^{(1)}_{\psi_2}(x' \mid x,a)}{\nu_h(x,a,x')}$$

$$\overset{(i)}{=} \sum_{\phi\in\Phi_h}\sum_{a\in\mathcal{A}}\sum_{\phi'\in\Phi_{h+1}\cup\{u_{h+1},v_{h+1}\}}\sum_{x\in\mathcal{X}(\phi)}\sum_{x'\in\mathcal{X}(\phi')} \frac{\check{\mu}_{h,\psi_1}(x,a)\check{\mu}_{h,\psi_2}(x,a)\check{T}^{(1)}_{\psi_1}(x' \mid x,a)\check{T}^{(1)}_{\psi_2}(x' \mid x,a)}{\sum_{z\in\phi}\sum_{z'\in\phi'}\frac{\mu_h(z,a)T^{(1)}(z'|z,a)}{|\mathcal{X}(\phi)||\mathcal{X}(\phi')|}}$$

$$\overset{(ii)}{=} \sum_{\phi\in\Phi_h}\sum_{a\in\mathcal{A}}\sum_{\phi'\in\Phi_{h+1}\cup\{u_{h+1},v_{h+1}\}}\sum_{z_1\in\phi}\sum_{z_1'\in\phi'}\sum_{z_2\in\phi}\sum_{z_2'\in\phi'} \mu_h(z_1,a)\mu_h(z_2,a)T^{(1)}(z_1 \mid z_1,a)T^{(1)}(z_2 \mid z_2,a)$$
$$\cdot \frac{\frac{|\phi|^2}{|\mathcal{X}(\phi)|^2}\cdot\frac{|\phi'|^2}{|\mathcal{X}(\phi')|^2}\cdot|\mathcal{X}_{\psi_1}(z_1')\cap\mathcal{X}_{\psi_2}(z_2')||\mathcal{X}_{\psi_1}(z_1)\cap\mathcal{X}_{\psi_2}(z_2)|}{\sum_{z\in\phi}\sum_{z'\in\phi'}\frac{\mu_h(z,a)T^{(1)}(z'|z,a)}{|\mathcal{X}(\phi)||\mathcal{X}(\phi')|}}$$

$$\overset{(iii)}{=} \sum_{\phi\in\Phi_h}\sum_{a\in\mathcal{A}}\sum_{\phi'\in\Phi_{h+1}\cup\{u_{h+1},v_{h+1}\}}\sum_{z_1\in\phi}\sum_{z_1'\in\phi'}\sum_{z_2\in\phi}\sum_{z_2'\in\phi'} \mu_h(z_1,a)\mu_h(z_2,a)T^{(1)}(z_1 \mid z_1,a)T^{(1)}(z_2 \mid z_2,a)$$
$$\cdot \frac{\frac{|\phi|^2}{|\mathcal{X}(\phi)|^2}\cdot\frac{|\phi'|^2}{|\mathcal{X}(\phi')|^2}\cdot\theta_h(\phi,z_1,z_2;\psi_1,\psi_2)\cdot\theta_{h+1}(\phi',z_1',z_2';\psi_1,\psi_2)\cdot|\mathcal{X}(\phi)|}{\sum_{z\in\phi}\sum_{z'\in\phi'}\frac{\mu_h(z,a)T^{(1)}(z'|z,a)}{|\mathcal{X}(\phi)||\mathcal{X}(\phi')|}}$$

$$\overset{(iv)}{=} \sum_{\phi\in\Phi_h}\sum_{a\in\mathcal{A}}\sum_{\phi'\in\Phi_{h+1}\cup\{u_{h+1},v_{h+1}\}}\sum_{z_1\in\phi}\sum_{z_1'\in\phi'}\sum_{z_2\in\phi}\sum_{z_2'\in\phi'} |\phi|^2|\phi'|^2\cdot\mu_h(z_1,a)\mu_h(z_2,a)T^{(1)}(z_1 \mid z_1,a)T^{(1)}(z_2 \mid z_2,a)$$
$$\frac{\theta_h(\phi,z_1,z_2;\psi_1,\psi_2)\cdot\theta_{h+1}(\phi',z_1',z_2';\psi_1,\psi_2)}{\sum_{z\in\phi}\sum_{z'\in\phi'}\mu_h(z,a)T^{(1)}(z' \mid z,a)}, \tag{38}$$

where in (i) we use the exact form of ν_h defined in (25), in (ii) we group those $x \in \mathcal{X}_{\psi_1}(z_1) \cap \mathcal{X}_{\psi_2}(z_2)$ and $x' \in \mathcal{X}_{\psi_1}(z_1') \cap \mathcal{X}_{\psi_2}(z_2')$ together (because the summand gives the same value) and use (37), in (iii) we use the definition of $\theta_h(\phi, z_1, z_2; \psi_1, \psi_2)$ in (34), and (iv) is just algebraic calculation.

Next, when $(\psi_1, \psi_2) \in \Gamma(\xi)$, according to the definition of $\gamma(\xi)$ in (36), we have

$$(\psi_1,\psi_2) \in \Gamma_h(\xi;\phi,z_1,z_2) \quad \forall h \in [H-1], \phi \in \Phi_h, z_1, z_2 \in \psi,$$

which implies that

$$\theta_h(\phi,z_1,z_2;\psi_1,\psi_2) \le \frac{1+\xi}{|\phi|^2} \quad \text{and} \quad \theta_{h+1}(\phi',z_1',z_2';\psi_1,\psi_2) \le \frac{1+\xi}{|\phi'|^2}. \tag{39}$$

Bringing this back to (38), we obtain that

$$\mathbb{E}_{(x,a,x')\sim\nu_h}\left[\frac{\check{\mu}_{h,\psi_1}(x,a)\check{\mu}_{h,\psi_2}(x,a)\check{T}^{(1)}_{\psi_1}(x' \mid x,a)\check{T}^{(1)}_{\psi_2}(x' \mid x,a)}{\nu_h(x,a,x')^2}\right]$$

$$\overset{(i)}{\le} (1+\xi)^2\cdot\sum_{\phi\in\Phi_h}\sum_{a\in\mathcal{A}}\sum_{\phi'\in\Phi_{h+1}\cup\{u_{h+1},v_{h+1}\}}\sum_{z_1\in\phi}\sum_{z_1'\in\phi'}\sum_{z_2\in\phi}\sum_{z_2'\in\phi'}\frac{\mu_h(z_1,a)\mu_h(z_2,a)T^{(1)}(z_1' \mid z_1,a)T^{(1)}(z_2' \mid z_2,a)}{\sum_{z\in\phi}\sum_{z'\in\phi'}\mu_h(z,a)T^{(1)}(z' \mid z,a)}$$

$$\overset{(ii)}{=} (1+\xi)^2 \cdot \sum_{\phi\in\Phi_h}\sum_{a\in\mathcal{A}}\sum_{\phi'\in\Phi_{h+1}\cup\{u_{h+1},v_{h+1}\}}\sum_{z_1\in\phi}\sum_{z_1'\in\phi'}\mu_h(z_1,a)T^{(1)}(z_1'\mid z_1,a)$$

$$\overset{(iii)}{=} (1+\xi)^2\sum_{z\in\mathcal{Z}_h}\sum_{a\in\mathcal{A}}\sum_{z'\in\mathcal{Z}'_{h+1}}\mu_h(z_1,a)T^{(1)}(z'\mid z,a) = (1+\xi)^2,$$

where in (i) we use (39), (ii) is just algebraic calculation, and in (iii) we use the fact that $\mu_h(\cdot)\in\Delta(\mathcal{Z}_h\times\mathcal{A})$ and $T^{(1)}(\cdot\mid z,a)\in\Delta(\mathcal{Z}_{h+1})$ for any $z\in\mathcal{Z}, a\in\mathcal{A}$. ■

Lemma 13. *For any $\psi_1,\psi_2\notin\Gamma(\xi)$, the term $\mathfrak{B}_h(\psi_1,\psi_2)$ satisfies*

$$\mathfrak{B}_h(\psi_1,\psi_2)\le\sup_{\phi\in\Phi}|\phi|^4.$$

Proof of Lemma 13. For $(\psi_1,\psi_2)\notin\Gamma(\xi)$, since for any $h\in[H-1]$, $\phi\in\Phi_h\cup\{u_h,v_h\}$ and $z_1,z_2\in\phi$, we always have

$$\mathcal{X}_{\psi_1}(z_1)\cap\mathcal{X}_{\psi_2}(z_2)\subset\mathcal{X}(\phi),$$

we have $\theta_h(\phi,z_1,z_2;\psi_1,\psi_2)\le 1$. Bringing this back to (38), we obtain that

$$\begin{aligned}
\mathfrak{B}_h(\psi_1,\psi_2) &= \mathbb{E}_{(x,a,x')\sim\nu_h}\left[\frac{\check{\mu}_{h,\psi_1}(x,a)\check{\mu}_{h,\psi_2}(x,a)\check{T}^{(1)}_{\psi_1}(x'\mid x,a)\check{T}^{(1)}_{\psi_2}(x'\mid x,a)}{\nu_h(x,a,x')^2}\right]\\
&\le \sup_{\phi\in\Phi_h}|\phi|^2\sup_{\phi'\in\Phi'\cup\{u_{h+1},v_{h+1}\}}|\phi'|^2\\
&\quad\cdot\sum_{\phi\in\Phi_h}\sum_{a\in\mathcal{A}}\sum_{\phi'\in\Phi_{h+1}\cup\{u_{h+1},v_{h+1}\}}\sum_{z_1\in\phi}\sum_{z_1'\in\phi'}\sum_{z_2\in\phi}\sum_{z_2'\in\phi'}\frac{\mu_h(z_1,a)\mu_h(z_2,a)T^{(1)}(z_1'\mid z_1,a)T^{(1)}(z_2'\mid z_2,a)}{\sum_{z\in\phi}\sum_{z'\in\phi'}\mu_h(z,a)T^{(1)}(z'\mid z,a)}\\
&= \sup_{\phi\in\Phi_h}|\phi|^2\sup_{\phi'\in\Phi'\cup\{u_{h+1},v_{h+1}\}}|\phi'|^2\cdot\sum_{\phi\in\Phi_h}\sum_{a\in\mathcal{A}}\sum_{\phi'\in\Phi_{h+1}\cup\{u_{h+1},v_{h+1}\}}\sum_{z_1\in\phi}\sum_{z_1'\in\phi'}\mu_h(z_1,a)T^{(1)}(z_1'\mid z_1,a)\\
&= \sup_{\phi\in\Phi_h}|\phi|^2\sup_{\phi'\in\Phi'\cup\{u_{h+1},v_{h+1}\}}|\phi'|^2\sum_{z\in\mathcal{Z}_h}\sum_{a\in\mathcal{A}}\sum_{z'\in\mathcal{Z}'_{h+1}}\mu_h(z_1,a)T^{(1)}(z_1'\mid z_1,a)\\
&= \sup_{\phi\in\Phi_h}|\phi|^2\sup_{\phi'\in\Phi'\cup\{u_{h+1},v_{h+1}\}}|\phi'|^2.
\end{aligned}$$

Further notice that for any $\phi'\in\{u_{h+1},v_{h+1}\}$, we have $|\phi'|\le 1\le\sup_{\phi'\in\Phi'}|\phi'|$, which indicates that $\mathfrak{B}_h(\psi_1,\psi_2)\le\sup_{\phi\in\Phi_h}|\phi|^2\cdot\sup_{\phi'\in\Phi_h}|\phi'|^2\le\sup_{\phi\in\Phi}|\phi|^4$. ■

Lemma 14. *For any $h\in[H-1]$, $\phi\in\Phi_h$ and $z_1,z_2\in\Phi$ and $\xi\in(0,1)$, we have*

$$\frac{1}{|\Psi|^2}\sum_{\psi_1,\psi_2\in\Psi}\mathbb{I}\{(\psi_1,\psi_2)\notin\Gamma_h(\xi;\phi,z_1,z_2)\}\le e^{-2\xi^2\frac{|\mathcal{X}(\phi)|}{|\phi|^3}},$$

where $\Gamma_h(\xi;\phi,z_1,z_2)$ is defined in (35)

Proof of Lemma 14. We denote $S=|\mathcal{X}(\phi)|$. Without loss of generality, we assume $\mathcal{X}(\phi)=[S]=\{1,2,\cdots,S\}$. Since ψ_1 and ψ_2 are samples i.i.d. according to $\mathrm{Unif}(\Psi)$, without loss of generality we assume $\mathcal{X}_{\psi_1}(z_1)=[S/|\phi|]$. And to prove this lemma, we only need to verify that when $\psi_2\sim\mathrm{Unif}(\Psi)$, we have

$$\mathbb{P}\left(\left|\mathcal{X}_{\psi_2}(z_2)\cap\left[\frac{S}{|\phi|}\right]\right|\ge\frac{S(1+\xi)}{|\phi|^2}\right)\le e^{-2\xi^2\frac{S}{|\phi|^3}}. \tag{40}$$

Next, we notice that sampling $\psi_2 \sim \mathrm{Unif}(\Psi)$ is equivalent of sampling $\mathcal{X}_{\psi_2}(z_2)$ uniformly from all subsets of $[S]$ with exact $S/|\phi|$ elements. Hence we obtain that

$$\mathbb{P}\left(\left|\mathcal{X}_{\psi_2}(z_2) \cap \left[\frac{S}{|\phi|}\right]\right| \geq \frac{S(1+\xi)}{|\phi|^2}\right) = \sum_{t \geq S(1+\xi)/|\phi|^2} \frac{\binom{S/|\phi|}{t}\binom{S-S/|\phi|}{S/|\phi|-t}}{\binom{S}{S/|\phi|}}.$$

We further notice that according to Lemma D.7 in Foster et al. (2022) (also in Hoeffding (1994)), we get

$$\sum_{t \geq S(1+\xi)/|\phi|^2} \frac{\binom{S/|\phi|}{t}\binom{S-S/|\phi|}{S/|\phi|-t}}{\binom{S}{S/|\phi|}} = \mathbb{P}\left[X \geq \left(\frac{K}{N} + \frac{\xi}{|\phi|}\right)N'\right] \leq e^{-2\xi^2 \frac{S}{|\phi|^3}},$$

where $X \sim \mathrm{Hypergeometric}(K, N, N')$ with $K = N' = S/|\phi|, N = S$. This verifies (40). ■

Lemma 15 (Lemma D.7 in Foster et al. (2022))**.** *Let* $X \sim \mathrm{Hypergeometric}(K, N, N')$ *and define* $p = K/N$. *Then for any* $0 < \varepsilon < pN'$, *we have*

$$\mathbb{P}[X \geq (p+\varepsilon)N'] \leq \exp(-2\varepsilon^2 N').$$

C.5.2 Proof of Theorem C.1

Proof of Theorem C.1. Given the original MDP M and distribution μ, we construct family $\mathfrak{G}$ of OPE problems in Appendix C.4.2. First of all, Lemma 8 indicates that function class $\mathcal{F}$ of these OPE problems in $\mathfrak{G}$ is 2. Next, bringing Lemma 11 into Lemma 10 and noticing that (21) satisfies the condition of Lemma 11, we have that any algorithm which takes $D_{h,n}$ for $h \in [H-1]$ must induce error at least $\varepsilon/8H$ in one case within $\mathfrak{G}$.

Additionally, Corollary 3 indicates that the standard concentrability of instances in $\mathfrak{G}$ equals to the standard concentrability of $\mathsf{C}(M^{(1)}, \mu')$ or $\mathsf{C}(M^{(2)}, \mu')$. Additionally, Corollary 1 indicates that $\mathsf{C}(M^{(1)}, \mu')$ or $\mathsf{C}(M^{(2)}, \mu')$ is no more than $6\mathsf{C}(M, \mu, \pi_e)$. This verifies the second condition in Theorem C.1. ■

D Missing Details from Section 3.1 and Section 4.1

In this subsection, we will prove Proposition 1 and Theorem 4.1.

D.1 Proof of Proposition 1

The following lemmas are some properties of the MDP defined in Example 1. Proposition 1 is a direct corollary of Lemma 19.

Lemma 16. *For* MDP M, *and policy* π_b *defined above, we have*

$$d_h^{\pi_b}(z_h^{[1]}; M) \geq \frac{1}{2^{h+2}}, \quad d_h^{\pi_b}(z_h^{[2]}; M) \geq \frac{1}{2^{h+2}H} \quad \text{and} \quad d_h^{\pi_b}(z_h^{[3]}; M) \geq \frac{1}{4}, \quad \forall 1 \leq h \leq H.$$

Proof of Lemma 16. Under policy $\pi_b(z) = \frac{1}{H^2}\delta_{\mathfrak{a}_1}(\cdot) + \frac{H^2-1}{H^2}\delta_{\mathfrak{a}_2}(\cdot)$, the transition satisfies

$$T(z_{h+1}^{[1]} \mid z_h^{[1]}, \pi_b(z_h^{[1]})) \geq \frac{H^2-1}{H^2} T(z_{h+1}^{[1]} \mid z_h^{[1]}, \mathfrak{a}_2) = \frac{H^2-1}{2H^2},$$

$$T(z_{h+1}^{[2]} \mid z_h^{[2]}, \pi_b(z_h^{[2]})) \geq \frac{H^2-1}{H^2} T(z_{h+1}^{[2]} \mid z_h^{[2]}, \mathfrak{a}_2) = \frac{H^2-1}{2H^2},$$
$$T(z_{h+1}^{[3]} \mid z_h^{[3]}, \pi_b(z_h^{[3]})) \geq \frac{H^2-1}{H^2} T(z_{h+1}^{[3]} \mid z_h^{[3]}, \mathfrak{a}_2) = \frac{H^2-1}{H^2}$$

Hence according to the choice initial distribution $\rho(\cdot)$, we have for all $1 \leq h \leq H$,

$$d_h^{\pi_b}(z_h^{[1]}; M) \geq \frac{H-1}{2H} \cdot \left(\frac{H^2-1}{2H^2}\right)^{h-1} \geq \frac{1}{2^{h+2}}, \quad d_h^{\pi_b}(z_h^{[2]}; M) = \frac{1}{2H} \cdot \left(\frac{H^2-1}{2H^2}\right)^{h-1} \geq \frac{1}{2^{h+2}H}$$
$$\text{and} \quad d_h^{\pi_b}(z_h^{[3]}; M) \geq \frac{1}{2}\left(\frac{H^2-1}{H^2}\right)^{h-1} \geq \frac{1}{4}.$$

■

Lemma 17. *For* MDP M *and policy* π_b *defined above, we have*

$$\frac{d_h^{\pi_b}(z_h^{[1]})}{d_h^{\pi_b}(z_h^{[2]})} \geq \frac{H-1}{3}.$$

Proof of Lemma 17. First we can write the dynamic programming formula for d_h^{π}:

$$d_{h+1}^{\pi_b}(z_{h+1}^{[1]}) = d_h^{\pi_b}(z_h^{[1]}) T(z_{h+1}^{[1]} \mid z_h^{[1]}, \pi_b(z_h^{[1]})) + d_h^{\pi_b}(z_h^{[2]}) T(z_{h+1}^{[1]} \mid z_h^{[2]}, \pi_b(z_h^{[2]})),$$
$$d_{h+1}^{\pi_b}(z_{h+1}^{[2]}) = d_h^{\pi_b}(z_h^{[1]}) T(z_{h+1}^{[2]} \mid z_h^{[1]}, \pi_b(z_h^{[1]})) + d_h^{\pi_b}(z_h^{[2]}) T(z_{h+1}^{[2]} \mid z_h^{[2]}, \pi_b(z_h^{[2]})).$$

According to our choice of π_b, we have

$$T(z_{h+1}^{[1]} \mid z_h^{[1]}, \pi_b(z_h^{[1]})) = \frac{H^2-1}{2H^2}, \quad T(z_{h+1}^{[2]} \mid z_h^{[1]}, \pi_b(z_h^{[1]})) = \frac{1}{H^2},$$
$$T(z_{h+1}^{[1]} \mid z_h^{[2]}, \pi_b(z_h^{[2]})) = 0, \quad T(z_{h+1}^{[2]} \mid z_h^{[2]}, \pi_b(z_h^{[2]})) = \frac{H^2-1}{2H^2},$$

which indicates that

$$\frac{d_{h+1}^{\pi_b}(z_{h+1}^{[2]})}{d_{h+1}^{\pi_b}(z_{h+1}^{[1]})} = \frac{T(z_{h+1}^{[2]} \mid z_h^{[1]}, \pi_b(z_h^{[1]}))}{T(z_{h+1}^{[1]} \mid z_h^{[1]}, \pi_b(z_h^{[1]}))} + \frac{T(z_{h+1}^{[2]} \mid z_h^{[2]}, \pi_b(z_h^{[2]}))}{T(z_{h+1}^{[1]} \mid z_h^{[1]}, \pi_b(z_h^{[1]}))} \cdot \frac{d_h^{\pi_b}(z_h^{[2]})}{d_h^{\pi_b}(z_h^{[1]})}$$
$$= \frac{2}{H^2-1} + \frac{d_h^{\pi_b}(z_h^{[2]})}{d_h^{\pi_b}(z_h^{[1]})} \leq \frac{2}{H^2-1} + \frac{d_h^{\pi_b}(z_h^{[2]})}{d_h^{\pi_b}(z_h^{[1]})}.$$

Additionally, after noticing that $\frac{d_h^{\pi_b}(z_1^{[2]})}{d_h^{\pi_b}(z_1^{[1]})} = \frac{\rho(z_1^{[2]})}{\rho(z_1^{[1]})} = \frac{1}{H-1}$, we have for all $1 \leq h \leq H$,

$$\frac{d_h^{\pi_b}(z_{h+1}^{[2]})}{d_h^{\pi_b}(z_{h+1}^{[1]})} \leq \frac{1}{H-1} + (h-1) \cdot \frac{2}{H^2-1} \leq \frac{1}{H-1} + \frac{2(H-1)}{H^2-1} \leq \frac{3}{H-1}.$$

■

Lemma 18. *For any policy* π*, and the* MDP M *defined above, we have*

$$d_h^{\pi}(z_h^{[1]}) + d_h^{\pi}(z_h^{[2]}) \leq \frac{1}{2^{h-1}} \quad \forall h \in [H].$$

Proof of Lemma 18. First we can write the dynamic programming formula for d_h^π:

$$d_{h+1}^\pi(z_{h+1}^{[1]}) = d_h^\pi(z_h^{[1]})T(z_{h+1}^{[1]} \mid z_h^{[1]}, \pi(z_h^{[1]})) + d_h^\pi(z_h^{[2]})T(z_{h+1}^{[1]} \mid z_h^{[2]}, \pi(z_h^{[2]}))$$
$$d_{h+1}^\pi(z_{h+1}^{[2]}) = d_h^\pi(z_h^{[1]})T(z_{h+1}^{[2]} \mid z_h^{[1]}, \pi(z_h^{[1]})) + d_h^\pi(z_h^{[2]})T(z_{h+1}^{[2]} \mid z_h^{[2]}, \pi(z_h^{[2]})).$$

We let $\pi(z_h^{[1]}) = p_1\delta_{a_1}(\cdot) + (1-p_1)\delta_{a_2}(\cdot)$, and $\pi(z_h^{[2]}) = p_2\delta_{a_1}(\cdot) + (1-p_2)\delta_{a_2}(\cdot)$. Then we have

$$T(z_{h+1}^{[1]} \mid z_h^{[1]}, \pi(z_h^{[1]})) = \frac{1}{2}(1-p_1), \quad T(z_{h+1}^{[2]} \mid z_h^{[1]}, \pi(z_h^{[1]})) = p_1$$
$$T(z_{h+1}^{[1]} \mid z_h^{[2]}, \pi(z_h^{[2]})) = 0, \quad T(z_{h+1}^{[2]} \mid z_h^{[2]}, \pi(z_h^{[2]})) = \frac{1}{2}(1-p_2).$$

This implies that

$$\begin{aligned} 2d_{h+1}^\pi(z_{h+1}^{[1]}) + d_{h+1}^\pi(z_{h+1}^{[2]}) &= d_h^\pi(z_h^{[1]}) \cdot (1-p_1p_1) + d_h^\pi(z_h^{[1]}) \cdot \left(2 \cdot 0 + \frac{1}{2}(1-p_2)\right) \\ &= \frac{1}{2} \cdot \left(2d^\pi(z_h^{[1]}) + d^\pi(z_h^{[2]})\right). \end{aligned}$$

Further noticing that $2d^\pi(z_1^{[1]}) + d^\pi(z_1^{[2]}) = 2\rho(z_1^{[1]}) + \rho(z_1^{[2]}) \le 1$, we obtain that for any $h \in [H]$,

$$2d_h^\pi(z_h^{[1]}) + d_h^\pi(z_h^{[2]}) \le \frac{1}{2^{h-1}}.$$

Therefore, for any $h \in [H]$, we have

$$d_h^\pi(z_h^{[1]}) + d_h^\pi(z_h^{[2]}) \le \frac{1}{2^{h-1}}.$$

■

Lemma 19. *For the* MDP M *and policy* π_b *defined above, the concentrability coefficient of all policies with respect to* $d_h^{\pi_b}(\cdot; M)$ *is upper bounded as*

$$\max_{\pi \in \Pi} \max_h \max_{z \in \mathcal{Z}_h, a \in \mathcal{A}} \frac{d_h^\pi(z, a; M)}{d_h^{\pi_b}(z, a; M)} \le 8H^3,$$

where Π *is the class of all policies. However, for* $\varepsilon \le 1/15$*, the aggregated concentrability coefficient is lower bounded as*

$$\bar{\mathsf{C}}_\varepsilon(M, \Phi, d_h^{\pi_b}(\cdot; M)) \ge 2^{H-7}.$$

Proof of Lemma 19. For our choice of π_b,

$$d_h^{\pi_b}(z, a; M) \ge \frac{d_h^{\pi_b}(z; M)}{H^2} \quad \text{and} \quad d_h^\pi(z, a; M) \le \frac{d_h^\pi(z; M)}{H^2} \quad \forall z \in \mathcal{Z}, a \in \mathcal{A},$$

which implies that

$$\frac{d_h^\pi(z, a; M)}{d_h^{\pi_b}(z, a; M)} \le H^2 \cdot \frac{d_h^\pi(z; M)}{d_h^{\pi_b}(z; M)} \qquad \forall z \in \mathcal{Z}.$$

Hence, Lemma 16 and Lemma 18 give that for any policy $\pi \in \Pi, 1 \le h \le H$ and $z \in \mathcal{Z}_h$,

$$\frac{d_h^\pi(z; M)}{d_h^{\pi_b}(z; M)} \le \max\left\{4, \frac{1/2^{h-1}}{1/(2^{h+2}H)}\right\} = 8H.$$

This implies that

$$\max_{\pi\in\Pi}\max_{h}\max_{z\in\mathcal{Z}_h,a\in\mathcal{A}}\frac{d_h^{\pi}(z,a;M)}{d_h^{\pi_b}(z,a;M)}\le 8H^3.$$

As for the lower bound of $\bar{\mathsf{C}}_\varepsilon(M,\Phi,d^{\pi_b}(\cdot;M))$, first we notice that according to (3), our choice of $\pi_{\mathrm{e}}(z)=\delta_{\mathfrak{a}_1}(\cdot)$ and Lemma 17, we have

$$\begin{aligned}\bar{T}(\phi_{h+1}^{[1]}\mid\phi_h^{[1]},\pi_{\mathrm{e}})&=\frac{\sum_{z_h\in\phi_h^{[1]},z'_{h+1}\in\phi_{h+1}^{[1]}}d_h^{\pi_b}(z_h,\mathfrak{a}_1;M)T(z'_{h+1}\mid z_h,\mathfrak{a}_1)}{\sum_{z\in\phi_h^{[1]}}d_h^{\pi_b}(z,\mathfrak{a}_1;M)}\\&\ge\frac{\sum_{z'_{h+1}\in\phi_{h+1}^{[1]}}d_h^{\pi_b}(z_h,\mathfrak{a}_1;M)T(z'_{h+1}\mid z_h^{[1]},\mathfrak{a}_1)}{\sum_{z\in\phi_h^{[1]}}d_h^{\pi_b}(z_h,\mathfrak{a}_1;M)}\\&=\frac{d_h^{\pi_b}(z_h,\mathfrak{a}_1;M)}{\sum_{z\in\phi_h^{[1]}}d_h^{\pi_b}(z_h,\mathfrak{a}_1;M)}=\frac{H-1}{H-1+3}=\frac{H-1}{H+2}.\end{aligned}$$

This implies that $\bar{d}_H^{\pi_{\mathrm{e}}}(\phi_H^{[1]};M)$ satisfies

$$\bar{d}_H^{\pi_{\mathrm{e}}}(\phi_H^{[1]};M)\ge\bar{d}_1^{\pi_{\mathrm{e}}}(\phi_1^{[1]};M)\prod_{h=1}^{H-1}\bar{T}(\phi_{h+1}^{[1]}\mid\phi_h^{[1]},\pi)\ge\frac{1}{2}\left(\frac{H-1}{H+2}\right)^{H-1}\ge\frac{1}{2e^3}\ge\frac{1}{41}.$$

Hence when $\varepsilon\le 1/15$, we have

$$\begin{aligned}\bar{\mathsf{C}}_\varepsilon(M,\Phi,\mu)&\ge\frac{\bar{d}_H^{\pi_{\mathrm{e}}}(\phi_H^{[1]};M)}{\sum_{z\in\phi_H^{[1]}}d_H^{\pi_b}(z,\pi_{\mathrm{e}}(z);M)}=\frac{\bar{d}_H^{\pi_{\mathrm{e}}}(\phi_H^{[1]};M)}{d_H^{\pi_b}(z_H^{[1]},\pi_{\mathrm{e}}(z_H^{[1]});M)+d_h^{\pi_b}(z_H^{[2]},\pi_{\mathrm{e}}(z_H^{[2]});M)}\\&\ge\frac{1/41}{1/(2^{H-1})}\ge 2^{H-7},\end{aligned}$$

where in the second inequality we adopt Lemma 18 with $\pi=\pi_b$. ∎

D.2 Proof of Theorem 4.1

In this section, we will prove the following stronger results that considers (Q,W)-realizability.

Theorem D.1. *For any $\varepsilon\le 1/15$ There exists a class $\mathcal{M}$ of H-layer MDPs with shared state spaces and action spaces, and two policies π_{e} and π_b, such that for any $N=o(H2^H/\varepsilon)$,*

- *The concentrability coefficient of all policies with respect to $d_h^{\pi_b}$ is upper bounded by $384H^3$;*
- *The (Q, W)-function class*

$$\{(Q^{\pi_{\mathrm{e}}}(\cdot,\cdot\,;M),W(\cdot;d^{\pi_b}(\cdot;M);M)):M\in\mathcal{M}\}$$

  has only two elements;

- *For size N admissible data $\mathcal{D}(M)$ with respect to policy π_{e} from* MDP *$M\in\mathcal{M}$: $\mathcal{D}(M)=\cup_{h=1}^H\mathcal{D}_h(M)$ where*

$$\mathcal{D}_h(M)=\{(x_h^i,a_h^i,r_h^i,x_{h+1}^i)_{i=1}^N:(x_h^i,a_h^i)\sim d_h^{\pi_b}(\cdot),r_h^i\sim R(\cdot\mid x,a;M),x_{h+1}^i\sim T(\cdot\mid x,a;M)\},$$

  any algorithm which takes $\mathcal{D}(M)$ as input and output the evaluation of value function $V^{\pi_{\mathrm{e}}}(\rho;M)$ cannot achieve ε/H-accurate for all $M\in\mathcal{M}$ with probability at least $1/2$.

Proof of Theorem D.1. We consider the class of offline RL problems constructed in Appendix C.4.2: $\mathfrak{G} = \bigcup_{\psi\in\Psi}\{\mathfrak{g}_\psi^{[1]}, \mathfrak{g}_\psi^{[2]}\}$. According to the construction, the sampling distributions over rich observations in $\mathfrak{g}_\psi^{[1]}$ or $\mathfrak{g}_\psi^{[2]}$ are $\check{\mu}_\psi(x,a) = \mu(\psi(z),a)/|\mathcal{X}_\psi(z)|$. And Theorem C.1 indicates that for $\varepsilon \le 1/41$, for any algorithm using less than

$$\frac{H\bar{\mathsf{C}}_\varepsilon(M,\phi,\mu)}{8\varepsilon} \ge 2^{H-10}\frac{H}{\varepsilon}$$

samples, there must exist some $\psi \in \Phi$ such that if the samples are according to $\check{\mu}_\psi$, the algorithm will have error greater than $\varepsilon/8H$ in $\mathfrak{g}_\psi^{[1]}$ or $\mathfrak{g}_\psi^{[2]}$ with probability at least $1/2$.

According to Lemma 4 (d), we have $d^{\pi_b}(z; M^{[1]}) \le d^{\pi_b}(z; M)$ for any $z \in \mathcal{Z}$. Hence we obtain that for any $x \in \mathcal{X}\backslash\{u,v,w\}_{h=1}^H$,

$$d^{\check{\pi}_b}(x; \check{M}_\psi^{[1]}) = \frac{d^{\pi_b}(\psi(x); M)}{|\mathcal{X}_\psi(z)|} \le \frac{d^{\pi_b}(\psi(x); M)}{|\mathcal{X}_\psi(z)|}.$$

This also implies

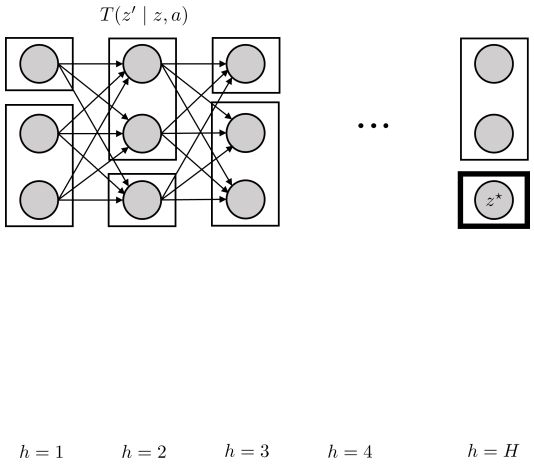
$$d^{\check{\pi}_b}(x,a; \check{M}_\psi^{[1]}) = d^{\check{\pi}_b}(x; \check{M}_\psi^{[1]})\pi_b(a \mid x) \le \frac{d^{\pi_b}(\psi(x); M)\pi_b(a \mid x)}{|\mathcal{X}_\psi(z)|} = \frac{d^{\pi_b}(\psi(x), a; M)}{|\mathcal{X}_\psi(z)|} = \check{\mu}_\psi(x,a).$$

Similarly, we can also obtain that

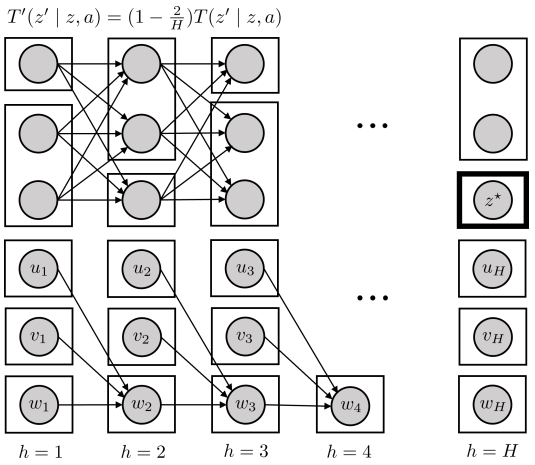
$$d^{\check{\pi}_b}(x,a; \check{M}_\psi^{[2]}) \le \check{\mu}_\psi(x,a).$$

Hence, for any algorithm using less than $2^{H-10}H^3/\varepsilon$ number of samples, there must exists some $\check{M} \in \{\check{M}_\psi^{[1]} : \psi \in \Psi\} \cup \{\check{M}_\psi^{[2]} : \psi \in \Psi\}$ such that if the samples are according to $d^{\check{\pi}_b}(\cdot \mid \check{M})$, the estimation error is at least ε/H.

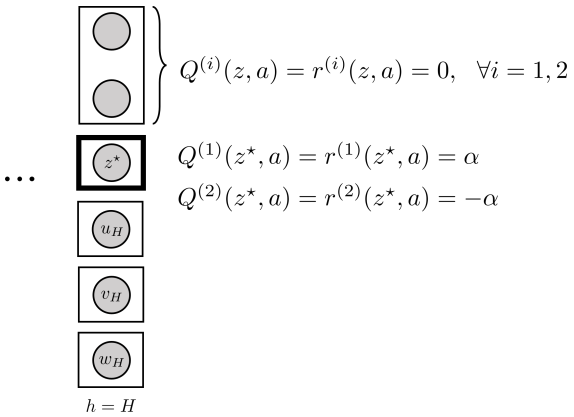
Finally, Lemma 19 indicates that the concentrability coefficient of M is bounded by $8H^3$. Hence according to Corollary 2 and Lemma 7 we have for any such $\check{M}$, the concentrability coefficient is upper bounded by $384H^3$. And similar to the proof of Theorem C.1, we can also show that the class of (Q,W)-functions contains only two items. ■



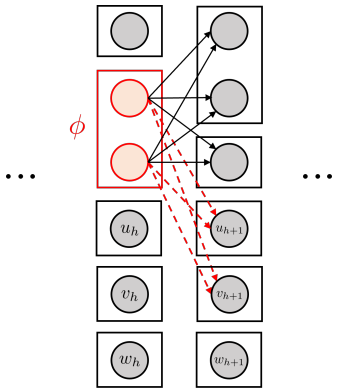
(a) Given Markov Transition Model M and aggregation scheme Φ.



(b) Augmented Markov Transition Model M' and aggregation scheme Φ'.



(c) Determine the reward and value functions in the layer H.



(d) Determine the value functions inductively assuming $Q^{(i)}$ are already determined on all layers $h' \geq h + 1$.

Figure 4: Lower bound construction used in the proof sketch of Theorem 3.1. States are represented with circles and the corresponding state aggregations are represented with rectangles. We use the bold rectangle to denote the set of aggregations $\mathcal{I}$ that attains the maximum in the definition of $\bar{\mathsf{C}}_\varepsilon(M, \Phi, \mu)$ (see Definition 3.1). For simplicity, in the above figure $\mathcal{I}$ only contains a single aggregation that contains a single latent state $z^\star$, while in general $\mathcal{I}$ may include multiple aggregations each with multiple latent states.

E Missing Details from Section 4.2

In this section, we will provide missing details from Section 4.2. In Section E.1 we present algorithms needed in Section 4.2. Finally, in Section 4.2 we present the proof of Theorem 4.2.

E.1 Algorithms

In this subsection, we provide algorithms mentioned in Section 4.2. Algorithm 1 will transform a hard-case OPE problem for admissible data into a hard-case OPE problem for trajectory data. Algorithm 2 is used in the proof of lower bound with trajectory data, where the algorithm can transform admissible data collected according to M (MDP in the hard-case OPE problem of admissible data) into trajectory data of $\widetilde{M}$ (MDP in the hard-case OPE problem of trajectory data). And finally, Algorithm 3 shows how to transform an algorithm for OPE with trajectory data to an algorithm for OPE with admissible data.

E.2 Proof of Theorem 4.2

For MDP $\widetilde{M}$ and policies $\widetilde{\pi}_{\mathrm{e}}, \widetilde{\pi}_b$, we have the following properties:

Lemma 20. *Suppose under* OPE *problem* $(\widetilde{M}, \widetilde{\pi}_{\mathrm{e}}, \widetilde{\pi}_{\mathrm{e}}, \mathcal{F})$ *is the output of Algorithm 1 after inputing* OPE *problem* $(M, \pi_{\mathrm{e}}, \pi_b, \mathcal{F})$ *and arbitrary integer* K*. Then,*

(a) *The standard concentrability of* $\widetilde{M}$ *and* M *satisfies that*

$$\mathsf{C}(\widetilde{M}, \widetilde{d}_h^{\widetilde{\pi}_b}(\cdot; \widetilde{M}) \le 2\mathsf{C}(M, d_h^{\pi_b}(\cdot; M)).$$

(b) $\widetilde{Q}$ *calculated in (41) equal to the* Q*-function of* MDP $\widetilde{M}$*. Especially, the value functions of* $\widetilde{M}$ *and* M *satisfies*

$$\widetilde{V}(\widetilde{\rho}; \widetilde{M}) = V(\rho; M), \tag{42}$$

where $\widetilde{\rho}$ *and* ρ *are initial distributions of* $\widetilde{M}$ *and* M *respectively.*

Proof of Lemma 20. We first prove Lemma 20 (a). We only need to verify that for any $1 \le h \le H$ and $1 \le k \le K$,

$$\frac{d_h^{\widetilde{\pi}_{\mathrm{e}}}((x_h, k), a; \widetilde{M})}{d_h^{\widetilde{\pi}_b}((x_h, k), a; \widetilde{M})} \le 2 \cdot \frac{d_h^{\pi_{\mathrm{e}}}(x_h, a; M)}{d_h^{\pi_b}(x_h, a; M)}. \tag{43}$$

Our first observation is that $d_h^{\widetilde{\pi}_{\mathrm{e}}}((x_h, k); \widetilde{M}) = d_h^{\pi_{\mathrm{e}}}(x_h; M)$. This can be proved via induction on $(h-1)K + k$. When $(h-1)K + k = 1$, i.e. $h = k = 1$, we have

$$d_h^{\widetilde{\pi}_{\mathrm{e}}}((x_h, k); \widetilde{M}) = \widetilde{\rho}(x_1) = \rho(x_1) = d_h^{\pi_{\mathrm{e}}}(x_h; M).$$

Next, we will do the induction. When $k \ge 2$, we have $\widetilde{T}((x_h, k) \mid (x_h, k-1), \widetilde{\pi}_{\mathrm{e}}((x_h, k-1))) = 1$, which implies

$$d_h^{\widetilde{\pi}_{\mathrm{e}}}((x_h, k); \widetilde{M}) = d_h^{\widetilde{\pi}_{\mathrm{e}}}((x_h, k-1); \widetilde{M}) = d^{\pi_{\mathrm{e}}}(x_h; M).$$

When $k = 1$, we have $\widetilde{T}((x_h, 1) \mid (x_{h-1}, K), \widetilde{\pi}_{\mathrm{e}}((x_{h-1}, K))) = T(x_h \mid x_{h-1}, \pi_{\mathrm{e}}(x_{h-1}))$. Hence by induction, we have

$$d_h^{\widetilde{\pi}_{\mathrm{e}}}((x_h, 1); \widetilde{M}) = \sum_{x_{h-1} \in \mathcal{X}_{h-1}} d_h^{\widetilde{\pi}_{\mathrm{e}}}((x_{h-1}, K); \widetilde{M}) \widetilde{T}((x_h, 1) \mid (x_{h-1}, K), \widetilde{\pi}_{\mathrm{e}}((x_{h-1}, K)))$$

Algorithm 1 REPLICATOR

Input:

- Offline policy evaluation problem $(M, \pi_e, \pi_b, \mathcal{F})$ with MDP $M = (\mathcal{X}, \mathcal{A}, T, r, H, \rho)$
- Parameter $K \geq 1$.

Define $\pi_c : \mathcal{X} \to \mathcal{A}$ to be an arbitrary mapping which satisfies $\pi_c(x) \neq \pi_e(x)$ for all $x \in \mathcal{X}$.

`/* Construct MDP` $\widetilde{M} = \mathtt{MDP}(\widetilde{\mathcal{X}}, \mathcal{A}, \widetilde{T}, \widetilde{r}, \widetilde{H}, \widetilde{\rho})$ `*/`

Horizon: $\widetilde{H} = K(H-1)+1$.

State space: $\widetilde{\mathcal{X}}_l = \{(x,k) : x \in \mathcal{X}_h, k \in [K]\}$ for $l = (h-1)K + k \leq \widetilde{H}$ with $1 \leq h \leq H$ and $1 \leq k \leq K$.

Initial distribution: $\widetilde{\rho}((x,1)) = \rho(x)$ for $x \in \mathcal{X}_1$.

Transition model: For $k \leq K-1$, define transition for $x, x' \in \mathcal{X}_h$ with $h \in [H]$ and $a \in \mathcal{A}$:

$$\widetilde{T}((x',k+1) \mid (x,k), a) = \begin{cases} \mathbb{I}(x' = x) & \text{if} \quad a = \pi_e(x), \\ d_h^{\pi_b}(x'; M) & \text{if} \quad a \neq \pi_e(x). \end{cases}$$

For $k = K$, define transition for $x \in \mathcal{X}_h, x' \in \mathcal{X}_{h+1}$ with $h = [H]$ and $a \in \mathcal{A}$:

$$\widetilde{T}((x',1) \mid (x,k), a) = T(x' \mid x, a).$$

Reward Functions: For $k \leq K-1$, define reward $\widetilde{r}((x,k),a) = 0$ for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$.

For $k = K$, define reward $\widetilde{r}((x,k),a) = r(x,a)$ for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$.

`/* Construct evaluation policy` $\widetilde{\pi}_e$ `*/`

Construct $\widetilde{\pi}_e$ as $\widetilde{\pi}_e((x,k)) = \pi_e(x)$ for all $x \in \mathcal{X}$ and $k \in [K]$.

`/* Construct offline policy` $\widetilde{\pi}_b$ `*/`

Construct $\widetilde{\pi}_b$ for all $x \in \mathcal{X}$ and $k \in [K]$ as

$$\widetilde{\pi}_b((x,k)) = \begin{cases} \pi_b(x) & \text{if } k = K, \\ \frac{1}{2}\pi_e(x) + \frac{1}{2}\pi_c(x) & \text{otherwise.} \end{cases}$$

`/* Construct state-action value function class` $\widetilde{\mathcal{F}}$ `*/`

Construct $\widetilde{\mathcal{F}}$ as $\widetilde{\mathcal{F}} = \{\widetilde{Q} : Q \in \mathcal{F}\}$, where $\widetilde{Q} : \widetilde{\mathcal{X}} \times \mathcal{A} \to \mathbb{R}$ is defined for all $x \in \mathcal{X}$ and $k \in [K]$ as

$$\widetilde{Q}((x_h,k),a) = \begin{cases} Q(x_h,a) & k = K, \\ Q(x_h, \pi_e(x_h)) & k < K, a = \pi_e(x_h), \\ \sum_{x'_h \in \mathcal{X}_h} d_h^{\pi_b}(x'_h; M) Q(x'_h, \pi_e(x'_h)) & k < K, a \neq \pi_e(x_h). \end{cases} \tag{41}$$

Return: offline policy evaluation problem $(\widetilde{M}, \widetilde{\pi}_e, \widetilde{\pi}_b, \widetilde{\mathcal{F}})$.

Algorithm 2 ADMISSIBLE-TO-TRAJECTORY

Input: Admissible datasets $\mathcal{D}_h^{\mathrm{ADM}} = \{(x_h^{(l)}, a_h^{(l)}, r_h^{(l)}, \bar{x}_h^{(l)})\}_{l \le K}$ for $h \in [H]$.
Set $\widetilde{x}_1 = x_1^{(1)}$, and initialize $\tau = (\widetilde{x}_1)$.
for $h = 1, \ldots, H-1$ **do**
    Set $l = 1$.
`/* Construct Trajectory within Block` h `*/`
    for $k = 1, \ldots, K-1$ **do**
        Let $\widetilde{h} := (h-1)K + k$.
        Sample $\widetilde{a}_{\widetilde{h}} \sim \mathrm{Uniform}(\{\pi_{\mathrm{e}}(\widetilde{x}_{\widetilde{h}}), \pi_c(\widetilde{x}_{\widetilde{h}})\})$, and update $l \leftarrow l+1$ if $\widetilde{a}_{\widetilde{h}} = \pi_c(\widetilde{x}_{\widetilde{h}})$.
        Set

$$\widetilde{x}_{\widetilde{h}+1} = \begin{cases} \widetilde{x}_{\widetilde{h}} & \text{if } \widetilde{a}_{\widetilde{h}} = \pi_{\mathrm{e}}(\widetilde{x}_{\widetilde{h}}) \quad \texttt{// Map to the Same State} \\ x_h^{(l)} & \text{if } \widetilde{a}_{\widetilde{h}} = \pi_c(\widetilde{x}_{\widetilde{h}}) \quad \texttt{// Read Fresh State from } \mathcal{D}_h^{\mathrm{ADM}} \end{cases}$$

        Update $\tau = \tau \circ (\widetilde{a}_{\widetilde{h}}, \widetilde{r}_{\widetilde{h}} = 0, \widetilde{x}_{\widetilde{h}+1})$.
    Update $\tau = \tau \circ (\widetilde{a}_{\widetilde{h}} = a_h^{(l)}, \widetilde{r}_{\widetilde{h}} = r_h^{(l)}, \widetilde{x}_{\widetilde{h}+1} = \bar{x}_h^{(l)})$.
Return trajectory τ of length $(H-1)K + 1$.

Algorithm 3 Reduction of OPE with Admissible Data to Trajectory Data

Input:

- Admissible dataset $\mathcal{D}_h^{\mathrm{ADM}}$ of size Kn for each $h \in [H]$.
- State-action value function class $\mathcal{F}$.
- OPE algorithm $\mathrm{ALG}_{\mathrm{TRAJ}}$ that takes evaluation policy, state-action value function class and trajectory data as input and returns a value estimation.

Initialize $\mathcal{D}^{\mathrm{TRAJ}} = \varnothing$.
for $j = 1, \ldots, n$ **do**
    For $h \in [K]$, construct $\mathcal{D}_h^{\mathrm{ADM},j} = \mathcal{D}_h^{\mathrm{ADM}}[K(j-1)+1 : Kj]$
    Get trajectory $\tau^j = \text{ADMISSIBLE-TO-TRAJECTORY}\left(\mathcal{D}_1^{\mathrm{ADM},j}, \ldots, \mathcal{D}_H^{\mathrm{ADM},j}\right)$.
    Update $\mathcal{D}^{\mathrm{TRAJ}} \leftarrow \mathcal{D}^{\mathrm{TRAJ}} \cup \{\tau^j\}$.
Construct Q-function class $\widetilde{\mathcal{F}}$ according to (41) based on $\mathcal{F}$.
Return $\widehat{V} \leftarrow \mathrm{ALG}_{\mathrm{TRAJ}}(\mathcal{D}^{\mathrm{TRAJ}}, \widetilde{\mathcal{F}})$.

$$= \sum_{x_{h-1} \in \mathcal{X}_{h-1}} d_{h-1}^{\pi_e}(x_{h-1}; M) T(x_h \mid x_{h-1}, \pi_e(x_{h-1})) = d_h^{\pi_e}(x_h; M).$$

Our next observation is that $d_h^{\widetilde{\pi}_b}((x_h, k); \widetilde{M}) = d_h^{\pi_b}(x_h; M)$. This can be proved via induction on $(h-1)K+k$. When $(h-1)K + k = 1$, i.e. $h = k = 1$, we have

$$d_h^{\widetilde{\pi}_b}((x_h, k); \widetilde{M}) = \widetilde{\rho}(x_1) = \rho(x_1) = d_h^{\pi_b}(x_h; M).$$

Next, we will do the induction. When $k \geq 2$, according to the transition model of $\widetilde{M}$, we have

$$\widetilde{T}((x_h, k) \mid (x'_h, k-1), \pi_c(x'_h))) = d_h^{\pi_b}(x_h),$$

which indicates that

$$\begin{aligned}
d_h^{\widetilde{\pi}_b}((x_h, k); \widetilde{M}) &= \sum_{x'_h \in \mathcal{X}_h} d_h^{\widetilde{\pi}_b}((x'_h, k-1); \widetilde{M}) \widetilde{T}((x_h, k) \mid (x'_h, k-1), \widetilde{\pi}_b((x'_h, k-1)))) \\
&= \frac{1}{2} \sum_{x'_h \in \mathcal{X}_h} d_h^{\pi_b}(x'_h; M) \widetilde{T}((x_h, k) \mid (x'_h, k-1), \pi_e(x'_h))) \\
&\quad + \frac{1}{2} \sum_{x'_h \in \mathcal{X}_h} d_h^{\pi_b}(x'_h; M) \widetilde{T}((x_h, k) \mid (x'_h, k-1), \pi_c(x'_h))) \\
&= \frac{1}{2} d_h^{\pi_b}(x_h; M) + \frac{1}{2} \sum_{x'_h \in \mathcal{X}_h} d_h^{\pi_b}(x'_h; M) d_h^{\pi_b}(x_h; M) = \frac{1}{2} d_h^{\pi_b}(x_h; M)
\end{aligned}$$

When $k = 1$, we have $\widetilde{T}((x_h, 1) \mid (x_{h-1}, K), \widetilde{\pi}_b((x_{h-1}, K))) = T(x_h \mid x_{h-1}, \pi_b(x_{h-1}))$. Hence by induction, we have

$$\begin{aligned}
d_h^{\widetilde{\pi}_b}((x_h, 1); \widetilde{M}) &= \sum_{x_{h-1} \in \mathcal{X}_{h-1}} d_h^{\widetilde{\pi}_b}((x_{h-1}, K); \widetilde{M}) \widetilde{T}((x_h, 1) \mid (x_{h-1}, K), \widetilde{\pi}_b((x_{h-1}, K))) \\
&= \sum_{x_{h-1} \in \mathcal{X}_{h-1}} d_{h-1}^{\pi_b}(x_{h-1}; M) T(x_h \mid x_{h-1}, \pi_e(x_{h-1})) = d_h^{\pi_b}(x_h; M).
\end{aligned}$$

Our final observation is that for $a \neq \pi_e(x_h)$, we always have

$$d_h^{\widetilde{\pi}_e}((x_h, k), a; \widetilde{M}) = 0,$$

hence (43) holds. For $a = \pi_e(x_h)$, we have

$$\begin{aligned}
\frac{d_h^{\widetilde{\pi}_e}((x_h, k), a; \widetilde{M})}{d_h^{\widetilde{\pi}_b}((x_h, k), a; \widetilde{M})} &= \frac{d_h^{\widetilde{\pi}_e}((x_h, k); \widetilde{M})}{d_h^{\widetilde{\pi}_b}((x_h, k); \widetilde{M})} \cdot \frac{\widetilde{\pi}_e(a \mid (x_h, k))}{\widetilde{\pi}_b(a \mid (x_h, k))} \\
\frac{d_h^{\pi_e}(x_h, a; M)}{d_h^{\pi_b}(x_h, a; M)} &= \frac{d_h^{\pi_e}(x_h; M)}{d_h^{\pi_b}(x_h; M)} \cdot \frac{\pi_e(a \mid x_h)}{\pi_b(a \mid x_h)}.
\end{aligned}$$

According to our construction of $\widetilde{\pi}_e$ and $\widetilde{\pi}_b$, we further have

$$\frac{\widetilde{\pi}_e(a \mid (x_h, K))}{\widetilde{\pi}_b(a \mid (x_h, K))} = \frac{\pi_e(a \mid x_h)}{\pi_b(a \mid x_h)}.$$

For $k \neq K$, we have

$$\frac{\widetilde{\pi}_{\mathrm{e}}(a \mid (x_h, k))}{\widetilde{\pi}_b(a \mid (x_h, k))} = 2.$$

After noticing that $\pi_b(\pi_{\mathrm{e}}(x_h) \mid x_h) \leq 1$, we obtain that for $a = \pi_{\mathrm{e}}(x_h)$,

$$\frac{\pi_{\mathrm{e}}(a \mid x_h)}{\pi_b(a \mid x_h)} \geq 1 = \frac{1}{2} \cdot \frac{\widetilde{\pi}_{\mathrm{e}}(a \mid (x_h, k))}{\widetilde{\pi}_b(a \mid (x_h, k))}.$$

Hence as long as $a = \pi_{\mathrm{e}}(x_h)$, we always have

$$\frac{\widetilde{\pi}_{\mathrm{e}}(a \mid (x_h, K))}{\widetilde{\pi}_b(a \mid (x_h, K))} \leq 2 \cdot \frac{\pi_{\mathrm{e}}(a \mid x_h)}{\pi_b(a \mid x_h)}.$$

This implies

$$\frac{d_h^{\widetilde{\pi}_{\mathrm{e}}}((x_h, k), a; \widetilde{M})}{d_h^{\widetilde{\pi}_b}((x_h, k), a; \widetilde{M})} \leq 2 \cdot \frac{d_h^{\pi_{\mathrm{e}}}(x_h, a; M)}{d_h^{\pi_b}(x_h, a; M)}.$$

Next, we will prove Lemma 20 Lemma (b) by induction on $l := (h-1)H + k$ from $\widetilde{H}$ to 1. When $l = \widetilde{H}$, we have

$$\widetilde{Q}((x_H, 1), a) = Q(x_H, a) = 0, \quad \forall x_H \in \mathcal{X}_H, a \in \mathcal{A},$$

which satisfies the induction hypothesis. Next, assuming the induction hypothesis holds for $l+1$, we will prove the induction hypothesis for l. We assume $l = (h-1)K + k$.

When $k = K$, according to Bellman equation and induction hypothesis, we have

$$\begin{aligned}\widetilde{Q}((x_h, K), a) &= \widetilde{r}((x_h, K), a) + \sum_{x'_{h+1} \in \mathcal{X}_{h+1}} \widetilde{T}((x'_{h+1}, 1) \mid (x_h, K), a) \widetilde{Q}((x'_{h+1}, 1), \widetilde{\pi}_{\mathrm{e}}((x'_{h+1}, 1))) \\ &= r(x_h, a) + \sum_{x'_{h+1} \in \mathcal{X}_{h+1}} T(x'_{h+1} \mid x_h, a) Q(x'_{h+1}, \pi_{\mathrm{e}}(x'_{h+1})) = Q(x_h, a).\end{aligned}$$

When $k < K$ and $a = \pi_{\mathrm{e}}(x_h)$, according to Bellman equation, we have

$$\begin{aligned}\widetilde{Q}((x_h, k), a) &= \widetilde{r}((x_h, k), a) + \sum_{x'_h \in \mathcal{X}_h} \widetilde{T}((x'_h, k+1) \mid (x_h, k), a) \widetilde{Q}((x'_h, k+1), \widetilde{\pi}_{\mathrm{e}}((x'_h, k+1))) \\ &= \widetilde{Q}((x_h, k+1), \widetilde{\pi}_{\mathrm{e}}((x_h, k+1))) = Q(x_h, \pi_{\mathrm{e}}(x_h)),\end{aligned}$$

where in the second last equation we use the fact that $\widetilde{T}((x_h, k+1) \mid (x_h, k), \pi_{\mathrm{e}}(x_h)) = 1$.

When $k < K$ and $a \neq \pi_{\mathrm{e}}(x_h)$, according to Bellman equation, we have

$$\begin{aligned}\widetilde{Q}((x_h, k), a) &= \widetilde{r}((x_h, k), a) + \sum_{x'_h \in \mathcal{X}_h} \widetilde{T}((x'_h, k+1) \mid (x_h, k), a) \widetilde{Q}((x'_h, k+1), \widetilde{\pi}_{\mathrm{e}}((x'_h, k+1))) \\ &= \sum_{x'_h \in \mathcal{X}_h} d_h^{\pi_b}(x'_h; M) \widetilde{Q}((x'_h, k+1), \widetilde{\pi}_{\mathrm{e}}((x'_h, k+1))) = d_h^{\pi_b}(x_h; M) Q(x_h, \pi_{\mathrm{e}}(x_h)),\end{aligned}$$

Finally for (42) follows from $\widetilde{\rho}((x_1, 1)) = \rho(x_1)$ for any $x_1 \in \mathcal{X}_1$. ∎

Lemma 21. *Suppose the law of output of Algorithm 2 given admissible data collected from M is $\widetilde{\mathbb{P}}_M$, and the law of collecting one trajectory according to $\widetilde{M}$ is $\mathbb{P}_{\widetilde{M}}$. If we let*

$$\mathcal{E}_0 := \{\exists 1 \leq h \leq H-1, \widetilde{a}_{(h-1)K+1} = \cdots = \widetilde{a}_{(h-1)K+K-1} = 1\}. \tag{44}$$

Then for trajectory $\widetilde{\tau} = (\widetilde{x}_1, \widetilde{a}_1, \widetilde{r}_1, \widetilde{x}_2, \widetilde{a}_2, \widetilde{r}_2, \cdots, \widetilde{x}_{\widetilde{H}}) \notin \mathcal{E}_0$, we have

$$\widetilde{\mathbb{P}}_M(\widetilde{\tau}) = \mathbb{P}_{\widetilde{M}}(\widetilde{\tau}).$$

Proof of Lemma 21. We denote $\widetilde{\tau}_l = (\widetilde{x}_1, \widetilde{a}_1, \widetilde{r}_1, \widetilde{x}_2, \widetilde{a}_2, \widetilde{r}_2, \cdots, \widetilde{x}_l)$, and will prove by induction on l that

$$\widetilde{\mathbb{P}}_M(\widetilde{\tau}_l) = \mathbb{P}_{\widetilde{M}}(\widetilde{\tau}_l). \tag{45}$$

When $l = 1$, this is true since $\widetilde{\mathbb{P}}_M(\widetilde{\tau}_1) = \widetilde{\rho}(\widetilde{x}_1) = \mathbb{P}_{\widetilde{M}}(\widetilde{\tau}_l)$. Next, to finish induction from l to $l+1$, by the chain rule of probability, we only need to show that

$$\widetilde{\mathbb{P}}_M(\widetilde{a}_l, \widetilde{r}_l, \widetilde{x}_{l+1} \mid \widetilde{\tau}_l) = \mathbb{P}_{\widetilde{M}}(\widetilde{a}_l, \widetilde{r}_l, \widetilde{x}_{l+1} \mid \widetilde{\tau}_l). \tag{46}$$

We write $l = hK + k$ with $1 \le k \le K$, and $\widetilde{x}_l = (x_h, k)$ and $\widetilde{x}_{l+1} = (x'_h, k+1)$ (or $(x'_{h+1}, 1)$). If $k = K$, we have

$$\widetilde{\mathbb{P}}_M(\widetilde{a}_l, \widetilde{r}_l, \widetilde{x}_{l+1} \mid \widetilde{\tau}_l) = \pi_b(\widetilde{a}_l \mid x_h) R(\widetilde{r}_l \mid x_h, \widetilde{a}_l) T(x'_{h+1} \mid x_h, \widetilde{a}_l).$$

According to Algorithm 2, as long as there exists some $(h-1)K + 1 \le t \le (h-1)K + K - 1$ with $\widetilde{a}_t \ne 1$, we will have $(\widetilde{x}_l, \widetilde{a}_l, \widetilde{r}_l, \widetilde{x}_{l+1}) \in \mathcal{D}_h^{\mathrm{ADM}}$, which indicates that

$$\mathbb{P}_{\widetilde{M}}(\widetilde{a}_l, \widetilde{r}_l, \widetilde{x}_{l+1} \mid \widetilde{\tau}_l) = \pi_b(\widetilde{a}_l \mid x_h) R(\widetilde{r}_l \mid x_h, \widetilde{a}_l) T(x'_{h+1} \mid x_h, \widetilde{a}_l) = \widetilde{\mathbb{P}}_M(\widetilde{a}_l, \widetilde{r}_l, \widetilde{x}_{l+1} \mid \widetilde{\tau}_l).$$

If $l = hK + k$ for some h and $1 \le k \le K - 1$, then the transition model of $\widetilde{M}$ gives that

$$\mathbb{P}_{\widetilde{M}}(\widetilde{a}_l, \widetilde{r}_l, \widetilde{x}_{l+1} \mid \widetilde{\tau}_l) = \begin{cases} \frac{1}{2} \cdot \delta(\widetilde{r}_l = 0) \cdot \delta(x'_h = x_h), & \text{if } \widetilde{a}_l = \pi_{\mathrm{e}}(x_l), \\ \frac{1}{2} \cdot \delta(\widetilde{r}_l = 0) \cdot d_h^{\pi_b}(x'_h; M), & \text{if } \widetilde{a}_l = \pi_c(x_l). \end{cases}$$

Additionally, according to Algorithm 2, action 1 is taken with probability $1/2$. Hence if $\widetilde{a}_l = \pi_{\mathrm{e}}(x_l)$, then the algorithm chooses $x'_h = x_h$ and $\widetilde{r} = 0$, we have

$$\text{LHS of (46)} = \frac{1}{2} \cdot \delta(\widetilde{r}_l = 0) \delta(x'_h = x_h).$$

And if $\widetilde{a}_l = \pi_c(x_l)$, the algorithm will sample x'_h from $d_h^{\pi_b}(\cdot; M)$. Hence we obtain

$$\text{LHS of (46)} = \frac{1}{2} \cdot \delta(\widetilde{r}_l = 0) d_h^{\pi_b}(x'_h; M).$$

This finishes the proof of induction at $l + 1$.

Finally, by induction, (45) holds for $l = \widetilde{H}$ as long as $\widetilde{\tau} \notin \mathcal{E}_0$. Hence for $\widetilde{\tau} \notin \mathcal{E}_0$ we always have $\widetilde{\mathbb{P}}_M(\widetilde{\tau}) = \mathbb{P}_{\widetilde{M}}(\widetilde{\tau})$. ∎

Theorem E.1. *Suppose for $\widetilde{M}$, algorithm $\mathrm{ALG}_{\mathrm{TRAJ}}$ taking n trajectories T_n and Q-function class $\widetilde{\mathcal{F}}$ can output $\widehat{V}_{T_n} \in [-1, 1]$ such that*

$$\mathbb{E}_{T_n \sim \widetilde{M}} \left[\left| \widehat{V}_{T_n} - V^{\widetilde{M}}(\widetilde{\rho}) \right| \right] \le \varepsilon.$$

Then taking $H^2 n$ admissible dataset $D_{H^2 n}$ and (Q, W)-tuple class, Algorithm 3 can output ε-close value to $\widehat{V}_{D_{H^2 n}}$ such that

$$\mathbb{E}_{T_n \sim M} \left[\left| \widehat{V}_{D_n} - V^M(\rho) \right| \right] \le \varepsilon + H 2^{-K+2}.$$

Proof of Theorem E.1. We use $\widetilde{\tau}_{1:n}$ to denote n trajectories $\widetilde{\tau}_1, \cdots, \widetilde{\tau}_n$, and let $e(\widetilde{\tau}_{1:n}) = |\widehat{V}_{\widetilde{\tau}_{1:n}} - V^{\widetilde{M}}(\widetilde{\rho})|$. We have

$$\sum_{\widetilde{\tau}_{1:n}} e(\widetilde{\tau}_{1:n}) \prod_{j=1}^{n} \mathbb{P}_{\widetilde{M}}(\widetilde{\tau}_j) = \mathbb{E}_{T_n \sim \widetilde{M}} \left[\left| \widehat{V}_{T_n} - V^{\widetilde{M}}(\widetilde{\rho}) \right| \right] \le \varepsilon. \tag{47}$$

We further notice $V(\widetilde{\rho};\widetilde{M}) = V(\rho;M)$ from Lemma 20. Furthermore, since Algorithm 3 first transit the admissible data D_{HKn} into trajectory data T_n according to Algorithm 2, and then output $\widehat{V}_{T_n}$ according to algorithm ALG_{TRAJ}, we have

$$\mathbb{E}_{T_n\sim M}\left[\left|\widehat{V}_{D_n} - V^M(\rho)\right|\right] = \sum_{\widetilde{\tau}_{1:n}} \left|\widehat{V}_{\widetilde{\tau}_{1:n}} - V^{\widetilde{M}}(\widetilde{\rho})\right| \prod_{j=1}^n \widetilde{\mathbb{P}}_M(\widetilde{\tau}_j) = \sum_{\widetilde{\tau}_{1:n}} e(\widetilde{\tau}_{1:n}) \prod_{j=1}^n \widetilde{\mathbb{P}}_M(\widetilde{\tau}_j),$$

where $\widetilde{\mathbb{P}}_M$ is defined in Lemma 21. Next, Lemma 21 indicates that if $\widetilde{\tau}_j \notin \mathcal{E}_0$ (where $\mathcal{E}_0$ is defined in (44)), then $\widetilde{\mathbb{P}}_M(\widetilde{\tau}_j) = \mathbb{P}_{\widetilde{M}}(\widetilde{\tau}_j)$. Therefore, noticing that $e(\widetilde{\tau}_{1:n}) = |\widehat{V}_{\widetilde{\tau}_{1:n}} - V^{\widetilde{M}}(\widetilde{\rho})| \le 2$, we have

$$\begin{aligned}
\sum_{\widetilde{\tau}_{1:n}} e(\widetilde{\tau}_{1:n}) \prod_{j=1}^n \widetilde{\mathbb{P}}_M(\widetilde{\tau}_j) &= \sum_{\widetilde{\tau}_{1:n}:\exists 1\le i\le n, \widetilde{\tau}_i\in\mathcal{E}_0} e(\widetilde{\tau}_{1:n}) \prod_{j=1}^n \widetilde{\mathbb{P}}_M(\widetilde{\tau}_j) + \sum_{\widetilde{\tau}_{1:n}:\forall 1\le i\le n, \widetilde{\tau}_i\notin\mathcal{E}_0} e(\widetilde{\tau}_{1:n}) \prod_{j=1}^n \widetilde{\mathbb{P}}_M(\widetilde{\tau}_j) \\
&= 2 \sum_{\widetilde{\tau}_{1:n}:\exists 1\le i\le n, \widetilde{\tau}_i\in\mathcal{E}_0} \prod_{j=1}^n \widetilde{\mathbb{P}}_M(\widetilde{\tau}_j) + \sum_{\widetilde{\tau}_{1:n}:\forall 1\le i\le n, \widetilde{\tau}_i\notin\mathcal{E}_0} e(\widetilde{\tau}_{1:n}) \prod_{j=1}^n \mathbb{P}_{\widetilde{M}}(\widetilde{\tau}_j) \\
&\le 2\sum_{i=1}^n \widetilde{\mathbb{P}}_M(\widetilde{\tau}_i \in \mathcal{E}_0) + \sum_{\widetilde{\tau}_{1:n}:\forall 1\le i\le n, \widetilde{\tau}_i\notin\mathcal{E}_0} e(\widetilde{\tau}_{1:n}) \prod_{j=1}^n \mathbb{P}_{\widetilde{M}}(\widetilde{\tau}_j) \\
&\le 2\sum_{i=1}^n \widetilde{\mathbb{P}}_M(\widetilde{\tau}_i \in \mathcal{E}_0) + \sum_{\widetilde{\tau}_{1:n}} e(\widetilde{\tau}_{1:n}) \prod_{j=1}^n \mathbb{P}_{\widetilde{M}}(\widetilde{\tau}_j)
\end{aligned}$$

Further notice that for any $1 \le j \le n$, we have

$$\widetilde{\mathbb{P}}_M(\widetilde{\tau} \in \mathcal{E}_0) = 1 - (1-2^{-H})^K \le H2^{-K+1}.$$

Therefore, using (47), we obtain

$$2\sum_{i=1}^n \widetilde{\mathbb{P}}_M(\widetilde{\tau}_i \in \mathcal{E}_0) + \sum_{\widetilde{\tau}_{1:n}} e(\widetilde{\tau}_{1:n}) \prod_{j=1}^n \mathbb{P}_{\widetilde{M}}(\widetilde{\tau}_j) \le 2nH2^{-K+1} + \varepsilon,$$

which indicates that $\mathbb{E}_{T_n\sim M}\left[\left|\widehat{V}_{D_n} - V^M(\rho)\right|\right] \le \varepsilon + H2^{-K+2}$. ■

This theorem has the following direct corollary, indicating that any algorithm taking trajectory data as input cannot do policy evaluation in polynomial time.

Proof of Theorem 4.2. According to Theorem D.1, there exists a class $\mathcal{M}$ of MDPs, where each $M \in \mathcal{M}$ has bounded coverage $384H^3$. And any algorithm which takes $o(H2^H/\varepsilon)$ number of admissible samples together with realizable (Q,W)-tuple class induces estimation error $\varepsilon/8H$ in at least one MDP. We further carry the lifting in this section for any MDP in $\mathcal{M}$, and suppose these lifting MDPs form the class $\widetilde{\mathcal{M}}$. Lemma Lemma 20 indicates that every instance in $\widetilde{\mathcal{M}}$ has bounded coverage $768H^3$.

Next, we will prove this corollary by contradiction. Suppose the algorithm ALG_{TRAJ} using $\tilde{o}(H2^H/\varepsilon)$ trajectories together with realizable Q-function class induces estimation error less than $\varepsilon/16H$ in every MDP in $\widetilde{\mathcal{M}}$. Then after inserting ALG_{TRAJ} into Algorithm 3, we form an algorithm which takes $\tilde{o}(H2^H/\varepsilon)$ admissible data for each layer, together with (Q,W)-function class, and outputs an estimation to the value function.

Theorem E.1 indicates that this algorithm will induce $\varepsilon/16H + H2^{-K+2}$ error for all MDPs in $\mathcal{M}$. Hence taking $K = 2 + \log_2 16H^2/\varepsilon = \log_2 64H^2/\varepsilon$, this algorithm will induces estimation error less than $\varepsilon/8H$ in all MDPs in $\mathcal{M}$. Notice that with this choice of K, we have $\tilde{o}(KH2^H/\varepsilon) \le \tilde{o}(H2^H/\varepsilon)$, which contradicts to Theorem D.1. ■

Algorithm 4 Batch Value-Function Tournament for Policy Evaluation (Xie and Jiang, 2021)

Input: Evaluation policy π_{e}, Offline Dataset $\mathcal{D}$ consisting of n tuples of the form (x, a, r, x').

1: Among all samples (x, a, r, x') in $\mathcal{D}$, only keep those such that $a = \pi_{\mathrm{e}}(x)$, and discard all others. We denote the new dataset as $\mathcal{D}' = \{(x_i, r_i, x_i')\}_{i=1}^{n'}$, where we omit a_i's since a_i is always equal to $\pi_{\mathrm{e}}(x_i)$ in this dataset.

2: Compute

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \max_{f' \in \mathcal{F}} \max_{h} \|f - \widehat{\mathcal{T}}_{\mathcal{G}(f,f')} f\|_{\hat{\nu},h} \tag{48}$$

where

$$\mathcal{G}(f, f') := \left\{ g : \mathcal{X} \to \mathbb{R} \;\middle|\; g(x) = g(y) \quad \text{if } f(x) = f(y) \text{ and } f'(x) = f'(y) \text{ and } h(x) = h(y) \right\},$$

($h(x) \in [H]$ denotes the layer at which x lies)

$$\widehat{\mathcal{T}}_{\mathcal{G}} f := \arg\min_{g \in \mathcal{G}} \sum_{i=1}^{n'} \left(g(x_i) - r_i - f(x_i') \right)^2, \tag{49}$$

$$\hat{\nu}(x) := \frac{1}{n} \sum_{i=1}^{n'} \mathbb{I}\{x_i = x\}. \tag{50}$$

3: **Return** $\widehat{f}$.

F Upper Bound for Offline Policy Evaluation

F.1 Setup

In previous sections we construct the lower bound assuming access to the $Q^{\pi_{\mathrm{e}}}$ function class. For the upper bound, we consider a slightly more challenging scenario, where the learner only has access to the $V^{\pi_{\mathrm{e}}}$ function class. It is not hard to see that learning with a $V^{\pi_{\mathrm{e}}}$ function class is more challenging than learning with $Q^{\pi_{\mathrm{e}}}$ because one can always reduce a $Q^{\pi_{\mathrm{e}}}$ function set to a V function set by redefining $f(x) \leftarrow \mathbb{E}_{a \sim \pi_{\mathrm{e}}(\cdot|x))}[f(x, a)]$. For simplicity, we assume that π_{e} is deterministic, though the extension to the stochastic case is straightforward. See Section F.2 for the discussion.

Similar to Xie and Jiang (2021), we first establish results for the case where the function set only *approximately* realizes the true $V^{\pi_{\mathrm{e}}}$. The result for the fully realizable setting can be easily deduced from it. Formally, we assume that the learner is given a function set $\mathcal{F}$ that consists of mappings $\mathcal{X} \to [-H, H]$ with the following approximate realizability guarantee.

Assumption F.1 (Approximate value function realizability)**.** *There exists an* $f^\star \in \mathcal{F}$ *such that* $\sup_{x \in \mathcal{X}} |V^{\pi_{\mathrm{e}}}(x) - f^\star(x)| \le \varepsilon_{\mathrm{appr}}$.

Besides, the learner is given an offline dataset $\mathcal{D}$, which consists of n tuples of (x, a, r, x') with (x, a) drawn from μ.

F.2 Algorithm

Our algorithm for this setting is presented in Algorithm 4, which is an adaptation of the BVFT algorithm (Xie and Jiang, 2021) to the case of policy evaluation. In the beginning of the algorithm, the dataset $\mathcal{D}$ is

pre-processed so that only (x, a, r, x') samples with $a = \pi_e(x)$ are kept (line 1). The core of the algorithm is to solve the min-max problem in (48). The high-level idea of it is that for every pair of functions $f, f' \in \mathcal{F}$, the algorithm creates a "tabular problem" by aggregating states with the same $(f(x), f'(x))$ value, and estimates the Bellman error for this tabular problem. Intuitively, this is probably the best the learner can do, since besides the value of $(f(x))_{f\in\mathcal{F}}$, the learner has no other ways to distinguish states in the large state space. The output function $\widehat{f}$ is the one that always attains a small Bellman error estimate no matter what the other function it is paired with. For more explanation on this min-max formulation, we refer the reader to Xie and Jiang (2021) or the amazing talk by Jiang (2021). The key differences with the algorithm of Xie and Jiang (2021) are the following:

- We deal with policy evaluation for π_e, and our function class consists of V^{π_e} functions, while Xie and Jiang (2021) deal with policy optimization, and their function class consists of $Q^\star$ functions. For this reason, we have a *preprocessing step* in line 1 of Algorithm 4, which removes data samples whose action is not generated by π_e. These samples are irrelevant to our policy evaluation task.
- We consider the finite-horizon setting while Xie and Jiang (2021) considers the discounted infinite-horizon setting. Therefore, different from theirs, Our aggregation is performed in a layer-by-layer manner, and only states in the same layer can be aggregated.

If π_e is stochastic, we perform the pre-processing step (line 1) in the following way: for each sample $(x, a, r, x') \in \mathcal{D}$, sample $a' \sim \pi_e(\cdot \mid x)$. If $a = a'$, then keep this sample; otherwise discard this sample.

Our upper bound result is stated in the following theorem, whose proof is provided in Appendix F.3 to Appendix F.5.

Theorem F.1. *Let $\widehat{f} \in \mathcal{F}$ be the output of the BVFT algorithm given in Algorithm 4. Let $\Phi(f, f')$ be the state aggregation scheme determined by f and f' (see Definition F.2 for the precise definition), and let $\bar{\mathsf{C}} = \max_{f,f'\in\mathcal{F}} \bar{\mathsf{C}}_{\varepsilon^2/H^2}(M, \Phi(f, f'), \mu)$. For given $\delta > 0$, if $n \geq \widetilde{\Omega}\left(\frac{\bar{\mathsf{C}}^2 H^6 \log(|\mathcal{F}|/\delta)}{\varepsilon^4}\right)$, then with probability at least $1 - \delta$,*

$$\left|\mathbb{E}_{x\sim\rho}[V^{\pi_e}(x) - \widehat{f}(x)]\right| \leq O(\varepsilon).$$

Because of the state aggregation procedure in BVFT, the sample complexity upper bound in Theorem F.1 depends on the concentrability coefficient of the aggregated MDP (i.e., $\bar{\mathsf{C}}$) rather than that of the original MDP. Notice also that the sample complexity scales with $\frac{1}{\varepsilon^4}$ instead of the more common $\frac{1}{\varepsilon^2}$. This is similar to Xie and Jiang (2021) and is because we divide the state space into $O(\frac{1}{\varepsilon^2})$ aggregations, each of which consists of states having the same value functions up to an accuracy of ε. Our bound have a smaller dependence on the horizon length H, but this is simply because we assume the range of the value function is $[-1, 1]$ while they assume it to be $[-H, H]$.

Finally, we provide some implications of Theorem F.1. First, as pointed out previously, we have $\bar{\mathsf{C}} \leq \mathsf{C}_{\text{pf}}$ (Lemma 29). Second, in the case that the data is admissible with offline policy π_b, and $\frac{1}{\pi_b(\pi_e(x)|x)} \leq \mathsf{C}_{\mathcal{A}}$ for all x, we have $\bar{\mathsf{C}} \leq (\mathsf{C}_{\mathcal{A}})^H$ (Lemma 30). Interestingly, a sample complexity bound of order $(\mathsf{C}_{\mathcal{A}})^H$ is also the case if we use *importance sampling* to perform offline policy evaluation. The difference is that importance sampling does not require access to any function class, but needs the data to be trajectories, while BVFT requires access to a function class, but only needs the data to be admissible.

For the remaining of this section, we provide a proof for Theorem F.1.

F.3 Definitions

Definition F.1 (partial offline data distribution $\nu(x)$)**.** *Given the offline data distribution $\mu \in \Delta(\mathcal{X} \times \mathcal{A})$, we define the partial distribution $\nu \in \Delta(\mathcal{X})$ such that $\nu(x) = \mu(x, \pi_e(x))$.*

Definition F.2 (aggregation schemes $\Phi(f, f')$, $\Phi_h(f, f')$ and maximum partition number $\Phi_{\max}$)**.** *Define $\Phi(f, f')$ as the state aggregation scheme (see Section 3.1) where x, y belongs to the same partition if and only if $f(x) = f(y)$ and $f'(x) = f'(y)$ and x, y are in the same layer. Let $\Phi_h(f, f') \subset \Phi(f, f')$ be the set of partitions in layer h. Define $\Phi_{\max} = \max_{f,f' \in \mathcal{F}} \max_h |\Phi_h(f, f')|$.*

Definition F.3 (aggregation $\Phi^\star$, aggregated transition $\bar{T}$, occupancy $\bar{d}$, and offline distribution $\bar{\nu}$)**.** *Consider the partition $\Phi^\star = \Phi(\widehat{f}, f^\star)$, where $\widehat{f}$ is the output of Algorithm 4 and $f^\star$ is defined in Assumption F.1. Let $\phi, \phi' \in \Phi^\star$, Define*

$$\bar{T}(\phi' \mid \phi) = \frac{\sum_{x \in \phi} \sum_{x' \in \phi'} \nu(x) T(x' \mid x, \pi_e)}{\sum_{x \in \phi} \nu(x)}.$$

Furthermore, let $\bar{d}(\phi)$ be the occupancy measure of π_e in the aggregated MDP. That is, $\bar{d}$ follows the recursive definition below:

$$\forall \phi' \in \Phi^\star_{h+1}, \quad \bar{d}(\phi') = \sum_{\phi \in \Phi^\star_h} \bar{d}(\phi) \bar{T}(\phi' \mid \phi), \quad \text{with } \bar{d}(\phi) = \frac{1}{H} \sum_{x \in \phi} \rho(x) \text{ for } \phi \in \Phi^\star_1.$$

Also, define $\bar{\nu}(\phi) = \sum_{x \in \phi} \nu(x)$.

Definition F.4 (aggregated concentrability $\bar{\mathsf{C}}^\star_\varepsilon$)**.** *The aggregated concentrability with respect to the aggregation $\Phi^\star = \Phi(\hat{f}, f^\star)$ (defined in Definition F.3) is defined as*

$$\bar{\mathsf{C}}^\star_\varepsilon = \max_h \max \left\{ \frac{\sum_{\phi \in \mathcal{I}} \bar{d}(\phi)}{\sum_{\phi \in \mathcal{I}} \bar{\nu}(\phi)} : \quad \mathcal{I} \subset \Phi^\star_h, \quad \sum_{\phi \in \mathcal{I}} \bar{d}(\phi) \geq \varepsilon \right\}. \tag{51}$$

Definition F.5 (projection operators $\mathcal{T}_\mathcal{G} f$ and $\hat{\mathcal{T}}_\mathcal{G} f$)**.** *Let $f : \mathcal{X} \to \mathbb{R}$, and let $\mathcal{G}$ be any function set that consists of functions of the form $\mathcal{X} \to \mathbb{R}$. Define*

$$\mathcal{T}_\mathcal{G} f = \arg\min_{g \in \mathcal{G}} \sum_{s, x'} \nu(x) T(x' \mid x, \pi_e) \left(g(x) - r(x, \pi_e) - f(x') \right)^2,$$

$$\widehat{\mathcal{T}}_\mathcal{G} f = \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n'} \left(g(x_i) - r_i - f(x'_i) \right)^2.$$

Definition F.6 (weighted norm $\|g\|_{w,h}$)**.** *Let $g \in \mathcal{G}(f, f')$ and $\Phi = \Phi(f, f')$ for some $f, f' \in \mathcal{F}$. Let $w : \Phi \to \mathbb{R}_{\geq 0}$ be arbitrary. With abuse of notation, define $\|g\|_{w,h} = \sqrt{\sum_{\phi \in \Phi_h} w(\phi) g(\phi)^2}$, where $g(\phi)$ is such that $g(x) = g(\phi)$ for all $x \in \phi$.*

Definition F.7 (estimation error $\varepsilon_{\text{stat}}$)**.** $\varepsilon_{\text{stat}} = \sqrt{\frac{\Phi_{\max} \log(n \Phi_{\max} |\mathcal{F}| / \delta)}{n}}$*, where n is the number of offline samples, and $\Phi_{\max}$ is defined in Definition F.2.*

We next establish a few properties of state aggregation.

Lemma 22. *Let $g = \mathcal{T}_{\mathcal{G}(f,f')} f$ for some $f, f' \in \mathcal{F}$. Fix a layer h. Let $g_{\alpha\beta}$ be the value of $g(x)$ for those $x \in \mathcal{X}_h$'s such that $f(x) = \alpha$ and $f'(x) = \beta$. Then $g_{\alpha\beta}$ has the following form:*

$$g_{\alpha\beta} = \frac{\sum_{x \in \mathcal{X}_h : f(x) = \alpha, f'(x) = \beta} \nu(x) \mathcal{T} f(x)}{\sum_{x \in \mathcal{X}_h : f(x) = \alpha, f'(x) = \beta} \nu(x)}$$

Proof of Lemma 22. Recall from Definition F.5 that g is the minimizer of $\mathbb{E}_{x\sim\nu}\mathbb{E}_{x'\sim T(\cdot|x,\pi_e)}\left(g(x) - r(x,\pi_e) - f(x')\right)^2$ within $\mathcal{G}(f,f')$. The derivative of this objective with respect to $g_{\alpha\beta}$ is

$$\begin{aligned}
&\mathbb{E}_{x\sim\nu}\mathbb{E}_{x'\sim T(\cdot|x,\pi_e)} 2(g_{\alpha\beta} - r(x,\pi_e) - f(x'))\mathbb{I}\{f(x)=\alpha, f'(x)=\beta, h(x)=h\} \\
&= 2g_{\alpha\beta} \sum_{x\in\mathcal{X}_h:\ f(x)=\alpha,f'(x)=\beta} \nu(x) - 2 \sum_{x\in\mathcal{X}_h:f(x)=\alpha,f'(x)=\beta} \nu(x)\left(r(x,\pi_e) + \sum_{x'} T(x' \mid x,\pi_e) f(x')\right) \\
&= 2g_{\alpha\beta} \sum_{x\in\mathcal{X}_h:\ f(x)=\alpha,f'(x)=\beta} \nu(x) - 2 \sum_{x\in\mathcal{X}_h:f(x)=\alpha,f'(x)=\beta} \nu(x)\mathcal{T}f(x).
\end{aligned}$$

Setting this to be zero gives the desired expression of $g_{\alpha\beta}$. ∎

Lemma 23. *For any $f\in\mathcal{F}$, we have that $\max_{f'\in\mathcal{F}} \|f - \mathcal{T}_{\mathcal{G}(f,f')}f\|_{\nu,h} \le \|f - \mathcal{T}f\|_{\nu,h}$.*

Proof of Lemma 23. Fix $f,f'\in\mathcal{F}$ and fix $h\in[H]$. Let $g = \mathcal{T}_{\mathcal{G}(f,f')}f$ and let $g_{\alpha\beta}$ be the value of $g(x)$ for $x\in\mathcal{X}_h$ such that $f(x)=\alpha$ and $f'(x)=\beta$.
Define $\nu_{\alpha\beta} = \sum_{x\in\mathcal{X}_h}\nu(x)\mathbb{I}\{f(x)=\alpha \text{ and } f'(x)=\beta\}$. Then by definition, we have

$$\begin{aligned}
\|f - \mathcal{T}_{\mathcal{G}(f,f')}f\|_{\nu,h}^2 &= \sum_{\alpha,\beta}\nu_{\alpha\beta}(\alpha - g_{\alpha\beta})^2 \\
&= \sum_{\alpha,\beta}\nu_{\alpha\beta}\left(\alpha - \frac{1}{\nu_{\alpha\beta}}\sum_{x\in\mathcal{X}_h:f(x)=\alpha,f'(x)=\beta}\nu(x)\mathcal{T}f(x)\right)^2 && \text{(using Lemma 22)} \\
&\le \sum_{\alpha,\beta}\nu_{\alpha\beta}\frac{1}{\nu_{\alpha\beta}}\sum_{x\in\mathcal{X}_h:f(x)=\alpha,f'(x)=\beta}\nu(x)\left(\alpha - \mathcal{T}f(x)\right)^2 && \text{(Jensen's inequality)} \\
&= \sum_{x\in\mathcal{X}_h}\nu(x)\left(f(x) - \mathcal{T}f(x)\right)^2 \\
&= \|f - \mathcal{T}f\|_{\nu,h}^2.
\end{aligned}$$

∎

F.4 Supporting Technical Results

Lemma 24. *With probability at least $1-\delta$, for all $f,f'\in\mathcal{F}$ and all $g\in\mathcal{G}(f,f')$ such that $\sup_x|g(x)|\le 1$, it holds that*

$$\begin{aligned}
\|g\|_{\nu,h} &\le \sqrt{2}\|g\|_{\hat{\nu},h} + O\left(\varepsilon_{\text{stat}}\right), \\
\|g\|_{\hat{\nu},h} &\le \sqrt{2}\|g\|_{\nu,h} + O\left(\varepsilon_{\text{stat}}\right).
\end{aligned}$$

(Recall the definition of $\varepsilon_{\text{stat}}$ in Definition F.7)

Proof of Lemma 24. Fix f,f' and g. By Bernstein's inequality, with probability at least $1-\delta'$,

$$\begin{aligned}
\left|\|g\|_{\nu,h}^2 - \|g\|_{\hat{\nu},h}^2\right| &= \left|\frac{1}{n}\sum_{i=1}^{n'}\left(g(x_i)^2 - \sum_s \nu(x)g(x)^2\right)\right| \\
&\le O\left(\sqrt{\frac{1}{n}\sum_s \nu(x)g(x)^4\log(1/\delta')} + \frac{\log(1/\delta')}{n}\right)
\end{aligned}$$

$$\leq \sum_s \nu(x) g(x)^2 + O\left(\frac{\log(1/\delta')}{n}\right) \qquad \text{(AM-GM)}$$
$$= \|g\|_{\nu,h}^2 + O\left(\frac{\log(1/\delta')}{n}\right).$$

Rearranging this gives

$$\|g\|_{\nu,h}^2 \leq 2\|g\|_{\hat{\nu},h}^2 + O\left(\frac{\log(1/\delta')}{n}\right) \qquad \text{and} \qquad \|g\|_{\hat{\nu},h}^2 \leq 2\|g\|_{\nu,h}^2 + O\left(\frac{\log(1/\delta')}{n}\right).$$

Next, we take union bounds over f, f' and g. Notice that for every pair of (f, f'), the value of $g \in \mathcal{G}(f, f')$ on $x \in \mathcal{X}_h$ is determined by the values of $\{g(\phi)\}_{\phi \in \Phi_h(f,f')}$, where $g(\phi)$ is the value of $g(x)$ for $x \in \phi$. Therefore, an ε-net of $\mathcal{G}(f, f')$ on layer h can be constructed by discretizing each value of $\{g(\phi)\}_{\phi \in \Phi_h(f,f')}$, and its size is at most $(1/\varepsilon)^{O(|\Phi_h(f,f')|)} \leq (1/\varepsilon)^{O(\Phi_{\max})}$. It suffices to pick $\varepsilon = \frac{1}{n}$ and bound the discretization error by $O(\frac{1}{n})$. Overall, the union bound is taken over $|\mathcal{F}|^2 n^{\Phi_{\max}}$ instances. Therefore, we pick $\delta' = \frac{\delta}{|\mathcal{F}|^2 n^{\Phi_{\max}}}$, which gives

$$\|g\|_{\nu,h}^2 \leq 2|g\|_{\hat{\nu},h}^2 + O\left(\frac{\Phi_{\max} \log(n|\mathcal{F}|/\delta)}{n}\right), \quad \text{and}$$
$$\|g\|_{\hat{\nu},h}^2 \leq 2\|g\|_{\nu,h}^2 + O\left(\frac{\Phi_{\max} \log(n|\mathcal{F}|/\delta)}{n}\right).$$

Finally, taking square root on both sides and recalling that $\varepsilon_{\text{stat}} = \sqrt{\frac{\Phi_{\max} \log(n\Phi_{\max}|\mathcal{F}|/\delta)}{n}}$ finishes the proof.

■

Lemma 25. *With probability at least $1 - \delta$, for all $f, f' \in \mathcal{F}$, and $h \leq H$,*

$$\|\mathcal{T}_{\mathcal{G}(f,f')} f - \widehat{\mathcal{T}}_{\mathcal{G}(f,f')} f\|_{\nu,h} \leq \mathcal{O}\left(\varepsilon_{\text{stat}}\right).$$

Proof of Lemma 25 . We first fix the function pair f, f' and define additional notation. Let $\Phi = \Phi(f, f')$ and $\mathcal{G} = \mathcal{G}(f, f')$. For every $\phi \in \Phi_h$, define

$$Y(\phi) = \sum_{x \in \phi} \nu(x) \left(r(x, \pi_e) + \sum_{x'} T(x' \mid x, \pi_e) f(x')\right), \qquad Z(\phi) = \sum_{x \in \phi} \nu(x),$$
$$\widehat{Y}(\phi) = \frac{1}{n} \sum_{i=1}^{n'} \mathbb{I}\{x_i \in \phi\} \left(r(x_i, \pi_e) + f(x_i')\right), \qquad \widehat{Z}(\phi) = \frac{1}{n} \sum_{i=1}^{n'} \mathbb{I}\{x_i \in \phi\}.$$

Additionally, let $\delta' = \frac{\delta}{|\mathcal{F}|^2 \Phi_{\max}}$, $\varepsilon_0 = \frac{\log(1/\delta')}{n}$, and $\Phi_h' = \{\phi \in \Phi_h : \ Z(\phi) \geq \varepsilon_0\}$.
Using Lemma 22 together with definitions above, we get that for all $x \in \phi$,

$$\mathcal{T}_{\mathcal{G}} f(x) = \frac{Y(\phi)}{Z(\phi)} \quad \text{and} \quad \widehat{\mathcal{T}}_{\mathcal{G}} f(x) = \frac{\widehat{Y}(\phi)}{\widehat{Z}(\phi)}.$$

Thus, we have

$$\|\mathcal{T}_{\mathcal{G}} f - \widehat{\mathcal{T}}_{\mathcal{G}} f\|_{\nu,h}^2 \qquad (52)$$
$$= \sum_{x \in \mathcal{X}_h} \nu(x) \left(\mathcal{T}_{\mathcal{G}} f(x) - \widehat{\mathcal{T}}_{\mathcal{G}} f(x)\right)^2$$

$$= \sum_{\phi\in\Phi_h}\sum_{x\in\phi}\nu(x)\left(\frac{Y(\phi)}{Z(\phi)}-\frac{\widehat{Y}(\phi)}{\widehat{Z}(\phi)}\right)^2$$

$$\le \sum_{\phi\in\Phi'_h}\sum_{x\in\phi}\nu(x)\left(\frac{Y(\phi)}{Z(\phi)}-\frac{\widehat{Y}(\phi)}{\widehat{Z}(\phi)}\right)^2 + 4\sum_{\phi\in\Phi_h\setminus\Phi'_h}\sum_{x\in\phi}\nu(x)H^2 \qquad \left(\tfrac{Y(\phi)}{Z(\phi)}-\tfrac{\widehat{Y}(\phi)}{\widehat{Z}(\phi)}\in[-2,2]\right)$$

$$\le 2\sum_{\phi\in\Phi'_h}Z(\phi)\left(\frac{Y(\phi)}{Z(\phi)}-\frac{\widehat{Y}(\phi)}{Z(\phi)}\right)^2 + 2\sum_{\phi\in\Phi'_h}Z(\phi)\left(\frac{\widehat{Y}(\phi)}{Z(\phi)}-\frac{\widehat{Y}(\phi)}{\widehat{Z}(\phi)}\right)^2 + 4\sum_{\phi\in\Phi_h\setminus\Phi'_h}Z(\phi) \tag{53}$$

$$\le 2\sum_{\phi\in\Phi'_h}\frac{(Y(\phi)-\widehat{Y}(\phi))^2}{Z(\phi)} + 2\sum_{\phi\in\Phi'_h}\frac{\widehat{Y}(\phi)^2}{\widehat{Z}(\phi)^2}\frac{(Z(\phi)-\widehat{Z}(\phi))^2}{Z(\phi)} + 4|\Phi_h|\varepsilon_0$$

(because $Z(\phi)\le\varepsilon_0$ for all $\phi\in\Phi_h\setminus\Phi'_h$)

$$\le 2\sum_{\phi\in\Phi'_h}\frac{(Y(\phi)-\widehat{Y}(\phi))^2+(Z(\phi)-\widehat{Z}(\phi))^2}{Z(\phi)} + \frac{4\Phi_{\max}\log(1/\delta')}{n}, \tag{54}$$

where the last line follows by observing that $\widehat{Y}(\phi)/\widehat{Z}(\phi)\le 1$ for any $\phi\in\Phi$.
Next, using Bernstein's inequality for any $\phi\in\Phi_h$, with probability at least $1-\delta'$, we have

$$|Y(\phi)-\widehat{Y}(\phi)|\le O\left(\sqrt{\frac{Z(\phi)}{n}\log(1/\delta')}+\frac{\log(1/\delta')}{n}\right),$$
$$|Z(\phi)-\widehat{Z}(\phi)|\le O\left(\sqrt{\frac{Z(\phi)}{n}\log(1/\delta')}+\frac{\log(1/\delta')}{n}\right).$$

Plugging the above in (54), and using a union bound over $\phi\in\Phi'_h$, we get that with probability at least $1-\delta'\Phi_{\max}$,

$$\|\mathcal{T}_{\mathcal{G}}f-\widehat{\mathcal{T}}_{\mathcal{G}}f\|^2_{\nu,h}\le\mathcal{O}\left(\sum_{\phi\in\Phi'_h}\frac{\log(1/\delta')}{n}+\frac{\log^2(1/\delta')}{n^2Z(\phi)}+\frac{\Phi_{\max}\log(1/\delta')}{n}\right)$$
$$\le\mathcal{O}\left(\frac{\Phi_{\max}\log(1/\delta')}{n}+\frac{\Phi_{\max}\log^2(1/\delta')}{n^2\varepsilon_0}+\frac{\Phi_{\max}\log(1/\delta')}{n}\right) \quad \text{(for } \phi\in\Phi'_h,\ Z(\phi)\ge\varepsilon_0\text{)}$$
$$=\mathcal{O}\left(\frac{\Phi_{\max}\log(|\mathcal{F}|\Phi_{\max}/\delta)}{n}\right). \quad \text{(by the definition of } \varepsilon_0 \text{ and } \delta'\text{)}$$

Finally, using a union bound over $(f,f')\in\mathcal{F}\times\mathcal{F}$ and recalling that $\varepsilon_{\text{stat}}=\sqrt{\frac{\Phi_{\max}\log(n|\mathcal{F}|\Phi_{\max}/\delta)}{n}}$ finishes the proof. ■

Lemma 26. *Let $\widehat{f}$ be the output of Algorithm 4 and $f^\star$ be defined in Assumption F.1.*

(a) $\|\widehat{f}-\mathcal{T}_{\mathcal{G}(\widehat{f},f^\star)}\widehat{f}\|_{\nu,h}\le\mathcal{O}\left(\varepsilon_{\text{appr}}+\varepsilon_{\text{stat}}\right)$.

(b) $\|\mathcal{T}_{\mathcal{G}(\widehat{f},f^\star)}\widehat{f}-\mathcal{T}_{\mathcal{G}(\widehat{f},f^\star)}f^\star\|_{\bar{d},h}\le\|\widehat{f}-f^\star\|_{\bar{d},h+1}$.

(recall the definition of $\bar{d}$ in Definition F.3)

Proof of Lemma 26. We prove the two parts separately below.

(a) Note that

$$
\begin{aligned}
&\max_h \|\widehat{f} - \mathcal{T}_{\mathcal{G}(\widehat{f},f^\star)}\widehat{f}\|_{\nu,h} \\
&\quad \le \max_h \|\widehat{f} - \widehat{\mathcal{T}}_{\mathcal{G}(\widehat{f},f^\star)}\widehat{f}\|_{\nu,h} + \|\widehat{\mathcal{T}}_{\mathcal{G}(\widehat{f},f^\star)}\widehat{f} - \mathcal{T}_{\mathcal{G}(\widehat{f},f^\star)}\widehat{f}\|_{\nu,h} && \text{(triangle inequality)} \\
&\quad \le \max_{f'\in\mathcal{F}} \max_h \|\widehat{f} - \widehat{\mathcal{T}}_{\mathcal{G}(\hat{f},f')}\widehat{f}\|_{\nu,h} + O\left(\varepsilon_{\text{stat}}\right) && \text{(by Lemma 25)} \\
&\quad \le \sqrt{2}\max_{f'\in\mathcal{F}} \max_h \|\widehat{f} - \widehat{\mathcal{T}}_{\mathcal{G}(\hat{f},f')}\widehat{f}\|_{\hat{\nu},h} + O\left(\varepsilon_{\text{stat}}\right) && \text{(by Lemma 24)} \\
&\quad \le \sqrt{2}\max_{f'\in\mathcal{F}} \max_h \|f^\star - \widehat{\mathcal{T}}_{\mathcal{G}(f^\star,f')}f^\star\|_{\hat{\nu},h} + O\left(\varepsilon_{\text{stat}}\right) && \text{(by the choice of } \widehat{f} \text{ in (48))} \\
&\quad \le 2\max_{f'\in\mathcal{F}} \max_h \|f^\star - \widehat{\mathcal{T}}_{\mathcal{G}(f^\star,f')}f^\star\|_{\nu,h} + O\left(\varepsilon_{\text{stat}}\right) && \text{(by Lemma 24)} \\
&\quad \le 2\max_{f'\in\mathcal{F}} \max_h \|f^\star - \mathcal{T}_{\mathcal{G}(f^\star,f')}f^\star\|_{\nu,h} + O\left(\varepsilon_{\text{stat}}\right) && \text{(by Lemma 25)} \\
&\quad \le 2\max_h \|f^\star - \mathcal{T}f^\star\|_{\nu,h} + O\left(\varepsilon_{\text{stat}}\right). && \text{(by Lemma 23)}
\end{aligned}
$$

(b) For the ease of notation, let $\mathcal{G} = \mathcal{G}(\widehat{f}, f^\star)$. Notice that by Lemma 22,

$$
\mathcal{T}_{\mathcal{G}}\widehat{f}(\phi) - \mathcal{T}_{\mathcal{G}}f^\star(\phi) = \frac{\sum_{x\in\phi}\nu(x)\sum_{x'\in\mathcal{X}_{h+1}} T(x' \mid x, \pi_{\text{e}})\left(\widehat{f}(x') - f^\star(x')\right)}{\sum_{x\in\phi}\nu(x)}.
$$

Therefore,

$$
\begin{aligned}
&\|\mathcal{T}_{\mathcal{G}}\widehat{f} - \mathcal{T}_{\mathcal{G}}f^\star\|_{\bar{d},h}^2 \\
&= \sum_{\phi\in\Phi_h} \bar{d}(\phi)\left(\mathcal{T}_{\mathcal{G}}\widehat{f}(\phi) - \mathcal{T}_{\mathcal{G}}f^\star(\phi)\right)^2 \\
&= \sum_{\phi\in\Phi_h} \bar{d}(\phi)\left(\frac{\sum_{s\in\phi}\nu(x)\sum_{x'\in\mathcal{X}_{h+1}} T(x' \mid x, \pi_{\text{e}})\left(\widehat{f}(x') - f^\star(x')\right)}{\sum_{s\in\phi}\nu(x)}\right)^2 \\
&\le \sum_{x'\in\mathcal{X}_{h+1}}\sum_{\phi\in\Phi_h} \bar{d}(\phi)\left(\frac{\sum_{s\in\phi}\nu(x)T(x' \mid x, \pi_{\text{e}})}{\sum_{s\in\phi}\nu(x)}\right)\left(\widehat{f}(x') - f^\star(x')\right)^2 && \text{(Jensen's inequality)} \\
&= \sum_{\phi'\in\Phi_{h+1}}\sum_{x'\in\phi'}\sum_{\phi\in\Phi_h} \bar{d}(\phi)\left(\frac{\sum_{s\in\phi}\nu(x)T(x' \mid x, \pi_{\text{e}})}{\sum_{s\in\phi}\nu(x)}\right)\left(\widehat{f}(\phi') - f^\star(\phi')\right)^2 \\
&= \sum_{\phi'\in\Phi_{h+1}}\sum_{\phi\in\Phi_h} \bar{d}(\phi)\bar{T}(\phi' \mid \phi)\left(\widehat{f}(\phi') - f^\star(\phi')\right)^2 \\
&\le \sum_{\phi'\in\Phi_{h+1}} \bar{d}(\phi')\left(\widehat{f}(\phi') - f^\star(\phi')\right)^2 \\
&= \|\widehat{f} - f^\star\|_{\bar{d},h+1}^2.
\end{aligned}
$$

■

Lemma 27. *With the $\Phi^\star$ and $\bar{\mathsf{C}}_\varepsilon^\star$ defined in Definition F.3 and Definition F.4, there exist $\mathcal{J}_1, \ldots, \mathcal{J}_H$ such that*

- $\mathcal{J}_h \subset \Phi_h^\star$
- $\sum_{\phi\in\mathcal{J}_h} \bar{d}(\phi) < \varepsilon$

- $\max_h \max_{\phi\in\Phi_h\setminus\mathcal{J}_h} \frac{\bar{d}(\phi)}{\bar{\nu}(\phi)} \le \bar{\mathsf{C}}^\star_\varepsilon$

Proof of Lemma 27. Define

$$\mathcal{I}_h = \arg\max \left\{ \frac{\sum_{\phi\in\mathcal{I}} \bar{d}(\phi)}{\sum_{\phi\in\mathcal{I}} \bar{\nu}(\phi)} : \quad \mathcal{I} \subset \Phi_h^\star, \quad \sum_{\phi\in\mathcal{I}} \bar{d}(\phi) \ge \varepsilon \right\}. \tag{55}$$

If $\mathcal{I}_h$ has more than one solution, we pick one such that $|\mathcal{I}_h|$ is the smallest. By the definition of $\bar{\mathsf{C}}^\star_\varepsilon$, we know that $\frac{\sum_{\phi\in\mathcal{I}_h} \bar{d}(\phi)}{\sum_{\phi\in\mathcal{I}_h} \bar{\nu}(\phi)} \le \bar{\mathsf{C}}^\star_\varepsilon$ for all h.

Assume $\mathcal{I}_h = \{\phi_{h,1}, \ldots, \phi_{h,N_h}\}$ where $N_h = |\mathcal{I}_h|$, and assume without loss of generality that

$$\frac{\bar{d}(\phi_{h,1})}{\bar{\nu}(\phi_{h,1})} \ge \frac{\bar{d}(\phi_{h,2})}{\bar{\nu}(\phi_{h,2})} \ge \cdots \ge \frac{\bar{d}(\phi_{h,N_h})}{\bar{\nu}(\phi_{h,N_h})}.$$

If $N_h = 1$, it is easy to see that $\phi_{h,1} = \arg\max_{\phi\in\Phi_h^\star} \frac{\bar{d}(\phi)}{\bar{\nu}(\phi)}$. This is because if not, then

$$\left\{ \phi_{h,1}, \ \arg\max_{\phi\in\Phi_h^\star} \frac{\bar{d}(\phi)}{\bar{\nu}(\phi)} \right\}$$

will be a better solution for $\mathcal{I}_h$ in (55) than $\{\phi_{h,1}\}$, contradicting that $\mathcal{I}_h = \{\phi_{h,1}\}$. Thus, in this case, $\mathcal{I}_h = \left\{\arg\max_{\phi\in\Phi_h^\star} \frac{\bar{d}(\phi)}{\bar{\nu}(\phi)}\right\}$. Then choosing $\mathcal{J}_h = \varnothing$ satisfies all conditions in the lemma.

If $N_h \ge 2$, we define $\mathcal{J}_h = \mathcal{I}_h \setminus \{\phi_{h,N_h}\} = \{\phi_{h,1}, \ldots, \phi_{h,N_h-1}\}$. Below we verify that it satisfies the two inequalities in the lemma.

First, we prove $\sum_{\phi\in\mathcal{J}_h} \bar{d}(\phi) < \varepsilon$ by contradiction. Suppose that $\sum_{\phi\in\mathcal{J}_h} \bar{d}(\phi) \ge \varepsilon$. By the assumption that

$$\frac{\bar{d}(\phi_{h,1})}{\bar{\nu}(\phi_{h,1})} \ge \cdots \ge \frac{\bar{d}(\phi_{h,N_h})}{\bar{\nu}(\phi_{h,N_h})},$$

we have

$$\frac{\sum_{\phi\in\mathcal{J}_h} \bar{d}(\phi)}{\sum_{\phi\in\mathcal{J}_h} \bar{\nu}(\phi)} \ge \frac{\sum_{\phi\in\mathcal{I}_h} \bar{d}(\phi)}{\sum_{\phi\in\mathcal{I}_h} \bar{\nu}(\phi)},$$

and thus $\mathcal{J}_h$ is also a solution of (55). However, $|\mathcal{J}_h| < |\mathcal{I}_h|$, contradicting the assumption that $|\mathcal{I}_h|$ is the smallest.

Next, we prove $\max_h \max_{\phi\in\Phi_h^\star\setminus\mathcal{J}_h} \frac{\bar{d}(\phi)}{\bar{\nu}(\phi)} \le \bar{\mathsf{C}}^\star_\varepsilon$. Define

$$\phi'_h := \arg\max_{\phi\in\Phi_h^\star\setminus\mathcal{J}_h} \frac{\bar{d}(\phi)}{\bar{\nu}(\phi)}.$$

If $\phi'_h = \phi_{h,N_h}$, then we have

$$\frac{\bar{d}(\phi'_h)}{\bar{\nu}(\phi'_h)} = \frac{\bar{d}(\phi_{h,N_h})}{\bar{\nu}(\phi_{h,N_h})} \le \frac{\sum_{\phi\in\mathcal{I}_h} \bar{d}(\phi)}{\sum_{\phi\in\mathcal{I}_h} \bar{\nu}(\phi)} \le \bar{\mathsf{C}}^\star_\varepsilon.$$

If $\phi'_h \ne \phi_{h,N_h}$ and $\frac{\bar{d}(\phi'_h)}{\bar{\nu}(\phi'_h)} > \bar{\mathsf{C}}^\star_\varepsilon$, then we have

$$\frac{\sum_{\phi\in\mathcal{I}_h} \bar{d}(\phi) + \bar{d}(\phi'_h)}{\sum_{\phi\in\mathcal{I}_h} \bar{\nu}(\phi) + \bar{\mu}(\phi'_h)} > \frac{\sum_{\phi\in\mathcal{I}_h} \bar{d}(\phi)}{\sum_{\phi\in\mathcal{I}_h} \bar{\nu}(\phi)}$$

because $\frac{\sum_{\phi\in\mathcal{I}_h} \bar{d}(\phi)}{\sum_{\phi\in\mathcal{I}_h} \bar{\nu}(\phi)} \le \bar{\mathsf{C}}^\star_\varepsilon$. This implies that $\mathcal{I}_h \cup \{\phi'_h\}$ is a better solution than $\mathcal{I}_h$ in (55), which is a contradiction. This concludes the proof. ∎

F.5 Proof of Theorem 5.1

Lemma 28. *Let* $\bar{\mathsf{C}}^\star = \bar{\mathsf{C}}^\star_{\varepsilon^2/H^2}$. *With probability* $\geq 1-\delta$,

$$\mathbb{E}_{x\sim\rho}[|\widehat{f}(x) - f^\star(x)|] \leq \mathcal{O}\left(H\varepsilon_{\mathrm{appr}}\sqrt{\bar{\mathsf{C}}^\star} + H\sqrt{\frac{\bar{\mathsf{C}}^\star \Phi_{\max}\log(n\Phi_{\max}|\mathcal{F}|/\delta)}{n}} + \varepsilon\right).$$

Proof of Lemma 28. In this proof, we denote $\mathcal{G} = \mathcal{G}(\widehat{f}, f^\star)$. Recall the definitions of $\Phi^\star$, $\bar{d}$, and $\bar{\nu}$ in Definition F.3. Notice that $\widehat{f}, f^\star, \mathcal{T}_{\mathcal{G}}\widehat{f}, \mathcal{T}_{\mathcal{G}}f^\star \in \mathcal{G}$. For any $g \in \mathcal{G}$ and $\phi \in \Phi^\star$, we use $g(\phi)$ to represent the value of $g(x)$ for those $x \in \phi$. Using Definition F.6, we have

$$\|\widehat{f} - f^\star\|_{\bar{d},h} \leq \|f^\star - \mathcal{T}_{\mathcal{G}}f^\star\|_{\bar{d},h} + \|\widehat{f} - \mathcal{T}_{\mathcal{G}}\widehat{f}\|_{\bar{d},h} + \|\mathcal{T}_{\mathcal{G}}\widehat{f} - \mathcal{T}_{\mathcal{G}}f^\star\|_{\bar{d},h}. \tag{56}$$

Before bounding the three terms above, we notice that by Lemma 27, there exist $\{\mathcal{J}_h\}_{h\in[H]}$ such that

- $\mathcal{J}_h \subset \Phi^\star_h$
- $\sum_{\phi\in\mathcal{J}_h} \bar{d}(\phi) < \frac{\varepsilon^2}{H^2}$
- $\max_h \max_{\phi\in\Phi^\star_h\setminus\mathcal{J}_h} \frac{\bar{d}(\phi)}{\bar{\nu}(\phi)} \leq \bar{\mathsf{C}}^\star_{\varepsilon^2/H^2}$

Below we denote $\bar{\mathsf{C}}^\star = \bar{\mathsf{C}}^\star_{\varepsilon^2/H^2}$. Now we bound the three terms in (56). The square of the first term can be upper bounded as

$$\begin{aligned}
\|f^\star - \mathcal{T}_{\mathcal{G}}f^\star\|^2_{\bar{d},h} &= \sum_{\phi\in\Phi^\star_h} \bar{d}(\phi)\left(f^\star(\phi) - \mathcal{T}_{\mathcal{G}}f^\star(\phi)\right)^2 \\
&\leq \sum_{\phi\in\Phi^\star_h\setminus\mathcal{J}_h} \bar{d}(\phi)\left(f^\star(\phi) - \mathcal{T}_{\mathcal{G}}f^\star(\phi)\right)^2 + \frac{\varepsilon^2}{H^2} && (\textstyle\sum_{\phi\in\mathcal{J}_h}\bar{d}(\phi) \leq \frac{\varepsilon^2}{H^2}) \\
&\leq \bar{\mathsf{C}}^\star \sum_{\phi\in\Phi^\star_h\setminus\mathcal{J}_h} \bar{\nu}(\phi)\left(f^\star(\phi) - \mathcal{T}_{\mathcal{G}}f^\star(\phi)\right)^2 + \frac{\varepsilon^2}{H^2} && (\text{by the definition of } \bar{\mathsf{C}}) \\
&\leq \bar{\mathsf{C}}^\star \|f^\star - \mathcal{T}_{\mathcal{G}}f^\star\|^2_{\bar{\nu},h} + \frac{\varepsilon^2}{H^2} \\
&\leq \bar{\mathsf{C}}^\star \|f^\star - \mathcal{T}_{\mathcal{G}}f^\star\|^2_{\nu,h} + \frac{\varepsilon^2}{H^2} && (\text{because } f^\star, \mathcal{T}_{\mathcal{G}}f^\star \in \mathcal{G}) \\
&\leq \bar{\mathsf{C}}\|f^\star - \mathcal{T}f^\star\|^2_{\nu,h} + \frac{\varepsilon^2}{H^2} && (\text{by Lemma 23}) \\
&\leq \bar{\mathsf{C}}^\star\varepsilon^2_{\mathrm{appr}} + \frac{\varepsilon^2}{H^2}. && (\text{by the definition of } \varepsilon_{\mathrm{appr}})
\end{aligned}$$

The square of the second term can be upper bounded as

$$\begin{aligned}
\|\widehat{f} - \mathcal{T}_{\mathcal{G}}\widehat{f}\|^2_{\bar{d},h} &= \sum_{\phi\in\Phi^\star_h} \bar{d}(\phi)\left(\widehat{f}(\phi) - \mathcal{T}_{\mathcal{G}}\widehat{f}(\phi)\right)^2 \\
&= \sum_{\phi\in\Phi^\star_h\setminus\mathcal{J}_h} \bar{d}(\phi)\left(\widehat{f}(\phi) - \mathcal{T}_{\mathcal{G}}\widehat{f}(\phi)\right)^2 + \frac{\varepsilon^2}{H^2} \\
&\leq \bar{\mathsf{C}}^\star \sum_{\phi\in\Phi^\star_h\setminus\mathcal{J}_h} \bar{\nu}(\phi)\left(\widehat{f}(\phi) - \mathcal{T}_{\mathcal{G}}\widehat{f}(\phi)\right)^2 + \frac{\varepsilon^2}{H^2} && (\text{by the definition of } \bar{\mathsf{C}}^\star)
\end{aligned}$$

$$
\begin{aligned}
&\le \bar{\mathsf{C}}^\star \|\widehat{f} - \mathcal{T}_{\mathcal{G}}\widehat{f}\|_{\bar{\nu},h}^2 + \frac{\varepsilon^2}{H^2} \\
&= \bar{\mathsf{C}}^\star \|\widehat{f} - \mathcal{T}_{\mathcal{G}}\widehat{f}\|_{\nu,h}^2 + \frac{\varepsilon^2}{H^2} && \text{(because } \widehat{f}, \mathcal{T}_{\mathcal{G}}\widehat{f} \in \mathcal{G}\text{)} \\
&\le \mathcal{O}\left(\bar{\mathsf{C}}^\star \varepsilon_{\text{appr}}^2 + \bar{\mathsf{C}}^\star \varepsilon_{\text{stat}}^2\right) + \frac{\varepsilon^2}{H^2}. && \text{(by Lemma 26-(a))}
\end{aligned}
$$

Next, again using Lemma 26-(b), the last term can be upper bounded as

$$
\|\mathcal{T}_{\mathcal{G}}\widehat{f} - \mathcal{T}_{\mathcal{G}} f^\star\|_{\bar{d},h} \le \|\widehat{f} - f^\star\|_{\bar{d},h+1}.
$$

Combining all above, we get

$$
\|\widehat{f} - f^\star\|_{\bar{d},h} \le \|\widehat{f} - f^\star\|_{\bar{d},h+1} + \mathcal{O}\left(\sqrt{\bar{\mathsf{C}}^\star}\varepsilon_{\text{appr}} + \sqrt{\bar{\mathsf{C}}^\star}\varepsilon_{\text{stat}} + \frac{\varepsilon}{H}\right),
$$

which gives $\|\widehat{f} - f^\star\|_{\bar{d},1} \le \mathcal{O}\left(H\sqrt{\bar{\mathsf{C}}^\star}\varepsilon_{\text{appr}} + H\sqrt{\bar{\mathsf{C}}^\star}\varepsilon_{\text{stat}} + \varepsilon\right)$ after recursively applying the inequality and using Cauchy-Schwarz inequality. The desired inequality follows by noticing that $\mathbb{E}_{x\sim\rho}[|\widehat{f}(x) - f^\star(x)|] \le \|\widehat{f} - f^\star\|_{\bar{d},1}$ by Cauchy-Schwarz. ■

Proof of Theorem F.1 (Theorem 5.1 in the main body). In the *fully realizable* setting (i.e., $V^{\pi_e} \in \mathcal{F}$), in order to control the magnitude of $\Phi_{\max}$, we discretize the function set $\mathcal{F}$, and make it an *approximate realizable* case.

For each $f \in \mathcal{F}$, we round the value of $f(s)$ to the nearest multiple of $\frac{\varepsilon}{H\sqrt{\bar{\mathsf{C}}}}$. This way, we have $\varepsilon_{\text{appr}} = \frac{\varepsilon}{H\sqrt{\bar{\mathsf{C}}}}$ and $\Phi_{\max} = O\left(\left(\frac{H}{\varepsilon_{\text{appr}}}\right)^2\right) = O\left(\frac{H^4\bar{\mathsf{C}}}{\varepsilon^2}\right)$. Thus, by Lemma 28,

$$
\begin{aligned}
\|\widehat{f} - f^\star\|_{\bar{d},1} &\le \mathcal{O}\left(H\sqrt{\bar{\mathsf{C}}^\star}\varepsilon_{\text{appr}} + H\sqrt{\frac{\bar{\mathsf{C}}^\star \Phi_{\max}\log(n|\mathcal{F}|\Phi_{\max}/\delta)}{n}} + \varepsilon\right) \\
&\le \mathcal{O}\left(H\sqrt{\frac{\bar{\mathsf{C}}^2 H^4 \log(n|\mathcal{F}|H\bar{\mathsf{C}}/\varepsilon\delta)}{\varepsilon^2 n}} + \varepsilon\right).
\end{aligned}
$$

In order to make the last expression to be $\mathcal{O}(\varepsilon)$, we need

$$
n \ge \widetilde{\Omega}\left(\frac{\bar{\mathsf{C}}^2 H^6 \log(|\mathcal{F}|/\delta)}{\varepsilon^4}\right).
$$

■

F.6 Implications of the BVFT Upper Bound

Lemma 29. *For any ε, $\bar{\mathsf{C}}_\varepsilon \le \mathsf{C}_{\text{pf}}$, where C_{pf} is defined in Definition B.1.*

Proof of Lemma 29. By the definition of $\bar{\mathsf{C}}_\varepsilon$,

$$
\begin{aligned}
\bar{\mathsf{C}}_\varepsilon &\le \max_h \max_{\phi\in\Phi_h} \frac{\bar{d}(\phi)}{\bar{\nu}(\phi)} \\
&= \max_h \max_{\phi\in\Phi_h} \frac{\sum_{\phi'\in\Phi_{h-1}} \bar{d}(\phi')\bar{T}(\phi \mid \phi', \pi_e)}{\sum_{x\in\phi}\mu(x, \pi_e(x))} && \text{(by the definitions of } \bar{d} \text{ and } \bar{\nu} \text{ and } \nu\text{)}
\end{aligned}
$$

$$
\begin{aligned}
&= \max_h \max_{\phi \in \Phi_h} \frac{\sum_{\phi' \in \Phi_{h-1}} \bar{d}(\phi') \bar{T}(\phi \mid \phi', \pi_e)}{\sum_{x \in \phi} \mu(x) \mu(\pi_e(x) \mid x)} && \text{(by the definition of } \mu(a \mid x)) \\
&\le \max_h \max_{\phi \in \Phi_h} \frac{\sum_{\phi' \in \Phi_{h-1}} \bar{d}(\phi') \bar{T}(\phi \mid \phi', \pi_e)}{\sum_{x \in \phi} \mu(x)} \cdot \mathsf{C}_{\mathcal{A}} \\
&\le \max_h \max_{\phi \in \Phi_h} \frac{1}{\sum_{x \in \phi} \mu(x)} \sum_{\phi' \in \Phi_{h-1}} \bar{d}(\phi') \frac{\sum_{x' \in \phi'} \nu(x') \sum_{x \in \phi} T(x \mid x', \pi_e)}{\sum_{x' \in \phi'} \nu(x')} \cdot \mathsf{C}_{\mathcal{A}} && \text{(by the definition of } \bar{T}) \\
&\le \max_h \max_{\phi \in \Phi_h} \max_{x' \in \mathcal{X}_{h-1}} \frac{\sum_{x \in \phi} T(x \mid x', \pi_e)}{\sum_{x \in \phi} \mu(x)} \cdot \mathsf{C}_{\mathcal{A}} \\
&\le \max_h \max_{x \in \mathcal{X}_h} \max_{x' \in \mathcal{X}_{h-1}} \frac{T(x \mid x', \pi_e)}{\mu(x)} \cdot \mathsf{C}_{\mathcal{A}} \\
&\le \mathsf{C}_{\mathcal{X}} \cdot \mathsf{C}_{\mathcal{A}} \\
&= \mathsf{C}_{\mathrm{pf}}.
\end{aligned}
$$

■

Lemma 30. *Let the offline data distribution μ be admissible, i.e., $\mu(x,a) = d^{\pi_b}(x)\pi_b(a \mid x)$ for some π_b. Suppose that $\frac{1}{\mu(\pi_e(x)|x)} = \frac{1}{\pi_b(\pi_e(x)|x)} \le \mathsf{C}_{\mathcal{A}}$ for all $x \in \mathcal{X}$. Then for any ε, $\bar{\mathsf{C}}_\varepsilon \le (\mathsf{C}_{\mathcal{A}})^H$.*

Proof of Lemma 30. For a fixed h, we have

$$
\begin{aligned}
&\max_{\phi \in \Phi_h} \frac{\bar{d}(\phi)}{\bar{\nu}(\phi)} \\
&= \max_{\phi \in \Phi_h} \frac{\sum_{\phi' \in \Phi_{h-1}} \bar{d}(\phi') \bar{T}(\phi \mid \phi', \pi_e)}{\sum_{x \in \phi} \mu(x, \pi_e(x))} && \text{(by the definitions of } \bar{d} \text{ and } \bar{\nu} \text{ and } \nu) \\
&= \max_{\phi \in \Phi_h} \frac{\sum_{\phi' \in \Phi_{h-1}} \bar{d}(\phi') \bar{T}(\phi \mid \phi', \pi_e)}{\sum_{x \in \phi} \mu(x) \mu(\pi_e(x) \mid x)} && \text{(by the definition of } \mu(a \mid x)) \\
&= \max_{\phi \in \Phi_h} \frac{\sum_{\phi' \in \Phi_{h-1}} \bar{d}(\phi') \bar{T}(\phi \mid \phi', \pi_e)}{\sum_{x \in \phi} \sum_{x' \in \mathcal{X}_{h-1}} \sum_{a' \in \mathcal{A}} \mu(x', a') T(x \mid x', a') \mu(\pi_e(x) \mid x)} \\
& && \text{(by the fact that } \mu \text{ is an occupancy measure)} \\
&\le \max_{\phi \in \Phi_h} \frac{\sum_{\phi' \in \Phi_{h-1}} \bar{d}(\phi') \bar{T}(\phi \mid \phi', \pi_e)}{\sum_{x \in \phi} \sum_{x' \in \mathcal{X}_{h-1}} \mu(x', \pi_e(x')) T(x \mid x', \pi_e(x'))} \cdot \mathsf{C}_{\mathcal{A}} \\
&= \max_{\phi \in \Phi_h} \frac{\sum_{\phi' \in \Phi_{h-1}} \bar{d}(\phi') \bar{T}(\phi \mid \phi', \pi_e)}{\sum_{\phi' \in \Phi_{h-1}} \sum_{x \in \phi} \sum_{x' \in \phi'} \mu(x', \pi_e(x')) T(x \mid x', \pi_e(x'))} \cdot \mathsf{C}_{\mathcal{A}} \\
&= \max_{\phi \in \Phi_h} \frac{\sum_{\phi' \in \Phi_{h-1}} \bar{d}(\phi') \bar{T}(\phi \mid \phi', \pi_e)}{\sum_{\phi' \in \Phi_{h-1}} \left(\sum_{x' \in \phi'} \mu(x', \pi_e(x')) \right) \bar{T}(\phi \mid \phi', \pi_e)} \cdot \mathsf{C}_{\mathcal{A}} && \text{(by the definition of } \bar{T}) \\
&= \max_{\phi \in \Phi_h} \frac{\sum_{\phi' \in \Phi_{h-1}} \bar{d}(\phi') \bar{T}(\phi \mid \phi', \pi_e)}{\sum_{\phi' \in \Phi_{h-1}} \bar{\nu}(\phi') \bar{T}(\phi \mid \phi', \pi_e)} \cdot \mathsf{C}_{\mathcal{A}} && \text{(by the definition of } \bar{\nu}) \\
&\le \max_{\phi' \in \Phi_{h-1}} \frac{\bar{d}(\phi')}{\bar{\nu}(\phi')} \cdot \mathsf{C}_{\mathcal{A}}.
\end{aligned}
$$

Recursively applying this we get $\max_h \max_{\phi \in \Phi_h} \frac{\bar{d}(\phi)}{\bar{\nu}(\phi)} \le (\mathsf{C}_{\mathcal{A}})^H$. Then by just noticing that $\bar{\mathsf{C}}_\varepsilon \le \max_h \max_{\phi \in \Phi_h} \frac{\bar{d}(\phi)}{\bar{\nu}(\phi)}$ finishes the proof. ■

G Role of Realizable Value Function Class in Offline RL

So far, we have considered offline policy evaluation problems where in addition to an offline data distribution μ (that satisfies concentrability), the learner is also given a function class $\mathcal{F}$ that contains Q^{π_e}—the state-action value function corresponding to the policy π_e that the learner wishes to evaluate. To understand the role of value function class in offline RL, in this section, we ask:

Is statistically efficient offline policy evaluation feasible without access to a realizable value function class?

We answer this question negatively for both admissible and trajectory data. Our first result in Theorem G.1 (below) shows that given only admissible offline data, offline policy evaluation is intractable without a realizable value function class, even when we have bounded pushforward concentrability coefficient.

Theorem G.1. *For any positive integer N, there exists a class $\mathcal{M}$ of* MDP*s with shared state space $\mathcal{X}_N$, action space $\mathcal{A} = \{\mathfrak{a}_1, \mathfrak{a}_2\}$ and horizon $H = 3$, a deterministic evaluation policy π_e, and an exploration policy π_b with $\Pr(\pi_e(x) = \pi_b(x)) \geq 1/2$ for all $x \in \mathcal{X}$ such that any algorithm that estimate the value $V^{\pi_e}(\rho, M)$ up to error $1/2$ for all* MDP*s $M \in \mathcal{M}$ must use $\Omega(N)$ many admissible samples in some* MDP *in $\mathcal{M}$.*

Theorem G.1 suggests the intractability of offline policy evaluation without a realizable function class since the result holds for any positive integer N.

Remark 1. *Note that since $H = 3$ in the construction of Theorem G.1, the property $\Pr(\pi_e(x) = \pi_b(x)) \geq 1/2$ indicates that the pushforward concentrability coefficient $\mathsf{C}_{\mathrm{pf}} \leq 8$ w.r.t. the admissible distribution $\mu_h(x, a; M) = d^{\pi_b}(x, a; M)$ (see Definition B.1). Thus, using Lemma 29, the aggregated concentrability coefficient $\bar{\mathsf{C}} \leq 8$ for any aggregation scheme on the underlying MDPs.*

On the other hand, under access to a realizable state-action value function class (Assumption F.1), BVFT algorithm (Theorem F.1 + Lemma 30) obtains a sample complexity upper bound of $O(\mathrm{poly}(2^H, \log(|\mathcal{F}|)))$ which is tractable for $H = 3$. This highlights the role of a realizable value function class in offline policy evaluation with admissible data. Our next result extends this to trajectory offline data.

Theorem G.2. *There exists a class $\mathcal{M}$ of* MDP*s with shared state space $\mathcal{X}$, action space $\mathcal{A} = \{\mathfrak{a}_1, \mathfrak{a}_2\}$, and horizon H, a deterministic evaluation policy π_e and an exploration policy π_b such that the pushforward concentrability coefficient $\mathsf{C}_{\mathrm{pf}} \leq 4$ for the offline distribution $\mu_h(x, a; M) = d_h^{\pi_b}(x, a; M)$ (see Definition B.1). Furthermore, any algorithm that estimate the value $V^{\pi_e}(\rho, M)$ up to precision $1/2$ for all* MDP*s $M \in \mathcal{M}$ must use $\Omega(2^H)$ many offline trajectories in some* MDP *in $\mathcal{M}$.*

The above shows that agnostic offline policy evaluation is not statistically tractable even when given trajectory offline data. On the other hand, recall that under access to a realizable state-action value function class (Assumption F.1) and bounded pushforward concentrability coefficient, the BVFT algorithm in Xie and Jiang (2021) enjoys a $\mathrm{poly}(\mathsf{C}_{\mathrm{pf}}, H, \log(|\mathcal{F}|), 1/\varepsilon)$ sample complexity (even without access to trajectory data).

Remark 2. *For the lower bound* MDP *construction in Theorem G.2, recall that $\pi_b(x) = \mathrm{Uniform}(\mathcal{A})$ for any $x \in \mathcal{X}$. Thus, given trajectory data, the classical importance sampling algorithm from Kearns et al. (1999); Agarwal et al. (2019) can evaluate the value of π_e upto precision ε after collecting $\mathcal{O}\left(\frac{2^H}{\varepsilon^2}\right)$ many offline trajectories from π_b.*

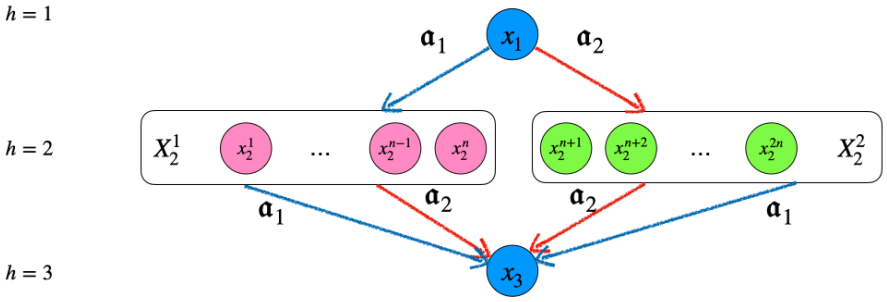


Figure 5: Lower bound construction in Theorem G.1. The blue arrows represent the transitions under the action $\mathfrak{a}_1$, and the red arrows represent the transitions under the action $\mathfrak{a}_2$. In the middle layer, the arrows to the blocks $\mathcal{X}_2^1$ and $\mathcal{X}_2^2$ denote uniform transitions to the states within those blocks.

G.1 Proof of Lower Bounds

To avoid redundancy, we only provide an informal construction of the lower bound here. A formal lower bound construction can be obtained by following arguments similar to that in Appendix C.

Proof of Theorem G.1. Let N be a positive integer, and consider a state space $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \mathcal{X}_3$ where $\mathcal{X}_1 = \{x_1\}$, $\mathcal{X}_3 = \{x_3\}$ and $\mathcal{X}_2$ is of size $4N^2$. Consider a partition ϕ of $\mathcal{X}_2$ that divides it into two parts $\mathcal{X}_2^1$ and $\mathcal{X}_2^2$, each of size $2N^2$. For any such partition ϕ, we define two MDPs $M^{(1)}$ and $M^{(2)}$, where the $M_\phi^{(i)} = MDP(\mathcal{X}, \mathcal{A}, H, T^{(i)}, r^{(i)}, \rho)$ is defined such that (this construction can be viewed in Figure 5):

- Horizon $H = 3$, action space $\mathcal{A} = \{\mathfrak{a}_1, \mathfrak{a}_2\}$, and initial distribution $\rho = \delta_{x_1}(\cdot)$.
- Transition dynamics $T^{(i)}$ is defined such that $T^{(i)}(\cdot \mid x_2) = \delta_{x_3}(\cdot)$ for any $x_2 \in \mathcal{X}_2$, and

$$T^{(i)}(\cdot \mid x_1, a) = \begin{cases} \text{Uniform}(\mathcal{X}_2^1) & \text{if} \quad a = \mathfrak{a}_1 \\ \text{Uniform}(\mathcal{X}_2^2) & \text{if} \quad a = \mathfrak{a}_2 \end{cases}.$$

- Reward function is defined for any $a \in \{\mathfrak{a}_1, \mathfrak{a}_2\}$,

$$r^{(i)}(x, a) = \begin{cases} 0 & \text{if} \quad x = x_1 \\ 1 & \text{if} \quad x \in \mathcal{X}_2^{(i)} \\ 0 & \text{otherwise} \end{cases}.$$

We thus define the class $\mathcal{M}$ as

$$\mathcal{M} = \bigcup_{\phi \in \Phi} (M_\phi^{(1)}, M_\phi^{(2)}),$$

where Φ denotes the set of all feasible partitions for $\mathcal{X}_2$ into $\mathcal{X}_2^1$ and $\mathcal{X}_2^2$ of the same size, and satisfies $|\Phi| = 2^{O(N \log(N))}$.

We further define the evaluation policy π_e and π_b such that for all $x \in \mathcal{X}$,

$$\pi_e(x) = \delta_{\mathfrak{a}_1}(\cdot), \qquad \text{and} \qquad \pi_b(x) = \text{Uniform}(\{\mathfrak{a}_1, \mathfrak{a}_2\}).$$

Then for any $M \in \cup_{\phi \in \Phi} \mathcal{M}^{(1)}$, we have $V^{\pi_e}(\rho; M) = 1$ and for any $M \in \cup_{\phi \in \Phi} \mathcal{M}^{(2)}$, we have $V^{\pi_e}(\rho; M) = 0$. Hence if the algorithm cannot tell whether $M \in \cup_{\phi \in \Phi} \mathcal{M}^{(2)}$ or $M \in \cup_{\phi \in \Phi} \mathcal{M}^{(2)}$, then the algorithm must fail to output $1/2$-accurate estimation in at least one case among $\mathcal{M}$.

The key intuition behind the proof is that given some dataset $\mathcal{D}_h = \{(x_h, a_h, r_h, x_{h+1})\}$, the learner can not identify which states belong to $\mathcal{X}_2^1$ vs $\mathcal{X}_2^2$ in the second layer. Since, the reward model depends on whether $x \in \mathcal{X}_2^1$ or $x \in \mathcal{X}_2^2$, inability to identify the partition ϕ which split $\mathcal{X}$ into $\mathcal{X}_2^1$ and $\mathcal{X}_2^2$, will lead to an error in evaluating $V^{\pi_e}(\rho; M)$ with probability $1/2$ since we consider Φ to be the set of all possible partitions of equal size.

To see the above, note that for any $M \in \mathcal{M}$, the marginal occupancy measure at layer 2 is

$$d_2^{\pi_b}(\cdot; M) = \text{Uniform}(\mathcal{X}_2 \times \{\mathfrak{a}_1, \mathfrak{a}_2\}).$$

Hence, samples (x_2, a_2, r_2, x_3) where $x_2 \in \mathcal{X}_2$ cannot provide useful any information unless we know whether $x_2 \in \mathcal{X}_2^1$ or $\mathcal{X}_2^2$. However, while collecting admissible samples of the second layer, the distribution we samples from is $d_2^{\pi_b}(\cdot; M)$, i.e. $\text{Uniform}(\mathcal{X}_2)$, which reveals no information on $\mathcal{X}_2^1$ and $\mathcal{X}_2^2$ unless some state x_2 appears both in samples (x_1, a_1, r_1, x_2) and (x_2, a_2, r_2). According to our choice of N, this happens with probability at most

$$1 - \prod_{i=1}^{2N} \left(1 - \frac{2N}{4N^2}\right)^N \le \frac{1}{2}.$$

Hence any algorithm must fail to output $1/2$-accurate estimation of $V^{\pi_e}(\rho; M)$ in at least one $M \in \mathcal{M}$ with probability at least $1/2$. ■

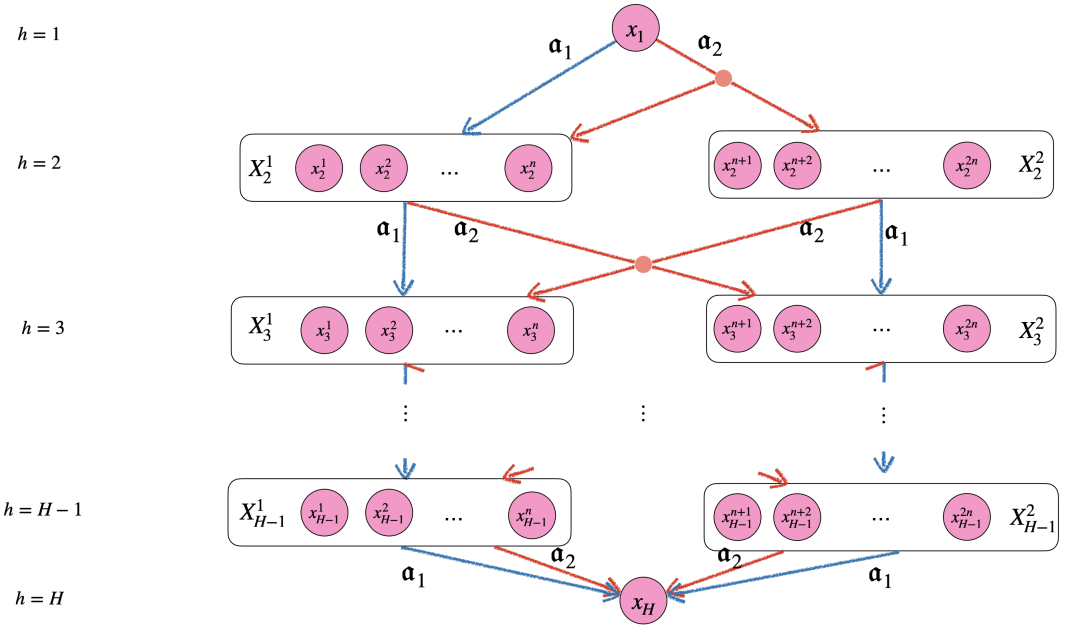


Figure 6: Lower bound construction in Theorem G.2. The blue arrows represent the transitions under the action $\mathfrak{a}_1$, and the red arrows represent the transitions under the action $\mathfrak{a}_2$. In layers for $h = 2$ to $H - 1$, the arrows to the blocks $\mathcal{X}_h^1$ and $\mathcal{X}_h^2$ denote uniform transitions to the states within those blocks.

Proof of Theorem G.2. Let $N = 2^H$ and consider a state space $\mathcal{X} = \cup_{h=1}^{H} \mathcal{X}_h$ where $\mathcal{X}_1 = \{x_1\}$ and $\mathcal{X}_2, \cdots, \mathcal{X}_H$ are of size $4N^2$ each. Consider ϕ to be a partition of $\mathcal{X}_2, \mathcal{X}_3, \cdots, \mathcal{X}_H$ that divides each $\mathcal{X}_h$ ($2 \le h \le H$) into two parts $\mathcal{X}_h^1$ and $\mathcal{X}_h^2$, each of size $2N^2$. For any such ϕ, we define two MDPs $M^{(1)}$ and $M^{(2)}$, where the $M_\phi^{(i)} = MDP(\mathcal{X}, \mathcal{A}, H, T^{(i)}, r^{(i)}, \rho)$ is defined such that (this construction can be viewed in Figure 6):

- Horizon H, action space $\mathcal{A} = \{\mathfrak{a}_1, \mathfrak{a}_2\}$, and initial distribution $\rho = \delta(x_1)$.
- Transition dynamics $T^{(i)} : \mathcal{X}_h \times \mathcal{A} \mapsto \mathcal{X}_{h+1}$ is defined such that

$$T^{(i)}(\cdot \mid x_1, a) = \begin{cases} \text{Uniform}(\mathcal{X}_2^1) & \text{if} \quad a = \mathfrak{a}_1 \\ \text{Uniform}(\mathcal{X}_2^1 \cup \mathcal{X}_2^2) & \text{if} \quad a = \mathfrak{a}_2, \end{cases}$$

  and for $x \in \mathcal{X}_h^1$ with $h \in [H]$,

$$T^{(i)}(\cdot \mid x, a) = \begin{cases} \text{Uniform}(\mathcal{X}_2^1) & \text{if} \quad a = \mathfrak{a}_1 \\ \text{Uniform}(\mathcal{X}_2^1 \cup \mathcal{X}_2^2) & \text{if} \quad a = \mathfrak{a}_2, \end{cases}$$

  for $x \in \mathcal{X}_h^2$ with $h \in [H]$,

$$T^{(i)}(\cdot \mid x, a) = \begin{cases} \text{Uniform}(\mathcal{X}_2^2) & \text{if} \quad a = \mathfrak{a}_1 \\ \text{Uniform}(\mathcal{X}_2^1 \cup \mathcal{X}_2^2) & \text{if} \quad a = \mathfrak{a}_2. \end{cases}$$

- Reward function: for any $a \in \{\mathfrak{a}_1, \mathfrak{a}_2\}$,

$$r^{(i)}(x, a) = \begin{cases} 1 & \text{if} \quad x \in \mathcal{X}_H^{(i)} \\ 0 & \text{otherwise} \end{cases}.$$

We thus define the class $\mathcal{M}$ as

$$\mathcal{M} = \bigcup_{\phi \in \Phi} (M_\phi^{(1)}, M_\phi^{(2)}),$$

where Φ denotes the set of all feasible partitions for $\mathcal{X}_2, \cdots, \mathcal{X}_H$ into $\mathcal{X}_h^1$ and $\mathcal{X}_h^2$ of the same size. We further define the evaluation policy π_e and π_b such that for all $x \in \mathcal{X}$,

$$\pi_e(x) = \delta_{\mathfrak{a}_1}(\cdot), \qquad \text{and} \qquad \pi_b(x) = \text{Uniform}(\{\mathfrak{a}_1, \mathfrak{a}_2\}).$$

For any MDP $M \in \mathcal{M}$, we observe that the occupancy measure $d_h^{\pi_b}(\cdot; M)$ is

$$d_h^{\pi_b}(\cdot; M) = \text{Uniform}(\mathcal{X}_h \times \mathcal{A}).$$

Hence it is easy to verify that the strong coverability of any instances in $\mathcal{M}$ is upper bounded by 4. Additionally, we also have for any MDP $M \in \bigcup_{\phi \in \Phi} M_\phi^{(1)}$, $V^{\pi_e}(\rho; M) = 1$ and for any $M \in \bigcup_{\phi \in \Phi} M_\phi^{(1)}$, $V^{\pi_e}(\rho; M) = 0$.

The only way to tell whether a case $M \in \bigcup_{\phi \in \Phi} M_\phi^{(1)}$ or $M \in \bigcup_{\phi \in \Phi} M_\phi^{(2)}$ is through the reward function in the last layer. However, if action $\mathfrak{a}_2$ is taken in any step within the whole trajectory, the last layer distribution will be $\text{Unif}(\mathcal{X}_H)$, which will induce the same reward distribution no matter the MDP is $M_\phi^{(1)}$ or $M_\phi^{(2)}$.

Hence as long as none of the trajectory collected takes only action 1 among the trajectory, the learner will fail to output $1/2$-accurate estimation in at least one MDP in $\mathcal{M}$ with probability at least $1/2$. And using $o(2^H)$ trajectories, the learner will fail to output $1/2$-accurate estimation in at least one MDP in $\mathcal{M}$ with probability at least $1/2 - o(2^H) \cdot 1/2^H \ge 1/4$. ■