OFFLINE RL OF THE UNDERLYING MDP FROM HETEROGENEOUS DATA SOURCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Most of the existing offline reinforcement learning (RL) studies assume the available dataset is sampled directly from the target environment. However, in some practical applications, the available data are often coming from several related but heterogeneous environments. A theoretical understanding of efficient learning from heterogeneous offline datasets remains lacking. In this work, we study the problem of learning a (hidden) underlying Markov decision process (MDP) based on heterogeneous offline datasets collected from multiple randomly perturbed data sources. A novel HetPEVI algorithm is proposed, which jointly considers two types of uncertainties: sample uncertainties from the finite number of data samples per data source, and source uncertainties due to a finite number of data sources. Building on HetPEVI, we further incorporate reference-advantage decompositions and Bernstein-type penalties to propose the HetPEVI-Adv algorithm. Theoretical analysis not only proves the effectiveness of both HetPEVI and HetPEVI-Adv but also demonstrates the advantage of the latter. More importantly, the results explicitly characterize the learning loss due to the finite heterogeneously realized environments compared with sampling directly from the underlying MDP. Finally, we extend the study to MDPs with linear function approximation and propose the HetPEVI-Lin algorithm that provides additional efficiency guarantees beyond the tabular case.

1 INTRODUCTION

Offline reinforcement learning (RL) (Levine et al., 2020), a.k.a. batch RL (Lange et al., 2012), has received growing interest in the 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 (Kidambi et al., 2020), recent studies have also established theoretical understandings of offline RL (Rashidinejad et al., 2021; Jin et al., 2021b; Duan et al., 2021; Uehara & Sun, 2021).

Despite these progresses, the majority of offline RL research focuses on learning via data collected exactly from the target task environment (Kumar et al., 2020). However, in practice, it is difficult to ensure that all such data are perfectly from one source environment. Instead, in many cases, it is more reasonable to assume that data are collected from different sources that are perturbed versions of the target model. For example, when training a chatbot (Jaques et al., 2020), the offline dialogue datasets typically consist of short conversations between different people, who naturally have varying language habits. The training objective is the common underlying language structure, e.g., basic grammar, which however cannot be reflected in any individual dialogue but must be holistically learned from the aggregation of them. More examples can be found in healthcare (Tang & Wiens, 2021), autonomous driving (Sallab et al., 2017) and others. While a few empirical investigations under the offline meta-RL framework have been reported (Dorfman et al., 2021; Lin et al., 2022; Mitchell et al., 2021), theoretical understandings of effectively and efficiently learning the underlying task using datasets from multiple heterogeneous sources are largely lacking.

Motivated by both practical and theoretical limitations, this work makes progress in the under-explored RL problem of *learning the common underlying structure from heterogeneous data sources*. In particular, we study the problem of learning an underlying Markov decision process (MDP) from offline datasets sampled from multiple heterogeneously realized source MDPs. Several provably efficient designs are proposed, targeting both tabular and linear MDPs. To the best of our

knowledge, this is the first work that proposes provably efficient offline RL algorithms to handle heterogeneous data sources, which can benefit relevant applications and further shed light on the theoretical understanding of offline meta-RL. The contributions are summarized as follows:

• We study a new offline RL problem where the datasets are sampled from multiple source MDPs, with possibly different reward and transition dynamics. Motivated by practical applications, the data sources are modeled as random perturbations of a hidden underlying MDP, which is the learning target. Compared with previous studies of offline RL from one single dataset sampled from the targeted MDP (Rashidinejad et al., 2021; Jin et al., 2021b), we face new challenges of jointly considering uncertainties caused by the finite number of data samples per source (referred to as the *sample uncertainties*) and by the finite number of data sources (referred to as the *source uncertainties*).

• A novel HetPEVI algorithm is proposed, which generalizes the idea of pessimistic value iteration (Jin et al., 2021b) and uses a two-part penalty term to address the sample and source uncertainties. Especially, for the sample uncertainty, a new Hoeffding-type penalty term is developed to aggregate uncertainties associated with heterogeneous data sources. Theoretical analysis proves the effectiveness of HetPEVI, which also demonstrates that even with finite randomly perturbed MDPs and finite data samples from each of them, it is feasible to efficiently learn the underlying MDP. Moreover, the result reveals that it is more important to involve more data sources than to collect more data from each source. This principle can help guide practical data collections.

• To further improve the performance, we propose HetPEVI-Adv by incorporating two new ingredients to account for the sample uncertainty: the reference-advantage decomposition technique and a new Bernstein-type penalty design. The latter relies on an enhanced empirical Bernstein inequality that is capable of jointly considering the variance of data from each source, which we believe has its individual merit. HetPEVI-Adv is theoretically proven to be superior to HetPEVI. More importantly, the analysis reveals that learning from heterogeneously realized environments does not increase the loss associated with finite data samples. Rather, the limited data sources incur an additional performance loss that cannot be reduced by increasing the amount of data samples. This result also confirms the aforementioned importance of the data source diversity.

• Lastly, we extend the study to linear function approximation where offline data are collected from linear MDPs with a shared feature mapping but heterogeneously realized system dynamics. The HetPEVI-Lin algorithm is developed to jointly consider the sample and source uncertainties while incorporating the linear structure, whose effectiveness is also verified by theoretical analysis.

Related Works. With the empirical success of offline RL (Levine et al., 2020), its theoretical understandings have been gradually established in recent years. In particular, the principle of "pessimism" is incorporated and proved efficient for offline RL (Jin et al., 2021b; Rashidinejad et al., 2021). Following this line, Xie et al. (2021b); Li et al. (2022); Shi et al. (2022) further fine-tune the designs for the tabular setting and Zanette et al. (2021); Min et al. (2021); Yin et al. (2022); Xiong et al. (2022) for linear MDPs (Jin et al., 2020). However, these theoretical advances are mainly focused on learning directly from the target data source, while practically, growing interests have been made to utilize data from heterogeneous sources, e.g., offline meta-RL (Mitchell et al., 2021; Dorfman et al., 2021; Lin et al., 2022; Li et al., 2020b). This work is thus motivated to provide a theoretical understanding of how to extract underlying information from the available data of multiple sources. A more detailed literature review regarding both online and offline RL with single or heterogeneous environments is provided in Appendix A.1.

2 Problem Formulation

Preliminaries of RL. We consider an RL problem characterized by an episodic MDP $\mathcal{M} := (H, \mathcal{S}, \mathcal{A}, \mathbb{P}, r)$. In this tuple, H is the length of each episode, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathbb{P}$ is the transition matrix so that $\mathbb{P}_h(s'|s,a)$ gives the probability of transiting to state s' if action a is taken for 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, in each episode, starting from an initial state s_1, at each step $h \in [H]$, the agent observes state $s_h \in \mathcal{S}$, picks action $a_h \in \mathcal{A}$, receives reward

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

$r_h(s_h, a_h)$, and then transits to a next state $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h)$. The episode ends after H steps. In the tabular RL setting, the state and action spaces are finite with $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$.

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 $\mathcal{A}$, whose value functions can be defined as $V_h^\pi(s) := \mathbb{E}_{\pi,\mathcal{M}}\left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})|s_h = s\right]$ and $Q_h^\pi(s,a) := \mathbb{E}_{\pi,\mathcal{M}}\left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})|s_h = s, a_h = a\right]$ for each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, 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 $\mathcal{M}$. The optimal policy π^* maximizes the value function, i.e., $\pi^* := \arg\max_\pi V_1^\pi(s)$. For convenience, we denote $V_h^*(s) := V_h^{\pi^*}(s)$ and $Q_h^*(s,a) := Q_h^{\pi^*}(s,a)$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, and use $\pi_h(s)$ to refer to the chosen action for state s at step h for a deterministic policy π.

2.1 Heterogeneous Data Sources

We consider that the agent has access to datasets from L sources, i.e., $\{\mathcal{D}_l, \forall l \in [L]\}$. Each dataset $\mathcal{D}_l := \{(s_{h,l}^k, a_{h,l}^k, r_{h,l}^k, s_{h+1,l}^k) : k \in [K], h \in [H]\}$ consists of K trajectories independently sampled by a (possibly different) behavior policy ρ_l on a MDP $\mathcal{M}_l = (H, \mathcal{S}, \mathcal{A}, \mathbb{P}_l, r_l)$. Note that different data source MDPs share the same episodic length, state space, and action space; however the transition and reward dynamics are not necessarily aligned, i.e., $\mathbb{P}_{h,l}(\cdot|s,a) \neq \mathbb{P}_{h,j}(\cdot|s,a)$, $r_{h,l}(s,a) \neq r_{h,j}(s,a)$ for $\mathcal{M}_l$ and $\mathcal{M}_j$, $l \neq j$.

2.2 The Underlying MDP

In practical applications, while being heterogeneous, the data source MDPs are often related and share some common hidden structures (e.g., the example of dialogue datasets in Sec. 1). In particular, each data source MDP in offline meta-RL is often assumed to be sampled from a common distribution (Mitchell et al., 2021), which motivates the following definition of the underlying MDP.

Definition 1 (Underlying MDP)**.** *An MDP $\mathcal{M} := (H, \mathcal{S}, \mathcal{A}, \mathbb{P}, r)$ is said to be an underlying MDP to MDPs $\{\mathcal{M}_l : l \in [L]\}$ if for each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, rewards $\{r_{h,l}(s,a) : l \in [L]\}$ are i.i.d. random variables with mean $r_h(s,a)$, and transition vectors $\{\mathbb{P}_{h,l}(\cdot|s,a) : l \in [L]\}$ are i.i.d. random vectors with mean $\mathbb{P}_h(\cdot|s,a)$.*

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

To facilitate the later discussions, the sample average MDP of the data source MDPs is denoted as $\bar{\mathcal{M}} := (H, \mathcal{S}, \mathcal{A}, \bar{\mathbb{P}}, \bar{r})$ where $\bar{\mathbb{P}}_h(\cdot|s,a) = \sum_{l \in [L]} \mathbb{P}_{h,l}(\cdot|s,a)/L$ and $\bar{r}_h(s,a) = \sum_{l \in [L]} r_{h,l}(s,a)/L$. Note that with finite data sources, the sample average MDP $\bar{\mathcal{M}}$ may not exactly match the true underlying MDP $\mathcal{M}$.

2.3 The Learning Target

As discussed in Sec. 1, while the agent is provided with datasets from individual source MDPs $\{\mathcal{M}_l : l \in [L]\}$, her goal is to learn the common underlying structure, especially the underlying MDP $\mathcal{M}$ in Definition 1. For an output policy $\hat{\pi}$, its sub-optimality gap is thus measured as:

$$\text{Gap}(\hat{\pi}; \mathcal{M}) := V_1^*(s_1) - V_1^{\hat{\pi}}(s_1),$$

where the value function and optimal policy π^* are both w.r.t. the underlying MDP $\mathcal{M}$. Then, the agent aims at outputting a policy $\hat{\pi}$ with a small gap, which is called ε-optimal if $\text{Gap}(\hat{\pi}; \mathcal{M}) \leq \varepsilon$.

Miscellaneous. Notations without subscripts l generally refer to the underlying MDP $\mathcal{M}$, while subscripts l are always added when discussing each individual source $\mathcal{M}_l$. For a clear presentation, the notation c is used throughout the paper with varing values to represent a poly-logarithmic term of order $O(\text{poly}(\log(LHSA/\delta)))$ for tabular RL and $O(\text{poly}(\log(LH/\delta)))$ for linear RL, where $\delta \in (0,1)$ is a confidence parameter. Additionally, for $x \in \mathbb{R}$, x^+ denotes $\max\{x, 0\}$.

For any function $f : \mathcal{S} \to \mathbb{R}$, the transition operator and Bellman operator of the underlying MDP $\mathcal{M}$ at each step $h \in [H]$ are defined, respectively, as $(\mathbb{P}_h f)(s,a) := \mathbb{E}[f(s')|s,a]$ and $(\mathbb{B}_h f)(s,a) :=$

$r_h(s,a) + (\mathbb{P}_h f)(s,a)$, where the expectation is w.r.t. the transition $s' \sim \mathbb{P}_h(\cdot|s,a)$. Moreover, $(\mathbb{V}_h f)(s,a)$ refers to the variance of $f(s')$ with $s' \sim \mathbb{P}_h(\cdot|s,a)$. Similar notations of $(\bar{\mathbb{P}}_h f)(s,a)$, $(\bar{\mathbb{B}}_h f)(s,a)$ and $(\bar{\mathbb{V}}_h f)(s,a)$ are defined w.r.t. the sample average MDP $\bar{\mathcal{M}}$. Expectation $\mathbb{E}_{\pi,\mathcal{M}}[\cdot]$ is over the trajectory induced by π on the underlying MDP $\mathcal{M}$ starting from s_1, and $d_h^\pi(s)$ (resp. $d_h^\pi(s,a)$) denotes the probability of visiting state s (resp. state-action pair (s,a)) at step h in the trajectory induced by policy π on the underlying MDP $\mathcal{M}$ starting from s_1. $\mathbb{E}_{\pi,\mathcal{M}_l}[\cdot]$, $d_{h,l}^\pi(s)$ and $d_{h,l}^\pi(s,a)$ are defined similarly but on $\mathcal{M}_l$.

Algorithm 1 HetPEVI

Input: Dataset $\mathcal{D} = \{D_l : l \in [L]\}$; $\hat{V}_{H+1}(s) = 0, \forall s \in \mathcal{S}$

1: Randomly split each dataset $\mathcal{D}_l$ into sub-datasets $\{\mathcal{D}_{h,l}\}_{h\in[H]}$ with $|\mathcal{D}_{h,l}| = K/H$

2: Estimate $\hat{r}_{h,l}(s,a)$ and $\hat{\mathbb{P}}_{h,l}(\cdot|s,a)$ via $\mathcal{D}_{h,l}$ as in Eqn. (1), $\forall (s,a,h,l) \in \mathcal{S}\times\mathcal{A}\times[H]\times[L]$

3: Obtain $\hat{r}_h(s,a)$ and $\hat{\mathbb{P}}_h(\cdot|s,a)$ as in Eqn. (2), $\forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$

4: **for** $h = H, H-1, \cdots, 1$ **do**

5: Perform updates for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ as in Eqn. (3) with the empirical Bellman operator $\hat{\mathbb{B}}_h$ in Eqn. (4) and the penalty Γ_h in Eqn. (6)

6: **end for**

Output: policy $\hat{\pi} = \{\hat{\pi}_h(s) : (s,h) \in \mathcal{S}\times[H]\}$, values $\{\hat{V}_h(s) : (s,h) \in \mathcal{S}\times[H]\}$

3 THE HETPEVI ALGORITHM

In this section, the HetPEVI algorithm is introduced for the tabular RL setting (presented in Alg. 1). HetPEVI follows the principle of pessimistic value iterations (PEVI) (Rashidinejad et al., 2021; Jin et al., 2021b), which is first briefly introduced, while the key challenges and our novel design in learning the underlying MDP from heterogeneous data sources are then illustrated.

At the beginning of HetPEVI, each dataset $\mathcal{D}_l$ from source $l \in [L]$ is randomly divided into H folds of sub-dataset $\mathcal{D}_{h,l}$ (i.e., each containing K/H trajectories) to ensure the statistical independence in the subsequent estimations. Then, empirical estimations of rewards and transitions are given for each $(s,a,h,s',l) \in \mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{S}\times[L]$ as

$$\hat{r}_{h,l}(s,a) = r_{h,l}(s,a)\mathbb{1}\{(s,a,h) \in \mathcal{D}_{h,l}\}, \qquad \hat{\mathbb{P}}_{h,l}(s'|s,a) = N_{h,l}(s,a,s')/(N_{h,l}(s,a) \vee 1), \quad (1)$$

where $N_{h,l}(s,a)$ and $N_{h,l}(s,a,s')$ are the amount of visitations on (s,a,h) and (s,a,h,s') in dataset $\mathcal{D}_{h,l}$, respectively. With empirical estimations of each individual data source, HetPEVI averages them for each $(s,a,h,s') \in \mathcal{S}\times\mathcal{A}\times[H]\times\mathcal{S}$ as

$$\hat{r}_h(s,a) = \sum\nolimits_{l\in[L]} \hat{r}_{h,l}(s,a)/L; \qquad \hat{\mathbb{P}}_h(s'|s,a) = \sum\nolimits_{l\in[L]} \hat{\mathbb{P}}_{h,l}(s'|s,a)/L. \quad (2)$$

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

$$\begin{aligned} \hat{Q}_h(s,a) &= \min\left\{\left(\hat{\mathbb{B}}_h\hat{V}_{h+1}\right)(s,a) - \Gamma_h(s,a), H-h+1\right\}^+, \\ \hat{V}_h(s) &= \max_{a\in\mathcal{A}} \hat{Q}_h(s,a), \qquad \hat{\pi}_h(s) = \arg\max_{a\in\mathcal{A}} \hat{Q}_h(s,a), \end{aligned} \quad (3)$$

with $\hat{V}_{H+1}(s) = 0, \forall s \in \mathcal{S}$, and the empirical Bellman operator $\hat{\mathbb{B}}_h$ is defined as

$$\left(\hat{\mathbb{B}}_h\hat{V}_{h+1}\right)(s,a) := \hat{r}_h(s,a) + \left(\hat{\mathbb{P}}_h\hat{V}_{h+1}\right)(s,a), \quad (4)$$

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 the above 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. Especially, the penalty term $\Gamma_h(s,a)$ should be carefully designed such that with a high probability, it holds that

$$\left|\left(\hat{\mathbb{B}}_h\hat{V}_{h+1}\right)(s,a) - \left(\mathbb{B}_h\hat{V}_{h+1}\right)(s,a)\right| \le \Gamma_h(s,a), \qquad \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]. \quad (5)$$

Technical Challenges. Previous offline RL studies (Jin et al., 2021b; Rashidinejad et al., 2021; Yin et al., 2022; Xie et al., 2021b) only deal with one single target data source and only one type of uncertainties (finite data samples) to ensure Eqn. (5). In this work, the agent needs to process multiple

heterogeneous datasets, while none of them individually characterizes the learning target (i.e., the underlying MDP). As a result, the agent faces two *coupled* challenges. First, the uncertainties due to the finite sample sizes still needs to be considered. This is referred to as the **sample uncertainties**, but the key difference is that now the agent needs to jointly consider and aggregate uncertainties associated with all data sources. Second, even with perfect knowledge of each data source, the underlying MDP may not be fully revealed. Thus, the agent also needs to consider the uncertainties from the limited number of data sources, which are referred to as the **source uncertainties**.

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

$$\Gamma_h(s,a) = \Gamma_h^{\alpha}(s,a) + \Gamma_h^{\beta}(s,a), \tag{6}$$

where $\Gamma_h^{\alpha}(s,a)$ aggregates the sample uncertainties from each data source while $\Gamma_h^{\beta}(s,a)$ accounts for the source uncertainties. Both of them are further elaborated on in the following.

Penalties to Aggregate Sample Uncertainties. For the first part of the penalty, i.e., $\Gamma_h^{\alpha}(s,a)$, since the agent faces data from multiple heterogeneous sources, the penalty term needs to jointly consider the individual uncertainties from all sources. In HetPEVI, the following penalty term is proposed, which originates from the Hoeffding inequality:

$$\Gamma_h^{\alpha}(s,a) = c\sqrt{\sum\nolimits_{l\in[L]} \frac{H^2}{L^2 N_{h,l}(s,a)}}. \tag{7}$$

We emphasize that this construction is feasible because of the random data splitting performed at the beginning of HetPEVI. In particular, the estimated value function $\hat{V}_{h+1}(\cdot)$ depends on datasets $\{\mathcal{D}_{h',l} : h' \geq h+1, l \in [L]\}$, while the estimated transition $\hat{\mathbb{P}}_h(\cdot|s,a)$ is based on datasets $\{\mathcal{D}_{h,l} : l \in [L]\}$. Thus, their estimations are independent and the Hoeffding inequality can be applied.

Penalties to Account for Source Uncertainties. The second part of the penalty $\Gamma_h^{\beta}(s,a)$ serves the important role of measuring the uncertainties due to the limited amount of data sources. With the observation that $(\mathbb{B}_{h,l}\hat{V}_{h+1})(s,a)$ is a bounded random variable with mean of $(\mathbb{B}_h\hat{V}_{h+1})(s,a)$, the following penalty term is proposed, which shrinks with more data sources, i.e, larger L:

$$\Gamma_h^{\beta}(s,a) = c\sqrt{\frac{H^2}{L}}. \tag{8}$$

Remark 1. Instead of treating each source individually and directly summing up their sample uncertainties (as $c\sum_{l\in[L]} \frac{1}{L}\sqrt{H^2/N_{h,l}(s,a)}$), the penalty term in Eqn. (7) is a joint measure of sample uncertainties from all sources. This would lead to a $1/\sqrt{L}$ speed-up in the performance guarantees.
Remark 2. With the two-part penalty term, with high probability, for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, it simultaneously holds that $|(\hat{\mathbb{B}}_h\hat{V}_{h+1})(s,a) - (\bar{\mathbb{B}}_h\hat{V}_{h+1})(s,a)| \leq \Gamma_h^{\alpha}(s,a)$ and $|(\bar{\mathbb{B}}_h\hat{V}_{h+1})(s,a) - (\mathbb{B}_h\hat{V}_{h+1})(s,a)| \leq \Gamma_h^{\beta}(s,a)$, which jointly ensure Eqn. (5).

3.1 Theoretical Analysis

Before establishing the performance guarantee of HetPEVI, a coverage assumption is first presented.

Assumption 1. *There exists a constant $C^* < \infty$ such that for each $l \in [L]$, the data collection policy ρ_l satisfies $d_h^{\pi^*}(s,a) \leq C^* d_{h,l}^{\rho_l}(s,a)$ for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.*

This assumption is intuitively reasonable as for each data source $\mathcal{M}_l$, the data collection policy ρ_l should be able to provide information for the state-action pairs that may show up with the optimal policy π^* on the target $\mathcal{M}$. The parameter C^* characterizes the distance between the data collection policy ρ_l and the targeted optimal policy π^*. Similar assumptions (on covering the optimal policy) are commonly adopted in offline RL studies, e.g., Rashidinejad et al. (2021); Xie et al. (2021b) for MDPs and Cui & Du (2022) for Markov Games. Also, when Assumption 1 does not hold w.r.t. covering the optimal policy, similar performance guarantees can still be achieved when comparing with the best policy covered by the datasets.

Theorem 1. *Under Assumption 1, w.p. at least $1-\delta$, the output policy $\hat{\pi}$ of HetPEVI satisfies*

$$\textit{Gap}(\hat{\pi};\mathcal{M}) = \tilde{O}\left(\sqrt{\frac{C^* H^5 S}{KL}} + \sqrt{\frac{H^4}{L}}\right). \tag{9}$$

The first term of the gap guarantee in Eqn. (9) comes from the finite sample size (i.e., the overall LK trajectories) while the second term is due to the availability of a limited number of data sources (i.e., the overall L data sources). This result illustrates that even with finite randomly perturbed MDPs and finite data samples from each of them, it is still feasible to efficiently learn the underlying MDP.

Additionally, Thm. 1 indicates that to obtain an ε-optimal policy, HetPEVI requires the overall number of samples $T = LK$ to be of order $\tilde{O}(C^* H^5 S/\varepsilon^2)$ and the available number of data sources L to be of order $\tilde{O}(H^4/\varepsilon^2)$. Note that the first requirement can be viewed as on the *sample complexity* while the second one is on the *source diversity*. Comparing these requirements, we see that it is more important to involve more data sources than to collect more data from each source. This is reasonable as involving more data sources provides additional population coverage (i.e., diversity) while also adding more data.

Remark 3. If the learning target is chosen to be the sample average MDP $\bar{\mathcal{M}}$ instead of the underlying MDP $\mathcal{M}$, the penalty term Γ_h^β for the source uncertainties in HetPEVI (and the later designs) is not necessary. Similar guarantees can be obtained with only the gap from sample uncertainties (e.g., the first term in Eqn. (9)). Consequently, only the sample complexity requirement is needed.

Algorithm 2 HetPEVI-Adv

Input: Dataset $\mathcal{D} = \{D_l : l \in [L]\}$; $\hat{V}_{H+1}(s) = 0, \forall s \in \mathcal{S}$

1: Randomly split each $\mathcal{D}_l$ into $\mathcal{D}_{\text{ref},l}$, $\mathcal{D}_{0,l}$ and $\mathcal{D}_{1,l} = \{\mathcal{D}_{h,1,l} : h \in [H]\}$ with $|\mathcal{D}_{\text{ref},l}| = K/3$, $|\mathcal{D}_{0,l}| = K/3$ and $|\mathcal{D}_{h,1,l}| = K/(3H)$

2: Obtain references $\{\hat{V}_h^{\text{ref}}(s), \forall (s,h) \in \mathcal{S} \times [H]\}$ from HetPEVI($\{\mathcal{D}_{\text{ref},l} : l \in [L]\}$)

3: Estimate $\hat{\mathbb{P}}_{h,0,l}(\cdot|s,a)$ via dataset $\mathcal{D}_{0,l}$, for all (s,a,h,l,s') in $\mathcal{S} \times \mathcal{A} \times [H] \times [L]$

4: Estimate $\hat{r}_{h,1,l}(s,a)$ and $\hat{\mathbb{P}}_{h,1,l}(\cdot|s,a)$ via dataset $\mathcal{D}_{h,1,l}$, for all (s,a,h,l) in $\mathcal{S} \times \mathcal{A} \times [H] \times [L]$

5: Set $\hat{\mathbb{P}}_{h,0}(\cdot|s,a) = \sum_{l\in[L]} \hat{\mathbb{P}}_{h,0,l}(\cdot|s,a)/L$, $\hat{r}_{h,1}(s,a) = \sum_{l\in[L]} \hat{r}_{h,1,l}(s,a)/L$ and $\hat{\mathbb{P}}_{h,1}(\cdot|s,a) = \sum_{l\in[L]} \hat{\mathbb{P}}_{h,1,l}(\cdot|s,a)/L$ for all (s,a,h) in $\mathcal{S} \times \mathcal{A} \times [H] \times [L]$

6: **for** $h = H, H-1, \cdots, 1$ **do**

7: Perform updates for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ as in Eqn. (3) with $\hat{\mathbb{B}}_h$ in Eqn. (10) and the penalty $\Gamma_h = \Gamma_{h,0}^\alpha + \Gamma_{h,1}^\alpha + \Gamma_h^\beta$ with $\Gamma_{h,0}^\alpha, \Gamma_{h,1}^\alpha$ in Eqns. (11), (12) and Γ_h^β in Eqn. (8)

8: **end for**

Output: policy $\hat{\pi} = \{\hat{\pi}_h(s) : (s,h) \in \mathcal{S} \times [H]\}$, values $\{\hat{V}_h(s) : (s,h) \in \mathcal{S} \times [H]\}$

4 THE HETPEVI-ADV ALGORITHM

We can further improve HetPEVI from the following two aspects. First, in HetPEVI, only one of the H folds of each dataset $\mathcal{D}_l$ is used for estimations at one step. While ensuring independence for estimations of different steps, such splitting is considerably wasteful. Second, it is widely recognized in RL studies that Hoeffding-type penalties (as in Eqn. (7)) are often overly conservative, as they ignore the variance information. In this section, these two drawbacks are tackled in the proposed HetPEVI-Adv algorithm (presented in Alg. 2) via the reference-advantage decomposition (Zhang et al., 2020) and novel Bernstein-type penalties (Azar et al., 2017), respectively.

Reference-Advantage Decomposition. Each dataset $\mathcal{D}_l$ is randomly divided into three parts: $\mathcal{D}_{\text{ref},l}$, $\mathcal{D}_{0,l}$ and $\mathcal{D}_{1,l}$ with $|\mathcal{D}_{\text{ref},l}| = |\mathcal{D}_{0,l}| = |\mathcal{D}_{1,l}| = K/3$. The dataset $\mathcal{D}_{1,l}$ is further split into H folds as $\{\mathcal{D}_{h,1,l} : h \in [H]\}$ with $|\mathcal{D}_{h,1,l}| = K/(3H)$. We denote $\mathcal{D}_{\text{ref}} := \{\mathcal{D}_{\text{ref},l} : l \in [L]\}$, $\mathcal{D}_0 := \{\mathcal{D}_{0,l} : l \in [L]\}$, $\mathcal{D}_1 := \{\mathcal{D}_{1,l} : l \in [L]\}$ and $\mathcal{D}_{h,1} := \{\mathcal{D}_{h,1,l} : l \in [L]\}$ for all $h \in [H]$.

The agent first applies HetPEVI with dataset $\mathcal{D}_{\text{ref}}$ and obtains the output values as the references, i.e., $\{\hat{V}_h^{\text{ref}}(s)\}$, which is guaranteed to be close to $\{V_h^*(s)\}$ as in Thm. 1. Then, the agent uses the *entire* dataset $\mathcal{D}_{0,l}$ to estimate the transition $\hat{\mathbb{P}}_{h,0,l}(s'|s,a)$ via empirical counts $N_{h,0,l}(s,a)$ and $N_{h,0,l}(s,a,s')$. Similarly, the agent performs another estimation of rewards $\hat{r}_{h,1,l}(s,a)$ and transitions $\hat{\mathbb{P}}_{h,1,l}(s'|s,a)$ with dataset $\mathcal{D}_{h,1,l}$ (which is split from $\mathcal{D}_{1,l}$) via empirical counts $N_{h,1,l}(s,a)$ and $N_{h,1,l}(s,a,s')$. Finally, the agent aggregates to have the overall estimation $\hat{\mathbb{P}}_{h,0}$, $\hat{r}_{h,1}$ and $\hat{\mathbb{P}}_{h,1}$.

Based on the estimations, PEVI in Eqn. (3) is performed with the following empirical Bellman operator $\hat{\mathbb{B}}_h$, where the second term is the reference and the third term is the advantage over it:

$$\big(\hat{\mathbb{B}}_h \hat{V}_{h+1}\big)(s,a) := \hat{r}_{h,1}(s,a) + \underbrace{\big(\hat{\mathbb{P}}_{h,0} V^{\text{ref}}_{h+1}\big)(s,a)}_{\text{reference}} + \underbrace{\big[\hat{\mathbb{P}}_{h,1}\big(\hat{V}_{h+1} - \hat{V}^{\text{ref}}_{h+1}\big)\big](s,a)}_{\text{advantage}}, \tag{10}$$

Bernstein-type Penalties for Sample Uncertainties. The penalty term $\Gamma^{\alpha}_h(s,a)$ in HetPEIV-Adv for sample uncertainties consists of two terms: $\Gamma^{\alpha}_{h,0}(s,a)$ and $\Gamma^{\alpha}_{h,1}(s,a)$, which consider the reference and advantage terms, respectively. In order to incorporate variance information, the empirical Bernstein inequality (Maurer & Pontil, 2009) is often adopted to design Bernstein-type penalties (Azar et al., 2017; Jin et al., 2018) as opposed to the Hoeffding-type penalties. However, its original form can only handle random variables from one source. To solve this problem, an enhanced version of the empirical Bernstein inequality is developed.

Lemma 1 (Enhanced Empirical Bernstein Inequality, Informal)**.** *Let $Z_l, Z_{1,l}, Z_{2,l}, \cdots, Z_{n_l,l}$ be i.i.d. random variables in $[0,1]$ for each $l \in [L]$. Then, with high probability, it holds that*

$$\left| \frac{1}{L}\sum\nolimits_{l\in[L]} \mathbb{E}[Z_l] - \frac{1}{L}\sum\nolimits_{l\in[L]} \frac{\sum_{i=1}^{n_l} Z_{i,l}}{n_l} \right| = \tilde{O}\left(\sqrt{\frac{1}{L^2}\sum\nolimits_{l\in[L]} \frac{\hat{\mathbb{V}}(Z_l)}{n_l}} + \sqrt{\sum\nolimits_{l\in[L]} \frac{1}{L^2 n_l^2}} \right),$$

where $\hat{\mathbb{V}}(Z_l)$ denotes the empirical variance of Z_l with samples $Z_{1,l}, Z_{2,l}, \cdots, Z_{n_l,l}$.

Based on Lemma 1, a new penalty term $\Gamma_{h,0}(s,a)$ for the reference part in Eqn. (10) is designed as:

$$\Gamma^{\alpha}_{h,0}(s,a) = c\sqrt{\sum\nolimits_{l\in[L]} \frac{\big(\hat{\mathbb{V}}_{h,0,l}\hat{V}^{\text{ref}}_{h+1}\big)(s,a)}{L^2 N_{h,0,l}(s,a)}} + c\sqrt{\sum\nolimits_{l\in[L]} \frac{H^2}{L^2 (N_{h,0,l}(s,a))^2}}, \tag{11}$$

where $(\hat{\mathbb{V}}_{h,0,l}\hat{V}^{\text{ref}}_{h+1})(s,a)$ denotes the empirical variance of $\hat{V}^{\text{ref}}_{h+1}(s')$ with $s' \sim \hat{\mathbb{P}}_{h,0,l}(\cdot|s,a)$.

Similarly, the following new penalty term $\Gamma_{h,1}(s,a)$ measures the uncertainty in the advantage term:

$$\Gamma^{\beta}_{h,1}(s,a) = c\sqrt{\sum\nolimits_{l\in[L]} \frac{\big[\hat{\mathbb{V}}_{h,1,l}\big(\hat{V}_{h+1} - \hat{V}^{\text{ref}}_{h+1}\big)\big](s,a)}{L^2 N_{h,1,l}(s,a)}} + c\sqrt{\sum\nolimits_{l\in[L]} \frac{H^2}{L^2 (N_{h,1,l}(s,a))^2}}, \tag{12}$$

where $[\hat{\mathbb{V}}_{h,1,l}(\hat{V}_{h+1} - \hat{V}^{\text{ref}}_{h+1})](s,a)$ is the empirical variance of $(\hat{V}_{h+1} - \hat{V}^{\text{ref}}_{h+1})(s')$ w.r.t. $\hat{\mathbb{P}}_{h,1,l}(\cdot|s,a)$. The penalty for sample uncertainties in HetPEVI-Adv is then $\Gamma^{\alpha}_h(s,a) = \Gamma^{\alpha}_{h,0}(s,a) + \Gamma^{\alpha}_{h,1}(s,a)$.

In terms of the source uncertainties, the same penalty term $\Gamma^{\beta}_h(s,a)$ as in Eqn. (8) is adopted in HetPEVI-Adv. It is a natural idea to also adopt Bernstein-type penalties here; however, it is challenging as we lack the exact information of $(\mathbb{B}_{h,l}\hat{V}_h)(s,a)$ to estimate the associated variance.

4.1 Theoretical Analysis

Theorem 2. *Under Assumption 1, w.p. at least $1-\delta$, the output policy $\hat{\pi}$ of HetPEVI-Adv satisfies*

$$\textit{Gap}(\hat{\pi};\mathcal{M}) = \tilde{O}\left(\sqrt{\frac{C^* H^3 S}{LK}} + \frac{C^* H^4 S}{\sqrt{L}K} + \sqrt{\frac{C^* H^7 S}{L^2 K}} + \sqrt{\frac{H^4}{L}} \right). \tag{13}$$

In Eqn. (13), the first three terms come from sample uncertainty and the last term comes from source uncertainty. Note that with sufficiently large L and K, the dominating gap term from sample uncertainty is the first one in Eqn. (13) of order $\tilde{O}(\sqrt{C^* H^3 S/(LK)})$, which improves that from HetPEVI by a factor of H, while the gap term from source uncertainties remains the same. The improvement in the first term comes from both the reference-advantage decomposition and the Bernstein-type penalty; see details in Appendix D.

Moreover, when learning directly with LK trajectories from the underlying MDP (thus only sample uncertainties exist), state-of-the-art works have a gap guarantee of order $\tilde{O}(\sqrt{C^* H^3 S/(LK)})$ (Xie et al., 2021b; Li et al., 2022; Shi et al., 2022). Thus, using HetPEVI-Adv, learning from heterogeneously realized environments does not increase the performance loss associated with finite data samples, but introduces an additional loss caused by the limited data sources (i.e., the last term in Eqn. (13)). This additional loss cannot be reduced by increasing the number of data samples, which again highlights the importance of ensuring source diversity.

5 EXTENSION TO OFFLINE LINEAR MDP

Lastly, we extend the study from tabular RL to incorporating function approximations. Especially, the following linear MDP model is considered.

Definition 2 (Linear MDP (Jin et al., 2020; 2021b))**.** *An MDP $\mathcal{M} = (H, \mathcal{S}, \mathcal{A}, \mathbb{P}, r)$ is said to be a linear MDP with a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if there exist d unknown measures $\mu_h = (\mu_h^{(1)}, \cdots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$ s.t. $\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \mu_h(s') \rangle$ and $r_h(s,a) = \langle \phi(s,a), \theta_h \rangle$ for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ at each step $h \in [H]$. Without loss of generality, it is assumed that $\|\phi(s,a)\|_2 \leq 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\max\{\|\mu_h(\mathcal{S})\|_2, \|\theta_h\|_2\} \leq \sqrt{d}$ for all $h \in [H]$, where $\|\mu_h(\mathcal{S})\|_2 := \int_{\mathcal{S}} \|\mu_h(s)\|_2 \, \mathrm{d}s$. For simplicity, we denote $\mathcal{M} = (H, \mathcal{S}, \mathcal{A}, \mu, \theta)$.*

With this definition, we consider the problem that each data source MDP $\mathcal{M}_l$ is a linear MDP of dimension d with the same feature map but different system dynamics, as formalized in the following assumption. We note that this shared feature mapping is commonly adopted in federated linear bandits (Huang et al., 2021; Li & Wang, 2022), which naturally extends to linear MDP.

Assumption 2. *For each $l \in [L]$, the data source MDP $\mathcal{M}_l = (H, \mathcal{S}, \mathcal{A}, \mathbb{P}_l, r_l)$ is a linear MDP with a common feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ but different measures $\{\mu_{h,l} : h \in [H]\}$ and vectors $\{\theta_{h,l} : h \in [H]\}$. In other words, it holds that $\mathcal{M}_l = (H, \mathcal{S}, \mathcal{A}, \phi, \mu_l, \theta_l)$ for all $l \in [L]$.*

Then, similar to Definition 1, an underlying linear MDP can be similarly defined as follows.

Definition 3 (Underlying Linear MDP)**.** *A linear MDP $\mathcal{M} := (H, \mathcal{S}, \mathcal{A}, \mathbb{P}, r) = (H, \mathcal{S}, \mathcal{A}, \phi, \mu, \theta)$ is said to be an underlying linear MDP to MDPs $\{\mathcal{M}_l = (H, \mathcal{S}, \mathcal{A}, \phi, \mu_l, \theta_l) : l \in [L]\}$ if for each $h \in [H]$, vectors $\{\theta_{h,l} : l \in [L]\}$ are i.i.d. random vectors with mean θ_h, and for each $(h,i) \in [H] \times [d]$, measures $\{\mu_{h,l}^{(i)} : l \in [L]\}$ are i.i.d. random measures over $\mathcal{S}$ with mean $\mu_h^{(i)}$.*

Note that when treating the tabular setting as $d = SA$ and $\phi(s,a) = \mathbf{e}_{(s,a)}$, where $\mathbf{e}_{(s,a)}$ denotes a canonical basis in $\mathbb{R}^d$, Definition 3 returns to Definition 1.

5.1 HETPEVI-LIN FOR LINEAR MDPS

From the linear MDP definition, it can be shown that the Bellman operators of each data source MDP and the underlying MDP are all linear (but with different weights).

Lemma 2. *Under Assumption 2 and Definition 3, with any function $f : \mathcal{S} \to \mathbb{R}$, there exists $\{w_{h,l}^f \in \mathbb{R}^d : h \in [H], l \in [L]\}$ and $\{w_h^f \in \mathbb{R}^d : h \in [H]\}$ s.t. for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, it holds that $(\mathbb{B}_{h,l} f)(s,a) = \langle \phi(s,a), w_{h,l}^f \rangle$ for each $l \in [L]$ and $(\mathbb{B}_h f)(s,a) = \langle \phi(s,a), w_h^f \rangle$.*

Furthermore, the following lemma can be established for the sample average MDP $\bar{\mathcal{M}} = \sum_{l \in [L]} \mathcal{M}_l / L$, which states that $\bar{\mathcal{M}}$ is also a linear MDP and the weights associated with its Bellman operator are the average of weights from each individual data source MDP.

Lemma 3. *The average MDP $\bar{\mathcal{M}} = \sum_{l \in [L]} \mathcal{M}_l / L$ is a linear MDP of dimension d with feature mapping ϕ. Also, for any function $f : \mathcal{S} \to \mathbb{R}$ and any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, it holds that $(\bar{\mathbb{B}}_h f)(s,a) = \langle \phi(s,a), \bar{w}_h^f \rangle$ with weights $\{\bar{w}_h^f = \sum_{l \in [L]} w_{h,l}^f / L : h \in [H]\}$.*

With the above observations, the HetPEVI-Lin algorithm is proposed (with detailed pseudocode in Appendix E due to the space limitation). First, as in HetPEVI, each dataset $\mathcal{D}_l$ is randomly split into H holds as $\{\mathcal{D}_{h,l} : h \in [H]\}$. Then, for each step $h \in [H]$ and each data source $l \in [L]$, using dataset $\mathcal{D}_{h,l}$, the weight $\hat{w}_{h,l}$ is estimated via a ridge regression using a regularization parameter λ:

$$\begin{aligned}\hat{w}_{h,l} &= \arg\min_{w \in \mathbb{R}^d} \sum\nolimits_{\tau \in \mathcal{D}_{h,l}} \left[r_{h,l}^\tau + \hat{V}_{h+1}(s_{h+1,l}^\tau) - \phi(s_{h,l}^\tau, a_{h,l}^\tau)^\top w \right]^2 + \lambda \|w\|_2^2 \\ &= \Lambda_{h,l}^{-1} \left[\sum\nolimits_{\tau \in \mathcal{D}_{h,l}} \phi(s_{h,l}^\tau, a_{h,l}^\tau) \left(r_{h,l}^\tau + \hat{V}_{h+1}(s_{h+1,l}^\tau) \right) \right],\end{aligned}$$

with $\Lambda_{h,l} := \sum_{\tau \in \mathcal{D}_{h,l}} \phi(s_{h,l}^\tau, a_{h,l}^\tau) \phi(s_{h,l}^\tau, a_{h,l}^\tau)^\top + \lambda I$. The agent aggregates $\hat{w}_h = \sum_{l \in [L]} \hat{w}_{h,l} / L$, which provides an estimation of the weight $\bar{w}_h$ associated with the average MDP $\bar{\mathcal{M}}$. The PEVI

in Eqn. (3) is then performed with the empirical Bellman operator defined as $(\hat{\mathbb{B}}_h \hat{V}_{h+1})(s,a) = \langle \phi(s,a), \hat{w}_h \rangle$ and a two-part penalty term $\Gamma_h(s,a) = \Gamma_h^{\alpha}(s,a) + \Gamma_h^{\beta}(s,a)$ illustrated in the following.

Penalties to Aggregate Sample Uncertainties. The first part of the penalty term aggregates the sample uncertainties from each data source, which is designed as

$$\Gamma_h^{\alpha}(s,a) = c\sqrt{(dH^2/L + dH^2\lambda) \sum\nolimits_{l\in[L]} \|\phi(s,a)\|_{\Lambda_{h,l}^{-1}}}. \tag{14}$$

While this penalty term shares a similar format as those in linear bandits (Li et al., 2010; Abbasi-Yadkori et al., 2011) and linear MDPs (Jin et al., 2020; 2021b), its novelty comes from jointly considering the uncertainties of different sources. Especially, this design avoids applying the self-normalized concentration (Abbasi-Yadkori et al., 2011) to each data source (which leads to an inferior dependency on L) by leveraging the statistical independence among the split datasets.

Penalties to Account for Source Uncertainties. Besides the sample uncertainties, the following penalty is designed to measure the source uncertainties:

$$\Gamma_h^{\beta}(s,a) = c\sqrt{dH^2/L}. \tag{15}$$

Note that compared with Eqn. (8), an additional d factor appears in Eqn. (15), which is from the covering argument over the d-dimensional feature space.

5.2 Theoretical Analysis

Similar to the discussions in the tabular setting, the following coverage assumption is first presented.

Assumption 3. *There exists a constant $D^* < \infty$ such that for each $l \in [L]$, the behavior policy ρ_l satisfies $\mathbb{E}_{\pi^*,\mathcal{M}}\left[\phi(s_h,a_h)\phi(s_h,a_h)^\top\right] \preceq D^* \mathbb{E}_{\rho_l,\mathcal{M}_l}\left[\phi(s_h,a_h)\phi(s_h,a_h)^\top\right]$ for all $h \in [H]$.*

Assumption 3 is a generalization of Assumption 1 by treating the tabular setting as using features of canonical bases. Similar (or stronger) assumptions are commonly adopted in offline linear RL (Jin et al., 2021b; Yin et al., 2022; Min et al., 2021).

Theorem 3. *Under Assumptions 2, 3 and assuming the matrices $\{\sum_{l\in[L]} \Lambda_{h,l}^{-1} : h \in [H]\}$ are invertible, with $\lambda = 1/L$, w.p. at least $1-\delta$, the output policy $\hat{\pi}$ of HetPEVI-Lin satisfies*

$$\textit{Gap}(\hat{\pi};\mathcal{M}) = \tilde{O}\left(\sqrt{D^* d^2 H^5/(KL)} + \sqrt{dH^4/L}\right). \tag{16}$$

Similar to Thm. 1, the two terms in Eqn. (16) originate from the finite data samples and the limited data sources, respectively. However, note that the gap guarantee in Thm. 3 does not have a dependency on the number of states S (which appears in Thms. 1 and 3), instead it depends on the feature dimension d, thanks to the careful design that incorporates linear function approximation. As a result, to output an ε-optimal policy, the sample complexity requirement is $T = KL = \tilde{O}(D^* d^2 H^5/\varepsilon^2)$ while the source diversity requirement is $L = \tilde{O}(dH^4/\varepsilon^2)$.

Remark 4. A natural but open question is how to adopt reference-advantage decompositions and variance information to HetPEVI-Lin, which may also improve the H-dependency in the sample complexity requirement. The recent work (Xiong et al., 2022) provides some insight into learning from a single source MDP, but major challenges exist in extending it to learning the underlying MDP, which requires careful aggregations of multiple sources (as in Eqn. (14)).

6 Conclusions

This work studied efficient offline RL with datasets from heterogeneous sources. A novel problem of learning a hidden underlying MDP via offline datasets collected from several realized MDPs perturbed from the target was investigated. The HetPEVI and HetPEVI-Adv algorithms were proposed, where novel penalty terms were designed to jointly consider the uncertainties from the finite data samples and the limited amount of data sources. Theoretical analyses proved their performance guarantees and in particular, highlighted the importance of ensuring source diversity. In particular, the result of HetPEVI-Adv illustrated that learning from heterogeneously realized environments only incurs additional performance losses due to the finite data sources. Lastly, linear function approximation was considered, and the HetPEVI-Lin algorithm was developed with penalties carefully designed to consider both uncertainties and incorporate the linear structure.

REFERENCES

Abbasi-Yadkori, Y., Pál, D., & Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

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

Azar, M. G., Osband, I., & Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning* (pp. 263–272).: PMLR.

Brunskill, E. & Li, L. (2013). Sample complexity of multi-task reinforcement learning. In *Uncertainty in Artificial Intelligence* (pp. 122).: Citeseer.

Chua, K., Lei, Q., & Lee, J. D. (2021). Provable hierarchy-based meta-reinforcement learning. *arXiv preprint arXiv:2110.09507*.

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

Dann, C., Mohri, M., Zhang, T., & Zimmert, J. (2021). A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34.

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

Duan, Y., Jin, C., & Li, Z. (2021). Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning* (pp. 2892–2902).: PMLR.

Fallah, A., Georgiev, K., Mokhtari, A., & Ozdaglar, A. (2021). On the convergence theory of debiased model-agnostic meta-reinforcement learning. *Advances in Neural Information Processing Systems*, 34.

Gupta, A., Mendonca, R., Liu, Y., Abbeel, P., & Levine, S. (2018). Meta-reinforcement learning of structured exploration strategies. *Advances in neural information processing systems*, 31.

Hallak, A., Di Castro, D., & Mannor, S. (2015). Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*.

Huang, R., Wu, W., Yang, J., & Shen, C. (2021). Federated linear contextual bandits. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*.

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

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., & Schapire, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning* (pp. 1704–1713).: PMLR.

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

Jin, C., Liu, Q., & Miryoosefi, S. (2021a). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34.

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

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

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

Johnson, R. & Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26.

Kidambi, R., Rajeswaran, A., Netrapalli, P., & Joachims, T. (2020). Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33, 21810–21823.

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

Kwon, J., Efroni, Y., Caramanis, C., & Mannor, S. (2021a). Reinforcement learning in reward-mixing mdps. *Advances in Neural Information Processing Systems*, 34.

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

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

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

Li, C. & Wang, H. (2022). Asynchronous upper confidence bound algorithms for federated linear bandits. In *International Conference on Artificial Intelligence and Statistics* (pp. 6529–6553).: PMLR.

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

Li, J., Vuong, Q., Liu, S., Liu, M., Ciosek, K., Christensen, H., & Su, H. (2020a). Multi-task batch reinforcement learning with metric learning. *Advances in Neural Information Processing Systems*, 33, 6197–6210.

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web* (pp. 661–670).

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

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

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

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

Maurer, A. & Pontil, M. (2009). Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*.

Min, Y., Wang, T., Zhou, D., & Gu, Q. (2021). Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems*, 34.

Mirsky, L. (1975). A trace inequality of john von neumann. *Monatshefte für mathematik*, 79(4), 303–306.

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

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

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

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

Sæmundsson, S., Hofmann, K., & Deisenroth, M. (2018). Meta reinforcement learning with latent variable gaussian processes. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, volume 34 (pp. 642–652).: Association for Uncertainty in Artificial Intelligence (AUAI).

Sallab, A. E., Abdou, M., Perot, E., & Yogamani, S. (2017). Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19), 70–76.

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

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

Si, N., Zhang, F., Zhou, Z., & Blanchet, J. (2020). Distributionally robust policy evaluation and learning in offline contextual bandits. In *International Conference on Machine Learning* (pp. 8884–8894).: PMLR.

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

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

Teh, Y., Bapst, V., Czarnecki, W. M., Quan, J., Kirkpatrick, J., Hadsell, R., Heess, N., & Pascanu, R. (2017). Distral: Robust multitask reinforcement learning. *Advances in neural information processing systems*, 30.

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

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

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

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

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

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

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

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

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

Yu, T., Kumar, A., Chebotar, Y., Hausman, K., Levine, S., & Finn, C. (2021). Conservative data sharing for multi-task offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34.

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

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

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

Zhang, K., Yang, Z., & Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, (pp. 321–384).

Zhang, Z., Zhou, Y., & Ji, X. (2020). Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 15198–15207.

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

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

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

A Additional Discussions

A.1 Related Works

RL has seen much progress over the past few years, particularly in its theoretical understandings (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.

Online RL with One Environment. The majority of RL studies focus on the online setting (Sutton & Barto, 2018), where the agent consistently improves her algorithm via online interactions with the environment. In recent years, many advances have deepened the theoretical understandings of online RL in interacting with a fixed and stable environment, with provably efficient designs (Azar et al., 2017; Jin et al., 2018; 2020; 2021a; Zhang et al., 2020; Jiang et al., 2017; Dann et al., 2021).

Online RL with Heterogeneous Environments. The importance of handling heterogeneous environments is well recognized. Especially, federated RL (Zhuo et al., 2019) and the general multi-agent RL (Zhang et al., 2021) study the setting with multiple agents in the system, while each of them may view an agent-depend environment (Jin et al., 2022). In contrast, meta-RL (Sæmundsson et al., 2018; Gupta et al., 2018; Fallah et al., 2021; Chua et al., 2021) focuses on extracting knowledge from previous tasks and adapting them to new (potentially different) tasks. Also, multi-task RL (Teh et al., 2017) attempts to handle multiple tasks together via information shared among tasks (Brunskill & Li, 2013; Zhang & Wang, 2021; Lu et al., 2021; Yang et al., 2020; 2022) or distinguishing the latent structure associated with each task via past observations (Hallak et al., 2015; Kwon et al., 2021a;b).

Offline RL with Single-source Datasets. Offline RL (Levine et al., 2020) attempts to avoid potentially expensive interactions with the environment as in online RL but instead use available offline datasets collected previously. 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., 2021b; Rashidinejad et al., 2021). Especially, it is illustrated that a nearly optimal policy can be learned via a dataset collected by a behavior policy that covers the trajectories of the optimal policy.

Following this line, Xie et al. (2021b); Li et al. (2022); Shi et al. (2022) further fine-tune the designs for the tabular setting. Especially, Xie et al. (2021b) adapts the idea of "reference-advantage" decomposition from the online setting (Zhang et al., 2020) (which further originates from the "variance reduction" technique in optimization (Johnson & Zhang, 2013)) and obtains a performance that approaches the lower bound.

For the linear setting, following Jin et al. (2021b), Zanette et al. (2021) incorporates the idea of pessimism to develop an actor-critic method. Depending (explicitly or implicitly) on a strong assumption that the collected data are independent across time steps, Min et al. (2021); Yin et al. (2022) introduce the idea of "variance reduction" via weighting the observations with the corresponding estimated variance. The recent work of Xiong et al. (2022) gets rid of such an assumption, the designs of which however are still limited to handling the single-source dataset. For the general function approximation, additional attempts are reported in Xie et al. (2021a); Uehara & Sun (2021). However, these results are mostly information-theoretical as computational intractable optimizations are required in general.

Offline Robust RL. Besides the above discussions, one particular line of research that relates to this work is under the category of offline robust RL (Zhou et al., 2021; Yang et al., 2021; Panaganti et al., 2022; Panaganti & Kalathil, 2022; Si et al., 2020). Especially, offline robust RL learns via data sampled from *one* MDP and tries to output a policy that is robustly good for a family of MDPs; however, our work attempts at learning the hidden common structure from data sampled from a family of MDPs. Nevertheless, we note that it would be an interesting question to investigate whether having access to multiple data source MDPs in the target family would make offline robust RL easier.

Offline RL with Multi-source Datasets. The aforementioned advances on offline RL are mainly focused on learning from a single data source, which limits their applicability. In the offline domain, growing interests have been made to utilize data from heterogeneous sources. The most related literature fall under the framework of "offline meta-RL" or "offline multi-task RL" (Mitchell et al., 2021; Dorfman et al., 2021; Lin et al., 2022; Li et al., 2020b;a; Yu et al., 2021). Especially, the underlying 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 the (currently lacking) theoretical understanding of offline meta-RL systems, especially the meta-training process, which may also serve as the foundation for studies on the meta-testing process.

A.2 Future Works

Coverage Assumptions. While the coverage assumption of Assumptions 1 and 3 are reasonable and similar ones are well adopted, it is still of major interest to further explore how to perform offline RL (especially with heterogeneous data sources) under weaker assumptions. This direction is particularly interesting with multiple data sources since the heterogeneous sources naturally enrich the data diversity.

Lower Bounds. For the offline RL with a single data source, various lower bounds are provided in different settings (e.g., Rashidinejad et al. (2021); Xie et al. (2021b) for the tabular setting, Jin et al. (2021b); Yin et al. (2022) for the linear setting). It would be interesting while considerably challenging to derive a generally applicable lower bound for offline RL of the underlying MDP with multiple heterogeneous data sources, which needs to consider both sample and source uncertainties simultaneously.

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 (e.g., the learned underlying MDP) have 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 Overview

In this section, we provide an overview with the steps shared in the proofs of each proposed algorithm. In other words, the following results are stated for all three proposed algorithms in general. The more challenging and specific proofs for our designs are illustrated in the following sections. The basic logistics here follow the seminal work of Jin et al. (2021b) in provably efficient offline RL.

The first step is to establish the validness of the pessimism. For different settings, various styles of pessimism are incorporated. Specifically, in this work, three penalty constructions are adopted in HetPEVI, HetPEVI-Adv and HetPEVI-Lin, respectively. As an abstraction, the validness of pessimism induced by penalties $\{\Gamma_h(s,a) : (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]\}$ is defined in the following.

Definition 4 (Validness of Pessimism)**.** *For pessimistic value iterations using an estimated Bellman operator $\hat{\mathbb{B}}_h$ and penalties $\{\Gamma_h(s,a) : (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]\}$, a valid pessimism is induced if the following event happens*

$$\mathcal{E} := \left\{ \left| \left(\hat{\mathbb{B}}_h \hat{V}_{h+1} \right)(s,a) - \left(\mathbb{B}_h \hat{V}_{h+1} \right)(s,a) \right| \leq \Gamma_h(s,a), \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H] \right\}.$$

With a valid pessimism, we can obtain the following lemma.

Lemma 4. *Suppose that event $\mathcal{E}$ in Definition 4 happens, it holds that*

$$\zeta_h(s,a) := \left(\mathbb{B}_h \hat{V}_{h+1} \right)(s,a) - \hat{Q}_h(s,a) \in [0, 2\Gamma_h(s,a)], \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H].$$

Proof. We can observe that for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, with the adopted pessimistic value iteration, it holds that

$$\zeta_h(s,a) = \left(\mathbb{B}_h \hat{V}_{h+1} \right)(s,a) - \hat{Q}_h(s,a)$$

$$
\begin{aligned}
&= \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a) - \min\left\{\left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \Gamma_h(s,a), H-h+1\right\}^+ \\
&\geq \min\left\{\left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a) - \left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) + \Gamma_h(s,a), \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a)\right\} \\
&\geq 0,
\end{aligned}
$$

where the last inequality is due to event $\mathcal{E}$. Similar, for the other direction, it holds that

$$
\begin{aligned}
\zeta_h(s,a) &= \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a) - \hat{Q}_h(s,a) \\
&= \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a) - \min\left\{\left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \Gamma_h(s,a), H-h+1\right\}^+ \\
&\leq \max\left\{\left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a) - \left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) + \Gamma_h(s,a), \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a) - (H-h+1)\right\} \\
&\leq 2\Gamma_h(s,a),
\end{aligned}
$$

where the last inequality is due to event $\mathcal{E}$ and the fact that $(\mathbb{B}_h \hat{V}_{h+1})(s,a) \leq H-h+1$. □

Furthermore, the suboptimality gap on the target MDP $\mathcal{M}$ between the output policy $\hat{\pi}$ and the optimal policy π^* can be generally bounded via the following lemma.

Lemma 5. *Suppose that event $\mathcal{E}$ in Definition 4 happens, it holds that*

$$
\textit{Gap}(\hat{\pi};\mathcal{M}) \leq 2\sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\Gamma_h(s_h,a_h)\right].
$$

Proof. It holds that

$$
\begin{aligned}
\text{Gap}(\hat{\pi};\mathcal{M}) &\overset{(a)}{=} -\sum_{h\in[H]} \mathbb{E}_{\hat{\pi},\mathcal{M}}\left[\zeta_h(s_h,a_h)\right] + \sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\zeta_h(s_h,a_h)\right] \\
&\quad + \sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\hat{Q}_h(s_h,\pi^*(s_h)) - \hat{Q}_h(s_h,\hat{\pi}_h(s_h))\right] \\
&\overset{(b)}{\leq} 2\sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\Gamma_h(s_h,a_h)\right],
\end{aligned}
$$

where equation (a) is from Lemma 3.1 in (Jin et al., 2021b) (provided as Lemma 30), and inequality (b) is due to Lemma 4 together with the fact that $\hat{Q}_h(s_h,\cdot)$ takes maximum at $\hat{\pi}_h(s_h)$ in the proposed algorithms. □

In other words, $\text{Gap}(\hat{\pi};\mathcal{M})$ is bounded via the sum of penalties $\Gamma_h(s_h,a_h)$ along the expected trajectory of optimal policy π^* on the target MDP $\mathcal{M}$.

C THE HETPEVI ALGORITHM

C.1 VALIDNESS OF PESSIMISM

Lemma 6. *With the penalties $\Gamma_h^\alpha(s,a) = c\sqrt{\sum_{l\in[L]} \frac{H^2}{L^2 N_{h,l}(s,a)}}$ in HetPEVI, with probability at least $1-\delta$, it holds that for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$,*

$$
\left|\left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \left(\bar{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a)\right| \leq \Gamma_h^\alpha(s,a).
$$

Proof. For a fixed (s,a,h), if there exists a $l \in [L]$ such that $N_{h,l}(s,a) = 0$, the claim naturally holds. Thus, we consider the scenarios with $N_{h,l}(s,a) \geq 1, \forall l \in [L]$, where it holds that

$$
\begin{aligned}
&\left|\left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \left(\bar{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a)\right| \\
&= \left|\hat{r}_h(s,a) - \bar{r}_h(s,a) + \left(\hat{\mathbb{P}}_h \hat{V}_{h+1}\right)(s,a) - \left(\bar{\mathbb{P}}_h \hat{V}_{h+1}\right)(s,a)\right|
\end{aligned}
$$

$$= \left| \left(\hat{\mathbb{P}}_h \hat{V}_{h+1} \right)(s,a) - \left(\bar{\mathbb{P}}_h \hat{V}_{h+1} \right)(s,a) \right|$$
$$= \left| \sum_{l \in [L]} \frac{1}{L} \left(\hat{\mathbb{P}}_{h,l} \hat{V}_{h+1} \right)(s,a) - \sum_{l \in [L]} \frac{1}{L} \left(\mathbb{P}_{h,l} \hat{V}_{h+1} \right)(s,a) \right|.$$

Since the data to estimate $\hat{V}_{h+1}(\cdot)$ and $\hat{\mathbb{P}}_{h,l}(\cdot|s,a)$ are independent, we can apply the Hoeffding inequality and a union bound to obtain the lemma. □

Lemma 7. *For the penalties $\Gamma_h^\beta(s,a) = c\sqrt{\frac{H^2}{L}}$ in HetPEVI, with probability at least $1-\delta$, it holds that for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$,*

$$\left| \left(\bar{\mathbb{B}}_h \hat{V}_{h+1} \right)(s,a) - \left(\mathbb{B}_h \hat{V}_{h+1} \right)(s,a) \right| \leq \Gamma_h^\beta(s,a).$$

Proof. For a fixed (s,a,h), since $(\mathbb{B}_{h,l}\hat{V}_{h+1})(s,a)$ is bounded between $[0,H]$ and has an expectation of $(\mathbb{B}_h \hat{V}_{h+1})(s,a)$, it is then a H-sub-Gaussian random variable (Vershynin, 2018). Thus, with an additional union bound, the lemma can be obtained. □

Lemma 8. *The penalties $\{\Gamma_h(s,a) = \Gamma_h^\alpha(s,a) + \Gamma_h^\beta(s,a) : (s,a,h) : (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]\}$ in HetPEVI induce a valid pessimism with probability at least $1-\delta$ with respect to the estimated Bellman operator $\hat{\mathbb{B}}_h$ in Eqn. (4).*

Proof. This result can be obtained by combining Lemmas 6 and 7. □

C.2 Good Event

Lemma 9. *The following event holds with probability at least $1-\delta$ for HetPEVI*

$$\mathcal{G} := \Bigg\{ \text{(i) The penalties } \{\Gamma_h(s,a)\} \text{ induce a valid pessimism;}$$
$$\text{(ii) } \frac{1}{N_{h,l}(s,a)} \leq c\frac{H}{K d_{h,l}^{\rho_l}(s,a)}, \forall (l,s,a,h) \in [L] \times \mathcal{S} \times \mathcal{A} \times [H] \Bigg\}.$$

Proof. Part (I) is from Lemma 8 and part (II) obtained via the following lemma. □

Lemma 10. *With probability at least $1-\delta$, it holds that,*

$$\frac{1}{N_{h,l}(s,a)} \leq c\frac{H}{K d_{h,l}^{\rho_l}(s,a)}, \quad \forall (l,s,a,h) \in [L] \times \mathcal{S} \times \mathcal{A} \times [H].$$

Proof. The proof can be done similarly as in Lemma B.3 of (Xie et al., 2021b). □

C.3 Suboptimaity Gap

Proof of Theorem 1. By Lemma 5 and 9, with probability at least $1-\delta$, it holds that

$$\text{Gap}(\hat{\pi}; \mathcal{M}) \leq 2 \sum_{h \in [H]} \mathbb{E}_{\pi^*, \mathcal{M}} \left[\Gamma_h(s_h, a_h) \right] = \tilde{O} \left(\sum_{h \in [H]} \mathbb{E}_{\pi^*, \mathcal{M}} \left[\sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 N_{h,l}(s,a)}} + \sqrt{\frac{H^2}{L}} \right] \right).$$

Furthermore, with Assumption 1, it holds that

$$\sum_{h \in [H]} \mathbb{E}_{\pi^*, \mathcal{M}} \left[\sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 N_{h,l}(s,a)}} \right] \leq \sum_{h \in [H]} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\pi^*}(s,a) \sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 N_{h,l}(s,a)}}$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) \sqrt{\sum_{l\in[L]} \frac{H^3}{L^2 K d_{h,l}^{\rho_l}(s,a)}} \\
&\overset{(b)}{\leq} \sqrt{\frac{C^* H^3}{LK}} \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{\mathbb{1}\{a=\pi_h^*(s)\} d_h^{\pi^*}(s,a)} \\
&\overset{(c)}{\leq} \sqrt{\frac{C^* H^3}{LK}} \sqrt{SH} \sqrt{\sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a)} \\
&\leq \sqrt{\frac{C^* H^3}{LK}} \sqrt{SH} \sqrt{H} \\
&= \tilde{O}\left(\sqrt{\frac{C^* H^5 S}{LK}}\right);
\end{aligned}
$$

$$
\sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\sqrt{\frac{H^2}{L}}\right] \leq \sqrt{\frac{H^4}{L}},
$$

where inequality (a) is from (ii) of event $\mathcal{G}$ in Lemma 9, inequality (b) is from Assumption 1, and inequality (c) is from the Cauchy-Schwarz inequality. □

D THE HETPEVI-ADV ALGORITHM

D.1 INTUITIONS FOR THE PERFORMANCE IMPROVEMENT OF HETPEVI-ADV

As mentioned in the main paper, an H factor is saved in the loss from sample uncertainties of HetPEVI-Adv, compared with HetPEVI. Some intuitions are provided in the following. Specifically, for the reference term and the associated penalty $\Gamma_{h,1}^{\alpha}$, the estimation of $\hat{\mathbb{P}}_{h,0}$ for each step $h \in [H]$ all utilizes the whole dataset $\mathcal{D}_0$ without any H-folding splitting, which saves one $\sqrt{H}$ factor. Moreover, the Bernstein-type penalty incorporates additional considerations of the variances, whose advantage becomes evident when summing over H steps, leading to an improvement of a $\sqrt{H}$ factor. For the advantage term, although the estimation of $\hat{\mathbb{P}}_{h,1}$ is from dataset $\mathcal{D}_{h,1}$, which is one of the H-fold splits of $\mathcal{D}_1$, the range of $\hat{V}_{h+1} - \hat{V}_{h+1}^{\text{ref}}$ is small as $\hat{V}_{h+1}^{\text{ref}}$ is learned via HetPEVI, which compensates for the inefficiency of H-fold splitting.

D.2 VALIDNESS OF PESSIMISM

Lemma 11. *For the penalties $\Gamma_h^{\alpha}(s,a) = \Gamma_{h,0}^{\alpha}(s,a) + \Gamma_{h,1}^{\alpha}(s,a)$ in HetPEVI-Adv, where*

$$
\Gamma_{h,0}^{\alpha}(s,a) = c\sqrt{\sum_{l\in[L]} \frac{\left(\hat{\mathbb{V}}_{h,0,l}\hat{V}_{h+1}^{ref}\right)(s,a)}{L^2 N_{h,0,l}(s,a)}} + c\sqrt{\sum_{l\in[L]} \frac{H^2}{L^2 (N_{h,0,l}(s,a))^2}}
$$

and

$$
\Gamma_{h,1}^{\alpha}(s,a) = c\sqrt{\sum_{l\in[L]} \frac{\left[\hat{\mathbb{V}}_{h,1,l}\left(\hat{V}_{h+1} - \hat{V}_{h+1}^{ref}\right)\right](s,a)}{L^2 N_{h,1,l}(s,a)}} + c\sqrt{\sum_{l\in[L]} \frac{H^2}{L^2 (N_{h,1,l}(s,a))^2}},
$$

with probability of at least $1-\delta$, it holds that for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$,

$$
\left|\left(\hat{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \left(\bar{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a)\right| \leq \Gamma_h^{\alpha}(s,a).
$$

Proof. For a fixed (s,a,h), if there exists a $l \in [L]$ such that $N_{h,0,l}(s,a) = 0$ or $N_{h,1,l}(s,a) = 0$, the claim naturally holds. Thus, we consider the scenarios with $N_{h,0,l}(s,a), N_{h,1,l}(s,a) \geq 1, \forall l \in [L]$,

and thus

$$
\begin{aligned}
&\left|(\hat{\mathbb{B}}_h \hat{V}_{h+1})(s,a) - (\bar{\mathbb{B}}_h \hat{V}_{h+1})(s,a)\right| \\
&= \left|\hat{r}_h(s,a) - \bar{r}_h(s,a) + \left(\hat{\mathbb{P}}_{h,0} V_{h+1}^{\text{ref}}\right)(s,a) + \left[\hat{\mathbb{P}}_{h,1}\left(\hat{V}_{h+1} - \hat{V}_{h+1}^{\text{ref}}\right)\right](s,a) - \bar{\mathbb{P}}_h \hat{V}_{h+1}(s,a)\right| \\
&\leq \left|\left(\hat{\mathbb{P}}_{h,0} V_{h+1}^{\text{ref}}\right)(s,a) - \left(\bar{\mathbb{P}}_h V_{h+1}^{\text{ref}}\right)(s,a)\right| \\
&+ \left|\left[\hat{\mathbb{P}}_{h,1}\left(\hat{V}_{h+1} - \hat{V}_{h+1}^{\text{ref}}\right)\right](s,a) - \left[\bar{\mathbb{P}}_{h,1}\left(\hat{V}_{h+1} - \hat{V}_{h+1}^{\text{ref}}\right)\right](s,a)\right| \\
&\leq \underbrace{\left|\sum_{l\in[L]} \frac{1}{L}\left(\hat{\mathbb{P}}_{h,0,l} \hat{V}_{h+1}^{\text{ref}}\right)(s,a) - \sum_{l\in[L]} \frac{1}{L}\left(\mathbb{P}_{h,l} \hat{V}_{h+1}^{\text{ref}}\right)(s,a)\right|}_{\text{term (I)}} \\
&+ \underbrace{\left|\sum_{l\in[L]} \frac{1}{L}\left[\hat{\mathbb{P}}_{h,1,l}\left(\hat{V}_{h+1} - \hat{V}_{h+1}^{\text{ref}}\right)\right](s,a) - \sum_{l\in[L]} \frac{1}{L}\left[\mathbb{P}_{h,l}\left(\hat{V}_{h+1} - \hat{V}_{h+1}^{\text{ref}}\right)\right](s,a)\right|}_{\text{term (II)}}
\end{aligned}
$$

Since the estimations of $\hat{V}_{h+1}^{\text{ref}}(\cdot)$ and $\hat{\mathbb{P}}_{h,0,l}(\cdot|s,a)$ are independent, we can apply the enhanced version of empirical Bernstein inequality in Lemma 33 and with a union bound over $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, get that with probability $1-\delta$, it holds that for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$

$$\text{term (I)} \leq \Gamma_{h,0}^{\alpha}(s,a).$$

Also, since the estimations of $\left(\hat{V}_{h+1} - \hat{V}_{h+1}^{\text{ref}}\right)(\cdot)$ and $\hat{\mathbb{P}}_{h,1,l}(\cdot|s,a)$ are independent, similarly via Lemma 33, with probability $1-\delta$, it holds that for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$

$$\text{term (II)} \leq \Gamma_{h,1}^{\alpha}(s,a).$$

□

Lemma 12. *For the penalties $\Gamma_h^{\beta}(s,a) = c\sqrt{\frac{H^2}{L}}$ in HetPEVI, with probability at least $1-\delta$, it holds that for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$,*

$$\left|\left(\bar{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a)\right| \leq \Gamma_h^{\beta}(s,a).$$

Proof. Similar to Lemma 7. □

Lemma 13. *The penalties $\Gamma_h(s,a) = \Gamma_h^{\alpha}(s,a) + \Gamma_h^{\beta}(s,a)$ in HetPEVI-Adv, induce a valid pessimism with a probability of at least $1-\delta$ w.r.t. the estimated Bellman operator $\hat{\mathbb{B}}_h$ in Eqn. (10).*

Proof. This result can be obtained by combining Lemmas 11 and 12. □

D.3 Good Event

Lemma 14. *Under Assumption 1, with probability at least $1-\delta$, the following good event $\mathcal{G}$ happens, where*

$$\mathcal{G} := \Bigg\{ \text{(i) The penalties } \{\Gamma_h(s,a) : \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]\} \text{ induce a valid pessimism;}$$

$$\text{(ii) } V_h^*(s) - \hat{V}_h^{ref}(s) = \tilde{O}\left(\sqrt{\frac{C^* H^5 S}{KL}} + \sqrt{\frac{H^4}{L}}\right), \forall (s,h) \in \mathcal{S}\times[H];$$

$$\text{(iii) } V_h^*(s) - \hat{V}_h(s) = \tilde{O}\left(\sqrt{\frac{C^* H^5 S}{KL}} + \sqrt{\frac{H^4}{L}}\right), \forall (s,h) \in \mathcal{S}\times[H];$$

(iv) $\frac{1}{N_{h,0,l}(s,a)} \leq \frac{c}{K d_{h,l}^{\rho_l}(s,a)}, \forall (l,s,a,h) \in [L] \times \mathcal{S} \times \mathcal{A} \times [H]$;

(v) $\frac{1}{N_{h,1,l}(s,a)} \leq \frac{cH}{K d_{h,l}^{\rho_l}(s,a)}, \forall (l,s,a,h) \in [L] \times \mathcal{S} \times \mathcal{A} \times [H]$;

(vi) $\left(\hat{\mathbb{V}}_{h,0,l} f\right)(s,a) \leq \left(\mathbb{V}_{h,l} f\right)(s,a) + c\sqrt{\frac{H^4}{N_{h,0,l}(s,a)}}, \forall (l,s,a,h) \in [L] \times \mathcal{S} \times \mathcal{A} \times [H]$;

(vii) $\left(\bar{\mathbb{V}}_h f\right)(s,a) \leq \left(\mathbb{V}_h f\right)(s,a) + \sqrt{\frac{H^4}{L}}, \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]\Big\}$,

where f is any fixed function: $\mathcal{S} \times \mathcal{A} \to [-H, H]$.

Proof. Part (i) is from Lemma 13, and the rest parts are established in the following lemmas. □

Lemma 15. *Under Assumption 1, with probability at least $1-\delta$, the following bound, regarding the reference values from HetPEVI using dataset $\mathcal{D}_{ref}$, happen*

$$V_h^*(s) - \hat{V}_h^{ref}(s) = \tilde{O}\left(\sqrt{\frac{C^* H^5 S}{KL}} + \sqrt{\frac{H^4}{L}}\right), \forall (s,h) \in \mathcal{S} \times [H].$$

Proof. The bound can be obtained similarly via similar steps of Theorem 1. □

Lemma 16. *Under Assumption 1, with a probability of at least $1-\delta$, the following crude bound of the output value of HetPEVI-Adv holds*

$$V_h^*(s) - \hat{V}_h(s) = \tilde{O}\left(\sqrt{\frac{C^* H^5 S}{KL}} + \sqrt{\frac{H^4}{L}}\right), \qquad \forall (s,a) \in \mathcal{S} \times [H]$$

Proof. Since $\hat{V}_h^{\text{ref}}(s) \in [0, H]$ and $\hat{V}_h(s) - \hat{V}_h^{\text{ref}}(s) \in [-H, H]$, it holds that

$$\Gamma_{h,0}^{\alpha}(s,a) \leq c\sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 N_{h,0,l}(s,a)}} + c\sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 (N_{h,0,l}(s,a))^2}} \leq c\sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 N_{h,0,l}(s,a)}},$$

and

$$\Gamma_{h,1}^{\alpha}(s,a) \leq c\sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 N_{h,1,l}(s,a)}} + c\sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 (N_{h,1,l}(s,a))^2}} \leq c\sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 N_{h,1,l}(s,a)}},$$

which implies that

$$\Gamma_h^{\alpha}(s,a) = \Gamma_{h,0}^{\alpha}(s,a) + \Gamma_{h,1}^{\alpha}(s,a) \leq c\sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 N_{h,0,l}(s,a)}} + c\sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 N_{h,1,l}(s,a)}}.$$

With

$$\Gamma_h^{\beta}(s,a) = c\sqrt{\frac{H^2}{L}},$$

we have

$$\Gamma_h(s,a) = \Gamma_h^{\alpha}(s,a) + \Gamma_h^{\beta}(s,a) \leq c\sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 N_{h,0,l}(s,a)}} + c\sqrt{\sum_{l \in [L]} \frac{H^2}{L^2 N_{h,1,l}(s,a)}} + c\sqrt{\frac{H^2}{L}}$$

Using the above upper bound of $\Gamma_h(s,a)$, the lemma can be obtained similar via similar steps of Theorem 1. □

Lemma 17. *With probability at least $1-\delta$, the following events happen,*

$$\frac{1}{N_{h,0,l}(s,a)} \leq \frac{c}{K d_{h,l}^{\rho_l}(s,a)}, \qquad \forall (l,s,a,h) \in [L] \times \mathcal{S} \times \mathcal{A} \times [H],$$
$$\frac{1}{N_{h,1,l}(s,a)} \leq \frac{cH}{K d_{h,l}^{\rho_l}(s,a)}, \qquad \forall (l,s,a,h) \in [L] \times \mathcal{S} \times \mathcal{A} \times [H].$$

Proof. The proof can be done similarly as in Lemma B.3 of (Xie et al., 2021b) while recognizing $|\mathcal{D}_{0,l}| = K/3$ and $|\mathcal{D}_{1,h,l}| = K/(3H)$ for all $(l,h) \in [L] \times [H]$. □

Lemma 18. *With probability at least $1-\delta$, for a fix functions $f : \mathcal{S} \times \mathcal{A} \to [-H, H]$, the following events happen*

$$\left(\hat{\mathbb{V}}_{h,0,l} f\right)(s,a) \leq \left(\mathbb{V}_{h,l} f\right)(s,a) + c\sqrt{\frac{H^4}{N_{h,0,l}(s,a)}}, \qquad \forall (l,s,a,h) \in [L] \times \mathcal{S} \times \mathcal{A} \times [H].$$

Proof. With a union bound over $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, according to Hoeffding inequality, with probability at least $1-\delta$, it holds that

$$\begin{aligned}
&\left(\hat{\mathbb{V}}_{h,0,l} f\right)(s,a) - \left(\mathbb{V}_{h,l} f\right)(s,a) \\
&= \left(\hat{\mathbb{P}}_{h,0,l} f^2\right)(s,a) - \left[\left(\hat{\mathbb{P}}_{h,0,l} f\right)(s,a)\right]^2 - \left(\mathbb{P}_{h,l} f^2\right)(s,a) + \left[\left(\mathbb{P}_{h,l} f\right)(s,a)\right]^2 \\
&= \left[\left(\hat{\mathbb{P}}_{h,0,l} - \mathbb{P}_{h,l}\right) f^2\right](s,a) + \left[\left(\mathbb{P}_{h,l} - \hat{\mathbb{P}}_{h,0,l}\right) f\right](s,a) \cdot \left[\left(\mathbb{P}_{h,l} + \hat{\mathbb{P}}_{h,0,l}\right) f\right](s,a) \\
&\leq c\sqrt{\frac{H^4}{N_{h,0,l}(s,a)}}.
\end{aligned}$$

□

Lemma 19. *With probability at least $1-\delta$, for a fix functions $f : \mathcal{S} \times \mathcal{A} \to [-H, H]$, the following events happen*

$$\left(\bar{\mathbb{V}}_h f\right)(s,a) \leq \left(\mathbb{V}_h f\right)(s,a) + c\sqrt{\frac{H^4}{L}}, \qquad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H].$$

Proof. With a union bound over $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, according to Hoeffding inequality, with probability at least $1-\delta$, it holds that

$$\begin{aligned}
&\left(\bar{\mathbb{V}}_{h,l} f\right)(s,a) - \left(\mathbb{V}_{h,l} f\right)(s,a) \\
&= \left(\bar{\mathbb{P}}_h f^2\right)(s,a) - \left[\left(\bar{\mathbb{P}}_h f\right)(s,a)\right]^2 - \left(\mathbb{P}_h f^2\right)(s,a) + \left[\left(\mathbb{P}_h f\right)(s,a)\right]^2 \\
&= \left[\left(\bar{\mathbb{P}}_h - \mathbb{P}_h\right) f^2\right](s,a) + \left[\left(\bar{\mathbb{P}}_h - \mathbb{P}_h\right) f\right](s,a) \cdot \left[\left(\bar{\mathbb{P}}_h + \mathbb{P}_h\right) f\right](s,a) \\
&\leq c\sqrt{\frac{H^4}{L}}.
\end{aligned}$$

□

D.4 Suboptimaity Gap

Lemma 20. *It holds that*

$$\sum_{h \in [H]} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} d_h^{\pi^*}(s,a) [\mathbb{V}_h V_{h+1}^*](s,a) \leq H^2.$$

Proof. The proof can be found in Lemma C.4 in Xie et al. (2021b). □

Lemma 21. *It holds that*

$$\sum_{l \in [L]} \left(\mathbb{V}_{h,l} V_{h+1}^*\right)(s,a) \leq L \left(\bar{\mathbb{V}}_h V_{h+1}^*\right)(s,a).$$

Proof. This lemma is a direct consequence of Lemma 35. □

Lemma 22. *With event $\mathcal{G}$ happening, it holds that*

$$\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{l\in[L]} d_h^{\pi^*}(s,a)\left(\mathbb{V}_{h,l}\hat{V}_{h+1}^{ref}\right)(s,a) \le LH^2 + c\sqrt{\frac{C^* H^9 SL}{K}} + c\sqrt{H^8 L}.$$

Proof. First, the left-hand side can be decomposed as

$$\begin{aligned}
&\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{l\in[L]} d_h^{\pi^*}(s,a)\left(\mathbb{V}_{h,l}\hat{V}_{h+1}^{\text{ref}}\right)(s,a)\\
&= \underbrace{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) L\left(\mathbb{V}_h V_{h+1}^*\right)(s,a)}_{:=\text{term (I)}}\\
&+ \underbrace{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) L\left(\left(\bar{\mathbb{V}}_h V_{h+1}^*\right)(s,a) - \left(\mathbb{V}_h V_{h+1}^*\right)(s,a)\right)}_{:=\text{term (II)}}\\
&+ \underbrace{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a)\left(\sum_{l\in[L]}\left(\mathbb{V}_{h,l}V_{h+1}^*\right)(s,a) - L\left(\bar{\mathbb{V}}_h V_{h+1}^*\right)(s,a)\right)}_{:=\text{term (III)}}\\
&+ \underbrace{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{l\in[L]} d_h^{\pi^*}(s,a)\left[\left(\mathbb{V}_{h,l}\hat{V}_{h+1}^{\text{ref}}\right)(s,a) - \left(\mathbb{V}_{h,l}V_{h+1}^*\right)(s,a)\right]}_{:=\text{term (IV)}}
\end{aligned}$$

Then, for term (I), with Lemma 20, it holds that

$$\text{term (I)} \le LH^2.$$

For term (II), with event (vii) in Lemma 14 it holds that

$$\begin{aligned}
\text{term (II)} &= \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) L\left(\left(\bar{\mathbb{V}}_h V_{h+1}^*\right)(s,a) - \left(\mathbb{V}_h V_{h+1}^*\right)(s,a)\right)\\
&\le c\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) L\sqrt{\frac{H^4}{L}}\\
&= c\sqrt{H^6 L}
\end{aligned}$$

For term (III), with Lemma 21, it holds that

$$\text{term (III)} \le 0.$$

For term (IV), we can obtain that

$$\begin{aligned}
\text{term (IV)} &= \sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{l\in[L]} d_h^{\pi^*}(s,a)\left[\left(\mathbb{V}_{h,l}\hat{V}_{h+1}^{\text{ref}}\right)(s,a) - \left(\mathbb{V}_{h,l}V_{h+1}^*\right)(s,a)\right]\\
&\overset{(a)}{\le} 4H\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{l\in[L]} d_h^{\pi^*}(s,a)\left\|\hat{V}_{h+1}^{\text{ref}}(\cdot) - V_{h+1}^*(\cdot)\right\|_\infty\\
&\overset{(b)}{\le} cH\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{l\in[L]} d_h^{\pi^*}(s,a)\cdot\left(\sqrt{\frac{C^* H^5 S}{LK}} + \sqrt{\frac{H^4}{L}}\right)\\
&= c\sqrt{\frac{C^* H^9 SL}{K}} + c\sqrt{H^8 L},
\end{aligned}$$

where inequality (a) is from simple algebraic based on the definition of variance and inequality (a) is from part (ii) of even $\mathcal{G}$ in Lemma 14. □

Proof of Theorem 2. By Lemma 5 and 14, with probability at least $1-\delta$, it holds that

$$\text{Gap}(\hat{\pi};\mathcal{M}) \le 2\sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\Gamma_h^{\alpha}(s_h,a_h)+\Gamma_h^{\beta}(s,a)\right].$$

For the first term, we have

$$\sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\Gamma_h^{\alpha}(s_h,a_h)\right] = \underbrace{\sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\Gamma_{h,0}^{\alpha}(s_h,a_h)\right]}_{:=\text{term (I)}} + \underbrace{\sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\Gamma_{h,1}^{\alpha}(s_h,a_h)\right]}_{:=\text{term (II)}}.$$

For term (I), we have

$$\begin{aligned}
\Gamma_{h,0}^{\alpha}(s_h,a_h) &= c\sqrt{\sum_{l\in[L]}\frac{\left(\hat{\mathbb{V}}_{h,0,l}\hat{V}_{h+1}^{\text{ref}}\right)(s,a)}{L^2N_{h,0,l}(s,a)}} + c\sqrt{\sum_{l\in[L]}\frac{H^2}{L^2(N_{h,0,l}(s,a))^2}} \\
&\overset{(a)}{\le} c\sqrt{\sum_{l\in[L]}\frac{\left(\mathbb{V}_{h,0,l}\hat{V}_{h+1}^{\text{ref}}\right)(s,a)}{L^2N_{h,0,l}(s,a)}} + c\sqrt{\sum_{l\in[L]}\frac{H^2}{L^2(N_{h,0,l}(s,a))^{3/2}}} + \sqrt{\sum_{l\in[L]}\frac{H^2}{L^2(N_{h,0,l}(s,a))^2}} \\
&\le c\sqrt{\sum_{l\in[L]}\frac{\left(\mathbb{V}_{h,0,l}\hat{V}_{h+1}^{\text{ref}}\right)(s,a)}{L^2N_{h,0,l}(s,a)}} + c\sqrt{\sum_{l\in[L]}\frac{1}{L^2N_{h,0,l}(s,a)} + \sum_{l\in[L]}\frac{H^4}{L^2(N_{h,0,l}(s,a))^2}} \\
&+ c\sqrt{\sum_{l\in[L]}\frac{H^2}{L^2(N_{h,0,l}(s,a))^2}} \\
&\le c\sqrt{\sum_{l\in[L]}\frac{\left(\mathbb{V}_{h,0,l}\hat{V}_{h+1}^{\text{ref}}\right)(s,a)}{L^2N_{h,0,l}(s,a)}} + \sqrt{\sum_{l\in[L]}\frac{1}{L^2N_{h,0,l}(s,a)}} + \sqrt{\sum_{l\in[L]}\frac{H^4}{L^2(N_{h,0,l}(s,a))^2}},
\end{aligned}$$

where inequality (a) is from event (vi) of Lemma 14. Then, for each separate term, it holds that

$$\begin{aligned}
\text{term (I.a)} &:= c\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a)\sqrt{\sum_{l\in[L]}\frac{\left(\mathbb{V}_{h,0,l}\hat{V}_{h+1}^{\text{ref}}\right)(s,a)}{L^2N_{h,0,l}(s,a)}} \\
&\overset{(a)}{\le} c\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a)\sqrt{\sum_{l\in[L]}\frac{\left(\mathbb{V}_{h,0,l}\hat{V}_{h+1}^{\text{ref}}\right)(s,a)}{L^2Kd_{h,l}^{\rho_l}(s,a)}} \\
&\le c\sqrt{\frac{C^*}{L^2K}}\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{d_h^{\pi^*}(s,a)\mathbb{1}\{a=\pi_h^*(s)\}\sum_{l\in[L]}\left(\mathbb{V}_{h,0,l}\hat{V}_{h+1}^{\text{ref}}\right)(s,a)} \\
&\le c\sqrt{\frac{C^*HS}{L^2K}}\sqrt{\sum_{h\in[H]}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{l\in[L]} d_h^{\pi^*}(s,a)\mathbb{1}\{a=\pi_h^*(s)\}\left(\mathbb{V}_{h,0,l}\hat{V}_{h+1}^{\text{ref}}\right)(s,a)} \\
&\overset{(b)}{\le} c\sqrt{\frac{C^*HS}{L^2K}}\sqrt{LH^2+\sqrt{\frac{C^*H^9SL}{K}}+\sqrt{H^8L}} \\
&\le c\sqrt{\frac{C^*HS}{L^2K}}\sqrt{LH^2+\frac{C^*H^3SL}{K}+H^6} \\
&\le c\sqrt{\frac{C^*H^3S}{LK}} + c\frac{C^*H^2S}{\sqrt{L}K} + c\sqrt{\frac{C^*H^7S}{L^2K}};
\end{aligned}$$

$$
\begin{aligned}
\text{term (I.b)} &:= c \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) \sqrt{\sum_{l\in[L]} \frac{1}{L^2 N_{h,0,l}(s,a)}} \\
&\overset{(c)}{\leq} c \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) \sqrt{\sum_{l\in[L]} \frac{1}{K L^2 d_{h,l}^{\rho_l}(s,a)}} \\
&\leq c\sqrt{\frac{C^*}{KL}} \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{d_h^{\pi^*}(s,a) \mathbb{1}\{a=\pi_h^*(s)\}} \\
&\leq c\sqrt{\frac{C^* H^2 S}{LK}};
\end{aligned}
$$

$$
\begin{aligned}
\text{term (I.c)} &:= c \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) \sqrt{\sum_{l\in[L]} \frac{H^4}{L^2 (N_{h,0,l}(s,a))^2}} \\
&\overset{(d)}{\leq} c \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) \sqrt{\sum_{l\in[L]} \frac{H^4}{L^2 (K d_{h,l}^{\rho_l}(s,a))^2}} \\
&\leq c \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathbb{1}\{\pi_h^*(s)=a\} \sqrt{\sum_{l\in[L]} \frac{(C^*)^2 H^4}{L^2 K^2}} \\
&\leq c \frac{C^* S H^3}{\sqrt{L} K},
\end{aligned}
$$

where inequalities (a), (c) and (d) are from event (iv) of Lemma 14 and inequality (b) is from Lemma 22.

By aggregating these three terms together, it can be observed that

$$
\text{term (I)} := \tilde{O}\left(\sqrt{\frac{C^* H^3 S}{LK}} + \frac{C^* H^3 S}{\sqrt{L} K} + \sqrt{\frac{C^* H^7 S}{L^2 K}}\right).
$$

Now, for term (II), we have

$$
\Gamma_{h,1}^{\alpha}(s,a) = c\sqrt{\sum_{l\in[L]} \frac{\left[\hat{\mathbb{V}}_{h,1,l}\left(\hat{V}_{h+1} - \hat{V}_{h+1}^{\text{ref}}\right)\right](s,a)}{L^2 N_{h,1,l}(s,a)}} + c\sqrt{\sum_{l\in[L]} \frac{H^2}{L^2 (N_{h,1,l}(s,a))^2}}.
$$

Then, for each separate term, it holds that

$$
\begin{aligned}
\text{term (II.a)} &:= c \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) \sqrt{\sum_{l\in[L]} \frac{\left[\hat{\mathbb{V}}_{h,l}\left(\hat{V}_{h+1} - \hat{V}_{h+1}^{\text{ref}}\right)\right](s,a)}{L^2 N_{h,1,l}(s,a)}} \\
&\overset{(a)}{\leq} c \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) \sqrt{\sum_{l\in[L]} \frac{H\left[\hat{\mathbb{V}}_{h,l}\left(\hat{V}_{h+1} - \hat{V}_{h+1}^{\text{ref}}\right)\right](s,a)}{L^2 K d_{h,l}^{\rho_l}(s,a)}} \\
&\overset{(b)}{\leq} c \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) \sqrt{\sum_{l\in[L]} \frac{H\left(\sqrt{\frac{C^* H^5 S}{KL}} + \sqrt{\frac{H^4}{L}}\right)^2}{L^2 K d_{h,l}^{\rho_l}(s,a)}} \\
&\leq c \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sqrt{d_h^{\pi^*}(s,a) \frac{C^* H\left(\frac{C^* H^5 S}{KL} + \frac{H^4}{L}\right)}{LK}}
\end{aligned}
$$

$$\leq c\frac{C^* H^4 S}{LK} + c\sqrt{\frac{C^* H^7 S}{L^2 K}};$$

$$\text{term (II.b)} := c \sum_{h\in[H]} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\pi^*}(s,a) \sqrt{\sum_{l\in[L]} \frac{H^2}{L^2 (N_{h,1,l}(s,a))^2}} \leq c\frac{C^* S H^3}{\sqrt{L}K},$$

where inequality (a) is from event (v) and inequality (b) is from events (ii) and (iii) of Lemma 14, respectively.

By aggregating these two terms together, it can be observed that

$$\text{term (II)} = \tilde{O}\left(\sqrt{\frac{C^* H^7 S}{L^2 K}} + \frac{C^* H^4 S}{LK} + \frac{C^* S H^3}{\sqrt{L}K}\right).$$

Also, with

$$\sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\Gamma_h^\beta(s,a)\right] \leq c\sqrt{\frac{H^4}{L}},$$

it can be obtained that

$$\text{Gap}(\hat{\pi};\mathcal{M}) = \tilde{O}\left(\sqrt{\frac{C^* H^3 S}{LK}} + \frac{C^* H^3 S}{\sqrt{L}K} + \sqrt{\frac{C^* H^7 S}{L^2 K}} + \sqrt{\frac{H^4}{L}}\right).$$

□

Algorithm 3 HetPEVI-Lin

Input: Dataset $\mathcal{D} = \{D_l : l \in [L]\}$; $\hat{V}_{H+1}(s) = 0, \forall s \in \mathcal{S}$

1: Randomly split each dataset $\mathcal{D}_l$ into sub-datasets $\{\mathcal{D}_{h,l}\}_{h\in[H]}$ with $|\mathcal{D}_{h,l}| = K/H$

2: Estimate $\hat{w}_{h,l} \leftarrow \Lambda_{h,l}^{-1}\left[\sum_{\tau\in\mathcal{D}_{h,l}} \phi(s_{h,l}^\tau, a_{h,l}^\tau)\left(r_{h,l}^\tau + \hat{V}_{h+1}(s_{h+1,l}^\tau)\right)\right], \forall (h,l) \in [H]\times[L]$

3: Set $\hat{w}_h \leftarrow \sum_{l\in[L]} \hat{w}_{h,l}/L, \forall h \in [H]$

4: **for** $h = H, H-1, \cdots, 1$ **do**

5: Perform updates for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ as in Eqn. (3) with the empirical Bellman operator $\hat{\mathbb{B}}_h$ as $(\hat{\mathbb{B}}_h \hat{V}_{h+1})(s,a) = \hat{w}_h^\top \phi(s,a)$ and the penalty $\Gamma_h(s,a) = \Gamma_h^\alpha(s,a) + \Gamma_h^\beta(s,a)$ in Eqns. (14) and(15)

6: **end for**

Output: policy $\hat{\pi} = \{\hat{\pi}_h(s) : (s,h) \in \mathcal{S}\times[H]\}$, values $\{\hat{V}_h(s) : (s,h) \in \mathcal{S}\times[H]\}$

E Extension to Offline Linear RL

The complete algorithm of HetPEVI-Lin is provided in Algorithm 3.

E.1 Properties of Linear MDPs

Lemma 23 (Restatement of Lemma 2)**.** *Under Assumption 2 and Definition 3, with any function $f : \mathcal{S} \to \mathbb{R}$, there exists $\{w_{h,l}^f \in \mathbb{R}^d : h \in [H], l \in [L]\}$ and $\{w_h^f \in \mathbb{R}^d : h \in [H]\}$ s.t. for any $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, it holds that $(\mathbb{B}_{h,l} f)(s,a) = \langle \phi(s,a), w_{h,l}^f\rangle$ for each $l \in [L]$ and $(\mathbb{B}_h f)(s,a) = \langle \phi(s,a), w_h^f\rangle$.*

Proof. This proof is standard for studies in linear MDPs (Jin et al., 2020; 2021b). We include it here for completeness and to facilitate the analysis of the next lemma. Based on the Bellman equation, it holds that

$$(\mathbb{B}_{h,l} f)(s,a) = r_{h,l}(s,a) + (\mathbb{P}_{h,l} f)(s,a) = \langle \phi(s,a), \theta_{h,l}\rangle + \int_{\mathcal{S}} f(s') \cdot \langle \phi(s,a), \mathrm{d}\mu_h(s')\rangle,$$

where means that $(\mathbb{B}_{h,l}f)(s,a) = \langle \phi(s,a), w_{h,l}^f \rangle$ with $w_{h,l}^f = \theta_{h,l} + \int_{\mathcal{S}} f(s')\,\mathrm{d}\mu_{h,l}(s')$. Similar arguments hold for $(\mathbb{B}_h f)(s,a) = \langle \phi(s,a), w_h^f \rangle$ with $w_h^f = \theta_h + \int_{\mathcal{S}} f(s')\,\mathrm{d}\mu_h(s')$. □

Lemma 24 (Restatement of Lemma 3)**.** *The average MDP $\bar{\mathcal{M}} = \sum_{l\in[L]} \mathcal{M}_l/L$ is a linear MDP of dimension d with feature mapping ϕ. Also, for any function $f : \mathcal{S} \to \mathbb{R}$ and any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, it holds that $(\bar{\mathbb{B}}_h f)(s,a) = \langle \phi(s,a), \bar{w}_h^f \rangle$ with weights $\{\bar{w}_h^f = \sum_{l\in[L]} w_{h,l}^f/L : h \in [H]\}$.*

Proof. First, it can be observed that the transition and reward of the sample average MDP $\bar{\mathcal{M}} = \sum_{l\in[L]} \mathcal{M}_l/L$ is linear since

$$\bar{\mathbb{P}}_h(s'|s,a) = \sum_{l\in[L]} \mathbb{P}_{h,l}(s'|s,a)/L = \sum_{l\in[L]} \langle \phi(s,a), \mu_{h,l}(s') \rangle/L = \langle \phi(s,a), \bar{\mu}_h(s') \rangle;$$

$$\bar{r}_h(s,a) = \sum_{l\in[L]} r_{h,l}(s,a)/L = \sum_{l\in[L]} \langle \phi(s,a), \theta_{h,l} \rangle/L = \langle \phi(s,a), \bar{\theta}_h \rangle,$$

where

$$\bar{\mu}_h(s') := \sum_{l\in[L]} \mu_{h,l}(s')/L; \qquad \bar{\theta}_h := \sum_{l\in[L]} \theta_{h,l}/L.$$

Then, we can verify the constraints in Definition 2 as

$$\|\phi(s,a)\|_2 \le 1, \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

and

$$\begin{aligned}
\max\{\|\bar{\mu}_h(\mathcal{S})\|_2, \|\bar{\theta}_h\|_2\} &= \max\left\{ \left\| \sum_{l\in[L]} \mu_{h,l}(\mathcal{S})/L \right\|_2, \left\| \sum_{l\in[L]} \theta_{h,l}/L \right\|_2 \right\} \\
&\le \max\left\{ \sum_{l\in[L]} \|\mu_{h,l}(\mathcal{S})\|_2/L, \sum_{l\in[L]} \|\theta_{h,l}\|_2/L \right\} \\
&\le \sqrt{d}, \qquad \forall h \in [H].
\end{aligned}$$

Moreover, it holds that

$$\begin{aligned}
(\bar{\mathbb{B}}_h f)(s,a) &= \bar{r}_h(s,a) + (\bar{\mathbb{P}}_h f)(s,a) \\
&= \langle \phi(s,a), \bar{\theta}_h \rangle + \int_{\mathcal{S}} f(s') \cdot \langle \phi(s,a), \mathrm{d}\bar{\mu}_h(s') \rangle \\
&= \left\langle \phi(s,a), \sum_{l\in[L]} \theta_{h,l}/L \right\rangle + \int_{\mathcal{S}} f(s') \cdot \left\langle \phi(s,a), \sum_{l\in[L]} \mathrm{d}\mu_{h,l}(s')/L \right\rangle \\
&= \left\langle \phi(s,a), \sum_{l\in[L]} \left(\theta_{h,l} + \int_{\mathcal{S}} f(s')\,\mathrm{d}\mu_{h,l}(s') \right)/L \right\rangle \\
&= \left\langle \phi(s,a), \sum_{l\in[L]} w_{h,l}^f/L \right\rangle \\
&= \left\langle \phi(s,a), \bar{w}_h^f \right\rangle,
\end{aligned}$$

which proves the claim. □

E.2 Validness of Pessimism

In the following, for all $h \in [H]$, based on the assumption that the matrix $\sum_{l\in[L]} \Lambda_{h,l}^{-1}/L^2$ is invertible, we denote that

$$\Upsilon_h := L^2 \left(\sum_{l\in[L]} \Lambda_{h,l}^{-1} \right)^{-1}; \qquad \Upsilon_h^{-1} := \frac{1}{L^2} \sum_{l\in[L]} \Lambda_{h,l}^{-1}.$$

Lemma 25. *With the penalties* $\Gamma_h^\alpha(s,a) = cH\left(\sqrt{d}+\sqrt{dL\lambda}\right)\|\phi(s,a)\|_{\Upsilon_h^{-1}}$ *in HetPEVI-Lin, with probability at least* $1-\delta$*, it holds that for all* $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$

$$\left|\left(\hat{\mathbb{B}}_h\hat{V}_{h+1}\right)(s,a)-\left(\bar{\mathbb{B}}_h\hat{V}_{h+1}\right)(s,a)\right|\le\Gamma_h^\alpha(s,a).$$

Proof. For a fixed h and a fixed function $\hat{V}_{h+1}(\cdot):\mathcal{S}\to\mathbb{R}$, with Lemma 3, there exists $w_h\in\mathbb{R}^d$ such that

$$\left(\bar{\mathbb{B}}_h\hat{V}_{h+1}\right)(s,a)=\langle\phi(s,a),\bar{w}_h\rangle,\qquad\forall(s,a)\in\mathcal{S}\times\mathcal{A}.$$

With $(\hat{\mathbb{B}}_h\hat{V}_{h+1})(s,a):=\langle\phi(s,a),\hat{w}_h\rangle$, it holds that

$$\left|\left(\bar{\mathbb{B}}_h\hat{V}_{h+1}\right)(s,a)-\left(\hat{\mathbb{B}}_h\hat{V}_{h+1}\right)(s,a)\right|=|\langle\phi(s,a),\bar{w}_h-\hat{w}_h\rangle|\le\|\phi(s,a)\|_{\Upsilon_h^{-1}}\|\bar{w}_h-\hat{w}_h\|_{\Upsilon_h},$$

and thus we can instead control $\|\bar{w}_h-\hat{w}_h\|_{\Upsilon_h}$. For this purpose, we can observe that

$$\|\bar{w}_h-\hat{w}_h\|_{\Upsilon_h}=\langle\bar{w}_h-\hat{w}_h,(\Upsilon_h)^{1/2}X\rangle,$$

where

$$X:=\frac{(\Upsilon_h)^{1/2}(\bar{w}_h-\hat{w}_h)}{\|\bar{w}_h-\hat{w}_h\|_{\Upsilon_h}}\in\mathbb{S}^{d-1}.$$

With Lemma 36, we can find a ε-covering $\mathcal{C}_\varepsilon$ over $\mathbb{S}^{d-1}$ with $|\mathcal{C}_\varepsilon|\le(3/\varepsilon)^d$. Using Lemma 26 and a union bound, we can have that with probability $1-\delta$, $\forall(y,h)\in\mathcal{C}_\varepsilon\times[H]$, it holds that

$$\begin{aligned}\left|\langle(\Upsilon_h)^{1/2}y,\bar{w}_h-\hat{w}_h\rangle\right|&\le cH\left(\sqrt{\log(H|\mathcal{C}_\varepsilon|/\delta)}+\sqrt{dL\lambda}\right)\left\|(\Upsilon_h)^{1/2}y\right\|_{(\Upsilon_h)^{-1}}\\&\le cH\left(\sqrt{d\log(H/(\varepsilon\delta))}+\sqrt{dL\lambda}\right).\end{aligned}$$

With this event happening, for any $x\in\mathbb{S}^{d-1}$, there exists $y\in\mathcal{C}_\varepsilon$ such that $\|x-y\|_2\le\varepsilon$ and further it holds that

$$\begin{aligned}\|\bar{w}_h-\hat{w}_h\|_{\Upsilon_h}&=\max_{x\in\mathbb{S}^{d-1}}\langle\bar{w}_h-\hat{w}_h,(\Upsilon_h)^{1/2}x\rangle\\&=\max_{x\in\mathbb{S}^{d-1}}\min_{y\in\mathcal{C}_\varepsilon}\left[\langle\bar{w}_h-\hat{w}_h,(\Upsilon_h)^{1/2}y\rangle+\langle\bar{w}_h-\hat{w}_h,(\Upsilon_h)^{1/2}(x-y)\rangle\right]\\&\le\max_{x\in\mathbb{S}^{d-1}}\min_{y\in\mathcal{C}_\varepsilon}\left[cH\left(\sqrt{d\log(H/(\varepsilon\delta))}+\sqrt{dL\lambda}\right)+\|\bar{w}_h-\hat{w}_h\|_{\Upsilon_h}\|x-y\|_2\right]\\&\le cH\left(\sqrt{d\log(H/(\varepsilon\delta))}+\sqrt{dL\lambda}\right)+\varepsilon\|\bar{w}_h-\hat{w}_h\|_{\Upsilon_h}.\end{aligned}$$

Thus, with $\varepsilon=1/2$, with probability at least $1-\delta$, $\forall h\in[H]$, it holds that

$$\|w_h-\hat{w}_h\|_{\Upsilon_h}\le cH\left(\sqrt{2d}+\sqrt{dL\lambda}\right).$$

Thus, for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, it holds that

$$\begin{aligned}\left|\left(\mathbb{B}_h\hat{V}_{h+1}\right)(s,a)-\left(\hat{\mathbb{B}}_h\hat{V}_{h+1}\right)(s,a)\right|&\le\|\phi(s,a)\|_{\Upsilon_h^{-1}}\|w_h-\hat{w}_h\|_{\Upsilon_h}\\&\le cH\left(\sqrt{d}+\sqrt{dL\lambda}\right)\|\phi(s,a)\|_{\Upsilon_h^{-1}}.\end{aligned}$$

With a union bound over $h\in[H]$, the lemma is proved. □

Lemma 26. *For a fixed vector* $x\in\mathbb{R}^d$ *and a fixed* $h\in[H]$*, with probability at least* $1-\delta$*, it holds that*

$$|\langle x,\bar{w}_h-\hat{w}_h\rangle|\le H\left(\sqrt{2\log(2/\delta)}+\sqrt{dL\lambda}\right)\|x\|_{\Upsilon_h^{-1}}.$$

Proof. It holds that

$$
\begin{aligned}
x^\top(\bar{w}_h - \hat{w}_h) &= x^\top\left(\sum_{l\in[L]} w_{h,l}/L - \sum_{l\in[L]} \hat{w}_{h,l}/L\right) \\
&\overset{(a)}{=} x^\top \sum_{l\in[L]} \frac{1}{L}\Lambda_{h,l}^{-1}\left(\sum_{\tau\in\mathcal{D}_{h,l}} \phi(s_{h,l}^\tau, a_{h,l}^\tau)\phi(s_{h,l}^\tau, a_{h,l}^\tau)^\top + \lambda I\right) w_{h,l} \\
&- x^\top \sum_{l\in[L]} \frac{1}{L}\Lambda_{h,l}^{-1}\left(\sum_{\tau\in\mathcal{D}_{h,l}} \phi(s_{h,l}^\tau, a_{h,l}^\tau)\left(r_{h,l}^\tau + \hat{V}_{h+1}(s_{h+1,l}^\tau)\right)\right) \\
&= \underbrace{\lambda x^\top \sum_{l\in[L]} \frac{1}{L}\Lambda_{h,l}^{-1} w_{h,l}}_{\text{term (I)}} \\
&\underbrace{+ x^\top \sum_{l\in[L]} \frac{1}{L}\Lambda_{h,l}^{-1}\left(\sum_{\tau\in\mathcal{D}_{h,l}} \phi(s_{h,l}^\tau, a_{h,l}^\tau)\left(\phi(s_{h,l}^\tau, a_{h,l}^\tau)^\top w_{h,l} - \left(r_{h,l}^\tau + \hat{V}_{h+1}(s_{h+1,l}^\tau)\right)\right)\right)}_{\text{term (II)}},
\end{aligned}
$$

where the definition of $\Lambda_{h,l}$ and $\hat{w}_{h,l}$ is used in equation (a).

For term (I), we have

$$
\begin{aligned}
\lambda\left|x^\top \sum_{l\in[L]} \frac{1}{L}\Lambda_{h,l}^{-1} w_{h,l}\right| &= \lambda\left|\sum_{l\in[L]} x^\top \Lambda_{h,l}^{-1} w_{h,l}/L\right| \\
&= \frac{\lambda}{L}\left|\left[x^\top\Lambda_{h,1}^{-1/2}, \cdots, x^\top\Lambda_{h,L}^{-1/2}\right]\left[w_{h,1}^\top\Lambda_{h,1}^{-1/2}, \cdots, w_{h,L}^\top\Lambda_{h,L}^{-1/2}\right]^\top\right| \\
&\le \frac{\lambda}{L}\sqrt{\sum_{l\in[L]} x^\top\Lambda_{h,l}^{-1}x}\sqrt{\sum_{l\in[L]} w_{h,l}^\top\Lambda_{h,l}^{-1}w_{h,l}} \\
&\overset{(a)}{\le} \frac{\lambda}{L}\sqrt{\sum_{l\in[L]} x^\top\Lambda_{h,l}^{-1}x}\sqrt{\sum_{l\in[L]} 4H^2 d/\lambda} \\
&= c\sqrt{dH^2L\lambda}\sqrt{x^\top\Upsilon_h^{-1}x},
\end{aligned}
$$

where inequality (a) is based on the fact that $\|w_{h,l}\|_2 \le 2H\sqrt{d}, \forall(l,h) \in [L]\times[H]$ (which can be proved similarly to Lemma B.1 in Jin et al. (2020)) and the following observation:

$$
w_{h,l}^\top\Lambda_{h,l}^{-1}w_{h,l} \le \|w_h\|_2^2\|\Lambda_{h,l}^{-1}\|_2 \le 4dH^2/\lambda.
$$

For term (II), with

$$
t := H\sqrt{2x^\top\Upsilon_h^{-1}x\log(2/\delta)},
$$

conditioned on the state-action pairs at step h, we have

$$
\begin{aligned}
&\mathbf{P}\left(\left|x^\top \sum_{l\in[L]} \frac{1}{L}\Lambda_{h,l}^{-1}\left(\sum_{\tau\in\mathcal{D}_{h,l}} \phi(s_{h,l}^\tau, a_{h,l}^\tau)\left(\phi(s_{h,l}^\tau, a_{h,l}^\tau)^\top w_{h,l} - \left(r_{h,l}^\tau + \hat{V}_{h+1}(s_{h+1,l}^\tau)\right)\right)\right)\right| \ge t\right) \\
&\le 2\exp\left(-\frac{2t^2}{4H^2\sum_{l\in[L]}\sum_{\tau\in\mathcal{D}_{h,l}}\frac{1}{L^2}\left(x^\top\Lambda_{h,l}^{-1}\phi(s_{h,l}^\tau, a_{h,l}^\tau)\right)^2}\right) \\
&\le 2\exp\left(-\frac{x^\top\Upsilon_h^{-1}x\log(2/\delta)}{\sum_{l\in[L]}\sum_{\tau\in\mathcal{D}_{h,l}}\frac{1}{L^2}x^\top\Lambda_{h,l}^{-1}\phi(s_{h,l}^\tau, a_{h,l}^\tau)\phi(s_{h,l}^\tau, a_{h,l}^\tau)^\top\Lambda_{h,l}^{-1}x}\right)
\end{aligned}
$$

$$\leq 2\exp\left(-\frac{x^\top \Upsilon_h^{-1} x \log(2/\delta)}{\sum_{l\in[L]}\sum_{\tau\in\mathcal{D}_{h,l}} \frac{1}{L^2} x^\top \Lambda_{h,l}^{-1}\left(\phi(s_{h,l}^\tau, a_{h,l}^\tau)\phi(s_{h,l}^\tau, a_{h,l}^\tau)^\top + \lambda I\right)\Lambda_{h,l}^{-1} x}\right)$$
$$\leq 2\exp\left(-\frac{L^2 x^\top \Upsilon_h^{-1} x \log(2/\delta)}{x^\top \Upsilon_h^{-1} x}\right)$$
$$= \delta,$$

which leads to the claim. □

Lemma 27. *For the penalties $\Gamma_h^\beta(s,a) = c\sqrt{\frac{dH^2}{L}}$ in HetPEVI, with probability at least $1-\delta$, it holds that for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$,*

$$\left|\left(\bar{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a)\right| \leq \Gamma_h^\beta(s,a).$$

Proof. For a fixed $\phi(s,a) \in \mathbb{R}^d$ and a fixed $\hat{V}_{h+1}$, since $(\mathbb{B}_{h,l}\hat{V}_{h+1})(s,a)$ is bounded between $[0,H]$ and has an expectation of $(\mathbb{B}_h\hat{V}_{h+1})(s,a)$, it is then a H-sub-Gaussian random variable (Vershynin, 2018). Thus, it holds that

$$\mathbf{P}\left(\left|\left(\bar{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a)\right| \geq c\sqrt{\frac{H^2\log(2H/\delta)}{L}}\right) \leq \delta, \qquad \forall h \in [H].$$

Similarly as in Lemma 25, using a covering argument, it holds that

$$\mathbf{P}\left(\left|\left(\bar{\mathbb{B}}_h \hat{V}_{h+1}\right)(s,a) - \left(\mathbb{B}_h \hat{V}_{h+1}\right)(s,a)\right| \geq c\sqrt{\frac{dH^2}{L}}\right) \leq \delta, \qquad \forall (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H].$$

□

Lemma 28. *The penalties $\{\Gamma_h(s,a) = \Gamma_h^\alpha(s,a) + \Gamma_h^\beta(s,a) : (s,a,h) : (s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]\}$ in HetPEVI induce a valid pessimism with probability at least $1-\delta$ with respect to the estimated Bellman operator $\hat{\mathbb{B}}_h$ as $(\hat{\mathbb{B}}_h\hat{V}_{h+1})(s,a) := \langle\phi(s,a), \hat{w}_h\rangle$.*

Proof. This result can be obtained by combining Lemmas 25 and 27. □

E.3 Good Event

Lemma 29. *The following event holds with probability at least $1-\delta$ for HetPEVI-Lin*

$$\mathcal{G} := \Bigg\{\text{(i) The penalties } \{\Gamma_h(s,a)\} \text{ induce a valid pessimism;}$$
$$\text{(ii) } \left\|\Xi_{h,l} - \hat{\Xi}_{h,l}\right\|_2 \leq c\sqrt{\frac{H}{L}}, \forall (l,h) \in [L]\times[H]\Bigg\},$$

where

$$\Xi_{h,l} := \mathbb{E}_{\rho_l,\mathcal{M}_l}\left[\phi(s_h,a_h)\phi(s_h,a_h)^\top\right]$$
$$\hat{\Xi}_{h,l} := \frac{H}{K}\sum_{\tau\in\mathcal{D}_{h,l}} \phi(s_{h,l}^\tau, a_{h,l}^\tau)\phi(s_{h,l}^\tau, a_{h,l}^\tau)^\top.$$

Proof. The part (i) holds according to Lemma 28 and the part (ii) can be observed via the Lemma 37. □

E.4 SUBOPTIMALITY GAP

Proof of Theorem 3. By Lemma 5 and 29, with probability at least $1-\delta$, it holds that

$$\text{Gap}(\hat{\pi};\mathcal{M}) \leq 2\sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\Gamma_h^{\alpha}(s_h,a_h)+\Gamma_h^{\beta}(s_h,a_h)\right].$$

For the first term, it holds that

$$\begin{aligned}\Gamma_h^{\alpha}(s_h,a_h) &= cH\left(\sqrt{d}+\sqrt{dL\lambda}\right)\|\phi(s_h,a_h)\|_{\Upsilon_h^{-1}}\\ &\overset{(a)}{=} c\sum_{h\in[H]}\mathbb{E}_{\pi^*,\mathcal{M}}\left[\sqrt{dH^2}\|\phi(s_h,a_h)\|_{\Upsilon_h^{-1}}\right]\end{aligned}$$

where λ is set to be $1/L$ in equality (a). Furthermore, we denote

$$\Sigma_{h,l} = \mathbb{E}_{\pi^*,\mathcal{M}}\left[\phi(s_h,a_h)\phi(s_h,a_h)^{\top}\right],$$

which is positive semi-definite as $\Xi_{h,l}$ and $\hat{\Xi}_{h,l}$ for all $h\in[H]$. Correspondingly, we denote $\text{Rank}_{h,l}=\text{Rank}(\Sigma_{h,l})$ for all $(l,h)\in[L]\times[H]$. Also, for a positive semi-definite matrix $\Sigma\in\mathbb{R}^{d\times d}$, we denote its ordered eigenvalues as

$$\gamma_1(\Sigma)\geq\gamma_1(\Sigma)\cdots\geq\gamma_d(\Sigma)\geq 0.$$

With Assumption 3, when $K\geq c\max_{h,l}\left\{\frac{(D^*)^2H}{(\gamma_{\text{Rank}_{h,l}}(\Sigma_{h,l}))^2}\right\}$, it holds that

$$\begin{aligned}\sum_{h\in[H]}\mathbb{E}_{\pi^*,\mathcal{M}}\left[\Gamma_h^{\alpha}(s_h,a_h)\right] &\leq c\sum_{h\in[H]}\mathbb{E}_{\pi^*,\mathcal{M}}\left[\sqrt{dH^2}\|\phi(s_h,a_h)\|_{\Upsilon_h^{-1}}\right]\\
&= c\sqrt{\frac{dH^2}{L^2}}\sum_{h\in[H]}\mathbb{E}_{\pi^*,\mathcal{M}}\left[\sqrt{\sum_{l\in[L]}\text{Tr}\left(\phi(s_h,a_h)\phi(s_h,a_h)^{\top}\Lambda_{h,l}^{-1}\right)}\right]\\
&\overset{(a)}{\leq} c\sqrt{\frac{dH^2}{L^2}}\sum_{h\in[H]}\sqrt{\sum_{l\in[L]}\text{Tr}\left(\mathbb{E}_{\pi^*,\mathcal{M}}\left[\phi(s_h,a_h)\phi(s_h,a_h)^{\top}\right]\Lambda_{h,l}^{-1}\right)}\\
&\overset{(b)}{=} c\sqrt{\frac{dH^2}{L^2}}\sum_{h\in[H]}\sqrt{\sum_{l\in[L]}\text{Tr}\left(\Sigma_{h,l}\Lambda_{h,l}^{-1}\right)}\\
&\overset{(c)}{\leq} c\sqrt{\frac{dH^2}{L^2}}\sum_{h\in[H]}\sqrt{\sum_{l\in[L]}\sum_{i\in[d]}\frac{\gamma_i(\Sigma_{h,l})}{\gamma_i(\Lambda_{h,l})}}\\
&\overset{(d)}{\leq} c\sqrt{\frac{dH^2}{L^2}}\sum_{h\in[H]}\sqrt{\sum_{l\in[L]}\sum_{i\in[\text{Rank}_{h,l}]}\frac{\gamma_i(\Sigma_{h,l})}{\gamma_i(\Lambda_{h,l})}}\\
&= c\sqrt{\frac{dH^2}{L^2}}\sum_{h\in[H]}\sqrt{\sum_{l\in[L]}\sum_{i\in[\text{Rank}_{h,l}]}\frac{H\gamma_i(\Sigma_{h,l})/K}{H/(KL)+\gamma_i(\hat{\Xi}_{h,l})}}\\
&\overset{(e)}{\leq} c\sqrt{\frac{dH^2}{L^2}}\sum_{h\in[H]}\sqrt{\sum_{l\in[L]}\sum_{i\in[\text{Rank}_{h,l}]}\frac{H\gamma_i(\Sigma_{h,l})/K}{H/(KL)+\gamma_i(\Sigma_{h,l})/D^*}}\\
&\leq c\sqrt{\frac{dH^2}{L^2}}\sum_{h\in[H]}\sqrt{\sum_{l\in[L]}\sum_{i\in[\text{Rank}_{h,l}]}\frac{HD^*}{K}}\\
&\leq c\sqrt{\frac{d^2H^5D^*}{LK}}\end{aligned}$$

where inequality (a) is from the Cauchy-Schwarz inequality, inequality (b) is from the definition of $\Sigma_{h,l}$, inequality (c) is from the Von Neumann's trace inequality (Mirsky, 1975), and equality (d) removes the zero eigenvalues of $\Sigma_{h,l}$ from the sum. Furthermore, for $i \leq r_{h,l}$ (implying $\gamma_i(\Sigma_{h,l}) > 0$), inequality (e) is from the following observation

$$\begin{aligned}
\gamma_i(\hat{\Xi}_{h,l}) &\overset{(a)}{\geq} \gamma_i(\Xi_{h,l}) - \|\Xi_{h,l} - \hat{\Xi}_{h,l}\|_2 \\
&\overset{(b)}{\geq} \gamma_i(\Xi_{h,l}) - c\frac{\sqrt{H}}{\sqrt{K}} \\
&\overset{(c)}{\geq} \frac{\gamma_i(\Sigma_{h,l})}{D^*} - c\frac{\sqrt{H}}{\sqrt{K}} \\
&\overset{(d)}{\geq} c\frac{\gamma_i(\Sigma_{h,l})}{D^*},
\end{aligned}$$

where inequality (a) is from the Weyl's inequality (see e.g., Chapter 8 in (Wainwright, 2019)), inequality (b) is from (ii) of event $\mathcal{G}$ in Lemma 29, inequality (c) is from the Assumption 3 (which implies $\gamma_i(\Sigma_{h,l}) \leq D^*\gamma_i(\Xi_{h,l})$ from the Weyl's inequality), and inequality (d) is due to $\gamma_i(\Gamma_{h,l}) > 0, \forall i \in [r_{h,l}]$ and the sufficiently large K. □

F Supporting Lemmas

F.1 Suboptimality Decomposition

The suboptimality gap between an output policy $\hat{\pi}$ from an offline RL algorithm and the optimal policy π^* can be decomposed as follows.

Lemma 30 (Lemma 3.1 in Jin et al. (2021b))**.** *Let $\hat{\pi} = \{\hat{\pi}_h\}_{h\in[H]}$ be the policy such that $\hat{V}_h(s) = \langle \hat{Q}_h(s,\cdot), \hat{\pi}(\cdot|s)\rangle$. For any $\hat{\pi}$ and $s \in \mathcal{S}$ it holds that*

$$\begin{aligned}
\textit{Gap}(\hat{\pi};\mathcal{M}) = &-\sum_{h\in[H]} \mathbb{E}_{\hat{\pi},\mathcal{M}}\left[\zeta_h(s_h,a_h)|s_1 = s\right] + \sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\zeta_h(s_h,a_h)|s_1 = s\right] \\
&+ \sum_{h\in[H]} \mathbb{E}_{\pi^*,\mathcal{M}}\left[\left\langle \hat{Q}_h(s_h,\cdot), \pi^*(\cdot|s_h) - \hat{\pi}_h(\cdot|s_h)\right\rangle |s_1 = s\right],
\end{aligned}$$

where the expectations $\mathbb{E}_{\hat{\pi},\mathcal{M}}$ and $\mathbb{E}_{\pi^*,\mathcal{M}}$ are with respect to the trajectories induced by $\hat{\pi}$ and π^*, and

$$\zeta_h(s,a) := (\mathbb{B}_h\hat{V}_{h+1})(s,a) - \hat{Q}_h(s,a)$$

is the model evaluation error at $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$.

F.2 Enhanced Empirical Bernstein Inequality

In following, an enhanced version of empirical Bernstein inequality is derived. First, the famous Bernstein inequality is presented, which serves as a starting point of the later generalization.

Lemma 31 (Bernstein Inequality)**.** *Let $Z_1, Z_2, \cdots, Z_n$ be independent random variables with values $|Z_i - \mathbb{E}[Z_i]| \leq c_0$. Then, it holds that*

$$\mathbf{P}\left(\left|\sum_{i=1}^n Z_i - \sum_{i=1}^n \mathbb{E}\left[Z_i\right]\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \mathbb{V}(Z_i) + \frac{2}{3}c_0 t}\right),$$

where $\mathbb{V}(Z_i)$ denotes the variance of random variable Z_i.

It is noted that the true variance is required in the original form of Bernstein inequality, which is hard to be obtained in practice. Thus, Maurer & Pontil (2009) established an empirical version of Bernstein inequality, where the estimated variance is used instead of the true variance.

Lemma 32 (Empirical Bernstein Inequality; Theorem 4 in Maurer & Pontil (2009))**.** *Let $Z, Z_1, Z_2, \cdots, Z_n$ be i.i.d. random variables and let $\delta > 0$ with values $|Z_i - \mathbb{E}[Z]| \leq c_0$. Then, with probability at least $1-\delta$, it holds that*

$$\left|\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right| \leq \sqrt{\frac{2\hat{\mathbb{V}}(Z)\ln(4/\delta)}{n}} + \frac{7c_0\ln(4/\delta)}{3n},$$

where $\hat{\mathbb{V}}(Z)$ denotes the empirical variance of Z with samples $Z_1, \cdots, Z_n$

However, in Lemma 32, the samples $Z_1, \cdots, Z_n$ are i.i.d, i.e., from the same source, which do not meet the need of handling data from heterogeneous sources in this work. Thus, a further enhanced version is provided in the following.

Lemma 33 (Enhanced Empirical Bernstein Inequality)**.** *Let $Z_l, Z_{1,l}, Z_{2,l}, \cdots, Z_{n_l,l}$ be i.i.d. random variables with values $Z_{i,l} \in [0,1]$ respectively for each $l \in [L]$, and let $\delta > 0$. Then, denoting*

$$\eta := \frac{1}{L}\sum_{l\in[L]} \mathbb{E}[Z_l] - \frac{1}{L}\sum_{l\in[L]} \frac{1}{n_l}\sum_{i=1}^{n_l} Z_{i,l},$$

with probability at least $1-\delta$, it holds that

$$|\eta| \leq \sqrt{\sum_{l\in[L]} \frac{8\hat{\mathbb{V}}(Z_l)\log(\frac{2(L+1)}{\delta})}{L^2 n_l}} + \sqrt{\sum_{l\in[L]} \frac{36\left(\log(\frac{2(L+1)}{\delta})\right)^2}{L^2(n_l)^2}},$$

where $\hat{\mathbb{V}}(Z_l)$ denotes the empirical variance of Z_l with samples $Z_{1,l}, Z_{2,l}, \cdots, Z_{n_l,l}$

Proof. With $n_{\min} = \min_{l\in[L]}\{n_l\}$ and

$$t := \sqrt{\sum_{l\in[L]} \frac{4\mathbb{V}(Z_l)\log(\frac{2(L+1)}{\delta})}{L^2 n_l}} + \frac{4\log(\frac{2(L+1)}{\delta})}{3Ln_{\min}},$$

which implies that

$$\begin{aligned}
&\frac{t^2}{2\sum_{l\in[L]}\sum_{i=1}^{n_l}\mathbb{V}(Z_i^l)/(L^2 n_l^2) + 2t/(3Ln_{\min})} \\
\leq &\frac{t^2}{\max\{4\sum_{l\in[L]}\mathbb{V}(Z_l)/(L^2 n_l), 4t/(3Ln_{\min})\}} \\
\leq &\log\left(\frac{2(L+1)}{\delta}\right),
\end{aligned}$$

from Lemma 31, it holds that

$$\begin{aligned}
\mathbf{P}\left(|\eta| \geq t\right) &\leq 2\exp\left(-\frac{t^2}{2\sum_{l\in[L]}\sum_{i=1}^{n_l}\mathbb{V}(Z_i^l)/(L^2 n_l^2) + 2t/(3Ln_{\min})}\right) \\
&\leq \frac{\delta}{L+1}.
\end{aligned}$$

Furthermore, with Lemma 34 and a union bound, with probability at least $1 - \frac{L\delta}{L+1}$, it holds that

$$\sqrt{\mathbb{V}(Z_l)} \leq \sqrt{\hat{\mathbb{V}}(Z_l)} + \sqrt{\frac{2\ln(\frac{L+1}{\delta})}{n_l}}, \qquad \forall l \in [L].$$

If the above event happens, it holds that

$$t = \sqrt{\sum_{l\in[L]} \frac{4\mathbb{V}(Z_l)\log(\frac{2(L+1)}{\delta})}{L^2 n_l}} + \frac{4\log(\frac{2(L+1)}{\delta})}{3Ln_{\min}}$$

$$
\begin{aligned}
&\leq \sqrt{\sum_{l\in[L]} \frac{4\left(\sqrt{\hat{\mathbb{V}}(Z_l)} + \sqrt{\frac{2\log(\frac{L+1}{\delta})}{L^2 n_l}}\right)^2 \log(\frac{2(L+1)}{\delta})}{n_l}} + \frac{4\log(\frac{2(L+1)}{\delta}}{3Ln_{\min}} \\
&\leq \sqrt{\sum_{l\in[L]} \frac{8\hat{\mathbb{V}}(Z_l)\log(\frac{2(L+1)}{\delta})}{L^2 n_l}} + \sqrt{\sum_{l\in[L]} \frac{16\left(\log(\frac{2(L+1)}{\delta})\right)^2}{L^2 (n_l)^2}} + \frac{4\log(\frac{2(L+1)}{\delta})}{3Ln_{\min}} \\
&\leq \sqrt{\sum_{l\in[L]} \frac{8\hat{\mathbb{V}}(Z_l)\log(\frac{2(L+1)}{\delta})}{L^2 n_l}} + \sqrt{\sum_{l\in[L]} \frac{32\left(\log(\frac{2(L+1)}{\delta})\right)^2}{L^2 (n_l)^2} + \frac{32\left(\log(\frac{2(L+1)}{\delta})\right)^2}{9L^2 n_{\min}^2}} \\
&\leq \sqrt{\sum_{l\in[L]} \frac{8\hat{\mathbb{V}}(Z_l)\log(\frac{2(L+1)}{\delta})}{L^2 n_l}} + \sqrt{\sum_{l\in[L]} \frac{36\left(\log(\frac{2(L+1)}{\delta})\right)^2}{L^2 (n_l)^2}}.
\end{aligned}
$$

Combining this the bound of t with the concentration inequality proved above, the lemma is proved. □

Lemma 34 (Theorem 10 in (Maurer & Pontil, 2009))**.** *Let $Z, Z_1, \cdots, Z_n$ be i.i.d. random variables with values in $[0,1]$, and let $\delta > 0$, with probability at least $1-\delta$, it holds that*

$$
\sqrt{\mathbb{V}(Z)} < \sqrt{\hat{\mathbb{V}}(Z)} + \sqrt{\frac{2\ln(1/\delta)}{n}}.
$$

F.3 Variance of Mixture Models

Lemma 35. *With random variable X_l following distribution $\mathbb{P}_l$ for each $l \in [L]$, the random variable X is assumed to follow the distribution mixture, especially, $\mathbb{P} = \sum_{l\in[L]} \alpha_l \mathbb{P}_l$, where $\alpha = [\alpha_1, \cdots, \alpha_L] \in \Delta_L$, the followings holds*

$$
C\mathbb{V}(X) \geq \sum_{l\in[L]} \alpha_l^2 \mathbb{V}(X_l),
$$

if

$$
C \geq \frac{\sum_{l\in[L]} \alpha_l^2 \mathbb{V}(X_l)}{\sum_{l\in[L]} \alpha_l \mathbb{V}(X_l)}.
$$

Proof. For each side of the desired inequality, we have

$$
\begin{aligned}
\text{LHS} &:= C\mathbb{V}(X) = C\left(\sum_{l\in[L]} \alpha_l \mathbb{E}[X_l^2] - \left(\sum_{l\in[L]} \alpha_l \mathbb{E}[X_l]\right)^2\right) \\
\text{RHS} &:= \sum_{l\in[L]} \alpha_l^2 \mathbb{V}(X_l) = \sum_{l\in[L]} \alpha_l^2 \mathbb{E}[X_l^2] - \sum_{l\in[L]} \alpha_l^2 [\mathbb{E}X_l]^2.
\end{aligned}
$$

With the above expressions, we can further obtain that

$$
\begin{aligned}
\text{LHS} - \text{RHS} &= \sum_{l\in[L]} \alpha_l (C-\alpha_l)\mathbb{E}[X_l^2] - C\left(\sum_{l\in[L]} \alpha_l \mathbb{E}[X_l]\right)^2 + \sum_{l\in[L]} \alpha_l^2 [\mathbb{E}X_l]^2 \\
&= \sum_{l\in[L]} \alpha_l (C-\alpha_l)\mathbb{E}[X_l^2] + \sum_{l\in[L]} \alpha_l^2 (1-C)[\mathbb{E}X_l]^2 - C\sum_{l\neq n} \alpha_l \alpha_n \mathbb{E}[X_l]\mathbb{E}[X_n] \\
&\geq \sum_{l\in[L]} \alpha_l (C-\alpha_l)\mathbb{E}[X_l^2] + \sum_{l\in[L]} \alpha_l^2 (1-C)[\mathbb{E}X_l]^2 - C\sum_{l\neq n} \alpha_l \alpha_n \frac{[\mathbb{E}X_l]^2 + [\mathbb{E}X_n]^2}{2}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{l\in[L]} \alpha_l(C-\alpha_l)\mathbb{E}[X_l^2] + \sum_{l\in[L]} \alpha_l^2(1-C)[\mathbb{E}X_l]^2 - C\sum_{l\in[L]} \alpha_l(1-\alpha_l)[\mathbb{E}X_l]^2 \\
&= \sum_{l\in[L]} \alpha_l(C-\alpha_l)\mathbb{E}[X_l^2] + \sum_{l\in[L]} \alpha_l(\alpha_l-C)[\mathbb{E}X_l]^2 \\
&= \sum_{l\in[L]} \alpha_l(C-\alpha_l)\mathbb{V}(X_l),
\end{aligned}
$$

which leads to the lemma. □

F.4 MISCELLANEOUS

Lemma 36 (Covering Number of a Euclidean Ball; Lemma D.5 of Jin et al. (2020))**.** *There exists a set $\mathcal{C}_\varepsilon \in \mathbb{R}^d$ with $|\mathcal{C}_\varepsilon| \leq (1+2R/\varepsilon)^d$ such that for all $x \in \mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$ there exists a $y \in \mathcal{C}_\varepsilon$ with $\|x-y\|_2 \leq \varepsilon$.*

Lemma 37 (Lemma H.4 of Min et al. (2021))**.** *Let $\psi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ satisfying $\|\psi(s,a)\|_2 \leq C$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. For any $T > 0$ and $\kappa > 0$, define $\bar{\mathbf{G}}_T = \sum_{\tau\in[T]} \psi(s^\tau,a^\tau)\psi(s^\tau,a^\tau)^\top + \kappa I_d$ where (s^τ,a^τ)'s are i.i.d. samples from some distribution ν over $\mathcal{S}\times\mathcal{A}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that*

$$
\left\| \frac{\bar{\mathbf{G}}_T}{T} - \mathbb{E}_\nu\left[\frac{\bar{\mathbf{G}}_T}{T}\right] \right\|_2 \leq \frac{4\sqrt{2}C^2}{\sqrt{T}} \left(\log\frac{2d}{\delta}\right)^{1/2}.
$$