Learning from Demonstration: Provably Efficient Adversarial Policy Imitation with Linear Function Approximation

Zhihan Liu [1] **Yufeng Zhang** [1] **Zuyue Fu** [1] **Zhuoran Yang** [2] **Zhaoran Wang** [1]

Abstract

In generative adversarial imitation learning (GAIL), the agent aims to learn a policy from an expert demonstration so that its performance cannot be discriminated from the expert policy on a certain predefined reward set. In this paper, we study GAIL in both online and offline settings with linear function approximation, where both the transition and reward function are linear in the feature maps. Besides the expert demonstration, in the online setting the agent can interact with the environment, while in the offline setting the agent only accesses an additional dataset collected by a prior. For online GAIL, we propose an optimistic generative adversarial policy imitation algorithm (OGAPI) and prove that OGAPI achieves $\widetilde{\mathcal{O}}(\sqrt{H^4d^3K}+\sqrt{H^3d^2K^2/N_1})$ regret. Here N_1 represents the number of trajectories of the expert demonstration, d is the feature dimension, K is the number of episodes, and $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic terms and constants. For offline GAIL, we propose a pessimistic generative adversarial policy imitation algorithm (PGAPI). We also obtain the optimality gap of PGAPI, achieving the minimax lower bound in the utilization of the additional dataset. Assuming sufficient coverage on the additional dataset, we show that PGAPI achieves $\widetilde{\mathcal{O}}(\sqrt{H^4d^2/K}+\sqrt{H^4d^3/N_2}+\sqrt{H^3d^2/N_1})$ optimality gap. Here N_2 represents the number of trajectories of the additional dataset with sufficient coverage.

1. Introduction

In imitation learning (IL, Hussein et al. (2017)) (a.k.a apprenticeship learning), the agent remains unknown of the reward, but can learn from an expert demonstration so that the agent learns a policy as good as the expert one. To solve IL problem, there exist mostly three types of methods: behavior cloning (BC, Argall et al. (2009)), inverse reinforcement learning (IRL, Abbeel & Ng (2004)), and online generative adversarial imitation learning (online GAIL). BC regards IL as a supervised learning problem of predicting actions based on states. While appealingly simple, BC suffers from compounding error caused by covariate shift (Ross & Bagnell, 2010). IRL explicitly solves the true reward function and then accordingly fully solves an RL subproblem at every iteration (Abbeel & Ng, 2004; Ng & Russell, 2000). Though it has succeeded in tasks involving continuous spaces (Finn et al.), IRL lacks computational efficiency and the desired true reward function may not be unique. To address these issues, online GAIL (Ho & Ermon, 2016) solves IL through minimax optimization with alternating updates to learn a policy whose performance cannot be discriminated from the expert policy on a certain predefined reward set. The alternating updates in online GAIL mirror the training of generative adversarial networks (Goodfellow et al., 2014; Arjovsky et al., 2017). Specifically, at every iteration, online GAIL first minimizes the discrepancy in expected cumulative reward between the expert policy and the learned policy and then maximizes such a discrepancy over a given reward function class in adversary. Online GAIL achieves tremendous empirical success in various fields, such as autonomous driving (Kuefler et al., 2017), human behavior modeling (Merel et al., 2017), natural language processing (Chen et al., 2017), and robotics control (Tsurumine et al., 2019).

Despite the state-of-art empirical performance of online GAIL, the agent requires a huge amount of interactions with the environment during the training. For some practical problems, it is inconvenient, costly, or risky to get expert data or labeled data, especially when collecting clinical data or developing autonomous driving. Meanwhile, it is available to get other sources of offline data, which may be originated from historical experiments, non-labeled data,

[1] Northwestern University [2] Yale University. Correspondence to: Zhihan Liu <zhihanliu2027@u.northwestern.edu>, Yufeng Zhang <yufengzhang2023@u.northwestern.edu>, Zuyue Fu <zuyue.fu@u.northwestern.edu>, Zhuoran Yang <zhuoran.yang@yale.edu>, Zhaoran Wang <zhaoranwang@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

and published datasets, etc. Naturally, we desire to utilize these offline data to alleviate the shortage of expert demonstration and aid the agent to mimic the expert policy. To this end, besides online GAIL, we consider the offline generative adversarial imitation learning (offline GAIL) setting. In offline GAIL, we assume that the agent is accessible to an additional dataset besides the expert demonstration, without further interaction with the environment. Some related works (Zolna et al., 2020; Zhang & Wu, 2021) study this setting, providing empiricial methods.

Furthermore, previous theoretical analyses on GAIL either focus on the tabular case (Shani et al., 2021), where the state and action spaces are discrete, or relies on strong assumptions, including access to a well-explored dataset (Zhang et al., 2020), linear-quadratic regulators (Cai et al., 2019), or kernelized nonlinear regulators (Chang et al., 2021). Theoretical analysis for GAIL with linear function approximation either in online or offline settings still remains an open problem, which is crucial for the application of GAIL in the continuous or high dimensional state and action spaces. The cruxes of such an analysis involve: (i) Different from RL, both online GAIL and offline GAIL are minimax optimization problems with respect to the policy and reward function, especially with linear reward set. (ii) For offline GAIL, without assuming the well-exploredness of the additional dataset, the agent may be misled by distribution shift in the additional dataset and shares the suffering with offline RL (Jin et al., 2021; Wang et al., 2020a); (iii) For offline GAIL, we are incapable to update the reward function based on the trajectory of present policy.

Hence in this paper, we aim at tackling these issues and answering the following question:

Can we design provably efficient algorithms for online and offline GAIL with linear function approximation?

To answer the above question, we present a unified framework and specialize it as **O**ptimistic **G**enerative **A**dversarial **P**olicy **I**mitation (OGAPI) for online GAIL and **P**essimistic **G**enerative **A**dversarial **P**olicy **I**mitation (PGAPI) for offline GAIL with linear function approximation. This framework is motivated by the alternating update process of GANs and involves two main stages: policy update stage and reward update stage. (i) In the policy update stage, we apply mirror descent (Beck & Teboulle, 2003; Hazan, 2019) to update the policy and evaluate policy online optimistically for OGAPI and offline pessimistically for PGAPI. (ii) In the reward update stage, we first estimate the gradient of GAIL objective function with respect to the reward parameter through the collected trajectory induced by the present policy for OGAPI. While for PGAPI, we build the estimate through estimated action-value functions during the stage of policy update. Then we use projected gradient ascent to update reward parameters via such an estimated gradient.

Contribution Particularly, we conclude our contributions in the following three aspects. First, for online GAIL with linear function approximation, we propose a new algorithm OGAPI and prove that OGAPI achieves $\widetilde{\mathcal{O}}(\sqrt{H^4d^3K} + \sqrt{H^3d^2K^2/N_1})$ regret when applying linear function approximation, demonstrating that OGAPI is provably efficient. Here N_1 represents the number of trajectories of the expert demonstration, d is the feature dimension, H is the horizon, K is the number of episodes, and $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic terms and constants Second, for offline GAIL with linear function approximation, we design a new algorithm PGAPI and obtain the optimality gap of the output policy under the minimal assumption on the additional dataset. Then we decompose the optimality gap into three sources: optimization error, Monte Carlo (MC) estimation error, and intrinsic error. We prove that optimization error and MC estimation error respectively scale to $\widetilde{\mathcal{O}}(K^{-1/2})$ and $\widetilde{\mathcal{O}}(N_1^{-1/2})$, while intrinsic error depends on how well the additional dataset $\mathbb{D}^{\mathrm{A}}$ covers the expert policy and attains the minimax optimality in the utilization of the additional dataset. Third, we demonstrate that if we further assume that the additional dataset $\mathbb{D}^{\mathrm{A}}$ has sufficient coverage on the expert policy, we prove PGAPI achieves $\widetilde{\mathcal{O}}(\sqrt{H^4d^2/K} + \sqrt{H^4d^3/N_2} + \sqrt{H^3d^2/N_1})$ optimality gap, thus PGAPI has global convergence. Here N_2 represents the number of trajectories of the additional dataset. Furthermore, we discuss the effect of the additional offline dataset $\mathbb{D}^{\mathrm{A}}$. In particular, facilitated with an additional dataset $\mathbb{D}^{\mathrm{A}}$ with sufficient coverage, we decrease the dependency for H and d in the optimality gap.

Related Works. Our work adds to the body of analysis on GAIL (Cai et al., 2019; Chen et al., 2020; Zhang et al., 2020; Xu et al., 2020; Shani et al., 2021; Chang et al., 2021). Shani et al. (2021) study online GAIL and obtain $\widetilde{\mathcal{O}}(\sqrt{H^4|\mathcal{S}|^2|\mathcal{A}|K} + \sqrt{H^3|\mathcal{S}||\mathcal{A}|K^2/N_1})$ regret in the tabular case with bounded reward functions but we apply linear function approximation on the transition kernels without assuming the state space or the action space is discrete and we adopt linear reward set. Chen et al. (2020) only study the convergence of offline GAIL to a stationary point instead of global convergence (optimality gap) as in this paper. Xu et al. (2020); Zhang et al. (2020) analyze the global convergence of GAIL with neural networks respectively in the tabular case and the continuous case but assume that a well-explored dataset is available (concentrability coefficients are uniformly upper bounded), while our analysis need not such a strict and impractical assumption. Cai et al. (2019) study the global convergence of offline GAIL in the setting of linear-quadratic regulators, which is unnecessary for this paper. Chang et al. (2021) study offline GAIL with bounded reward functions in the continuous kernelized nolinear regulator (KNR, Kakade et al. (2020)) and gaussian process (GP, Fisac et al. (2018)) setting. We point out that the KNR (resp. GP) setting is different from linear kernel MDP as analyzed in this paper and each one does not imply the other, which leads to the difference in model estimation and later analysis. In addition, we study the linear reward set instead of bounded reward set and the former case is difficult to handle with Shani et al. (2021).

Our work is also related to BC (Ross & Bagnell, 2010; Rajaraman et al., 2020; 2021; Rashidinejad et al., 2021). BC does not solve a minimax problem as GAIL, but directly mimics the expert policy extracted from the expert demonstration. Rajaraman et al. (2021) propose a BC method which achieves $\widetilde{\mathcal{O}}(|\mathcal{S}|H^2/N_1)$ suboptimality, attaining $\Omega(|\mathcal{S}|H^2/N_1)$ the lower bound of BC (Rajaraman et al., 2020), when the transition model is unknown. To best of our knowledge, present analysis of BC only focus on the tabular case and would fail in the continuous state and action space with horizon $H \geq 2$, since BC is considered as a classification problem and always faces unseen states in the continuous state space. See the remaining discussions of related works in §B.

2. Preliminary

In this section, we introduce the notion of the episodic Markov decision process (MDP), generative adversarial imitation learning in the online and offline settings, and linear function approximation.

2.1. Episodic Markov Decision Process

We consider an episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively, H is the length of each episode, $\mathcal{P}_h$ is the Markov transition kernel of the h-th step of each episode for any $h \in [H]$, and $r_h \colon \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function at the h-th step of each episode for any $h \in [H]$. We assume without loss of generality that the reward function r_h is deterministic.

In the episodic MDP, the agent interacts with the environment as follows. At the beginning of each episode, the agent determines a policy $\pi = \{\pi_h\}_{h\in[H]} \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)$. Then the agent takes the action $a_h \sim \pi_h(\cdot \,|\, s_h)$ at the h-th step of the k-th episode, observes the reward $r_h(s_h, a_h)$, and transits to the next state $s_{h+1} \sim \mathcal{P}_h(\cdot \,|\, s_h, a_h)$. The episode terminates when the agent reaches the state s_{H+1}. Without loss of generality, we assume that the initial state $s_1 = x$ is fixed across different episodes. We remark that our analyses readily generalize to the setting where the initial state s_1 is sampled from a fixed distribution.

We now define the value functions in the episodic MDP. For any policy $\pi = \{\pi_h\}_{h\in[H]}$ and reward function $r = \{r_h\}_{h\in[H]}$, the state- and action-value functions are defined for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$ as follows,

$$V^r_{h,\pi}(s) = \mathbb{E}_\pi\Big[\sum_{i=h}^{H} r_i(s_i, a_i) \big| s_h = s\Big], \tag{1}$$

$$Q^r_{h,\pi}(s, a) = \mathbb{E}_\pi\Big[\sum_{i=h}^{H} r_i(s_i, a_i) \big| s_h = s, a_h = a\Big], \tag{2}$$

where the expectation $\mathbb{E}_\pi[\cdot]$ is taken with respect to the action $a_i \sim \pi_i(\cdot \,|\, s_i)$ and the state $s_{i+1} \sim \mathcal{P}_i(\cdot \,|\, s_i, a_i)$ for any $i \in \{h, h+1, \ldots, H\}$. With slight abuse of notations, we also denote by $\mathcal{P}_h$ the operator form of the transition kernel such that $(\mathcal{P}_h f)(s, a) = \mathbb{E}_{s' \sim \mathcal{P}_h(\cdot \,|\, s,a)}[f(s')]$ for any $f \colon \mathcal{S} \to \mathbb{R}$. By the definitions of the value functions in (2), for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, any policy π, and any reward function r, we have

$$V^r_{h,\pi}(s) = \langle Q^r_{h,\pi}(s, \cdot), \pi_h(\cdot, s)\rangle_{\mathcal{A}}, \tag{3}$$

$$Q^r_{h,\pi}(s, a) = r_h(s, a) + \mathcal{P}_h V^r_{h+1,\pi}(s, a), \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product over the action space $\mathcal{A}$ and $V^r_{H+1,\pi}(s)$ is set to zero for any $s \in \mathcal{S}$. We further define the expected cumulative reward as follows,

$$J(\pi, r) = V^r_{1,\pi}(x). \tag{5}$$

In this paper, we characterize the performance of the agent via the expected cumulative reward $J(\pi, r)$ defined in (5).

2.2. Generative Adversarial Imitation Learning

Given an expert demonstration with N_1 trajectories of state-action pairs $\mathbb{D}^{\mathrm{E}} = \{(s^{\mathrm{E}}_{h,\tau}, a^{\mathrm{E}}_{h,\tau})\}_{h\in[H],\tau\in[N_1]}$, which generated following the underlying MDP and the expert policy π^{E}, the goal of GAIL is to find a policy whose performance is close to that of the expert policy π^{E} for any reward function in a given set $\mathcal{R}$ (Ho & Ermon, 2016). Here the set $\mathcal{R}$ is specified later in §2.3. We assume that the trajectories in the expert demonstration $\mathbb{D}^{\mathrm{E}}$ are independent, which is a standard assumption in the literature (Abbeel & Ng, 2004; Shani et al., 2021). In GAIL, we consider the following minimax optimization problem,

$$\min_{\pi\in\Delta(\mathcal{S}|\mathcal{A},H)} \max_{r\in\mathcal{R}} J(\pi^{\mathrm{E}}, r) - J(\pi, r), \tag{6}$$

where $J(\pi, r)$ is defined in (5).

Online GAIL. In online GAIL, the agent interacts with the environment to collect state-action pairs following the underlying MDP and the current policy. For online GAIL, we are interested in the performance of the algorithm during learning. To this end, we compare the expected cumulative reward corresponding to the algorithm during learning with the expected cumulative reward corresponding to the expert policy under the worst-case scenario, which is defined as follows (Shani et al., 2021),

$$\mathrm{Regret}(K) = \max_{r\in\mathcal{R}} \sum_{k=1}^{K} \big[J(\pi^{\mathrm{E}}, r) - J(\pi^{k}, r)\big], \qquad (7)$$

where π^k is the policy of the agent at the k-th episode.

Offline GAIL. To simultaneously utilize non-expert data without further interaction with the environment, we consider offline GAIL, which involves an additional dataset to benefit the policy learning. Specifically, except for the expert demonstration $\mathbb{D}^{\mathrm{E}} = \{(s^{\mathrm{E}}_{h,\tau}, a^{\mathrm{E}}_{h,\tau})\}_{h\in[H],\tau\in[N_1]}$ collected by the expert policy π^{E} in the underlying MDP, the agent has access to an additional dataset $\mathbb{D}^{\mathrm{A}} = \{(s^{\tau}_{h}, a^{\tau}_{h})\}_{h\in[H],\tau\in[N_2]}$, which is collected a priori by an experimenter in the underlying MDP. In particular, at each step $h \in [H]$ of each trajectory $\tau \in [N_2]$, the experimenter takes the action a^{τ}_{h} at the state s^{τ}_{h} and observes the next state $s^{\tau}_{h+1} \sim \mathcal{P}_h(\cdot \,|\, s^{\tau}_{h}, a^{\tau}_{h})$. Here a^{τ}_{h} is arbitrarily chosen by the experimenter given the filtration

$$\mathcal{F}_{h,\tau} = \sigma\big(\{(s^{n}_{i}, a^{n}_{i})\colon (n-1)H + i \le (\tau-1)H + h\}\big),$$

In other words, in the τ-th trajectory, the action the experiment takes is only determined by the historical information with randomness. For offline GAIL, we measure the performance of a policy π by the $\mathcal{R}$-distance (Chen et al., 2020) between the expert policy π^{E} and π, which is defined as,

$$\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}}, \pi) = \max_{r\in\mathcal{R}} \big[J(\pi^{\mathrm{E}}, r) - J(\pi, r)\big]. \qquad (8)$$

Here $\mathcal{R}$ is the reward set, which is specified later in §2.3. Optimality gap defined in (8) can be considered as one episode regret defined in (7). When optimality gap of policy π approaches zero, it implies that the performance difference between the policy π and the expert policy π^{E} tends to be undistinguishable by the reward set $\mathcal{R}$, which implies that the performance of π is measured by both the optimality gap $\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}}, \pi)$ and the richness of the reward set $\mathcal{R}$.

2.3. Linear Function Approximation

We consider the linear setting where the transition kernel is linear in a feature map, which is formalized as follows.

Assumption 2.1 (Linear Kernel Episodic MDP)**.** Given measurable sets $\mathcal{S}$ and $\mathcal{A}$ with finite measure, the episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r^{\mu})$ is a linear MDP with a feature map $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, that is, for any $h \in [H]$, there exists $\theta_h \in \mathbb{R}^d$ with $\|\theta_h\|_2 \le \sqrt{d}$ such that $\mathcal{P}_h(s' \,|\, s, a) = \phi(s, a, s')^{\top}\theta_h$ for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Also, there exists an absolute constant $R > 0$ such that

$$R^{-2} \cdot \sup_{s'\in\mathcal{S}} |\phi(s, a, s')^{\top} y|^2 \le \int_{\mathcal{S}} |\phi(s, a, s')^{\top} y|^2 \, \mathrm{d}s' \le d,$$

for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $y \in \mathbb{R}^d$ with $\|y\|_2 \le 1$.

Under Assumption 2.1, we further assume that there exists a feature map $\psi\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that the reward set $\mathcal{R}$ in (6) takes the following form,

$$\big\{r^{\mu}\colon r^{\mu}_{h}(\cdot,\cdot) = \psi(\cdot,\cdot)^{\top}\mu_h \text{ for any } (h, \mu) \in [H]\times S\big\}, \quad (9)$$

where $r^{\mu} = \{r^{\mu}_{h}(\cdot,\cdot)\}_{h\in[H]}$ is the reward function and $\mu = \{\mu_h\}_{h\in[H]}$ is the reward parameter. Here S is the reward parameter domain, which is defined as follows,

$$S = \{\mu\colon \mu_h \in B \text{ for any } h \in [H]\}, \qquad (10)$$

where $B = \{u \in \mathbb{R}^d\colon \|u\|_2 \le \sqrt{d}\}$. We assume that $\|\psi(s, a)\|_2 \le 1$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, which ensures that $r^{\mu}_{h}(s, a) \in [0, \sqrt{d}\,]$ for any $(s, a, h, \mu) \in \mathcal{S} \times \mathcal{A} \times [H] \times S$. For notational convenience, for any reward function r^{μ}, we denote by the GAIL objective function $L(\pi, \mu)$ as follow,

$$L(\pi, \mu) = J(\pi^{\mathrm{E}}, r^{\mu}) - J(\pi, r^{\mu}), \qquad (11)$$

where $J(\pi, r^{\mu})$ is defined in (5).

Assumption 2.1 corresponds to the linear kernel MDP model in RL. See Ayoub et al. (2020); Zhou et al. (2021); Cai et al. (2020) for various examples of linear kernel MDPs. We remark that the existence of R in Assumption 2.1 can be guaranteed if for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, the feature map $\phi(s, a, \cdot)$ is upper bounded and Lipschitz continuous. Particularly, a tabular MDP where the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are both finite, is a special case of the linear kernel MDP in Assumption 2.1 with $d = |\mathcal{S}|^2|\mathcal{A}|$ and the feature map $\phi(s, a, s')$ being the canonical basis $e_{(s,a,s')}$ of $\mathbb{R}^{|\mathcal{S}|^2|\mathcal{A}|}$. It implies that our analysis for GAIL with linear function approximation also covers the tabular case. We also remark that the range of the reward function is $[0, \sqrt{d}]$ instead of $[0, 1]$. With the increasing d, we can enrich the reward set $\mathcal{R}$ and then capture the performance of the policy more meticulously using the optimality gap defined in (8). Analysis in GAIL with linear reward is more challenging than the case with bounded reward, as studied in the previous literature (Shani et al., 2021; Chang et al., 2021).

3. Algorithms

We first propose a unified framework in Algorithm 1 to solve GAIL in both online and offline settings. Then we specify the framework for online and offline settings in optimistic generative adversarial policy Imitation (OGAPI in Algorithm 2 of §3.1) and pessimistic generative adversarial policy imitation (PGAPI in Algorithm 3 of §3.2), respectively.

This framework in Algorithm 1 involves two stages: policy update stage and reward update stage. (i) In the policy update stage, we use mirror descent to update the policy based on the estimated action-value function constructed in

Algorithm 1 A Unified Framework for OGAPI and PGAPI

```
Initialize $\{Q_h^0\}_{h\in[H]}$ as zero functions over $\mathcal{S}\times\mathcal{A}$ and $\{\pi_h^0\}_{h\in[H]}$ as uniform distributions over $\mathcal{A}$.
(PGAPI) Construct estimated transition kernels $\{\widehat{\mathcal{P}}_h\}_{h\in[H]}$ and uncertainty qualifiers $\{\Gamma_h\}_{h\in[H]}$ based on $\mathbb{D}^{\mathrm{A}}$.
for $k=1,\ldots,K$ do
    Update policy $\pi^k=\{\pi_h^k\}_{h\in[H]}$ by mirror descent with estimated action-value function $\{\widehat{Q}_h^{k-1}\}_{h\in[H]}$.
    (OGAPI) Rollout a trajectory following $\pi^k$, and construct empirically estimated transition kernels $\{\widehat{\mathcal{P}}_h^k\}_{h\in[H]}$ and bonus functions $\{\Gamma_h^k\}_{h\in[H]}$.
    (OGAPI/PGAPI) Optimistically/Pessimistically estimate action-value function $\{\widehat{Q}_h^k\}_{h\in[H]}$.
    (OGAPI/PGAPI) Estimate $\nabla_\mu L(\pi^k,\mu^k)$ via (25)/(28).
    Update reward parameter $\mu^{k+1}$ by projected gradient ascent with estimated $\nabla_\mu L(\pi^k,\mu^k)$.
end for
(PGAPI) Output the mixed policy $\widehat{\pi}$ of $\{\pi^k\}_{k\in[K]}$.
```

the previous iteration. For OGAPI, we sample a trajectory following the updated policy. Then we construct estimated action-value functions with optimism based on the finite historical data for OGAPI or pessimism based on the additional dataset for PGAPI. (ii) In the reward update stage, we first construct an estimate of the gradient based on the collected trajectory induced by the present policy and the finite historical data for OGAPI or estimated action-value functions for PGAPI, and then we use projected gradient ascent to update reward parameters via such an estimate of gradient. We further detail OGAPI and PGAPI as follows.

3.1. Optimistic Generative Adversarial Policy Imitation

To specialize Algorithm 1 to solve online GAIL, we propose OGAPI in Algorithm 2, which is detailed as follows.

3.1.1. Policy Update Stage

The policy update stage (Lines 4–12 of Algorithm 2) consists of two steps: (i) policy improvement (Lines 4–6) and (ii) policy evaluation (Lines 8–12). In policy improvement, we apply mirror descent in its proximal form to update the current policy via estimated action-value functions, which is specified in policy evaluation stage. In policy evaluation, we employ the optimism principle to construct the estimated action-value functions, which further utilize estimated transition kernels and bonus functions.

Policy Improvement. To generate a policy whose performance is close to the expert policy π^{E}, we update the policy π^k to minimize the GAIL objective function $L(\pi,\mu^{k-1})=J(\pi^{\mathrm{E}},r^{k-1})-J(\pi,r^{k-1})$ in (11) under the current reward function $r^{k-1}=r^{\mu^{k-1}}$. Note that π^{E} is fixed, then we only need to maximize $J(\pi,r^{k-1})$. Applying online mirror descent (Beck & Teboulle, 2003; Hazan, 2019), a standard algorithm to solve online learning problem, we update π as follows,

$$\pi^k=\underset{\pi\in\Delta(\mathcal{A}\,|\,\mathcal{S},H)}{\operatorname{argmax}}\ \{\mathcal{L}_{k-1}(\pi)-\alpha^{-1}D(\pi,\pi^{k-1})\},\quad(12)$$

where α is the step size, Bregman divergence regularizer $D(\pi,\pi^{k-1})$ is chosen as the expected KL divergence $\mathbb{E}_{\pi^{k-1}}[\sum_{h=1}^H D_{\mathrm{KL}}(\pi_h(\cdot\,|\,s_h)\|\pi_h^{k-1}(\cdot\,|\,s_h))\,|\,s_1{=}x]$, and $\mathcal{L}_{k-1}(\pi)$ takes the form as

$$\mathcal{L}_{k-1}(\pi)=J(\pi^{k-1},r^{k-1})+\mathbb{E}_{\pi^{k-1}}\Big[\sum_{h=1}^H\big\langle\widehat{Q}_h^{k-1}(s_h,\cdot),\\ \pi_h(\cdot\,|\,s_h)-\pi_h^{k-1}(\cdot\,|\,s_h)\big\rangle_{\mathcal{A}}\,\Big|\,s_1=x\Big].\quad(13)$$

Here expectation $\mathbb{E}_\pi[\cdot]$ is taken with respect to the trajectory induced by π and $\widehat{Q}_h^{k-1}$ is an estimator of $Q_{h,\pi^{k-1}}^{r^{k-1}}$, which is specified later in (20). Such policy update formulation defined in (12) also corresponds to the policy optimization in online RL (Kakade, 2001; Schulman et al., 2015; 2017; Geist et al., 2019; Shani et al., 2020a; Cai et al., 2020).

By solving (12), we obtain the following closed-form solution for any $(s,h)\in\mathcal{S}\times[H]$,

$$\pi_h^k(\cdot\,|\,s)\propto\pi_h^{k-1}(\cdot\,|\,s)\cdot\exp\{\alpha\cdot\widehat{Q}_h^{k-1}(s,\cdot)\},\quad(14)$$

which gives Line 5 of Algorithm 2.

Policy Evaluation. To evaluate the policy π^k under the reward function r^k, we first construct estimated transition kernels $\widehat{\mathcal{P}}^k=\{\widehat{\mathcal{P}}_h^k\}_{h\in[H]}$, through value-target regression (Ayoub et al., 2020) on finite historical data in Line 9, and then construct an estimator of the action-value functions by the Bellman equation in (4) with an extra bonus term to incorporate exploration in Line 10.

Specifically, in the k-th episode, we construct our estimated transition kernels $\widehat{\mathcal{P}}^k=\{\widehat{\mathcal{P}}_h^k\}_{h\in[H]}$ as

$$\widehat{\mathcal{P}}_h^k(s'\,|\,s,a)=\phi(s,a,s')^\top\widehat{\theta}_h^k,\quad(15)$$

for any $(h,s,a,s')\in[H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}$, where $\widehat{\theta}_h^k$ is the minimizer of the regularized empirical mean-squared Bellman error defined as follows,

$$\min_{\theta\in\mathbb{R}^d}\sum_{\tau=1}^{k-1}\big|\varphi_h^\tau(s_h^\tau,a_h^\tau)^\top\theta-\widehat{V}_h^\tau(s_h^\tau)\big|^2+\lambda\|\theta\|_2^2,\quad(16)$$

where $\varphi_h^\tau(\cdot,\cdot)=\int_{\mathcal{S}}\phi\,(\cdot,\cdot,s')\,\widehat{V}_{h+1}^\tau\,(s')\,\mathrm{d}s'$. Here $\widehat{V}_h^\tau$ is constructed in Line 11 of Algorithm 2 and $\lambda>0$ is the regularization parameter, which is specified later in Theorem 4.1.

By solving (16), we obtain the closed-form update of $\widehat{\theta}_h^k$ as

$$\widehat{\theta}_h^k = (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \varphi_h^\tau(s_h^\tau, a_h^\tau) \widehat{V}_{h+1}^\tau(s_{h+1}^\tau), \qquad (17)$$

where $\Lambda_h^k = \sum_{\tau=1}^{k-1} \varphi_h^\tau(s_h^\tau, a_h^\tau) \varphi_h^\tau(s_h^\tau, a_h^\tau)^\top \mathrm{d}s' + \lambda I$. We use $(r_h^k + \widehat{\mathcal{P}}_h^k \widehat{V}_{h+1}^k)(s,a)$ as an estimator of $Q_{h,\pi^k}^{r^k}(s,a)$. To further handle the uncertainty incurred by finite historical data and balance between exploration and exploitation, we employ optimism to incentivize exploration as many no-regret online RL algorithms do (Auer et al., 2002; 2009; Azar et al., 2017; Jin et al., 2018; 2019; Yang & Wang, 2020). Specifically, we first define the bonus term as

$$\Gamma_h^k(s,a) = H\sqrt{d} \cdot \min\left\{\kappa \cdot \|\varphi_h^k(s,a)\|_{(\Lambda_h^k)^{-1}}, 1\right\}, \quad (18)$$

where $\kappa > 0$ is a scaling parameter. Then we incorporate such a bonus term into the estimator $(r_h^k + \widehat{\mathcal{P}}_h^k \widehat{V}_{h+1}^k)(s,a)$ of $Q_{h,\pi^k}^{r^k}(s,a)$, i.e.,

$$\bar{Q}_h^k(\cdot,\cdot) = (r_h^k + \widehat{\mathcal{P}}_h^k \widehat{V}_{h+1}^k + \Gamma_h^k)(\cdot,\cdot), \qquad (19)$$

$$\widehat{Q}_h^k(\cdot,\cdot) = \min\left\{\bar{Q}_h^k(\cdot,\cdot), (H-h+1)\sqrt{d}\right\}_+. \qquad (20)$$

We highlight that the policy update stage of OGAPI (Lines 4–12 of Algorithm 2) corresponds to the no-regret policy optimization in adversarial MDP with full information feedback (Shani et al., 2020b; Rosenberg & Mansour, 2019; Cai et al., 2020; Jin et al., 2020). Such tolerance of arbitrarily chosen reward function every episode paves the way for the alternate update between the policy and reward function.

3.1.2. Reward Update Stage

To discriminate the discrepancy between the expert policy π^{E} and the current policy π^k, we update the reward parameter μ^{k+1} by maximizing GAIL objective function $L(\pi^k, \mu)$ defined in (11). By projected gradient ascent, we obtain the update of the reward parameter as follows,

$$\mu_h^{k+1} = \mathrm{Proj}_B\{\mu_h^k + \eta \widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k)\}, \qquad (21)$$

where η is the stepsize, $\widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k)$ is an estimator of $\nabla_{\mu_h} L(\pi^k, \mu^k)$, and $\mathrm{Proj}\colon \mathbb{R}^d \to B$ is the projection operator to restrict the updated reward parameter μ_h^{k+1} within the ball B for any $h \in [H]$. Here B is defined in (10). Without accessing to the true transition kernels of the underlying MDP and the expert policy π^{E}, we need to obtain an estimator $\widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k)$ in (21).

Specifically, to construct an estimator of $\nabla_{\mu_h} L(\pi^k, \mu^k)$, we first construct a Monte Carlo (MC) estimator $\widehat{L}(\pi^k, \mu^k)$ of $L(\pi^k, \mu^k)$ as follows,

$$\widehat{L}(\pi^k, \mu^k) = \widetilde{J}(\pi^{\mathrm{E}}, r^k) - \widetilde{J}(\pi^k, r^k). \qquad (22)$$

Here $\widetilde{J}(\pi^{\mathrm{E}}, r^k)$ and $\widetilde{J}(\pi^k, r^k)$ are MC estimators of $J(\pi^{\mathrm{E}}, r^k)$ and $J(\pi^k, r^k)$, which are defined as,

$$\widetilde{J}(\pi^{\mathrm{E}}, r^k) = \frac{1}{N_1} \sum_{\tau=1}^{N_1} \sum_{h=1}^{H} \psi(s_{h,\tau}^{\mathrm{E}}, a_{h,\tau}^{\mathrm{E}})^\top \mu_h, \qquad (23)$$

$$\widetilde{J}(\pi^k, r^k) = \sum_{h=1}^{H} \psi(s_h^k, a_h^k)^\top \mu_h, \qquad (24)$$

where we use N_1 trajectories in $\widetilde{J}(\pi^{\mathrm{E}}, r^k)$ and one trajectory in $\widetilde{J}(\pi^k, r^k)$. Combining (22) and (23), we obtain that

$$\widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k) = \frac{1}{N_1} \sum_{\tau=1}^{N_1} \psi(s_{h,\tau}^{\mathrm{E}}, a_{h,\tau}^{\mathrm{E}}) - \psi(s_h^k, a_h^k). \quad (25)$$

We use $\widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k)$ as an estimator of $\nabla_{\mu_h} L(\pi^k, \mu^k)$, which gives Lines 13–17 of Algorithm 2.

3.2. Pessimistic Generative Adversarial Policy Imitation

To specialize Algorithm 1 to solve offline GAIL, we propose PGAPI in Algorithm 3. Besides the policy update stage and reward update stage, PGAPI further contains an initial construction stage, which constructs estimated transition kernels and bonus functions at the beginning of the algorithm. We detail the initial construction, the policy update, and the reward update stage as follows.

3.2.1. Initial Construction Stage

In Line 3 of PGAPI, we construct estimated transition kernels $\{\widehat{\mathcal{P}}_h\}_{h\in[H]}$ and uncertainty quantifiers $\{\Gamma_h\}_{h\in[H]}$ via the additional dataset $\mathbb{D}^{\mathrm{A}}$. Before we detail such construct, we first introduce the following definition of uncertainty quantifiers (Jin et al., 2021) with the confidence parameter $\xi \in (0,1)$, which quantifies the uncertainty.

Definition 3.1 (ξ-Uncertainty Quantifier)**.** We say $\{\Gamma_h\}_{h\in[H]}$ with $\Gamma_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are ξ-uncertainty quantifiers for estimated kernels $\widehat{\mathcal{P}} = \{\widehat{\mathcal{P}}_h\}_{h\in[H]}$ with respect to $\mathbb{P}_{\mathbb{D}}$ if the event

$$\begin{aligned}\mathcal{E} = \big\{&|\widehat{\mathcal{P}}_h \widehat{V}(s,a) - \mathcal{P}_h \widehat{V}(s,a)| \le \Gamma_h(s,a), \text{ for any} \\ &(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H] \text{ and any } \widehat{V} : \mathcal{S} \to [0, H\sqrt{d}\,]\big\}\end{aligned}$$

satisfies $\mathbb{P}_{\mathbb{D}}(\mathcal{E}) \ge 1 - \xi/2$. Here $\mathbb{P}_{\mathbb{D}}$ is with respect to the joint distribution of $\mathbb{D}^{\mathrm{A}} \cup \mathbb{D}^{\mathrm{E}}$.

We remark that the ξ-uncertainty quantifiers in Definition 3.1 is a counterpart of the bonus functions in OGAPI. Recalling that $|r_h(s,a)| \le \sqrt{d}$ for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, Definition 3.1 implies that with probability at least $1 - \xi/2$, the deviation between the true Bellman equation in (4) with $\mathcal{P}$ and the estimated Bellman equation in (4) with $\widehat{\mathcal{P}}$ is upper bounded by the ξ-uncertainty quantifier $\{\Gamma_h\}_{h\in[H]}$. Due to the page limit, we postpone the construction process in §E.

3.2.2. Policy Update Stage

As a pessimistic variant of OGAPI in Algorithm 2, the policy update stage of PGAPI (Lines 5–11 of Algorithm 3) consists of two steps: (i) policy improvement (Lines 5–7) and (ii) policy evaluation (Lines 8–11). In the stage of policy improvement, we adopt the same idea as in OGAPI, which employs mirror descent to update the current policy. In the stage of policy evaluation, instead of the optimism principle, we employ the pessimism principle to construct the estimated action-value functions via the additional dataset $\mathbb{D}^{\mathrm{A}}$, which is not assumed to be well-explored as specified later in §4.2. The principle pessimism-in-face-of-uncertainty guides the agent to be conservative to visit the states and actions that are less covered by the additional dataset $\mathbb{D}^{\mathrm{A}}$ (Kumar et al., 2020; Jin et al., 2021; Liu et al., 2020; Yu et al., 2020; 2021; Buckman et al., 2020). Specifically, we construct the estimated action-value functions as follows,

$$\widehat{Q}_h^k(\cdot,\cdot) = \max\big\{(r_h^k + \widehat{\mathcal{P}}_h \widehat{V}_{h+1}^k - \Gamma_h)(\cdot,\cdot), 0\big\},$$

where $\{\widehat{\mathcal{P}}_h\}_{h\in[H]}$ are the estimated transition kernels and $\{\Gamma_h\}_{h\in[H]}$ are the uncertainty quantifiers constructed in Line 3 of PGAPI, which satisfies Definition 3.1.

3.2.3. Reward Update Stage

Similar to the reward update stage of OGAPI, in the reward update stage of PGAPI, we update the reward parameter as,

$$\mu_h^{k+1} = \mathrm{Proj}_B\{\mu_h^k + \eta \widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k)\}. \tag{26}$$

Here η is the stepsize, $\widehat{\nabla}_{\mu_h} L(\pi^k,\mu^k)$ is an estimator of $\nabla_{\mu_h} L(\pi^k,\mu^k)$, and $\mathrm{Proj}\colon \mathbb{R}^d \to B$ is the projection operator to restrict the updated reward parameter μ_h^{k+1} within the ball B for any $h \in [H]$. Here B is defined in (10). To achieve (26), we also need to obtain the estimated gradient $\widehat{\nabla}_{\mu_h} L(\pi^k,\mu^k)$ in (26). However, since the agent in offline GAIL cannot interact with the environment to collect the state-action pairs following current policy π^k, the estimator in (25) for OGAPI is not applicable to PGAPI. Instead, we construct an estimator $\widehat{L}(\pi^k,\mu^k)$ for $L(\pi^k,\mu^k)$ and use its gradient $\nabla_{\mu_h}\widehat{L}(\pi^k,\mu^k)$ to estimate $\nabla_{\mu_h} L(\pi^k,\mu^k)$. Specifically, we define the estimator $\widehat{L}(\pi^k,\mu^k)$ as

$$\widehat{L}(\pi^k,\mu^k) = \widetilde{J}(\pi^{\mathrm{E}}, r^k) - \widehat{J}(\pi^k, r^k), \tag{27}$$

where $\widetilde{J}(\pi^{\mathrm{E}}, r^k)$ is a MC estimator of $J(\pi^{\mathrm{E}}, r^k)$. Here we estimate $J(\pi^k, r^k)$ with $\widehat{J}(\pi^k, r^k) = \widehat{V}_1^k(x)$, which is constructed in Line 9 of PGAPI. Based on (27), we construct an estimator $\widehat{\nabla}_{\mu_h} L(\pi^k,\mu^k)$ of $\nabla_{\mu_h} L(\pi^k,\mu^k)$ by taking gradient on $\widehat{L}(\pi^k,\mu^k)$ w.r.t μ_h as follows,

$$\widehat{\nabla}_{\mu_h} L(\pi^k,\mu^k) = \nabla_{\mu_h}\widetilde{J}(\pi^{\mathrm{E}}, r^k) - \nabla_{\mu_h}\widehat{J}(\pi^k, r^k). \tag{28}$$

4. Main Results

In this section, we present the theoretical analysis for OGAPI and PGAPI in §4.1 and §4.2, respectively. Specifically, in §4.1 we upper bound the regret of OGAPI. In §4.2, we upper bound the optimality gap of PGAPI under no coverage assumption and propose a lower bound to show that PGAPI achieves minimax optimality in the utilization of the additional dataset $\mathbb{D}^{\mathrm{A}}$. Moreover, under the assumption that the additional dataset $\mathbb{D}^{\mathrm{A}}$ has sufficient coverage, we establish the global convergence guarantee for PGAPI.

4.1. Analysis of OGAPI

We derive an upper bound of the regret of OGAPI in the following theorem, whose proof sketch is in §H.1

Theorem 4.1 (Regret of OGAPI)**.** *In Algorithm 2, we set $\alpha = (2\log|\mathcal{A}|/(H^2\sqrt{d}K))^{1/2}$, $\lambda = 1$, $\kappa = C\sqrt{d\log(HdK/\xi)}$, $\eta = 1/\sqrt{HK}$, where $C > 0$ is a constant. Under Assumption 2.1, it holds with probability at least $1-\xi$ that*

$$\mathrm{Regret}(K) \leq \mathcal{O}\big(\sqrt{H^4 d^3 K}\log(HdK/\xi)\big) + K\Delta_{N_1}, \tag{29}$$

where $\Delta_{N_1} = \mathcal{O}(\sqrt{H^3 d^2/N_1}\log(N_1/\xi))$.

The first term on the right-hand side of (29) scales with $\sqrt{K}$, which attains the optimal dependency on K for online RL. The second term on the right-hand side of (29) is linear in K and depends on the MC estimation error Δ_{N_1}. As the statistical error from the MC estimation, the error term Δ_{N_1} is inevitable and independent of GAIL algorithm, since we cannot access the expert policy but expert demonstration with N_1 trajectories. When the number of trajectories N_1 in the expert demonstration is sufficiently large such that $N_1 = \Omega(K)$, the first term on the right-hand side of (29) dominates the regret upper bound so that the regret of OGAPI scales with $\sqrt{K}$. The dependency of H, K, and N_1 correspond to $\widetilde{\mathcal{O}}(\sqrt{H^4|\mathcal{S}|^2|\mathcal{A}|K} + \sqrt{H^3|\mathcal{S}||\mathcal{A}|K^2/N_1})$ regret in the tabular case, established by Shani et al. (2021). If we consider the case $K = \Omega(N_1^{3/2})$, then the average regret decays at a rate of $N_1^{-1/2}$ and the dependency for H turns from H^2 into $H^{3/2}$. As K and N_1 both tend to infinity, the average regret also shrinks to zero, meaning that the output policy has the same performance on average with the expert policy with respect to the linear reward set $\mathcal{R}$.

According to Assumption 2.1, if we constrain the reward set $\mathcal{R}$ to a fixed reward function $r = \{r_h\}_{h\in[H]}$, then GAIL (6) is reduced to RL, with respect to an episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$, where $\mathcal{S}, \mathcal{A}, H, \mathcal{P}$ are the same as the ones in Assumption 2.1. Hence OGAPI can also be considered as an RL algorithm for episodic MDP with linear function approximation. From the aspect for information-theory, the lower bound of the regret of any online RL algorithm is of the square-root order with respect to K even in tabular case (Jin et al., 2018). Since the reward set is singleton, OGAPI needs not MC estimation, whose regret is also of the square-root order with respect to K, achieving such a lower bound in online RL.

4.2. Analysis of PGAPI

We upper bound the optimality gap of PGAPI in the following theorem, whose proof sketch can be found in §I.1.

Theorem 4.2. *(Optimalty Gap of PGAPI). In Algorithm 3, we set* $\lambda = 1$, $\kappa = cR\sqrt{d\log(HdK)/\xi}$, $\alpha = (2\log(\mathrm{vol}(\mathcal{A}))/(H^2\sqrt{d}K))^{1/2}$, $\eta = 1/\sqrt{HK}$, *where* $c{>}0$ *is a constant. Under Assumption 2.1,* $\{\Gamma_h\}_{h=1}^H$ *constructed in §3.2.1 are* ξ*-uncertainty qualifiers defined in Definition 3.1. it holds with probability at least* $1-\xi$ *that*

$$\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}}, \widehat{\pi}) {\le} \mathcal{O}\big(\sqrt{H^4d^2/K}\big) {+} \Delta_{N_1} {+} \mathrm{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}}, \tag{30}$$

where $\widehat{\pi}$ *is the output policy of Algorithm 3,* $\mathrm{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}} {=} 2\sum_{h=1}^H \mathbb{E}_{\pi^{\mathrm{E}}}[\Gamma_h(s_h, a_h) \,|\, s_1 = x]$, *and* $\Delta_{N_1} = \mathcal{O}(\sqrt{H^3d^2/N_1}\log(N_1/\xi))$.

In Theorem 4.2, the first term on the right-hand side of (30) is an optimization error term, which is independent of both expert demonstration $\mathbb{D}^{\mathrm{E}}$ and the additional dataset $\mathbb{D}^{\mathrm{A}}$. The optimization error term decays at a rate of $K^{-1/2}$. The second term on the right-hand side of (30) is related to the MC estimation error and also occurs in the upper bound of the regret of OGAPI as in Theorem 4.1. The third term on the right-hand side of (30) is an intrinsic error $\mathrm{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}}$, which arises from the uncertainty of estimating Bellman equation (4) based on the additional dataset $\mathbb{D}^{\mathrm{A}}$. In the structure of the intrinsic error $\mathrm{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}}$, we note that the expectation is taken with respect to the trajectory induced by the expert policy π^{E}, which measures the quality of the additional dataset $\mathbb{D}^{\mathrm{A}}$ and is irrelevant to the training process. We clarify that the occurrence of $\mathrm{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}}$ implies that the optimality gap only depends on how well the additional dataset $\mathbb{D}^{\mathrm{A}}$ covers the trajectories of the expert policy π^{E} and it is not necessary to assume that the additional dataset $\mathbb{D}^{\mathrm{A}}$ is well-explored. Hence, Theorem 4.2 relies on no assumption on the coverage of the additional dataset $\mathbb{D}^{\mathrm{A}}$, such as uniformly lower bound of densities of visitation measures, the behavior policy to be upper bounded uniformly over the state-action space, the concentrability coefficients are uniformly upper bounded, or even the partial coverage assumption (Antos et al., 2007; Munos & Szepesvári, 2008; Yang et al., 2020b;a; Levine et al., 2020; Uehara et al., 2020; Siegel et al., 2020; Wang et al., 2020b; Zhang et al., 2020; Xu et al., 2020). We also highlight that Theorem 4.2 can be generalized to the case when the transition kernel is non-linear, only if we explicitly find proper uncertainty quantifiers $\{\Gamma_h\}_{h=1}^H$ satisfying Definition 3.1 for the estimated transition kernel.

Next we show that PGAPI is provably efficient and attains global convergence under the assumption that the additional dataset $\mathbb{D}^{\mathrm{A}}$ has sufficient coverage. We first impose such an assumption on the additional dataset $\mathbb{D}^{\mathrm{A}}$ as follows.

Assumption 4.3 (Sufficient Coverage)**.** The additional dataset $\mathbb{D}^{\mathrm{A}}$ has sufficient coverage with the expert policy π^{E}, that is, there exists an absolute constant $c^\dagger > 0$ such that the event $\mathcal{E}^\dagger = \big\{ \frac{1}{N_2}\sum_{\tau=1}^{N_2} \int_{\mathcal{S}} \phi(s_h^\tau, a_h^\tau, s')\phi(s_h^\tau, a_h^\tau, s')^\top \mathrm{d}s' \ge c^\dagger \cdot \mathbb{E}_{\pi^{\mathrm{E}}}\big[\int_{\mathcal{S}} \phi(s_h, a_h, s')\phi(s_h, a_h, s')^\top \mathrm{d}s'\big]$, for any $(s_1, h) \in \mathcal{S} \times [H]\big\}$ satisfies $\mathbb{P}_{\mathbb{D}}(\mathcal{E}^\dagger) \ge 1 - \xi/2$. Here the expectation $\mathbb{E}_{\pi^{\mathrm{E}}}[\cdot]$ is taken w.r.t. the trajectory induced by π^{E} and $\xi \in (0,1)$ is the confidence level.

Assumption 4.3 implies that the additional dataset $\mathbb{D}^{\mathrm{A}}$ with sufficient coverage covers the trajectories of the expert policy π^{E} averagely in the sense of the feature map outer product $\int_{\mathcal{S}} \phi(\cdot,\cdot,s')\phi(\cdot,\cdot,s')^\top \mathrm{d}s'$. We highlight that sufficient coverage does not assume that the additional dataset $\mathbb{D}^{\mathrm{A}}$ to be well-explored dataset (Zhang et al., 2020; Xu et al., 2020; Yang et al., 2020b;a; Levine et al., 2020), e.g. restricting that the densities of visitation measures of the behavior policy generating the dataset are uniformly lower bounded, i.e. $\inf_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \big[\rho_h^{\pi^{\mathrm{A}}}(s,a)\big] = c > 0$, where ρ_h^π is the density of visitation measure on $\mathcal{S}\times\mathcal{A}$ induced by policy π at the h-step and π^{A} is the policy of the experimenter who collected the additional dataset $\mathbb{D}^{\mathrm{A}}$. We also remark that sufficient coverage is a weaker restriction than the partial coverage assumption in offline RL (Uehara et al., 2020; Siegel et al., 2020; Wang et al., 2020b), which assumes that $\sup_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]} \big[\rho_h^{\pi^{\mathrm{E}}}(s,a)/\rho_h^{\pi^{\mathrm{A}}}(s,a)\big] = C^{\pi_{\mathrm{E}}} < \infty$.

Under Assumption 4.3, we present the following corollary, whose proof can be found in §I.3.

Corollary 4.4. *Under Assumption 4.3 and the same assumptions as in Theorem 4.2, it holds with probability at least* $1-\xi$ *that*

$$\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}}, \widehat{\pi}) {\le} \widetilde{\mathcal{O}}\big(\sqrt{H^4d^2/K} {+} \sqrt{H^4d^3/N_2} {+} \sqrt{H^3d^2/N_1}\big),$$

where $\widehat{\pi}$ *is the output of Algorithm 3.*

Corollary 4.4 proves that under Assumption 4.3 the intrinsic error $\mathrm{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}}$ in Theorem 4.2 decays at a rate of $N_2^{-1/2}$, showing that PGAPI attains global convergence. We remark that this result does not require the additional dataset $\mathbb{D}^{\mathrm{A}}$ to be well-explored and only relies on a much weaker assumption as sufficient coverage in Assumption 4.3. This improved result also responds to the information-theoretical lower bound $\Omega(H^2N_2^{-1/2})$ for offline policy evaluation (Duan et al., 2020). When K, N_1, and N_2 all tend to infinity, the optimality gap shrinks to zero as a negative square-root rate, meaning that the output policy has the same performance with the expert policy with respect to the reward set $\mathcal{R}$. More discussions about PGAPI can be found in §F, where we show that how pessimism guarantees the minimax utilization of the additional dataset $\mathbb{D}^{\mathrm{A}}$ and how the additional dataset $\mathbb{D}^{\mathrm{A}}$ contributes to our policy learning.

5. Experiment

Besides the theoretical analysis, we also conduct an experiment for PGAPI in the offline setting. Results show that PGAPI can converge fast and exceeds the performance of BC method. See Appendix G for detailed discussions and experimental results.

6. Conclusion

In this paper, we study provably efficient algorithms for GAIL in the online and offline setting with linear function approximation, where both the transition kernels and reward functions are linearly parameterized. We present a unified framework and specialize it as optimistic generative adversarial policy imitation (OGAPI) for online GAIL and pessimistic generative adversarial policy imitation (PGAPI) for offline GAIL, respectively. With linear function approximation, we derive the upper bound of the regret of OGAPI as $\widetilde{\mathcal{O}}(H^2 d^{3/2} K^{1/2} + K H^{3/2} d N_1^{-1/2})$ and the decomposition of optimality gap of PGAPI, without any assumption on the additional dataset. Facilitated with an additional dataset with sufficient coverage, we demonstrate that PGAPI also attains global convergence, achieving $\widetilde{\mathcal{O}}(H^2 d K^{-1/2} + H^2 d^{3/2} N_2^{-1/2} + H^{3/2} d N_1^{-1/2})$ optimality gap. Furthermore, we discuss that OGAPI can be reduced to an online RL algorithm whose online regret achieves an information-theoretic lower bound. Besides, we show that pessimism in PGAPI guarantees the minimax utilization of the additional dataset and how an additional dataset with sufficient coverage contributes to our policy learning. However, how to design provably efficient GAIL algorithms with general function approximation on both the transition kernels and reward functions still remains an open problem, which is a challenging but important future direction.

Acknowledgement

Zhaoran Wang acknowledges National Science Foundation (Awards 2048075, 2008827, 2015568, 1934931), Amazon, J.P. Morgan, and Two Sigma for their supports. Zhihan Liu would like to thank Miao Lu for valuable discussions.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *International Conference on Machine Learning*, 2011.

Abbeel, P. and Ng, A. Apprenticeship learning via inverse reinforcement learning. *International Conference on Machine Learning*, 2004.

Antos, A., Munos, R., and Szepesvári, C. Fitted q-iteration in continuous action-space mdps. 2007.

Argall, B. D., Chernova, S., Veloso, M., and Browning, B. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2009.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, 2020.

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

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 05 2003.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

Buckman, J., Gelada, C., and Bellemare, M. G. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

Cai, Q., Hong, M., Chen, Y., and Wang, Z. On the global convergence of imitation learning: A case for linear quadratic regulator. *arXiv preprint arXiv:1901.03674*, 2019.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, 2020.

Chang, J. D., Uehara, M., Sreenivas, D., Kidambi, R., and Sun, W. Mitigating covariate shift in imitation learning via offline data without great coverage. *arXiv preprint arXiv:2106.03207*, 2021.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051, 2019.

Chen, M., Wang, Y., Liu, T., Yang, Z., Li, X., Wang, Z., and Zhao, T. On computation and generalization of generative adversarial imitation learning. *ArXiv 2001.02792*, 2020.

Chen, Z., Zhang, X., Boedihardjo, A. P., Dai, J., and Lu, C.-T. Multimodal storytelling via generative adversarial imitation learning. In *International Joint Conference on Artificial Intelligence*, 2017.

Demiris*, Y. and Johnson, M. Distributed, predictive perception of actions: A biologically inspired robotics architecture for imitation and learning. *Connection Science*, 15(4):231–243, 2003.

Duan, Y., Jia, Z., and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, 2020.

Finn, C., Levine, S., and Abbeel, P. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*.

Fisac, J. F., Akametalu, A. K., Zeilinger, M. N., Kaynama, S., Gillula, J., and Tomlin, C. J. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7): 2737–2752, 2018.

Fujimoto, S., Conti, E., Ghavamzadeh, M., and Pineau, J. Benchmarking batch deep reinforcement learning algorithms. *arXiv preprint arXiv:1910.01708*, 2019a.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, 2019b.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Hazan, E. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Ho, J. and Ermon, S. Generative adversarial imitation learning. *arXiv preprint arXiv:1606.03476*, 2016.

Hussein, A., Gaber, M. M., Elyan, E., and Jayne, C. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

Jalali, S. M. J., Kebria, P. M., Khosravi, A., Saleh, K., Nahavandi, D., and Nahavandi, S. Optimal autonomous driving through deep imitation learning and neuroevolution. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp. 1215–1220. IEEE, 2019.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, 2018.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Annual Conference on Learning Theory*, 2019.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, 2020.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, 2021.

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. Information theoretic regret bounds for online nonlinear control. In *Advances in Neural Information Processing Systems*, 2020.

Kakade, S. M. A natural policy gradient. *Advances in Neural Information Processing Systems*, 2001.

Kebria, P. M., Khosravi, A., Salaken, S. M., and Nahavandi, S. Deep imitation learning for autonomous vehicles based on convolutional neural networks. *IEEE/CAA Journal of Automatica Sinica*, 7(1):82–95, 2019.

Kuefler, A., Morton, J., Wheeler, T., and Kochenderfer, M. Imitating driver behavior with generative adversarial networks. In *IEEE Intelligent Vehicles Symposium (IV)*, pp. 204–211. IEEE, 2017.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative Q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

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

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

Merel, J., Tassa, Y., TB, D., Srinivasan, S., Lemmon, J., Wang, Z., Wayne, G., and Heess, N. Learning human behaviors from motion capture by adversarial imitation. *arXiv preprint arXiv:1707.02201*, 2017.

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

Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.

Neu, G. and Szepesvári, C. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2007.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 2000.

Rajaraman, N., Yang, L., Jiao, J., and Ramchandran, K. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33: 2914–2924, 2020.

Rajaraman, N., Han, Y., Yang, L. F., Ramchandran, K., and Jiao, J. Provably breaking the quadratic error compounding barrier in imitation learning, optimally. *arXiv preprint arXiv:2102.12948*, 2021.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pp. 5478–5486, 2019.

Ross, S. and Bagnell, D. Efficient reductions for imitation learning. In *International Conference on Artificial Intelligence and Statistics*, 2010.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shani, L., Efroni, Y., and Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020a.

Shani, L., Efroni, Y., Rosenberg, A., and Mannor, S. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613, 2020b.

Shani, L., Zahavy, T., and Mannor, S. Online apprenticeship learning. *arXiv preprint arXiv:2102.06924*, 2021.

Siegel, N. Y., Springenberg, J. T., Berkenkamp, F., Abdolmaleki, A., Neunert, M., Lampe, T., Hafner, R., Heess, N., and Riedmiller, M. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

Syed, U. and Schapire, R. E. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems*, 2008.

Syed, U., Bowling, M., and Schapire, R. Apprenticeship learning using linear programming. In *International Conference on Machine Learning*, 2008.

Tsurumine, Y., Cui, Y., Yamazaki, K., and Matsubara, T. Generative adversarial imitation learning with deep P-network for robotic cloth manipulation. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pp. 274–280. IEEE, 2019.

Uehara, M. and Sun, W. Pessimistic model-based offline RL: Pac bounds and posterior sampling under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and Q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, 2020.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020a.

Wang, Z., Novikov, A., Zolna, K., Springenberg, J. T., Reed, S., Shahriari, B., Siegel, N., Merel, J., Gulcehre, C., Heess, N., et al. Critic regularized regression. *arXiv preprint arXiv:2006.15134*, 2020b.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021a.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *arXiv preprint arXiv:2106.04895*, 2021b.

Xu, T., Li, Z., and Yu, Y. Error bounds of imitating policies and environments. In *Advances in Neural Information Processing Systems*, 2020.

Yang, L. and Wang, M. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, 2019a.

Yang, L. and Wang, M. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, 2019b.

Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756, 2020.

Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*, 2020a.

Yang, Z., Jin, C., Wang, Z., Wang, M., and Jordan, M. I. Bridging exploration and general function approximation in reinforcement learning: Provably efficient kernel and neural value iterations. *arXiv preprint arXiv:2011.04622*, 2020b.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., and Finn, C. Combo: Conservative offline model-based policy optimization. *arXiv preprint arXiv:2102.08363*, 2021.

Zanette, A. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. In *International Conference on Machine Learning*, 2021.

Zhang, J. and Wu, F. A method of offline reinforcement learning virtual reality satellite attitude control based on generative adversarial network. *Wireless Communications and Mobile Computing*, 2021, 2021.

Zhang, Y., Cai, Q., Yang, Z., and Wang, Z. GAIL with neural network parameterization: Global optimality and convergence rate. In *International Conference on Machine Learning*, 2020.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, 2021.

Zolna, K., Novikov, A., Konyushkova, K., Gulcehre, C., Wang, Z., Aytar, Y., Denil, M., de Freitas, N., and Reed, S. Offline learning from demonstrations and unlabeled experience. *arXiv preprint arXiv:2011.13885*, 2020.

A. Notations

We denote by $a = \mathcal{O}(b)$ if there exists an absolute constant c such that $a \leq cb$ when a and b are both large enough. We use $\widetilde{\mathcal{O}}(\cdot)$ to hide the constants term and logarithmic terms in $\widetilde{\mathcal{O}}(\cdot)$. We denote by $[N] = \{1, \ldots, N\}$. We also denote by $a = \Omega(b)$ if there exists an absolute constant c such that $a \geq cb$ when a and b are both large enough. We denote by $\|\cdot\|_2$ the ℓ_2-norm of a vector and denote by $\|\cdot\|_A$ the spectral norm of a matrix A. We denote by $\Delta(\mathcal{X})$ the set of probability distributions on a set $\mathcal{X}$ and correspondingly define $\Delta(\mathcal{A} \,|\, \mathcal{S}, H) = \{\{\pi_h(\cdot \,|\, \cdot)\}_{h\in[H]} : \pi_h(\cdot \,|\, s) \in \Delta(\mathcal{A})$ for any $(s, h) \in \mathcal{S} \times [H]\}$ for all set $\mathcal{S}$ and $H \in \mathbb{N}_+$. For $p_1, p_2 \in \Delta(\mathcal{A})$, we denote by $D_{\mathrm{KL}}(p_1 \| p_2)$ the KL-divergence, that is, $D_{\mathrm{KL}}(p_1 \| p_2) = \int_{\mathcal{A}} p_1(a) \log \frac{p_1(a)}{p_2(a)} \mathrm{d}a$. And $\langle \cdot, \cdot \rangle_{\mathcal{A}}$ is the inner product taken over the action space $\mathcal{A}$. We also denote by δ_x Dirac function centered at x. We denote by $\mathrm{Vol}(\mathcal{X})$ by the measure of set $\mathcal{X}$. We also denote by $\{x\}_+ = \max\{x, 0\}$.

B. More Discussions about Related Works

Our work is also related to IRL, (Abbeel & Ng, 2004; Neu & Szepesvári, 2007; Syed et al., 2008; Syed & Schapire, 2008) study the convergence of IRL in the tabular case, while they require to solve an RL subproblem every iteration, inefficiently.

Besides, our work is related to the vast body of existing literature on online RL cooperated with optimism (Auer et al., 2002; 2009; Azar et al., 2017; Jin et al., 2018; 2019; Yang & Wang, 2020), offline RL (Fujimoto et al., 2019b; Kumar et al., 2020; Fujimoto et al., 2019a; Duan et al., 2020; Levine et al., 2020; Jin et al., 2021), policy optimization (Beck & Teboulle, 2003; Hazan, 2019; Cai et al., 2020; Nemirovskij & Yudin, 1983), adversarial MDP (Shani et al., 2020b; Rosenberg & Mansour, 2019; Jin et al., 2020), and linear function approximation (Duan et al., 2020; Bradtke & Barto, 1996; Yang & Wang, 2019a; Jin et al., 2019; Ayoub et al., 2020; Yang & Wang, 2019b; Zhou et al., 2021) while they study minimization or maximization problem with known reward through value-based or policy-based method, instead of minimax problem with respect to policy and reward function as GAIL.

Our work is related to a line of study on pessimism. Specifically, the uncertainty quantification for estimated model in PGAPI is motivated by the pessimism in offline RL (Chen & Jiang, 2019; Xie et al., 2021a;b; Kumar et al., 2020; Jin et al., 2021; Liu et al., 2020; Yu et al., 2020; 2021; Buckman et al., 2020; Rashidinejad et al., 2021; Uehara & Sun, 2021). (Liu et al., 2020) propose a pessimistic variant of fitted Q-learning algorithm (Antos et al., 2007; Munos & Szepesvári, 2008) achieving the optimal policy within a restricted class of policies without assuming the dataset to be well-explored. (Jin et al., 2021) propose a provably efficient algorithm with the spirit of pessimism to solve offline RL with linear function approximation, under no coverage assumption on the dataset. (Xie et al., 2021a) propose a refined pessimistic estimate and obtain a tighter suboptimality in d compared with (Jin et al., 2021). (Rashidinejad et al., 2021) study the offline RL in the tabular case through lower confidence bound (LCB), relining on the partial coverage assumption on the dataset. (Uehara & Sun, 2021) analyze the constrained pessimistic policy optimization with general function approximation and with the partial coverage assumption of the dataset, then they specialize the case in the KNR setting and give a refined upper bound. The importance of pessimism in offline RL is characterized by (Buckman et al., 2020; Zanette, 2021) through discussing the lower bound of offline RL when the dataset has no restriction.

C. Pseudocode of OGPAI

Algorithm 2 Optimistic Generative Adversarial Policy Imitation (OGAPI)

```
Input: Expert demonstration $\mathbb{D}^{\mathrm{E}}$, scaling factor $\kappa$, and step size $\eta$ and $\alpha$.
Initialize $\{\widehat{Q}_h^0\}_{h\in[H]}$ as zero functions over $\mathcal{S} \times \mathcal{A}$, $\{\pi_h^0\}_{h\in[H]}$ as uniform distributions over $\mathcal{A}$, $\mu^1 = \{\mu_h^1\}_{h\in[H]}$ as zero vectors, and $\{\widehat{V}_{H+1}^k\}_{k\in\{0,1,\ldots,K\}}$ as zero functions over $\mathcal{S}$.
for $k = 1, \ldots, K$ do
  for $h = 1, \ldots, H$ do
    $\pi_h^k(\cdot\,|\,\cdot) \propto \pi_h^{k-1}(\cdot\,|\,\cdot) \cdot \exp\{\alpha \cdot \widehat{Q}_h^{k-1}(\cdot,\cdot)\}$.          //Policy Improvement
  end for
  Rollout a trajectory $\{(s_h^k, a_h^k)\}_{h\in[H]}$ following $\pi^k$.
  for $h = H, \ldots, 1$ do
    Set $\{\widehat{\mathcal{P}}_h^k\}_{h=1}^H$ and $\{\Gamma_h^k\}_{h=1}^H$ via (15) and (18), respectively.          //Policy Evaluation
    $\widehat{Q}_h^k(\cdot,\cdot) \leftarrow \min\{(r_h^k + \widehat{\mathcal{P}}_h^k \widehat{V}_{h+1}^k + \Gamma_h^k)(\cdot,\cdot), (H-h+1)\sqrt{d}\}_+$.
    $\widehat{V}_h^k(\cdot) \leftarrow \langle \widehat{Q}_h^k(\cdot,\cdot), \pi_h^k(\cdot\,|\,\cdot)\rangle_{\mathcal{A}}$.
  end for
  Set $\{\nabla_{\mu_h} \widetilde{J}(\pi^{\mathrm{E}}, r^\mu)\}_{h\in[H]}$ via (25).          //Reward Update
  for $h = 1, \ldots, H$ do
    $\widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k) \leftarrow \nabla_{\mu_h} \widetilde{J}(\pi^{\mathrm{E}}, r^\mu)\,|_{\mu=\mu^k} - \psi(s_h^k, a_h^k)$.
    $\mu_h^{k+1} \leftarrow \mathrm{Proj}_B(\mu_h^k + \eta \widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k))$.
  end for
end for
```

D. Pseudocode of PGPAI

Algorithm 3 Pessimistic Generative Adversarial Policy Imitation (PGAPI)

```
Input: Expert demonstration $\mathbb{D}^{\mathrm{E}}$, the additional dataset $\mathbb{D}^{\mathrm{A}}$, step size $\eta, \alpha$
Initialize $\{\widehat{Q}_h^0\}_{h\in[H]}$ as zero functions, $\{\pi_h^0\}_{h\in[H]}$ as uniform distribution, $\mu^1 = \{\mu_h^1\}_{h\in[H]}$ as zero vectors, and $\{\widehat{V}_{H+1}^k\}_{k\in\{0,1,\ldots,K\}}$ as zero functions over $\mathcal{S}$.
Construct $\{\widehat{\mathcal{P}}_h\}_{h\in[H]}$ and $\{\Gamma_h\}_{h\in[H]}$ from $\mathbb{D}^{\mathrm{A}}$ via (39) and (36), respectively.          //Initial Construction
for $k = 1, \ldots, K$ do
  for $h = 1, \ldots, H$ do
    $\pi_h^k(\cdot\,|\,\cdot) \propto \pi_h^{k-1}(\cdot\,|\,\cdot) \exp\{\alpha \cdot \widehat{Q}_h^{k-1}(\cdot,\cdot)\}$.          //Policy Improvement
  end for
  for $h = H, \ldots, 1$ do
    $\widehat{Q}_h^k(\cdot,\cdot) \leftarrow \max\{(r_h^k + \widehat{\mathcal{P}}_h \widehat{V}_{h+1}^k - \Gamma_h)(\cdot,\cdot), 0\}$.          //Policy Evaluation
    $\widehat{V}_h^k(\cdot) \leftarrow \langle \widehat{Q}_h^k(\cdot,\cdot), \pi_h^k(\cdot\,|\,\cdot)\rangle_{\mathcal{A}}$.
  end for
  Construct $\{\nabla_{\mu_h} \widetilde{J}(\pi^{\mathrm{E}}, r^\mu)\}_{h\in[H]}$ via (25).          //Reward Update
  Construct $\{\nabla_{\mu_h} \widehat{J}(\pi^k, r^\mu)\}_{h\in[H]}$ via Proposition D.1.
  for $h = 1, \ldots, H$ do
    $\nabla_{\mu_h} \widehat{L}(\pi^k, \mu^k) \leftarrow \nabla_{\mu_h} \widetilde{J}(\pi^{\mathrm{E}}, r^\mu)\,|_{\mu=\mu^k} - \nabla_{\mu_h} \widehat{J}(\pi^k, r^\mu)\,|_{\mu=\mu^k}$.
    $\mu_h^{k+1} \leftarrow \mathrm{Proj}_B[\mu_h^k + \eta \nabla_{\mu_h} \widehat{L}(\pi^k, \mu^k)]$.
  end for
end for
Output: $\widehat{\pi} = \mathrm{Unif}(\{\pi^k\}_{k\in[K]})$.
```

Proposition D.1. *If we define $\{\widehat{Q}_h^{k,r^\mu}\}_{h\in[H]}$ and $\{\widehat{V}_{h+1}^{k,r^\mu}\}_{h\in[H]}$ as*

$$\begin{aligned} &\widehat{V}_{H+1}^{k,r^\mu}(\cdot) = 0, \\ &\widehat{Q}_h^{k,r^\mu}(\cdot,\cdot) = \max\big\{(r_h^\mu + \widehat{\mathcal{P}}_h \widehat{V}_{h+1}^{k,r^\mu} - \Gamma_h)(\cdot,\cdot), 0\big\}, \\ &\widehat{V}_h^{k,r^\mu}(\cdot,\cdot) = \big\langle \widehat{Q}_h^{k,r^\mu}(\cdot,\cdot), \pi_h^k(\cdot\,|\,\cdot)\big\rangle_{\mathcal{A}}, \end{aligned} \tag{31}$$

for all $h\in[H]$ and $\mu\in S$. It suggests that $\widehat{Q}_h^k = \widehat{Q}_h^{k,r^k}$ and $\widehat{V}_h^k = \widehat{V}_h^{k,r^k}$, where $\widehat{Q}_h^k$ and $\widehat{V}_h^k$ are constructed in the policy evaluation stage in PGAPI (Algorithm 3 Lines 8–11). We can solve $\nabla_{\mu_h}\widehat{V}_1^{k,r^\mu}(x)$ recursively as follows,

$$\nabla_{\mu_h}\widehat{V}_t^{k,r^\mu}(s_t) = \begin{cases} \Big\langle \pi_h^k(\cdot\,|\,s_t) g_t^k(s_t,\cdot), \big[\widehat{\mathcal{P}}_t(\nabla_{\mu_h}\widehat{V}_{t+1}^{k,r^\mu})\big](s_t,\cdot)\Big\rangle_{\mathcal{A}} & \text{if } 1\le t\le h-1, \\ \big\langle \pi_h^k(\cdot\,|\,s_h) g_h^k(s_h,\cdot), \nabla_{\mu_h} r_h^\mu(s_h,\cdot)\big\rangle_{\mathcal{A}} & \text{if } t=h, \end{cases}$$

where $s_1 = x$, $[\widehat{\mathcal{P}}_h f](s,a)$ is a shorthand of $\int_{\mathcal{S}} f(s')\widehat{\mathcal{P}}_h(s'\,|\,s,a)\mathrm{d}s'$ and $g_h^k : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ is defined as

$$g_h^k(s,a) = \mathbf{1}\big\{\widehat{Q}_h^{k,r^\mu}(s,a) > 0\big\}, \tag{32}$$

for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. Here $\mathbf{1}\{\cdot\}$ is the indicator function.

Proof. Taking gradient toward μ_h and applying chain rule on (31), we can obtain Proposition D.1. □

E. Initial Construction of PGAPI

Now, we construct the estimated transition kernels and uncertainty quantifiers as follows. Specifically, we first construct the initial estimated transition kernels $\widetilde{\mathcal{P}} = \{\widetilde{\mathcal{P}}_h\}_{h\in[H]}$ as

$$\widetilde{\mathcal{P}}_h(s'\,|\,s,a) = \phi(s,a,s')^\top \widetilde{\theta}_h, \tag{33}$$

for any $(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$, where $\widetilde{\theta}_h$ is the solution of the following optimization problem,

$$\min_{\theta\in\mathbb{R}^d}\sum_{\tau=1}^{N_2}\int_{\mathcal{S}}\big|\phi(s_h^\tau,a_h^\tau,s')^\top\theta - \delta_{s_{h+1}^\tau}(s')\big|^2\,\mathrm{d}s' + \lambda\|\theta\|_2^2. \tag{34}$$

Here $\lambda>0$ is the regularization parameter and $\delta_x(y)$ is Dirac function. By solving (34), we obtain the closed-form solution of $\widetilde{\theta}_h$ as follows,

$$\widetilde{\theta}_h = \Lambda_h^{-1}\sum_{\tau=1}^{N_2}\phi(s_h^\tau,a_h^\tau,s_{h+1}^\tau), \tag{35}$$

where $\Lambda_h = \sum_{\tau=1}^{N_2}\int_{\mathcal{S}}\phi(s_h^\tau,a_h^\tau,s')\phi(s_h^\tau,a_h^\tau,s')^\top\mathrm{d}s' + \lambda I$. Given (33) and (35), we further construct ξ-uncertainty quantifiers $\{\Gamma_h\}_{h\in[H]}$ as follows,

$$\Gamma_h(s,a) = H\sqrt{d}\int_{\mathcal{S}}\Gamma_h^{\mathcal{P}}(s,a,s')\mathrm{d}s', \tag{36}$$

where $\Gamma_h^{\mathcal{P}}(s,a,s') = \min\big\{\kappa\cdot\|\phi(s,a,s')\|_{\Lambda_h^{-1}}, 1\big\}$. Here $\kappa>0$ is a scaling parameter. In the following lemma, we show that $\{\Gamma_h\}_{h\in[H]}$ in (36) are ξ-uncertainty qualifiers for $\widetilde{\mathcal{P}}$ in (33) if κ is properly chosen.

Lemma E.1. *In (36), we set $\lambda=1$ and $\kappa = c\cdot R\sqrt{d\log(dHN_2/\xi)}$, where $c>0$ is an absolute constant and $\xi\in(0,1)$ is the confidence parameter. Under Assumption 2.1, $\{\Gamma_h\}_{h\in[H]}$ in (36) are ξ-uncertainty qualifiers for $\widetilde{\mathcal{P}}$, defined in Definition 3.1.*

Proof. See Appendix K.1 for a detailed proof. □

However, given $(s,a)\in\mathcal{S}\times\mathcal{A}$, the initial estimated transition kernels $\widetilde{\mathcal{P}}_h(\cdot\,|\,s,a)$ in (33) is not guaranteed to lie within $\Delta(\mathcal{S})$. Different from OGAPI, we are incapable to update reward functions based on the newly sampled trajectory in offline GAIL. This difference is crucial to the analysis of PGAPI (Theorem 4.2), which relies on the fact that estimated GAIL objective function $\widehat{L}(\pi,\mu)$ defined in (27) is concave for μ (we prove it in Lemma I.3). To address this issue, we define a feasible estimation parameter domain Θ and choose the estimated transition kernel parameter $\widehat{\theta}$ from the feasible domain Θ (Zhou et al., 2021). Formally, we take $\Theta=\Theta_1\cap\Theta_2$ with Θ_1 and Θ_2 defined as follows,

$$\begin{aligned}\Theta_1=\big\{\widehat{\theta}\colon&\text{ if }\mathcal{E}\text{ holds, then it satisfies that }|\widehat{\mathcal{P}}_h\widehat{V}(s,a)-\widetilde{\mathcal{P}}_h\widehat{V}(s,a)|\le\Gamma_h(s,a)\\&\text{for any }(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]\text{ and any }\widehat{V}:\mathcal{S}\to[0,H\sqrt{d}]\big\},\\\Theta_2=\big\{\widehat{\theta}\colon&\,\widetilde{\mathcal{P}}_h(\cdot\,|\,s,a)\in\Delta(\mathcal{S})\text{ for any }(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]\big\},\end{aligned}\tag{37}$$

where $\mathcal{E}$ and $\{\widetilde{\mathcal{P}}_h\}_{h\in[H]}$ are defined in Definition 3.1 and (33), respectively. We remark that under Assumption 2.1, the true transition kernel parameter $\theta=\{\theta_h\}_{h\in[H]}$ lies within the feasible estimation parameter domain Θ, which implies that Θ is not empty. Thus, similar to (34) but enforcing the estimated transition kernel parameter to lie within Θ, we define $\widehat{\theta}=\{\widehat{\theta}_h\}_{h\in[H]}$ as follows,

$$\widehat{\theta}_h=\operatorname*{argmin}_{\theta\in\Theta}\sum_{\tau=1}^{N_2}\int_{\mathcal{S}}\big|\phi(s_h^\tau,a_h^\tau,s')^\top\theta-\delta_{s_{h+1}^\tau}(s')\big|^2\,\mathrm{d}s'+\lambda\|\theta\|_2^2,\tag{38}$$

where the minimization is taken over Θ. Similarly, we construct the estimated transition kernel $\widehat{\mathcal{P}}=\{\widehat{\mathcal{P}}_h\}_{h\in[H]}$ as follows,

$$\widehat{\mathcal{P}}_h(s'\,|\,s,a)=\phi(s,a,s')^\top\widehat{\theta}_h,\tag{39}$$

for any $(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]$, where $\widehat{\theta}_h$ is defined in (38).

F. More Discussions about PGAPI

Pessimism Guarantees Minimax Utilization. With a well-explored and large enough dataset, the full information about the transition kernel can be extracted by the agent and supports the agent to make correct decision. But when we assert no restriction on the dataset, it is challenging to do the same because of the distribution shift and extrapolation error on the states and actions that are less covered by the dataset. This problem has been studied widely in offline RL (Fujimoto et al., 2019a; Kumar et al., 2020; Fujimoto et al., 2019b; Levine et al., 2020; Jin et al., 2021) and (Wang et al., 2020a) even propose that the lower bound of offline RL can grow exponentially with the horizon under linear approximation and no assumption on the dataset. Hence how to cooperate the additional dataset $\mathbb{D}^{\mathrm{A}}$ to aid the agent to improve the performance in offline GAIL is also difficult since the additional dataset $\mathbb{D}^{\mathrm{A}}$ is not assumed to be well-explored. Inspired by the spirit of being conservative in offline RL (Fujimoto et al., 2019b; Kumar et al., 2020; Jin et al., 2021), we propose a pessimistic variant of policy optimization in the policy update stage of PGAPI (Lines 5–11 of Algorithm 3), which ensures that PGAPI utilize the information of the additional dataset $\mathbb{D}^{\mathrm{A}}$ in the sense of minimax optimality. To illustrate it, we present the following proposition, which is adapted from Theorem 4.7 in (Jin et al., 2021).

Proposition F.1 (Minimax Optimality in Utilizing Additional Dataset)**.** *For the output policy* $\texttt{Algo}(\overline{\mathbb{D}})$ *of any offline algorithm only based on the dataset* $\overline{\mathbb{D}}$*, there exists a linear kernel MDP* $\mathcal{M}$ $(\mathcal{S},\mathcal{A},H,\mathcal{P},r)$ *with an initial state* $x\in\mathcal{S}$*, a dataset* $\overline{\mathbb{D}}$ *compliant with* $\mathcal{M}$*, and a reward set* $\mathcal{R}$ *, such that*

$$\max_{\pi^{\mathrm{E}}\in\Delta(\mathcal{A}\,|\,\mathcal{S},H)}\mathbb{E}_{\overline{\mathbb{D}}}\bigg[\frac{\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}},\texttt{Algo}(\overline{\mathbb{D}}))}{\text{Information}_{\overline{\mathbb{D}}}^{\pi^{\mathrm{E}}}}\bigg]\ge c,$$

where $c>0$ *is an absolute constant,* $\mathbb{E}_{\overline{\mathbb{D}}}[\cdot]$ *is taken expectation with respect to randomness of the dataset* $\overline{\mathbb{D}}$*, and* $\text{Information}_{\overline{\mathbb{D}}}^{\pi^{\mathrm{E}}}$ *is defined as*

$$\text{Information}_{\overline{\mathbb{D}}}^{\pi^{\mathrm{E}}}=(\mathit{Vol}(\mathcal{S}))^{-1}\cdot\mathbb{E}_{\pi^{\mathrm{E}}}\Big[\sum_{h=1}^{H}\int_{\mathcal{S}}\|\phi(s_h,a_h,s')\|_{\Lambda_h^{-1}}\mathrm{d}s'\Big|\,s_1=x\Big],$$

where Λ_h *is only determined by the dataset* $\overline{\mathbb{D}}$ *and takes the same form as in* (35)*.*

Proof. See §I.2 for a detailed proof. □

According to Proposition F.1, if we consider the additional dataset $\mathbb{D}^{\mathrm{A}}$ as $\overline{\mathbb{D}}$, it reveals that $\mathrm{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}}$ in the upper bound of optimality gap of PGAPI (Theorem 4.2) matches the lower bound up to H, $\sqrt{d}$, $\mathrm{Vol}(\mathcal{S})$, and the scaling parameter κ defined in (36). Though we do not assume any restriction on the additional dataset $\mathbb{D}^{\mathrm{A}}$, owing to the pessimism principle, PGAPI ensures the good utilization of the additional dataset $\mathbb{D}^{\mathrm{A}}$ even in the worst case.

The Additional dataset Contributes. We illustrate the contribution of the additional dataset $\mathbb{D}^{\mathrm{A}}$ by considering the following two cases.

1. Without accessing an additional dataset, PGAPI is also applicable by simply treating the given expert demonstration $\mathbb{D}^{\mathrm{E}}$ as the additional dataset $\mathbb{D}^{\mathrm{A}}$ in PGAPI (Algorithm 3), that is, $\mathbb{D}^{\mathrm{A}} = \mathbb{D}^{\mathrm{E}}$, which satisfies Assumption 4.3. If we set $K = \Omega(N_1)$, then by Corollary 4.4, we have

$$\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}}, \widehat{\pi}) = \widetilde{\mathcal{O}}(H^2 d^{3/2} N_1^{-1/2}). \tag{40}$$

2. If we have access to a large enough additional dataset with sufficient coverage, taking $N_2 = \Omega(d^2 H N_1)$ for instance, then by setting $K = \Omega(N_2)$, we upper bound the optimality gap of PGAPI as follows,

$$\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}}, \widehat{\pi}) = \widetilde{\mathcal{O}}(H^{3/2} d N_1^{-1/2}). \tag{41}$$

By comparing (40) and (41), we observe that the additional dataset $\mathbb{D}^{\mathrm{A}}$ helps decrease the dependency for H and d in the optimality gap by $H^{1/2}$ and $d^{1/2}$. It implies that we can use a much smaller expert demonstration $\mathbb{D}^{\mathrm{E}}$ to learn a policy as good as the expert policy π^{E}, especially when horizon H and feature space dimension d are sufficiently large. This improvement is meaningful in the imitation learning tasks, such as autonomous driving and robotics (Demiris* & Johnson, 2003; Hussein et al., 2017; Kebria et al., 2019; Jalali et al., 2019).

G. Experiment

To verify our theoretical analysis of PGAPI, we provide the experiment of PGAPI here, choosing the simulation environment as a MDP with a finite state space, a finite action space, and a linear reward function. Next we will first provide the descriptions of environment setup, then introduce the implementation details, and finally discuss the results. The codes are available on https://github.com/YSLIU627/Adversarial-Policy-Imitation-with-LFA.

Environment Setup. In the following experiment, we consider a MDP with a linear reward function, which is adapted from a simple 3×3 GridWorld deterministic environment. The original GridWorld is a $n \times n$ grid network with a state space $\mathcal{S} = \{(i,j)\}_{i,j=1}^{n}$, an action space $\mathcal{A} = \{\text{stay, up, down, left, right}\}$, and a horizon $H = n^2$. The agent always starts from $x_0 = (1,1)$ and collects rewards by taking actions. Whenever the agent reaches (n,n), she gets reward 1; whenever the agent reaches $(n,1)$, she gets reward 0; otherwise, she gets reward 0.1.

In our test environment, we let $n = 3$ and make two adaptations from the original GridWorld: (i) At each time timestep, the agent will get stuck on the current state for one timestep with probability 0.1 and move successfully according to the original GridWorld with probability 0.9, which induces the *transition kernel*. (ii) *The reward function* $r(s,a)$ is defined by $r(s,a) = \psi(s,a)^\top \mu$, where the reward parameter μ is a $|\mathcal{S}||\mathcal{A}|$-dimensional vector and a known feature map ψ maps from $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}^d$. We design $\psi(s,a)$ by first adding 0.1 to the each entry of the canonical basis $\mathbf{e}_{(s,a)}$ of the space $\mathcal{S} \times \mathcal{A}$ and then normalizing it. Since each entry of $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ corresponds to a state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, the reward parameter $\mu \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ can be determined by the reward function of the original GridWorld.

We can verify that our test environment satisfies Assumption 2.1. Given the test environment, we solve the expert policy π^{E} by conducting a model-based reinforcement learning method with a known transition kernel and reward function for 20 iterations such that π^{E} has converged to a near optimal policy. The expert demonstration $\mathbb{D}^{\mathrm{E}}$ is obtained by sampling 5 trajectories from π^{E} while the additional dataset $\mathbb{D}^{\mathrm{A}}$ is collected by sampling 1000 trajectories from the policy which samples uniformly random actions.

Implementation of Algorithms. Our proposed method PGAPI are implemented according to Algorithm 3, where we can simplify the forms of uncertainty quantification in the tabular setting: $\Gamma_h(s,a) = CH|\mathcal{S}|\sqrt{\log(H|\mathcal{S}||\mathcal{A}|/\delta)/\big(N(h,s,a)\vee 1\big)}$. Here C is a hyper-parameter, δ is the confidence level, and $N(h,s,a)$ is the visitation count of (s,a) within the additional dataset $\mathbb{D}^{\mathrm{A}}$ at the timestep h. We implement BC on a given dataset by $\pi_h^{BC}(a\,|\,s) = \mathbf{1}\{N(h,s)=0\}/|\mathcal{A}| + \mathbf{1}\{N(h,s)>0\}N(h,s,a)/N(h,s)$, where $\mathbf{1}\{\cdot\}$ is the indicator function, $N(h,s,a)$ and $N(h,s)$ are the visitation count of (s,a) and s at the timestep h within the given dataset , respectively.

Results and Conclusions. We conduct our proposed method PGAPI for 20 iterations and compare the average return of PGAPI with the performance of expert π^{E}, BC method on $\mathbb{D}^{\mathrm{E}}$, and BC method on the mixture of $\mathbb{D}^{\mathrm{E}}$ and $\mathbb{D}^{\mathrm{A}}$. From the result in Figure 1, we show that PGAPI converges fast and exceeds the performance of BC method conducted on $\mathbb{D}^{\mathrm{E}}$ after the sixth iteration, while BC conducted on the mixture of $\mathbb{D}^{\mathrm{E}}$ and $\mathbb{D}^{\mathrm{A}}$ has the nearly the same performance as uniformly random policy. Although the additional dataset $\mathbb{D}^{\mathrm{A}}$ only involves the trajectories of uniformly random policy, PGAPI utilizes it effectively given only very limited expert trajectories.

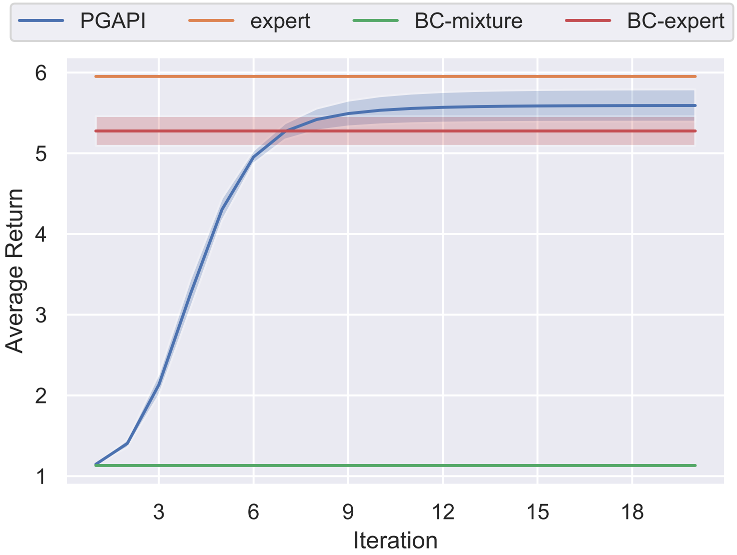

Figure 1. Average return of PGAPI (our proposed method), expert, BC method conducted on the mixed dataset, and BC conducted method on the expert demonstration. Results are averaged for 5 random seeds.

H. Proof Sketch: Analysis of OGAPI

H.1. Proof of Theorem 4.1

Proof. Recall the definition of regret in (7) and GAIL objective function $L(\pi,\mu)$ in (6), we decompose the regret as follows,

$$\begin{aligned}\mathrm{Regret}(K) &= \sup_{\mu\in S}\sum_{k=1}^{K} L(\pi^k,\mu)\\ &\le \underbrace{\sum_{k=1}^{K}[J(\pi^{\mathrm{E}},r^k) - J(\pi^k,r^k)]}_{\text{(A)}} + \underbrace{\sup_{\mu\in S}\sum_{k=1}^{K}\big[L(\pi^k,r^\mu) - L(\pi^k,r^k)\big]}_{\text{(B)}}.\end{aligned} \tag{42}$$

The intuition of decomposition in (42) is to respectively deal with regret occurring in the stage of policy update and reward update, which are denoted by term (A) and term (B).

Upper bound of term (A) in (42). In what follows, we upper bound term (A) in (42). For the simplicity of later discussion, we define the model prediction error for estimating Bellman equation (4) in the h-th step of k-th episode in Algorithm 3 with reward function r^μ as follows,

$$\iota_h^k(s,a) := r_h^k(s,a) + [\mathcal{P}_h \widehat{V}_{h+1}^k](s,a) - \widehat{Q}_h^k(s,a), \tag{43}$$

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

First, we introduce a regret decomposition lemma to decompose term (A) in (42).

Lemma H.1 (Regret Decomposition for Policy Update)**.** *It holds for any initial state $x \in \mathcal{S}$ that*

$$\begin{aligned} \sum_{k=1}^{K} \big(V_{1,\pi^{\mathrm{E}}}^{r^k}(x) - V_{1,\pi^k}^{r^k}(x)\big) =& \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^{\mathrm{E}}}\big[\langle \widehat{Q}_h^k(s_h,\cdot), \pi_h^k(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle \,\big|\, s_1 = x\big] \\ &+ \mathcal{M}_{K,H,2} + \sum_{k=1}^{K}\sum_{h=1}^{H}\Big(\mathbb{E}_{\pi^{\mathrm{E}}}[\iota_h^k(s_h^k, a_h^k)|s_1 = x] - \iota_h^k(s_h^k, a_h^k)\Big). \end{aligned} \tag{44}$$

Here ι_h^k is the model prediction error defined in (43), and $\{\mathcal{M}_{K,H,m}\}_{(k,h,m)\in[K]\times[H]\times[2]}$ is a martingale adapted to the filtration $\{\mathcal{F}_{k,h,m}\}_{(k,h,m)\in[K]\times[H]\times[2]}$, with respect to the timestep index $t(k,h,m) = (k-1)\cdot 2H + (h-1)\cdot 2 + m$.

Proof. See Appendix J.1 for a detailed proof. □

Lemma H.1 shows that term (A) in (42) can be decomposed into three terms as follows,

$$\begin{aligned} \text{(A)} &= \sum_{k=1}^{K} \big(V_{1,\pi^{\mathrm{E}}}^{r^k}(x) - V_{1,\pi^k}^{r^k}(x)\big) \\ &= \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^{\mathrm{E}}}\big[\langle \widehat{Q}_h^k(s_h,\cdot), \pi_h^{\mathrm{E}}(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle \big|\, s_1 = x\big]}_{\text{(A1)}} + \underbrace{\mathcal{M}_{K,H,2}}_{\text{(A2)}} \\ &\quad + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\big(\mathbb{E}_{\pi^{\mathrm{E}}}[\iota_h^k(s_h, a_h)|s_1 = x] - \iota_h^k(s_h^k, a_h^k)\big)}_{\text{(A3)}}. \end{aligned} \tag{45}$$

To upper bound term (A1) and term (A2) in (45), we introduce the following two lemmas, respectively.

Lemma H.2 (Performance Improvement)**.** *If we set $\alpha = \sqrt{2\log(\mathrm{vol}(\mathcal{A}))/(H^2K\sqrt{d}\,)}$ in the policy update stage of OGAPI (Line 5 of Algorithm 2), then under Assumption 2.1, for any initial state $x \in \mathcal{S}$, it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^{\mathrm{E}}}\big[\langle \widehat{Q}_h^{k-1}, \pi_h^{\mathrm{E}} - \pi_h^{k-1}\rangle_{\mathcal{A}} \,\big|\, s_1 = x\big] \le \sqrt{2H^4\sqrt{d}K\log(\mathrm{vol}(\mathcal{A}))}.$$

Proof. See Appendix J.2 for a detailed proof. □

Lemma H.3. *It holds that*

$$|\mathcal{M}_{K,H,2}| \le 4\sqrt{H^3dK\log(8/\xi)}.$$

with probability at least $1 - \xi/4$, where $\mathcal{M}_{K,H,2}$ is the martingale defined in (44).

Proof. See Appendix J.3 for a detailed proof. □

To upper bound the term (A3) in (45), we introduce the following two lemmas.

Lemma H.4 (Optimism). *Under Assumption 2.1, it holds with probability at least $1-\xi/4$ that*

$$-2\Gamma_h^k(s,a) \le \iota_h^k(s,a) \le 0 \quad \textit{for any } (h,k,s,a) \in [H]\times[K]\times\mathcal{S}\times\mathcal{A},$$

where ι_h^k is the model prediction error defined in (44).

Proof. See Appendix J.4 for a detailed proof. □

Lemma H.5. *Under Assumption 2.1, it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\Gamma_h^k(s_h^k,a_h^k) \le C'\sqrt{H^4d^3K}\cdot\log(HdK/\xi),$$

where $C'>0$ is an absolute constant.

Proof. See Appendix J.5 for a detailed proof. □

Lemma H.4 implies that $\mathbb{E}_{\pi^{\mathrm{E}}}[\iota_h^k(s_h,a_h)\,|\,s_1=x]\ge 0$ with high probability. Combining Lemmas H.4 and H.5, it holds with probability at least $1-\xi/4$ that

$$\text{(A3)} \le \sum_{k=1}^{K}\sum_{h=1}^{H}\iota_h^k(s_h^k,a_h^k) \le 2\sum_{k=1}^{K}\sum_{h=1}^{H}\Gamma_h^k\left(s_h^k,a_h^k\right) \le 2C'\sqrt{H^4d^3K}\cdot\log(HdK/\xi), \tag{46}$$

with probability at least $1-\xi/4$.

Now, by plugging Lemma H.2, Lemma H.3, and (46) into the formulation of term (A) in (42), we obtain that

$$\begin{aligned}\text{(A)} &\le \sqrt{2H^4\sqrt{d}K\log(\mathrm{vol}(\mathcal{A}))} + 4\sqrt{H^3dK\log(8/\xi)} + 2C'\sqrt{H^4d^3K}\cdot\log(HdK/\xi)\\ &\le C_1\sqrt{H^4d^3K}\log(HdK/\xi),\end{aligned}\tag{47}$$

with probability at least $1-\xi/2$, where C_1 is an absolute constant.

Upper bound of term (B) in (42). We decompose term (B) in (42) by the following lemma, which characterizes the regret occuring in the reward update.

Lemma H.6. *For any $\mu=\{\mu_h\}_{h=1}^H\in S$, it holds that*

$$\begin{aligned}\sum_{k=1}^{K}\left[L(\pi^k,\mu)-L(\pi^k,\mu^k)\right] \le &\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{2\eta}\left(\|\mu_h^k-\mu_h\|_2^2-\|\mu_h^{k+1}-\mu_h\|_2^2-\|\mu_h^{k+1}-\mu_h^k\|_2^2\right)\\ &+\sum_{k=1}^{K}\sum_{h=1}^{H}\left[(\mu_h^{k+1}-\mu_h^k)^\top\widehat{\nabla}_{\mu_h}L(\pi^k,\mu^k)\right]+K\left[\widetilde{J}(\pi^{\mathrm{E}},r^\mu)-J(\pi^{\mathrm{E}},r^\mu)\right]\\ &+\sum_{k=1}^{K}\sum_{h=1}^{H}\left[(\mu_h^k-\mu_h)^\top(\psi(s_h^k,a_h^k)-\nabla_{\mu_h}J(\pi^k,\mu^k))\right],\end{aligned}$$

for any $(k,h)\in[K]\times[H]$. Here $\widehat{\nabla}_{\mu_h}L(\pi,\mu^k)$ and $\widetilde{J}(\pi^{\mathrm{E}},r^\mu)$ are defined in (25) and (23), respectively.

Proof. See Appendix J.6 for a detailed proof. □

Applying Lemma H.6, we decompose term (B) in (42) into four terms as follows,

$$\begin{aligned}\text{(B)} \leq & \underbrace{\sup_{\mu\in S}\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{2\eta}\big(\|\mu_h^k-\mu_h\|_2^2-\|\mu_h^{k+1}-\mu_h\|_2^2-\|\mu_h^{k+1}-\mu_h^k\|_2^2\big)}_{\text{(B1)}}\\ & +\underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\big[(\mu_h^{k+1}-\mu_h^k)^\top\widehat{\nabla}_{\mu_h}L(\pi^k,\mu^k)}_{\text{(B2)}}\\ & +\underbrace{K\cdot\sup_{\mu\in S}\big[\widetilde{J}(\pi^{\mathrm{E}},r^\mu)-J(\pi^{\mathrm{E}},r^\mu)\big]}_{\text{(B3)}}\\ & +\underbrace{\sup_{\mu\in S}\sum_{k=1}^{K}\sum_{h=1}^{H}\big[(\mu_h^k-\mu_h)^\top\big(\psi(s_h^k,a_h^k)-\nabla_{\mu_h}J(\pi^k,\mu^k)\big)\big]}_{\text{(B4)}}.\end{aligned} \tag{48}$$

We upper bound terms (B1), (B2), (B3), and (B4) as follows.

By telescoping the summand in term (B1) of (48) with respect to $k\in[K]$, we have

$$\begin{aligned}\text{(B1)} &= \sup_{\mu\in S}\frac{1}{2\eta}\Big[\sum_{h=1}^{H}\big((\|\mu_h^1-\mu_h\|_2^2-\|\mu_h^{K+1}-\mu_h\|_2^2-\sum_{k=1}^{K}\|\mu_h^{K+1}-\mu_h^k\|_2^2\big)\Big]\\ &\leq \sup_{\mu\in S}\frac{1}{2\eta}\sum_{h=1}^{H}\|\mu_h^1-\mu_h\|_2^2\leq\frac{2}{\eta}Hd,\end{aligned} \tag{49}$$

where the last inequality relies on the fact that $\|\mu_h\|_2\leq\sqrt{d}$ for all $\mu=\{\mu_h\}_{h=1}^H\in S$. This upper bounds term (B1).

By the update process $\mu_h^{k+1}=\mathrm{Proj}_B\{\mu_h^k+\eta\widehat{\nabla}_{\mu_h}L(\pi^k,\mu^k)\}$ in OGAPI (Line 16 of Algorithm 2), we have

$$\|\mu_h^{k+1}-\mu_h^k\|_2\leq\|\eta\widehat{\nabla}_{\mu_h}L(\pi^k,\mu^k)\|_2. \tag{50}$$

Then we upper bound (B2) in (48) as follows,

$$\begin{aligned}\text{(B2)} &= \sum_{k=1}^{K}\sum_{h=1}^{H}\big[(\mu_h^{k+1}-\mu_h^k)^\top\widehat{\nabla}_{\mu_h}L(\pi^k,\mu^k)\big]\\ &\leq\sum_{k=1}^{K}\sum_{h=1}^{H}\|\mu_h^{k+1}-\mu_h^k\|_2\cdot\|\widehat{\nabla}_{\mu_h}L(\pi^k,\mu^k)\|_2\\ &\leq\sum_{k=1}^{K}\sum_{h=1}^{H}\eta\cdot\|\widehat{\nabla}_{\mu_h}L(\pi^k,\mu^k)\|_2^2\leq 4\eta HK,\end{aligned} \tag{51}$$

where the first inequality follows from Cauchy-Schwartz inequality, the second inequality follows from (50) and the last inequality follows from the fact that $\|\widehat{\nabla}_{\mu_h}L(\pi^k,\mu^k)\|\leq 2\|\psi(\cdot,\cdot)\|_2\leq 2$. This upper bounds term (B2).

We upper bound term (B3) via the following lemma.

Lemma H.7 (Monte Carlo Estimation)**.** *Since reward function class $\mathcal{R}$ is linear as defined in (9) and the estimator $\widetilde{J}(\pi^{\mathrm{E}},r^\mu)$ constructed in (25), it holds that*

$$\sup_{\mu\in S}\big|\widetilde{J}(\pi^{\mathrm{E}},r^\mu)-J(\pi^{\mathrm{E}},r^\mu)\big|\leq 4\sqrt{H^3d^2/N_1}\log(6N_1/\xi),$$

with probability at least $1-\xi$.

Proof. See Appendix J.7 for detailed proof. □

By Lemma H.7, it holds with probability at least $1-\xi/4$ that

$$\text{(B3)} \leq 4K\sqrt{H^3d^2/N_1}\log(24N_1/\xi), \tag{52}$$

which upper bounds term (B3).

To upper bound term (B4) in (48), we introduce the following lemma.

Lemma H.8 (Unbiased Estimation)**.** *It holds that*

$$\sup_{\mu\in S}\sum_{k=1}^{K}\sum_{h=1}^{H}\big[(\mu_h^k-\mu_h)^\top(\psi(s_h^k,a_h^k)-\nabla_{\mu_h}J(\pi^k,\mu^k))\big] \leq 32\sqrt{H^3d^2K\log(9/\xi)},$$

with probability at least $1-\xi$.

Proof. See Appendix J.8 for a detailed proof. □

By Lemma H.8, it holds with probability at least $1-\xi/4$ that

$$\text{(B4)} \leq \sup_{\mu\in S}\sum_{k=1}^{K}\sum_{h=1}^{H}\big[(\mu_h^k-\mu_h)^\top(\psi(s_h^k,a_h^k)-\nabla_{\mu_h}J(\pi^k,\mu^k)) \leq 32\sqrt{H^3d^2K\log(36/\xi)}, \tag{53}$$

which upper bounds term (B4).

Plugging (49), (51), (53), and (52) into (48), it holds with probability at least $1-\xi/2$ that

$$\begin{aligned}\text{(B)} &\leq 2\sqrt{H^3d^2K}+4K\sqrt{HK}+32\sqrt{H^3d^2K\log(36/\xi)}+4\sqrt{H^3d^2/N_1}\log(24N_1/\xi)\\ &\leq 32\sqrt{H^3d^2K\log(36/\xi)}+4K\sqrt{H^3d^2/N_1}\log(24N_1/\xi),\end{aligned} \tag{54}$$

where we recall that $\eta = 1/\sqrt{HK}$. Combining (42), (47), and (54), we obtain that

$$\begin{aligned}\text{Regret}(K) \leq& C_1\sqrt{H^4d^3K}\log(HdK/\xi)+32\sqrt{H^3d^2K\log(24/\xi)}+4\sqrt{H^3d^2/N_1}\log(36N_1/\xi)\\ \leq& C_2\big(H^2d^{3/2}K^{1/2}\log(HdK/\xi)+KH^{3/2}dN_1^{-1/2}\log(N_1/\xi)\big),\end{aligned}$$

with probability at least $1-\xi$, where C_2 is an absolute constant. This concludes the proof of Theorem 4.1. □

I. Proof Sketch: Analysis of PGAPI

I.1. Proof of Theorem 4.2

Proof. By the property of mixed policy, we can rewrite the optimality gap as

$$\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}},\widehat{\pi}) = \sup_{\mu\in S}\big[J(\pi^{\mathrm{E}},r^\mu)-J(\widehat{\pi},r^\mu)\big] = \frac{1}{K}\sup_{\mu\in S}\sum_{k=1}^{K}L(\pi^k,\mu), \tag{55}$$

where $L(\pi^k,\mu)=J(\pi^{\mathrm{E}},r^\mu)-J(\pi^k,r^\mu)$. Recall that $J(\pi^k,r^\mu)=V_{1,\pi^k}^{r^\mu}(x)$ and $\widehat{J}(\pi^k,r^k)=\widehat{V}_1^k(x)$, where x is the initial state, we upper bound $\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}},\widehat{\pi})$ as follows,

$$\begin{aligned}\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}},\widehat{\pi}) \leq \frac{1}{K}\Big\{&\underbrace{\sum_{k=1}^{K}\big[J(\pi^{\mathrm{E}},r^k)-\widehat{J}(\pi^k,r^k)\big]}_{\text{(A)}}\\ &+\underbrace{\sup_{\mu\in S}\sum_{k=1}^{K}\big[J(\pi^{\mathrm{E}},r^\mu)-J(\pi^k,r^\mu)-J(\pi^{\mathrm{E}},r^k)+\widehat{J}(\pi^k,r^k)\big]}_{\text{(B)}}\Big\}.\end{aligned} \tag{56}$$

We upper bound terms (A) and (B) in (56)as follows, respectively.

Upper Bound of Term (A) in (56). To upper bound term (A) in (56), we introduce the following lemma.

Lemma I.1 (Extended Value Difference (Cai et al., 2020)). *Let $\pi = \{\pi_h\}_{h=1}^H$ and $\pi' = \{\pi'_h\}_{h=1}^H$ be any two policies and let $\{\widehat{Q}_h\}_{h=1}^H$ be any estimated Q-functions. For any $h \in [H]$, we define the estimated V-function $\widehat{V}_h : \mathcal{S} \to \mathbb{R}$ by setting $\widehat{V}_h(x) = \big\langle \widehat{Q}_h(x, \cdot), \pi_h(\cdot \,|\, x) \big\rangle_{\mathcal{A}}$ for any $x \in \mathcal{S}$. For any initial state $x \in \mathcal{S}$, we have*

$$\widehat{V}_1(x) - V_1^{\pi'}(x) = \sum_{h=1}^H \mathbb{E}_{\pi'}\big[\langle \widehat{Q}_h(s_h, \cdot), \pi_h(\cdot \,|\, s_h) - \pi'_h(\cdot \,|\, s_h)\rangle_{\mathcal{A}} \,\big|\, s_1 = x\big]$$
$$+ \sum_{h=1}^H \mathbb{E}_{\pi'}\big[\widehat{Q}_h(s_h, a_h) - r(s_h, a_h) - \mathcal{P}_h \widehat{V}_{h+1}(s_h, a_h) \,\big|\, s_1 = x\big],$$

where $\mathbb{E}_{\pi'}$ is taken with respect to the trajectory generated by π' and r is the reward function.

Proof. See Appendix B.1 in (Cai et al., 2020) for a detailed proof. □

For the simplicity of later discussion, at the h-th step of k-th episode, we define the error for estimating Bellman equation in (4) in the policy evaluation stage of PGAPI (Lines 8–11 of Algorithm 3) with any reward function r^μ as follows,

$$\iota_h^{k,r^\mu}(s,a) := r_h^\mu(s,a) + [\mathcal{P}_h \widehat{V}_{h+1}^{k,r^\mu}](s,a) - \widehat{Q}_h^{k,r^\mu}(s,a),$$

for any $(s,a,h,\mu) \in \mathcal{S} \times \mathcal{A} \times [H] \times S$, where $\widehat{Q}_h^{k,r^\mu}$ and $\widehat{V}_{h+1}^{k,r^\mu}$ are defined as

$$\begin{aligned} &\widehat{V}_{H+1}^{k,r^\mu}(\cdot) = 0, \\ &\widehat{Q}_h^{k,r^\mu}(\cdot,\cdot) = \max\big\{(r_h^\mu + \widehat{\mathcal{P}}_h \widehat{V}_{h+1}^{k,r^\mu} - \Gamma_h)(\cdot,\cdot), 0\big\}, \\ &\widehat{V}_h^{k,r^\mu}(\cdot,\cdot) = \big\langle \widehat{Q}_h^{k,r^\mu}(\cdot,\cdot), \pi_h^k(\cdot \,|\, \cdot)\big\rangle_{\mathcal{A}}, \quad \text{for } h \in [H]. \end{aligned} \tag{57}$$

It implies that $\widehat{Q}_h^k = \widehat{Q}_h^{k,r^k}$ and $\widehat{V}_h^k = \widehat{V}_h^{k,r^k}$, where $\widehat{Q}_h^k$ and $\widehat{V}_h^k$ are constructed in the policy evaluation stage in PGAPI (Lines 8–11 of Algorithm 3).

By applying Lemma I.1, we decompose term (A) in (56) as follows,

$$\begin{aligned} \text{(A)} &= \underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^{\mathrm{E}}}\big[\langle \widehat{Q}_h^k(s_h, \cdot), \pi_h^{\mathrm{E}}(\cdot \,|\, s_h) - \pi_h^k(\cdot \,|\, s_h)\rangle_{\mathcal{A}} \big| s_1 = x\big]}_{\text{(A1)}} \\ &\quad + \underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^{\mathrm{E}}}\big[\iota_h^{k,r^k}(s_h, a_h) | s_1 = x\big]}_{\text{(A2)}}. \end{aligned} \tag{58}$$

We upper bound terms (A1) and (A2) in (58) as follows.

By Lemma H.2, we have

$$\text{(A1)} \le \sqrt{2H^4 d\sqrt{d}K \log(\mathrm{vol}(\mathcal{A}))}. \tag{59}$$

We upper bound term (A2) in (58) using the following lemma.

Lemma I.2 (Pessimism). *If $\{\Gamma_h\}_{h\in[H]}$ are ξ-uncertainty qualifiers defined in Definition 3.1, when conditioned on $\mathcal{E}$ defined in Definition 3.1, it holds for any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ and any reward function r^μ with $\mu \in S$ that*

$$0 \le \iota_h^{k,r^\mu}(s,a) \le 2\Gamma_h(s,a).$$

Proof. See Appendix K.2 for a detailed proof. □

By Lemma I.2, conditioned on $\mathcal{E}$, we upper bound term (A2) in (58) as follows,

$$\text{(A2)} \le \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^{\mathrm{E}}}\big[2\Gamma_h(s_h,a_h)\,\big|\,s_1=x\big] = K\cdot \mathrm{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}}, \tag{60}$$

where we denote by $\mathrm{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}} = \sum_{h=1}^{H}\mathbb{E}_{\pi^{\mathrm{E}}}\big[2\Gamma_h(s_h,a_h)\,\big|\,s_1=x\big]$ for notational convenience. Combining (59) and (60), we derive an upper bound on term (A) in (56) conditioned on $\mathcal{E}$ as

$$\text{(A)} \le \sqrt{2H^4\sqrt{d}\log(\mathrm{vol}(\mathcal{A}))} + K\cdot \mathrm{IntUncert}_{\mathbb{D}}^{\pi^{\mathrm{E}}}. \tag{61}$$

Upper Bound of Term (B) in (56). We $\widehat{L}(\pi,\mu)$ as follows,

$$\widehat{L}(\pi,\mu) = \widetilde{J}(\pi^{\mathrm{E}},r^{\mu}) - \widehat{J}(\pi^k,r^{\mu}).$$

Here $\widetilde{J}(\pi,r^{\mu})$ is the MC estimation defined in (23) and $\widehat{J}(\pi^k,r^{\mu})$ is the estimated cumulative reward defined as $\widehat{J}(\pi^k,r^{\mu}) = \widehat{V}_1^{k,r^{\mu}}(x)$, where $\widehat{V}_1^{k,r^{\mu}}$ is defined in (57) and x is the initial state. By this, we upper bound term (B) in (56) as

$$\begin{aligned}
\text{(B)} =& \sup_{\mu\in S}\sum_{k=1}^{K}\big[J(\pi^{\mathrm{E}},r^{\mu}) - J(\pi^k,r^{\mu}) - J(\pi^{\mathrm{E}},r^k) + \widehat{J}(\pi^k,r^k)\big] \\
\le& \underbrace{\sup_{\mu\in S}\sum_{k=1}^{K}\big[J(\pi^{\mathrm{E}},r^{\mu}) - \widetilde{J}(\pi^{\mathrm{E}},r^{\mu})\big]}_{\text{(B1)}} + \underbrace{\sup_{\mu\in S}\sum_{k=1}^{K}\big[-J(\pi^{\mathrm{E}},r^k) + \widetilde{J}(\pi^{\mathrm{E}},r^k)\big]}_{\text{(B2)}} \\
&+ \underbrace{\sup_{\mu\in S}\sum_{k=1}^{K}\big[\widehat{J}(\pi^k,r^{\mu}) - J(\pi^k,r^{\mu})\big]}_{\text{(B3)}} + \underbrace{\sup_{\mu\in S}\sum_{k=1}^{K}\big[\widehat{L}(\pi^k,\mu) - \widehat{L}(\pi^k,\mu^k)\big]}_{\text{(B4)}}.
\end{aligned} \tag{62}$$

We upper bounds terms (B1)–(B4) in (62) as follows.

By applying Lemma H.7 on term (B1) and term (B2) in (62), it holds with probability at least $1-\xi/2$ that

$$\text{(B1) + (B2)} \le 8K\sqrt{H^3d^2/N_1}\log(24N_1/\xi). \tag{63}$$

To upper bound term (B3) in (62), we invoke Lemmas I.1 and I.2, which imply

$$\text{(B3)} = \sup_{\mu\in S}\Big[\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^k}\big[-\iota_h^{k,r^{\mu}}(s_h^k,a_h^k)\,\big|\,s_1=x\big]\Big] \le 0. \tag{64}$$

To upper bound term (B4) in (62), we first introduce the following lemma.

Lemma I.3. *The function $\widehat{L}(\pi^k,\mu)$ defined in (27) is concave in μ_h for any $h\in[H]$, where $\mu = \{\mu_h\}_{h=1}^{H}\in S$ and $\widehat{L}(\pi^k,\mu)$.*

Proof. See Appendix K.3 for a detailed proof. □

By Lemma I.3, we establish the following lemma to upper bound term (B4) in (62), which corresponds to the reward update stage in PGAPI (Lines 12–17 of Algorithm 3).

Lemma I.4. *For any $\mu \in S$, it holds that*

$$\sum_{k=1}^{K} \big[\widehat{L}(\pi^k, \mu) - \widehat{L}(\pi^k, \mu^k)\big] \le \sum_{k=1}^{K}\sum_{h=1}^{H} \big[\frac{1}{2\eta}\|\mu_h^{k+1} - \mu_h\|_2^2 + \frac{1}{2\eta}\|\mu_h^{k+1} - \mu_h\|_2^2 \\ - \frac{1}{2\eta}\|\mu_h^{k+1} - \mu_h^k\|_2^2 + \eta\|\nabla_{\mu_h}\widehat{L}(\pi^k, \mu^k)\|_2^2\big],$$

Proof. See Appendix K.4 for a detailed proof. □

By Lemma I.4, we have

$$\text{(B4)} \le \sup_{\mu\in S}\Big[\sum_{h=1}^{H}\big(\frac{1}{2\eta}\|\mu_h^1 - \mu_h\|_2^2 - \frac{1}{2\eta}\|\mu_h^{K+1} - \mu_h\|_2^2 - \frac{1}{2\eta}\sum_{k=1}^{K}\|\mu_h^k - \mu_h^{k+1}\|_2^2 \\ + \sum_{k=1}^{K}\eta\|\nabla_{\mu_h}\widehat{L}(\pi^k, \mu^k)\|_2^2\big)\Big],$$

which implies that

$$\text{(B4)} \le \sup_{\mu\in S}\sum_{h=1}^{H}\big[\frac{1}{2\eta}\|\mu_h^1 - \mu_h\|_2^2\big] + \sup_{\mu\in S}\sum_{h=1}^{H}\sum_{k=1}^{K}\big[\eta\|\nabla_{\mu_h}\widehat{L}(\pi^k, \mu^k)\|_2^2\big]. \tag{65}$$

Based on (65), we upper bound $\|\nabla_{\mu_h}\widehat{L}(\pi^k, r^\mu)\|_2^2$ for any $\mu \in S$ and $h \in [H]$ as follows,

$$\begin{aligned}\|\nabla_{\mu_h}\widehat{L}(\pi^k, r^\mu)\|_2^2 &\le \|\nabla_{\mu_h}\widetilde{J}(\pi^{\mathrm{E}}, r^\mu) - \nabla_{\mu_h}\widehat{J}(\pi^k, r^\mu)\|_2^2 \\ &\le 2\|\nabla_{\mu_h}\widetilde{J}(\pi^{\mathrm{E}}, r^\mu)\|_2^2 + 2\|\nabla_{\mu_h}\widehat{J}(\pi^k, r^\mu)\|_2^2.\end{aligned} \tag{66}$$

Recall that $\nabla_{\mu_h} J(\pi^{\mathrm{E}}, r^\mu) = {N_1}^{-1}\sum_{\tau=1}^{N_1}\psi(s_{h,\tau}^{\mathrm{E}}, a_{h,\tau}^{\mathrm{E}})$ and $\nabla_{\mu_h}\widehat{J}(\pi^k, r^\mu)$ is characterized in Proposition D.1, then we have that

$$\|\nabla_{\mu_h}\widetilde{J}(\pi^{\mathrm{E}}, r^\mu)\|_2^2 + \|\nabla_{\mu_h}\widehat{J}(\pi^k, r^\mu)\|_2^2 \le 2\|\psi(\cdot,\cdot)\|_2^2. \tag{67}$$

Since it holds that $\|\mu_h\|_2 \le \sqrt{d}$ for any $\mu \in S$ and $\|\psi(\cdot,\cdot)\|_2 \le 1$, it also yields that

$$\|\mu_h' - \mu_h\|_2^2 \le (\|\mu_h'\|_2 + \|\mu_h\|_2)^2 \le 4d \tag{68}$$

for any $\mu, \mu' \in S$. By setting $\eta = 1/\sqrt{KH}$ and combining (65), (66), (67), and (68), we attain that

$$\begin{aligned}\text{(B4)} \le& H \cdot \frac{\sqrt{HK}}{2}\cdot 4d + HK \cdot \frac{1}{\sqrt{HK}}\cdot 2 \cdot 4d \\ \le& 2H^{3/2}dK^{1/2} + 8H^{1/2}dK^{1/2} \le 8H^{3/2}dK^{1/2}.\end{aligned} \tag{69}$$

Plugging (63), (64), and (69) into (62), conditioned on $\mathcal{E}$, it holds with probability at least $1 - \xi/2$ that

$$\text{(B)} \le 8K\sqrt{H^3d^2/N_1}\log(24N_1/\xi) + 8dH^{3/2}K^{1/2}. \tag{70}$$

Recall that Definition 3.1 implies $\mathbb{P}_{\mathbb{D}}(\mathcal{E}) > 1 - \xi/2$. Combining (55), (61), and (70), with probability at least $1 - \xi$, we have

$$\begin{aligned}\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}}, \widehat{\pi}) \le& \sqrt{2H^4\sqrt{d}\log(\mathrm{vol}(\mathcal{A}))/K} + \ \mathrm{IntUncert}_{\mathbb{D}}^{\pi^{\mathrm{E}}} \\ &+ 8\sqrt{H^3d^2/N_1}\log(24N_1/\xi) + 8H^{3/2}dK^{-1/2} \\ \le& 8H^2dK^{-1/2} + \ \mathrm{IntUncert}_{\mathbb{D}}^{\pi^{\mathrm{E}}} + 8H^{3/2}dN_1^{-1/2}\log(24N_1/\xi).\end{aligned}$$

Thus, we conclude the proof of Theorem 4.2. □

I.2. Proof of Proposition F.1

Proof. Our proof is based on the following theorem that presents the information-theoretic lower bound.

Theorem I.5 (Information-Theoretic Lower Bound, Theorem 4.6 in (Jin et al., 2021))**.** *For the output $\texttt{Algo}(\overline{\mathbb{D}})$ of any offline RL algorithm based on the dataset $\overline{\mathbb{D}}$, there exists a tabular MDP $\mathcal{M}$ $(\mathcal{S},\mathcal{A},H,\mathcal{P},r)$ with initial state $x\in\mathcal{S}$ and a dataset $\overline{\mathbb{D}}$ which is compliant with $\mathcal{M}$, such that*

$$\mathbb{E}_{\overline{\mathbb{D}}}\left[\frac{J(\pi^\star)-J(\texttt{Algo}(\overline{\mathbb{D}}))}{\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[1/\sqrt{1+n_h\left(s_h,a_h\right)}\,\middle|\,s_1=x\right]}\right]\geq c$$

where c is an absolute constant,$\pi^\star$ is the optimal policy statisfying that $\pi^\star:=\operatorname{argmax}_{\pi\in\Delta(\mathcal{A}\,|\,\mathcal{S},H)}J(\pi)$, and $n_h\left(s_h,a_h\right)=\sum_{\tau=1}^{N}\mathbf{1}\left\{s_h^\tau=s_h,a_h^\tau=a_h\right\}$ for $\left(s_h,a_h\right)\in\mathcal{S}\times\mathcal{A}$.

First we show the linear kernel MDP defined in Assumption 2.1 can be reduced to tabular MDP. If we set $d=|\mathcal{S}|^2|\mathcal{A}|$ and take the feature map as the canonical basis

$$\phi(s,a,s')=\mathbf{e}_{(s,a,s')},\quad \psi(s,a)=\mathbf{e}_{(s,a)},\quad \text{for all } (s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S},$$

then Assumption 2.1 is satisfied with $R=1$.

Applying Theorem I.5, we derive a hard instance in tabular MDP with known reward r. When we choose the reward set $\mathcal{R}$ as the singleton reward r, it yields that

$$\max_{\pi^{\mathrm{E}}\in\Delta(\mathcal{A}\,|\,\mathcal{S},H)}\mathbb{E}_{\overline{\mathbb{D}}}\left[\frac{\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}},\texttt{Algo}(\overline{\mathbb{D}}))}{\text{Information}_{\overline{\mathbb{D}}}^{\pi^{\mathrm{E}}}}\right]\geq\mathbb{E}_{\overline{\mathbb{D}}}\left[\frac{\mathbf{D}_{\mathcal{R}}(\pi^\star,\texttt{Algo}(\overline{\mathbb{D}}))}{\text{Information}_{\overline{\mathbb{D}}}^{\pi^\star}}\right]=\mathbb{E}_{\overline{\mathbb{D}}}\left[\frac{J(\pi^\star)-J(\texttt{Algo}(\overline{\mathbb{D}}))}{\text{Information}_{\overline{\mathbb{D}}}^{\pi^\star}}\right],\tag{71}$$

where the last equality is originated from the definition of optimality gap in (8).

Next we handle $\text{IntUncert}_{\overline{\mathbb{D}}}^{\pi^\star}$, which takes the form as

$$\begin{aligned}\text{Information}_{\overline{\mathbb{D}}}^{\pi^\star}&=(\text{Vol}(\mathcal{S}))^{-1}\cdot\mathbb{E}_{\pi^\star}\left[\sum_{h=1}^{H}\int_{\mathcal{S}}\|\phi(s_h,a_h,s')\|_{\Lambda_h^{-1}}\mathrm{d}s'\,\middle|\,s_1=x\right]\\ \text{where }\Lambda_h&=\lambda I+\sum_{\tau=1}^{N}\int_{\mathcal{S}}\phi(s_h^\tau,a_h^\tau,s')\phi(s_h^\tau,a_h^\tau,s')^\top\mathrm{d}s',\end{aligned}\tag{72}$$

where N is the number of the trajectories in the dataset $\overline{\mathbb{D}}$. In the tabular setting, we obtain that

$$\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau,s')\phi(s_h^\tau,a_h^\tau,s')^\top=\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}n_h(s,a)W_{(s,a,s')},$$

where $W_{(s,a,s')}$ is a symmetric matrix whose non-zero entry is at $((s,a,s')(s,a,s'))$ and equals to 1. Summing s' over $\mathcal{S}$, in the tabular case we have that

$$\Lambda_h=\lambda I+\sum_{\tau=1}^{N}\sum_{s'\in\mathcal{S}}(s_h^\tau,a_h^\tau,s')\phi(s_h^\tau,a_h^\tau,s')^\top\mathrm{d}s'=\lambda I+\sum_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}n_h(s,a)W_{(s,a,s')}.$$

Choosing $\lambda=1$, we obtain that

$$\phi\left(s,a,s'\right)^\top\Lambda_h^{-1}\phi\left(s,a,s'\right)=\frac{1}{1+n_h(s,a)},$$

for all $(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$. Hence we have that

$$\sum_{h=1}^{H}\sum_{s'\in\mathcal{S}}\|\phi(s,a,s')\|_{\Lambda_h^{-1}}=\sum_{h=1}^{H}\frac{|\mathcal{S}|}{\sqrt{1+n_h(s,a)}},\tag{73}$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Taking expectation on (73) with respect to the optimal policy $\pi^\star$ and according to (72), it holds that

$$\text{Information}_{\mathbb{D}}^{\pi^\star} = |\mathcal{S}|^{-1} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\big[1/\sqrt{1+n_h\left(s_h, a_h\right)} \,\big|\, s_1 = x\big]. \tag{74}$$

Plugging (74) into (71), under the hard instance in Theorem I.5, we obtain that

$$\max_{\pi^{\mathrm{E}} \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)} \mathbb{E}_{\overline{\mathbb{D}}}\bigg[\frac{\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}}, \texttt{Algo}(\overline{\mathbb{D}}))}{\text{Information}_{\mathbb{D}}^{\pi^{\mathrm{E}}}}\bigg] \geq c,$$

where $c > 0$ is a positive constant. Then we conclude the proof of Proposition F.1

□

I.3. Proof of Corollary 4.4

Proof. By the property of trace, we have that

$$\begin{aligned}
&\mathbb{E}_{\pi^{\mathrm{E}}}\Big[\int_{\mathcal{S}} \sqrt{(\phi(s_h, a_h, s')^\top \Lambda_h^{-1} \phi(s_h, a_h, s')} \mathrm{d}s' \Big| \, s_1 = x\Big] \\
&\quad = \mathbb{E}_{\pi^{\mathrm{E}}}\Big[\int_{\mathcal{S}} \sqrt{\mathrm{Tr}(\phi(s_h, a_h, s')^\top \Lambda_h^{-1} \phi(s_h, a_h, s'))} \mathrm{d}s' \Big| \, s_1 = x\Big] \\
&\quad = \mathbb{E}_{\pi^{\mathrm{E}}}\Big[\int_{\mathcal{S}} \sqrt{\mathrm{Tr}(\phi(s_h, a_h, s')\phi(s_h, a_h, s')^\top \Lambda_h^{-1})} \mathrm{d}s' \Big| \, s_1 = x\Big].
\end{aligned} \tag{75}$$

Applying Cauchy-Schwartz inequality on (75), we derive that

$$\begin{aligned}
&\mathbb{E}_{\pi^{\mathrm{E}}}\Big[\int_{\mathcal{S}} \sqrt{\mathrm{Tr}(\phi(s_h, a_h, s')\phi(s_h, a_h, s')^\top \Lambda_h^{-1})} \mathrm{d}s' \Big| \, s_1 = x\Big] \\
&\quad \leq \big(\mathrm{Vol}(\mathcal{S})\big)^{1/2} \cdot \Big(\mathbb{E}_{\pi^{\mathrm{E}}}\Big[\int_{\mathcal{S}} \mathrm{Tr}(\phi(s_h, a_h, s')\phi(s_h, a_h, s')^\top \Lambda_h^{-1}) \mathrm{d}s' \Big| \, s_1 = x\Big]\Big)^{1/2}
\end{aligned} \tag{76}$$

for any $x, s' \in \mathcal{S}$ and all $h \in [H]$.

To utilize Theorem 4.2, we define the event $\mathcal{E}^\sharp$ as follows,

$$\mathcal{E}^\sharp = \Big\{\mathbf{D}_{\mathcal{R}}(\pi^{\mathrm{E}}, \widehat{\pi}) \leq \mathcal{O}\big(H^2 d K^{-1/2}\big) + \Delta_{N_1} + \text{ IntUncert}_{\mathbb{D}}^{\pi^{\mathrm{E}}}\Big\},$$

where $\text{IntUncert}_{\mathbb{D}}^{\pi^{\mathrm{E}}} = 2\sum_{h=1}^{H} \mathbb{E}_{\pi^{\mathrm{E}}}[\Gamma_h(s_h, a_h) \,|\, s_1 = x]$ and Γ_h is defined in (36). Conditioned on the event $\mathcal{E}^\sharp \cap \mathcal{E}^\dagger$, where $\mathcal{E}^\dagger$ is defined in Assumption 4.3, we obtain that

$$\text{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}} \leq 2\kappa H\sqrt{d} \sum_{h=1}^{H} \mathbb{E}_{\pi^{\mathrm{E}}}\Big[\int_{\mathcal{S}} \|\phi(s_h, a_h, s')\|_{\Lambda_h^{-1}} \, \mathrm{d}s' \,\big|\, s_1 = x\Big]. \tag{77}$$

By plugging (75) and (76) into (77), we have

$$\text{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}} \leq 2\kappa H\sqrt{d}\sqrt{\mathrm{Vol}(\mathcal{S})} \cdot \sum_{h=1}^{H} \sqrt{\mathrm{Tr}\Big(\mathbb{E}_{\pi^{\mathrm{E}}}\Big[\int_{\mathcal{S}} \phi(s_h, a_h, s')\phi(s_h, a_h, s')^\top \mathrm{d}s' \,|\, s_1 = x\Big] \, \Lambda_h^{-1}\Big)}, \tag{78}$$

where $\mathrm{Vol}(\mathcal{S})$ is the finite measure of the state space $\mathcal{S}$. For notational simplicity, we define

$$\Sigma_h(x) = \mathbb{E}_{\pi^{\mathrm{E}}}\Big[\int_{\mathcal{S}} \phi(s_h, a_h, s')\phi(s_h, a_h, s')^\top \mathrm{d}s' \,|\, s_1 = x\Big], \tag{79}$$

for all $x \in \mathcal{S}$ and all $h \in [H]$.

By Assumption 4.3 and the definition of $\Sigma_h(x)$ in (79), we know that the matrix $\big(I + c^\dagger N_2 \Sigma_h(x)\big)^{-1} - \Lambda_h^{-1}$ is positive definite conditioned on $\mathcal{E}^\dagger$. Combining (78) and (79), we obtain that

$$\begin{aligned}\text{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}} &\le 2\big(\text{Vol}(\mathcal{S})\big)^{1/2} \cdot \kappa H \sqrt{d} \sum_{h=1}^{H} \sqrt{\text{Tr}\left(\Sigma_h(x) \cdot (I + c^\dagger \cdot N_2 \cdot \Sigma_h(x))^{-1}\right)} \\ &= 2\big(\text{Vol}(\mathcal{S})\big)^{1/2} \cdot \kappa H \sqrt{d} \sum_{h=1}^{H} \sqrt{\sum_{j=1}^{d} \frac{\lambda_{h,j}(x)}{1 + c^\dagger \cdot N_2 \cdot \lambda_{h,j}(x)}}. \end{aligned} \tag{80}$$

Here $\{\lambda_{h,j}(x)\}_{j=1}^{d}$ are the eigenvalues of $\Sigma_h(x)$ for any $x \in \mathcal{S}$ and $h \in [H]$. Meanwhile, under Assumption 2.1, we have $\|\phi(\cdot,\cdot,\cdot)\|_2 \le dR$, which is shown in (109). By applying Cauchy-Schwartz inequality to (79), it holds that

$$\|\Sigma_h(x)\|_{\text{op}} \le \mathbb{E}_{\pi^{\mathrm{E}}}\left[\left\| \int_{\mathcal{S}} \phi(s_h, a_h, s')\phi(s_h, a_h, s')^\top \mathrm{d}s' \right\|_{\text{op}} \,\middle|\, s_1 = x\right] \le d^{3/2} R^2 \cdot (\text{Vol}(\mathcal{S}))^{1/2},$$

for any $x \in \mathcal{S}$ and $h \in [H]$, where $\|\cdot\|_{\text{op}}$ is the operator norm. As $\Sigma_h(x)$ is positive semidefinite, we have $\lambda_{h,j}(x) \in [0, \|\Sigma_h(x)\|_{\text{op}}]$ for any $x \in \mathcal{S}$ and $(h,j) \in [H] \times [d]$. Hence, conditioned on $\mathcal{E}^\dagger \cap \mathcal{E}^\dagger$, combining (80), it holds for any $x \in \mathcal{S}$ that

$$\begin{aligned}\text{IntUncert}_{\mathbb{D}^{\mathrm{A}}}^{\pi^{\mathrm{E}}} &\le 2\big(\text{Vol}(\mathcal{S})\big)^{1/2} \cdot \kappa H \sqrt{d} \sum_{h=1}^{H} \sqrt{\sum_{j=1}^{d} \frac{\lambda_{h,j}(x)}{1 + c^\dagger \cdot N_2 \cdot \lambda_{h,j}(x)}} \\ &\le 2\big(\text{Vol}(\mathcal{S})\big)^{1/2} \cdot \kappa H \sqrt{d} \sum_{h=1}^{H} \sqrt{\sum_{j=1}^{d} \frac{1}{c^\dagger \cdot N_2}} \\ &\le c'\big(\text{Vol}(\mathcal{S})\big)^{1/2} d^{3/2} H^2 N_2^{-1/2} \log(HdN_2/\xi), \end{aligned} \tag{81}$$

where the second inequality follows from the fact that $\lambda_{h,j}(x) \in [0, \|\Sigma_h(x)\|_{\text{op}}]$ for any $(x,h,j) \in \mathcal{S} \times [H] \times [d]$, while the third inequality follows from the choice of the scaling parameter $\kappa > 0$ stated in Theorem 4.2. Here c' is an absolute constant dependent on c and $c^\dagger$. By the condition in Corollary 4.4, we have $\mathbb{P}_{\mathbb{D}}(\mathcal{E}^\dagger) \ge 1 - \xi/2$. Also, by Theorem 4.2, we have $\mathbb{P}_{\mathbb{D}}(\mathcal{E}^\sharp) \ge 1 - \xi/2$. Hence, by the union bound, we derive that $\mathbb{P}_{\mathbb{D}}(\mathcal{E}^\dagger \cap \mathcal{E}^\sharp) \ge 1 - \xi$. Combining (81), we finish the proof of Corollary 4.4. □

J. Proofs of Supporting Lemmas: Analysis of OGAPI

J.1. Proof of Lemma H.1

Proof. For notational simplicity, we define operators $\mathbb{J}_h$ and $\mathbb{J}_h^k$ as

$$(\mathbb{J}_h f)(s) = \langle f(s,\cdot), \pi_h^{\mathrm{E}}(\cdot \,|\, s)\rangle, \;\; (\mathbb{J}_h^k f)(s) = \langle f(s,\cdot), \pi_h^k(\cdot \,|\, s)\rangle, \tag{82}$$

for any $s \in \mathcal{S}$, $(k,h) \in [K] \times [H]$, and any function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. We define $\mathcal{F}_{k,h,1}, \mathcal{F}_{k,h,2}$ as follows,

$$\begin{aligned}\mathcal{F}_{k,h,1} &= \sigma\big(\{(s_i^\tau, a_i^\tau)\}_{(\tau,i)\in[k-1]\times[H]} \cup \{(s_i^k, a_i^k)\}_{i\in[h]}\big) \\ \mathcal{F}_{k,h,2} &= \sigma\big(\{(s_i^\tau, a_i^\tau)\}_{(\tau,i)\in[k-1]\times[H]} \cup \{(s_i^k, a_i^k)\}_{i\in[h]} \cup \{s_{h+1}^k\}\big), \end{aligned} \tag{83}$$

where s_{H+1}^k is defined as a null state for any $k \in [K]$. We define the time index as follows,

$$t(k,h,m) = (k-1)\cdot 2H + (h-1)\cdot 2 + m, \tag{84}$$

which imples that $\{\mathcal{F}_{k,h,m}\}_{(k,h,m)\in[K]\times[H]\times[2]}$ is a filtration with respect to $t(k,h,m)$.

Now we are ready to prove Lemma H.1. First we note that for any initial state $x \in \mathcal{S}$, it holds that

$$V_{1,\pi^{\mathrm{E}}}^{r^k}(x) - V_{1,\pi^k}^{r^k}(x) = \underbrace{\big(V_{1,\pi^{\mathrm{E}}}^{r^k}(x) - V_1^k(x)\big)}_{\text{(i)}} + \underbrace{\big(\widehat{V}_1^k(x) - V_{1,\pi^k}^{r^k}(x)\big)}_{\text{(ii)}}, \tag{85}$$

where V_1^k is the estimated state value function in the stage of policy evaluation of OGAPI (Lines 2–2 of Algorithm 2). We calculate terms (i) and (ii) separately.

Term (i). By (4), we have

$$V^{r^k}_{h,\pi^{\mathrm{E}}}(s) = \langle Q^{r^k}_{h,\pi^{\mathrm{E}}}(s,\cdot), \pi^{\mathrm{E}}_h(\cdot\,|\,s)\rangle = \mathbb{J}_h Q^{r^k}_{h,\pi^{\mathrm{E}}}(s), \;\; \widehat{V}^k_h(s) = \langle \widehat{Q}^k_h(s,\cdot), \pi^k_h(\cdot\,|\,s)\rangle = \mathbb{J}^k_h \widehat{Q}^k_h(s), \tag{86}$$

for any $(k,h) \in [K] \times [H]$. We then have

$$\begin{aligned} V^{r^k}_{h,\pi^{\mathrm{E}}} - \widehat{V}^k_h &= \mathbb{J}_h Q^{r^k}_{h,\pi^{\mathrm{E}}} - \mathbb{J}^k_h \widehat{Q}^k_h \\ &= \mathbb{J}_h (Q^{r^k}_{h,\pi^{\mathrm{E}}} - \widehat{Q}^k_h) + (\mathbb{J}_h - \mathbb{J}^k_h)\widehat{Q}^k_h, \end{aligned} \tag{87}$$

where $\mathbb{J}_h$ and $\mathbb{J}^k_h$ are defined in (82). By the property of state-value function and the definition of ι^k_h, we have

$$Q^{r^k}_{h,\pi^{\mathrm{E}}} = r^k_h + \mathcal{P}_h V^{r^k}_{h+1,\pi^{\mathrm{E}}}, \; \widehat{Q}^k_h = r^k_h + \mathcal{P}_h \widehat{V}^k_{h+1} - \iota^k_h, \tag{88}$$

Define $\zeta^k_h = (\mathbb{J}_h - \mathbb{J}^k_h)\widehat{Q}^k_h$ and plug (88) into (87), we have

$$V^{r^k}_{h,\pi^{\mathrm{E}}} - \widehat{V}^k_h = \mathbb{J}_h \mathcal{P}_h (V^{r^k}_{h+1,\pi^{\mathrm{E}}} - \widehat{V}^k_{h+1}) + \mathbb{J}_h \iota^k_h + \zeta^k_h, \tag{89}$$

for any $(k,h) \in [K] \times [H]$. Here ι^k_h is the prediction error defined in (43). For any $k \in [H]$, note that $V^{r^k}_{h+1,\pi^{\mathrm{E}}} = \widehat{V}^k_{H+1} = 0$, we expand (89) across $h \in [H]$ to obtain that

$$V^{r^k}_{1,\pi^{\mathrm{E}}} - \widehat{V}^k_1 = \sum_{h=1}^{H} (\prod_{i=1}^{h-1} \mathbb{J}_i \mathcal{P}_i) \mathbb{J}_h \iota^k_h + \sum_{h=1}^{H} (\prod_{i=1}^{h-1} \mathbb{J}_i \mathcal{P}_i) \zeta^k_h. \tag{90}$$

The effect of composite operator $\mathcal{P}_h \mathbb{J}_h$ on function f is to calculate one-step expectation of f following policy π^{E}_h. Hence we rewrite (90) as

$$\begin{aligned} V^{r^k}_{1,\pi^{\mathrm{E}}}(x) - \widehat{V}^k_1(x) = \sum_{h=1}^{H} &\big(\mathbb{E}_{\pi^{\mathrm{E}}}\big[\iota^k_h(s^k_h, a^k_h) | s_1 = x\big]\big) \\ &+ \sum_{h=1}^{H} \mathbb{E}_{\pi^{\mathrm{E}}}\big[\langle \widehat{Q}^k_h(s_h,\cdot), \pi^{\mathrm{E}}_h(\cdot\,|\,s_h) - \pi^k_h(\cdot\,|\,x_h)\rangle \,\big|\, s_1 = x\big]. \end{aligned} \tag{91}$$

This characterize term (i).

Term (ii). By (85), we have

$$\begin{aligned} \iota^k_h &= r^k_h + \mathcal{P}_h \widehat{V}^k_{h+1} - \widehat{Q}^k_h \\ &= r^k_h + \mathcal{P}_h \widehat{V}^k_{h+1} - Q^{r^k}_{h,\pi^{\mathrm{E}}} + (Q^{r^k}_{h,\pi^{\mathrm{E}}} - \widehat{Q}^k_h) \\ &= \mathcal{P}_h (\widehat{V}^k_{h+1} - V^{r^k}_{h+1,\pi^{\mathrm{E}}}) + (Q^{r^k}_{h,\pi^{\mathrm{E}}} - \widehat{Q}^k_h). \end{aligned} \tag{92}$$

By (92), we obtain that

$$\begin{aligned} \widehat{V}^k_h - V^{r^k}_{h,\pi^k} &= \mathbb{J}^k_h (\widehat{Q}^k_h - Q^{r^k}_{h,\pi^{\mathrm{E}}}) + \iota^k_h - \iota^k_h \\ &= \big(\mathbb{J}^k_h (\widehat{Q}^k_h - Q^{r^k}_{h,\pi^{\mathrm{E}}}) - (\widehat{Q}^k_h - Q^{r^k}_{h,\pi^{\mathrm{E}}})\big) + \mathcal{P}_h \big(\widehat{V}^k_{h+1} - V^{r^k}_{h+1,\pi^{\mathrm{E}}}\big) - \iota^k_h. \end{aligned} \tag{93}$$

We define $D_{k,h,1}$ and $D_{k,h,2}$ as follows,

$$\begin{aligned} D_{k,h,1} &= \big(\mathbb{J}^k_h (\widehat{Q}^k_h - Q^{r^k}_{h,\pi^{\mathrm{E}}})\big)(s^k_h) - (\widehat{Q}^k_h - Q^{r^k}_{h,\pi^{\mathrm{E}}})(s^k_h, a^k_h) \\ D_{k,h,2} &= \big(\mathcal{P}_h (\widehat{V}^k_{h+1} - V^{r^k}_{h+1,\pi^{\mathrm{E}}})\big)(s^k_h, a^k_h) - (\widehat{V}^k_{h+1} - V^{r^k}_{h+1,\pi^{\mathrm{E}}})(s^k_{h+1}). \end{aligned} \tag{94}$$

By plugging (94) into (93), we obtain that

$$\widehat{V}_h^k(s_h^k) - V_{h,\pi^k}^{r^k}(s_h^k) = D_{k,h,1} + D_{k,h,2} + (\widehat{V}_{h+1}^k - V_{h+1,\pi^{\mathrm{E}}}^{r^k})(s_{h+1}^k) - \iota_h^k(s_h^k, a_h^k). \tag{95}$$

By telescoping (95) with respect to $h \in [H]$, we have

$$\widehat{V}_1^k(x) - V_{h,\pi^k}^{r^k}(x) = \sum_{h=1}^{H}(D_{k,h,1} + D_{k,h,2}) - \sum_{h=1}^{H} \iota_h^k(s_h^k, a_h^k). \tag{96}$$

By the definition of $\mathcal{F}_{k,h,1}$ and $\mathcal{F}_{k,h,2}$ in (83), we have

$$D_{k,h,1} \in \mathcal{F}_{k,h,1},\ D_{k,h,2} \in \mathcal{F}_{k,h,1},\ \mathbb{E}[D_{k,h,1} | \mathcal{F}_{k,h-1,1}] = 0,\ \mathbb{E}[D_{k,h,2} | \mathcal{F}_{k,h,1}] = 0. \tag{97}$$

Following from (97), we define the martingale

$$\mathcal{M}_{k,h,m} = \sum_{\substack{(\tau,i,l)\in[K]\times[H]\times[2] \\ t(\tau,i,l)\le t(k,h,m)}} D_{\tau,i,l}, \tag{98}$$

with respect to the time index $t(k, h, m)$ defined in (84). It is obvious that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}(D_{k,h,1} + D_{k,h,2}) = \mathcal{M}_{K,H,2} \tag{99}$$

Combining (85), (94), and (96) we obtain that

$$\begin{aligned}\sum_{k=1}^{K}\big(V_{1,\pi^{\mathrm{E}}}^{r^k}(x) - V_{1,\pi^k}^{r^k}(x)\big) &= \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^{\mathrm{E}}}\Big[\langle \widehat{Q}_h^k(s_h,\cdot), \pi_h^{\mathrm{E}}(\cdot\,|\,s_h) - \pi_h^k(\cdot\,|\,x_h)\rangle | s_1 = x\Big] \\ &\quad + \mathcal{M}_{K,H,2} + \sum_{k=1}^{K}\sum_{h=1}^{H}\Big(\mathbb{E}_{\pi^{\mathrm{E}}}\big[\iota_h^k(s_h^k, a_h^k) | s_1 = x\big] - \iota_h^k(s_h^k, a_h^k)\Big).\end{aligned} \tag{100}$$

By this, we conclude the proof of Lemma H.1. □

J.2. Proof of Lemma H.2

Proof. By the update rule of OGAPI in (14) (Lines 2–2 of Algorithm 2) and the property of the mirror descent, we have

$$\mathcal{L}_{k-1}(\pi^k) - \alpha^{-1}\cdot D(\pi^k, \pi^{k-1}) \ge \mathcal{L}_{k-1}(\pi^{\mathrm{E}}) - \alpha^{-1}\cdot D(\pi^{\mathrm{E}}, \pi^{k-1}) + \alpha^{-1}\cdot D(\pi^{\mathrm{E}}, \pi^k).$$

Recalling the definition of $\mathcal{L}_{k-1}(\pi)$ in (13) and rearranging the above inequality, we derive that

$$\begin{aligned}\sum_{h=1}^{H}\big\langle \widehat{Q}_h^{k-1}, \pi_h^{\mathrm{E}} - \pi_h^{k-1}\big\rangle_{\mathcal{A}} &\le \alpha^{-1}\cdot D(\pi^{\mathrm{E}}, \pi^{k-1}) - \alpha^{-1}\cdot D(\pi^{\mathrm{E}}, \pi^k) \\ &\quad + \sum_{h=1}^{H}\langle \widehat{Q}_h^{k-1}, \pi_h^k - \pi_h^{k-1}\rangle_{\mathcal{A}} - \alpha^{-1}\cdot D(\pi^k, \pi^{k-1}),\end{aligned} \tag{101}$$

where $D(\pi^k, \pi^{k-1}) = \sum_{h=1}^{H} D_{\mathrm{KL}}(\pi_h^k \| \pi_h^{k-1})$. For the last two terms on the right-hand side of (101), we have

$$\begin{aligned}&\sum_{h=1}^{H}\big\langle \widehat{Q}_h^{k-1}, \pi_h^k - \pi_h^{k-1}\big\rangle_{\mathcal{A}} - \alpha^{-1}\cdot D(\pi^k, \pi^{k-1}) \\ &\quad\le \sum_{h=1}^{H}\big(\|\widehat{Q}_h^{k-1}\|_{\mathcal{A},\infty}\cdot\|\pi_h^k - \pi_h^{k-1}\|_{\mathcal{A},1} - (2\alpha)^{-1}\cdot\|\pi_h^k - \pi_h^{k-1}\|_{\mathcal{A},1}^2\big) \\ &\quad\le \frac{\alpha}{2}\cdot\sum_{h=1}^{H}\|\widehat{Q}_h^{k-1}\|_{\mathcal{A},\infty}^2 \le \alpha H^3\sqrt{d}/2,\end{aligned}$$

where the first inequality follows from Holder's inequality and Pinsker's inequality, and the last inequality derives from the fact that $|r_h^{\mu}(\cdot,\cdot)| \leq \sqrt{d}$ for any $h \in [H]$ and $\mu \in S$. Since π^0 is a uniform distribution on $\mathcal{A}$, it holds that $D(\pi^{\mathrm{E}}, \pi^0) \leq H \log(\mathrm{vol}(\mathcal{A}))$. Telescoping (J.2) with respect to $k \in [K]$, we have

$$\begin{aligned} \sum_{k=1}^{K}\sum_{h=1}^{H}\big[\langle \widehat{Q}_h^{k-1}, \pi_h^{\mathrm{E}} - \pi_h^{k-1}\rangle_{\mathcal{A}}\big] &\leq \alpha K H^3\sqrt{d}/2 + \alpha^{-1} H D(\pi^{\mathrm{E}}, \pi^0) \\ &\leq \alpha K H^3\sqrt{d}/2 + \alpha^{-1} H \log(|\mathcal{A}|). \end{aligned} \tag{102}$$

Recalling that $\alpha = \sqrt{2\log(|\mathcal{A}|)/(H^2K\sqrt{d})}$ and taking expectation on both side of (102), we have

$$\begin{aligned} \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^{\mathrm{E}}}\big[\langle \widehat{Q}_h^{k-1}, \pi_h^{\mathrm{E}} - \pi_h^{k-1}\rangle_{\mathcal{A}} \,\big|\, s_1 = x\big] &\leq \alpha K H^3\sqrt{d}/2 + \alpha^{-1} H \log(|\mathcal{A}|) \\ &\leq \sqrt{2H^4\sqrt{d}K\log(\mathrm{vol}(\mathcal{A}))}. \end{aligned}$$

Then we conclude the proof of Lemma H.2. □

J.3. Proof of Lemma H.3

Proof. Recalling that we define $D_{k,h,1}$ and $D_{k,h,2}$ in (94) and the fact that $|r_h^{\mu}(\cdot,\cdot)| \leq \sqrt{d}$ for any $\mu \in S$, we derive that $|D_{k,h,1}| \leq 2H\sqrt{d}$ and $|D_{k,h,2}| \leq 2H\sqrt{d}$ for any $(k,h) \in [K]\times[H]$. Now by Azuma-Hoeffding inequality, we have

$$\mathbb{P}(|\mathcal{M}_{K,H,2}| > t) \leq 2\exp\Big(\frac{-t^2}{16H^3Kd}\Big), \tag{103}$$

for any $t > 0$. Setting $t = \sqrt{16H^3dK \cdot \log(8/\xi)}$ with $\xi \in (0,1)$ in (103), we have

$$\mathcal{M}_{K,H,2} \leq \sqrt{16H^3dK \cdot \log(8/\xi)},$$

with probability at least $1 - \xi/4$. □

J.4. Proof of Lemma H.4

Proof. For notational simplicity, we write $\bar{Q}_h^k(s,a) = r_h^k(s,a) + \widehat{\mathcal{P}}_h\widehat{V}_{h+1}^k(s,a) + \Gamma_h^k(s,a)$. Then, from the policy evaluation stage in Lines 2–2 of Algorithm 3, we have

$$\widehat{Q}_h^k(s,a) = \min\Big\{\max\big\{\bar{Q}_h^k(s,a), 0\big\}, (H-h+1)\sqrt{d}\Big\}. \tag{104}$$

We introduce the following lemma.

Lemma J.1. *Let $\lambda = 1$ in the construction of estimated kernels (17) and $\kappa = C\sqrt{d\log(HdK/\xi)}$ in the construction of bonus (18). Then it holds with probability at least $1-\xi/4$ that*

$$\big|\mathcal{P}_h\widehat{V}_{h+1}^k(s,a) - \widehat{\mathcal{P}}_h^k\widehat{V}_{h+1}^k(s,a)\big| \leq \Gamma_h^k(s,a)$$

for any $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Proof. See Appendix J.9 for a detailed proof. □

By Lemma J.1, we obtain that $r_h^k + \mathcal{P}_h\widehat{V}_{h+1}^k \leq \bar{Q}_h^k$. Moreover, by the fact that $|r_h^k(s,a)| \leq \sqrt{d}$ and $\widehat{V}_{h+1}^k(s) = \langle \widehat{Q}_h^k(s,\cdot), \pi_h^k(\cdot\,|\,s)\rangle_{\mathcal{A}} \in [0, (H-h)\sqrt{d}]$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have $r_h^k + \mathcal{P}_h V_{h+1,\pi^k}^k \in [0, (H-h+1)\sqrt{d}]$. Thus, we have

$$\begin{aligned} \widehat{Q}_h^k(s,a) &= \min\Big\{\max\big\{\bar{Q}_h^k(s,a), 0\big\}, (H-h+1)\sqrt{d}\Big\} \\ &\geq \min\Big\{\max\big\{r_h^k(s,a) + \mathcal{P}_h\widehat{V}_{h+1}^k(s,a), 0\big\}, (H-h+1)\sqrt{d}\Big\} \\ &= r_h^k(s,a) + \mathcal{P}_h\widehat{V}_{h+1}^k(s,a), \end{aligned}$$

which implies that $\iota_h^k \leq 0$.

It remains to establish the lower bound of $\iota_h^k(s,a)$. By Lemma J.1, we have

$$\begin{aligned}\bar{Q}_h^k(s,a) &= r_h^k(s,a) + \widehat{\mathcal{P}}_h \widehat{V}_{h+1}^k(s,a) + \Gamma_h^k(s,a) \\ &\geq r_h^k(s,a) + \mathcal{P}_h \widehat{V}_{h+1}^k(s,a) \geq 0,\end{aligned} \tag{105}$$

where the last inequality follows from the fact that $\widehat{V}_{h+1}^k(s,a) \geq 0$ and $r_h^k(s,a) \geq 0$. By (104) and (105), we obtain that $\widehat{Q}_h^k(s,a) \leq \bar{Q}_h^k(s,a)$, which implies that

$$\begin{aligned}\iota_h^k(s,a) &= (r_h^k + \mathcal{P}_h \widehat{V}_h^k)(s,a) - \widehat{Q}_h^k(s,a) \\ &\geq (\mathcal{P}_h - \widehat{\mathcal{P}}_h)\widehat{V}_h^k(s,a) - \Gamma_h^k(s,a) \\ &\geq -2\Gamma_h^k(s,a).\end{aligned}$$

Here the last inequality follows from Lemma J.1. Thus, we conclude the proof of Lemma I.2. □

J.5. Proof of Lemma H.5

Proof. By the construction of bonus Γ_h^k in (18), we have

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{K}\Gamma_h^k(s_h^k,a_h^k) &= H\sqrt{d}\cdot\sum_{h=1}^{H}\sum_{k=1}^{K}\min\left\{1, \kappa\cdot\varphi_h^k\left(s_h^k,a_h^k\right)^\top(\Lambda_h^k)^{-1}\varphi_h^k(s_h^k,a_h^k)\right\} \\ &\leq H\sqrt{d}\kappa\cdot\sum_{h=1}^{H}\left(K\cdot\sum_{k=1}^{K}\varphi_h^k\left(s_h^k,a_h^k\right)^\top(\Lambda_h^k)^{-1}\varphi_h^k(s_h^k,a_h^k)\right)^{1/2},\end{aligned} \tag{106}$$

where the last inequality comes from Cauchy-Schwarz inequality. To upper bound the right-hand side of (106), we introduce the following lemma.

Lemma J.2 (Elliptical Potential (Abbasi-Yadkori et al., 2011)). *Let $\{\phi_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued sequence. Meanwhile, let $\Lambda_0 \in \mathbb{R}^{d\times d}$ be a positive-definite matrix and $\Lambda_t = \Lambda_0 + \sum_{j=1}^{t-1}\phi_j\phi_j^\top$. It holds for any $t \in \mathbb{Z}_+$ that*

$$\sum_{j=1}^{t}\min\left\{1, \phi_j^\top\Lambda_j^{-1}\phi_j\right\} \leq 2\log\left(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\right).$$

Moreover, assuming that $\|\phi_j\|_2 \leq 1$ for any $j \in \mathbb{Z}_+$ and $\lambda_{\min}(\Lambda_0) \geq 1$, it holds for any $t \in \mathbb{Z}_+$ that

$$\log\left(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\right) \leq \sum_{j=1}^{t}\phi_j^\top\Lambda_j^{-1}\phi_j \leq 2\log\left(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\right).$$

Proof. See proof of Lemma 11 in (Abbasi-Yadkori et al., 2011) for a detailed proof. □

For any fixed $h \in [H]$, by Lemma J.2, we have

$$\sum_{k=1}^{K}\varphi_h^k\left(s_h^k,a_h^k\right)^\top(\Lambda_h^k)^{-1}\varphi_h^k(s_h^k,a_h^k) \leq 2\log\left(\frac{\det(\Lambda_h^{K+1})}{\det(\Lambda_h^1)}\right), \tag{107}$$

where $\Lambda_h^1 = \lambda\cdot I$ and $\Lambda_h^{K+1} \in \mathcal{F}_{K,H,2}$, which is defined in (83). By Assumption 2.1, we obtain that

$$\begin{aligned}\left\|\varphi_h^k(s,a)\right\|_2 &= \left\|\int_{\mathcal{S}}\phi(s,a,s')\widehat{V}_{h+1}^k(s')\mathrm{d}s'\right\|_2 \\ &\leq H\sqrt{d}\cdot\left\|\int_{\mathcal{S}}\phi(s,a,s')\mathrm{d}s'\right\|_2 \\ &\leq H\sqrt{d}\cdot\mathrm{Vol}(\mathcal{S})\cdot\sup_{s'\in\mathcal{S}}\|\phi(s,a,s')\|_2 \leq Hd^{3/2}R\cdot\mathrm{Vol}(\mathcal{S}).\end{aligned} \tag{108}$$

Here the first inequality comes from the fact that $\widehat{V}_h^k \in [0, H\sqrt{d}]$ for any $(k,h) \in [K] \times [H]$, and the last inequality comes from the fact that $\|\phi(\cdot,\cdot,\cdot)\|_2 \le dR$, which can be verified as follows,

$$\sup_{(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}} \|\phi(s,a,s')\|_2 = \sup_{(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}} \sqrt{\sum_{i=1}^d \|\phi(s,a,s')^\top \boldsymbol{e}_i\|_2^2} \le dR. \tag{109}$$

Here $\{\boldsymbol{e}_i\}_{i=1}^d$ is a group of orthonormal basis of $\mathbb{R}^d$ and the last inequality follows from Assumption 2.1. By the definition of Λ_h^k in (17), we can upper bound $\det(\Lambda_h^{K+1})$ by (108) as follows,

$$\begin{aligned} \det(\Lambda_h^{K+1}) &= \det\Big(\sum_{k=1}^K \varphi(s_h^k, a_h^k)\varphi(s_h^k, a_h^k)^\top + I\Big) \\ &\le \Big(\det\big((Hd^{3/2}R \cdot \mathrm{Vol}(\mathcal{S}) + 1)\cdot I\big)\Big)^d, \end{aligned} \tag{110}$$

which implies that

$$\log\left(\frac{\det\left(\Lambda_h^{K+1}\right)}{\det\left(\Lambda_h^1\right)}\right) \le 2d \cdot \log(H^2 d^3 R^2 K \cdot \mathrm{Vol}(\mathcal{S})^2). \tag{111}$$

Recalling that $\kappa = C\sqrt{d\log(HdK/\xi)}$, combining (106), (107), and (111), we have

$$\begin{aligned} \sum_{h=1}^H \sum_{k=1}^K \Gamma_h^k(s_h^k, a_h^k) &\le 4H\sqrt{d}\kappa \cdot H\sqrt{dK \cdot \log(H^2 d^3 R^2 K \cdot \mathrm{Vol}(\mathcal{S})^2)} \\ &\le C'\sqrt{H^4 d^3 K} \cdot \log(HdK/\xi), \end{aligned}$$

where C' is an absolute constant determined by C, R, and $\log(\mathrm{Vol}(\mathcal{S}))$. By this, we conclude the proof of Lemma H.5. □

J.6. Proof of Lemma H.6

Proof. By the definition of cumulative reward in (5), we observe that

$$\begin{aligned} J(\pi,\mu) &= \mathbb{E}_\pi \sum_{h=1}^H \left[r_h^\mu(s_h, a_h)\right] \\ &= \sum_{h=1}^H \mathbb{E}_\pi\left[r_h^\mu(s_h, a_h)\right] \\ &= \sum_{h=1}^H \int_{\mathcal{S}\times\mathcal{A}} \rho_h^\pi(s,a) \cdot r_h^\mu(s,a)\mathrm{d}s\mathrm{d}a, \end{aligned} \tag{112}$$

where $\rho_h^\pi(s,a) = \mathbb{P}(s_h = s, a_h = a)$ is the density of state-action visition measure on $\mathcal{S}\times\mathcal{A}$. Recall that under Assumption 2.1, we have $r_h^\mu(s,a) = \psi(s,a)^\top \mu_h$, hence we have

$$\nabla_{\mu_h} J(\pi^k, \mu^k) = \int_{\mathcal{S}\times\mathcal{A}} \rho_h^{\pi^k}(s,a)\cdot\psi(s,a)\mathrm{d}s\mathrm{d}a. \tag{113}$$

By (6), we obtain that

$$\begin{aligned} L(\pi^k,\mu) - L(\pi^k,\mu^k) = \sum_{h=1}^H (\mu_h - \mu_h^k)^\top \nabla_{\mu_h} L(\pi^k, \mu^k), \\ \text{where } \nabla_{\mu_h} L(\pi^k,\mu^k) = \nabla_{\mu_h} J(\pi^{\mathrm{E}}, \mu^k) - \nabla_{\mu_h} J(\pi^k,\mu^k). \end{aligned} \tag{114}$$

Combining (113) and (114), we know that $L(\pi,\mu)$ is a linear function in μ for any π. Recall that $\mu_h^{k+1} = \mathrm{Proj}_B\{\mu_h^k + \eta\widehat{\nabla}_{\mu_h} L(\pi^k,\mu^k)\}$ in OGAPI (Lines 2–2 of Algorithm 2), by the definition of the projection operator $\mathrm{Proj}_B(\cdot)$, it holds that

$$\left[\mu_h^{k+1} - \mu_h^k - \eta\widehat{\nabla}_{\mu_h} L(\pi^k,\mu^k)\right]^\top (\mu_h - \mu_h^{k+1}) \ge 0. \tag{115}$$

Rearranging terms in (115), we obtain that

$$\begin{aligned}\eta(\mu_h - \mu_h^{k+1})^\top \widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k) &\le (\mu_h^{k+1} - \mu_h^k)^\top (\mu_h - \mu_h^{k+1}) \\ &= \frac{1}{2}\Big(\|\mu_h^k - \mu_h\|_2^2 - \|\mu_h^{k+1} - \mu_h\|_2^2 - \|\mu_h^{k+1} - \mu_h^k\|_2^2\Big),\end{aligned} \tag{116}$$

which also implies that

$$\frac{1}{2}\Big(\|\mu_h^k - \mu_h\|_2^2 - \|\mu_h^{k+1} - \mu_h\|_2^2 - \|\mu_h^{k+1} - \mu_h^k\|_2^2\Big) - \eta(\mu_h - \mu_h^{k+1})^\top \widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k) \ge 0. \tag{117}$$

By adding a term $\eta \nabla_{\mu_h} L(\pi^k, \mu^k)^\top (\mu_h - \mu_h^k)$ on both sides of (117) and combining (114), we obtain that

$$\begin{aligned}L(\pi^k, \mu) - L(\pi^k, \mu^k) \le \sum_{h=1}^H \frac{1}{2\eta}(\|\mu_h^k - \mu_h\|_2^2 - \|\mu_h^{k+1} - \mu_h\|_2^2 - \|\mu_h^{k+1} - \mu_h^k\|_2^2) \\ + \sum_{h=1}^H \big[(\mu_h^{k+1} - \mu_h^k)^\top \widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k)\big] \\ + \sum_{h=1}^H \big[(\mu_h^k - \mu_h)^\top (\widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k) - \nabla_{\mu_h} L(\pi^k, \mu^k))\big],\end{aligned} \tag{118}$$

where we take the summation on h from 1 to H. By the fact that $\widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k) = \nabla_{\mu_h} \widetilde{J}(\pi^{\mathrm{E}}, r^\mu) - \psi(s_h^k, a_h^k)$ and the definition of the GAIL objective function $L(\pi, \mu)$ in (6), we rewrite the third term on the right-hand side of (118) as

$$\begin{aligned}\sum_{h=1}^H \big[(\mu_h^k - \mu_h)^\top (\widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k) - \nabla_{\mu_h} L(\pi^k, \mu^k))\big] &= \sum_{h=1}^H \big[\mu_h^\top (\nabla_{\mu_h} \widetilde{J}(\pi^k, \mu^k) - \nabla_{\mu_h} J(\pi^{\mathrm{E}}, r^k))\big] \\ &+ \sum_{h=1}^H \big[(\mu_h^k - \mu_h)^\top (\psi(s_h^k, a_h^k) - \nabla_{\mu_h} J(\pi^k, \mu^k))\big].\end{aligned} \tag{119}$$

By (113) and the definition of $\widetilde{J}(\pi^{\mathrm{E}}, r^\mu)$ in (23), we derive from (118) and (119) that

$$\begin{aligned}L(\pi^k, \mu) - L(\pi^k, \mu^k) \le \sum_{h=1}^H \frac{1}{2\eta}\big(\|\mu_h^k - \mu_h\|_2^2 - \|\mu_h^{k+1} - \mu_h\|_2^2 - \|\mu_h^{k+1} - \mu_h^k\|_2^2\big) \\ + \sum_{h=1}^H \big[(\mu_h^{k+1} - \mu_h^k)^\top \widehat{\nabla}_{\mu_h} L(\pi^k, \mu^k)\big] + \big[\widetilde{J}(\pi^{\mathrm{E}}, r^\mu) - J(\pi^{\mathrm{E}}, r^\mu)\big] \\ + \sum_{h=1}^H \big[(\mu_h^k - \mu_h)^\top (\psi(s_h^k, a_h^k) - \nabla_{\mu_h} J(\pi^k, \mu^k))\big].\end{aligned}$$

Upon telescoping sum on the below inequality for the index $k \in [K]$, we complete the proof of Lemma H.6. □

J.7. Proof of Lemma H.7

Proof. For any fixed reward parameter $\mu \in S$, we define $J^\tau(\pi^{\mathrm{E}}, r^\mu) = \sum_{h=1}^H \psi(s_{h,\tau}^{\mathrm{E}}, a_{h,\tau}^{\mathrm{E}})^\top \mu_h$ for any $\tau \in [N_1]$. Since the expert demonstration $\mathbb{D}^{\mathrm{E}} = \{(s_{h,k}^{\mathrm{E}}, a_{h,k}^{\mathrm{E}})\}_{(k,h)\in[N_1]\times[H]}$ involves N_1 independent trajectories induced by the expert policy π^{E}, we apply Monte Carlo method to estimate $J(\pi^{\mathrm{E}}, r^\mu)$ by N_1 i.i.d. samples $\{J^\tau(\pi^{\mathrm{E}}, r^\mu)\}_{\tau=1}^{N_1}$.

Let $Z_n = \sum_{\tau=1}^n \big(J^\tau(\pi^{\mathrm{E}}, r^\mu) - J(\pi^{\mathrm{E}}, r^\mu)\big)$ and we have $|Z_n - Z_{n-1}| \le 2H\sqrt{d}$, since $|r_h(\cdot,\cdot)| \le \sqrt{d}$ for all $h \in [H]$. Note that $\{Z_n\}$ is a martingale with zero mean with respect to the filtration $\mathcal{F}_n = \sigma\big(\{s_h^i, a_h^i\}_{(h,i)\in[H]\times[n]}\big)$, by Azuma-Hoffeding inequality, we have

$$\mathbb{P}_{\mathbb{D}}(|Z_n| > t) \le 2\exp\left(\frac{-2t^2}{4H^2 dn}\right),$$

which implies that

$$\mathbb{P}_{\mathbb{D}}\left(\left|\frac{Z_{N_1}}{N_1}\right| > m\right) \leq 2\exp\left(\frac{-2m^2N_1}{4H^2d}\right).$$

Let $\delta = 2\exp\{-m^2N_1/(2H^2d)\}$, it holds with probability at least $1-\delta$ that

$$\left|\widetilde{J}(\pi^{\mathrm{E}}, r^{\mu}) - J(\pi^{\mathrm{E}}, r^{\mu})\right| = \left|\frac{Z_{N_1}}{N_1}\right| \leq H\sqrt{2d\log(2/\delta)/N_1}. \tag{120}$$

We union bound $\left|\widetilde{J}(\pi^{\mathrm{E}}, r^{\mu}) - J(\pi^{\mathrm{E}}, r^{\mu})\right|$ for any $\mu \in S$ as follows. Since the reward parameter domain S defined in (10) is not a finite set, we apply discretization on S to derive a union bound on $\left|\widetilde{J}(\pi^{\mathrm{E}}, r^{\mu}) - J(\pi^{\mathrm{E}}, r^{\mu})\right|$. If we define a normed space $(\mathbb{R}^{Hd}, \|\cdot\|_\star)$, where $\|\cdot\|_\star$ is defined as

$$\|\mu\|_\star = \sup_{h\in[H]} \|\mu_h\|_2, \tag{121}$$

then parameter domain S belongs to this normed space. Before we continue, we first introduce the definitions of ϵ-covering and covering number as follows.

Definition J.3 (ϵ-covering)**.** Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \subset V$. We say that $\{V_1, \ldots, V_N\}$ is an ϵ-covering of Θ if $\Theta \subset \cup_{i=1}^{N} B(V_i, \epsilon)$, or equivalently, $\forall\theta \in \Theta$, $\exists i$ such that $\|\theta - V_i\| \leq \epsilon$. Here $B(V_i, \epsilon)$ denotes a ball centering V_i with radius ϵ.

We define the covering number as follows,

$$\mathcal{N}(\Theta, \|\cdot\|, \epsilon) := \min\left\{n : \exists\epsilon \text{ -covering over } \Theta \text{ of size } n, \Theta \in (V, \|\cdot\|)\right\}.$$

With Definition J.3, we introduce the following lemma to upper bound the covering number.

Lemma J.4. *If $(V, \|\cdot\|)$ is a normed space, and (i) $\Theta \subset V = \mathbb{R}^d$, (ii) Θ is convex, (iii) $\epsilon B_{\mathrm{unit}} \in \Theta$, where $\epsilon > 0$ and B_{unit} is the unit ball in $\mathbb{R}^d$, then it holds that*

$$\mathcal{N}(\Theta, \|\cdot\|, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^d \frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(B_{\mathrm{unit}})}.$$

Proof. See Lemma 5.2 of (Vershynin, 2010) for proof. □

Note that S is convex as a subset of $\mathbb{R}^{Hd}$, we apply Lemma J.4 with $V = \mathbb{R}^{Hd}$, $\Theta = S$, $\|\cdot\| = \|\cdot\|_\star$, and an appropriate $\epsilon > 0$ satisfing condition (iii) in Lemma J.4, which implies that

$$\mathcal{N}(S, \|\cdot\|_\star, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^{Hd} d^{Hd/2}.$$

By the definition of covering number, there exists an ϵ-covering $\mathcal{V}_\epsilon = \{\mu^1, ..., \mu^{\mathcal{N}(S,\|\cdot\|_\star,\epsilon)}\} \subset S$. For each $\mu \in \mathcal{V}_\epsilon$, by (120), it holds that

$$\left|\widetilde{J}(\pi^{\mathrm{E}}, r^{\mu}) - J(\pi^{\mathrm{E}}, r^{\mu})\right| \leq H\sqrt{2d\log(2\mathcal{N}(S, \|\cdot\|_\star, \epsilon)/\xi)/N_1},$$

with probability at least $1 - \xi/\mathcal{N}(S, \|\cdot\|_\star, \epsilon)$. By the union bound, it yields that

$$\begin{aligned}\sup_{\mu\in\mathcal{V}_\epsilon} \left|\widetilde{J}(\pi^{\mathrm{E}}, r^{\mu}) - J(\pi^{\mathrm{E}}, r^{\mu})\right| &\leq H\sqrt{2d\log(2\mathcal{N}(S, \|\cdot\|_\star, \epsilon)/\xi)/N_1} \\ &\leq H\sqrt{\big(Hd^2\log(d) + 2Hd^2\log(3/\epsilon) + 2d\log(2/\xi)\big)/N_1},\end{aligned} \tag{122}$$

with probability at least $1-\xi$. Note that for any $\mu', \mu'' \in S$ satisfying $\|\mu' - \mu''\|_\star \leq \epsilon$, it holds that

$$\left|\left[\widetilde{J}(\pi^{\mathrm{E}}, r^{\mu'}) - J(\pi^{\mathrm{E}}, r^{\mu'})\right] - \left[\widetilde{J}(\pi^{\mathrm{E}}, r^{\mu''}) - J(\pi^{\mathrm{E}}, r^{\mu''})\right]\right| \leq 4H\epsilon. \tag{123}$$

Combining (122) and (123) and applying triangle inequality, we derive that

$$\sup_{\mu\in S}\big|\widetilde{J}(\pi^{\mathrm{E}},r^{\mu})-J(\pi^{\mathrm{E}},r^{\mu})\big|\le H\sqrt{\big(Hd^2\log(d)+2Hd^2\log(3/\epsilon)+2d\log(2/\xi)\big)/N_1}+4H\epsilon, \tag{124}$$

with probability at least $1-\xi$. By taking $\epsilon=\sqrt{6d/N_1}$ in (124), which satisfies the conditions in Lemma J.4, it holds with probability at least $1-\xi$ that

$$\begin{aligned}\sup_{\mu\in S}\big|\widetilde{J}(\pi^{\mathrm{E}},r^{\mu})-J(\pi^{\mathrm{E}},r^{\mu})\big| &\le H\sqrt{\big(Hd^2\log(d)+Hd^2\log(\frac{2N_1}{d})+2d\log(2/\xi)\big)/N_1}+4H\sqrt{\frac{d}{N_1}},\\ &\le 4\sqrt{H^3d^2/N_1}\log(6N_1/\xi).\end{aligned}$$

We conclude the proof of Lemma H.7. □

J.8. Proof of Lemma H.8

Proof. First, we arbitrarily fix $\mu\in S$. We define for $(k,h)\in[K]\times[H]$ that

$$\begin{aligned}X_h^k&=(\mu_h^k-\mu_h)^\top(-\psi(s_h^k,a_h^k)+\nabla_{\mu_h}J(\pi^k,\mu^k)),\\ Y^k&=\sum_{i=1}^{k}\sum_{h=1}^{H}X_h^i,\\ S_h^k&=\sigma\big((s_h^1,a_h^1),(s_h^2,a_h^2),\ldots,(s_k^h,a_k^h)\big)\big),\\ E_h&=\sigma\big((s_{h,1}^{\mathrm{E}},a_{h,1}^{\mathrm{E}}),(s_{h,2}^{\mathrm{E}},a_{h,2}^{\mathrm{E}}),\cdots,(s_{h,N_1}^{\mathrm{E}},a_{h,N_1}^{\mathrm{E}})\big),\\ G_h^k&=\sigma(S_h^k,E_h),\\ G^k&=\sigma(G_1^k,G_2^k,\ldots,G_H^k),\end{aligned}\tag{125}$$

where $\sigma(\cdot)$ denotes the generated σ-algebra. It holds that $\{G^k\}_{k\in[K]}$ is a filtration with respect to the time index k, since $G^{k_1}\subseteq G^{k_2}$ for $k_1\le k_2$.

We first show that $X_h^k\in G^k$ holds for any $(k,h)\in[K]\times[H]$. By the definition of X_h^k in (125), it only suffices to prove that $\mu_h^k\in G^k$ for any $(k,h)\in[K]\times[H]$. Here we show this by induction with index k. Since $\mu_h^1=\mathbf{0}$ for any $h\in[H]$, the base case where $k=1$ is trival. We assume that $\mu_h^k\in G^k$ where $k\ge 1$ is a given integer, then we consider the case $k+1$. Recall that the update process of reward parameter in OGAPI (Lines 2–2 of Algorithm 2) takes the following form,

$$\begin{aligned}\mu_h^{k+1}&=\mathrm{Proj}_B\{\mu_h^k+\eta\widehat{\nabla}_{\mu_h}L(\pi^k,\mu^k)\}\\ &=\mathrm{Proj}_B\Big\{\mu_h^k+\eta\cdot\Big[\frac{1}{N_1}\sum_{\tau=1}^{N_1}\psi(s_{h,\tau}^{\mathrm{E}},a_{h,\tau}^{\mathrm{E}})-\psi(s_h^k,a_h^k)\Big]\Big\}.\end{aligned}$$

First, according to the induction hypothesis, we have $\mu_h^k\in G^k\subseteq G^{k+2}$, which implies that $(s_h^k,a_h^k)\in G^{k+1}$. Then it holds that

$$\mu_h^k+\frac{1}{N_1}\sum_{\tau=1}^{N_1}\psi(s_{h,\tau}^{\mathrm{E}},a_{h,\tau}^{\mathrm{E}})-\psi(s_h^k,a_h^k)\in G^{k+1},$$

for any $h\in[H]$. As Proj_B is a continous operator, we obtain that $\mu_h^{k+1}\in G^{k+1}$. Thus we complete the induction.

Now we construct a martingale to upper bound Y^K. Note that conditioning on the filtration G^{k-1}, the term $\mu_h^k-\mu_h$ is a constant. Recall that in (113) we show that

$$\nabla_{\mu_h}J(\pi^k,\mu^k)=\int_{\mathcal{S}\times\mathcal{A}}\psi(s,a)\cdot\rho_h^{\pi^k}(s,a)\mathrm{d}s\mathrm{d}a,$$

which implies that

$$\mathbb{E}_k(X_h^k|G^{k-1})=(\mu_h^k-\mu_h)^\top\mathbb{E}_k(\psi(s_h^k,a_h^k)-\nabla_{\mu_h}J(\pi^k,\mu^k)|G^{k-1})=0 \tag{126}$$

for any $(k,h) \in [K] \times [H]$. Here the expectation $\mathbb{E}_k$ is taken with respect to $a_i^k \sim \pi_i^k(\cdot \,|\, s_i^k)$ and $s_{i+1}^k \sim \mathcal{P}_i(\cdot \,|\, s_i^k, a_i^k)$, corresponding to the expectation taken with respect to the state-action visitation measure $\rho_h^{\pi^k}$ defined in (112). Also, we have $Y^k \in G^k$, since $X_h^k \in G^k$ for any $(k,h) \times [K] \times [H]$. Moreover, we obtain for any $k \in [K]$ that

$$\begin{aligned}\mathbb{E}_k(Y^k|G^{k-1}) &= \mathbb{E}_k\big(\sum_{i=1}^{k}\sum_{h=1}^{H} X_h^i \,\big|\, G^{k-1}\big)\\ &= \mathbb{E}_k\big(\sum_{h=1}^{H} X_h^k + \sum_{i=1}^{k-1}\sum_{h=1}^{H} X_h^i \,\big|\, G^{k-1}\big)\\ &= \mathbb{E}_k\big(\sum_{h=1}^{k} X_h^k \,\big|\, G^{k-1}\big) + Y^{k-1} = Y^{k-1},\end{aligned}$$

where the last equality follows from (126). Thus $\{Y^k\}_{k=1}^K$ is a martingale. Furthermore, by the definition of X_h^k in (125), it holds that

$$|Y^k - Y^{k-1}| = \big|\sum_{h=1}^{H} X_h^k\big| \le \sum_{h=1}^{H} \|\mu_h^k - \mu_h\|_2 \|\psi(s_h^k, a_h^k) - \nabla_{\mu_h} J(\pi^k, \mu^k)\|_2 \le 8\sqrt{d}H,$$

where the last inequality follows from the facts that $\|\psi(\cdot,\cdot)\|_2 \le 1$ and $\|\mu_h^k\|_2 \le \sqrt{d}$. Therefore, by Azuma-Hoeffding inequality, we obtain that

$$\mathbb{P}(|Y^K| \ge t) \le \exp\left(\frac{-t^2}{2\sum_{k=1}^K (8\sqrt{d}H)^2}\right) = \exp\Big(\frac{-t^2}{128KdH^2}\Big)$$

for any $t > 0$. Setting $t = \sqrt{128H^2dK\log(2/\xi)}$ with $\xi \in (0,1)$ and by the definition of Y^k in (125), it holds with probability at least $1-\xi$ that

$$\Big|\sum_{k=1}^{K}\sum_{h=1}^{H} (\mu_h^k - \mu_h)^\top(-\psi(s_h^k, a_h^k) + \nabla_{\mu_h} J(\pi^k, \mu^k))\Big| \le 8\sqrt{2H^2dK\log(2/\xi)}. \tag{127}$$

Now we union bound (127). This is similar to the proof of Lemma H.7 in Appendix J.7. By applying Lemma J.4 in the same normed space $(\mathbb{R}^{Hd}, \|\cdot\|_\star)$, it holds with probability at least $1-\xi$ that

$$\begin{aligned}\sup_{\mu\in\mathcal{V}_\epsilon} |M(\mu)| &= \Big|\sum_{k=1}^{K}\sum_{h=1}^{H} (\mu_h^k - \mu_h)^\top(-\psi(s_h^k, a_h^k) + \nabla_{\mu_h} J(\pi^k, \mu^k))\Big|\\ &\le 8\sqrt{\big(H^3d^2\log(d) + 2Hd^2\log(3/\epsilon) + 2d\log(2/\xi)\big)K},\end{aligned} \tag{128}$$

where $\mathcal{V}_\epsilon$ is the ϵ-covering for S in Definition J.3 and $\|\cdot\|_\star$ is defined in (121). Here for notational convenience, we denote by $M(\mu) = \sum_{k=1}^K\sum_{h=1}^H (\mu_h^k - \mu_h)^\top(-\psi(s_h^k, a_h^k) + \nabla_{\mu_h} J(\pi^k, \mu^k))$. For any $\mu', \mu'' \in S$ satisfying $\|\mu' - \mu''\|_\star \le \epsilon$, it holds that

$$|M(\mu') - M(\mu'')| \le 4HK\epsilon. \tag{129}$$

Combining (128) and (129) and applying triangle inequality, we have that

$$\sup_{\mu\in S} |M(\mu)| \le 8\sqrt{\big(H^3d^2\log(d) + 2Hd^2\log(3/\epsilon) + 2d\log(2/\xi)\big)K} + 4HK\epsilon, \tag{130}$$

with probability at least $1-\xi$. By taking $\epsilon = \sqrt{d/K}$ in (130), which satisfies that $\epsilon B_{\text{unit}} \subset S$, we derive that

$$\begin{aligned}\sup_{\mu\in S} |M(\mu)| &\le 8\sqrt{\big(H^3d^2\log(d) + 2Hd^2\log(9K/d) + 2d\log(2/\xi)\big)K} + 4H\sqrt{dK},\\ &\le 32\sqrt{H^3d^2K}\log(9K/\xi),\end{aligned}$$

with probability at least $1-\xi$. Hence we conclude the proof of Lemma H.8. □

J.9. Proof of Lemma J.1

Proof. Under Assumption 2.1 and by the definition of Λ_h^k in (17), we have

$$\begin{aligned}(\mathcal{P}_h \widehat{V}_{h+1}^k)(x,a) &= \varphi_h^k(s,a)^\top (\Lambda_h^k)^{-1}\big(\sum_{\tau=1}^{k-1}\varphi_h^\tau(s_h^\tau,a_h^\tau)\varphi_h^\tau(s_h^\tau,a_h^\tau)^\top\theta_h + \lambda\cdot\theta_h\big)\\ &= \varphi_h^k(s,a)^\top (\Lambda_h^k)^{-1}\big(\sum_{\tau=1}^{k-1}\varphi_h^\tau(s_h^\tau,a_h^\tau)\cdot(\mathcal{P}_h \widehat{V}_{h+1}^\tau)(s_h^\tau,a_h^\tau) + \lambda\cdot\theta_h\big).\end{aligned} \tag{131}$$

Note that $\widehat{\mathcal{P}}_h \widehat{V}_h^k(s,a) = \varphi(s,a)^\top\widehat{\theta}_h^k$ by the closed form of $\widehat{\theta}_h^k$ in (138), we obtain that

$$\begin{aligned}&\varphi_h^k(s,a)^\top\widehat{\theta}_h^k - (\mathcal{P}_h \widehat{V}_{h+1}^k)(s,a)\\ &\quad= \underbrace{\varphi_h^k(s,a)^\top(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\varphi_h^\tau(s_h^\tau,a_h^\tau)\cdot\big(\widehat{V}_{h+1}^\tau(s_{h+1}^\tau) - (\mathcal{P}_h\widehat{V}_{h+1}^\tau)(s_h^\tau,a_h^\tau)\big)\Big)}_{\text{(i)}}\\ &\qquad - \underbrace{\lambda\cdot\varphi_h^k(s,a)^\top(\Lambda_h^k)^{-1}\theta_h}_{\text{(ii)}},\end{aligned} \tag{132}$$

for any $(s,a)\in\mathcal{S}\times\mathcal{A}$. To upper bound the norm of term (i) in (132), we introduce the following lemma.

Lemma J.5. *Let $\lambda = 1$ in the construction of estimated kernels (17). It holds with probability at least $1-\delta/4$ that*

$$\Big\|\sum_{\tau=1}^{k-1}\varphi_h^\tau\left(s_h^\tau,a_h^\tau\right)\cdot\Big(\widehat{V}_{h+1}^\tau(s_{h+1}^\tau) - \Big(\mathcal{P}_h\widehat{V}_{h+1}^\tau\Big)\left(s_h^\tau,a_h^\tau\right)\Big)\Big\|_{(\Lambda_h^k)^{-1}} \le C\sqrt{H^2d^2\cdot\log(HdK/\delta)}$$

for any $(k,h)\in[K]\times[H]$, where $C>0$ is an absolute constant and $\delta\in(0,1)$.

Proof. See Appendix J.10 for a detailed proof. □

By applying Cauchy-Schwarz inequality and Lemma J.5 on term (i) in (132), we have

$$\begin{aligned}|\text{(i)}| &\le \|\varphi(s,a)\|_{(\Lambda_h^k)^{-1}}\cdot\Big\|\sum_{\tau=1}^{k-1}\varphi_h^\tau\left(s_h^\tau,a_h^\tau\right)\cdot\Big(\widehat{V}_{h+1}^\tau(s_{h+1}^\tau) - \Big(\mathcal{P}_h\widehat{V}_{h+1}^\tau\Big)\left(s_h^\tau,a_h^\tau\right)\Big)\Big\|_{(\Lambda_h^k)^{-1}}\\ &\le H\sqrt{d}\cdot C\sqrt{d\log(HdK/\xi)}\cdot\|\varphi(s,a)\|_{(\Lambda_h^k)^{-1}}\end{aligned} \tag{133}$$

with probability at least $1-\xi/4$.

For term (ii) in (132), by setting $\lambda=1$, we obtain that

$$\begin{aligned}|\text{(ii)}| &\le \|\varphi(s,a)\|_{(\Lambda_h^k)^{-1}}\cdot\|\theta_h\|_{(\Lambda_h^k)^{-1}}\\ &\le \sqrt{d}\cdot\|\varphi(s,a)\|_{(\Lambda_h^k)^{-1}},\end{aligned} \tag{134}$$

where the last inequality follows from the fact that $\|(\Lambda_h^k)^{-1}\|_2 \le 1$ and $\|\theta_h\|_2\le\sqrt{d}$ for any $h\in[H]$. Combining (131), (132), (133), and (134), it holds with probability at least $1-\xi/4$ that

$$\begin{aligned}\big|\mathcal{P}_h\widehat{V}_{h+1}^k(s,a) - \widehat{\mathcal{P}}_h^k\widehat{V}_{h+1}^k(s,a)\big| &\le C\sqrt{d\log(HdK/\xi)}\cdot\|\varphi(s,a)\|_{(\Lambda_h^k)^{-1}}\\ &\le H\sqrt{d}\kappa\cdot\|\varphi(s,a)\|_{(\Lambda_h^k)^{-1}} \le \Gamma_h^k(s,a)\end{aligned}$$

for any $h\in[H]$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$. Here $\kappa = C\sqrt{d\log(HdK/\xi)}$ is the scaling parameter in (18) with an absolute constant $C>0$. Then we conclude the proof of Lemma J.1. □

J.10. Proof of Lemma J.5

Proof. The proof of Lemma J.5 is adapted from that of Lemma D.1 in (Cai et al., 2020).

Lemma J.6 (Concentration of Self-Normalized Process). *Let $\{\widetilde{\mathcal{F}}_t\}_{t=0}^{\infty}$ be a filtration and $\{\eta_t\}_{t=1}^{\infty}$ be an $\mathbb{R}$ -valued stochastic process such that η_t is $\widetilde{\mathcal{F}}_t$ -measurable for any $t \geq 0$. Moreover, we assume that, for any $t \geq 0$, conditioning on $\widetilde{\mathcal{F}}_t, \eta_t$ is a zero-mean and σ-sub-Gaussian random variable with the variance proxy $\sigma^2 > 0$, that is,*

$$\mathbb{E}\big[\exp(\lambda\eta_t)\,|\,\widetilde{\mathcal{F}}_t\big] \leq e^{\lambda^2\sigma^2/2}$$

for any $\lambda \in \mathbb{R}$. Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$ -valued stochastic process such that X_t is $\widetilde{\mathcal{F}}_t$ -measurable for any $t \geq 0$. Also, let $Y \in \mathbb{R}^{d\times d}$ be a deterministic and positive-definite matrix. For any $t \geq 0$, we define

$$\bar{Y}_t = Y + \sum_{s=1}^{t} X_s X_s^{\top}, \quad S_t = \sum_{s=1}^{t} \eta_s \cdot X_s.$$

For any $\delta > 0$, it holds with probability at least $1-\delta$ that

$$\|S_t\|_{\bar{Y}_t^{-1}}^2 \leq 2\sigma^2 \cdot \log\left(\frac{\det\left(\bar{Y}_t\right)^{1/2}\det(Y)^{-1/2}}{\delta}\right)$$

for any $t \geq 0$.

Proof. See Theorem 1 of (Abbasi-Yadkori et al., 2011) for a detailed proof. □

By the definition of filtration $\{\mathcal{F}_{k,h,m}\}_{(k,h,m)\in[K]\times[H]\times[2]}$ in (83) and Markov property, we have

$$\mathbb{E}\big[\widehat{V}_{h+1}^{\tau}(s_{h+1}^{\tau})\,\big|\,\mathcal{F}_{\tau,h,1}\big] = \big(\mathcal{P}_h\widehat{V}_{h+1}^{\tau}\big)(s_h^{\tau}, a_h^{\tau}). \tag{135}$$

Conditioning on $\mathcal{F}_{\tau,h,1}$, the only randomness comes from s_{h+1}^{τ}, while $\widehat{V}_{h+1}^{\tau}$ is a deterministic function determined by $\widehat{Q}_{h+1}^{\tau}$ and π_{h+1}^{τ}, which are further determined by the historical data in $\mathcal{F}_{\tau,h,1}$. For simipliciity of notations, we define $\eta_{\tau,h} = \widehat{V}_{h+1}^{\tau}\left(s_{h+1}^{\tau}\right) - \left(\mathcal{P}_h\widehat{V}_{h+1}^{\tau}\right)\left(s_h^{\tau}, a_h^{\tau}\right)$. By (135), conditioning on $\mathcal{F}_{\tau,h,1}, \eta_{\tau,h}$ is a zero-mean random variable. Moreover, as $\widehat{V}_{h+1}^{\tau} \in [0, H\sqrt{d}]$, conditioning on $\mathcal{F}_{\tau,h,1}$, $\eta_{\tau,h}$ is an $(H\sqrt{d}/2)$-sub-Gaussian random variable defined in Lemma J.6. Meanwhile, $\eta_{\tau,h}$ is $\mathcal{F}_{k,h,2}$ -measurable, since $\mathcal{F}_{\tau,h,1} \subseteq \mathcal{F}_{k,h,2}$ for any $\tau \in [k-1]$. Hence, for any fixed $h \in [H]$, by Lemma J.6, it holds with probability at least $1 - \delta/(4H)$ that

$$\begin{aligned}&\Big\|\sum_{\tau=1}^{k-1}\varphi_h^{\tau}\left(s_h^{\tau}, a_h^{\tau}\right)\cdot\Big(\widehat{V}_{h+1}^{\tau}(s_{h+1}^{\tau}) - \big(\mathcal{P}_h\widehat{V}_{h+1}^{\tau}\big)(s_h^{\tau}, a_h^{\tau})\Big)\Big\|_{\left(\Lambda_h^k\right)^{-1}}^2 \\ &\qquad \leq \frac{H^2 d}{2}\big(\frac{1}{2}\log(\det(\Lambda_h^k)) - \frac{1}{2}\log(\det(I)) + \log(4H/\delta)\big).\end{aligned} \tag{136}$$

Recall that in (110) we derive that

$$\det(\Lambda_h^{K+1}) \leq \Big(\det\big((Hd^{3/2}R\cdot\mathrm{Vol}(\mathcal{S})+1)\cdot I\big)\Big)^d. \tag{137}$$

By plugging (137) into (136) and a union bound argument, we obtain with probability at least $1-\delta/2$ that

$$\begin{aligned}&\Big\|\sum_{\tau=1}^{k-1}\varphi_h^{\tau}\left(s_h^{\tau}, a_h^{\tau}\right)\cdot\Big(\widehat{V}_{h+1}^{\tau}(s_{h+1}^{\tau}) - \big(\mathcal{P}_h\widehat{V}_{h+1}^{\tau}\big)(s_h^{\tau}, a_h^{\tau})\Big)\Big\|_{\left(\Lambda_h^k\right)^{-1}}^2 \\ &\qquad \leq \frac{H^2 d}{2}\big(d\cdot\log((Hd^{3/2}R\cdot\mathrm{Vol}(\mathcal{S})+1) + \log(4H/\delta)\big),\end{aligned}$$

which implies that

$$\Big\|\sum_{\tau=1}^{k-1}\varphi_h^{\tau}\left(s_h^{\tau}, a_h^{\tau}\right)\cdot\Big(\widehat{V}_{h+1}^{\tau}(s_{h+1}^{\tau}) - \big(\mathcal{P}_h\widehat{V}_{h+1}^{\tau}\big)(s_h^{\tau}, a_h^{\tau})\Big)\Big\|_{\left(\Lambda_h^k\right)^{-1}}^2 \leq C''\sqrt{H^2d^2\cdot\log(HdK/\delta)},$$

for any $(k,h) \in [K]\times[H]$ with probability at least $1-\delta/4$. Here $C'' > 0$ is an absolute constant. By this, we conclude the proof of Lemma J.5. □

K. Proofs of Supporting Lemmas: PGAP

K.1. Proof of Lemma E.1

Proof. We show that $\{\Gamma_h\}_{h=1}^H$ constructed in Lemma E.1 are the ξ-uncertainty qualifiers for the initially estimated transition kernels $\{\widetilde{\mathcal{P}}_h\}_{h=1}^H$ constructed in (35). By the definition of Λ_h in (35), we have

$$\begin{aligned}\mathcal{P}_h(s'\,|\,s,a) &= \phi(s,a,s')^\top\theta_h \\ &= \phi(s,a,s')^\top\Lambda_h^{-1}(\sum_{\tau=1}^{N_2}\int_{\mathcal{S}}\phi(s_h^\tau,a_h^\tau,s')\mathcal{P}_h(s'\,|\,s_h^\tau,a_h^\tau)\mathrm{d}s' + \lambda\cdot\theta_h).\end{aligned} \tag{138}$$

By (138), we have

$$\begin{aligned}&\mathcal{P}_h(s'\,|\,s,a) - \widetilde{\mathcal{P}}_h(s'\,|\,s,a) \\ &\quad = \mathcal{P}_h(s'\,|\,s,a) - \phi(s,a,s')^\top\widetilde{\theta}_h \\ &\quad = \underbrace{\phi(s,a,s')^\top\Lambda_h^{-1}\Big(\sum_{\tau=1}^{N_2}\big(\int_{\mathcal{S}}\phi(s_h^\tau,a_h^\tau,s')\mathcal{P}_h(s'\,|\,s_h^\tau,a_h^\tau)\mathrm{d}s' - \phi(s_h^\tau,a_h^\tau,s_{h+1}^\tau)\big)\Big)}_{\text{(i)}} \\ &\qquad + \underbrace{\lambda\cdot\phi(s,a,s')^\top\Lambda_h^{-1}\theta_h}_{\text{(ii)}}.\end{aligned} \tag{139}$$

We introduce the following lemma to upper bound term (i) on the RHS of (139).

Lemma K.1. *Let $\lambda = 1$ in the construction of $\widetilde{\mathcal{P}}_h$ and Γ_h in (35) and (36). By Assumption 2.1 , the event that*

$$\Big\|\sum_{\tau=1}^{N_2}\big(\int_{\mathcal{S}}\phi(s_h^\tau,a_h^\tau,s')\mathcal{P}_h(s'\,|\,s_h^\tau,a_h^\tau)\mathrm{d}s' - \phi(s_h^\tau,a_h^\tau,s_{h+1}^\tau)\big)\Big\|_{\Lambda_h^{-1}} \le \sqrt{c_1R^2\cdot(d\log(HdN_2/\delta))}$$

holds for all $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$ with probability at least $1-\delta$. Here c_1 is an absolute constant.

Proof. See proof in Appendix K.5. □

For term (i) on the RHS of (139), by Cauchy-Schwartz inequality, it holds with probability at least $1-\xi/2$ that

$$\begin{aligned}|\text{(i)}| &\le \|\phi(s,a,s')\|_{\Lambda_h^{-1}}\cdot\Big\|\sum_{\tau=1}^{N_2}\big(\int_{\mathcal{S}}\phi(s_h^\tau,a_h^\tau,s')\mathcal{P}_h(s'\,|\,s_h^\tau,a_h^\tau)\mathrm{d}s' - \phi(s_h^\tau,a_h^\tau,s_{h+1}^\tau)\big)\Big\|_{\Lambda_h^{-1}} \\ &\le c_1R\cdot\sqrt{d\log(dHN_2/\xi)}\cdot\|\phi(s,a,s')\|_{\Lambda_h^{-1}},\end{aligned} \tag{140}$$

where the last inequality follows from Lemma K.1.

For term (ii) in (139), setting $\lambda = 1$, we have

$$|\text{(ii)}| \le \|\phi(s,a,s')\|_{\Lambda_h^{-1}}\cdot\|\theta_h\|_{\Lambda_h^{-1}} \le \sqrt{d}\cdot\|\phi(s,a,s')\|_{\Lambda_h^{-1}}, \tag{141}$$

where the last inequality follows from the facts that $\|\Lambda_h^{-1}\|_2 \le 1$ and $\|\theta_h\|_2 \le \sqrt{d}$ for all $h\in[H]$. Plugging (140) and (141) into (139), it holds with probability at least $1-\xi/2$ that

$$\begin{aligned}|\mathcal{P}_h(s'\,|\,s,a) - \widetilde{\mathcal{P}}_h(s'\,|\,s,a)| &\le cR\sqrt{d\log(HdN_2/\xi)}\cdot\|\phi(s,a,s')\|_{\Lambda_h^{-1}} \\ &\le \kappa\cdot\|\phi(s,a,s')\|_{\Lambda_h^{-1}}\end{aligned}$$

for any $h\in[H]$ and $(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$. Here $\kappa = cR\sqrt{d\log(dN_2)}$ is the scaling parameter with an absolute constant $c>0$. We conclude the proof of Lemma E.1. □

K.2. Proof of Lemma I.2

Proof. We establish the lower and upper bounds for ι_h^{k,r^μ} as follows, respectively.

Lower Bound. First we prove by backward induction that $\widehat{V}_h^{k,r^\mu} \in [0, (H-h+1)\sqrt{d}]$ for any $h \in [H]$. The base case $h = H$ holds, since $\widehat{V}_{H+1}^{k,r^\mu} = 0$ and $r_h^\mu \in [0, \sqrt{d}]$. We assume that $\widehat{V}_{h+1}^{k,r^\mu} \in [0, H-h]$. For the case for h, recall that $\{\widehat{\mathcal{P}}_h(\,\cdot \,|\, s', a')\}_{h=1}^H$ is a set of probability measures on the state space $\mathcal{S}$ for $(s', a') \in \mathcal{S} \times \mathcal{A}$, which implies that $\widehat{\mathcal{P}}_h \widehat{V}_{h+1}^{k,r^\mu} \in [0, (H-h)\sqrt{d}\ ]$. Note that $\Gamma_h \geq 0$, hence we have that $\widehat{Q}_h^{k,r^\mu}(s,a) \in [0, (H-h+1)\sqrt{d}]$. Then it holds that $\widehat{V}_h^{k,r^\mu} \in [0, H-h]$, since $\widehat{V}_h^{k,r^\mu}(s) = \langle \widehat{Q}_h^{k,r^\mu}(s,\cdot), \pi_h^k(\cdot \,|\, s)\rangle_{\mathcal{A}}$ for any $s \in \mathcal{S}$. By induction, it holds that $\widehat{V}_h^{k,r^\mu} \in [0, (H-h+1)\sqrt{d}]$ for any $h \in [H]$.

For notational simplicity, we write $\bar{Q}_h^{k,r^\mu}(s,a) = r_h^\mu(s,a) + \widehat{\mathcal{P}}_h \widehat{V}_{h+1}^{k,r^\mu}(s,a) - \Gamma_h(s,a)$. From the policy evaluation stage in Algorithm 3, we have

$$\widehat{Q}_h^{k,r^\mu}(s,a) = \max\big\{\bar{Q}_h^{k,r^\mu}(s,a), 0\big\}. \tag{142}$$

Meanwhile, by the definition of the ξ-uncertainty qualifiers in Definition 3.1, we have $r_h^\mu + \mathcal{P}_h \widehat{V}_{h+1}^k \geq \bar{Q}_h^{k,r^\mu}$. Moreover, by the fact that $r_h^\mu \in [0, \sqrt{d}]$ and $\widehat{V}_{h+1}^{k,r^\mu} \in [0, (H-h)\sqrt{d}]$, we have $r_h^\mu + \mathcal{P}_h V_{h+1,\pi^k}^{k,r^\mu} \in [0, (H-h+1)\sqrt{d}]$. Thus, we derive that

$$\begin{aligned}
\widehat{Q}_h^k(s,a) &= \max\{\bar{Q}_h^{k,r^\mu}(s,a), 0\} \\
&\leq \max\{r_h^\mu(s,a) + \mathcal{P}_h \widehat{V}_{h+1}^{k,r^\mu}(s,a), 0\} \\
&= r_h^\mu(s,a) + \mathcal{P}_h \widehat{V}_{h+1}^{k,r^\mu}(s,a),
\end{aligned}$$

which implies that $\iota_h^{k,r^\mu} \geq 0$.

Upper Bound. Since we condition on the event $\mathcal{E}$ defined in Definition 3.1, we have

$$\begin{aligned}
\bar{Q}_h^{k,r^\mu}(s,a) &= r_h^\mu(s,a) + \widehat{\mathcal{P}}_h \widehat{V}_{h+1}^{k,r^\mu}(s,a) - \Gamma_h(s,a)) \\
&\leq r_h^\mu(s,a) + \mathcal{P}_h \widehat{V}_{h+1}^{k,r^\mu}(s,a) \leq H-h+1,
\end{aligned}$$

where the last inequality follows from the facts that $\widehat{V}_{h+1}^{k,r^\mu}(s,a) \leq (H-h)\sqrt{d}$ and $r_h^\mu(s,a) \leq \sqrt{d}$. By (142) we have that $\widehat{Q}_h^{k,r^\mu}(s,a) \geq \bar{Q}_h^{k,r^\mu}(s,a)$. Thus, we obtain that

$$\begin{aligned}
\iota_h^{k,r^\mu}(s,a) &= (r_h^\mu + \mathcal{P}_h \widehat{V}_h^k)(s,a) - \widehat{Q}_h^{k,r^\mu}(s,a) \\
&\leq r_h^\mu(s,a) + (\mathcal{P}_h - \widehat{\mathcal{P}}_h)\widehat{V}_h^{k,r^\mu}(s,a) + \Gamma_h(s,a) \\
&\leq 2\Gamma_h(s,a),
\end{aligned}$$

where the last inequality follows from the definition of $\mathcal{E}$. Then we complete the proof of Lemma I.2. □

K.3. Proof of Lemma I.3

Proof. Recall that $\widehat{L}(\pi,\mu) = \widetilde{J}(\pi^{\mathrm{E}},\mu) - \widehat{J}(\pi^k,\mu)$. By Assumption 2.1, we know that the function $\widetilde{J}(\pi^{\mathrm{E}},\mu) = \frac{1}{N_1}\sum_{\tau=1}^{N_1}\sum_{h=1}^H \psi(s_{h,\tau}^{\mathrm{E}}, a_{h,\tau}^{\mathrm{E}})^\top \mu_h$ is a linear combination of $\{\mu_h\}_{h=1}^H$ and concave. Therefore, to prove that $\widehat{L}(\pi,\mu)$ is concave, it suffices to prove that $\widehat{J}(\pi^k, r^\mu)$ is convex for any μ_h with $\mu = \{\mu_h\}_{h=1}^H \in S$.

Recall that $\widehat{J}(\pi^k, r^\mu) = \widehat{V}_1^{k,r^\mu}(x)$, where x is the fixed initial state and $\widehat{V}_1^{k,r^\mu}$ defined in (57) is solved by

$$\begin{aligned}
&\widehat{V}_{H+1}^{k,r^\mu}(\cdot) = 0 \\
&\widehat{Q}_h^{k,r^\mu}(\cdot,\cdot) = \max\big\{(r_h^\mu + \widehat{\mathcal{P}}_h \widehat{V}_{h+1}^{k,r^\mu} - \Gamma_h)(\cdot,\cdot), 0\big\} \\
&\widehat{V}_h^{k,r^\mu}(\cdot,\cdot) = \big\langle \widehat{Q}_h^{k,r^\mu}(\cdot,\cdot), \pi_h^k(\cdot \,|\, \cdot)\big\rangle_{\mathcal{A}}, \text{ for } h \in [H].
\end{aligned} \tag{143}$$

Our proof relies on the following three basic properties of convex functions:

(i) If $f(u)$ and $g(u)$ are both convex function for u, then $f(u)+g(u)$ is also convex.

(ii) If $f(u)$ and $g(u)$ are both convex function for u, then $\max\big(f(u),g(u)\big)=\big(|f(u)+g(u)|+|f(u)-g(u)|\big)/2$ is also convex.

(iii) If $f(u,s)$ is a convex function for u, then $\mathbb{E}_{s\sim p}f(u,s)$ is also convex function for u, where p is a distribution.

Now we are ready to prove that $\widehat{J}(\pi^k,\mu)$ is convex for $\{\mu_h\}_{h=1}^H$. For the base case where $h=1$, observing that $\widehat{J}(\pi^k,\mu)=\big\langle \widehat{Q}_1^{k,r^\mu}(s_1,\cdot),\pi_1^k(\cdot\,|\,s_1)\big\rangle_{\mathcal{A}}$, by property (ii) and (iii) and (143), it suffices to prove that $r_h^\mu+\widehat{\mathcal{P}}_1\widehat{V}_2^{k,r^\mu}-\Gamma_1$ is convex for μ_1. Note that $\{\mu_h\}_{h=1}^H$ are separate reward parameters and $r_h^\mu(\cdot,\cdot)$ is only determined by μ_h, it shows that $\widehat{\mathcal{P}}_1\widehat{V}_2^{k,r^\mu}-\Gamma_1$ is a constant regardless of μ_h, which implies that $\widehat{\mathcal{P}}_1\widehat{V}_2^{k,r^\mu}-\Gamma_1$ is convex for μ_1. Meanwhile, since $\mathcal{R}$ is linear to ψ as shown in (9), we know that $r_h^\mu=\psi^\top\mu_h$ is also convex for μ_h. By property (i), we know that $\widehat{J}(\pi^k,\mu)=\widehat{V}_1^{k,r^\mu}(s_1)$ is convex for μ_1.

For the case when $h=H'$, where $2\le H'\le H$, similar to the analysis in the case when $h=1$, we can prove that $\widehat{V}_{H'}^{k,r^\mu}(s_h)$ is convex for $\mu_{H'}$. Note that $\{\Gamma_h\}_{h=1}^H$ defined in (3.1) is independent of $\{\mu_h\}_{h=1}^H$, we know that $r_{H'-1}^\mu+\widehat{\mathcal{P}}_{H'-1}\widehat{V}_{H'}^{k,r^\mu}-\Gamma_{H'-1}$ is convex for $\mu_{H'}$. By property (ii) and (iii) and (143), we know that $\widehat{V}_{H'-1}^{k,r^\mu}$ is also convex for $\mu_{H'}$. By repeating the analysis, we know that $\widehat{J}(\pi^k,\mu)$ is convex for $\mu_{H'}$. Therefore, we conclude the proof of Lemma I.3. □

K.4. Proof of Lemma I.4

Proof. Since Lemma I.3 shows that $\widehat{L}(\pi^k,\mu)$ is concave for μ_h for any $h\in[H]$, by the property of concave function, we have

$$\widehat{L}(\pi^k,\mu)-\widehat{L}(\pi^k,\mu^k)\le\sum_{h=1}^H\big[\nabla_{\mu_h}\widehat{L}(\pi^k,\mu^k)^\top(\mu_h-\mu_h^k)\big]. \tag{144}$$

Recall that we apply projected gradient ascent method to update $\{\mu_h^k\}_{h=1}^H$ in PGAP (Line 3 of Algorithm 3) as

$$\mu_h^{k+1}=\mathrm{Proj}_S\big[\mu_h^k+\eta\nabla_{\mu_h}\widehat{L}(\pi^k,\mu^k)\big], \tag{145}$$

we obtain that

$$\big[\mu_h^{k+1}-\mu_h^k-\eta\nabla_{\mu_h}\widehat{L}(\pi^k,\mu^k)\big]^\top(\mu_h-\mu_h^{k+1})\ge 0. \tag{146}$$

Rearranging terms in (146), we have

$$\begin{aligned}\nabla_{\mu_h}\widehat{L}(\pi^k,\mu^k)^\top(\mu_h-\mu_h^{k+1})&\le-\frac{1}{2\eta}\big((\mu_h^{k+1}-\mu_h^k)^\top(\mu_h-\mu_h^{k+1})\big)\\&=\frac{1}{2\eta}\big(\|\mu_h^k-\mu_h\|_2^2-\|\mu_h^{k+1}-\mu_h\|_2^2-\|\mu_h^{k+1}-\mu_h^k\|_2^2\big).\end{aligned} \tag{147}$$

By adding the term $\nabla_{\mu_h}\widehat{L}(\pi^{k+1},\mu^k)^\top(\mu_h^{k+1}-\mu_h^k)$ on both sides of (147), we obtain that

$$\begin{aligned}\nabla_{\mu_h}\widehat{L}(\pi^k,\mu^k)^\top(\mu_h-\mu_h^k)=\frac{1}{2\eta}\big(\|\mu_h^k-\mu_h\|_2^2-\|\mu_h^{k+1}-\mu_h\|_2^2-\|\mu_h^{k+1}-\mu_h^k\|_2^2\big)\\+\nabla_{\mu_h}\widehat{L}(\pi^k,\mu^k)^\top(\mu_h^{k+1}-\mu_h^k).\end{aligned} \tag{148}$$

Note that η is positive and by applying Cauchy-Schwartz inequality on the second term of the right-hand side of (148), we derive that

$$\nabla_{\mu_h}\widehat{L}(\pi^k,\mu^k)^\top(\mu_h^{k+1}-\mu_h^k)\le\|\nabla_{\mu_h}\widehat{L}(\pi^k,\mu^k)\|_2\|\mu_h^{k+1}-\mu_h^k\|_2. \tag{149}$$

From the reward update process in (145), we observe that

$$\|\mu_h^{k+1}-\mu_h^k\|_2\le\|\nabla_{\mu_h}\widehat{L}(\pi^k,\mu^k)\|_2. \tag{150}$$

By plugging (148), (149), and (150) into (144), we have

$$\begin{aligned}\sum_{k=1}^{K}\big[\widehat{L}(\pi^k,\mu)-\widehat{L}(\pi^k,\mu^k)\big] \le \sum_{k=1}^{K}\sum_{h=1}^{H}\Big[&\frac{1}{2\eta}||\mu_h^{k+1}-\mu_h||_2^2+\frac{1}{2\eta}||\mu_h^{k+1}-\mu_h||_2^2\\ &-\frac{1}{2\eta}||\mu_h^{k+1}-\mu_h^k||_2^2+\eta\|\nabla_{\mu_h}\widehat{L}(\pi^k,\mu^k)\|_2^2\Big],\end{aligned}$$

which concludes the proof of Lemma I.4. □

K.5. Proof of Lemma K.1

Proof. Before we prove Lemma K.1, we introduce the following lemma to generalize the concentration of self-normalized vector-valued process in (Abbasi-Yadkori et al., 2011) to function-valued process.

Lemma K.2. *Let Ω be a probability space and $\{\eta_t\}_{t=1}^{\infty}$ be a function-valued stochastic process with a filtration $\{\mathcal{M}_t\}_{t=0}^{\infty}$, i.e. $\eta_t:\mathcal{S}\times\Omega\to\mathbb{R}$. We assume that $\eta_t\,|\,\mathcal{G}_{t-1}$ is zero-mean and σ-sub-Gaussian, that is,*

$$\mathbb{E}[\eta_t(s)\,|\,\mathcal{G}_{t-1}]=0,\qquad \log\Big(\mathbb{E}\big[\exp\big(\int_{\mathcal{S}}g(s)\eta_t(s)\mathrm{d}s\big)\,|\,\mathcal{G}_{t-1}\big]\Big)\le\frac{1}{2}\|g\|_\infty^2\cdot\sigma^2,$$

for any $s\in\mathcal{S}$ and function $g:\mathcal{S}\to\mathbb{R}$. Let $\{X_t\}_{t=0}^{\infty}$ be an vector-function-valued stochastic process where $X_t:\mathcal{S}\times\Omega\to\mathbb{R}^d$, $X_t\in\mathcal{M}_{t-1}$. We also assume that $\|\lambda^\top X_t\|_{\infty,\mathcal{S}}\le R\cdot\|\lambda^\top X_t\|_{2,\mathcal{S}}$ a.s. for all $\lambda\in\mathbb{R}^d$. Let $V\in\mathbb{R}^{d\times d}$ be a positive definite matrix and $\bar{V}_t=\sum_{\tau=1}^{t}\int X_\tau(s)X_\tau(s)^\top\mathrm{d}s$. We also define

$$S_t=\sum_{\tau=1}^{t}\int_{\mathcal{S}}X_\tau(s)\eta_\tau(s)\mathrm{d}s.$$

Then for any $\delta>0$ and $t>0$, it holds with probability at least $1-\delta$ that

$$\|S_t\|_{\bar{V}_t^{-1}}^2\le 2(\sigma R)^2\cdot\log\bigg(\frac{\det(\bar{V}_t)^{1/2}}{\delta\det(V)^{1/2}}\bigg).$$

Proof. See Appendix K.5.1 for a detailed proof. □

We consider the filtration $\{\mathcal{F}_{h,\tau}\}_{h\in[H],\tau\in[N_2]}$ defined in §2.2. For any function $f:\mathcal{S}\to\mathbb{R}$, by Holder inequality, it holds that

$$\Big|\int_{\mathcal{S}}f(s')\big(\mathcal{P}_h(s'\,|\,s_h^\tau,a_h^\tau)-\delta_{s_{h+1}^\tau}(s')\big)\mathrm{d}s'\Big|\le 2\|f\|_\infty. \tag{151}$$

By the property of Dirac function, we have

$$\mathbb{E}[\delta_{s_{h+1}^\tau}(s')\,|\,\mathcal{F}_{h,\tau}]=\mathcal{P}_h(s\,|\,s_h^\tau,a_h^\tau). \tag{152}$$

Combining (151) and (152), we verify the condition of Lemma K.2 as follows,

$$\log\bigg(\mathbb{E}\Big[\exp\Big(\int_{\mathcal{S}}f(s')\big(\mathcal{P}_h(s'\,|\,s_h^\tau,a_h^\tau)-\delta_{s_{h+1}^\tau}(s')\big)\mathrm{d}s'\Big)\Big|\,\mathcal{F}_{h,\tau}\Big]\bigg)\le 2\|f\|_\infty^2.$$

Note that $\phi(s_h^\tau,a_h^\tau,s')$ is $\mathcal{F}_{h,\tau}$-measurable and $\delta_{s_h^\tau+1}$ is $\mathcal{F}_{h+1,\tau}$-measurable, we apply Lemma K.2 with $X_\tau=\phi(s_h^\tau,a_h^\tau,\cdot)$ and $\eta_\tau=\mathcal{P}_h(\cdot\,|\,s_h^\tau,a_h^\tau)-\delta_{s_{h+1}^\tau}$, which implies that

$$\begin{aligned}&\Big\|\sum_{\tau=1}^{N_2}\Big[\int_{\mathcal{S}}\phi(s_h^\tau,a_h^\tau,s')\mathcal{P}_h(s'\,|\,s_h^\tau,a_h^\tau)\mathrm{d}s'-\phi(s_h^\tau,a_h^\tau,s_{h+1}^\tau)\Big]\Big\|_{\Lambda_h^{-1}}^2\\ &\quad\le 8R^2\cdot\log\big(H/\delta\cdot\det(\Lambda_h)^{1/2}\det(\lambda I)^{-1/2}\big),\end{aligned} \tag{153}$$

with probability at least $1-\delta/H$.

We now upper bound the term $\det(\Lambda_h)$. By the definition of Λ_h in (35), it holds for any $y \in \mathbb{R}^d$ that

$$y^\top \Lambda_h y = \lambda \|y\|_2^2 + \sum_{\tau=1}^{N_2} \int_{\mathcal{S}} |y^\top \phi(s_h^\tau, a_h^\tau, s')|^2 \mathrm{d}s' \le (\lambda + dN_2)\|y\|_2^2,$$

where the last inequality follows from Assumption 2.1. Hence we derive that $\|\Lambda_h\|_2 \le \lambda + dN_2$, which implies that

$$\det(\Lambda_h) \le \|\Lambda_h\|_2^d \le (\lambda + dN_2)^d. \tag{154}$$

Setting $\lambda = 1$, combining (153) and (154), it holds with probability at least $1 - p/H$ that

$$\begin{aligned} &\Big\| \sum_{\tau=1}^{N_2} \Big(\int_{\mathcal{S}} \phi(s_h^\tau, a_h^\tau, s') \mathcal{P}_h(s' \,|\, s_h^\tau, a_h^\tau) \mathrm{d}s' - \phi(s_h^\tau, a_h^\tau, s_{h+1}^\tau) \Big) \Big\|_{\Lambda_h^{-1}}^2 \\ &\quad \le 8R^2 \cdot \big(1/2 \cdot d \log(1 + dN_2) + \log(H/p) \big) \le c_1 R^2 \cdot \big(d \log(HdN_2/p) \big). \end{aligned} \tag{155}$$

Here $c_1 > 0$ is an absolute constant. By the union bound for $h \in [H]$, we know that (155) holds for all $h \in [H]$ with probability at least $1 - p$. Thus, we complete the proof of Lemma K.1. □

K.5.1. Proof of Lemma K.2

Proof. The proof is a generalization of that in (Abbasi-Yadkori et al., 2011). For notational simplicity, we denote by $\langle f, g \rangle = \int_{\mathcal{S}} f(s) g(s) \mathrm{d}s$ the inner product of any functions f and g. We use the same definitions and notations as Lemma K.1. First, we introduce the following lemmas.

Lemma K.3. *Let $\beta \in \mathbb{R}^d$ be a vector and*

$$M_t^\beta = \exp\bigg\{ \sum_{\tau=1}^t \bigg(\frac{\langle \beta^\top X_\tau, \eta_\tau \rangle}{\sigma R} - \frac{\langle \beta^\top X_\tau, \beta^\top X_\tau \rangle}{2} \bigg) \bigg\}.$$

Let T be a stopping time with respect to the filtration $\{\mathcal{M}_t\}_{t=1}^\infty$. Then M_T^β is almost surely well-defined and $\mathbb{E}[M_T^\beta] \le 1$.

Proof. We first show that $\{M_t^\beta\}_{t=0}^\infty$ is a supermartingale. Let

$$G_\tau^\beta = \exp\bigg(\frac{\langle \beta^\top X_\tau, \eta_\tau \rangle}{\sigma R} - \frac{\|\beta^\top X_\tau\|_{2,\mathcal{S}}^2}{2} \bigg).$$

By the conditional sub-Gaussian property of η_τ and the fact that $\|\beta^\top X_t\|_{\infty,\mathcal{S}} \le R \cdot \|\beta^\top X_t\|_{2,\mathcal{S}}$, we have

$$\mathbb{E}[G_\tau^\beta \,|\, \mathcal{M}_{t-1}] \le \exp\bigg(\frac{\|\beta^\top X_\tau\|_{\infty,\mathcal{S}}^2}{2R} - \frac{\|\beta^\top X_\tau\|_{\infty,\mathcal{S}}^2}{2R} \bigg) = 1.$$

Thus, we have $\mathbb{E}[M_t^\beta \,|\, \mathcal{M}_{t-1}] = M_{t-1}^\beta \cdot \mathbb{E}[G_\tau^\beta \,|\, \mathcal{M}_{t-1}] \le M_{t-1}^\beta$, which implies that $\{M_t^\beta\}_{t=0}^\infty$ is a supermartingale and $\mathbb{E}[M_t^\beta] \le 1$. We then show that M_T^β is well-defined, where T is a stopping time. By the convergence theorem of nonnegative supermartingales, it holds that $M_\infty^\beta = \lim_{t\to\infty} M_t^\beta$. Thus, M_T^β is well-defined whether $T < \infty$ or not. Finally, to show that $\mathbb{E}[M_T^\beta] \le 1$, we apply Fatou's lemma and obtain that

$$\mathbb{E}[M_\tau^\beta] = \mathbb{E}[\lim_{t\to\infty} M_{T\wedge t}^\beta] \le \liminf_{t\to\infty} \mathbb{E}[M_{T\wedge t}^\beta] \le 1.$$

Thus, we conclude the proof of Lemma K.3. □

Lemma K.4. *Let T be a stopping time with respect to $\{\mathcal{M}_t\}_{t=0}^\infty$, then it holds with probability at least $1 - \delta$ that*

$$\|S_T\|_{\bar{V}_T^{-1}}^2 > 2(\sigma R)^2 \cdot \log\bigg(\frac{\det(\bar{V}_T)^{1/2}}{\delta \det(V)^{1/2}} \bigg).$$

Proof. Without loss of generality, we assume that $\sigma \cdot R = 1$. We define

$$V_t = \sum_{\tau=1}^{t} \int X_\tau(s) X_\tau(s)^\top.$$

Then, we have

$$M_t^\beta = \exp(\beta^\top S_t - \|\beta\|_{V_t}^2/2).$$

By Lemma K.3, we have that $\mathbb{E}[M_t^\beta] \leq 1$. Let Λ be an $\mathbb{R}^d$-valued Gaussian random variable with covariance matrix V^{-1}. Moreover, we assume that Λ is independent of $\{\mathcal{M}_t\}_{t=0}^\infty$. Let $M_t = \mathbb{E}[M_t^\Lambda \mid \mathcal{M}_\infty]$, where $\mathcal{M}_\infty = \sigma(\cup_{\tau=0}^\infty \mathcal{M}_\tau)$. Notice that $\mathbb{E}[M_T] = \mathbb{E}[\mathbb{E}[M_T^\Lambda \mid \Lambda]] \leq 1$. We denote by p the density of Λ and by $v(A) = \int \exp(-x^\top A x)\mathrm{d}x = \sqrt{(2\pi)^d / \det(A)}$ for positive definite matrix $A \in \mathbb{R}^{d\times d}$. Then we obtain that

$$\begin{aligned} M_t &= \int \exp(\beta^\top S_t - \|\beta\|_{V_t}^2/2) p(\beta) \mathrm{d}\beta \\ &= \int \exp(-\|\beta - V_t^{-1} S_t\|_{V_t}^2/2 + \|S_t\|_{V_t^{-1}}^2) p(\beta) \mathrm{d}\beta \\ &= v(V)^{-1} \cdot \exp(\|S_t\|_{V_t^{-1}}^2/2) \cdot \int \exp(-\|\beta - V_t^{-1} S_t\|_{V_t}^2/2 - \|\beta\|_V^2/2) \mathrm{d}\beta. \end{aligned} \tag{156}$$

Note that

$$\begin{aligned} \|\beta - V_t^{-1} S_t\|_{V_t}^2 + \|\beta\|_V^2/2 &= \|\beta - \bar{V}_t^{-1} S_t\|_{V_t}^2 + \|V_t^{-1} S_t\|_{V_t}^2 - \|S_t\|_{V_t}^2 \\ &= \|\beta - \bar{V}_t^{-1} S_t\|_{V_t}^2 + \|S_t\|_{V_t^{-1}}^2 - \|S_t\|_{V_t}^2. \end{aligned} \tag{157}$$

Plugging (157) into (156), we have that

$$\begin{aligned} M_t &= v(V)^{-1} \cdot \exp(\|S_t\|_{\bar{V}_t^{-1}}^2/2) \cdot \int \exp(-\|\beta - \bar{V}_t^{-1} S_t\|_{\bar{V}_t}^2/2) \mathrm{d}\beta \\ &= \frac{v(\bar{V}_t)}{v(V_t)} \cdot \exp(\|S_t\|_{\bar{V}_t^{-1}}^2/2) \\ &= \sqrt{\det(V)/\det(\bar{V}_t)} \cdot \exp(\|S_t\|_{\bar{V}_t^{-1}}^2/2). \end{aligned}$$

Thus, we have

$$\mathbb{P}\bigg\{\|S_T\|_{\bar{V}_T^{-1}}^2 > 2\log\bigg(\frac{\det(\bar{V}_T)^{1/2}}{\delta \det(V)^{1/2}}\bigg)\bigg\} = \mathbb{P}(\delta M_T > 1) \leq \mathbb{E}[\delta M_T] \leq \delta,$$

which completes the proof of Lemma K.4. □

We now prove Lemma K.2 as follows. Define

$$T = \inf\bigg\{t \geq 0 \colon 2\log\bigg(\frac{\det(\bar{V}_t)^{1/2}}{\delta \det(V)^{1/2}}\bigg) < \|S_t\|_{\bar{V}_t^{-1}}^2\bigg\}$$

for a fixed $\delta > 0$. Then it holds that

$$\begin{aligned} \mathbb{P}\bigg\{\exists t \geq 0, \|S_t\|_{\bar{V}_t^{-1}}^2 > 2\log\bigg(\frac{\det(\bar{V}_t)^{1/2}}{\delta \det(V)^{1/2}}\bigg)\bigg\} &= \mathbb{P}(T < \infty) \\ &= \mathbb{P}\bigg\{\|S_T\|_{\bar{V}_T^{-1}}^2 > 2\log\bigg(\frac{\det(\bar{V}_T)^{1/2}}{\delta \det(V)^{1/2}}\bigg), T < \infty\bigg\} \\ &\leq \mathbb{P}\bigg\{\|S_T\|_{\bar{V}_T^{-1}}^2 > 2\log\bigg(\frac{\det(\bar{V}_T)^{1/2}}{\delta \det(V)^{1/2}}\bigg)\bigg\} \leq \delta, \end{aligned}$$

which completes the proof of Lemma K.2. □