Provably Efficient Offline Reinforcement Learning with Perturbed Data Sources

Chengshuai Shi [1] Wei Xiong [2] Cong Shen [1] Jing Yang [3]

Abstract

Existing theoretical studies on offline reinforcement learning (RL) mostly consider a dataset sampled directly from the target task. In practice, however, data often come from several heterogeneous but related sources. Motivated by this gap, this work aims at rigorously understanding offline RL with multiple datasets that are collected from *randomly perturbed versions* of the target task instead of from itself. An information-theoretic lower bound is derived, which reveals a necessary requirement on the number of involved sources in addition to that on the number of data samples. Then, a novel HetPEVI algorithm is proposed, which simultaneously considers the *sample uncertainties* from a finite number of data samples per data source and the *source uncertainties* due to a finite number of available data sources. Theoretical analyses demonstrate that HetPEVI can solve the target task as long as the data sources *collectively* provide a good data coverage. Moreover, HetPEVI is demonstrated to be optimal up to a polynomial factor of the horizon length. Finally, the study is extended to offline Markov games and offline robust RL, which demonstrates the generality of the proposed designs and theoretical analyses.

1. Introduction

Offline reinforcement learning (RL) (Levine et al., 2020), a.k.a. batch RL (Lange et al., 2012), has received growing interest in recent years. It aims at training RL agents using accessible datasets collected *a priori* and thus avoids expensive online interactions. Along with its tremendous empirical successes, recent studies have also advanced the theoretical understandings of offline RL (Rashidinejad et al., 2021; Jin et al., 2021; Xie et al., 2021b).

Despite these progresses, most theoretical studies on offline RL focus on learning via data collected exactly from the target task. In practice, however, it is difficult to ensure such a perfect match. Instead, it is more reasonable to model that data are collected from different sources that are perturbed versions of the target task in some applications. For example, when training a chatbot (Jaques et al., 2020), the offline dialogue datasets typically consist of conversations from different people with naturally varying language habits. The training objective is the common underlying language structure, e.g., the basic grammar, which cannot be completely reflected in any individual dialogue but can be holistically learned from the aggregation of them. Similar examples can be found in healthcare with records from different hospitals (Tang & Wiens, 2021), recommender systems with histories from different customers (Afsar et al., 2022), and others; more discussions are provided in Appendix A.1.

While a few empirical investigations have been reported (in particular, under the offline meta-RL framework, e.g., Dorfman et al. (2021); Lin et al. (2022); Mitchell et al. (2021)), theoretical understandings of effectively and efficiently learning with heterogeneous while related data sources are lacking. Motivated by this limitation, this work makes progress in answering the following open question:

Can we design provably efficient offline RL for a target task with multiple randomly perturbed data sources?

Challenges. Existing offline RL studies typically deal with one type of uncertainty, i.e., the *sample uncertainty* associated with the finite data sampled directly from the target task, which results in distributional shift and partial coverage. In addition to these, randomly perturbed data sources bring new challenges. First, since multiple data sources are involved, it is important to *jointly aggregate* their sample uncertainties, and to leverage their *collective information*. Moreover, even if every data source is perfectly known, the target task may *not* be fully revealed as the data sources are perturbations of the target. Thus, importantly, an additional type of uncertainty due to a finite number of available data sources should be jointly considered, which is referred to as

[1] Department of Electrical and Computer Engineering, University of Virginia [2] Department of Mathematics, The Hong Kong University of Science and Technology [3] Department of Electrical Engineering, The Pennsylvania State University. Correspondence to: Chengshuai Shi <cs7ync@virginia.edu>, Cong Shen <cong@virginia.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

the *source uncertainty*.

Contributions. To the best of our knowledge, this is the first theoretical work that studies the fundamental limits of offline RL with multiple perturbed data sources and develops provably efficient algorithm designs, which can benefit relevant applications of RL using multiple heterogeneous data sources (e.g., offline meta-RL). The contributions are summarized as follows:

- We study a new offline RL problem where the datasets are collected from multiple heterogeneous source Markov Decision Processes (MDPs), with possibly different reward and transition dynamics, as opposed to directly from the target MDP. Motivated by practical applications, the data source MDPs are modeled as randomly perturbed versions of the target MDP.
- A novel information-theoretic lower bound is derived. It illustrates that in addition to ensuring sufficient sample complexity (i.e., the amount of collected data samples), it is equally (if not more) important to guarantee sufficient source diversity (i.e., the number of involved data sources). This observation is new in the offline RL study and provides useful guidance for practical data collection.
- A novel HetPEVI algorithm is proposed with a carefully designed two-part penalty term to ensure pessimistic estimations during learning. Especially, the first part of the penalty jointly aggregates the sample uncertainties associated with each dataset, while the second one provides additional compensations for the source uncertainties, which is uniquely required to handle randomly perturbed data sources.
- Theoretical analysis demonstrates that as long as the perturbed data sources collectively provide a good data coverage, HetPEVI can learn the target task efficiently. This *collective* coverage requirement is more practical than the previous *individual* coverage requirement. In particular, it only requires that for each state-action pair induced by the optimal policy on the target task, there *exists* a (potentially different) data source that can provide data samples for it. More importantly, compared with the lower bound, HetPEVI is shown to be optimal up to a polynomial factor of the horizon length regarding its requirements of sample complexity and source diversity. Additional experimental results further corroborate the effectiveness of HetPEVI.
- The design principle in HetPEVI is further extended to offline Markov games and offline robust RL with perturbed data sources, which showcases its generality. Importantly, these extensions further validate that learning with perturbed data sources is feasible given a good collective data coverage, while it requires guarantees of both sample complexity and source diversity.

Related Works. Theoretical understandings of offline RL (Levine et al., 2020) have been gaining increased interest in recent years, where the principle of "pessimism" plays an important role. In particular, Xie et al. (2021b); Li et al. (2022); Shi et al. (2022); Rashidinejad et al. (2021) have investigated the standard tabular setting, while Jin et al. (2021); Yin et al. (2022); Xie et al. (2021a); Uehara & Sun (2021) studied function approximations. Most of these theoretical advances assume that data are collected directly from the target task. On the other hand, practical RL research has seen growing interest in how to utilize data from heterogeneous sources, e.g., meta-RL (Mitchell et al., 2021; Dorfman et al., 2021; Lin et al., 2022; Li et al., 2020) and federated RL (Zhuo et al., 2019; Jin et al., 2022). As a first step to filling this theoretical gap, this work aims at understanding offline RL with multiple randomly perturbed sources. Some particularly related research domains in theoretical RL are discussed in Table 1, which are compared with this work in the studied data sources and evaluation criteria. A detailed literature review can be found in Appendix A.2.

2. Problem Formulation

2.1. Preliminaries of Episodic MDPs

We consider an episodic MDP $\mathcal{M} := (H, \mathcal{S}, \mathcal{A}, \mathbb{P} := \{\mathbb{P}_h : h \in [H]\}, r := \{r_h : h \in [H]\})$. In this tuple, H is the length of each episode, $\mathcal{S}$ is the state space with $S := |\mathcal{S}|$, $\mathcal{A}$ is the action space with $A := |\mathcal{A}|$, $\mathbb{P}_h(s'|s, a)$ gives the probability of transiting to state s' if action a is taken upon state s at step h, and $r_h(s, a)$ is the deterministic reward in the interval of $[0, 1]$ of taking action a for state s at step h.[1] Specifically, at each step $h \in [H]$, the agent observes state $s_h \in \mathcal{S}$, picks action $a_h \in \mathcal{A}$, receives reward $r_h(s_h, a_h)$, and then transits to a next state $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h)$.

A policy $\pi := \{\pi_h(\cdot|s) : (s, h) \in \mathcal{S} \times [H]\}$ consists of distributions $\pi_h(\cdot|s)$ over the action space $\mathcal{A}$. For convenience, we use $\pi_h(s)$ to refer to the chosen action at $(s, h) \in \mathcal{S} \times [H]$ by a deterministic policy π. To measure the performance, the value function of the policy π is defined as $V_h^{\pi,\mathcal{M}}(s) := \mathbb{E}_{\pi,\mathcal{M}}[\sum_{i=h}^{H} r_i(s_i, a_i)|s_h = s]$ for all $(s, h) \in \mathcal{S} \times [H]$, where the expectation $\mathbb{E}_{\pi,\mathcal{M}}[\cdot]$ is with respect to (w.r.t.) the random trajectory induced by policy π on MDP $\mathcal{M}$. Similarly, the Q-function of π can be defined as $Q_h^{\pi,\mathcal{M}}(s, a) := \mathbb{E}_{\pi,\mathcal{M}}[\sum_{i=h}^{H} r_i(s_i, a_i)|s_h = s, a_h = a]$ for all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. If the initial state is drawn from a distribution $\xi \in \Delta(\mathcal{S})$, the following notation is adopted: $V_1^{\pi,\mathcal{M}}(\xi) := \mathbb{E}_{s\sim\xi}[V_1^{\pi,\mathcal{M}}(s)]$.

To characterize the state and state-action occupancy distribution induced by policy π on MDP $\mathcal{M}$ at each step, we denote that $d_h^{\pi,\mathcal{M}}(s; \xi) := \mathbb{E}_{\pi,\mathcal{M}}\left[\mathbb{1}\{s_h = s\}|s_1 \sim \xi\right]$ and $d_h^{\pi,\mathcal{M}}(s, a; \xi) := \mathbb{E}_{\pi,\mathcal{M}}\left[\mathbb{1}\{s_h = s, a_h = a\}|s_1 \sim \xi\right]$, where the expectation is conditioned on $s_1 \sim \xi$. Whenever it is clear from the context, we simplify the notations as

[1] The assumption of deterministic rewards is standard in theoretical RL (Jin et al., 2018; 2020) as the uncertainties in estimating rewards are dominated by those in estimating transitions.

Table 1. Related works and their studied settings; see Appendix A.2 and Fig. 4 for more discussions and graphical illustrations.

	Data source	Evaluation of the learned policy
Canonical offline RL (Xie et al., 2021b; Li et al., 2022)	The target MDP	Performance on the target MDP
Offline robust RL (Shi & Chi, 2022; Zhou et al., 2021)	The nominal MDP	Worst-case performance in an uncertainty set around the nominal MDP
Offline latent RL (Offline version of Kwon et al. (2021))	A set of potential MDPs	Average performance on unknown MDPs randomly selected from the data source set
Offline federated/multi-task RL (Zhou et al., 2022a; Lu et al., 2021)	A set of task MDPs	Performance of the learned task-dependent policy on each task MDP
Offline RL with perturbed data sources (this work)	A set of MDPs perturbed from the target MDP (Assumption 2.1)	Performance on the target MDP

$d_h^{\pi,\mathcal{M}}(s) := d_h^{\pi,\mathcal{M}}(s;\xi)$ and $d_h^{\pi,\mathcal{M}}(s,a) := d_h^{\pi,\mathcal{M}}(s,a;\xi)$.

2.2. Learning Goal

This work considers a target task modeled by an MDP $\mathcal{M} = (H,\mathcal{S},\mathcal{A},\mathbb{P},r)$ as introduced above. The goal is to find a good policy for this target MDP using certain existing datasets, i.e., offline learning. Especially, the sub-optimality gap of a policy $\hat{\pi}$ on $\mathcal{M}$ w.r.t. $s_1 \sim \xi$ is defined as follows:

$$\text{Gap}(\hat{\pi};\mathcal{M},\xi) := V_1^{\pi^*,\mathcal{M}}(\xi) - V_1^{\hat{\pi},\mathcal{M}}(\xi), \qquad (1)$$

where $\pi^* := \arg\max_\pi V_1^{\pi,\mathcal{M}}(\xi)$ is the optimal (deterministic) policy on the target MDP $\mathcal{M}$. Correspondingly, an output policy $\hat{\pi}$ is called ε-optimal if $\text{Gap}(\hat{\pi};\mathcal{M},\xi) \le \varepsilon$.

2.3. Data Sources and The Task-Source Relationship

Instead of assuming data sampled directly from the unknown target task MDP $\mathcal{M}$, this work considers that the learning agent has access to datasets from L different data sources. Each data source is also an unknown MDP, and the l-th data source can be represented as $\mathcal{M}_l = (H,\mathcal{S},\mathcal{A},\mathbb{P}_l,r_l)$. To capture heterogeneity, each data source $\mathcal{M}_l$ may not exactly match the target task $\mathcal{M}$. Concretely, despite the same episodic length, state space, and action space, their transition and reward dynamics are not necessarily aligned, i.e., possibly, $\mathbb{P}_{h,l}(\cdot|s,a) \neq \mathbb{P}_h(\cdot|s,a)$ and $r_{h,l}(s,a) \neq r_h(s,a)$.

In practical applications, while being heterogeneous, the data source MDPs are often still related to the target task (e.g., the dialogue dataset example in Section 1). In particular, data sources in offline meta-RL are often assumed to be sampled from one certain distribution (Mitchell et al., 2021). Thus, the following relationship is considered between the target MDP and the data source MDPs.

Assumption 2.1 (Task–source relationship)**.** Data source MDPs $\{\mathcal{M}_l = \{H,\mathcal{S},\mathcal{A},\mathbb{P}_l,r_l\} : l \in [L]\}$ are generated from an unknown set of distributions $g = \{g_h : h \in [H]\}$ such that for each $(l,h) \in [L] \times [H]$, the reward and transition $\{r_{h,l},\mathbb{P}_{h,l}\}$ are independently sampled from the distribution $g_h(\cdot)$ whose expectation is $\{r_h,\mathbb{P}_h\}$ of the target MDP $\mathcal{M} = \{H,\mathcal{S},\mathcal{A},\mathbb{P},r\}$.

The requirement that rewards are random samples with the expectation as the target task is commonly adopted in bandits literature (Shi & Shen, 2021; Zhu & Kveton, 2022), and the same requirement on the transition dynamics is a natural extension, where one representative example is to follow a Dirichlet distribution (Marchal & Arbel, 2017).

Remark 2.2. This work essentially considers a "worst-case" scenario in the sense that our proposed designs and obtained results are for any generation process satisfying Assumption 2.1, i.e., the generation process g exists and has an expectation as the target task $\mathcal{M}$; the other properties of g (e.g., its variance) are not specified but our designs and results still hold.

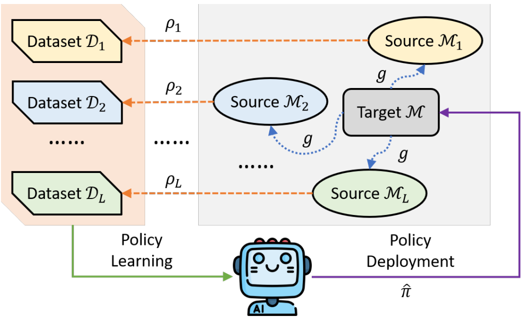


Figure 1. Problem overview: dotted (blue and orange) lines indicate the collection of datasets (with randomness from both source generation and data sampling), while the solid (green and purple) lines are for policy learning and deployment. Especially, the agent aims at solving a target MDP $\mathcal{M}$ but lacks direct access. Instead, available datasets $\{\mathcal{D}_l : l \in [L]\}$ are collected via behavior policies $\{\rho_l : l \in [L]\}$ from data source MDPs $\{\mathcal{M}_l : l \in [L]\}$ that are randomly perturbed from the target MDP $\mathcal{M}$ (through distribution g in Assumption 2.1). With such datasets, the agent learns a policy $\hat{\pi}$ offline, which is deployed (potentially in the future) on the target MDP $\mathcal{M}$ with its performance gap measured by Eqn. (1). See Fig. 3 for additional graphical illustrations.

2.4. Collections of Datasets

We consider that from each data source $\mathcal{M}_l$, a dataset $\mathcal{D}_l := \{(s^k_{1,l}, a^k_{1,l}, r^k_{1,l}, \cdots, s^k_{H,l}, a^k_{H,l}, r^k_{H,l}) : k \in [K]\}$ is collected, which consists of K independent trajectories sampled by a (possibly different) unknown behavior policy ρ_l with a (possibly different) initial state distribution ξ_l. More specifically, the k-th trajectory in dataset $\mathcal{D}_l$ is generated according to $s^k_{1,l} \sim \xi_l(\cdot)$, $a^k_{h,l} \sim \rho_{h,l}(\cdot|s^k_{h,l})$, $r^k_{h,l} = r_{h,l}(s^k_{h,l}, a^k_{h,l})$, and $s^k_{h+1,l} \sim \mathbb{P}_{h,l}(\cdot|s^k_{h,l}, a^k_{h,l})$.

It can be observed that although the trajectories in the collected datasets are independently collected, the sampled transitions from the same episode are still correlated. A two-fold sub-sampling technique is developed in Li et al. (2022) to alleviate such temporal dependencies and re-create datasets where the sampled transitions is independent of each other. To ease the presentation, we denote $\{\mathcal{D}'_l : l \in [L]\}$ as the datasets re-created from the original $\{\mathcal{D}_l : l \in [L]\}$ with the two-fold sub-sampling. Details on the sub-sampling technique are provided in Appendix C.1.

A compact overview of the studied offline learning problem is provided in Fig. 1, whose complete step-by-step version ca be found in Fig. 3.

Miscellaneous. Notations without subscripts l generally refer to the target MDP $\mathcal{M}$, while subscript l is added when discussing each individual data source $\mathcal{M}_l$. For any function $f : \mathcal{S} \to \mathbb{R}$, the transition operator and Bellman operator of the target MDP $\mathcal{M}$ at each step $h \in [H]$ are defined, respectively, as $(\mathbb{P}_h f)(s,a) := \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)}[f(s')|s,a]$ and $(\mathbb{B}_h f)(s,a) := r_h(s,a) + (\mathbb{P}_h f)(s,a)$. The notation c is used throughout the paper with varying values to represent a constant of order $O(1)$. Lastly, the notation $y \gtrsim x$ compactly denotes that $y \geq x \log(KHSAL/\delta)$, where δ is a constant in $(0,1)$, while $x \vee y := \max\{x,y\}$ and $x \wedge y := \min\{x,y\}$.

3. Lower Bound Analysis

With $\mathcal{L}_h(s,a) := \{l \in [L] : d^{\rho_l, \mathcal{M}_l}_h(s,a;\xi_l) > 0\} \subseteq [L]$ as the set of data sources that can visit (s,a,h), the following two quantities are introduced: the minimum number of sources that cover each possible visitations of the optimal policy π^* on the target MDP $\mathcal{M}$ is defined as

$$L^\dagger := \min\left\{|\mathcal{L}_h(s,a)| : (s,a,h) \text{ s.t. } d^{\pi^*,\mathcal{M}}_h(s,a;\xi) > 0\right\},$$

and the collective coverage parameter is defined as

$$C^\dagger := \max_{(s,a,h)}\left\{\sum_{l \in \mathcal{L}_h(s,a)} \frac{\min\left\{d^{\pi^*,\mathcal{M}}_h(s,a;\xi), \frac{1}{S}\right\}}{|\mathcal{L}_h(s,a)| \cdot d^{\rho_l,\mathcal{M}_l}_h(s,a;\xi_l)}\right\},$$

where we adopt the convention that $0/0 = 0$. Note that $L^\dagger$ captures how many sources provide information on the optimal policy by counting the useful ones, while $C^\dagger$ further characterizes how well these data sources collectively provide information by comparing their aggregated occupancy probability with the optimal policy. The clipping with $1/S$ follows the definition of the single-policy clipped coverage parameter recently proposed in Li et al. (2022).

The following novel information-theoretic lower bound is established to provide fundamental limits.

Theorem 3.1. *For any $(H, S, L^\dagger, C^\dagger, \varepsilon)$ obeying $H \geq 4$, $C^\dagger \geq 4/S$ and $\varepsilon \leq c_0 H$, if either of the following two conditions is not satisfied:*

$$L^\dagger K \geq c_1 \frac{C^\dagger H^2 S}{\varepsilon^2}, \qquad L^\dagger \geq c_2 \frac{H^2}{\varepsilon^2},$$

one can construct two target MDPs $\{\mathcal{M}^0, \mathcal{M}^1\}$, an initial state distribution ξ, a data source generation distribution set g, and datasets $\{\mathcal{D}_l : l \in [L]\}$, such that

$$\inf_{\hat{\pi}} \max_{\phi \in \{0,1\}} \left\{\mathbf{P}_\phi\left(\text{Gap}(\hat{\pi}; \mathcal{M}^\phi, \xi) > \varepsilon\right)\right\} \geq \frac{1}{8},$$

where c_0, c_1 and c_2 are universal constants, the infimum is taken over all estimators $\hat{\pi}$, and $\mathbf{P}_0$ (resp. $\mathbf{P}_1$) denotes the probability when the target MDP is $\mathcal{M}^0$ (resp. $\mathcal{M}^1$).

It can be observed that this lower bound has two requirements: (1) sample complexity, i.e., $L^\dagger K = \Omega(C^\dagger H^2 S/\varepsilon^2)$; (2) source diversity, i.e., $L^\dagger = \Omega(H^2/\varepsilon^2)$. While similar requirements on sample complexity have appeared in previous studies (Li et al., 2022; Rashidinejad et al., 2021), the one derived here is established collectively through $C^\dagger$ and $L^\dagger$ on the aggregation of heterogeneous data sources with different behavior policies. To the best of our knowledge, the second requirement on source diversity appears in offline RL studies for the first time. It provides a key observation that without enough data sources that provide useful information, the target MDP cannot be efficiently learned even with infinite data samples from each data source. The combination of these two requirements indicates that it is equally (if not more) important to involve sufficient high-quality sources as to sample adequate data from each of them, which is a helpful principle to guide practical data collection.

4. The HetPEVI Algorithm

In this section, we present a novel model-based algorithm, termed HetPEVI, to perform offline RL with perturbed data sources, which is summarized in Algorithm 1.

4.1. Constructing Empirical Estimations

The HetPEVI algorithm begins by counting the number of visitations in the available datasets. Especially, we denote

Algorithm 1 HetPEVI

```
Input: Dataset $\mathcal{D} = \{D_l : l \in [L]\}$
Obtain $\mathcal{D}'_l \leftarrow$ subsampling$(\mathcal{D}_l), \forall l \in [L]$
For each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, first obtain $\hat{\mathcal{L}}_h(s,a)$; then for each $l \in \hat{\mathcal{L}}_h(s,a)$, estimate $\hat{r}_{h,l}(s,a)$ and $\hat{\mathbb{P}}_{h,l}(\cdot|s,a)$; lastly, aggregate $\hat{r}_h(s,a)$ and $\hat{\mathbb{P}}_h(\cdot|s,a)$ {See Section 4.1}
Initialize $\hat{V}_{H+1}(s) \leftarrow 0, \forall s \in \mathcal{S}$    {See Section 4.2}
for $h = H, H-1, \cdots, 1$ do
    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ do
        $\Gamma^{\alpha}_h(s,a) \leftarrow c\sqrt{\sum_{l \in \hat{\mathcal{L}}_h(s,a)} \frac{H^2 \log(SAH/\delta)}{(\hat{L}_h(s,a))^2 N_{h,l}(s,a)}}$
        $\Gamma^{\beta}_h(s,a) \leftarrow c\sqrt{H^2 \log(SAH/\delta)/\hat{L}_h(s,a)}$
        $\Gamma_h(s,a) \leftarrow \min\{\Gamma^{\alpha}_h(s,a) + \Gamma^{\beta}_h(s,a), H\}$
        $\hat{Q}_h(s,a) \leftarrow \max\{(\hat{\mathbb{B}}_h \hat{V}_{h+1})(s,a) - \Gamma_h(s,a), 0\}$
    end for
    for $s \in \mathcal{S}$ do
        $\hat{\pi}_h(s) \leftarrow \arg\max_{a \in \mathcal{A}} \hat{Q}_h(s,a)$
        $\hat{V}_h(s) \leftarrow \hat{Q}_h(s, \hat{\pi}_h(s))$
    end for
end for
Output: policy $\hat{\pi} = \{\hat{\pi}_h(s) : (s,h) \in \mathcal{S} \times [H]\}$
```

$N_{h,l}(s,a)$ and $N_{h,l}(s,a,s')$ as the amount of visitations on each tuple (s,a,h) and (s,a,h,s') in dataset $\mathcal{D}'_l$, respectively. Then, the subset of datasets that have non-zero visitations on tuple (s,a,h) can be found as $\hat{\mathcal{L}}_h(s,a) := \{l \in [L] : N_{h,l}(s,a) > 0\}$, whose size is denoted as $\hat{L}_h(s,a) := |\hat{\mathcal{L}}_h(s,a)|$. Empirical estimations of rewards and transitions are then obtained for each tuple $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and each source $l \in \hat{\mathcal{L}}_h(s,a)$ as $\hat{r}_{h,l}(s,a) = r_{h,l}(s,a)$ and $\hat{\mathbb{P}}_{h,l}(s'|s,a) = N_{h,l}(s,a,s')/N_{h,l}(s,a)$. These individual estimates are further aggregated into overall estimates for each tuple $(s,a,h,s') \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{S}$ as $\hat{r}_h(s,a) = \sum_{l \in \hat{\mathcal{L}}_h(s,a)} \hat{r}_{h,l}(s,a)/(\hat{L}_h(s,a) \vee 1)$ and $\hat{\mathbb{P}}_h(s'|s,a) = \sum_{l \in \hat{\mathcal{L}}_h(s,a)} \hat{\mathbb{P}}_{h,l}(s'|s,a)/(\hat{L}_h(s,a) \vee 1)$. Note that in these estimations of transitions and rewards, only the data sources that provide non-zero visitations are counted, which may differ for each tuple (s,a,h).

4.2. Considering Two Types of Uncertainties

With the obtained estimations, HetPEVI iterates backward from the last step to the first step as

$$\hat{Q}_h(s,a) = \max\left\{\left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \Gamma_h(s,a), 0\right\}, \qquad (2)$$
$$\hat{\pi}_h(s) = \arg\max_{a \in \mathcal{A}} \hat{Q}_h(s,a), \quad \hat{V}_h(s) = \hat{Q}_h(s, \hat{\pi}_h(s)),$$

with $\hat{V}_{H+1}(s) = 0, \forall s \in \mathcal{S}$, and the empirical Bellman operator $\hat{\mathbb{B}}_h$ defined as $(\hat{\mathbb{B}}_h \hat{V}_{h+1})(s,a) := \hat{r}_h(s,a) + (\hat{\mathbb{P}}_h \hat{V}_{h+1})(s,a)$, where $(\hat{\mathbb{P}}_h \hat{V}_{h+1})(s,a)$ is the empirical version of $(\mathbb{P}_h \hat{V}_{h+1})(s,a)$ using the estimated $\hat{\mathbb{P}}_h(\cdot|s,a)$. The essence of this procedure is that instead of directly setting $\hat{Q}_h(s,a)$ as $(\hat{\mathbb{B}}_h \hat{V}_{h+1})(s,a)$ (as in the standard value iteration), a penalty term $\Gamma_h(s,a)$ is subtracted, which serves the important role of keeping the estimations $\hat{V}_h(s)$ and $\hat{Q}_h(s,a)$ pessimistic and providing conservative actions.

Previous offline RL studies (Jin et al., 2021; Rashidinejad et al., 2021; Xie et al., 2021b; Li et al., 2022) only deal with one single data source (i.e., the target MDP) and thus only one type of uncertainty due to a finite number of data samples. Instead, the agent in this work needs to process multiple heterogeneous datasets, while none of them individually characterize the target task. Thus, as mentioned in Section 1, the agent faces two *coupled* uncertainties. First, the *sample uncertainties* associated with each data source need to be jointly aggregated instead of being measured individually to leverage collective information. Second, even with perfect knowledge of each data source, the target MDP may not be fully revealed. As a result, the agent also needs to consider the uncertainties from the limited number of data sources, i.e., the *source uncertainties*.

To address the two uncertainties, the penalty term is designed to have two parts as follows:

$$\Gamma_h(s,a) = \min\left\{\Gamma^{\alpha}_h(s,a) + \Gamma^{\beta}_h(s,a), H\right\},$$

where $\Gamma^{\alpha}_h(s,a)$ aggregates the sample uncertainties while $\Gamma^{\beta}_h(s,a)$ accounts for the source uncertainties.

Penalties to Aggregate Sample Uncertainties. The first part of the penalty, i.e., $\Gamma^{\alpha}_h(s,a)$, is designed as

$$\Gamma^{\alpha}_h(s,a) = c\sqrt{\frac{1}{(\hat{L}_h(s,a))^2} \sum_{l \in \hat{\mathcal{L}}_h(s,a)} \frac{H^2 \log(SAH/\delta)}{N_{h,l}(s,a)}}.$$

Note that this design avoids the data sources that have zero visitations on this tuple (s,a,h) and is a joint measure of sample uncertainties from the other sources (instead of directly summing up their individual sample uncertainties as $\tilde{O}(\frac{1}{\hat{L}_h(s,a)} \sum_{l \in \hat{\mathcal{L}}_h(s,a)} \sqrt{\frac{H^2}{N_{h,l}(s,a)}})$). These designs are important to accelerate learning and obtain the near-optimal performance illustrated later in Section 5.

Penalties to Account for Source Uncertainties. The second part of the penalty $\Gamma^{\beta}_h(s,a)$ serves the important role of measuring the uncertainties due to the limited amount of data sources, which is designed as:

$$\Gamma^{\beta}_h(s,a) = c\sqrt{\frac{H^2 \log(SAH/\delta)}{\hat{L}_h(s,a)}}.$$

Intuitively, it shrinks with the number of datasets that provides information on the tuple (s,a,h), i.e., $\hat{L}_h(s,a)$, which thus may differ among state-action pairs. Jointly using the two penalty terms, the overall uncertainties in the datasets can be compensated. Especially, with high probability, for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, it can be ensured that $|(\hat{\mathbb{B}}_h\hat{V}_{h+1})(s,a)-(\mathbb{B}_h\hat{V}_{h+1})(s,a)|\leq\Gamma_h(s,a)$.

Remark 4.1. The adopted penalty $\Gamma^{\beta}_h(s,a)$ for source uncertainty is intended to accommodate any unknown variance between sources and the task, i.e., a worst-case consideration as mentioned in Remark 2.2. However, if there is prior knowledge of the variance, it is feasible to incorporate such information. In particular, if the rewards and transition vectors (at each (s,a,h)) are generated via σ_g-sub-Gaussian distributions, the penalty for source uncertainties can be designed as $\Gamma^{\beta}_h(s,a)=c\sqrt{\frac{H^2\sigma_g^2\log(SAH/\delta)}{\hat{L}_h(s,a)}}$.

5. Performance Analysis

This section provides a theoretical analysis of HetPEVI. In particular, the following performance guarantee can be established.

Theorem 5.1 (HetPEVI)**.** *Under Assumption 2.1, with probability at least* $1-\delta$*, the output policy* $\hat{\pi}$ *of HetPEVI satisfies*

$$\mathrm{Gap}(\hat{\pi};\mathcal{M},\xi)=\tilde{O}\left(\sqrt{\frac{C^{\dagger}H^4S}{L^{\dagger}K}}+\sqrt{\frac{H^4}{L^{\dagger}}}\right),$$

when $K\gtrsim c/d^{\min}$*, where* $d^{\min}:=\min\{d_h^{\rho_l,\mathcal{M}_l}(s,a):(s,a,h,l)$ *s.t.* $d_h^{\pi^*,\mathcal{M}}(s,a)>0, d_h^{\rho_l,\mathcal{M}_l}(s,a)>0\}$.

It can be observed that as long as $C^{\dagger}<\infty$ and $L^{\dagger}>0$, a meaningful performance gap can be provided by Theorem 5.1. To ensure these two conditions, it is equivalent to have the following assumption of collective coverage.

Assumption 5.2 (Collective coverage, this work)**.** For each $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$ that $d_h^{\pi^*,\mathcal{M}}(s,a;\xi)>0$, there *exists* $l\in[L]$ such that the behavior policy ρ_l satisfy that $d_h^{\rho_l,\mathcal{M}_l}(s,a;\xi_l)>0$.

It is beneficial to compare Assumption 5.2 with those from previous offline RL studies. Especially, with a dataset $\mathcal{D}$ sampled with a behavior policy ρ and an initial state distribution ξ' directly from the target MDP $\mathcal{M}$, the following individual coverage assumption is often required.

Assumption 5.3 (Individual coverage, Rashidinejad et al. (2021); Xie et al. (2021b); Li et al. (2022))**.** For each $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$ that $d_h^{\pi^*,\mathcal{M}}(s,a;\xi)>0$, the behavior policy ρ satisfy that $d_h^{\rho,\mathcal{M}}(s,a;\xi')>0$.

Assumption 5.3 is strong as the behavior policy needs to individually cover the unknown optimal policy. Instead, Assumption 5.2 is more practical because it leverages collective information: different parts of the optimal trajectory can be covered by different behavior policies and data sources. In particular, it may be easier to reach some states and actions in certain data sources with their corresponding behavior policies.

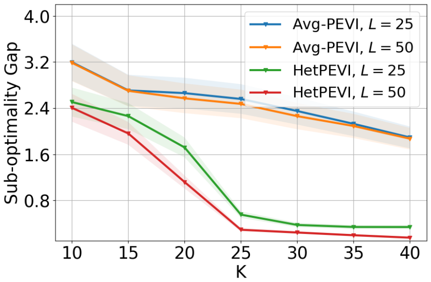


Figure 2. Performance comparisons between HetPEVI and the baseline with varying amounts of data samples and data sources.

Moreover, besides the burn-in cost of $K\gtrsim c/d^{\min}$ which does not scale with ε, Theorem 5.1 illustrates that to obtain an ε-optimal policy, HetPEVI only needs

$$L^{\dagger}K=\tilde{O}\left(\frac{C^{\dagger}H^4S}{\varepsilon^2}\right)\quad\text{and}\quad L^{\dagger}=\tilde{O}\left(\frac{H^4}{\varepsilon^2}\right),$$

where the first requirement is on the *sample complexity* while the second one is on the *source diversity*. Compared with the lower bound in Theorem 3.1, it can be observed that HetPEVI is optimal (up to logarithmic factors) on the dependency of $L^{\dagger},C^{\dagger},K,S,\varepsilon$, and only incurs an additional H^2 factor, which demonstrates its effectiveness and efficiency. Note that if we further leverage prior variance information and adopt the adaptive penalty as illustrated in Remark 4.1, a corresponding variance-adaptive performance guarantee can be established following the proof of Theorem 5.1.

6. Experimental Results

To further empirically validate the effectiveness of HetPEVI, experimental results are reported in Fig. 2. In particular, simulations are performed on a target MDP with $S=2$, $A=20$ and $H=20$. The data source MDPs are randomly generated through a set of independent Dirichlet distributions (Marchal & Arbel, 2017). The baseline, labeled as 'Avg-PEVI', is an aggregated policy. In particular, with each individual dataset $\mathcal{D}_l$, a policy $\hat{\pi}_l$ is learned via PEVI (Xie et al., 2021b; Jin et al., 2021). Then, at each $(s,h)\in\mathcal{S}\times[H]$, Avg-PEVI selects an action uniformly among $\{\hat{\pi}_{h,l}(s):l\in[L]\}$. Additional experimental details can be found in Appendix F.

From Fig. 2, it can be observed that HetPEVI consistently outperforms the baseline policy, and is capable of approaching the optimal performance particularly when sufficient data sources and data samples are available.

7. Extension to Offline Markov Game

We extend the study to the multi-player regime. In particular, the most representative two-player zero-sum Markov game (MG) is considered, which can be characterized by $\mathcal{G} := \{H, \mathcal{S}, \mathcal{A} := \mathcal{A}^1 \times \mathcal{A}^2, \mathbb{P} := \{\mathbb{P}_h : h \in [H]\}, r := \{r_h : h \in [H]\}\}$. Besides H as episode length and $\mathcal{S}$ as the state space, the major distinction in MG compared with MDP is that the entire action space $\mathcal{A}$ is a product of the individual action spaces of two players, i.e., $\mathcal{A}^1$ for the max-player and $\mathcal{A}^2$ for the min-player, respectively. We further denote that $A^1 := |\mathcal{A}^1|$ and $A^2 := |\mathcal{A}^2|$. Consequently, the transitions and rewards depend on the action pair $a = (a^1, a^2) \in \mathcal{A}^1 \times \mathcal{A}^2$. Value functions of MG $\mathcal{G}$ can be defined similarly as those of MDP to be $V_h^{\pi,\mathcal{G}}(s)$ and $Q_h^{\pi,\mathcal{G}}(s,a)$ for a product policy $\pi = \mu \times \nu$ from the max-player's policy μ and the min-player's policy ν.

With $\mathcal{G}$ as the target MG, we are interested in approximating its Nash equilibrium (NE) policy pair $\pi^* = (\mu^*, \nu^*)$, where μ^* and ν^* are the best responses to each other. In particular, from the perspective of the max-player[2], the performance gap of a learned policy $\hat{\mu}$ with an initial state distribution ξ is defined as $\text{MGGap}(\hat{\mu}; \mathcal{G}, \xi) := V_1^{\mu^* \times \nu^*, \mathcal{G}}(\xi) - V_1^{\hat{\mu} \times \text{br}(\hat{\mu}), \mathcal{G}}(\xi)$, where $\text{br}(\mu)$ denotes the best response of policy μ, i.e., $\text{br}(\mu) = \arg\min_\nu V_1^{\mu \times \nu, \mathcal{G}}(\xi)$.

However, instead of having data directly sampled from the target MG as in Cui & Du (2022a;b); Zhong et al. (2022); Yan et al. (2022), we consider that there are L data source MGs $\{\mathcal{G}_l : l \in [L]\}$ while K trajectories being independently sampled from each data source MG $\mathcal{G}_l$ by a behavior policy ρ_l and an initial state distribution ξ_l, where $\rho_l := \rho_l^1 \times \rho_l^2$ with ρ_l^1 for the max-player and ρ_l^2 for the min-player. A task-source relationship similar to Assumption 2.1 is considered between $\{\mathcal{G}_l : l \in [L]\}$ and $\mathcal{G}$, which is rigorously stated in Assumption D.1.

7.1. Algorithm Design and Analysis

The HetPEVI-Game algorithm is generalized from HetPEVI to perform efficient offline learning in MG with multiple perturbed data sources. With algorithm details in Appendix D.2, we note that HetPEVI-Game inherits most parts of HetPEVI while the major distinction being in the value iteration. In particular, the following is performed instead of Eqn. (2):

$$\begin{aligned}
&\hat{Q}_h(s,a) = \max\left\{\left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \Gamma_h^g(s,a), 0\right\}, \\
&(\hat{\mu}_h(\cdot|s), \hat{\nu}_h(\cdot|s)) = \text{NE}(\hat{Q}_h(s,\cdot)), \\
&\hat{V}_h(s) = \mathbb{E}_{a \sim \hat{\mu}_h(\cdot|s) \times \hat{\nu}_h(\cdot|s)}[\hat{Q}_h(s,a)],
\end{aligned}$$

where $\text{NE}(\cdot)$ finds the NE policy pair regarding the input and $\Gamma_h^g(s,a)$ is designed to have the same two-part structure as $\Gamma_h(s,a)$ of HetPEVI. Again, the two parts in $\Gamma_h^g(s,a)$ jointly capture the sample uncertainties and source uncertainties associated with the available datasets.

Similar to $C^\dagger$ and $L^\dagger$, we introduce the following quantities, $L_g^\dagger$ and $C_g^\dagger$, to measure the collective data coverage in MG:

$$L_g^\dagger := \min\left\{|\mathcal{L}_h(s,a)| : (s,a,h) \text{ s.t. } \exists \nu, d_h^{\mu^* \times \nu, \mathcal{G}}(s,a;\xi) > 0\right\},$$

$$C_g^\dagger := \max_{(s,a,h)} \max_{\nu} \left\{ \sum_{l \in \mathcal{L}_h(s,a)} \frac{\min\left\{d_h^{\mu^* \times \nu, \mathcal{G}}(s,a;\xi), \frac{1}{SA_1}\right\}}{|\mathcal{L}_h(s,a)| \cdot d_h^{\rho_l, \mathcal{G}_l}(s,a;\xi_l)} \right\},$$

where we reload the notation that $\mathcal{L}_h(s,a) := \{l \in [L] : d_h^{\rho_l, \mathcal{G}_l}(s,a;\xi_l) > 0\} \subseteq [L]$. Then, the performance guarantee for HetPEVI-Game is established in the following theorem.

Theorem 7.1 (HetPEVI-Game)**.** *Under Assumption D.1, with probability at least $1 - \delta$, the output policy $\hat{\mu}$ of HetPEVI-Game satisfies*

$$\text{MGGap}(\hat{\mu}; \mathcal{M}, \xi) = \tilde{O}\left(\sqrt{\frac{C_g^\dagger H^4 S A_1}{L_g^\dagger K}} + \sqrt{\frac{H^4}{L_g^\dagger}}\right).$$

when $K \gtrsim c/d_g^{\min}$, where $d_g^{\min} := \min\{d_h^{\rho_l, \mathcal{G}_l}(s,a) : (s,a,h,l)$ s.t. $\exists \nu, d_h^{\mu^* \times \nu, \mathcal{G}}(s,a) > 0, d_h^{\rho_l, \mathcal{G}_l}(s,a) > 0\}$.

Thus, to obtain an ε-optimal policy, HetPEVI-Game requires $L_g^\dagger K = \tilde{O}(C_g^\dagger H^4 S A_2/\varepsilon^2)$ and $L_g^\dagger = \tilde{O}(H^4/\varepsilon^2)$ besides the burn-in requirement $K \gtrsim c/d_g^{\min}$. The conditions that $L_g^\dagger > 0$ and $C_g^\dagger < \infty$ are further implied by the following collective unilateral coverage assumption.

Assumption 7.2 (Collective unilateral coverage, this work)**.** At each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, if there exists ν that $d_h^{\mu^* \times \nu, \mathcal{G}}(s,a;\xi) > 0$, then there *exists* $l \in [L]$ such that the behavior policy ρ_l satisfies that $d_h^{\rho_l, \mathcal{G}_l}(s,a;\xi_l) > 0$.

The following individual coverage assumption is quoted, where dataset $\mathcal{D}$ is collected by behavior policy ρ directly from the target MG $\mathcal{G}$ and initial state distribution ξ'.

Assumption 7.3 (Individual unilateral coverage, Cui & Du (2022a;b); Yan et al. (2022))**.** At each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, if there exists ν that $d_h^{\mu^* \times \nu, \mathcal{G}}(s,a;\xi) > 0$, then the behavior policy ρ satisfies that $d_h^{\rho, \mathcal{G}}(s,a;\xi') > 0$.

It can be observed that efficient learning in MG requires a stronger coverage assumption than that in MDP, i.e., the dataset needs to cover not only the NE pair (μ^*, ν^*) but also the policy pair (μ^*, ν) for any policy ν of the min-player. This requirement stated in Assumption 7.3 can be stringent as the only policy ρ needs to satisfy it individually. On the other hand, Assumption 7.2 is more practical as multiple data sources can collectively provide coverage.

[2] The min-player's perspective is symmetrical and thus can be similarly solved with minor modifications on notations.

8. Extension to Offline Robust RL

In addition to learning a single target task, it is often essential to learn a robustly good policy in many practical applications. We are thus motivated to consider the distributionally robust RL problem (Zhou et al., 2021; Shi & Chi, 2022). In particular, it can be characterized by a nominal (i.e., center) MDP $\mathcal{M}^c := \{H, \mathcal{S}, \mathcal{A}, \mathbb{P}^c, r\}$ and an associated uncertainty set $\mathcal{U}^\sigma(\mathbb{P}^c) = \otimes_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}\mathcal{U}^\sigma(\mathbb{P}^c_h(\cdot|s,a))$ for an uncertainty level $\sigma > 0$, where $\otimes$ denotes the Cartesian product and $\mathcal{U}^\sigma(\mathbb{P}^c_h(\cdot|s,a)) := \{\mathbb{P}^\sigma_h(\cdot|s,a) \in \Delta(\mathcal{S}) : \mathrm{KL}\left(\mathbb{P}^\sigma_h(\cdot|s,a)||\mathbb{P}^c_h(\cdot|s,a)\right) \le \sigma\}$. In other words, the uncertainty set contains the transition distributions whose KL-divergence from that of the nominal MDP at each $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ is at most σ. We further denote $\mathcal{R} := \{\mathcal{M}^\sigma = \{H, \mathcal{S}, \mathcal{A}, \mathbb{P}^\sigma, r\} : \mathbb{P}^\sigma \in \mathcal{U}^\sigma(\mathbb{P}^c)\}$ as the collection of MDPs with transitions contained in the uncertainty set. Moreover, the robust value functions of a policy π at $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$ can be defined as $V^{\pi,\mathcal{R}}_h(s) = \inf_{\mathcal{M}^\sigma\in\mathcal{R}} V^{\pi,\mathcal{M}^\sigma}_h(s)$ and $Q^{\pi,\mathcal{R}}_h(s,a) = \inf_{\mathcal{M}^\sigma\in\mathcal{R}} Q^{\pi,\mathcal{M}^\sigma}_h(s,a)$, which provide worst-case characterizations among the uncertainty set.

For this problem of robust RL, it has been established that there exists an optimal policy π^* that is deterministic and maximizes the above value functions (Iyengar, 2005). Thus, the quality of a learned policy $\hat{\pi}$ is measured by the following gap for a given initial state distribution ξ: $\mathrm{RGap}(\hat{\mu}; \mathcal{R}, \xi) := V^{\pi^*,\mathcal{R}}_1(\xi) - V^{\hat{\pi},\mathcal{R}}_1(\xi)$.

Existing works considering the offline version of this robust RL problem typically assume a dataset sampled from the nominal MDP $\mathcal{M}^c$ (Shi & Chi, 2022; Zhou et al., 2021). Instead, we consider L available datasets, i.e., $\{\mathcal{D}_l : l \in [L]\}$, and each $\mathcal{D}_l$ contains K trajectories sampled independently by a behavior policy ρ_l and an initial state distribution ξ_l on the data source MDP $\mathcal{M}_l$. A stochastic relationship between the data sources $\{\mathcal{M}_l : l \in [L]\}$ and the nominal MDP $\mathcal{M}^c$ similar to Assumption 2.1 is further considered, which is rigorously stated in Assumption E.1.

8.1. Algorithm Design and Analysis

A variant of HetPEVI is developed to find a robustly good policy with datasets from multiple perturbed data sources, termed HetPEVI-Robust. While the complete description of HetPEVI-Robust is deferred to Appendix E.2, it particularly performs the value iteration as follows:

$$\hat{Q}_h(s,a) = \max\left\{\hat{r}_h(s,a) + \left(\hat{\mathbb{P}}^{\inf}_h \hat{V}_{h+1}\right)(s,a) - \Gamma^\sigma_h(s,a), 0\right\},$$

$$\hat{\pi}_h(s) = \arg\max_{a\in\mathcal{A}} \hat{Q}_h(s,a), \quad \hat{V}_h(s) = \hat{Q}_h(s, \hat{\pi}_h(s)),$$

where we define that

$$\begin{aligned}&(\hat{\mathbb{P}}^{\inf}_h \hat{V}_{h+1})(s,a)\\ &:= \inf_{\hat{\mathbb{P}}^\sigma_h(\cdot|s,a)\sim\mathcal{U}^\sigma(\hat{\mathbb{P}}_h(\cdot|s,a))}\left(\hat{\mathbb{P}}^\sigma_h \hat{V}_{h+1}\right)(s,a)\\ &= \sup_{\lambda\ge 0}\left\{-\lambda\log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\hat{V}_{h+1}/\lambda\right)\right](s,a)\right) - \lambda\sigma\right\}.\end{aligned}$$

The above last equation holds due to the strong duality (Hu & Hong, 2013) and can be efficiently solved (Panaganti & Kalathil, 2022; Yang et al., 2021). More importantly, the penalty term $\Gamma^\sigma_h(s,a)$ is specifically designed as

$$\begin{aligned}\Gamma^\sigma_h(s,a) := \min\Bigg\{H,\ &\frac{c}{\sigma\hat{\mathbb{P}}^{\min}_h(s,a)}\sqrt{\sum_{l\in\hat{\mathcal{L}}_h(s,a)}\frac{H^2\log(SAH/\delta)}{(\hat{L}_h(s,a))^2 N_{h,l}(s,a)}}\\ &+ \frac{c}{\sigma\hat{\mathbb{P}}^{\min}_h(s,a)}\sqrt{\frac{H^2\log(SAH/\delta)}{\hat{L}_h(s,a)}} + c\sqrt{\frac{\log(SAH/\delta)}{\hat{L}_h(s,a)}}\Bigg\},\end{aligned}$$

where $\hat{\mathbb{P}}^{\min}_h(s,a) = \min\{\hat{\mathbb{P}}_h(s'|s,a) : s' \text{ s.t. } \hat{\mathbb{P}}_h(s'|s,a) > 0\}$. First, it is noted that this penalty term once again has a two-part structure: the first term to aggregate sample uncertainties and the last two terms to compensate source uncertainties. Moreover, compared with $\Gamma_h(s,a)$, an additional factor $1/(\sigma\hat{\mathbb{P}}^{\min}_h(s,a))$ appears in the design of $\Gamma^\sigma_h(s,a)$, which is carefully crafted to maintain pessimism during the non-linear value iteration.

Following the same steps in the analyses of HetPEVI and HetPEVI-Game, the following two quantities are introduced as data coverage measurements regarding robust MDP:

$$\begin{aligned}L^\dagger_\sigma &:= \min\{|\mathcal{L}_h(s,a)| : (s,a,h) \text{ s.t.}\\ &\qquad\quad \exists\mathcal{M}^\sigma \in \mathcal{R}, d^{\pi^*,\mathcal{M}^\sigma}_h(s,a;\xi) > 0\},\\ C^\dagger_\sigma &:= \max_{(s,a,h)}\max_{\mathcal{M}^\sigma\in\mathcal{R}}\left\{\sum_{l\in\mathcal{L}_h(s,a)}\frac{\min\left\{d^{\pi^*,\mathcal{M}^\sigma}_h(s,a;\xi), \frac{1}{S}\right\}}{|\mathcal{L}_h(s,a)|\cdot d^{\rho_l,\mathcal{M}_l}_h(s,a;\xi_l)}\right\},\end{aligned}$$

where $\mathcal{L}_h(s,a) := \{l \in [L] : d^{\rho_l,\mathcal{M}_l}_h(s,a;\xi_l) > 0\}$. With the following additional notations,

$$\begin{aligned}L^{\min}_\sigma &:= \min\{|\mathcal{L}_h(s,a)| : (s,a,h) \text{ s.t. } |\mathcal{L}_h(s,a)| > 0\},\\ d^{\min}_\sigma &:= \min\{d^{\rho_l,\mathcal{M}_l}_h(s,a;\xi_l) : (s,a,h,l)\\ &\qquad \text{s.t. } d^{\rho_l,\mathcal{M}_l}_h(s,a;\xi_l) > 0\},\\ \mathbb{P}^{\min}_* &:= \min\{\mathbb{P}^c_h(s'|s,a) : (s,a,h,s')\\ &\qquad \text{s.t. } d^{\pi^*,\mathcal{M}^c}_h(s,a;\xi) > 0, \mathbb{P}^c_h(s'|s,a) > 0\},\\ \mathbb{P}^{\min}_\sigma &:= \min\{\mathbb{P}^c_h(s'|s,a) : (s,a,h,s')\\ &\qquad \text{s.t. } \exists l \in [L], d^{\rho_l,\mathcal{M}_l}_h(s,a;\xi_l) > 0, \mathbb{P}^c_h(s'|s,a) > 0\},\end{aligned}$$

the following Theorem 8.1 provides a characterization of the performance of HetPEVI-Robust.

Theorem 8.1 (HetPEVI-Robust)**.** *Under Assumption E.1, with probability at least* $1-\delta$*, the output policy* $\hat{\mu}$ *of HetPEVI-Robust satisfies*

$$\mathrm{RGap}(\hat{\pi}; \mathcal{R}, \xi) = \tilde{O}\left(\frac{\sqrt{C^\dagger_\sigma H^4 S}}{\sigma\mathbb{P}^{\min}_*\sqrt{L^\dagger_\sigma K}} + \frac{H + H^2/(\sigma\mathbb{P}^{\min}_*)}{\sqrt{L^\dagger_\sigma}}\right),$$

when $K \gtrsim c/(d_\sigma^{\min}(\mathbb{P}_\sigma^{\min})^2)$ *and* $L_\sigma^{\min} \gtrsim c/(\mathbb{P}_\sigma^{\min})^2$.

It can be observed that besides the burn-in requirements on K and $L_\sigma^{\min}$, with guarantees $L_\sigma^\dagger K = \tilde{O}(C_\sigma^\dagger H^4 S(\sigma\mathbb{P}_*^{\min})^{-2}/\varepsilon^2)$ and $L_\sigma^\dagger = \tilde{O}((H^2 + H^4(\sigma\mathbb{P}_*^{\min})^{-2})/\varepsilon^2)$, HetPEVI-Robust can find an ε-optimal policy for the target robust RL. To ensure $L_\sigma^\dagger > 0$ and $C_\sigma^\dagger < \infty$, it is sufficient to have the following Assumption 8.2 of collective robust coverage, which is also compared with the previously required Assumption 8.3 on individual robust coverage (Shi & Chi, 2022).

Assumption 8.2 (Collective robust coverage, this work)**.** At each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, if there exists $\mathcal{M}^\sigma \in \mathcal{R}$ that $d_h^{\pi^*,\mathcal{M}^\sigma}(s,a;\xi) > 0$, then there *exists* $l \in [L]$ such that the behavior policy ρ_l satisfies that $d_h^{\rho_l,\mathcal{M}_l}(s,a;\xi_l) > 0$.

Assumption 8.3 (Individual robust coverage, Shi & Chi (2022))**.** At each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, if there exists $\mathcal{M}^\sigma \in \mathcal{R}$ that $d_h^{\pi^*,\mathcal{M}^\sigma}(s,a;\xi) > 0$, then the behavior policy ρ satisfies that $d_h^{\rho,\mathcal{M}^c}(s,a;\xi') > 0$.

Once again, it can be observed that Assumption 8.2 is more practical than Assumption 8.3 as it leverages collective information from all data sources.

9. Discussions

Target-source Relationships. This work mainly targets one basic target-source relationship formulated in Assumption 2.1: the source MDPs are randomly perturbed versions of the target MDPs, where the expectation of the source generation process exactly matches the target MDP. This scenario itself captures key features of many applications as mentioned in Sections 1 and A.1, and the design ideas in HetPEVI can be similarly extended to other different task-source relationships. For example, instead of the stochastic relationship in Assumption 2.1, a static relationship can be considered such that the target MDP is a weighted average of source MDPs (Agarwal et al., 2022): $r_h(s,a) = \sum_{l\in[L]} \omega_l(s,a) r_{h,l}(s,a)$ and $\mathbb{P}_h(s,a) = \sum_{l\in[L]} \omega_l(s,a)\mathbb{P}_{h,l}(s,a)$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. In this case, there is no need to consider source uncertainties and the penalty for sample uncertainties can be designed as $c\sqrt{\sum_{l\in[L]}(\omega_l(s,a))^2 H^2/N_{h,l}(s,a)}$, and a corresponding performance guarantee can also be established following the procedure of Theorem 5.1.

We hope this work can be a starting point for further investigations into different target-source relationships. Especially, one interesting direction is to consider function approximation to accommodate large state/action spaces. For example, in linear MDP (Jin et al., 2020), the random perturbations could potentially happen on the overall linear structures.

Information Aggregation Schemes. Another interesting direction to be further explored is how to aggregate information from different sources more effectively. As the first step to investigating this problem, we start with the simple approach of *equally weighting* estimates from all sources that provide information. One limitation of this approach is that adding a data source with poor coverage would not necessarily improve the performance of HetPEVI since it is equally treated in information aggregation. It is conceivable that some other fine-tuned approaches might be more efficient in aggregating information. One promising idea is to aggregate the estimate from each source with weights according to their uncertainties, i.e., a small weight for a high-uncertainty source. This approach, intuitively, would be able to deal with sources with poor coverage more efficiently (by assigning a small weight to them) while additional investigations are needed to concretely design and analyze such algorithms.

10. Conclusions

This work studied a novel problem of offline RL with data sources that are randomly perturbed versions of the target task. An information-theoretic lower bound was derived, which reveals that guarantees on sample complexity and source diversity are simultaneously required for finding a good policy on the target task. Then, a novel HetPEVI algorithm was proposed, which adopts a two-part penalty term to jointly consider the uncertainties from the finite number of data samples and the limited amount of data sources. Theoretical analyses proved that as long as a good collective (as opposed to individual) data coverage can be provided by the data sources, HetPEVI can effectively solve the target task. Moreover, the required sample complexity and source diversity of HetPEVI is optimal up to a polynomial factor of the horizon length. Experimental results further illustrated the effectiveness of HetPEVI. At last, we extended the study to offline Markov games and offline robust RL with perturbed data sources with two generalized versions of HetPEVI proposed. These extensions further corroborated that offline RL with perturbed data sources is feasible given a good collective data coverage, while it requires sufficient source diversity besides adequate sample complexity.

Acknowledgements

The work of CSs was supported in part by the US National Science Foundation (NSF) under awards 2143559, 2029978, 2002902, and 2132700, and the Bloomberg Data Science Ph.D. Fellowship. The work of JY was supported in part by the US NSF under awards 2003131, 1956276, 2030026, and 2133170.

References

Afsar, M. M., Crump, T., and Far, B. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

Agarwal, A., Song, Y., Sun, W., Wang, K., Wang, M., and Zhang, X. Provable benefits of representational transfer in reinforcement learning. *arXiv preprint arXiv:2205.14571*, 2022.

Cui, Q. and Du, S. S. When is offline two-player zero-sum markov game solvable? *arXiv preprint arXiv:2201.03522*, 2022a.

Cui, Q. and Du, S. S. Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. *arXiv preprint arXiv:2206.00159*, 2022b.

Dorfman, R., Shenfeld, I., and Tamar, A. Offline meta reinforcement learning–identifiability challenges and effective data collection strategies. *Advances in Neural Information Processing Systems*, 34, 2021.

Dubey, A. and Pentland, A. Provably efficient cooperative multi-agent reinforcement learning with function approximation. *arXiv preprint arXiv:2103.04972*, 2021.

Hu, Z. and Hong, L. J. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, pp. 1695–1724, 2013.

Iyengar, G. N. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

Jaques, N., Shen, J. H., Ghandeharioun, A., Ferguson, C., Lapedriza, À., Jones, N., Gu, S., and Picard, R. W. Human-centric dialog training via offline reinforcement learning. In *EMNLP (1)*, 2020.

Jeong, Y. and Rothenhäusler, D. Calibrated inference: statistical inference that accounts for both sampling uncertainty and distributional uncertainty. *arXiv preprint arXiv:2202.11886*, 2022.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Jin, H., Peng, Y., Yang, W., Wang, S., and Zhang, Z. Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 18–37. PMLR, 2022.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34, 2021.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

Li, L., Yang, R., and Luo, D. Focal: Efficient fully-offline meta-reinforcement learning via distance metric learning and behavior regularization. *arXiv preprint arXiv:2010.01112*, 2020.

Lin, S., Wan, J., Xu, T., Liang, Y., and Zhang, J. Model-based offline meta-reinforcement learning with regularization. *arXiv preprint arXiv:2202.02929*, 2022.

Lu, R., Huang, G., and Du, S. S. On the power of multitask representation learning in linear mdp. *arXiv preprint arXiv:2106.08053*, 2021.

Ma, X., Liang, Z., Xia, L., Zhang, J., Blanchet, J., Liu, M., Zhao, Q., and Zhou, Z. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*, 2022.

Marchal, O. and Arbel, J. On the sub-gaussianity of the beta and dirichlet distributions. *Electronic Communications in Probability*, 22:1–14, 2017.

Mitchell, E., Rafailov, R., Peng, X. B., Levine, S., and Finn, C. Offline meta-reinforcement learning with advantage weighting. In *International Conference on Machine Learning*, pp. 7780–7791. PMLR, 2021.

Panaganti, K. and Kalathil, D. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.

Panaganti, K., Xu, Z., Kalathil, D., and Ghavamzadeh, M. Robust reinforcement learning using offline data. *arXiv preprint arXiv:2208.05129*, 2022.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34, 2021.

Shi, C. and Shen, C. Federated multi-armed bandits. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Shi, L. and Chi, Y. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. *arXiv preprint arXiv:2202.13890*, 2022.

Shrestha, A., Lee, S., Tadepalli, P., and Fern, A. Deepaveragers: offline reinforcement learning by solving derived non-parametric mdps. *arXiv preprint arXiv:2010.08891*, 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Tang, S. and Wiens, J. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pp. 2–35. PMLR, 2021.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, 2009.

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

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021a.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34, 2021b.

Xiong, W., Zhong, H., Shi, C., Shen, C., Wang, L., and Zhang, T. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*, 2022.

Yan, Y., Li, G., Chen, Y., and Fan, J. Model-based reinforcement learning is minimax-optimal for offline zero-sum markov games. *arXiv preprint arXiv:2206.04044*, 2022.

Yang, J., Hu, W., Lee, J. D., and Du, S. S. Impact of representation learning in linear bandits. In *International Conference on Learning Representations*, 2020.

Yang, J., Lei, Q., Lee, J. D., and Du, S. S. Nearly minimax algorithms for linear bandits with shared representation. *arXiv preprint arXiv:2203.15664*, 2022.

Yang, W., Zhang, L., and Zhang, Z. Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*, 2021.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Zhang, C. and Wang, Z. Provably efficient multi-task reinforcement learning with model transfer. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhong, H., Xiong, W., Tan, J., Wang, L., Zhang, T., Wang, Z., and Yang, Z. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *International Conference on Machine Learning*, pp. 27117–27142. PMLR, 2022.

Zhou, D., Zhang, Y., Sonabend-W, A., Wang, Z., Lu, J., and Cai, T. Federated offline reinforcement learning. *arXiv preprint arXiv:2206.05581*, 2022a.

Zhou, R., Wang, R., and Du, S. S. Horizon-free reinforcement learning for latent markov decision processes. *arXiv preprint arXiv:2210.11604*, 2022b.

Zhou, Z., Zhou, Z., Bai, Q., Qiu, L., Blanchet, J., and Glynn, P. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, 2021.

Zhu, R. and Kveton, B. Random effect bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 3091–3107. PMLR, 2022.

Zhuo, H. H., Feng, W., Lin, Y., Xu, Q., and Yang, Q. Federated deep reinforcement learning. *arXiv preprint arXiv:1901.08277*, 2019.

A. Additional Discussions

A.1. The Motivation and Setting of This Work

The work is largely motivated by the need of finding generally applicable strategies in many real-world applications. Especially, concrete motivating examples include the chatbot training discussed in Section 1, and the following ones (among many others) from the healthcare and recommendation systems:

- In healthcare applications, standardized procedures for medical diagnosis and treatment are often valuable as they provide general guidelines for medical personnel to follow. The output policy in this work can serve this purpose as it is able to aggregate common information in historical medical records from different patients;
- In recommendation systems, when dealing with a new customer, online shopping platforms would need a generally effective mechanism for advertising and promoting. Our proposed approach can then leverage existing histories from different past customers to find the desired generic mechanism.

In other words, it would be helpful to interpret the target MDP in this work as the population-level response dynamics in these real-world applications (e.g., common reactions to certain medical treatments and promoting strategies), while the source MDPs characterize the sampled individuals. Motivated by the aforementioned practical needs, this work focuses on finding such generally applicable strategies via individually perturbed data sources.

For a more detailed comparison with related works, we also recall the setting of this work and provide a step-by-step overview in Fig. 3. In particular, we note that the learning goal is the target MDP $\mathcal{M}$ but there is no direct access to it. Instead, a few data source MDPs $\{\mathcal{M}_l : l \in [L]\}$ are available, which are randomly perturbed from the target MDP $\mathcal{M}$. Datasets $\{\mathcal{D}_l : l \in [L]\}$ are collected from these data source MDPs via behavior policies $\{\rho_l : l \in [L]\}$. With these datasets, the agent finds (e.g., via the proposed HetPEVI algorithm) an output policy $\hat{\pi}$. Lastly, this learned policy $\hat{\pi}$ is intended to be deployed back to the target MDP $\mathcal{M}$, where its performance is measured.

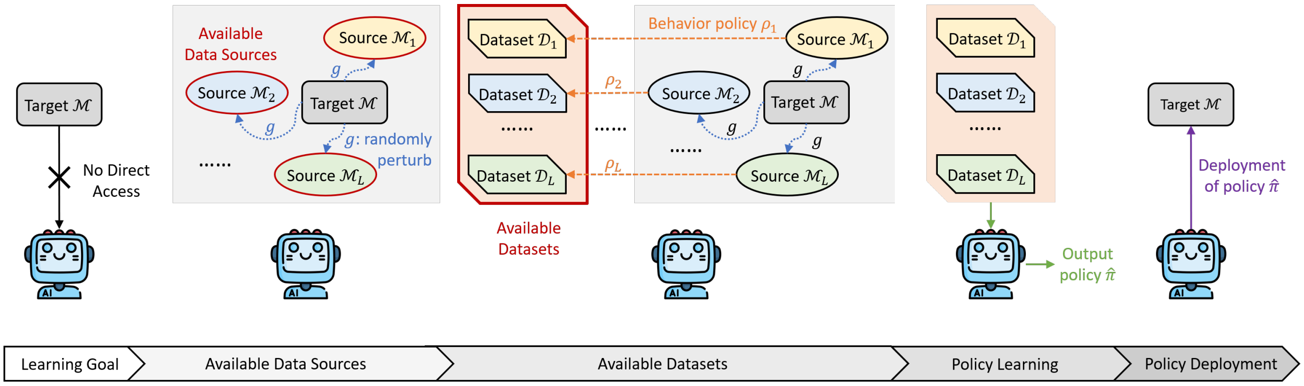
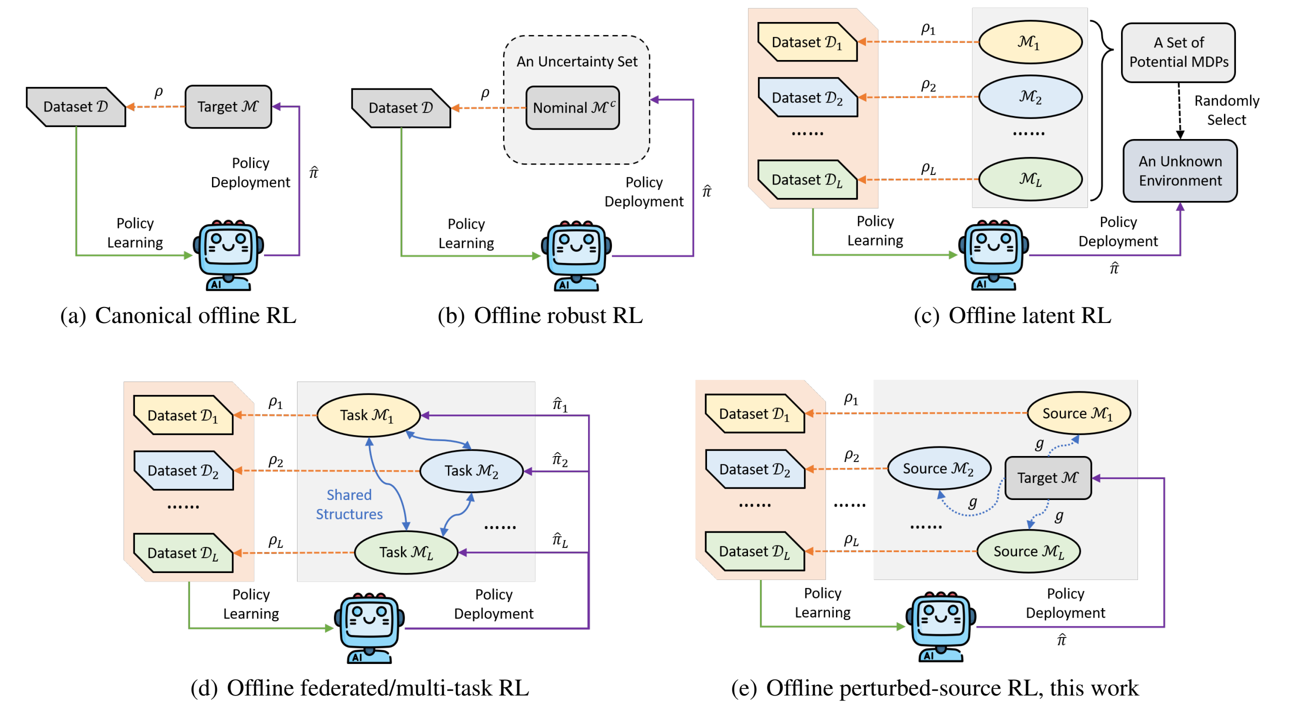


Figure 3. A step-by-step problem overview.

To further illustrate the considered setting, we specify a learning problem as follows.

- **Target MDP.** Consider the target MDP to be the dynamics of picking up a bowl of a certain shape and weight.
- **Source MDPs.** One possible scenario is that we have four other bowls: one larger than the target bowl, one smaller, one heavier, and one lighter. Each data source is randomly given one bowl from these four and then its picking-up trajectories are collected.
- **Target-source Relationship.** The most basic setting to be considered is that the averaged shape and weight of the four other bowls match those of the target bowl. Then, the averaged picking-up dynamics at each location and with each movement of these four bowls should conceivably match those of picking up the target bowl. Finally, since data sources are randomly selected from the four bowls, the expected source dynamics match the target dynamics, which is now stated as Assumption 2.1. With the proposed HetPEVI, one can learn how to pick the original bowl using these collected trajectories by picking other bowls.



(a) Canonical offline RL (b) Offline robust RL (c) Offline latent RL

(d) Offline federated/multi-task RL (e) Offline perturbed-source RL, this work

Figure 4. Graphical illustrations of related research topics, which differ from this work in their studied data sources and evaluation criteria.

A.2. Related Works

Reinforcement learning (Sutton & Barto, 2018) has seen much progress over the past few years, particularly in its theoretical understanding; see the recent monograph (Agarwal et al., 2019) for an overview. We will discuss the most related papers in the following, with a particular focus on the theoretical aspect as well as the offline setting. A compact summary of these topics and their comparisons with this work can be found in Table 1. Graphical illustrations can be found in Fig. 4.

Canonical offline RL. Inspired by empirical advances (Yu et al., 2020; Kumar et al., 2020), the principle of "pessimism" is incorporated and proved efficient for offline RL (Jin et al., 2021; Rashidinejad et al., 2021; Xie et al., 2021b; Li et al., 2022; Shi et al., 2022; Yin et al., 2022; Xiong et al., 2022; Xie et al., 2021a; Uehara & Sun, 2021; Zanette et al., 2021), which is also adopted in the design of HetPEVI and its generalizations. However, these works focus on the classical setting of learning with a dataset sampled exactly from the target task, which is rather restricted for practical applications.

Robust RL. Recently, a series of work (Zhou et al., 2021; Yang et al., 2021; Panaganti et al., 2022; Panaganti & Kalathil, 2022; Shi & Chi, 2022; Ma et al., 2022) has made theoretical advances on the topic of offline robust RL. In particular, a dataset from one data source, i.e., a nominal MDP, is collected, which is used by the learning agent to find an output policy. Then, the output policy is deployed to an uncertainty set around the nominal MDP, and its worst-case performance is adopted as the evaluation criteria. However, this work mainly considers multiple data sources while the learned policy is intended for deployment on the target task. Furthermore, in Section 8, we generalize the study of offline robust RL to consider that the available data are not from multiple perturbed versions of the nominal MDP instead of itself.

Latent RL. Another related topic is latent RL (Kwon et al., 2021; Zhou et al., 2022b). These existing studies are mainly in the online setting. Following the same spirit, the corresponding offline version would require datasets from a set of potential MDPs. Then, the learning agent aims to find a good policy that performs well on average in an unknown environment randomly selected from the aforementioned set of potential MDPs (which is often modeled to be related to a latent variable). Thus, although both latent RL and this work need to deal with multiple data sources, this work considers data sources that are perturbed versions of a target MDP while latent RL poses no relationships among data sources. Moreover, this work evaluates the learned policy on the target MDP while latent RL targets at performing well on the potential MDPs on average.

Federated and multi-task RL. Federated RL has attracted much attention recently (Dubey & Pentland, 2021; Jin et al., 2022), and Zhou et al. (2022a) studies its offline version. In particular, Zhou et al. (2022a) considers datasets from different MDPs at different sites, which share certain representations. The design is to leverage the shared structure to accelerate learning of each individual site MDP, i.e., find a good policy for each site MDP. Similarly, multi-task RL attempts to leverage common structures of multiple tasks to facilitate learning each individual task (Zhang & Wang, 2021; Lu et al., 2021; Yang et al., 2020; 2022). However, this work considers a stochastic relationship between data source MDPs (Assumption 2.1) instead of explicitly shared structures, while aiming to find one good policy for the target MDP (but not for the data source MDPs).

Meta-RL. The most related literature of this work falls in the research domain of "offline meta-RL" (Mitchell et al., 2021; Dorfman et al., 2021; Lin et al., 2022; Li et al., 2020), which however lacks rigorous theoretical understanding currently. Especially, the target MDP can be viewed as a learning target for the "meta-training" process of offline meta-RL (Mitchell et al., 2021), which aims to extract information from the available data of multiple sources. In addition to "meta-training", the empirically studied offline meta-RL systems often feature another step of "meta-testing", which further utilizes the learned information and applies them to a specific task. Thus, we believe this work may contribute to the theoretical understanding of offline meta-RL systems, especially the meta-training process, which may also serve as the foundation for studies of the meta-testing process.

Other related works. Another conceptually related work is Shrestha et al. (2020). In particular, it looks for similar state-action pairs with small distances in the dataset, which can be thought of as available data sources in this work. Then, the Lipschitz continuity assumption is posed, which serves a similar role as Assumption 2.1 to establish the connection between desired task information with the available datasets. From this perspective, the first term in Theorem 3.1 (Shrestha et al., 2020) can be interpreted as coming from the source uncertainty while the second term is from the sample uncertainty. However, we also note that the Lipschitz continuity assumption is a worst-case consideration that would not characterize the concentration of involving more data sources, which however is the key of this work.

Moreover, Jeong & Rothenhäusler (2022) provided a set of results to jointly characterize the sample and source uncertainties. One distinction is that it focuses on leveraging multiple estimators on one randomly perturbed data source, while this work targets aggregating information from multiple heterogeneously perturbed data sources. Despite the difference, the methods proposed in Jeong & Rothenhäusler (2022) may still be of value in the future study of offline RL with randomly perturbed data sources, especially its utilization of between-dataset information in quantifying uncertainties.

A.3. Future Works

Some discussions on future works are included in Section 9. A few other potential directions are discussed as follows.

Coverage Assumptions. While the collective coverage requirements of Assumptions 5.2, 7.2 and 8.2 are relatively weak, it is still of major interest to further explore how to perform offline RL (especially with heterogeneous data sources) under weaker conditions. This direction is particularly interesting with multiple data sources since the heterogeneity naturally enriches the data diversity.

Unknown Source Identities. This work considers the scenario where each data sample is known to belong to a particular source. One interesting direction is to investigate the scenarios without such information, i.e., unknown source identities. A potential solution is to first cluster the data samples and then adopt the algorithms proposed in this work. However, it is challenging to design clustering algorithms with provable performance guarantees. One candidate clustering technique is developed in (Kwon et al., 2021) for the study of latent MDP, which however relies on strong assumptions of prior knowledge about the source MDPs.

Personalization. As mentioned in the discussions of related work, this work can be viewed as targeting at the "meta-training" process of offline meta-RL (Mitchell et al., 2021), which extracts common knowledge from available data of multiple sources. While the extracted common knowledge has individual values, in many applications, an additional step of personalization is performed to further use such knowledge to benefit a specific task, which is called the "meta-testing" process of offline meta-RL (Mitchell et al., 2021). Based on this work, it would be valuable to further study how to perform such a personalization step with theoretical guarantees.

B. Proof of Theorem 3.1

In this section, we provide the proof of Theorem 3.1, which is inspired by Li et al. (2022) and Shi & Chi (2022) but more complicated as the datasets are collected from data sources instead of the target task itself.

B.1. Construction of hard problem instances

Let us first introduce two MDPs to be used in the following proofs:

$$\{\mathcal{N}^{\chi} = (H, \mathcal{S}, \mathcal{A}, \mathbb{Q}^{\chi} = \{\mathbb{Q}_h^{\chi} : h \in [H]\}, r = \{r_h : h \in [H]\}) : \chi \in \{0,1\}\},$$

where the state space is $\mathcal{S} = \{0, 1, \cdots, S-1\}$, and the action space is $\mathcal{A} = \{0, 1, 2\}$.

The transition kernel $\mathbb{Q}^0$ is defined as

$$\mathbb{Q}_1^0(s'|s,a) = \begin{cases} p'\mathbb{1}\{s'=0\} + (1-p')\mathbb{1}\{s'=1\} & \text{if } (s,a) = (0,0) \\ q'\mathbb{1}\{s'=0\} + (1-q')\mathbb{1}\{s'=1\} & \text{if } (s,a) = (0,1) \\ q\mathbb{1}\{s'=0\} + (1-q)\mathbb{1}\{s'=1\} & \text{if } (s,a) = (0,2) \\ \mathbb{1}\{s'=s\} & \text{if } s \geq 1 \end{cases}$$
$$\mathbb{Q}_h^0(s'|s,a) = \mathbb{1}\{s'=s\}, \qquad \forall (h,s,,a) \in \{2,\cdots,H\} \times \mathcal{S} \times \mathcal{A}.$$

The transition kernel $\mathbb{Q}^1$ is defined as

$$\mathbb{Q}_1^1(s'|s,a) = \begin{cases} q'\mathbb{1}\{s'=0\} + (1-q')\mathbb{1}\{s'=1\} & \text{if } (s,a) = (0,0) \\ p'\mathbb{1}\{s'=0\} + (1-p')\mathbb{1}\{s'=1\} & \text{if } (s,a) = (0,1) \\ q\mathbb{1}\{s'=0\} + (1-q)\mathbb{1}\{s'=1\} & \text{if } (s,a) = (0,2) \\ \mathbb{1}\{s'=s\} & \text{if } s \geq 1 \end{cases}$$
$$\mathbb{Q}_h^1(s'|s,a) = \mathbb{1}\{s'=s\}, \qquad \forall (h,s,,a) \in \{2,\cdots,H\} \times \mathcal{S} \times \mathcal{A}.$$

The parameters p', p, q and q' are set to be

$$p' = \frac{3}{4} - \frac{1}{H} + \Delta; \qquad p = p' - \alpha\Delta; \qquad q' = p' - \Delta = \frac{3}{4} - \frac{1}{H}; \qquad q = q' + \alpha\Delta$$

for some H, Δ and α obeying

$$\frac{1}{H} < \frac{1}{4}; \qquad \Delta \leq \frac{1}{8}; \qquad \alpha \leq \frac{1}{2}.$$

Thus,

$$\frac{7}{8} > p' > p > q > q' \geq \frac{1}{2}.$$

Moreover, for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, the reward function is defined as

$$r_h(s,a) = \begin{cases} 1 & \text{if } s = 0 \\ 0 & \text{otherwise.} \end{cases}$$

B.1.1. Construction of a collection of hard target MDPs

Let us introduce another two MDPs as target MDPs:

$$\left\{\mathcal{M}^{\phi} = \left(H, \mathcal{S}, \mathcal{A}, \mathbb{P}^{\phi} = \{\mathbb{P}_h^{\phi} : h \in [H], \}, r = \{r_h : h \in [H]\}\right) : \phi \in \{0,1\}\right\},$$

where the state space is $\mathcal{S} = \{0, 1, \cdots, S-1\}$, and the action space is $\mathcal{A} = \{0, 1, 2\}$.

The transition kernel $\mathbb{P}^0$ is defined as

$$\mathbb{P}^0_1(s'|s,a) = \begin{cases} p\mathbb{1}\{s'=0\} + (1-p)\mathbb{1}\{s'=1\} & \text{if } (s,a)=(0,0) \\ q\mathbb{1}\{s'=0\} + (1-q)\mathbb{1}\{s'=1\} & \text{if } (s,a)=(0,1) \\ q\mathbb{1}\{s'=0\} + (1-q)\mathbb{1}\{s'=1\} & \text{if } (s,a)=(0,2) \\ \mathbb{1}\{s'=s\} & \text{if } s \geq 1 \end{cases}$$

$$\mathbb{P}^0_h(s'|s,a) = \mathbb{1}\{s'=s\}, \qquad \forall (h,s,,a) \in \{2,\cdots,H\} \times \mathcal{S} \times \mathcal{A}.$$

The transition kernel $\mathbb{P}^1$ is defined as

$$\mathbb{P}^1_1(s'|s,a) = \begin{cases} q\mathbb{1}\{s'=0\} + (1-q)\mathbb{1}\{s'=1\} & \text{if } (s,a)=(0,0) \\ p\mathbb{1}\{s'=0\} + (1-p)\mathbb{1}\{s'=1\} & \text{if } (s,a)=(0,1) \\ q\mathbb{1}\{s'=0\} + (1-q)\mathbb{1}\{s'=1\} & \text{if } (s,a)=(0,2) \\ \mathbb{1}\{s'=s\} & \text{if } s \geq 1 \end{cases}$$

$$\mathbb{P}^1_h(s'|s,a) = \mathbb{1}\{s'=s\}, \qquad \forall (h,s,,a) \in \{2,\cdots,H\} \times \mathcal{S} \times \mathcal{A}.$$

It can be observed that

$$\mathcal{M}^0 = (1-\alpha)\cdot\mathcal{N}^0 + \alpha\cdot\mathcal{N}^1, \qquad \mathcal{M}^1 = \alpha\cdot\mathcal{N}^0 + (1-\alpha)\cdot\mathcal{N}^1,$$

where the weighted average is w.r.t. rewards and transition kernels.

B.1.2. Construction of a source MDP generation distribution

If the target MDP is $\mathcal{M}^0$, then with probability $1-\alpha$, the generated source MDP is $\mathcal{N}^0$, and with probability α, the generated source MDP is $\mathcal{N}^1$. If the target MDP is $\mathcal{M}^0$, then with probability α, the generated source MDP is $\mathcal{N}^0$, and with probability $1-\alpha$, the generated source MDP is $\mathcal{N}^1$.

B.1.3. Construction of the offline dataset

In the environment $\mathcal{N}^\chi$, a batch dataset is generated consisting of K independent sample trajectories each of length H based on an initial distribution

$$\xi^d(s) = \mu(s),$$

where

$$\mu(s) = \frac{1}{CS}\mathbb{1}\{s=0\} + \left(1 - \frac{1}{CS}\right)\mathbb{1}\{s=1\}, \qquad \text{with } \frac{1}{CS} \leq \frac{1}{4},$$

and a behavior policy, which is specified in the following.

Good behavior policy. The good behavior policy ρ^g uniformly selects actions $\{0,1\}$ as follows:

$$\rho^g_h(a|s) = \frac{1}{2}, \qquad \forall (s,a,h) \in \mathcal{S} \times \{0,1\} \times [H].$$

It turns out that for any MDP $\mathcal{N}^\chi$, the occupancy distributions of the above batch dataset admit the following characterization:

$$d_1^{\rho^g,\mathcal{N}^\chi}(0,a;\xi^d) = \frac{1}{2}\mu(0), \qquad \forall a \in \{0,1\}; \qquad \frac{\mu(s)}{4} \leq d_h^{\rho^g,\mathcal{N}^\chi}(s,a;\xi^d) \leq \mu(s), \qquad \forall (s,a,h) \in \mathcal{S} \times \{0,1\} \times [H].$$

In particular, for any $\mathcal{N}^\chi$ with $\chi \in \{0,1\}$ and the initial distribution as $\xi^d(s) = \mu(s)$, we have that

$$d_1^{\rho^g,\mathcal{N}^\chi}(s;\xi^d) = \mu(s), \qquad \forall s \in \mathcal{S},$$

which leads to

$$d_1^{\rho^g,\mathcal{N}^\chi}(0,0;\xi^d) = \mu(0)\rho_1^g(0|0) = \frac{\mu(0)}{2}; \qquad d_1^{\rho^g,\mathcal{N}^\chi}(0,1;\xi^d) = \mu(0)\rho_1^g(1|0) = \frac{\mu(0)}{2}.$$

The state occupancy distribution at step $h = 2$ obeys that

$$d_2^{\rho^g,\mathcal{N}^\chi}(0;\xi^d) = \mu(0)\left[\rho_1^g(\chi|0)p' + \rho_1^g(1-\chi|0)q' + \rho_1^g(2|0)q\right] = \frac{\mu(0)(p'+q')}{2},$$

and

$$d_2^{\rho^g,\mathcal{N}^\chi}(1;\xi^d) = \mu(1) + \mu(0)\left[\rho_1^g(\chi|0)(1-p') + \rho_1^g(1-\chi|0)(1-q') + \rho_1^g(2|0)(1-q)\right] = \mu(1) + \frac{\mu(0)(2-p'-q')}{2}.$$

The above results can be further bounded as

$$\frac{\mu(0)}{2} \le d_2^{\rho^g,\mathcal{N}^\chi}(0;\xi^d) \le \mu(0), \qquad \mu(1) \le d_2^{\rho^g,\mathcal{N}^\chi}(1;\xi^d) \le 2\mu(1),$$

which leads to that

$$\begin{aligned}
\frac{\mu(0)}{4} &\le d_h^{\rho^g,\mathcal{N}^\chi}(0,a;\xi^d) \le \frac{\mu(0)}{2}, \qquad \forall (a,h) \in \{0,1\} \times [2,H];\\
\frac{\mu(1)}{2} &\le d_h^{\rho^g,\mathcal{N}^\chi}(1,a;\xi^d) \le \mu(1), \qquad \forall (a,h) \in \{0,1\} \times [2,H].
\end{aligned}$$

Bad behavior policy. The bad behavior policy ρ^b always selects action 2 as follows:

$$\rho_h^b(2|s) = 1, \qquad \forall (s,a,h) \in \mathcal{S} \times \{2\} \times [H].$$

Correspondingly, for any MDP $\mathcal{N}^\chi$, it is easy to observe that the occupancy distributions of the above batch dataset follow that

$$d_h^{\rho^b,\mathcal{N}^\chi}(s,a;\xi^d) = 0, \qquad \forall (s,a,h) \in \mathcal{S} \times \{0,1\} \times [H].$$

We then specify the initial distributions of all data sources as ξ^d, the behavior policies of the first $L^\ddagger$ data sources as the good ones, i.e., ρ^g, and the behavior policies of the other $L - L^\ddagger$ as the bad ones, i.e., ρ^b.

B.1.4. Value functions and optimal policies

We choose the initial state distribution to be tested on as

$$\xi(s) = \begin{cases} 1, & \text{if } s = 0\\ 0, & \text{otherwise.}\end{cases}$$

Then, the following lemma can be established.

Lemma B.1. *For any $\phi \in \{0,1\}$ and any policy π, it holds that*

$$V_1^{\pi,\mathcal{M}^\phi}(0) = 1 + \pi_1(\phi|0)p(H-1) + \pi_1(1-\phi|0)q(H-1) + \pi_1(2|0)q(H-1).$$

In addition, there exists an optimal policy $\pi^{*,\mathcal{M}^\phi}$ such that its optimal value functions obey

$$\begin{aligned}
& V_1^{\pi^*,\mathcal{M}^\phi}(0) = 1 + p(H-1),\\
\forall h \in [2,H]: \quad & V_h^{\pi^*,\mathcal{M}^\phi}(0) = H - h + 1,\\
\forall h \in [H]: \quad & \pi_h^{*,\mathcal{M}^\phi}(\phi|0) = 1, \qquad \pi_h^{*,\mathcal{M}^\phi}(\phi|1) = 1, \qquad V_h^{\pi^*,\mathcal{M}^\phi}(1) = 0,
\end{aligned}$$

where we denote $V_h^{\pi^*,\mathcal{M}^\phi}(s) := V_h^{\pi^{*,\mathcal{M}^\phi},\mathcal{M}^\phi}(s)$ for simplicity.

Furthermore, it holds that

$$L^\dagger = L^\ddagger; \qquad C^\dagger \in [C, 4C].$$

Proof. For any policy ϕ, it can be easily observed that

$$\begin{aligned} V_h^{\pi,\mathcal{M}^\phi}(0) &= H-h+1, \qquad \forall h \in [2,H]; \\ V_h^{\pi,\mathcal{M}^\phi}(s) &= 0, \qquad \forall (s,h) \in \{1,\cdots,S-1\} \times [H], \end{aligned}$$

which immediately indicates that

$$\begin{aligned} & V_1^{\pi^*,\mathcal{M}^\phi}(0) = 1 + p(H-1), \\ \forall h \in [2,H]: \qquad & V_h^{\pi^*,\mathcal{M}^\phi}(0) = H-h+1, \\ \forall h \in [H]: \qquad & \pi_h^{*,\mathcal{M}^\phi}(\phi|0) = 1, \qquad \pi_h^{*,\mathcal{M}^\phi}(\phi|1) = 1, \qquad V_h^{\pi^*,\mathcal{M}^\phi}(1) = 0. \end{aligned}$$

For a policy π, it further holds that

$$\begin{aligned} V_1^{\pi,\mathcal{M}^\phi}(0) &= \mathbb{E}_{a\sim\pi_1(\cdot|0)}\left[r_h(0,a) + \left(\mathbb{P}_1 V_1^{\pi,\mathcal{M}^\phi}\right)(0,a)\right] \\ &= 1 + \pi_1(\phi|0)\left(pV_2^{\pi,\mathcal{M}^\phi}(0) + (1-p)V_2^{\pi,\mathcal{M}^\phi}(1)\right) \\ &+ \pi_1(1-\phi|0)\left(qV_2^{\pi,\mathcal{M}^\phi}(0) + (1-q)V_2^{\pi,\mathcal{M}^\phi}(1)\right) \\ &+ \phi_1(2|0)\left(qV_2^{\pi,\mathcal{M}^\phi}(0) + (1-q)V_2^{\pi,\mathcal{M}^\phi}(1)\right) \\ &= 1 + \pi_1(\phi|0)p(H-1) + \pi_1(1-\phi|0)q(H-1) + \pi_1(2|0)q(H-1). \end{aligned}$$

Moreover, it holds that

$$\begin{aligned} d_h^{\pi^*,\mathcal{M}^\phi,\mathcal{M}^\phi}(0,\phi;\xi) = d_h^{\pi^*,\mathcal{M}^\phi,\mathcal{M}^\phi}(0;\xi) &= \mathbf{P}\left\{s_h = 0 | s_{h-1} \sim d_{h-1}^{\pi^*,\mathcal{M}^\phi,\mathcal{M}^\phi}(\cdot|\xi), a_h \sim \pi_h^{*,\mathcal{M}^\phi}(\cdot|s_h)\right\} \\ &= d_{h-1}^{\pi^*,\mathcal{M}^\phi,\mathcal{M}^\phi}(0;\xi) = \cdots = d_2^{\pi^*,\mathcal{M}^\phi,\mathcal{M}^\phi}(0;\xi) \ge p\xi(0) \ge \frac{1}{2}. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{\min\left\{d_h^{\pi^*,\mathcal{M}^\phi}(0,\phi;\xi), \frac{1}{S}\right\}}{d_h^{\rho^g,\mathcal{N}^\chi}(0,\phi;\xi^d)} &= \frac{1/S}{d_h^{\rho^g,\mathcal{N}^\chi}(0,\phi;\xi^d)} \in \left[\frac{1/S}{\mu(0)}, \frac{1/S}{\mu(0)/4}\right] = [C, 4C]; \\ \frac{\min\left\{d_h^{\pi^*,\mathcal{M}^\phi}(1,\phi;\xi), \frac{1}{S}\right\}}{d_h^{\rho^g,\mathcal{N}^\chi}(1,\phi;\xi^d)} &\le \frac{1/S}{\mu(1)/2} = \frac{2}{S(1-\frac{1}{CS})} \le \frac{8}{3S} \le \frac{2C}{3}, \end{aligned}$$

which concludes the proof. □

B.2. Establishing the minimax lower bound

B.2.1. Converting the goal to estimate ϕ

We choose α and Δ such that

$$(H-1)(1-2\alpha)\Delta \ge 2\varepsilon.$$

Then, with the selected ξ, it holds that

$$\begin{aligned} & V_1^{\pi^*,\mathcal{M}^\phi}(\xi) - V_1^{\hat{\pi},\mathcal{M}^\phi}(\xi) \\ &= V_1^{\pi^*,\mathcal{M}^\phi}(0) - V_1^{\hat{\pi},\mathcal{M}^\phi}(0) \\ &= 1 + p(H-1) - [1 + \hat{\pi}_1(\phi|0)p(H-1) + \hat{\pi}_1(1-\phi|0)q(H-1) + \hat{\pi}_1(2|0)q(H-1)] \\ &= p(H-1) - \hat{\pi}_1(\phi|0)p(H-1) - \hat{\pi}_1(1-\phi|0)q(H-1) - \hat{\pi}_1(2|0)q(H-1) \end{aligned}$$

$$
\begin{aligned}
&= (H-1)(p-q)(1-\hat{\pi}_1(\phi|0)) \\
&= (H-1)(1-2\alpha)\Delta(1-\hat{\pi}_1(\phi|0)) \\
&\geq 2\varepsilon(1-\hat{\pi}_1(\phi|0)).
\end{aligned}
$$

Suppose that for any $\phi \in \{0,1\}$,

$$\mathbf{P}_\phi\left\{V_1^{\pi^*,\mathcal{M}^\phi}(\xi) - V_1^{\hat{\pi},\mathcal{M}^\phi}(\xi) \leq \varepsilon\right\} \geq \frac{7}{8},$$

we necessarily have that

$$\mathbf{P}_\phi\left\{\hat{\phi} = \phi\right\} = \mathbf{P}_\phi\left\{\hat{\pi}_1(\phi|0) \geq \frac{1}{2}\right\} \geq \frac{7}{8},$$

where

$$\hat{\phi} = \begin{cases} \phi, & \text{if } \hat{\pi}_1(\phi|0) \geq \frac{1}{2} \\ 1-\phi & \text{if } \hat{\pi}_1(\phi|0) < \frac{1}{2}. \end{cases}$$

B.2.2. Probability of error in testing two hypotheses

We focus on differentiating the two hypotheses $\phi \in \{0,1\}$. Towards this, consider the minimax probability of error defined as follows:

$$p_e := \inf_{\psi} \max\left\{\mathbf{P}_0(\psi \neq 0), \mathbf{P}_1(\psi \neq 1)\right\},$$

where the infimum is taken over all possible tests ψ constructed from the batch dataset. Then, following the standard results (Theorem 2.2, Tsybakov (2009)), it holds that

$$p_e \geq \frac{1}{4}\exp\left(-\mathrm{KL}\left(\upsilon^{\mathcal{M}^0} || \upsilon^{\mathcal{M}^1}\right)\right),$$

where $\upsilon^{\mathcal{M}^\phi}(\cdot)$ is the distribution of the sampled dataset with the target MDP as $\mathcal{M}^\phi$. Furthermore, it holds that

$$
\begin{aligned}
\mathrm{KL}\left(\upsilon^{\mathcal{M}^0} || \upsilon^{\mathcal{M}^1}\right) &= \sum_{\mathcal{D}} \upsilon^{\mathcal{M}^0}(\mathcal{D}) \log\left(\frac{\upsilon^{\mathcal{M}^0}(\mathcal{D})}{\upsilon^{\mathcal{M}^1}(\mathcal{D})}\right) \\
&\overset{(i)}{=} \sum_{l\in[L]}\sum_{\mathcal{D}_l} \upsilon_l^{\mathcal{M}^0}(\mathcal{D}_l) \log\left(\frac{\upsilon_l^{\mathcal{M}^0}(\mathcal{D}_l)}{\upsilon_l^{\mathcal{M}^1}(\mathcal{D}_l)}\right) \\
&\overset{(ii)}{=} \sum_{l\in[L^\dagger]}\sum_{\mathcal{D}_l} \upsilon_l^{\mathcal{M}^0}(\mathcal{D}_l) \log\left(\frac{\upsilon_l^{\mathcal{M}^0}(\mathcal{D}_l)}{\upsilon_l^{\mathcal{M}^1}(\mathcal{D}_l)}\right) \\
&\overset{(iii)}{=} \sum_{l\in[L^\dagger]}\sum_{\mathcal{D}_l} \left(\upsilon_l^{\mathcal{N}^0}(\mathcal{D}_l)(1-\alpha) + \upsilon_l^{\mathcal{N}^1}(\mathcal{D}_l)\alpha\right) \log\left(\frac{\upsilon_l^{\mathcal{N}^0}(\mathcal{D}_l)(1-\alpha) + \upsilon_l^{\mathcal{N}^1}(\mathcal{D}_l)\alpha}{\upsilon_l^{\mathcal{N}^0}(\mathcal{D}_l)\alpha + \upsilon_l^{\mathcal{N}^1}(\mathcal{D}_l)(1-\alpha)}\right) \\
&\overset{(iv)}{\leq} \min\left\{\text{term (I)}, \text{term (II)}\right\}.
\end{aligned}
$$

where equation (i) is from the definition of $\upsilon_l^{\mathcal{M}^\phi}(\cdot)$ as the distribution of the sampled dataset at source l with the target MDP as $\mathcal{M}^\phi$ and the property of KL-divergence for product measures. Equation (ii) is from the fact that the distribution of the last $L - L^\ddagger$ datasets are the same. Equation(iii) leverages the definition of $\upsilon_l^{\mathcal{N}^\psi}(\cdot)$ as the distribution of the sampled dataset at source l with the source MDP as $\mathcal{N}^\psi$ and the two-step dataset generation procedure (i.e., first randomly perturb target MDP as source MDP, and then randomly sample data from the source MDP). Inequality (iv) is from the log-sum inequality and the following definition:

$$\text{term (I)} := \sum_{l\in[L^\dagger]}\sum_{\mathcal{D}_l} \upsilon_l^{\mathcal{N}^0}(\mathcal{D}_l)(1-\alpha)\log\left(\frac{\upsilon_l^{\mathcal{N}^0}(\mathcal{D}_l)(1-\alpha)}{\upsilon_l^{\mathcal{N}^0}(\mathcal{D}_l)\alpha}\right) + \sum_{l\in[L^\dagger]}\sum_{\mathcal{D}_l} \upsilon_l^{\mathcal{N}^1}(\mathcal{D}_l)\alpha\log\left(\frac{\upsilon_l^{\mathcal{N}^1}(\mathcal{D}_l)\alpha}{\upsilon_l^{\mathcal{N}^1}(\mathcal{D}_l)(1-\alpha)}\right)$$

$$\begin{aligned}
&= \sum_{l\in[L^\dagger]} (1-\alpha)\log\left(\frac{(1-\alpha)}{\alpha}\right) + \sum_{l\in[L^\dagger]} \alpha\log\left(\frac{\alpha}{(1-\alpha)}\right)\\
&= L^\dagger(1-2\alpha)\log\left(\frac{(1-\alpha)}{\alpha}\right);\\
\text{term (II)} := &\sum_{l\in[L^\dagger]}\sum_{\mathcal{D}_l} \upsilon_l^{\mathcal{N}^0}(\mathcal{D}_l)(1-\alpha)\log\left(\frac{\upsilon_l^{\mathcal{N}^0}(\mathcal{D}_l)(1-\alpha)}{\upsilon_l^{\mathcal{N}^1}(\mathcal{D}_l)(1-\alpha)}\right) + \sum_{l\in[L^\dagger]}\sum_{\mathcal{D}_l} \upsilon_l^{\mathcal{N}^1}(\mathcal{D}_l)\alpha\log\left(\frac{\upsilon_l^{\mathcal{N}^1}(\mathcal{D}_l)\alpha}{\upsilon_l^{\mathcal{N}^0}(\mathcal{D}_l)\alpha}\right)\\
&= \sum_{l\in[L^\dagger]}\sum_{\mathcal{D}_l} \upsilon_l^{\mathcal{N}^0}(\mathcal{D}_l)(1-\alpha)\log\left(\frac{\upsilon_l^{\mathcal{N}^0}(\mathcal{D}_l)}{\upsilon_l^{\mathcal{N}^1}(\mathcal{D}_l)}\right) + \sum_{l\in[L^\dagger]}\sum_{\mathcal{D}_l} \upsilon_l^{\mathcal{N}^1}(\mathcal{D}_l)\alpha\log\left(\frac{\upsilon_l^{\mathcal{N}^1}(\mathcal{D}_l)}{\upsilon_l^{\mathcal{N}^0}(\mathcal{D}_l)}\right)\\
&= \alpha L^\dagger K\cdot \mathrm{KL}\left(\upsilon_1^{\mathcal{N}^0}(\tau)||\upsilon_1^{\mathcal{N}^1}(\tau)\right) + (1-\alpha)L^\dagger K\cdot \mathrm{KL}\left(\upsilon_1^{\mathcal{N}^1}(\tau)||\upsilon_1^{\mathcal{N}^0}(\tau)\right),
\end{aligned}$$

where notation τ refers to a complete H-step trajectory.

Furthermore, it holds that

$$\begin{aligned}
\mathrm{KL}\left(\upsilon_1^{\mathcal{N}^0}(\tau)||\upsilon_1^{\mathcal{N}^1}(\tau)\right) &= \frac{1}{2}\mu(0)\sum_{a\in\{0,1\}}\mathrm{KL}\left(\mathbb{Q}^0(\cdot|0,a)||\mathbb{Q}^1(\cdot|0,a)\right) \le \mu(0)\frac{(p'-q')^2}{p'(1-p')} = \mu(0)\frac{\Delta^2}{p'(1-p')}\\
\mathrm{KL}\left(\upsilon_1^{\mathcal{N}^1}(\tau)||\upsilon_1^{\mathcal{N}^0}(\tau)\right) &= \frac{1}{2}\mu(0)\sum_{a\in\{0,1\}}\mathrm{KL}\left(\mathbb{Q}^1(\cdot|0,a)||\mathbb{Q}^0(\cdot|0,a)\right) \le \mu(0)\frac{(p'-q')^2}{p'(1-p')} = \mu(0)\frac{\Delta^2}{p'(1-p')},
\end{aligned}$$

where the inequality is from the fact that $\mathcal{N}^0$ and $\mathcal{N}^1$ only differ at state-action pairs $(0,0)$ and $(0,1)$, and the inequality is from a basic property of KL divergence (see Lemma 10 in (Li et al., 2022)).

With $\varepsilon < \frac{H}{64}$, if it is designed that

$$\alpha = \frac{1}{2} - \frac{16\varepsilon}{H} \in (\frac{1}{4},\frac{1}{2}), \qquad \Delta = \frac{1}{8}$$

it holds that

$$(H-1)(1-2\alpha)\Delta = (H-1)\cdot\frac{32\varepsilon}{H}\cdot\frac{1}{8} \ge 2\varepsilon,$$

and

$$(1-2\alpha)\log\left(\frac{1-\alpha}{\alpha}\right) = (1-2\alpha)\log\left(1+\frac{1-2\alpha}{\alpha}\right) \le \frac{(1-2\alpha)^2}{\alpha} \le \frac{4096\varepsilon^2}{H^2}.$$

Thus, if

$$L^\dagger \le \frac{H^2\log(2)}{4096\varepsilon^2},$$

it holds that

$$p_e \ge \frac{1}{4}\exp\left(-\mathrm{KL}\left(v^0||v^1\right)\right) \ge \frac{1}{4}\exp\left(-L^\dagger\cdot\mathrm{KL}\left(1-\alpha||\alpha\right)\right) \ge \frac{1}{4}\exp\left(-\frac{H^2\log(2)}{4096\varepsilon^2}\cdot\frac{4096\varepsilon^2}{H^2}\right) = \frac{1}{8}.$$

Similarly with $\varepsilon \le \frac{H}{64}$, if it is designed that

$$\alpha = \frac{1}{4}, \qquad \Delta = \frac{8\varepsilon}{H} \le \frac{1}{8}$$

it holds that

$$(H-1)(1-2\alpha)\Delta = (H-1)\cdot\frac{1}{2}\cdot\frac{8\varepsilon}{H} \ge 2\varepsilon,$$

and

$$\frac{\Delta^2}{p'(1-p')} \le 16\Delta^2 = \frac{1024\varepsilon^2}{H^2}.$$

Thus, if

$$L^\dagger K \le \frac{H^2\log(2)}{1024\varepsilon^2\mu(0)},$$

it holds that

$$\begin{aligned}
p_e &\ge \frac{1}{4}\exp\left(-\mathrm{KL}\left(\upsilon^0||\upsilon^1\right)\right)\\
&\ge \frac{1}{4}\exp\left(-\alpha L^\dagger K\cdot \mathrm{KL}\left(\upsilon_1^{\mathcal{N}^0}(\tau)||\upsilon_1^{\mathcal{N}^1}(\tau)\right) + (1-\alpha)L^\dagger K\cdot \mathrm{KL}\left(\upsilon_1^{\mathcal{N}^1}(\tau)||\upsilon_1^{\mathcal{N}^0}(\tau)\right)\right)\\
&\ge \frac{1}{4}\exp\left(-\frac{H^2\log(2)}{1024\varepsilon^2\mu(0)}\cdot\mu(0)\frac{1024\varepsilon^2}{H^2}\right) = \frac{1}{8}.
\end{aligned}$$

B.2.3. PUTTING THINGS TOGETHER

Finally, suppose that there exists an estimator $\hat{\pi}$ such that

$$\mathbf{P}_0\left\{V_1^{\pi^*,\mathcal{M}^0}(\xi) - V_1^{\hat{\pi},\mathcal{M}^0}(\xi) \le \varepsilon\right\} \ge \frac{7}{8} \quad \text{and} \quad \mathbf{P}_1\left\{V_1^{\pi^*,\mathcal{M}^1}(\xi) - V_1^{\hat{\pi},\mathcal{M}^1}(\xi) \le \varepsilon\right\} \ge \frac{7}{8}.$$

The estimator $\hat{\phi}$ must satisfy that

$$\mathbf{P}_0\left\{\hat{\phi}\neq 0\right\} \le \frac{1}{8} \quad \text{and} \quad \mathbf{P}_1\left\{\hat{\phi}\neq 1\right\} \le \frac{1}{8},$$

which cannot happen if

$$L^\dagger \le \frac{H^2\log(2)}{4096\varepsilon^2}$$

or

$$L^\dagger K \le \frac{H^2 C^\dagger S\log(2)}{4096\varepsilon^2} \le \frac{H^2 CS\log(2)}{1024\varepsilon^2} = \frac{H^2\log(2)}{1024\varepsilon^2\mu(0)}$$

under the correspondingly designed scenarios.

C. Details of HetPEVI and Proof of Theorem 5.1

C.1. The Subsampling Procedure

The detailed subsampling procedure can be found in Li et al. (2022). Here we state their obtained main result as follows.

Lemma C.1. *With probability at least $1-8\delta$, the output dataset from the two-fold subsampling scheme in Li et al. (2022) is distributionally equivalent to independently sampled from the data source MDP and*

$$N_{h,l}(s,a) \ge \frac{Kd_h^{\rho_l,\mathcal{M}_l}(s,a)}{8} - 5\sqrt{Kd_h^{\rho_l,\mathcal{M}_l}(s,a)\log\left(\frac{KHL}{\delta}\right)}.$$

for all $(h,s,a,l)\in[H]\times\mathcal{S}\times\mathcal{A}\times[L]$.

C.2. Core Lemmas

Lemma C.2. *For all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, and any function $V : \mathcal{S} \to [0, H]$ independent of $\hat{\mathbb{P}}_h$, with probability at least $1 - 4\delta$, it holds that*

$$\left|\left(\hat{\mathbb{B}}_h V\right)(s,a) - \left(\mathbb{B}_h V\right)(s,a)\right| \leq \Gamma_h(s,a) := c\sqrt{\sum_{l\in\hat{\mathcal{L}}_h(s,a)} \frac{H^2\log(SAH/\delta)}{\left(\hat{L}_h(s,a)\right)^2 N_{h,l}(s,a)}} + c\sqrt{\frac{H^2\log(SAH/\delta)}{\hat{L}_h(s,a)}}.$$

Proof. For a fixed (s, a, h), it holds that

$$\begin{aligned}
&\left|\left(\hat{\mathbb{B}}_h V\right)(s,a) - \left(\mathbb{B}_h V\right)(s,a)\right| \\
&= \left|\hat{r}_h(s,a) + \left(\hat{\mathbb{P}}_h V\right)(s,a) - r_h(s,a) - \left(\mathbb{P}_h V\right)(s,a)\right| \\
&= \left|\sum_{l\in\hat{\mathcal{L}}_h(s,a)} \frac{1}{\hat{L}_h(s,a)}\left(r_{h,l}(s,a) + \left(\hat{\mathbb{P}}_{h,l}V\right)(s,a)\right) - \left(r_h(s,a) + \mathbb{P}_h V(s,a)\right)\right| \\
&\leq \left|\sum_{l\in\hat{\mathcal{L}}_h(s,a)} \frac{1}{\hat{L}_h(s,a)}\left(r_{h,l}(s,a) + \left(\hat{\mathbb{P}}_{h,l}V\right)(s,a)\right) - \sum_{l\in\hat{\mathcal{L}}_h(s,a)} \frac{1}{\hat{L}_h(s,a)}\left(r_{h,l}(s,a) + \left(\mathbb{P}_{h,l}V\right)(s,a)\right)\right| \\
&+ \left|\sum_{l\in\hat{\mathcal{L}}_h(s,a)} \frac{1}{\hat{L}_h(s,a)}\left(r_{h,l}(s,a) + \left(\mathbb{P}_{h,l}V\right)(s,a)\right) - \left(r_h(s,a) + \mathbb{P}_h V(s,a)\right)\right| \\
&\leq \sqrt{\sum_{l\in\hat{\mathcal{L}}_h(s,a)} \frac{2H^2\log(SAH/\delta)}{\left(\hat{L}_h(s,a)\right)^2 N_{h,l}(s,a)}} + \sqrt{\frac{2H^2\log(SAH/\delta)}{\hat{L}_h(s,a)}}
\end{aligned}$$

where the last step holds with probability at least $1 - 4\delta/(SAH)$ due to Hoeffding inequality. The lemma can then be established via a union bound over $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$. □

C.3. Main Proofs

In the following, we establish Theorem 5.1. The proof framework is inspired by Li et al. (2022) but is uniquely adapted to handle randomly perturbed data sources.

Step 1: establishing the pessimism property. Armed with Lemma C.2, with probability at least $1 - \delta$, the following relation holds

$$\hat{Q}_h(s,a) \leq Q_h^{\hat{\pi},\mathcal{M}}(s,a), \qquad \hat{V}_h(s,a) \leq V_h^{\hat{\pi},\mathcal{M}}(s,a), \qquad \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]. \tag{3}$$

Towards this, it is first observed that

$$\hat{Q}_{H+1}(s,a) = Q_{H+1}^{\hat{\pi},\mathcal{M}}(s,a) = 0, \qquad \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

Then, suppose that $\hat{Q}_{h+1}(s,a) \leq Q_{h+1}^{\hat{\pi},\mathcal{M}}(s,a)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ at some step $h \in [H]$, we can observe that by the update rule in HetPEVI, it holds that

$$0 \leq \hat{V}_{h+1}(s) = \max_{a\in\mathcal{A}} \hat{Q}_{h+1}(s,a) \leq \max_{a\in\mathcal{A}} Q_{h+1}^{\hat{\pi},\mathcal{M}}(s,a) = V_{h+1}^{\hat{\pi},\mathcal{M}}(s) \leq H, \qquad \forall s \in \mathcal{S},$$

If $\hat{Q}_h(s,a) = 0$, the claim naturally holds. If not, we can obtain that

$$\begin{aligned}
\hat{Q}_h(s,a) &\leq \left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \Gamma_h(s,a) \\
&\leq \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a) + \left|\left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a)\right| - \Gamma_h(s,a)
\end{aligned}$$

$$\overset{(i)}{\leq} \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a) \overset{(ii)}{\leq} \left(\mathbb{B}_h V_{h+1}^{\hat{\pi},\mathcal{M}}\right)(s,a) = Q_h^{\hat{\pi},\mathcal{M}}(s,a).$$

The above inequality (i) is from Lemma C.2 and leverages the fact that $\hat{V}_{h+1}(\cdot)$ is independent of $\hat{\mathbb{P}}_h$ and takes value in $[0,H]$. Inequality (ii) is from the obtained fact that $\hat{V}_{h+1}(s) \leq V_{h+1}^{\hat{\pi},\mathcal{M}}(s)$. The desired claim Eqn. (3) can be verified by induction.

Step 2: bounding the performance difference. From Eqn. (3), we can observe that

$$0 \leq V_h^{\pi^*,\mathcal{M}}(s) - V_h^{\hat{\pi},\mathcal{M}}(s) \leq V_h^{\pi^*,\mathcal{M}}(s) - \hat{V}_h(s) \leq Q_h^{\pi^*,\mathcal{M}}(s,\pi_h^*(s)) - \hat{Q}_h(s,\pi_h^*(s)).$$

With

$$\begin{aligned} Q_h^{\pi^*,\mathcal{M}}(s,\pi_h^*(s)) &= \left(\mathbb{B}_h V_{h+1}^{*,\mathcal{M}}\right)(s,\pi_h^*(s)) \\ \hat{Q}_h(s,\pi_h^*(s)) &= \max\left\{\left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,\pi_h^*(s)) - \Gamma_h(s,\pi_h^*(s)), 0\right\}, \end{aligned}$$

we can further obtain that

$$\begin{aligned} V_h^{\pi^*,\mathcal{M}}(s) - \hat{V}_h(s) &\leq \left(\mathbb{B}_h V_{h+1}^{*,\mathcal{M}}\right)(s,\pi_h^*(s)) - \left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,\pi_h^*(s)) + \Gamma_h(s,\pi_h^*(s)) \\ &= \left(\mathbb{B}_h V_{h+1}^{*,\mathcal{M}}\right)(s,\pi_h^*(s)) - \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,\pi_h^*(s)) \\ &\quad + \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,\pi_h^*(s)) - \left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,\pi_h^*(s)) + \Gamma_h(s,\pi_h^*(s)) \\ &\overset{(i)}{\leq} \left(\mathbb{B}_h V_{h+1}^{*,\mathcal{M}}\right)(s,\pi_h^*(s)) - \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,\pi_h^*(s)) + 2\Gamma_h(s,\pi_h^*(s)) \\ &= \left(\mathbb{P}_h V_{h+1}^{*,\mathcal{M}}\right)(s,\pi_h^*(s)) - \left(\mathbb{P}_h \hat{V}_{h+1}\right)(s,\pi_h^*(s)) + 2\Gamma_h(s,\pi_h^*(s)) \end{aligned}$$

where inequality (i) holds with probability at least $1-\delta$ according to Lemma C.2. If applying the above argument iteratively, we can further obtain that

$$V_1^{\pi^*,\mathcal{M}}(\xi) - \hat{V}_1(\xi) \leq 2\sum_{h\in[H]}\sum_{s\in\mathcal{S}} d_h^{\pi^*,\mathcal{M}}(s)\Gamma_h(s,\pi_h^*(s)).$$

Step 3: completing the proof with concentrability. Let us consider $(s,h) \in \mathcal{S}\times[H]$ such that $d_h^{\pi^*,\mathcal{M}}(s) > 0$. We can then obtain that for all $l \in \mathcal{L}_h(s,\pi_h^*(s))$,

$$N_{h,l}(s,\pi_h^*(s)) \overset{(i)}{\geq} \frac{K d_h^{\rho_l,\mathcal{M}_l}(s,\pi_h^*(s))}{8} - 5\sqrt{K d_h^{\rho_l,\mathcal{M}_l}(s,\pi_h^*(s))\log\left(\frac{KHL}{\delta}\right)} \overset{(ii)}{\geq} \frac{K d_h^{\rho_l,\mathcal{M}_l}(s,\pi_h^*(s))}{16} \overset{(iii)}{\geq} 1.$$

where inequality (i) is from Lemma C.1; and inequalities (ii) and (iii) are from the condition that

$$K \geq \frac{c\log\left(\frac{KH}{\delta}\right)}{d^{\min}} \geq \frac{c\log\left(\frac{KH}{\delta}\right)}{d_h^{\rho_l,\mathcal{M}_l}(s,\pi_h^*(s))}.$$

Thus, it holds that

$$\hat{L}_h(s,\pi_h^*(s)) = \sum_{l\in[L]} \mathbb{1}\{N_{h,l}(s,\pi_h^*(s)) \geq 1\} = L_h(s,\pi_h^*(s)).$$

As a result, it holds that

$$\Gamma_h(s,\pi_h^*(s)) \leq c\sqrt{\sum_{l\in\hat{\mathcal{L}}_h(s,\pi_h^*(s))} \frac{H^2\log(SAH/\delta)}{\left(\hat{L}_h(s,\pi_h^*(s))\right)^2 N_{h,l}(s,\pi_h^*(s))}} + c\sqrt{\frac{H^2\log(SAH/\delta)}{\hat{L}_h(s,\pi_h^*(s))}}$$

$$\leq c\sqrt{\sum_{l\in\mathcal{L}_h(s,\pi^*_h(s))}\frac{H^2\log(SAH/\delta)}{(L_h(s,\pi^*_h(s)))^2Kd_h^{\rho_l,\mathcal{M}_l}(s,\pi^*_h(s))}}+c\sqrt{\frac{H^2\log(SAH/\delta)}{L_h(s,\pi^*_h(s))}}.$$

We can then obtain that

$$\begin{aligned}&\sum_{h\in[H]}\sum_{s\in\mathcal{S}}d_h^{\pi^*,\mathcal{M}}(s)\Gamma_h(s,\pi^*_h(s))\\&\leq c\sum_{h\in[H]}\sum_{s\in\mathcal{S}}d_h^{\pi^*,\mathcal{M}}(s)\sqrt{\frac{C^\dagger H^2\log(SAH/\delta)}{L^\dagger K\min\left\{d_h^{\pi^*,\mathcal{M}}(s),\frac{1}{S}\right\}}}+c\sum_{h\in[H]}\sum_{s\in\mathcal{S}}d_h^{\pi^*,\mathcal{M}}(s)\sqrt{\frac{H^2\log(SAH/\delta)}{L^\dagger}}\\&\leq c\sum_{h\in[H]}\sqrt{\sum_{s\in\mathcal{S}}d_h^{\pi^*,\mathcal{M}}(s)\frac{C^\dagger H^2\log(SAH/\delta)}{L^\dagger K\min\left\{d_h^{\pi^*,\mathcal{M}}(s),\frac{1}{S}\right\}}}\sqrt{\sum_{s\in\mathcal{S}}d_h^{\pi^*,\mathcal{M}}(s)}+c\sqrt{\frac{H^4\log(SAH/\delta)}{L^\dagger}}\\&\leq cH\sqrt{\frac{C^\dagger SH^2\log(SAH/\delta)}{L^\dagger K}}+c\sqrt{\frac{H^4\log(SAH/\delta)}{L^\dagger}}.\end{aligned}$$

Putting these results together, it can then be established that

$$\begin{aligned}V_1^{\pi^*,\mathcal{M}}(\xi)-V_1^{\hat{\pi},\mathcal{M}}(\xi)\leq V_1^{\pi^*,\mathcal{M}}(\xi)-\hat{V}_1(\xi)&\leq 2\sum_{h\in[H]}\sum_{s\in\mathcal{S}}d_h^{\pi^*,\mathcal{M}}(s)\Gamma_h(s,\pi^*_h(s))\\&\leq cH\sqrt{\frac{C^\dagger SH^2\log(SAH/\delta)}{L^\dagger K}}+cH\sqrt{\frac{2H^2\log(SAH/\delta)}{L^\dagger}},\end{aligned}$$

which concludes the proof.

D. Markov Game

D.1. Problem Formulation

The following task–source relationship is considered for MG. It shares the same content as Assumption 2.1 while we note that the overall action here consists of two individual actions from the max-player and min-player, i.e., $a=(a^1,a^2)$.

Assumption D.1 (Task–source Relationship)**.** Data source MGs $\{\mathcal{G}_l=\{H,\mathcal{S},\mathcal{A},\mathbb{P}_l,r_l\}:l\in[L]\}$ are generated from an unknown set of distributions $g=\{g_h:h\in[H]\}$ such that for each $(l,h)\in[L]\times[H]$, the reward and transition $\{r_{h,l},\mathbb{P}_{h,l}\}$ are independently sampled from the distribution $g_h(\cdot)$ whose expectation is $\{r_h,\mathbb{P}_h\}$ of the target MDP $\mathcal{G}:=\{H,\mathcal{S},\mathcal{A},\mathbb{P},r\}$.

D.2. Algorithm Details of HetPEVI-Game

The complete description of HetPEVI-Game can be found in Algorithm 2.

D.3. Core Lemmas

Following the same steps in the proof of Lemma C.2, the following lemma can be established.

Lemma D.2. *For all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$, and any function $V:\mathcal{S}\to[0,H]$ independent of $\hat{\mathbb{P}}_h$, with probability at least $1-\delta$, it holds that*

$$\left|\left(\hat{\mathbb{B}}_hV\right)(s,a)-\left(\mathbb{B}_hV\right)(s,a)\right|\leq\Gamma_h^g(s,a):=c\sqrt{\sum_{l\in\mathcal{L}_h(s,a)}\frac{H^2\log(SAH/\delta)}{\left(\hat{L}_h(s,a)\right)^2N_{h,l}(s,a)}}+c\sqrt{\frac{H^2\log(SAH/\delta)}{\hat{L}_h(s,a)}}.$$

We especially note that the action a in the above lemma and the following proofs stand for an action pair (a^1,a^2).

Algorithm 2 HetPEVI-Game

```
Input: Dataset $\mathcal{D} = \{D_l : l \in [L]\}$
Obtain $\mathcal{D}'_l \leftarrow$ subsampling$(\mathcal{D}_l), \forall l \in [L]$
for $\forall (s, a, h, s') \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{S}$ do
    $\forall l \in [L], N_{h,l}(s,a) \leftarrow$ number of visitations on $(s, a, h)$ in $\mathcal{D}'_l$
    $\forall l \in [L], N_{h,l}(s,a,s') \leftarrow$ number of visitations on $(s, a, h, s')$ in $\mathcal{D}'_l$
    $\hat{\mathcal{L}}_h(s,a) \leftarrow \{l \in [L] : N_{h,l}(s,a) > 0\}$
    $\forall l \in \hat{\mathcal{L}}_h(s,a), \hat{r}_{h,l}(s,a) \leftarrow r_{h,l}(s,a), \hat{\mathbb{P}}_{h,l}(s'|s,a) \leftarrow N_{h,l}(s,a,s')/N_{h,l}(s,a)$
    $\hat{r}_h(s,a) \leftarrow \sum_{l \in \hat{\mathcal{L}}_h(s,a)} \hat{r}_{h,l}(s,a)/(\hat{L}_h(s,a) \vee 1), \hat{\mathbb{P}}_h(s'|s,a) \leftarrow \sum_{l \in \hat{\mathcal{L}}_h(s,a)} \hat{\mathbb{P}}_{h,l}(s'|s,a)/(\hat{L}_h(s,a) \vee 1)$
end for
Initialize $\hat{V}_{H+1}(s) \leftarrow 0, \forall s \in \mathcal{S}$
for $h = H, H-1, \cdots, 1$ do
    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ do
        $\Gamma^g_h(s,a) \leftarrow \min\left\{c\sqrt{\sum_{l \in \mathcal{L}_h(s,a)} \frac{H^2 \log(SAH/\delta)}{(\hat{L}_h(s,a))^2 N_{h,l}(s,a)}} + c\sqrt{\frac{H^2 \log(SAH/\delta)}{\hat{L}_h(s,a)}}, H\right\}$
        $\hat{Q}_h(s,a) \leftarrow \max\left\{(\hat{\mathbb{B}}_h \hat{V}_{h+1})(s,a) - \Gamma^g_h(s,a)), 0\right\}$
    end for
    for $s \in \mathcal{S}$ do
        $(\hat{\mu}_h(\cdot|s), \hat{\nu}_h(\cdot|s)) \leftarrow$ NE$(\hat{Q}_h(s,\cdot))$
        $\hat{V}_h(s) \leftarrow \mathbb{E}_{a \sim \hat{\mu}_h(\cdot|s) \times \hat{\nu}_h(\cdot|s)}[\hat{Q}_h(s,a)]$
    end for
end for
Output: policy $\hat{\pi} = \{\hat{\pi}_h(s) : (s,h) \in \mathcal{S} \times [H]\}$
```

D.4. Main Proofs

Proof of Theorem 7.1. In the following, we establish Theorem 7.1. The proof framework is inspired by Yan et al. (2022) but is uniquely adapted to handle randomly perturbed data sources.

Step 1: establishing the pessimism property. Armed with Lemma D.2, with probability at least $1 - \delta$, the following relation holds

$$\hat{Q}_h(s,a) \leq Q_h^{\hat{\mu} \times \text{br}(\hat{\mu}), \mathcal{G}}(s,a), \qquad \hat{V}_h(s,a) \leq V_h^{\hat{\mu} \times \text{br}(\hat{\mu}), \mathcal{G}}(s,a), \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]. \tag{4}$$

Towards this, it is first observed that

$$\hat{Q}_{H+1}(s,a) = Q_{H+1}^{\hat{\mu} \times \text{br}(\hat{\mu}), \mathcal{G}}(s,a) = 0, \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Then, suppose that $\hat{Q}_{h+1}(s,a) \leq Q_{h+1}^{\hat{\mu} \times \text{br}(\hat{\mu}), \mathcal{G}}(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ at some step $h \in [H]$, we can observe that

$$0 \leq \hat{V}_{h+1}(s) = \hat{Q}_{h+1}(s, \hat{\mu} \times \hat{\nu}) \leq \hat{Q}_{h+1}(s, \hat{\mu} \times \text{br}(\hat{\mu})) \leq Q_{h+1}^{\hat{\mu} \times \text{br}(\hat{\mu}), \mathcal{G}}(s, \hat{\mu} \times \text{br}(\hat{\mu})) = V_{h+1}^{\hat{\mu} \times \text{br}(\hat{\mu}), \mathcal{G}}(s) \leq H, \qquad \forall s \in \mathcal{S}.$$

If $\hat{Q}_h(s,a) = 0$, the claim naturally holds. If not, we can obtain that

$$\begin{aligned}
\hat{Q}_h(s,a) &= \left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \Gamma_h(s,a) \\
&\leq \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a) + \left|\left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a)\right| - \Gamma^g_h(s,a) \\
&\overset{(i)}{\leq} \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a) \overset{(ii)}{\leq} \left(\mathbb{B}_h V_{h+1}^{\hat{\mu} \times \text{br}(\hat{\mu}), \mathcal{G}}\right)(s,a) = Q_{h+1}^{\hat{\mu} \times \text{br}(\hat{\mu}), \mathcal{G}}(s,a).
\end{aligned}$$

The above inequality (i) is from Lemma D.2 and leverages the fact that $\hat{V}_{h+1}(\cdot)$ is independent of $\hat{\mathbb{P}}_h$ and takes value in $[0, H]$. Inequality (ii) is from the obtained fact that $\hat{V}_{h+1}(s) \leq V_{h+1}^{\hat{\mu} \times \text{br}(\hat{\mu}), \mathcal{G}}(s)$. The desired claim Eqn. (4) can be verified by induction. □

Step 2: bounding the performance difference. From Eqn. (4), we can observe that

$$\begin{aligned}0 &\leq V_h^{\mu^*\times\nu^*,\mathcal{G}}(s) - V_h^{\hat{\mu}\times\mathrm{br}(\hat{\mu}),\mathcal{G}}(s) \leq V_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s) - \hat{V}_h(s)\\ &\leq Q_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s,\mu^*\times\bar{\nu}) - \hat{Q}_h(s,\mu^*\times\hat{\nu}) \leq Q_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s,\mu^*\times\bar{\nu}) - \hat{Q}_h(s,\mu^*\times\bar{\nu}),\end{aligned}$$

where

$$\bar{\nu} = \{\bar{\nu}_h : h\in[H]\}, \qquad \bar{\nu}_h(s) = \operatorname*{arg\,min}_{a^2\in\mathcal{A}^2} \mathbb{E}_{a^1\sim\mu_h^*(\cdot|s)}\hat{Q}_h(s,(a^1,a^2)).$$

We particularly note that $\bar{\nu}$ is a deterministic policy.

With

$$\begin{aligned}Q_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s,\mu^*\times\bar{\nu}) &= \mathbb{E}_{a\sim\mu_h^*(\cdot|s)\times\bar{\nu}_h(s)}\left[\left(\mathbb{B}_h V_{h+1}^{\mu^*\times\bar{\nu},\mathcal{G}}\right)(s,a)\right]\\ \hat{Q}_h(s,\mu^*\times\bar{\nu}) &= \mathbb{E}_{a\sim\mu_h^*(\cdot|s)\times\bar{\nu}_h(s)}\left[\max\left\{\left(\hat{\mathbb{B}}_h\hat{V}_{h+1}\right)(s,a) - \Gamma_h^g(s,a), 0\right\}\right],\end{aligned}$$

we can further obtain that

$$\begin{aligned}V_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s) - \hat{V}_h(s) &\leq \mathbb{E}_{a\sim\mu_h^*(\cdot|s)\times\bar{\nu}_h(s)}\left[\left(\mathbb{B}_h V_{h+1}^{\mu^*\times\bar{\nu},\mathcal{G}}\right)(s,a) - \left(\hat{\mathbb{B}}_h\hat{V}_{h+1}\right)(s,a) + \Gamma_h^g(s,a)\right]\\ &= \mathbb{E}_{a\sim\mu_h^*(\cdot|s)\times\bar{\nu}_h(s)}\left[\left(\mathbb{B}_h V_{h+1}^{\mu^*\times\bar{\nu},\mathcal{G}}\right)(s,a) - \left(\mathbb{B}_h\hat{V}_{h+1}\right)(s,a)\right]\\ &\quad + \mathbb{E}_{a\sim\mu_h^*(\cdot|s)\times\bar{\nu}_h(s)}\left[\left(\mathbb{B}_h\hat{V}_{h+1}\right)(s,a) - \left(\hat{\mathbb{B}}_h\hat{V}_{h+1}\right)(s,a) + \Gamma_h^g(s,a)\right]\\ &\overset{(i)}{\leq} \mathbb{E}_{a\sim\mu_h^*(\cdot|s)\times\bar{\nu}_h(s)}\left[\left(\mathbb{B}_h V_{h+1}^{\mu^*\times\bar{\nu},\mathcal{G}}\right)(s,a) - \left(\mathbb{B}_h\hat{V}_{h+1}\right)(s,a) + 2\Gamma_h^g(s,a)\right]\\ &= \mathbb{E}_{a\sim\mu_h^*(\cdot|s)\times\bar{\nu}_h(s)}\left[\left(\mathbb{P}_h V_{h+1}^{\mu^*\times\bar{\nu},\mathcal{G}}\right)(s,a) - \left(\mathbb{P}_h\hat{V}_{h+1}\right)(s,a) + 2\Gamma_h^g(s,a)\right]\end{aligned}$$

where inequality (i) holds with probability at least $1-\delta$ according to Lemma D.2. If applying the above argument iteratively, we can further obtain that

$$\sum_{s\in\mathcal{S}} d_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s)\left(V_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s) - \hat{V}_h(s)\right) \leq 2\sum_{h'=h}^{H}\sum_{s\in\mathcal{S}} d_{h'}^{\mu^*\times\bar{\nu},\mathcal{G}}(s)\Gamma_{h'}^g(s,\mu^*\times\bar{\nu}),$$

which indicates that

$$V_1^{\mu^*\times\bar{\nu},\mathcal{G}}(\xi) - \hat{V}_1(\xi) \leq 2\sum_{h\in[H]}\sum_{s\in\mathcal{S}} d_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s)\Gamma_h(s,\mu^*\times\bar{\nu}).$$

Step 3: completing the proof with concentrability. Let us consider $(s,a^1,h)\in\mathcal{S}\times\mathcal{A}^1\times[H]$ such that $d_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s,(a^1,\bar{\nu}_h(s)))>0$. We can then obtain that for all $l\in\mathcal{L}_h(s,(a^1,\bar{\nu}_h(s)))$,

$$\begin{aligned}N_{h,l}(s,(a^1,\bar{\nu}_h(s))) &\overset{(i)}{\geq} \frac{Kd_h^{\rho_l,\mathcal{G}_l}(s,(a^1,\bar{\nu}_h(s)))}{8} - 5\sqrt{Kd_h^{\rho_l,\mathcal{G}_l}(s,(a^1,\bar{\nu}_h(s)))\log\left(\frac{KHL}{\delta}\right)}\\ &\overset{(ii)}{\geq} \frac{Kd_h^{\rho_l,\mathcal{G}_l}(s,(a^1,\bar{\nu}_h(s)))}{16} \overset{(iii)}{\geq} 1.\end{aligned}$$

where inequality (i) is from Lemma C.1; and inequalities (ii) and (iii) are from the condition that

$$K \geq \frac{c\log\left(\frac{KH}{\delta}\right)}{d_g^{\min}} \geq \frac{c\log\left(\frac{KH}{\delta}\right)}{d_h^{\rho_l,\mathcal{G}_l}(s,(a^1,\bar{\nu}_h(s)))}.$$

Thus, it holds that

$$\hat{L}_h(s,(a^1,\bar{\nu}_h(s))) = \sum_{l\in[L]}\mathbb{1}\{N_{h,l}(s,(a^1,\bar{\nu}_h(s)))\geq 1\} = L_h(s,(a^1,\bar{\nu}_h(s))).$$

As a result, it holds that

$$\begin{aligned}\Gamma_h(s,(a^1,\bar{\nu}_h(s))) &\leq c\sqrt{\sum_{l\in\hat{\mathcal{L}}_h(s,(a^1,\bar{\nu}_h(s)))}\frac{H^2\log(SAH/\delta)}{\left(\hat{L}_h(s,(a^1,\bar{\nu}_h(s)))\right)^2N_{h,l}(s,(a^1,\bar{\nu}_h(s)))}}+c\sqrt{\frac{H^2\log(SAH/\delta)}{\hat{L}_h(s,(a^1,\bar{\nu}_h(s)))}}\\ &\leq c\sqrt{\sum_{l\in\mathcal{L}_h(s,(a^1,\bar{\nu}_h(s)))}\frac{H^2\log(SAH/\delta)}{(L_h(s,(a^1,\bar{\nu}_h(s))))^2Kd_h^{\rho_l,\mathcal{G}_l}(s,a)}}+c\sqrt{\frac{H^2\log(SAH/\delta)}{L_h(s,(a^1,\bar{\nu}_h(s)))}}.\end{aligned}$$

We can then obtain that

$$\begin{aligned}&\sum_{h\in[H]}\sum_{(s,a^1)\in\mathcal{S}\times\mathcal{A}^1}d_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s,(a^1,\bar{\nu}_h(s)))\Gamma_h(s,(a^1,\bar{\nu}_h(s)))\\ &\leq c\sum_{h\in[H]}\sum_{(s,a^1)\in\mathcal{S}\times\mathcal{A}^1}d_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s,(a^1,\bar{\nu}_h(s)))\sqrt{\frac{C_g^\dagger H^2\log(SAH/\delta)}{L_g^\dagger K\min\left\{d_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s,(a^1,\bar{\nu}_h(s))),\frac{1}{SA_1}\right\}}}\\ &+c\sum_{h\in[H]}\sum_{(s,a^1)\in\mathcal{S}\times\mathcal{A}^1}d_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s,(a^1,\bar{\nu}_h(s)))\sqrt{\frac{H^2\log(SAH/\delta)}{L_g^\dagger}}\\ &\leq c\sum_{h\in[H]}\sqrt{\sum_{(s,a^1)\in\mathcal{S}\times\mathcal{A}^1}\frac{d_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s,(a^1,\bar{\nu}_h(s)))C_g^\dagger H^2\log(SAH/\delta)}{L_g^\dagger K\min\left\{d_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s,(a^1,\bar{\nu}_h(s))),\frac{1}{SA_1}\right\}}}\sqrt{\sum_{(s,a^1)\in\mathcal{S}\times\mathcal{A}^1}d_h^{\pi^*,\mathcal{M}}((a^1,\bar{\nu}_h(s)))}\\ &+c\sqrt{\frac{H^4\log(SAH/\delta)}{L_g^\dagger}}\\ &\leq cH\sqrt{\frac{C_g^\dagger SA_1H^2\log(SAH/\delta)}{L_g^\dagger K}}+c\sqrt{\frac{H^4\log(SAH/\delta)}{L_g^\dagger}}.\end{aligned}$$

Putting these results together, it can then be established that

$$\begin{aligned}V_1^{\mu^*\times\nu^*,\mathcal{G}}(\xi)-V_1^{\hat{\mu}\times\mathrm{br}(\hat{\mu}),\mathcal{G}}(\xi)\leq V_1^{\mu^*\times\bar{\nu},\mathcal{G}}(\xi)-\hat{V}_1(\xi)&\leq 2\sum_{h\in[H]}\sum_{s\in\mathcal{S}}d_h^{\mu^*\times\bar{\nu},\mathcal{G}}(s)\Gamma_h(s,\mu^*\times\bar{\nu})\\ &\leq cH\sqrt{\frac{C_g^\dagger SA_1H^2\log(SAH/\delta)}{L_g^\dagger K}}+c\sqrt{\frac{H^4\log(SAH/\delta)}{L_g^\dagger}},\end{aligned}$$

which concludes the proof.

E. Robust RL

E.1. Problem Formulation

The following task–source relationship is considered for offline robust RL with perturbed data sources. It shares a similar content as Assumption 2.1 while an additional mild constraint is added to have the transition probabilities of data source MDPs bounded in a regime around that of the nominal MDP. This constraint simplifies later analysis while it is left for future works to investigate its necessity.

Assumption E.1 (Task–source Relationship, Robust MDP)**.** Data source MDPs $\{\mathcal{M}_l=\{H,\mathcal{S},\mathcal{A},\mathbb{P}_l,r_l\}:l\in[L]\}$ are generated from an unknown set of distributions $g=\{g_h:h\in[H]\}$ such that for each $(l,h)\in[L]\times[H]$, the reward and transition $\{r_{h,l},\mathbb{P}_{h,l}\}$ are independently sampled from the distribution $g_h(\cdot)$ whose expectation is $\{r_h,\mathbb{P}^c_h\}$ of the nominal MDP $\mathcal{M}^c=\{H,\mathcal{S},\mathcal{A},\mathbb{P}^c,r\}$. In addition, $\{r'_h,\mathbb{P}'_h\}\sim g_h(\cdot)$ satisfies that $\mathbb{P}'_h(s'|s,a)\in[T_l\cdot\mathbb{P}_h(s'|s,a),T_u\cdot\mathbb{P}_h(s'|s,a)]$ for constants $T_u<1,T_l>1$ at each $(s,a,h,s')\in\mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{S}$.

E.2. Algorithm Details of HetPEVI-Robust

The complete description of HetPEVI-Game can be found in Algorithm 3.

Algorithm 3 HetPEVI-Robust

```
**Input:** Dataset $\mathcal{D} = \{D_l : l \in [L]\}$
Obtain $\mathcal{D}'_l \leftarrow$ subsampling$(\mathcal{D}_l), \forall l \in [L]$
**for** $\forall (s, a, h, s') \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{S}$ **do**
    $\forall l \in [L], N_{h,l}(s, a) \leftarrow$ number of visitations on $(s, a, h)$ in $\mathcal{D}'_l$
    $\forall l \in [L], N_{h,l}(s, a, s') \leftarrow$ number of visitations on $(s, a, h, s')$ in $\mathcal{D}'_l$
    $\hat{\mathcal{L}}_h(s, a) \leftarrow \{l \in [L] : N_{h,l}(s, a) > 0\}$
    $\forall l \in \hat{\mathcal{L}}_h(s, a), \hat{r}_{h,l}(s, a) \leftarrow r_{h,l}(s, a), \hat{\mathbb{P}}_{h,l}(s'|s, a) \leftarrow N_{h,l}(s, a, s')/N_{h,l}(s, a)$
    $\hat{r}_h(s, a) \leftarrow \sum_{l \in \hat{\mathcal{L}}_h(s,a)} \hat{r}_{h,l}(s, a)/(\hat{L}_h(s, a) \vee 1), \hat{\mathbb{P}}_h(s'|s, a) \leftarrow \sum_{l \in \hat{\mathcal{L}}_h(s,a)} \hat{\mathbb{P}}_{h,l}(s'|s, a)/(\hat{L}_h(s, a) \vee 1)$
**end for**
Initialize $\hat{V}_{H+1}(s) \leftarrow 0, \forall s \in \mathcal{S}$
**for** $h = H, H - 1, \cdots, 1$ **do**
    **for** $s \in \mathcal{S}$ **do**
        $\Gamma^\sigma_h(s, a) \leftarrow \min\left\{\frac{c}{\sigma\hat{\mathbb{P}}^{\min}_h(s,a)}\sqrt{\sum_{l \in \hat{\mathcal{L}}_h(s,a)} \frac{H^2 \log(SAH/\delta)}{(\hat{L}_h(s,a))^2 N_{h,l}(s,a)}} + \frac{c}{\sigma\hat{\mathbb{P}}^{\min}_h(s,a)}\sqrt{\frac{H^2 \log(SAH/\delta)}{\hat{L}_h(s,a)}} + c\sqrt{\frac{\log(SAH/\delta)}{\hat{L}_h(s,a)}}, H\right\}$
        $\hat{Q}_h(s, a) \leftarrow \max\left\{\hat{r}_h(s, a) + \sup_{\lambda \geq 0}\left\{-\lambda \log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\hat{V}_{h+1}/\lambda\right)\right](s, a)\right) - \lambda\sigma\right\} - \Gamma^\sigma_h(s, a), 0\right\}$
    **end for**
    **for** $s \in \mathcal{S}$ **do**
        $\hat{\pi}_h(s) \leftarrow \arg\max_{a \in \mathcal{A}}(s, a)$
        $\hat{V}_h(s) \leftarrow \hat{Q}_h(s, \hat{\pi}_h(s))$
    **end for**
**end for**
**Output:** policy $\hat{\pi} = \{\hat{\pi}_h(s) : (s, h) \in \mathcal{S} \times [H]\}$
```

E.3. Core Lemmas

First, we introduce the following notations:

$$\begin{aligned}
\mathcal{C} &:= \left\{(s, a, h) : \exists l \in [L] \text{ s.t. } d^{\rho_l, \mathcal{M}_l}_h(s, a) > 0\right\};\\
\mathbb{P}^{\min}_h(s, a) &:= \min\left\{\mathbb{P}^c_h(s'|s, a) : s' \text{ s.t. } \mathbb{P}^c_h(s'|s, a) > 0\right\};\\
\mathbb{P}^{\min}_{h,l}(s, a) &:= \min\left\{\mathbb{P}_{h,l}(s'|s, a) : s' \text{ s.t. } \mathbb{P}_{h,l}(s'|s, a) > 0\right\};\\
\mathbb{P}^{\min}_\sigma &:= \min\left\{\mathbb{P}^c_h(s'|s, a) : (s, a, h, s') \text{ s.t. } \exists l \in [L], d^{\rho_l, \mathcal{M}_l}_h(s, a) > 0, \mathbb{P}^c_h(s'|s, a) > 0\right\}\\
&= \min\left\{\mathbb{P}^c_h(s'|s, a) : (s, a, h, s') \text{ s.t. } (s, a, h) \in \mathcal{C}, \mathbb{P}^c_h(s'|s, a) > 0\right\};\\
\mathbb{P}^{\min}_* &:= \min\left\{\mathbb{P}^c_h(s'|s, a) : (s, a, h, s') \text{ s.t. } d^{\pi^*, \mathcal{M}^c}_h(s, a) > 0, \mathbb{P}^c_h(s'|s, a) > 0\right\}\\
d^{\pi^*, \mathcal{R}}_h(s) &:= \left\{d^{\pi^*, \mathcal{M}^\sigma}_h(s) : \mathcal{M}^\sigma \in \mathcal{R}\right\};\\
d^{\min}_\sigma &:= \min\left\{d^{\rho_l, \mathcal{M}_l}_h(s, a) : (s, a, h, l) \text{ s.t. } d^{\rho_l, \mathcal{M}_l}_h(s, a) > 0\right\}.
\end{aligned}$$

A core lemma is then presented in the following.

Lemma E.2. *For all $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, and any function $V : \mathcal{S} \to [0, H]$ independent of $\hat{\mathbb{P}}_h$, with probability at least $1 - \delta$, it holds that*

$$\left|\hat{r}_h(s, a) + \inf_{\hat{\mathbb{P}}^\sigma_h(\cdot|s,a) \in \mathcal{U}^\sigma(\hat{\mathbb{P}}_h(\cdot|s,a))}\left(\hat{\mathbb{P}}^\sigma_h V\right)(s, a) - r_h(s, a) - \inf_{\mathbb{P}^\sigma_h(\cdot|s,a) \in \mathcal{U}^\sigma(\mathbb{P}^c_h(\cdot|s,a))}\left(\mathbb{P}^\sigma_h V\right)(s, a)\right| \leq \Gamma^\sigma_h(s, a). \tag{5}$$

Moreover, for all $(s, a, h) \in \mathcal{C}$, with probability at least $1 - \delta$, it holds that

$$\mathbb{P}^{\min}_h(s, a) \geq \frac{1}{T_u}\frac{\hat{\mathbb{P}}^{\min}_h(s, a)}{e^2} \geq \frac{T_l}{T_u}\frac{\mathbb{P}^{\min}_h(s, a)}{8e^2 \log(KHLSA/\delta)}. \tag{6}$$

Proof of Lemma E.2. We first prove the following fact that

$$\forall (s,a,h)\in\mathcal{C}, l\in\mathcal{L}_h(s,a): \qquad N_{h,l}(s,a) \geq \frac{cT_l^2\log(KHLSA/\delta)}{16(\mathbb{P}^{\min}_{h,l}(s,a))^2} \geq -\frac{\log(2KHLSA/\delta)}{\log(1-\mathbb{P}^{\min}_{h,l}(s,a))} \tag{7}$$

In particular, with

$$K \geq \frac{c\log(KHLSA/\delta)}{d_\sigma^{\min}(\mathbb{P}_\sigma^{\min})^2},$$

it holds that

$$Kd_h^{\rho_l,\mathcal{M}_l}(s,a) \geq \frac{cd_h^{\rho_l,\mathcal{M}_l}(s,a)\log(KHLSA/\delta)}{d_\sigma^{\min}(\mathbb{P}_\sigma^{\min})^2} \geq \frac{c\log(KHLSA/\delta)}{(\mathbb{P}_\sigma^{\min})^2} \geq \frac{c\log(KHLSA/\delta)}{(\mathbb{P}_h^{\min}(s,a))^2} \geq \frac{cT_l^2\log(KHLSA/\delta)}{(\mathbb{P}_{h,l}^{\min}(s,a))^2}.$$

Lemma C.1 then indicates that with probability at least $1-8\delta$,

$$N_{h,l}(s,a) \geq \frac{Kd_h^{\rho_l,\mathcal{M}_l}(s,a)}{8} - 5\sqrt{Kd_h^{\rho_l,\mathcal{M}_l}(s,a)\log\left(\frac{KHL}{\delta}\right)} \geq \frac{Kd_h^{\rho_l,\mathcal{M}_l}(s,a)}{16} \geq \frac{cT_l^2\log(KHLSA/\delta)}{16(\mathbb{P}_{h,l}^{\min}(s,a))^2}.$$

Furthermore, with $x\leq -\log(1-x)$ for all $x\in[0,1]$ and a suitable c, it holds that

$$\frac{cT_l^2\log(KHLSA/\delta)}{16(\mathbb{P}_{h,l}^{\min}(s,a))^2} \geq \frac{cT_l^2\log(KHLSA/\delta)}{16\mathbb{P}_{h,l}^{\min}(s,a)} \geq -\frac{\log(2KHLSA/\delta)}{\log(1-\mathbb{P}_{h,l}^{\min}(s,a))},$$

which completes the proof of Eqn. (7).

Then, with Lemma E.3, we can obtain that

$$\mathbb{P}_{h,l}(s'|s,a) \geq \frac{\hat{\mathbb{P}}_{h,l}(s'|s,a)}{e^2} \geq \frac{\mathbb{P}_{h,l}(s'|s,a)}{8e^2\log(KHLSA/\delta)},$$

This result further indicates that

$$\begin{aligned}
\mathbb{P}_h^{\min}(s,a) = \mathbb{P}_h^c(s_1|s,a) &\geq \frac{1}{T_u}\sum_{l\in\hat{\mathcal{L}}_h(s,a)}\frac{\mathbb{P}_{h,l}(s_1|s,a)}{\hat{L}_h(s,a)} \geq \frac{1}{T_u}\sum_{l\in\hat{\mathcal{L}}_h(s,a)}\frac{\hat{\mathbb{P}}_{h,l}(s_1|s,a)}{e^2\hat{L}_h(s,a)}\\
&\geq \frac{1}{T_u}\sum_{l\in\hat{\mathcal{L}}_h(s,a)}\frac{\hat{\mathbb{P}}_{h,l}(s_2|s,a)}{e^2\hat{L}_h(s,a)} = \frac{1}{T_u}\frac{\hat{\mathbb{P}}_h^{\min}(s,a)}{e^2} \geq \frac{1}{T_u}\sum_{l\in\hat{\mathcal{L}}_h(s,a)}\frac{\mathbb{P}_{h,l}(s_2|s,a)}{8e^2\hat{L}_h(s,a)\log(KHLSA/\delta)}\\
&\geq \frac{T_l}{T_u}\sum_{l\in\hat{\mathcal{L}}_h(s,a)}\frac{\mathbb{P}_h^c(s_2|s,a)}{8e^2\hat{L}_h(s,a)\log(KHLSA/\delta)} \geq \frac{T_l}{T_u}\frac{\mathbb{P}^c_{\min,h}(s,a)}{8e^2\log(KHLSA/\delta)},
\end{aligned}$$

where $\mathbb{P}_h^c(s_1|s,a) = \mathbb{P}_h^{\min}(s,a)$ and $\hat{\mathbb{P}}_h(s_2|s,a) = \hat{\mathbb{P}}_h^{\min}(s,a)$. Eqn. (6) is thus proved.

Then, we prove the first part in this lemma, i.e., Eqn (5). It can be first observed that

$$\begin{aligned}
\left|\hat{\mathbb{P}}_h(s'|s,a) - \mathbb{P}_h(s'|s,a)\right| &= \left|\sum_{l\in\hat{\mathcal{L}}_h(s,a)}\frac{1}{\hat{L}_h(s,a)}\hat{\mathbb{P}}_{h,l}(s'|s,a) - \mathbb{P}_h(s'|s,a)\right|\\
&\leq \left|\sum_{l\in\hat{\mathcal{L}}_h(s,a)}\frac{1}{\hat{L}_h(s,a)}\hat{\mathbb{P}}_{h,l}(s'|s,a) - \sum_{l\in\hat{\mathcal{L}}_h(s,a)}\frac{1}{\hat{L}_h(s,a)}\mathbb{P}_{h,l}(s'|s,a)\right|\\
&\quad + \left|\sum_{l\in\hat{\mathcal{L}}_h(s,a)}\frac{1}{\hat{L}_h(s,a)}\mathbb{P}_{h,l}(s'|s,a) - \mathbb{P}_h(s'|s,a)\right|
\end{aligned}$$

$$\leq \sqrt{\sum_{l \in \hat{\mathcal{L}}_h(s,a)} \frac{2\log(S^2AH/\delta)}{(\hat{L}_h(s,a))^2 N_{h,l}(s,a)}} + \sqrt{\frac{2\log(S^2AH/\delta)}{\hat{L}_h(s,a)}}. \tag{8}$$

where the last step holds with probability at least $1 - 2\delta/(S^2AH)$ according to the Hoeffidng inequality. This inequality thus holds with probability at least $1 - 2\delta$ for all $(s,a,h,s') \in \mathcal{S} \times \mathcal{A} \times [H] \times \mathcal{S}$. It is further indicated that for a function $V : \mathcal{S} \to [0,H]$

$$\begin{aligned}
&\frac{\left|\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda}\right)\right](s,a) - \left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right|}{\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)} \\
&\leq \max_{s' \in \mathrm{supp}(\mathbb{P}_h^c(\cdot|s,a))} \frac{\left|\hat{\mathbb{P}}_h(s'|s,a) - \mathbb{P}_h^c(s'|s,a)\right|}{\mathbb{P}_h^c(s'|s,a)} \\
&\leq \frac{1}{\mathbb{P}_h^{\min}(s,a)} \left(\sqrt{\sum_{l \in \hat{\mathcal{L}}_h(s,a)} \frac{2\log(S^2AH/\delta)}{(\hat{L}_h(s,a))^2 N_{h,l}(s,a)}} + \sqrt{\frac{2\log(S^2AH/\delta)}{\hat{L}_h(s,a)}}\right) \\
&\leq \frac{1}{2}
\end{aligned}$$

where the last inequality holds due to that with a suitable c,

$$\begin{aligned}
&N_{h,l}(s,a) \geq \frac{cT_l^2 \log(KHLSA/\delta)}{16(\mathbb{P}_{h,l}^{\min}(s,a))^2} \geq \frac{8\log(S^2AH/\delta)}{(\mathbb{P}_{h,l}^{\min}(s,a))^2} \geq 1, \qquad \forall l \in \mathcal{L}_h(s,a) \qquad \Rightarrow \mathcal{L}_h(s,a) = \hat{\mathcal{L}}_h(s,a); \\
&|\hat{\mathcal{L}}_h(s,a)| = |\mathcal{L}_h(s,a)| \geq \frac{8\log(S^2AH/\delta)}{(\mathbb{P}_h^{\min}(s,a))^2}.
\end{aligned}$$

From Lemma E.4, it can then be observed that

$$\begin{aligned}
&\left|\inf_{\mathbb{P}_h^\sigma(\cdot|s,a) \in \mathcal{U}^\sigma(\hat{\mathbb{P}}_h(\cdot|s,a))} (\mathbb{P}_h^\sigma V)(s,a) - \inf_{\mathbb{P}_h^\sigma(\cdot|s,a) \in \mathcal{U}^\sigma(\mathbb{P}_h^c(\cdot|s,a))} (\mathbb{P}_h^\sigma V)(s,a)\right| \\
&= \left|\sup_{\lambda>0}\left\{-\lambda \log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right) - \lambda\sigma\right\} - \sup_{\lambda>0}\left\{-\lambda \log\left(\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right) - \lambda\sigma\right\}\right|.
\end{aligned}$$

Denote

$$\begin{aligned}
\hat{\lambda}_h(s,a) &= \arg\max_{\lambda \geq 0}\left\{-\lambda \log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right) - \lambda\sigma\right\} \\
\lambda_h(s,a) &= \arg\max_{\lambda \geq 0}\left\{-\lambda \log\left(\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right) - \lambda\sigma\right\},
\end{aligned}$$

with Lemma E.5, we can further obtain that

$$\hat{\lambda}_h(s,a) \in \left[0, \frac{H}{\sigma}\right], \qquad \lambda_h(s,a) \in \left[0, \frac{H}{\sigma}\right].$$

In the following, we consider several different cases.

Case (I): $\hat{\lambda}_h(s,a) > 0$ and $\lambda_h(s,a) > 0$. In this case, it follows that

$$\begin{aligned}
&\left|\sup_{\lambda>0}\left\{-\lambda \log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right) - \lambda\sigma\right\} - \sup_{\lambda>0}\left\{-\lambda \log\left(\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right) - \lambda\sigma\right\}\right| \\
&\leq \max\Bigg\{ - \hat{\lambda}_h(s,a) \log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\hat{\lambda}_h(s,a)}\right)\right](s,a)\right) + \hat{\lambda}_h(s,a) \log\left(\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\hat{\lambda}_h(s,a)}\right)\right](s,a)\right),
\end{aligned}$$

$$
\begin{aligned}
&- \lambda_h(s,a)\log\left(\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda_h(s,a)}\right)\right](s,a)\right) + \lambda_h(s,a)\log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda_h(s,a)}\right)\right](s,a)\right)\Bigg\} \\
&\le \max_{\lambda\in\{\hat{\lambda}_h(s,a),\lambda_h(s,a)\}} \lambda\left|\log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right) - \log\left(\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right)\right| \\
&\le \max_{\lambda\in\{\hat{\lambda}_h(s,a),\lambda_h(s,a)\}} \lambda\left|\log\left(1+\frac{\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda}\right)\right](s,a) - \left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)}{\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)}\right)\right| \\
&\le \max_{\lambda\in\{\hat{\lambda}_h(s,a),\lambda_h(s,a)\}} 2\lambda\cdot\frac{\left|\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda}\right)\right](s,a) - \left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right|}{\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)} \\
&\le \frac{2H}{\sigma}\cdot\frac{1}{\mathbb{P}_h^{\min}(s,a)}\left(\sqrt{\sum_{l\in\hat{\mathcal{L}}_h(s,a)}\frac{2\log(S^2AH/\delta)}{(\hat{L}_h(s,a))^2N_{h,l}(s,a)}}+\sqrt{\frac{2\log(S^2AH/\delta)}{\hat{L}_h(s,a)}}\right) \\
&\le \frac{2H}{\sigma}\cdot\frac{T_ue^2}{\hat{\mathbb{P}}_h^{\min}(s,a)}\left(\sqrt{\sum_{l\in\mathcal{L}_h(s,a)}\frac{2\log(S^2AH/\delta)}{(\hat{L}_h(s,a))^2N_{h,l}(s,a)}}+\sqrt{\frac{2\log(S^2AH/\delta)}{\hat{L}_h(s,a)}}\right).
\end{aligned}
$$

Case (II): $\hat{\lambda}_h(s,a)>0$ and $\lambda_h(s,a)=0$; $\hat{\lambda}_h(s,a)=0$ and $\lambda_h(s,a)>0$. We consider the sub-case that $\hat{\lambda}_h(s,a)>0$ and $\lambda_h(s,a)=0$ while the other sub-case can be proved similarly. In particular, with Lemma E.5 and Lemma E.6, we can obtain that

$$
\begin{aligned}
\sup_{\lambda\ge 0}\left\{-\lambda\log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right)-\lambda\sigma\right\} &\ge \lim_{\lambda\to 0}\left\{-\lambda\log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right)-\lambda\sigma\right\} \\
&= \operatorname*{essinf}_{s'\sim\hat{\mathbb{P}}_h(\cdot|s,a)} V(s') = \inf \operatorname*{essinf}_{s'\sim\hat{\mathbb{P}}_{h,l}(\cdot|s,a)} V(s') \\
&= \inf \operatorname*{essinf}_{s'\sim\mathbb{P}_{h,l}(\cdot|s,a)} V(s') = \operatorname*{essinf}_{s'\sim\mathbb{P}_h^c(\cdot|s,a)} V(s') \\
&= \sup_{\lambda\ge 0}\left\{-\lambda\log\left(\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right)-\lambda\sigma\right\}.
\end{aligned}
$$

As a result, it holds that

$$
\begin{aligned}
&\left|\sup_{\lambda>0}\left\{-\lambda\log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right)-\lambda\sigma\right\}-\sup_{\lambda>0}\left\{-\lambda\log\left(\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right)-\lambda\sigma\right\}\right| \\
&= \sup_{\lambda>0}\left\{-\lambda\log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right)-\lambda\sigma\right\}-\sup_{\lambda>0}\left\{-\lambda\log\left(\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right)-\lambda\sigma\right\} \\
&\le -\hat{\lambda}_h(s,a)\log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\hat{\lambda}_h(s,a)}\right)\right](s,a)\right)-\hat{\lambda}_h(s,a)\sigma \\
&+\hat{\lambda}_h(s,a)\log\left(\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\hat{\lambda}_h(s,a)}\right)\right](s,a)\right)+\hat{\lambda}_h(s,a)\sigma \\
&= \hat{\lambda}_h(s,a)\left[\log\left(\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\hat{\lambda}_h(s,a)}\right)\right](s,a)\right)-\hat{\lambda}_h(s,a)\log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\hat{\lambda}_h(s,a)}\right)\right](s,a)\right)\right],
\end{aligned}
$$

which can be bounded via the same steps in Case (I).

Case (III): $\hat{\lambda}_h(s,a)=\lambda_h(s,a)=0$. With Lemma E.5 and Lemma E.6, it holds that

$$
\left|\sup_{\lambda>0}\left\{-\lambda\log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right)-\lambda\sigma\right\}-\sup_{\lambda>0}\left\{-\lambda\log\left(\left[\mathbb{P}_h^c \exp\left(-\frac{V}{\lambda}\right)\right](s,a)\right)-\lambda\sigma\right\}\right|
$$

$$= \left| \operatorname*{essinf}_{s' \sim \hat{\mathbb{P}}_h(\cdot|s,a)} V(s') - \operatorname*{essinf}_{s' \sim \mathbb{P}^c_h(\cdot|s,a)} V(s') \right| = \left| \inf_{l\in[L]} \operatorname*{essinf}_{s' \sim \hat{\mathbb{P}}_{h,l}(\cdot|s,a)} V(s') - \operatorname*{essinf}_{s' \sim \mathbb{P}^c_h(\cdot|s,a)} V(s') \right|$$
$$= \left| \inf_{l\in[L]} \operatorname*{essinf}_{s' \sim \mathbb{P}_{h,l}(\cdot|s,a)} V(s') - \operatorname*{essinf}_{s' \sim \mathbb{P}^c_h(\cdot|s,a)} V(s') \right| = 0.$$

Together with the fact that with probability at least $1-\delta$,

$$\hat{r}_h(s,a) - r_h(s,a) \le \sqrt{\frac{\log(1/(HSA\delta))}{\hat{L}_h(s,a)}}, \qquad \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H].$$

Eqn. (5) is shown to be valid. □

E.4. Main Proofs

Proof of Theorem 8.1. In the following, we establish Theorem 8.1. The proof framework is inspired by Shi & Chi (2022) but is uniquely adapted to handle randomly perturbed data sources.

Step 1: establishing the pessimism property. With Lemma E.2, we first show that the following inequalities hold with probability at least $1-\delta$:

$$\hat{Q}_h(s,a) \le Q_h^{\hat{\pi},\mathcal{R}}(s,a), \qquad \hat{V}_h(s,a) \le V_h^{\hat{\pi},\mathcal{R}}(s,a), \qquad \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]. \tag{9}$$

Towards this, it is first observed that

$$\hat{Q}_{H+1}(s,a) = Q_{H+1}^{\hat{\pi},\mathcal{R}}(s,a) = 0, \qquad \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

Then, suppose that $\hat{Q}_{h+1}(s,a) \le Q_{h+1}^{\hat{\pi},\mathcal{R}}(s,a)$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ at some step $h\in[H]$, we can observe that by the update rule in HetPEVI-Game, it holds that

$$0 \le \hat{V}_{h+1}(s) = \max_{a\in\mathcal{A}} \hat{Q}_{h+1}(s,a) \le \max_{a\in\mathcal{A}} Q_{h+1}^{\hat{\pi},\mathcal{R}}(s,a) = V_{h+1}^{\hat{\pi},\mathcal{R}}(s) \le H, \qquad \forall s\in\mathcal{S},$$

If $\hat{Q}_h(s,a)=0$, the claim naturally holds. If not, we can obtain that

$$\begin{aligned}
\hat{Q}_h(s,a) &= \hat{r}_h(s,a) + \sup_{\lambda>0}\left\{ -\lambda \log\left(\left[\hat{\mathbb{P}}_h \exp\left(-\frac{\hat{V}_{h+1}}{\lambda}\right)\right](s,a)\right) - \lambda\sigma \right\} - \Gamma_h^\sigma(s,a) \\
&= \hat{r}_h(s,a) + \inf_{\hat{\mathbb{P}}_h^\sigma(\cdot|s,a)\in\mathcal{U}^\sigma(\hat{\mathbb{P}}_h(\cdot|s,a))} \left(\hat{\mathbb{P}}_h^\sigma \hat{V}_{h+1}\right)(s,a) - \Gamma_h^\sigma(s,a) \\
&\le r_h(s,a) + \inf_{\mathbb{P}_h^\sigma(\cdot|s,a)\in\mathcal{U}^\sigma(\mathbb{P}^c_h(\cdot|s,a))} \left(\mathbb{P}_h^\sigma \hat{V}_{h+1}\right)(s,a) - \Gamma_h^\sigma(s,a) \\
&\quad + \left| r_h(s,a) + \inf_{\mathbb{P}_h^\sigma(\cdot|s,a)\in\mathcal{U}^\sigma(\mathbb{P}^c_h(\cdot|s,a))} \left(\mathbb{P}_h^\sigma \hat{V}_{h+1}\right)(s,a) - \hat{r}_h(s,a) - \inf_{\hat{\mathbb{P}}_h^\sigma(\cdot|s,a)\in\mathcal{U}^\sigma(\hat{\mathbb{P}}_h(\cdot|s,a))} \left(\hat{\mathbb{P}}_h^\sigma \hat{V}_{h+1}\right)(s,a) \right| \\
&\overset{(i)}{\le} r_h(s,a) + \inf_{\mathbb{P}_h^\sigma(\cdot|s,a)\in\mathcal{U}^\sigma(\mathbb{P}^c_h(\cdot|s,a))} \left(\mathbb{P}_h^\sigma \hat{V}_{h+1}\right)(s,a) \\
&\overset{(ii)}{\le} r_h(s,a) + \inf_{\mathbb{P}_h^\sigma(\cdot|s,a)\in\mathcal{U}^\sigma(\mathbb{P}^c_h(\cdot|s,a))} \left(\mathbb{P}_h^\sigma V_{h+1}^{\hat{\pi},\mathcal{R}}\right)(s,a) = Q_h^{\hat{\pi},\mathcal{R}}(s,a).
\end{aligned}$$

The above inequality (i) is from Lemma E.2 and leverages the fact that $\hat{V}_{h+1}(\cdot)$ is independent of $\hat{\mathbb{P}}_h$ and takes value in $[0,H]$. Inequality (ii) is from the obtained fact that $\hat{V}_{h+1}(s) \le V_{h+1}^{\hat{\pi},\mathcal{R}}(s)$. The desired claim Eqn. (9) can be verified by induction.

Step 2: bounding the performance difference. From Eqn. (9), we can observe that

$$0 \le V_h^{\pi^*,\mathcal{R}}(s) - V_h^{\hat{\pi},\mathcal{R}}(s) \le V_h^{\pi^*,\mathcal{R}}(s) - \hat{V}_h(s) \le Q_h^{\pi^*,\mathcal{R}}(s,\pi_h^*(s)) - \hat{Q}_h(s,\pi_h^*(s)).$$

With

$$Q_h^{\pi^*,\mathcal{R}}(s,\pi_h^*(s)) = r_h(s,\pi_h^*(s)) + \inf_{\mathbb{P}_h^\sigma(\cdot|s,a)\in\mathcal{U}^\sigma(\mathbb{P}_h^c(\cdot|s,a)} \left(\mathbb{P}_h^\sigma V_{h+1}^{\pi^*,\mathcal{R}}\right)(s,a)$$

$$\hat{Q}_h(s,\pi_h^*(s)) = \max\left\{\hat{r}_h(s,\pi_h^*(s)) + \inf_{\hat{\mathbb{P}}_h^\sigma(\cdot|s,a)\in\mathcal{U}^\sigma(\hat{\mathbb{P}}_h(\cdot|s,a)} \left(\hat{\mathbb{P}}_h^\sigma \hat{V}_{h+1}\right)(s,a) - \Gamma_h^\sigma(s,\pi_h^*(s)), 0\right\},$$

we can further obtain that

$$\begin{aligned}
V_h^{\pi^*,\mathcal{R}}(s) - \hat{V}_h(s) &\le r_h(s,\pi_h^*(s)) + \inf_{\mathbb{P}_h^\sigma(\cdot|s,\pi_h^*(s))\in\mathcal{U}^\sigma(\mathbb{P}_h^c(\cdot|s,\pi_h^*(s))} \left(\mathbb{P}_h^\sigma V_{h+1}^{\pi^*,\mathcal{R}}\right)(s,\pi_h^*(s)) \\
&\quad - \hat{r}_h(s,\pi_h^*(s)) - \inf_{\hat{\mathbb{P}}_h^\sigma(\cdot|s,\pi_h^*(s))\in\mathcal{U}^\sigma(\hat{\mathbb{P}}_h(\cdot|s,\pi_h^*(s))} \left(\hat{\mathbb{P}}_h^\sigma \hat{V}_{h+1}\right)(s,\pi_h^*(s)) + \Gamma_h^\sigma(s,\pi_h^*(s)) \\
&= r_h(s,\pi_h^*(s)) + \inf_{\mathbb{P}_h^\sigma(\cdot|s,\pi_h^*(s))\in\mathcal{U}^\sigma(\mathbb{P}_h^c(\cdot|s,\pi_h^*(s))} \left(\mathbb{P}_h^\sigma V_{h+1}^{\pi^*,\mathcal{R}}\right)(s,\pi_h^*(s)) \\
&\quad - r_h(s,\pi_h^*(s)) - \inf_{\mathbb{P}_h^\sigma(\cdot|s,\pi_h^*(s))\in\mathcal{U}^\sigma(\mathbb{P}_h^c(\cdot|s,\pi_h^*(s))} \left(\mathbb{P}_h^\sigma \hat{V}_{h+1}\right)(s,\pi_h^*(s)) \\
&\quad + r_h(s,\pi_h^*(s)) + \inf_{\mathbb{P}_h^\sigma(\cdot|s,\pi_h^*(s))\in\mathcal{U}^\sigma(\mathbb{P}_h^c(\cdot|s,\pi_h^*(s))} \left(\mathbb{P}_h^\sigma \hat{V}_{h+1}\right)(s,\pi_h^*(s)) \\
&\quad - \hat{r}_h(s,\pi_h^*(s)) - \inf_{\hat{\mathbb{P}}_h^\sigma(\cdot|s,\pi_h^*(s))\in\mathcal{U}^\sigma(\hat{\mathbb{P}}_h(\cdot|s,\pi_h^*(s))} \left(\hat{\mathbb{P}}_h^\sigma \hat{V}_{h+1}\right)(s,\pi_h^*(s)) + \Gamma_h^\sigma(s,\pi_h^*(s)) \\
&\overset{(i)}{\le} \inf_{\mathbb{P}_h^\sigma(\cdot|s,\pi_h^*(s))\in\mathcal{U}^\sigma(\mathbb{P}_h^c(\cdot|s,\pi_h^*(s))} \left(\mathbb{P}_h^\sigma V_{h+1}^{\pi^*,\mathcal{R}}\right)(s,\pi_h^*(s)) \\
&\quad - \inf_{\mathbb{P}_h^\sigma(\cdot|s,\pi_h^*(s))\in\mathcal{U}^\sigma(\mathbb{P}_h^c(\cdot|s,\pi_h^*(s))} \left(\mathbb{P}_h^\sigma \hat{V}_{h+1}\right)(s,\pi_h^*(s)) + 2\Gamma_h^\sigma(s,\pi_h^*(s)) \\
&\overset{(ii)}{\le} \left(\mathbb{P}_h^{\inf} V_{h+1}^{\pi^*,\mathcal{R}}\right)(s,\pi_h^*(s)) - \left(\mathbb{P}_h^{\inf} \hat{V}_{h+1}\right)(s,\pi_h^*(s)) + 2\Gamma_h^\sigma(s,\pi_h^*(s))
\end{aligned}$$

where inequality (i) holds with probability at least $1-\delta$ according to Lemma E.2 and inequality (ii) holds with the notation

$$\mathbb{P}_h^{\inf}(\cdot|s,a) = \operatorname*{arg\,min}_{\mathbb{P}_h^\sigma\in\mathcal{U}^\sigma(\mathbb{P}_h^c(\cdot|s,a))} \left(\mathbb{P}_h^\sigma \hat{V}_{h+1}\right)(s,a).$$

If applying the above argument iteratively, we can further obtain that

$$\sum_{s\in\mathcal{S}} d_h^{\inf}(s)\left(V_h^{\pi*,\mathcal{R}}(s) - \hat{V}_h(s)\right) \le 2\sum_{h'=h}^{H}\sum_{s\in\mathcal{S}} d_{h'}^{\inf}(s)\Gamma_{h'}(s,\pi_h^*(s)),$$

where $d_h^{\inf}(s)$ denotes the visitation probability induced by optimal policy π^* and $\mathbb{P}^{\inf} = \{\mathbb{P}_h^{\inf} : h\in[H]\}$. Finally, it can be obtained that

$$V_1^{\pi*,\mathcal{R}}(\xi) - \hat{V}_1(\xi) \le 2\sum_{h\in[H]}\sum_{s\in\mathcal{S}} d_h^{\inf}(s)\Gamma_h(s,\pi_h^*(s)),$$

and $d_h^{\inf}(s)\in d_h^{\pi^*,\mathcal{R}}(s)$.

Step 3: completing the proof with concentrability. Let us consider $(s,h)\in\mathcal{S}\times[H]$ such that $d_h^{\inf}(s)>0$. We can then obtain that for all $l\in\mathcal{L}_h(s,\pi_h^*(s))$,

$$N_{h,l}(s,\pi_h^*(s)) \overset{(i)}{\ge} \frac{Kd_h^{\rho_l,\mathcal{M}_l}(s,\pi_h^*(s))}{8} - 5\sqrt{Kd_h^{\rho_l,\mathcal{M}_l}(s,\pi_h^*(s))\log\left(\frac{KHL}{\delta}\right)} \overset{(ii)}{\ge} \frac{Kd_h^{\rho_l,\mathcal{M}_l}(s,\pi_h^*(s))}{16} \overset{(iii)}{\ge} 1.$$

where inequality (i) is from Lemma C.1; and inequalities (ii) and (iii) are from the condition that

$$K \ge \frac{c\log(KHLSA/\delta)}{d_\sigma^{\min}(\mathbb{P}_\sigma^{\min})^2} \ge \frac{c\log(KHL/\delta)}{d_h^{\rho_l,\mathcal{M}_l}(s,a)}.$$

Thus, it holds that

$$\hat{L}_h(s,\pi_h^*(s)) = \sum_{l\in[L]} \mathbb{1}\{N_{h,l}(s,\pi_h^*(s)) \geq 1\} = L_h(s,\pi_h^*(s)).$$

As a result, it holds that

$$\begin{aligned}
\Gamma_h(s,\pi_h^*(s)) &\leq \frac{c}{\sigma\hat{\mathbb{P}}_h^{\min}(s,\pi_h^*(s))}\sqrt{\sum_{l\in\hat{\mathcal{L}}_h(s,a)} \frac{H^2\log(SAH/\delta)}{(\hat{L}_h(s,\pi_h^*(s)))^2 N_{h,l}(s,\pi_h^*(s))}} \\
&\quad + \frac{c}{\sigma\hat{\mathbb{P}}_h^{\min}(s,\pi^*(s))}\sqrt{\frac{H^2\log(SAH/\delta)}{\hat{L}_h(s,\pi_h^*(s))}} + c\sqrt{\frac{\log(SAH/\delta)}{\hat{L}_h(s,\pi_h^*(s))}} \\
&\leq \frac{c}{\sigma\hat{\mathbb{P}}_h^{\min}(s,\pi_h^*(s))}\sqrt{\sum_{l\in\mathcal{L}_h(s,a)} \frac{H^2\log(SAH/\delta)}{(L_h(s,\pi_h^*(s)))^2 K d^{\rho_l,\mathcal{M}_l}(s,\pi_h^*(s))}} \\
&\quad + \frac{c}{\sigma\hat{\mathbb{P}}_h^{\min}(s,\pi^*(s))}\sqrt{\frac{H^2\log(SAH/\delta)}{L_h(s,\pi_h^*(s))}} + c\sqrt{\frac{\log(SAH/\delta)}{L_h(s,\pi_h^*(s))}}.
\end{aligned}$$

We can then obtain that

$$\begin{aligned}
&\sum_{h\in[H]}\sum_{s\in\mathcal{S}} d_h^{\inf}(s)\Gamma_h(s,\pi_h^*(s)) \\
&\leq c\sum_{h\in[H]}\sum_{s\in\mathcal{S}} d_h^{\inf}(s)\frac{1}{\sigma\cdot\hat{\mathbb{P}}_h^{\min}(s,\pi_h^*(s))}\sqrt{\frac{C_\sigma^\dagger H^2\log(SAH/\delta)}{L_\sigma^\dagger K\min\left\{d_h^{\inf}(s),\frac{1}{S}\right\}}} \\
&+ \sum_{h\in[H]}\sum_{s\in\mathcal{S}} d_h^{\inf}(s)\left(1+\frac{H}{\sigma\cdot\hat{\mathbb{P}}_h^{\min}(s,\pi_h^*(s))}\right)\sqrt{\frac{H^2\log(SAH/\delta)}{L_\sigma^\dagger}} \\
&\leq c\sum_{h\in[H]}\sum_{s\in\mathcal{S}} d_h^{\inf}(s)\frac{H\log(KHLSA/\delta)}{\sigma\cdot\mathbb{P}_h^{\min}(s,\pi_h^*(s))}\sqrt{\frac{C_\sigma^\dagger\log(SAH/\delta)}{L_\sigma^\dagger K\min\left\{d_h^{\inf}(s),\frac{1}{S}\right\}}} \\
&+ c\sum_{h\in[H]}\sum_{s\in\mathcal{S}} d_h^{\inf}(s)\left(1+\frac{H\log(KHLSA/\delta)}{\sigma\cdot\mathbb{P}_h^{\min}(s,\pi_h^*(s))}\right)\sqrt{\frac{\log(SAH/\delta)}{L_\sigma^\dagger}} \\
&\leq c\sum_{h\in[H]}\frac{H^2}{\sigma\cdot\mathbb{P}_*^{\min}}\sqrt{\sum_{s\in\mathcal{S}} d_h^{\inf}(s)\frac{C_\sigma^\dagger\log(SAH/\delta)}{L_\sigma^\dagger K\min\left\{d_h^{\inf}(s),\frac{1}{S}\right\}}}\sqrt{\sum_{s\in\mathcal{S}} d_h^{\inf}(s)} \\
&+ cH\left(1+\frac{H\log(KHLSA/\delta)}{\sigma\cdot\mathbb{P}_*^{\min}}\right)\sqrt{\frac{\log(SAH/\delta)}{L_\sigma^\dagger}} \\
&\leq c\frac{H^2}{\sigma\cdot\mathbb{P}_*^{\min}}\sqrt{\frac{C_\sigma^\dagger S\log(SAH/\delta)}{L_\sigma^\dagger K}} + cH\left(1+\frac{H\log(KHLSA/\delta)}{\sigma\cdot\mathbb{P}_*^{\min}}\right)\sqrt{\frac{\log(SAH/\delta)}{L_\sigma^\dagger}}.
\end{aligned}$$

Putting these results together, it can then be established that

$$\begin{aligned}
V_1^{\pi^*,\mathcal{R}}(\xi) - V_1^{\hat{\pi},\mathcal{R}}(\xi) \leq V_1^{\pi^*,\mathcal{R}}(\xi) - \hat{V}_1(\xi) &\leq 2\sum_{h\in[H]}\sum_{s\in\mathcal{S}} d_h^{\inf}(s)\Gamma_h(s,\pi_h^*(s)) \\
&\leq c\frac{H^2}{\sigma\cdot\mathbb{P}_*^{\min}}\sqrt{\frac{C_\sigma^\dagger S\log(SAH/\delta)}{L_\sigma^\dagger K}} + cH\left(1+\frac{H\log(KHLSA/\delta)}{\sigma\cdot\mathbb{P}_*^{\min}}\right)\sqrt{\frac{\log(SAH/\delta)}{L_\sigma^\dagger}},
\end{aligned}$$

which concludes the proof. □

E.5. Auxiliary Lemmas

Lemma E.3 (Lemma 8, Shi et al. (2022))**.** *Suppose $N \sim$ Binomial(n,p), where $n \geq 1$ and $p \in [0,1]$. For any $\delta \in [0,1]$, we have*

$$N \geq \frac{np}{8\log(1/\delta)}, \qquad \textit{if } np \geq 8\log(1/\delta); \qquad N \leq \begin{cases} e^2 np & \textit{if } np \geq \log(1/\delta), \\ 2e^2\log(1/\delta) & \textit{if } np \leq 2\log(1/\delta) \end{cases}$$

hold with probability at least $1-\delta$.

Lemma E.4 (Theorem 1, Hu & Hong (2013))**.** *Suppose $f(x)$ has a finite moment generating function in some neighborhood around $x=0$, then for any $\sigma > 0$ and a nominal distribution $\mathbb{P}^c$, we have*

$$\sup_{\mathbb{P}\in\mathcal{U}^\sigma(\mathbb{P}^c)} \mathbb{E}_{X\sim\mathbb{P}}[f(X)] = \inf_{\lambda=0}\left\{\lambda\log\left(\mathbb{E}_{X\sim\mathbb{P}^c}\left[\exp\left(\frac{f(X)}{\lambda}\right)\right]\right) + \lambda\sigma\right\}.$$

Lemma E.5 (Lemma 4, Zhou et al. (2021))**.** *Let $X \sim \mathbb{P}$ be a bounded random variable with $X \in [0,M]$. Let $\sigma > 0$ be any uncertainty level and the corresponding optimal dual variable be*

$$\lambda^* \in \arg\max_{\lambda>0} f(\lambda,\mathbb{P}), \qquad \textit{where } f(\lambda,\mathbb{P}) := -\lambda\log\left(\mathbb{E}_{X\sim\mathbb{P}}\left[\exp\left(-\frac{X}{\lambda}\right)\right]\right) - \lambda\sigma.$$

Then the optimal value λ^* obeys

$$\lambda^* \in \left[0, \frac{M}{\sigma}\right].$$

Moreover, when $\lambda^* = 0$, we have

$$\lim_{\lambda\to 0} f(\lambda,\mathbb{P}) = \operatorname{essinf} X.$$

Lemma E.6 (Zhou et al. (2021))**.** *Let $X \sim \mathbb{P}$ be a discrete bounded random variable with $X \in [0,M]$. Let $\mathbb{P}_n$ denote the empirical distribution constructed from n independent samples $X_1, X_2, \cdots, X_n$ and let $\hat{X} \sim \mathbb{P}_n$. Denote $\mathbb{P}_{\min} := \min\{\mathbb{P}_{X=x} : x \in supp(X)\}$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\min_{i\in[n]} X_i = \operatorname{essinf} \hat{X} = \operatorname{essinf} X,$$

as long as

$$n \geq -\frac{\log(2/\delta)}{\log(1-\mathbb{P}_{\min})}.$$

F. Experimental Setups

The target MDP is set to have $H = 20$ steps, $S = 2$ states (labelled as $\{1,2\}$), $A = 20$ actions (labelled as $\{1,2,\cdots,20\}$). The reward and transitions are specified as follows:

$$\forall h \in [H], r_h(s,a) = \begin{cases} 0.9 & \text{if } (s,a) = (1,1) \\ 0.1 & \text{otherwise} \end{cases}$$

$$\forall h \in [H], \mathbb{P}_h(1|s,a) = \begin{cases} 0.9 & \text{if } (s,a) = (1,1) \\ 0.5 & \text{otherwise.} \end{cases}$$

$$\forall h \in [H], \mathbb{P}_h(1|s,a) = \begin{cases} 0.1 & \text{if } (s,a) = (1,1) \\ 0.5 & \text{otherwise.} \end{cases}$$

The rewards of the data source are independently sampled from Bernoulli distributions, i.e., $r_{h,l}(s,a) \sim$ Bernoulli$(r_h(s,a))$, while the transitions are independently generated with standard Dirichlet distributions (Marchal & Arbel, 2017), i.e., $\mathbb{P}_{h,l}(\cdot|s,a) \sim$ Dirichlet$(\mathbb{P}_h(\cdot|s,a))$. The behavior policy is shared by all data sources, which at each $(s,h) \in \mathcal{S}\times[H]$, selects action 1 with probability 0.2 and otherwise randomly chooses from other actions. The results plotted in Fig. 2 are averaged from 100 independently repeated experiments.