Learning Efficiently Function Approximation for Contextual MDP

Orin Levy orinlevy@mail.tau.ac.il
Tel Aviv University
Israel

Yishay Mansour mansour.yishay@gmail.com
Tel Aviv University and Google Research
Israel

Abstract

We study learning contextual MDPs using a function approximation for both the rewards and the dynamics. We consider both the case that the dynamics dependent or independent of the context. For both models we derive polynomial sample and time complexity (assuming an efficient ERM oracle). Our methodology gives a general reduction from learning contextual MDP to supervised learning.

Keywords: Reinforcement Leaning, Sample Complexity, Contextual MDP, Function Approximation

1. Introduction

Markov decision processes (MDPs) are commonly used to describe dynamic environments. MDPs characterize many real-life tasks in a variety of applications including: advertising, healthcare, games, robotics and more. In those applications, at each episode an agent arrives and interacts with the environment with the goal of maximizing her return. (See, e.g., Sutton and Barto (2018).)

In many applications, in each episode, there are additional exogenous factors that affect the environment, which we refer to as the *context*. One can extend the state space to include the context, but this has the disadvantage of greatly increasing the state space, and hence the complexity of learning and even the representation of a policy. An alternative approach is to keep a small state space, and regard the context as an additional side-information. Contextual Markov Decision Process (CMDP) describes such a model, where for each context there is a potential different optimal policy.

CMDPs are useful to model many user-driven applications, where the context is a user-related information which influences the optimal decision making. One natural application is in healthcare. We can model the interaction with a given patient using an MDP. For a given medical treatment, the expected outcome of a patient is highly dependent on his medical history and other personal parameters, which we model as her context. For example, the success probability of a given treatment might heavily depend on the patient's age and weight.

We abstract the patient's medical history and any other relevant information as the context. The benefit of using a CMDP is the fact that most patients behave similarly, although the context space may be large, and there might be unforeseen connection between the context and the outcomes. CMDPs allow to share information and behavior between different contexts in a natural way.

Our contributions. We present efficient learning algorithms for CMDP, given an access to an ERM oracle. We consider a finite horizon CMDP, where the rewards are an arbitrary function of the context and the state-action. The dynamics may be either *context-free*, where the context does not influence the dynamics, or *context-dependent*, where different contexts induce different dynamics. Clearly, the most challenging model is the unknown context-dependent dynamics. Our method induces an efficient reduction from learning contextual MDP (a Reinforcement Learning problem) to supervised learning.

The learning process outputs an explicit function approximation of the rewards and the dynamics. Following the learning phase, our learner receives the current context, builds a related MDP for that context, computes an optimal policy for that MDP, and later runs that policy. Both the construction of the MDP and computing the optimal policy are done in polynomial time in the number of states, actions, and horizon.

For *context free* dynamics, we give an efficient algorithm that creates an unbiased sample of the context-reward and context-next state pairs, for each significantly-reachable state and action. We use the unbiased sample to approximate both the context-free dynamics and the state-action rewards.

The most challenging case is unknown *context dependent* dynamics. Here, we are unable to define an unbiased sample at the state-action level, since we do not know the probability of a state-action pair for given a context and policy. However, we give an efficient algorithm that constructs for each layer two unbiased data sets. Both function approximations are done once per an entire layer of the MDP.

Table 1 contains a summary of our sample complexity results, up to poly-logarithmic factors. In all cases we have a polynomial dependence in all our parameters. Specifically, d, the function approximation class pseudo-dimension or fat-shattering dimension (see Appendix A for more information regarding the dimensions), the number of states $|S|$, actions $|A|$, horizon H, inverse accuracy $1/\epsilon$ and poly-logarithmic dependence in the inverse confidence parameter $\log(1/\delta)$. The ERM oracle complexity in all cases is $|S||A|$ except for unknown context dependent case where it is only H. We remark that our definition bounds by ϵ^2 the squared error rather than ϵ. This essentially introduces additional ϵ^2 factors which do not exist using the standard definition. We use our definition mainly for convenience.

Table 1: summary of our results.

Dynamics	Absolute Loss	Square Loss
Known, context-free	$d\epsilon^{-2}H^2\|S\|^4\|A\|^3\log(1/\delta)$	$d\epsilon^{-4}H^4\|S\|^6\|A\|^5\log(1/\delta)$
Unknown, context-free	$d\epsilon^{-3}H^5\|S\|^5\|A\|^3\log(1/\delta)$	$d\epsilon^{-4}H^4\|S\|^6\|A\|^5\log(1/\delta)$
Known, context-dependent	$d\epsilon^{-6}H^5\|S\|^5\|A\|^3\log(1/\delta)$	$d\epsilon^{-8}H^7\|S\|^5\|A\|^3\log(1/\delta)$
Unknown, context-dependent	$d\epsilon^{-6}H^9\|S\|^{11}\|A\|^2\log(1/\delta)$	$d\epsilon^{-8}H^{13}\|S\|^{15}\|A\|^2\log(1/\delta)$

2. Related Work

Contextual Reinforcement Leaning. CMDP was introduce by Hallak et al. (2015). Modi et al. (2018) gives a general framework for deriving generalization bounds as a function of the covering number for smooth CMDPs and contextual linear combination of MDPs. For smooth CMDPs they obtain sample complexity upper bound of $\tilde{O}\left(NH^2|S|A|\epsilon^{-3}(|S|+\ln\frac{N|S||A|}{\delta}\ln\frac{N}{\delta})\right)$, and a lower bound of $\Omega\left(\frac{N|S||A|}{\epsilon^2}\right)$ where N is the covering number of the context space, which can be exponential in the dimension of it. For the contextual linear combination of MDPs, they obtain a sample complexity bound of $O\left(\epsilon^{-2}m^2H^4|S||A|\log\frac{1}{\delta}\max\{m^2,|S|^2\log^2(m|S||A|/\delta)\}\right)$ where m is the number of combined MDPs. In contrast, our bounds depend on the complexity dimension (VC,Pseudo etc.) which can be logarithmic in the covering number of the context space (see Subsection 27.2 in Shalev-Shwartz and Ben-David (2014)) or independent of it. For example, the γ-fat shattering dimension of of linear functions is $1/\gamma^2$. However, our results do not contradicts the above lower bound, as the pseudo and fat-shattering dimensions are known to be tightly upper bounded by the covering number of the function class input space (i.e., the domain). For smooth CMDP, the function classes used to approximate the rewards and dynamics are L_r and L_p-Lipschitz (respectively) and it is known that the γ-fat shattering dimension of L-Lipshcitz function class is (approximatly) linear in the covering number of the domain. Our work generalize the work of Modi et al. (2018) since we have no assumption regarding the CMDP or the function classes.

Modi and Tewari (2020) give a regret analysis for Generalized Linear Models (GLMs). Our function approximation framework is much more general than GLM.

Foster et al. (2021) present a statistical complexity measure for interactive decision making and present an application of it to contextual RL. They assume an access to an online estimation oracle with regret guarantees. Using it, they obtain $\tilde{O}(\sqrt{T})$ regret. However, this oracle is very strong and might be computationally inefficient. It is also unclear whether their algorithmic approach can be extended to offline oracles for estimation. In contrast, we use a standard ERM oracle.

Jiang et al. (2017) present OLIVE which is sample efficient for Contextual Decision Processes with a small Bellman rank. We do not make any assumptions on the Bellman rank.

Reward-Free exploration. The setting of unknown and context-free dynamics is closely related to Reward-free RL Jin et al. (2020); Zhang et al. (2021); Ménard et al. (2021); Chen et al. (2022); Qiu et al. (2021). Our main motivation for developing the context-free algorithms is to extend them later to the context-dependent case.

Contextual Bandits. Contextual bandits (CMAB) are a natural extension of the Multi-Arm Bandit (MAB), augmented by a context which influences the rewards Slivkins (2019); Lattimore and Szepesvári (2020). Agarwal et al. (2014) use efficiently an optimization oracle to derive an optimal regret bound. Regression based approaches appear in Agarwal et al. (2012); Foster et al. (2018); Foster and Rakhlin (2020); Simchi-Levi and Xu (2021); Xu and Zeevi (2020). We differ from CMAB, since our main challenge is the dynamics, and the need to optimize future rewards, which is the case in most RL settings.

3. Preliminaries and Notations

Markov Decision Process (MDP). We consider an episodic MDP with a finite horizon H, and assume w.l.o.g it is layered, loop free and has a unique start state. A *Markov Decision Process (MDP)* is a tuple (S, A, P, r, s_0, H), where (1) S is a finite state space decomposed into $H+1$ disjoint subsets (layers) $S_0, S_1, \ldots, S_H$ such that transitions are only possible between consecutive layers (i.e., loop-free), (2) A is a finite action space, (3) $s_0 \in S$ is the unique start state, (4) $P(\cdot|s,a)$ defines the transition probability function, i.e., $P(s'|s,a)$ is the probability that we reach state s' given that we are in state s and perform action a, (5) $R(s,a) \in [0,1]$ is a random variable for the reward of performing action a in state s, and $r(s,a)$ is its expectation, i.e., $r(s,a) = \mathbb{E}[R(s,a)]$, and (6) H is the finite horizon.

Policy. A *stochastic policy* π is a mapping from states to distribution over actions, i.e., $\pi : S \to \Delta(A)$. A *deterministic policy* π is a mapping from states to actions, i.e., $\pi : S \to A$.

Occupancy measure. Let $q_h(s|\pi, P)$ denote the probability of reaching state $s \in S_h$ at time $h \in [H]$ of an episode generated using policy π and dynamics P.

Episode and trajectory. At the start of each episode we select a policy π, run it, and observe a trajectory $\tau = (s_0, a_0, r_0, s_1, \ldots, s_{H-1}, a_{H-1}, r_{H-1}, s_H)$, where for all $h \in [H-1]$, $a_h \sim \pi(\cdot|s_h)$, $r_h \sim R(s_h, a_h)$ and $s_{h+1} \sim P(\cdot|s_h, a_h)$ [1].

Value function. Given a policy π and a MDP $M = (S, A, P, r, s_0, H)$, the $h \in [H-1]$ stage value function of a state $s \in S_h$ is defined as $V^{\pi}_{M,h}(s) = \mathbb{E}_{\pi,M}\left[\sum_{k=h}^{H-1} r(s_k, \pi(s_k))|s_h = s\right]$.
For brevity, when $h = 0$ we denote $V^{\pi}_{M,0}(s_0) := V^{\pi}_{M}(s_0)$.

Optimal policy and Bellman equations. A (deterministic) optimal policy $\pi^{\star}_M$ for MDP M satisfies, for every stage $h \in [H-1]$ and a state $s \in S_h$, $\pi^{\star}_{M,h}(s) \in \arg\max_{\pi:S\to A}\{V^{\pi}_{M,h}(s)\}$.

Planning. Given an MDP $M = (S, A, P, r, s_0, H)$ the procedure `Planning`(M) returns an optimal policy $\pi^{\star}_M$ and its value $V^{\star}_M(s_0)$ and runs in time $O(|S|^2\,|A|\,H)$.

Contextual MDP (CMDP) is a tuple $(\mathcal{C}, S, A, \mathcal{M})$ where $\mathcal{C} \subseteq \mathbb{R}^{d'}$ is the context space, S is the state space and A is the action space. The mapping $\mathcal{M}$ maps a context $c \in \mathcal{C}$ to a MDP $\mathcal{M}(c) = (S, A, P^c, r^c, s_0, H)$. There is an unknown distribution $\mathcal{D}$ over the context space $\mathcal{C}$, and for each episode a context c is sampled i.i.d. from $\mathcal{D}$. For mathematical convenience, we assume the context space is finite (but potentially huge). Our results naturally extend to infinite contexts space.

Context-free dynamics vs. context-dependent dynamics. A CMDP has *context-free* dynamics when the context effects only the rewards function, while the dynamics are identical for all contexts. i.e., there exits a dynamics P such that for all $c \in \mathcal{C}$, $P^c = P$. A *context-dependent* dynamics has a potentially different dynamics P^c for each context c. We consider both settings.

Context-dependent policy. A context-dependent policy $\pi = (\pi_c : S \to \Delta(A))_{c\in\mathcal{C}}$ maps a context $c \in \mathcal{C}$ to a policy $\pi_c : S \to \Delta(A)$. We similarly define a deterministic context-dependent policy.

Optimal context-dependent policy is a policy $\pi^{\star} = (\pi^{\star}_c)_{c\in\mathcal{C}}$ that satisfies, for every context $c \in \mathcal{C}$, $\pi^{\star}_c \in \arg\max_{\pi:S\to\Delta(A)} V^{\pi}_{\mathcal{M}(c)}(s_0)$.

1. W.l.o.g. we assume that $r(s_H, a_H) = 0$ for any $s_H \in S_H$ and $a_H \in A$ so we can omit it.

Losses. The square loss is $\ell_2(z,y) = (z-y)^2$ and absolute loss is $\ell_1(z,y) = |z-y|$.

Function approximation using functions class $\mathcal{F}$. The *squared error* (*absolute error*, respectively) of a function $f \in \mathcal{F}$ is $sqerr(f) = \mathbb{E}_x[\ell_2(f(x), f^\star(x))]$ ($aberr(f) = \mathbb{E}_x[\ell_1(f(x), f^\star(x))]$), where $f^*(x)$ is the target function (and we might have $f^* \notin \mathcal{F}$). The *squared approximation error* (*absolute approximation error*) of $\mathcal{F}$ is $\alpha_2^2(\mathcal{F}) = \inf_{f\in\mathcal{F}} sqerr(f)$ ($\alpha_1(\mathcal{F}) = \inf_{f\in\mathcal{F}} aberr(f)$). Note that for square loss we square the approximation error, while this is not standard, it is mainly for mathematical convenience. When clear from the context, we use α instead of α_1 or α_2.

ERM oracle. Let $\mathcal{X}$ be some domain, and let $\mathcal{F}$ be a function class that maps $\mathcal{X}$ to $[0,1]$. An Empirical Risk Minimization (ERM) oracle for $\mathcal{F}$ with respect to a loss function ℓ takes as input a data set $D = \{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathcal{X}$, $y_i \in [0,1]$ and computes $\widehat{f} = \arg\min_{f\in\mathcal{F}} \sum_{(x,y)\in D} \ell(f(x), y)$.

Function class complexity measures. Our sample complexity bounds are stated in the terms of the pseudo and fat-shattering dimension of the function class (see Anthony et al. (1999)), which are complexity measures for learning real-valued function classes. It is known that if the pseudo/fat-shattering dimension of the function class $\mathcal{F}$ is finite, then $\mathcal{F}$ has a uniform convergence property. Hence $\mathcal{F}$ is learnable using an ERM algorithm up to an ϵ error, with probability at least $1-\delta$. $m(\epsilon,\delta)$ is the required sample complexity. For more information regarding the dimensions definitions and the sample complexity requires for learning, please see Appendix A.

Reward function approximation. For every state $s \in S$ and action $a \in A$ we have a function class $\mathcal{F}_{s,a}^R = \{f_{s,a} : \mathcal{C} \to [0,1]\}$, which maps context c to (approximate) reward. The function $N_R(\mathcal{F}, \epsilon, \delta)$ maps a function class $\mathcal{F}$, required accuracy $\epsilon \in (0,1)$ and confidence $\delta \in (0,1)$ to the number of required samples for the ERM oracle to guarantee, with probability $1-\delta$, that $\mathbb{E}[\ell(\widehat{f}(x), f^\star(x))] \leq \epsilon + \alpha$, where α is the approximation error.

Dynamics function approximation. For the unknown context-free case we simply use a tabular approximation (see Section 4). For the unknown context-dependent case, we use a function approximation per layer, as we define in Section 5.

Reachability. The reachability of a state is the maximum probability of reaching it, by any policy. A state $s_h \in S_h$ is *β-reachable for dynamics P* if there exists a policy π such that $q_h(s_h|\pi, P) \geq \beta$. For a dynamics P and $s_h \in S_h$ let π_{s_h} denote the policy with the highest probability to visit s_h. Hence, a state s_h is β-reachable for dynamics P iff π_{s_h} satisfies that $q_h(s_h|\pi_{s_h}, P) \geq \beta$.

Learning objective. A mapping $\widehat{\pi}^\star$ from contexts c to a policy $\widehat{\pi}_c^\star$ is ϵ-optimal if $\mathbb{E}_{c\sim\mathcal{D}}[V_{\mathcal{M}(c)}^{\pi_c^\star}(s_0) - V_{\mathcal{M}(c)}^{\widehat{\pi}_c^\star}(s_0)] \leq \epsilon + O(\alpha H)$, where α is an agnostic approximation error. The goal of the learning algorithm is to compute a mapping from contexts to policies which are ϵ-optimal. The learning algorithm is sample efficient if it uses $T = poly(|S|, |A|, H, \frac{1}{\epsilon}, \log\frac{1}{\delta})$ samples, and is computationally efficient if it is sample efficient, its running time is $poly(T, |S|, |A|, H)$, and the number of oracle queries is $poly(|S|, |A|, |H|)$.

Mathematical notations. We denote expectation by $\mathbb{E}[\cdot]$ and probabilities by $\mathbb{P}[\cdot]$. The indicator function is $\mathbb{I}[G]$ returns 1 if event G holds and 0 otherwise.

4. Context-Free Dynamics

This section addresses the case of an unknown dynamics which do not depend on the context (i.e., context-free dynamics). The main goal of this section is to provide intuition for our approach, and develop algorithmic tools that we will later use to solve the unknown context-dependent dynamics case, which is the main contribution of the paper.

Our approach. Our goal is to collect "sufficient" i.i.d examples (c, r) of contexts and rewards for each state-action pair (s, a), to learn the context-dependent rewards function using ERM. This goal is not trivial even without the context, due to inner dependencies in a trajectory $\tau = (c, s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_H)$ generated by a policy π and the dynamics P. For simplicity, we sample for each state-action pair independently. To collect i.i.d samples efficiently for each state, we need to compute an exploration policy which (approximately) maximizes the probability to visit the target state. However, special care needs to be taken for states which are "hard" to reach, i.e., states which are not β-reachable, for some parameter β. Conceptually, we transition states which are not β-reachable to a sink state and avoid the need to

approximate their dynamics. Rather than using a fixed parameter β, we (later) introduce a more gradual transition which improves our dependency on ϵ in the resulted sample complexity bound.

Algorithm overview. We approximate the true dynamics P by $\widehat{P}$, one layer at a time. after we learned the first $h-1$ layers, we use $\widehat{P}$ to compute a policy $\widehat{\pi}_s$ to reach every state s in layer h (i.e., $s \in S_h$), and the probability of reaching it. Intuitively, if $\widehat{P} \approx P$ for the first $h-1$ layers, then $\widehat{\pi}_s$ will approximately maximizing the probability of reaching s, and would allow to sample it efficiently. Once we reach state s we use the various actions in a round-robin manner.

Algorithm EXPLORE-UCFD (Algorithms 1 and 7) works in phases, where in phase h we approximate the dynamics and rewards of layer h. We define the approximated dynamics $\widehat{P}$ as the empirical dynamics, when in addition we transition not β-reachable states to the sink state s_{sink} (a new state which we add to the approximated model).

We first collect samples for each (significantly reachable) state in layer h and then use them to approximate the dynamics, using simple tabular estimation. Using the same sample we also estimate the rewards using ERM oracle. The required accuracy for each state-action pair (s_h, a_h) is determined by the accuracy-per-state function $\epsilon_\star(\cdot)$ which gets $\widehat{p}_{s_h} := q_h(s_h|\widehat{\pi}_{s_h}, \widehat{P})$ as an input. After collecting sufficient number of samples for every (significantly reachable) state in layers up to $h-1$, we have a good approximation of the dynamics up to layer $h-1$. This yields a good approximation of the occupancy measure of layer h for any policy π.

Given a state $s_h \in S_h$ and the approximated dynamics $\widehat{P}$ we compute $\widehat{\pi}_{s_h} = \arg\max_\pi q_h(s_h|\pi, \widehat{P})$ using a planning algorithm. We run $\widehat{\pi}_{s_h}$ to generate the sample of s_h. In order to control the number of sampled episodes we define significantly reachable states as β-reachable for $\widehat{P}$, i.e., they have $q_h(s_h|\widehat{\pi}_{s_h}, \widehat{P}) \geq \beta$.

At the end of the sampling we have for each β-reachable for $\widehat{P}$ state $s_h \in S_h$ for $\widehat{P}$, and every action a_h a data set which contains tuples (s_h, a_h, s_{h+1}). Then, we use a tabular approximation to learn the context-free dynamics. For the rewards, we use the collected examples of tuples $((c, s_h, a_h), r_h)$ and run the ERM on that data set to compute a function approximation for the rewards [2]. Note that it is important that we first fix the approximation of layers up to h, which guarantee that we use the same $\widehat{\pi}_s$ and $\widehat{p}_s$ in each sampling state $s \in S_h$.

The approximated dynamics $\widehat{P}$. Let $n(s'|s,a)$ denote the number of times the triplet (s,a,s') was observed and $n(s,a)$ denote the number of times the pair (s,a) was observed. We have a threshold $N_P(\gamma, \delta_1) = O(\gamma^{-2}(|S| + \log(1/\delta)))$. If s is not β-reachable (given our learned dynamics) or (s,a) is sampled less than $N_P(\gamma, \delta_1)$ times it transition to the sink state s_{sink}. When (s,a) is sampled at least $N_P(\gamma, \delta_1)$ times, we use the empirical next state distribution, i.e., $\widehat{P}(s'|s,a) = \frac{n(s'|s,a)}{n(s,a)}$, to approximate the transition probability distribution of (s,a).

Accuracy per state function. We set a refined desired accuracy per state function $\epsilon_\star$, and saves a $1/\epsilon$ factors in the sample complexity. We do not approximate states which are very hard to reach. States which are very easy to reach, we want maximum accuracy. For intermediate levels we have a gradual accuracy dependency. This is captured in our definition of the accuracy-per-state function $\epsilon_\star$, which depends the probability to visit state s, i.e., $\widehat{P}_s := q_h(s_h|\widehat{\pi}_s, \widehat{P})$. We define it as follows: where $B > 0$ is a constant we determine later.

$$\epsilon_\star(\widehat{P}_s) = \begin{cases} 1 & , \text{ if } \widehat{P}_s < \frac{\epsilon}{B|S|} \\ \frac{\epsilon}{B\widehat{P}_s|S||A|} & , \text{ if } p_s \in [\frac{\epsilon}{B|S|}, \frac{1}{|S|}] \\ \frac{\epsilon}{BH|S||A|} & , \text{ if } \widehat{P}_s > \frac{\epsilon}{B|S|} \end{cases}$$

For the ℓ_2 loss we use $\epsilon^2_\star(\widehat{P}_s)$. We also denote $m_{s,a}(\widehat{P}_s) = \max\{N_R(\mathcal{F}^R_{s,a}, \epsilon_\star(\widehat{p}_s), \delta_1), N_P(\gamma, \delta_1)\}$.

Approximate optimal policy. For a context c, let the true MDP be $\mathcal{M}(c) = (S, A, P, r^c, s_0, H)$ and the approximated MDP be $\widehat{\mathcal{M}}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}, \widehat{r}^c, s_0, H)$. Let $\pi^\star_c$ and $\widehat{\pi}^\star_c$ be an optimal policy for $\mathcal{M}(c)$ and $\widehat{\mathcal{M}}(c)$ respectively. For both ℓ_1 and ℓ_2 loss function, we obtain the following.

Theorem 1 *With probability $1-\delta$ it holds that $\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha H$, after collecting $\tilde{O}\Big(d\epsilon^{-3}H^5|S|^5|A|^3 \log\frac{H|S||A|}{\delta}\Big)$ trajectories for ℓ_1 loss, and $\tilde{O}\Big(d\epsilon^{-4}H^4|S|^6|A|^5 \log\frac{H|S||A|}{\delta}\Big)$ for the ℓ_2 loss. α and d are the maximal approximation error and fat-shattering / pseudo dimension over all states and actions, respectively.*

2. When collecting samples for state s and action a, we update their sample only, to guarantee it is i.i.d.

Algorithm 1 EXPLORE Unknown Context Free Dynamics (sketch for the ℓ_1 loss)

initialize counters $n(s,a)=0, n(s'|s,a)=0$ for all $(s,a,s') \in S \times A \times S$.
for $h \in [H-1]$ **do**
  compute the approximated dynamics $\widehat{P}$ up to layer $h-1$
  for $s \in S_h$ **do**
    compute $\widehat{p}_s$, the highest probability to visit s in $\widehat{P}$, and a policy $\widehat{\pi}_s$ that reaches it
    if $\widehat{p}_{s_h} \geq \beta$ **then**
      for $a \in A$ **do**
        initialize $Sample(s,a) = \emptyset$
        set $\widehat{\pi}_s(s) \leftarrow a$
        for $t = 1, 2, \ldots, \lceil \frac{2}{\widehat{p}_s - \gamma h}(\ln(\frac{1}{\delta_1}) + m_{s,a}(\widehat{P}_s)\ ) \rceil$ **do**
          observe context c, run $\widehat{\pi}_s$
          observe trajectory, update sample and counters
        if $|Sample(s,a)| \geq m_{s,a}(\widehat{P}_s)$ **then**
          $f_{s,a} = \texttt{ERM}(\mathcal{F}^R_{s,a}, Sample(s,a), \ell_1)$
        else
          return FAIL
    else
      set for all $a \in A:\ f_{s,a} = 0$
return $F = \{f_{s,a}:\ \forall s \in S, a \in A\}, \widehat{P}$

Analysis outline. For every layer $h \in [H-1]$ we define the following good events. Event G_1^h states that for every state-action pair $(s_h, a_h) \in S_h \times A$, if s_h is β-reachable for $\widehat{P}$, then sufficient number of samples were collected for the pair (s_h, a_h). Event G_2^h states that for every state-action pair $(s_h, a_h) \in S_h \times A$ such that s_h is β-reachable, the learned dynamics $\widehat{P}$ approximate the true dynamics P up to a small error of γ, i.e., $\|\widehat{P}(\cdot|s_h, a_h) - P(\cdot|s_h, a_h)\|_1 \leq \gamma$.

Event G_3^h state that for every state-action pair $(s_h, a_h) \in S_h \times A$ such that s_h is β-reachable for $\widehat{P}$, the ERM oracle returns a function $f_{s_h,a_h}(c)$ with low generalization error. Let $G_i = \cap_{h \in [H-1]} G_i^h$ for all $i \in \{1,2,3\}$. We analyse the value error caused by both the dynamics and rewards approximation under the good events. We also show that the event $G_1 \cap G_2 \cap G_3$ holds with high probability.

For the analysis, we define an intermediate MDP $\widetilde{\mathcal{M}}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}, r^c, s_0, H)$, which differ from $\mathcal{M}(c)$ only in the dynamics and from $\widehat{\mathcal{M}}(c)$ only in the rewards function. Let α denote the maximal approximation error. For $\beta = \frac{\epsilon}{B|S|H}$, $B = 24$ and $\gamma = \frac{\epsilon}{48|S|H^2}$ we obtain the following bound on the value difference caused by the dynamics approximation.

Lemma 2 *If event $G_1 \cap G_2$ holds, then for every context $c \in \mathcal{C}$ and context-dependent policy $\pi = (\pi_c)_{c \in \mathcal{C}}$ it holds that $|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{M}(c)}(s_0)| \leq \epsilon/16$.*

Proof [sketch] Under the event $G_1 \cap G_2$, for every β-reachable state s for $\widehat{P}$ and an action a it holds that $\|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\| \leq \gamma$. We show in Lemma 47 that for any policy $\pi = (\pi_c)_{c \in \mathcal{C}}$ it holds that $\forall c \in \mathcal{C}\ \forall h \in [H], \sum_{s_h \in S_h} |q_h(s_h|\pi_c, P) - q_h(s_h|\pi_c, \widehat{P})| \leq \gamma h + \beta \sum_{k=0}^{h-1} |S_k|$. For our choice of β and γ, the latter yields that $|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{M}(c)}(s_0)| \leq \sum_{h=0}^{H-1} \sum_{s_h \in S_h} |q_h(s_h|\pi, P) - q_h(s_h|\pi, \widehat{P})| \leq \epsilon/16$. ■

The following lemma bounds the expected value difference caused by the rewards approximation.

Lemma 3 *If event G_3 holds, then for every policy $\pi = (\pi_c)_{c \in \mathcal{C}}$, it holds that*

$$\mathbb{E}_{c \sim \mathcal{D}}[|V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0) - V^{\pi_c}_{\widetilde{M}(c)}(s_0)|] \leq \epsilon/8 + \alpha H.$$

By combining Lemmas 2 and 3 we obtain an expected value difference bound for any policy.

Lemma 4 *If events G_1 ,G_2 and G_3 hold, then for every policy $\pi = (\pi_c)_{c\in\mathcal{C}}$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq 3\epsilon/16 + \alpha H.$$

The above lemma establishes Theorem 1. For detailed analysis, see Appendix C.

Known Dynamics. When the context free dynamics is known, we can achieve better sample complexity, as the following theorem states. (For more details, see Appendix B.)

Theorem 5 *With probability $1-\delta$ it holds that $\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha H$, after collecting $\tilde{O}\Big(d\epsilon^{-2}H^2|S|^4|A|^3\log\frac{|S||A|}{\delta}\Big)$ trajectories for the ℓ_1 loss, and $\tilde{O}\Big(d\epsilon^{-4}H^4|S|^6|A|^5\log\frac{|S||A|}{\delta}\Big)$ for the ℓ_2 loss.*

5. Context Dependent Dynamics

In this section we address the challenging model of context dependent dynamics, where each context induces a potentially different dynamics. Clearly, this implies that for any policy π, the occupancy measure is determined by the context. Hence, a state $s \in S$ that is highly-reachable for some context $c_1 \in \mathcal{C}$ might be poorly-reachable for a different context $c_2 \in \mathcal{C}$.

For the *unknown context-dependent dynamics* we move the approximation from being per state-action pair to being per layer. (While this is a slight modification of the assumption, it is still very reasonable.) Conceptually, we move from collecting samples per state-action, to collecting samples per layer, and those samples are index by context-state-action tuples (c, s, a). We construct an unbiased data set with respect to those tuples. However, we guarantee that the collected samples have the "right" marginal distribution over the entire layer. This requires a much more involved algorithm and analysis. Thus, we extend the definition of reachability.

Good contexts of a state. For a state $s \in S$ we define the set of β-*good contexts* with respect to P as $\mathcal{C}^\beta(s|P) := \{c \in \mathcal{C} : s \text{ is } \beta\text{-reachable for } P^c\}$. Given the approximated dynamics $\widehat{P}^c$, we define $\widehat{\mathcal{C}}^\beta(s) := \mathcal{C}^\beta(s|\widehat{P})$. Note that there might be no context c which is good for all states (unlike in the context-independent dynamics). The following defines the modification of the β-reachability.

(γ,β)-good states. Let $\gamma, \beta \in (0, 1]$. For each layer $h \in [H]$ we define the set of (γ, β)-*good states with respect to* P as $S^{\gamma,\beta}_{h,P} := \{s_h \in S_h : \mathbb{P}_{c\sim\mathcal{D}}[c \in \mathcal{C}^\beta(s_h|P)] \geq \gamma\}$. Given the approximated dynamics $\widehat{P}$, we define $\widehat{S}^{\gamma,\beta}_h = S^{\gamma,\beta}_{h,\widehat{P}}$. We define the *target domain* $\mathcal{X}^{\gamma,\beta}_h = \{(c, s, a) : s \in \widehat{S}^{\gamma,\beta}_h, c \in \widehat{\mathcal{C}}^\beta(s), a \in A\}$ of collected examples.

Function approximation for each layer. A major hurdle caused by the context-dependent dynamics is that for each (s_h, a_h) the probability of sampling $((c, s_h, a_h), r_h)$ is highly dependent on the context c through the dynamics P^c. Our goal is to create an unbiased sample, which we will perform for an entire level, but this seems very challenging to achieve at the individual state-action level. To exemplify that, assume we observe the context c and run some policy π to generate a trajectory $\tau = (c, s_0, a_0, r_0, s_1, \ldots, s_H)$. For every layer $h \in [H-1]$ the distribution of the example $((c, s_h, a_h), r_h)$ is $\mathcal{D}(c) \cdot q_h(s_h, a_h|P^c, \pi) \cdot \mathbb{P}[R^c(s_h, a_h) = r_h|c, s_h, a_h]$. If we aim to collect samples for each state-action pair separately, the appropriate distribution is $\mathcal{D}(c) \cdot \mathbb{P}[R^c(s_h, a_h) = r_h|c, s_h, a_h]$. Hence, we need to guarantee that the contexts are sampled in an unbiased way, i.e., the marginal context distribution for any state-action pair is $\mathcal{D}$. If the dynamics were known, we would overcome this using Importance Sampling. When the dynamics are unknown, we side step this issue, and create an unbiased sample at the layer level. The advantage of tuples $((c, s, a), r)$ is that we can sample them for the entire layer and obtain good estimates on average, and then claim that for the "important" states we have a good approximation. At the layer level, the occupancy measure determines the joint distribution over $(c, s_h, a_h) \in \mathcal{X}^{\gamma,\beta}_h$, which is the desired distribution. Hence, we approximate the rewards as a function of context, state and action. Similarly for the dynamics.

Dynamics and rewards function approximation. We slightly modify our assumption for the function approximation class, which works per layer and not per state-action.

For each layer $h \in [H-1]$ we have a function class for the dynamics $\mathcal{F}^P_h = \{f^P_h : \mathcal{C} \times S_h \times A \times S_{h+1} \to [0, 1]\}$ and for the rewards $\mathcal{F}^R_h = \{f^R_h : \mathcal{C} \times S_h \times A \to [0, 1]\}$. Intuitively, given that we are in state s, perform action a and the

context is c, the function $f_h^P \in \mathcal{F}_h^P$ and $f_h^R \in \mathcal{F}_h^R$, approximates the transition probability to state s', i.e., $P^c(s'|s,a)$, and the expected reward, $r^c(s,a)$, respectively. For the dynamics approximation, we also assume reliability for every layer h, i.e., $\alpha_1(\mathcal{F}_h^P) = \alpha_2^2(\mathcal{F}_h^P) = 0$.

The functions $N_P(\mathcal{F}_h^P, \epsilon, \delta)$ and $N_R(\mathcal{F}_h^R, \epsilon, \delta)$ map a function class, required accuracy ϵ and confidence δ to the required number of examples for the ERM oracle to have the desired guarantees. For the dynamics the ERM guarantee is that with probability $1-\delta$, $\mathbb{E}[\ell(f_h^P(x), y)] \leq \epsilon$. For the rewards the guarantee is that $\mathbb{E}[\ell(f_h^R(x), y)] \leq \epsilon + \alpha$, where α is the approximation error.

Layer dynamics realizability assumption. We assume that for each layer $h \in [H-1]$ we have a function $f_h^P \in \mathcal{F}_h^P$ such that $f_h^P(c, s, a, s') = P^c(s'|s,a)$. We relay on this assumption when approximating the dynamics, in order to properly estimate whether a state is (γ, β)-good.

Sample collection. In algorithm EXPLORE-UCDD (see Algorithm 2 or 15), we learn the dynamics and rewards for each layer, given the approximated dynamics of previous layers. When learning the dynamics associated with layer h, we collect examples of the form (c, s_h, a_h, s_{h+1}) from each trajectory $\tau = (c, s_0, a_0, r_0, \ldots, s_H)$ that contains (s_h, a_h). We add to our data set a sample $((c, s_h, a_h, s_{h+1}), 1)$ and samples $((c, s_h, a_h, s'), 0)$ for each $s' \in S_{h+1} \setminus \{s_{h+1}\}$. This reduces the learning dynamics to a regression problem. When learning the rewards associated with layer h, as before, we collect samples of the form $((c, s_h, a_h), r_h)$, and use them to approximate the rewards function for the layer.

Algorithm overview. Algorithm EXPLORE-UCDD (see Algorithms 2 and 15) runs in H phases, one per layer. In phase $h \in [H-1]$ we maintain an approximate dynamics for all previous layers $k \leq h-1$, which we already learned. In phase h we run multiple iterations, in each iteration, (1) we select at random a (approximately) (γ, β)-good state $s_h \in \widetilde{S}_h^{\gamma,\beta}$ and an action $a_h \in A$. (2) Given a context c and a state s_h we compute a policy $\widehat{\pi}_{s_h}^c$ which maximizes the probability of reaching state s_h under the approximated dynamics $\widehat{P}^c$. (3) We run $\widehat{\pi}_{s_h}^c$. If it reaches s_h we play a_h, get a reward r_h and transits to s_{h+1}, we add: (a) to the dynamics data set $Sample^P(h)$: $((c, s_h, a_h, s'), \mathbb{I}[s_{h+1} = s'])$ for each $s' \in S_{h+1}$, (b) to the reward data set $Sample^R(h)$: $((c, s_h, a_h), r_h)$. (4) After collecting sufficient number of samples, we use the ERM oracle to (a) approximate the transition probabilities of layer h, i.e., $f_h^P = \texttt{ERM}(\mathcal{F}_h^P, Sample^P(h), \ell)$. (b) approximate the rewards function of layer h, i.e., $f_h^R = \texttt{ERM}(\mathcal{F}_h^R, Sample^R(h), \ell)$. Consider the following algorithm sketch for the ℓ_1 loss and the parameters set $\delta_1 = \frac{\delta}{8H}$, $\delta_2 = \frac{\delta}{8|S|}$, $\epsilon_2 = \gamma/4$, $\beta = \gamma = \frac{\epsilon}{20|S|H}$, $\rho = \frac{\beta}{16|S|H}$, $\epsilon_P = \frac{\epsilon^2}{10 \cdot 16 \cdot 20 |A||S|^4 H^3}$ and $\epsilon_R = \frac{\epsilon^2}{20^2 |S||A| H^2}$, where $m_h = 2\max\{N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2)), N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2))\}$.

Algorithm 2 EXPLORE Unknown Context Dependent Dynamics (sketch for the ℓ_1 loss)

for $h \in [H-1]$ **do**
  compute $\widehat{P}^c$, the approximated context-dependent dynamics, up to layer $h-1$
  let $\widetilde{S}_h^{\gamma,\beta}$ denote the approximation of the set of (γ, β)-good states w.r.t $\widehat{P}^c$
  for $t = 1, 2, \ldots, \left\lceil \frac{8|S|}{\beta \cdot \gamma}(\ln(\frac{1}{\delta_1}) + m_h) \right\rceil$ **do**
    observe context c_t and choose $(s_h, a_h) \in \widetilde{S}_h^{\gamma,\beta} \times A$ uniformly at random
    compute $\widehat{p}_s^{c_t}$, the highest probability to visit s in $\widehat{P}^{c_t}$, and a policy $\widehat{\pi}_s^{c_t}$ that reaches it
    set $\widehat{\pi}_{s_h}^{c_t}(s_h) \leftarrow a_h$
    if $\widehat{p}_{s_h}^{c_t} \geq \beta$ **then**
      run $\widehat{\pi}_{s_h}^{c_t}$ and generate trajectory τ
      if (s_h, a_h, r_h, s_{h+1}) is in τ **then**
        add $((c_t, s_h, a_h), r_h)$ to the rewards sample
        add $\{((c_t, s_h, a_h, s'_{h+1}), \mathbb{I}[s_{h+1} = s'_{h+1}]) : s'_{h+1} \in S_{h+1}\}$ to the dynamics sample
  if $|Sample^P(h)| \geq m_h$ **then**
    $f_h^P = \texttt{ERM}(\mathcal{F}_h^P, Sample^P(h), \ell_1)$
    $f_h^R = \texttt{ERM}(\mathcal{F}_h^R, Sample^R(h), \ell_1)$
  else
    return FAIL
return $\{f_h^R, f_h^P, \widehat{S}_h^{\gamma,\beta} : \forall h \in [H-1]\}$

Approximate optimal policy. Given a context c, we define an MDP with the learned rewards and dynamics and compute its optimal policy. We define the approximated CMDP as $(\mathcal{C}, S \cup \{s_{sink}\}, A, \widehat{\mathcal{M}})$ where $\widehat{\mathcal{M}}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}^c, \widehat{r}^c, s_0, H)$. We define $\widehat{r}^c$ as $\widehat{r}^c(s_h, a_h) = f_h^R(c, s_h, a_h) \cdot \mathbb{I}[s_h \in \widetilde{S}_h^{\gamma,\beta}, c \in \widehat{\mathcal{C}}^\beta(s_h)]$ and $\widehat{r}^c(s_{sink}, a_h) = 0$. The dynamics $\widehat{P}^c$ uses the dynamics approximation functions normalized, and states which are not (γ, β)-good transition to the sink. For any other state-action we define a transition to the sink w.p. 1. For a context $c \in \mathcal{C}$, let $\widehat{\pi}_c^\star$ and $\pi_c^\star$ denote an optimal policy for $\widehat{\mathcal{M}}(c)$ and $\mathcal{M}(c)$, respectively. Let α denote the maximal approximation error over all layers h w.r.t. $\ell \in \{\ell_1, \ell_2\}$. We obtain the following.

Theorem 6 *With probability at least $1 - (\delta + \frac{\epsilon}{5})$ it holds that $\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha H$, after collecting $\tilde{O}\Big(d\epsilon^{-6}H^9|S|^{11}|A|^2 \log \frac{H}{\delta}\Big)$ trajectories for the ℓ_1 loss, and $\tilde{O}\Big(d\epsilon^{-8}H^{13}|S|^{15}|A|^2 \log \frac{H}{\delta}\Big)$ for the ℓ_2 loss.*

Analysis outline. In the analysis, we show that the following good events hold, with high probability, for every layer $h \in [H-1]$: (1) Every state $s_h \in \widehat{S}_h^{\gamma,\beta}$ we identify correctly. (2) We collect sufficient number of samples of $(c, s_h, a_h) \in \mathcal{X}_h^{\gamma,\beta}$. (3) Our approximation of the dynamics has low generalization error. (4) Our approximation of the rewards has low generalization error. The above form the good events G_1, G_2, G_3 and G_4, and there is a choice of parameters such that they all hold with high probability.

Our analysis (see Appendix E) shows that under these good events, our approximation of the dynamics and rewards for every layer h and $(c, s, a) \in \mathcal{X}_h^{\gamma,\beta}$ is accurate, with high probability. We also show that any $(c, s, a) \notin \mathcal{X}_h^{\gamma,\beta}$ adds only small error to our estimations. Hence, in expectation over $c \in \mathcal{C}$ we have small errors for both rewards and dynamics.

Lemma 7 *Under the good events G_1, G_2 and G_3, for all $(c, s_h, a_h) \in \mathcal{X}_h^{\gamma,\beta}$, there exist parameters choice such that $\mathbb{P}\Big[\|\widehat{P}^c(\cdot|s_h, a_h) - P^c(\cdot|s_h, a_h)\|_1 \leq \epsilon/40|S|H^2 \Big| (c, s_h, a_h) \in \mathcal{X}_h^{\gamma,\beta}\Big] \geq 1 - \epsilon/(10|S||A|H)$, where $\widehat{P}^c$ is the approximated dynamics and we ignore s_{sink}.*

Proof [sketch for the ℓ_1 loss] By the good event G_3 and Markov's inequality the following holds

$$\mathbb{P}\Big[|f_h^P(c, s_h, a_h, s_{h+1}) - P^c(s_{h+1}|s_h, a_h)| \geq \rho \Big| (c, s_h, a_h) \in \mathcal{X}_h^{\gamma,\beta}\Big] \leq \epsilon_P/\rho.$$

Since $\sum_{s_{h+1}\in S_{h+1}} P^c(s_{h+1}|s_h, a_h) = 1$, by union bound over $s_{h+1} \in S_{h+1}$ and $\widehat{P}^c$ definition we obtain

$$\mathop{\mathbb{P}}_{(c,s_h,a_h)}\Big[\forall s_{h+1} \in S_{h+1}.\ \frac{P^c(s_{h+1}|s_h, a_h) - \rho}{1 + \rho|S|} \leq \widehat{P}^c(s_{h+1}|s_h, a_h) \leq \frac{P^c(s_{h+1}|s_h, a_h) + \rho}{1 - \rho|S|} \Big| (c, s_h, a_h) \in \mathcal{X}_h^{\gamma,\beta}\Big] \geq 1 - \frac{\epsilon_P}{\rho}|S_{h+1}|.$$

Now, using simple calculation, we derive the lemma for our choice of β, ρ,ϵ_P. ■

In the analysis, we define an intermediate MDP $\widetilde{\mathcal{M}}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}^c, r^c, s_0, H)$, where $\widehat{P}^c$ is the approximated dynamics and r^c is the true rewards function extended to s_{sink} by defining $\forall c \in \mathcal{C},\ a \in A : r^c(s_{sink}, a) := 0$. We use it to estimate the influence of the error on the rewards separately from the error of the dynamics.

The following lemma states the expected value-difference caused by the dynamics approximation.

Lemma 8 *Under the good events G_1, G_2 and G_3, there exist a parameters choice such that for every policy $\pi = (\pi_c)_{c\in\mathcal{C}}$, it holds that $\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|] \leq 0.225\epsilon$.*

Proof [sketch for the ℓ_1 loss] For a context c, let $G(c)$ denote the following event

$$G(c) = \{\forall h \in [H].\quad \|q_h(\cdot|\pi_c, P^c) - q_h(\cdot|\pi_c, \widehat{P}^c)\|_1 \leq 3\epsilon/(40H)\}.$$

We show in Lemma 118 that

$$\mathbb{P}_c[G(c)] = \mathbb{P}_c\left[\forall h \in [H]. \quad \|q_h(\cdot|\pi_c, P^c) - q_h(\cdot|\pi_c, \widehat{P}^c)\|_1 \leq 3\epsilon/(40H)\right] \geq 1 - 3\epsilon/(20H),$$

yielding the lemma since,

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|] \leq \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=1}^{H-1} \|q_h(\cdot|\pi_c, P^c) - q_h(\cdot|\pi_c, \widehat{P}^c)\|_1 \middle| G(c)\right] + 3\epsilon/20 = 0.225\epsilon.$$

■

The next lemma states the expected value-difference caused by the rewards approximation.

Lemma 9 *Under the good events G_1, G_2 and G_3, there exist a parameters choice such that for every policy $\pi = (\pi_c)_{c\in\mathcal{C}}$ it holds that $\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\widetilde{M}(c)}(s_0) - V^{\pi_c}_{\widehat{M}(c)}(s_0)|] \leq 0.2\ \epsilon + \alpha H$.*

By combining Lemmas 8 and 10 we obtain the following lemma, which establishes Theorem 6.

Lemma 10 *Under the good events G_1, G_2,G_3 and G_4, for every policy $\pi = (\pi_c)_{c\in\mathcal{C}}$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|] \leq 0.5\epsilon + \alpha H.$$

Known context-dependent dynamics. For this case we continue in the approach of collecting examples for each (γ, β)-good state and action. We obtain the following result. For more details, see Appendix D.

Theorem 11 *With probability $1-\delta$ it holds that $\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha H$, after collecting $\tilde{O}\left(d\epsilon^{-6}H^5|S|^5|A|^3\log\frac{|S||A|}{\delta}\right)$ trajectories for the ℓ_1 loss, and $\tilde{O}\left(d\epsilon^{-8}H^7|S|^5|A|^3\log\frac{|S||A|}{\delta}\right)$ trajectories for the ℓ_2.*

6. Discussion

To the best of our knowledge, our work is the first to drive sample complexity bounds for CMDP, without assuming any additional assumptions regarding it. Our sample complexity bounds do not depend on the size of the context space, which allows it to be huge.

An interesting future research direction is to drive lower bounds. Clearly, the sample complexity is lower bounded by classical PAC lower bounds, of $\Omega(d\epsilon^{-2}\log\frac{1}{\delta})$, where d is the complexity dimension (i.e., VC, Natrajan, Fat-shattering, Pseudo dimension). Also, non-contextual MDP sample complexity lower bounds apply to our case and give $\Omega\left(\epsilon^{-2}|S||A|\log(|S|/\delta)\right)$. Deriving stronger lower bounds that are based on the special structure of the CMDP is an important open problem.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation(grant number 993/17), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University.

OL thanks Idan Attias for helpful discussions and patient explanations about ERM. OL thanks Aviv Rosenberg for helpful advices and his comments on a former version of the paper.

References

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert Schapire. Contextual bandit learning with predictable rewards. In *Artificial Intelligence and Statistics*, pages 19–26. PMLR, 2012.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

Martin Anthony, Peter L Bartlett, Peter L Bartlett, et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

Jinglin Chen, Aditya Modi, Akshay Krishnamurthy, Nan Jiang, and Alekh Agarwal. On the statistical efficiency of reward-free exploration in non-linear rl. *arXiv preprint arXiv:2206.10770*, 2022.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Dylan Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR, 2018.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

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

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pages 7599–7608. PMLR, 2021.

Aditya Modi and Ambuj Tewari. No-regret exploration in contextual reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 829–838. PMLR, 2020.

Aditya Modi, Nan Jiang, Satinder Singh, and Ambuj Tewari. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pages 597–618. PMLR, 2018.

Shuang Qiu, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. On reward-free RL with kernel and neural function approximations: Single-agent MDP and markov game. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8737–8747. PMLR, 2021. URL http://proceedings.mlr.press/v139/qiu21d.html.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 2021.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Found. Trends Mach. Learn.*, 12(1-2):1–286, 2019.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Yunbei Xu and Assaf Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*, 2020.

Zihan Zhang, Simon Du, and Xiangyang Ji. Near optimal reward-free reinforcement learning. In *International Conference on Machine Learning*, pages 12402–12412. PMLR, 2021.

Appendix A. Real-Valued Function Class Dimensions

Our sample complexity bounds are stated in the terms of the Pseudo and γ-fat-shattering dimension of the function class, which are complexity measured for learning real-valued function classes.

In the following, we define the Pseudo and γ-fat-shattering dimension of a function class $\mathcal{F}$. For more information, see Anthony et al. (1999).

A.1 Pseudo Dimension Definition

Definition 12 (pseudo-shattering, Definition 11.1 in Anthony et al. (1999)) *Let $\mathcal{F}$ be a set of function from a domain $\mathcal{X}$ to $\mathbb{R}$ and suppose that $\mathcal{S} = \{x_1, x_2, \ldots, x_m\} \subseteq \mathcal{X}$. Then $\mathcal{S}$ is pseudo-shattered by $\mathcal{F}$ if there are real numbers $r_1, r_2, \ldots, r_m$ such that for each $b \in \{0,1\}^m$ there is a function f_b in $\mathcal{F}$ with $sign(f_b(x_1) - r_i) = b_i$ for $1 \leq i \leq m$. We say that $r = (r_1, r_2, \ldots, r_m)$ witnesses the shattering.*

Definition 13 (pseudo-dimension, Definition 11.2 in Anthony et al. (1999)) *Suppose that $\mathcal{F}$ is a set of functions from a domain $\mathcal{X}$ to $\mathbb{R}$. Then $\mathcal{F}$ has pseudo-dimension d if d is the maximum cardinality of a subset $\mathcal{S}$ of $\mathcal{X}$ that is pseudo-shattered by $\mathcal{F}$. If no such maximum exists, we say that $\mathcal{F}$ has infinite pseudo-dimension. The pseudo-dimension of $\mathcal{F}$ is denoted $Pdim(\mathcal{F})$.*

A.2 Fat-Shattering Dimension Definition

Definition 14 (γ-shattering, Definition 11.10 in Anthony et al. (1999)) *Let $\mathcal{F}$ be a set of functions mapping from a domain $\mathcal{X}$ to $\mathbb{R}$ and suppose that $\mathcal{S} = \{x_1, x_2, \ldots, x_m\} \subseteq \mathcal{X}$. Suppose also that γ is a positive real number. Then $\mathcal{S}$ is γ-shattered by $\mathcal{F}$ if there are real numbers $r_1, r_2, \ldots, r_m$ such that for each $b \in \{0,1\}^m$ there is a function f_b in $\mathcal{F}$ with $f_b(x_i) \geq r_i + \gamma$ if $b_i = 1$, and $f_b(x_i) \leq r_i - \gamma$ if $b_i = 0$, for $1 \leq i \leq m$. We say that $r = (r_1, r_2, \ldots, r_m)$ witnesses the shattering.*

Thus, $\mathcal{S}$ is γ-shattered if it is shattered with a 'width of shattering' of at least γ. This notion of shattering leads to the following dimension.

Definition 15 (Fat shattering dimension, Definition 11.11 in Anthony et al. (1999)) *Suppose that $\mathcal{F}$ is a set of functions from a domain $\mathcal{X}$ to $\mathbb{R}$ and that $\gamma > 0$. Then $\mathcal{F}$ has γ-dimension d if d is the maximum cardinality of a subset $\mathcal{S}$ of $\mathcal{X}$ that is γ-shattered by $\mathcal{F}$. If no such maximum exists, we say that $\mathcal{F}$ has infinite γ-dimension. The γ-dimension of $\mathcal{F}$ is denoted $fat_{\mathcal{F}}(\gamma)$. This defined a function $fat_{\mathcal{F}} : \mathbb{R}^+ \to \mathbb{N} \cap \{0, \infty\}$, which we call the fat-shattering dimension of $\mathcal{F}$. We say that $\mathcal{F}$ has finite fat-shattering dimension whenever it is the case that for all $\gamma > 0$, $fat_{\mathcal{F}}(\gamma)$ is finite.*

Remark 16 *For every function class $\mathcal{F}$ and $\gamma > 0$ it holds that $fat_{\mathcal{F}}(\gamma) \leq Pdim(\mathcal{F})$.*

A.3 Sample Complexity Results

The following theorems state that if the Pseudo/fat-shattering dimension of the function class $\mathcal{F}$ is finite, then $\mathcal{F}$ has a uniform convergence property. Hence $\mathcal{F}$ is learnable using an ERM algorithm up to an ϵ error, with probability at least $1 - \delta$. $m(\epsilon, \delta)$ is the required sample complexity for the learning task.

Theorem 17 (Adaption of Theorem 19.2 in Anthony et al. (1999)) *Let $\mathcal{F}$ be a hypothesis space of real valued functions with a finite pseudo dimension, denoted $Pdim(\mathcal{F}) < \infty$. Then, $\mathcal{F}$ has a uniform convergence with*

$$m(\epsilon, \delta) = O\Big(\frac{1}{\epsilon^2}(Pdim(\mathcal{F}) \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})\Big).$$

Theorem 18 (Adaption of Theorem 19.1 in Anthony et al. (1999)) *Let $\mathcal{F}$ be a hypothesis space of real valued functions with a finite fat-shattering dimension, denoted $fat_{\mathcal{F}}(\gamma)$. Then, $\mathcal{F}$ has a uniform convergence with*

$$m(\epsilon, \delta) = O\Big(\frac{1}{\epsilon^2}(fat_{\mathcal{F}}(\epsilon/256) \ln^2 \frac{1}{\epsilon} + \ln \frac{1}{\delta})\Big).$$

Appendix B. Known and Context-Free Dynamics

In this section we assume a known context-independent transition probability function, i.e., $\forall c \in \mathcal{C} : P^c = P$ and P is known to the learner.

B.1 Algorithm

Let use first outline the main ideas of our algorithm EXPLORE-KCFD (Algorithm 5). Since the dynamics are context-free we have that for any policy π and state $s_h \in S_h$, the probability to visit s_h using $\pi : S \to A$ that is identical for any context c, i.e., for any context $c \in \mathcal{C}$, we have $q_h(s_h|\pi, P^c) = q_h(s_h|\pi, P)$. Given the dynamics P for each state $s_h \in S_h$ Algorithm EXPLORE-KCFD computes, using the planning oracle, a policy $\pi_{s_h} := \arg\max_{\pi:S\to A} q_h(s_h|\pi, P)$ and the probability p_{s_h} that π_{s_h} reaches s_h. Then, the probability that π_{s_h} reaches s_h is used to check whether the state is β-reachable with respect to the known dynamics P for β that will be determined later. We use π_{s_h,a_h} to generate trajectory τ. If τ contains state s_h we add $((c, s_h, a_h), r_h)$ to our reward sample of (s_h, a_h). Clearly the collected samples are i.i.d. After collecting "sufficient" number of contexts-rewards we use the ERM oracle to compute a function $f_{s_h,a_h}(c)$ that approximates $r^c(s_h, a_h)$.

We set a refined desired accuracy per state, which depends on its maximum probability, and saves a $1/\epsilon$ factors in the sample complexity. States which are very hard to reach, we do not approximate. States which are very easy to reach, we want maximum accuracy. For intermediate levels we have a gradual accuracy dependency. This is captured in our definition of the accuracy-per-state function $\epsilon_\star$, which depends the probability to visit state s, i.e., $p_s := q_h(s_h|\pi_{s_h}, P)$.

$$\epsilon_\star(p_s) = \begin{cases} 1 & , \text{ if } p_s < \frac{\epsilon}{B|S|} \\ \frac{\epsilon}{BH|S||A|} & , \text{ if } p_s > \frac{1}{|S|} \\ \frac{\epsilon}{Bp_s|S||A|} & , \text{ if } p_s \in [\frac{\epsilon}{B|S|}, \frac{1}{|S|}] \end{cases}$$

where $B > 0$ is a constant.

Since we sample only the β-reachable states for every action $a \in A$, Algorithm EXPLORE-KCFD (Algorithm 5) learns an approximation of the context-dependent reward function given P (i.e., the context-free dynamics is known to the learner) and $N_R(\mathcal{F}, \epsilon, \delta)$ (the sample complexity function of the ERM oracle) efficiently.

In more details:

Let $\texttt{Planning}(M)$ denote a planning algorithm which gets as input a MDP $M = (S, A, P, r, s_0, H)$ The planning algorithm returns an optimal policy for the H finite horizon return and the appropriate value function. It runs in time $O(|S||A|H)$.[3]

Algorithm PaP (Algorithm 4) returns for each state $s \in S$ a policy π_s that maximizes the probability to visit it, denoted p_s.

Algorithm EXPLORE-KCFD (Algorithm 5), uses π_s to sample each β-reachable state s for each action $a \in A$ for sufficiently many times to create a large enough sample $Sample(s, a)$ containing the tuples $((c, s, a), R^c(s, a))$. Then, we feed the ERM with that sample and output an approximation of the reward function $r^c(s, a)$ using $f_{s,a} = \texttt{ERM}(\mathcal{F}^R_{s,a}, Sample(s, a), \ell)$. For not-$\beta$-reachable state s we set $f_{s,a} = 0 \quad \forall a \in A$. The algorithm returns $F = \{f_{s,a}, \quad \forall (s, a) \in S \times A\}$, or *Fail* if insufficient number of samples have been collected for any β-reachable state.

To improve our overall sample complexity, we define the accuracy-per-state function, for both ℓ_1 and ℓ_2 :

For ℓ_1 we define it as

$$\epsilon^1_\star(p_s) := \epsilon_\star(p_s) = \begin{cases} 1 & , \text{ if } p_s < \frac{\epsilon}{6|S|} \\ \frac{\epsilon}{6H|S||A|} & , \text{ if } p_s > \frac{1}{|S|} \\ \frac{\epsilon}{6p_s|S||A|} & , \text{ if } p_s \in [\frac{\epsilon}{6|S|}, \frac{1}{|S|}] \end{cases}$$

3. For example, policy iteration is such a planning algorithm. It is finds an optimal policy (and its value) with respect to the finite horizon return, and can be computed in time polynomial in the MDPs parameters.

For ℓ_2 we define it as $\epsilon_\star^2(p_s) := (\epsilon_\star(p_s))^2$, or equivalently,

$$\epsilon_\star^2(p_s) = \begin{cases} 1 & , \text{ if } p_s < \frac{\epsilon}{6|S|} \\ \frac{\epsilon^2}{36H^2|S|^2|A|^2} & , \text{ if } p_s > \frac{1}{|S|} \\ \frac{\epsilon^2}{36(p_s)^2|S|^2|A|^2} & , \text{ if } p_s \in [\frac{\epsilon}{6|S|}, \frac{1}{|S|}] \end{cases}.$$

In both functions, where the required accuracy for a state s is 1, we do not sample it.

Algorithm EXPLOIT-KCFD (Algorithm 6) get as inputs the MDP parameters and the functions approximate the rewards (that computed using EXPLORE-KCFD algorithm). Given a context c it computed the approximated MDP $\widehat{\mathcal{M}}(c)$ and use it to compute a near optimal policy $\pi_c^\star$. Then, it run $\pi_c^\star$ to generate trajectory. Recall that $\widehat{\mathcal{M}}(c) = (S, A, P, s_0, \widehat{r}^c, H)$ where we define $\forall s \in S, a \in A : \widehat{r}^c(s,a) = f_{s,a}(c)$.

Algorithm 3 Find Fast Policy (FFP)

inputs:
- MDP parameters:S - the states space,A - a finite actions space,P - transition probabilities matrix,s_0 - the unique start state,H - the horizon length.
- s - the target state.

let $r(s', a) = \mathbb{I}[s' = s]$
$(p_s, \pi_s) \leftarrow \texttt{Planning}(M = (S, A, P, r, s_0, H))$
return: (p_s, π_s)

Algorithm 4 Policies and Probabilities(PaP)

inputs: MDP parameters:$S = \{S_0, S_1, \ldots, S_H\}$ - a layered states space,A - a finite actions space,P - transition probabilities matrix,s_0 - the unique start state, H - the horizon length.
for $h \in [H-1]$ **do**
  for $s \in S_h$ **do**
    $(p_s, \pi_s) \leftarrow \texttt{FFP}(S, A, P, s_0, H, s)$
return : $\{(p_s, \pi_s) \ \ \forall s \in S\}$

Remark 19 *Since the reward function defined in algorithm FFP has a reward of* 1 *for state* s *and* 0 *for any other state (regardless of the action), the value function of any policy computed using this rewards function is exactly the probability the policy visits state* s*.*

B.2 Analysis

B.2.1 Analysis Outline

In the following analysis, our goal is to bound the expected value difference between the true model $\mathcal{M}(c)$ and $\widehat{\mathcal{M}}(c)$, for any context-dependent policy $\pi = (\pi_c)_{c \in \mathcal{C}}$, with high probability. (See Lemmas 24 and 26, for the ℓ_2 and ℓ_1 losses, respectively).

Using that bound, we derive a bound on the expected value difference between the true optimal context-dependent policy $\pi^\star = (\pi_c^\star)_{c \in \mathcal{C}}$ and our approximated optimal policy $\widehat{\pi}^\star = (\hat{\pi}_c^\star)_{c \in \mathcal{C}}$, which holds with high probability. (See Theorems 25 and 27 for the ℓ_2 and ℓ_1 losses, respectively).

We present analysis for both ℓ_2 (see Sub-subsection B.2.3) and ℓ_1 (see Sub-subsection B.2.4 losses in the agnostic case.

Lastly, we derive sample complexity bound using known uniform convergence sample complexity bounds for the Pseudo dimension (See Theorem 28) and the fat-shattering dimension (See Theorem 29). For the sample complexity analysis, see Sub-subsection B.3.1 for the ℓ_2 loss, and B.3.2 for the ℓ_1 loss.

Algorithm 5 Explore Rewards Known Context-Free Dynamics (EXPLOR-KCFD)

inputs:

- MDP parameters: $S = \{S_0, S_1, \ldots, S_H\}$ - a layered states space,A - a finite actions space, P - transition probabilities matrix, s_0 - the unique start state, H - the horizon length.
- Accuracy and confidence parameters: ϵ,δ.
- ℓ - the loss function ($\ell \in \{\ell_1, \ell_2\}$).
- $\forall s \in S, a \in A: \ \mathcal{F}^R_{s,a}$ - the function classes use to approximate the rewards function.
- $N_R(\mathcal{F}, \epsilon, \delta)$ - sample complexity function for approximating the reward with respect to ℓ.
- $\epsilon^i_\star(\cdot)$ - the accuracy-per-state function, (assumed to be $\epsilon^1_\star$ or $\epsilon^2_\star$, with accordance to ℓ).

set $\delta_1 = \frac{\delta}{4|S||A|}$, $\beta = \frac{\epsilon}{6|S|}$
$\{(p_s, \pi_s)\} \leftarrow \texttt{PaP}(S, A, P, s_0, H)$
for $h \in [H-1]$ **do**
  for $s \in S_h$ **do**
    if $p_s \geq \beta$ **then**
      for $a \in A$ **do**
        set $\pi_s(s) \leftarrow a$
        compute the required number of episodes:

$$T_{s,a} = \left\lceil \frac{2}{p_s}(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}^R_{s,a}, \epsilon^i_\star(p_s), \delta_1)) \right\rceil$$

        initialize $Sample(s,a) = \emptyset$
        $\pi_s(s) \leftarrow a$
        for $t = 1, 2, \ldots, T_{s,a}$ **do**
          observe context c
          run π_s to generate trajectory τ_t
          if $(s, a, r) \in \tau_t$, for a reward $r \in [0, 1]$ **then**
            update $Sample(s,a) = Sample(s,a) + \{((c, s, a), r)\}$
        if $|Sample(s,a)| \geq N_R(\mathcal{F}^R_{s,a}, \epsilon^i_s(p_s), \delta_1)$ **then**
          call Oracle: $f_{s,a} = \texttt{ERM}(\mathcal{F}^R_{s,a}, Sample(s,a), \ell)$
        else
          return `FAIL`
    else
      set $\forall a \in A : f_{s,a} = 0$
return $F = \{f_{s,a} : \ \forall s \in S, a \in A\}$

Algorithm 6 Exploit-CMDP for Known and Context-Free-Dynamics (EXPLOIT-KCFD)

inputs:

- The MDP parameters: S, A, P, s_0, H.
- The functions approximate the rewards for each state-action pair: $\{f_{s,a} | \forall (s,a) \in S \times A\}$.

for $t = 1, 2, \ldots$ **do**
  observe context c_t
  define the approximated reward function $\forall s \in S, a \in A : \widehat{r}^{c_t}(s,a) = f_{s,a}(c_t)$
  define the approximated CMDP $\widehat{\mathcal{M}}(c_t) = (S, A, P, s_0, \widehat{r}^{c_t}, H)$
  compute an optimal policy of the approximated model $(\pi_t, V_t) \leftarrow \texttt{Planning}(\widehat{\mathcal{M}}(c_t))$
  run π_t in episode t.

Remark 20 *Throughout the analysis, we strongly use that we collect samples only for β-reachable states, where $\beta = \frac{\epsilon}{6|S|}$.*

B.2.2 GOOD EVENTS

We analyse algorithm EXPLORE-KCFD (Algorithm 5) under the following good events:

Event G_1. Let G_1 be the event that for every $\frac{\epsilon}{6|S|}$-reachable state s and each action $a \in A$ we have $|Sample(s,a)| \geq N_R(\mathcal{F}^R_{s,a}, \epsilon^i_\star(p_s), \delta_1)$ (where $i \in \{1,2\}$, in accordance to the used loss function).

Lemma 21 *It holds that $\mathbb{P}[G_1] \geq 1 - \frac{1}{4}\delta$.*

Proof Fix a pair (s,a) of a $\frac{\epsilon}{6|S|}$-reachable state $s \in S$ and an action $a \in A$.

Assume we run π_s for T episodes, and in state s the agent always plays action a.

Let $\mathbb{I}_t[(s,a)]$ be an indicator which indicates whether (s,a) was sampled in the t'th episode. Then, $\mathbb{E}[\mathbb{I}_t[(s,a)]] = p_s \geq \frac{\epsilon}{|S|}$.

We wish to collect $m_{s,a} := N_R(\mathcal{F}^R_{s,a}, \epsilon^i_\star(p_s), \delta_1)$ samples. For T such that $Tp_s \geq m_{s,a}$ we would like to lower bound the number of episodes T needed to collect at least $m_{s,a}$ samples with probability at least $1-\delta_1$. For that mission, we use multiplicative Chernoff bound. Thus, we need to find $\beta \in [0,1]$ such that $(1-\beta)Tp_s = m_{s,a}$. $\beta = \frac{Tp_s - m_{s,a}}{Tp_s}$ is satisfying the requirement.

Hence,

$$\begin{aligned}
\mathbb{P}[\sum_{t=1}^{T} \mathbb{I}_t[(s,a)] \leq m_{s,a}] &= \mathbb{P}[\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}_t[(s,a)] \leq (1-\beta)Tp_s] \\
&\leq \exp\left(-\frac{\beta^2 Tp_s}{2}\right) \\
&= \exp\left(-\frac{(Tp_s - m_{s,a})^2}{2Tp_s}\right) \\
&\leq \delta_1 \iff T \geq \frac{2}{p_s}\left(\ln\frac{1}{\delta_1} + m_{s,a}\right).
\end{aligned}$$

Thus, for any $\frac{\epsilon}{6|S|}$-reachable state $s \in S$ and an action $a \in A$, if we run

$$T_{s,a} = \left\lceil \frac{2}{p_s}(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}^R_{s,a}, \epsilon^i_\star(p_s), \delta_1)) \right\rceil$$

iterations, we collect at least $N_R(\mathcal{F}^R_{s,a}, \epsilon^i_\star(p_s), \delta_1)$ examples. Since $\delta_1 = \frac{\delta}{4|S||A|}$, the lemma follows using union bound. ■

Event G_2. Let G_2 be the event where for every for any pair (s,a) of a $\frac{\epsilon}{6|S|}$-reachable state s and an action a, we have

$$\mathbb{E}_{c\sim\mathcal{D}}[(f_{s,a}(c) - r^c(s,a))^2] \leq \epsilon^2_\star(p_s) + \alpha^2_2(\mathcal{F}^R_{s,a}).$$

where $f_{s,a} = \texttt{ERM}(\mathcal{F}_{s,a}, Sample(s,a), \ell_2)$.

We similarly define the event G_2 for the ℓ_1 loss where

$$\mathbb{E}_{c\sim\mathcal{D}}[|f_{s,a}(c) - r^c(s,a)|] \leq \epsilon^1_\star(p_s) + \alpha_1(\mathcal{F}^R_{s,a}),$$

and $f_{s,a} = \texttt{ERM}(\mathcal{F}_{s,a}, Sample(s,a), \ell_1)$.

Lemma 22 *It holds that* $\mathbb{P}[G_2|G_1] \geq 1 - \frac{1}{4}\delta$.

Proof Follows immediately form ERM guarantees 3 for every pair (s,a) of a $\frac{\epsilon}{6|S|}$-reachable state $s \in S$ and an action $a \in A$, when combined using union bound over each pair (s,a). ■

Lemma 23 *It holds that* $\mathbb{P}[G_1 \cap G_2] \geq 1 - \frac{\delta}{2}$.

Proof By the results of Lemmas 21 and 22 when combined using a union bound. ■

Bellow we present analysis for both ℓ_1 and ℓ_2 losses.

B.2.3 ANALYSIS FOR ℓ_2 LOSS

Let $\alpha_2^2 := \max_{(s,a)\in S\times A} \alpha_2^2(\mathcal{F}_{s,a}^R)$.

The following lemma shows that under the good events G_1 and G_2, the value of any context-dependent policy with respect to the approximated model $\widehat{\mathcal{M}}$ is similar to that with respect to the true model $\mathcal{M}$, in expectation over the context.

Lemma 24 *Assume the events G_1 and G_2 hold. Then for any context-dependent policy $\pi = (\pi_c : S \to \Delta(A))_{c\in\mathcal{C}}$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V_{\mathcal{M}(c)}^{\pi_c}(s_0) - V_{\widehat{\mathcal{M}}(c)}^{\pi_c}(s_0)|] \leq \frac{1}{2}\epsilon + \alpha_2 H.$$

Proof Recall that for every context $c \in \mathcal{C}$, state $s \in S$ and action $a \in A$, the expected reward is $r^c(s,a) \in [0,1]$. By construction of $\widehat{\mathcal{M}}(c)$, for any state $s \in S$ which are not $\frac{\epsilon}{6|S|}$-reachable, we set $f_{s,a}(c) = 0$ for any action a. Hence,

$$|r^c(s,a) - f_{s,a}(c)| \leq 1.$$

Since the good event G_2 holds, for every state-action pair (s,a), such that state s is $\frac{\epsilon}{6|S|}$-reachable, it holds that

$$\epsilon_\star^2(p_s) + \alpha_2^2(\mathcal{F}_{s,a}^R) \underbrace{\geq}_{G_2} \mathbb{E}_{c\sim\mathcal{D}}[(f_{s,a}(c) - r^c(s,a))^2] \underbrace{\geq}_{\text{Jensen's inequality}} \mathbb{E}_{c\sim\mathcal{D}}^2[|f_{s,a}(c) - r^c(s,a)|].$$

Using that for all $a, b \in [0,\infty)$ it holds that $\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$, we obtain

$$\sqrt{\epsilon_\star^2(p_s)} + \alpha_2 \geq \sqrt{\epsilon_\star^2(p_s)} + \alpha_2(\mathcal{F}_{s,a}^R) \geq \sqrt{\epsilon_\star(p_s) + \alpha_2^2(\mathcal{F}_{s,a}^R)} \geq \mathbb{E}_{c\sim\mathcal{D}}[|f_{s,a}(c) - r^c(s,a)|]. \tag{1}$$

The above in particular implies that

$$\sqrt{\epsilon_\star^2(p_s)} \geq \mathbb{E}_{c\sim\mathcal{D}}[|f_{s,a}(c) - r^c(s,a)| - \alpha_2]. \tag{2}$$

Fix any context-dependent policy $\pi = (\pi_c : S \to \Delta(A))_{c\in\mathcal{C}}$. By definition of the value function we have **for any context** c:

$$V_{\mathcal{M}(c)}^{\pi_c}(s_0) = \mathbb{E}_{\pi_c,\mathcal{M}(c)}\Big[\sum_{h=0}^{H-1} r^c(s_h,a_h)|s_0\Big] = \sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)r^c(s,a)$$

and

$$V_{\widehat{\mathcal{M}}(c)}^{\pi_c}(s_0) = \mathbb{E}_{\pi_c,\widehat{\mathcal{M}}(c)}\Big[\sum_{h=0}^{H-1} \widehat{r}^c(s_h,a_h)|s_0\Big] = \sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)f_{s,a}(c).$$

By inequality 2, linearity of expectation and triangle inequality we obtain the following derivation:

$$
\begin{aligned}
&\mathbb{E}_{c\sim\mathcal{D}}\left[\left|V^{\pi_c}_{\mathcal{M}(c)}(s_0)-V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)\right|\right]\\
=&\mathbb{E}_{c\sim\mathcal{D}}\left[\left|\sum_{h=0}^{H-1}\sum_{s\in S_h}q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)r^c(s,a)-\sum_{h=0}^{H-1}\sum_{s\in S_h}q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)f_{s,a}(c)\right|\right]\\
\leq&\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{s\in S_h}q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)|r^c(s,a)-f_{s,a}(c)|\right]\\
=&\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{s\in S_h}q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)(|r^c(s,a)-f_{s,a}(c)|-\alpha_2+\alpha_2)\right]\\
=&\alpha_2 H+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{s\in S_h}q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)(|r^c(s,a)-f_{s,a}(c)|-\alpha_2)\right]\\
=&\alpha_2 H+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s<\frac{\epsilon}{6|S|}}q_h(s|\pi_c,P)\pi_c(a|s)(|r^c(s,a)-f_{s,a}(c)|-\alpha_2)\right]\\
&+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s>\frac{1}{|S|}}q_h(s|\pi_c,P)\pi_c(a|s)(|r^c(s,a)-f_{s,a}(c)|-\alpha_2)\right]\\
&+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s\in[\frac{\epsilon}{6|S|},\frac{1}{|S|}]}q_h(s|\pi_c,P)\pi_c(a|s)(|r^c(s,a)-f_{s,a}(c)|-\alpha_2)\right]\\
\leq&\alpha_2 H+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\pi_c(a|s)\sum_{s\in S_h:p_s<\frac{\epsilon}{6|S|}}q_h(s|\pi_c,P)\cdot 1\right]\\
&+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\pi_c(a|s)\sum_{s\in S_h:p_s>\frac{1}{|S|}}p_s(|r^c(s,a)-f_{s,a}(c)|-\alpha_2)\right]\\
&+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\pi_c(a|s)\sum_{s\in S_h:p_s\in[\frac{\epsilon}{6|S|},\frac{1}{|S|}]}p_s(|r^c(s,a)-f_{s,a}(c)|-\alpha_2)\right]\\
&\leq\alpha_2 H+\frac{\epsilon}{6}\\
&+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s>\frac{1}{|S|}}p_s(|r^c(s,a)-f_{s,a}(c)|-\alpha_2)\right]\\
&+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s\in[\frac{\epsilon}{6|S|},\frac{1}{|S|}]}p_s(|r^c(s,a)-f_{s,a}(c)|-\alpha_2)\right]\\
&=\alpha_2 H+\frac{\epsilon}{6}
\end{aligned}
$$

$$+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s>\frac{1}{|S|}} p_s\mathbb{E}_{c\sim\mathcal{D}}\left[(|r^c(s,a)-f_{s,a}(c)|-\alpha_2)\right]$$

$$+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s\in[\frac{\epsilon}{6|S|},\frac{1}{|S|}]} p_s\mathbb{E}_{c\sim\mathcal{D}}\left[(|r^c(s,a)-f_{s,a}(c)|-\alpha_2)\right]$$

$$\underbrace{\leq}_{\text{By ineq 2}} \alpha_2 H+\frac{\epsilon}{6}+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s>\frac{1}{|S|}} p_s\sqrt{\epsilon_\star^2(p_s)}+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s\in[\frac{\epsilon}{6|S|},\frac{1}{|S|}]} p_s\sqrt{\epsilon_\star^2(p_s)}$$

$$=\alpha_2 H+\frac{\epsilon}{6}+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s>\frac{1}{|S|}} p_s\frac{\epsilon}{6H|S||A|}+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s\in[\frac{\epsilon}{6|S|},\frac{1}{|S|}]} p_s\frac{\epsilon}{6p_s|S||A|}$$

$$=\alpha_2 H+\frac{\epsilon}{2}.$$

■

Theorem 25 *With probability at least* $1-\delta$ *it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0)-V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)]\leq\epsilon+2\alpha_2 H,$$

where $\pi^\star=(\pi_c^\star)_{c\in\mathcal{C}}$ *is the optimal policy for the true model* $\mathcal{M}(c)$*, and* $\widehat{\pi}^\star=(\widehat{\pi}_c^\star)_{c\in\mathcal{C}}$ *is the optimal policy for* $\widehat{\mathcal{M}}(c)$*.*

Proof Assume the good events G_1 and G_2 hold. By Lemma 24 we have for $\pi^\star=(\pi_c^\star)_{c\in\mathcal{C}}$ that

$$\left|\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0)-V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)]\right|\leq\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0)-V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)|]\leq\frac{1}{2}\epsilon+\alpha_2 H,$$

yielding,

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0)]-\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)]\leq\frac{1}{2}\epsilon+\alpha_2 H.$$

Similarly, we have for $\widehat{\pi}^\star=(\widehat{\pi}_c^\star)_{c\in\mathcal{C}}$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)]-\mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)]\leq\frac{1}{2}\epsilon+\alpha_2 H.$$

Also, for all $c\in\mathcal{C}$, since $\widehat{\pi}_c^\star$ is the optimal policy for $\widehat{\mathcal{M}}(c)$ we have $V^{\widehat{\pi}_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)\geq V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)$, which implies that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)]-\mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)]\leq 0.$$

Since by Lemma 23 we have that G_1 and G_2 hold with probability at least $1-\delta/2$, the theorem follows by summing the above three inequalities. ■

B.2.4 ANALYSIS FOR ℓ_1 LOSS

Let $\alpha_1:=\max_{(s,a)\in S\times A}\alpha_1(\mathcal{F}^R_{s,a})$. The following lemma shows that under the good events G_1 and G_2, the value of any context-dependent policy with respect to the approximated model $\widehat{\mathcal{M}}$ is similar to that with respect to the true model $\mathcal{M}$, in expectation over the context.

Lemma 26 *Assume the events G_1 and G_2 hold. Then for any context-dependent policy $\pi = (\pi_c : S \to \Delta(A))_{c\in\mathcal{C}}$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \frac{1}{2}\epsilon + \alpha_1 H.$$

Proof Recall that for every context $c \in \mathcal{C}$, state $s \in S$ and action $a \in A$, the expected reward is $r^c(s,a) \in [0,1]$. By construction of $\widehat{\mathcal{M}}(c)$, for any $s \in S$ which are not $\frac{\epsilon}{6|S|}$-reachable, we set $f_{s,a}(c) = 0$ for any action a. Hence,

$$|r^c(s,a) - f_{s,a}(c)| \leq 1.$$

Since the good event G_2 holds, for every state-action pair (s,a) such that s is $\frac{\epsilon}{6|S|}$-reachable it holds that

$$\epsilon^1_\star(p_s) + \alpha_1 \geq \epsilon^1_\star(p_s) + \alpha_1(\mathcal{F}^R_{s,a}) \geq \mathbb{E}_{c\sim\mathcal{D}}[|f_{s,a}(c) - r^c(s,a)|]. \tag{3}$$

The above implies that

$$\epsilon^1_\star(p_s) \geq \mathbb{E}_{c\sim\mathcal{D}}[|f_{s,a}(c) - r^c(s,a)| - \alpha_1]. \tag{4}$$

Fix any context-dependent policy $\pi = (\pi_c : S \to \Delta(A))_{c\in\mathcal{C}}$. By definition of the value function we have for any fixed context c:

$$V^{\pi_c}_{\mathcal{M}(c)}(s_0) = \mathbb{E}_{\pi_c,\mathcal{M}(c)}\Big[\sum_{h=0}^{H-1} r^c(s_h,a_h)|s_0 = s_0\Big] = \sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)r^c(s,a),$$

and

$$V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0) = \mathbb{E}_{\pi_c,\widehat{\mathcal{M}}(c)}\Big[\sum_{h=0}^{H-1} \widehat{r}^c(s_h,a_h)|s_0\Big] = \sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,P)\sum_{a\in A}\pi(a|s)f_{s,a}(c).$$

By inequality 4, linearity of expectation and triangle inequality we derive the following.

$$\begin{aligned}
&\mathbb{E}_{c\sim\mathcal{D}}\left[\left|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)\right|\right]\\
=&\mathbb{E}_{c\sim\mathcal{D}}\left[\left|\sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)r^c(s,a) - \sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)f_{s,a}(c)\right|\right]\\
\leq&\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)|r^c(s,a) - f_{s,a}(c)|\right]\\
=&\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)(|r^c(s,a) - f_{s,a}(c)| - \alpha_1 + \alpha_1)\right]
\end{aligned}$$

$$\begin{aligned}
=&\alpha_1 H + \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,P)\sum_{a\in A}\pi_c(a|s)(|r^c(s,a) - f_{s,a}(c)| - \alpha_1)\right]\\
=&\alpha_1 H + \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s<\frac{\epsilon}{6|S|}} q_h(s|\pi_c,P)\pi_c(a|s)(|r^c(s,a) - f_{s,a}(c)| - \alpha_1)\right]\\
&+ \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s>\frac{1}{|S|}} q_h(s|\pi_c,P)\pi_c(a|s)(|r^c(s,a) - f_{s,a}(c)| - \alpha_1)\right]
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s\in[\frac{\epsilon}{6|S|},\frac{1}{|S|}]} q_h(s|\pi_c,P)\pi_c(a|s)(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
\leq&\alpha_1 H+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\pi_c(a|s)\sum_{s\in S_h:p_s<\frac{\epsilon}{6|S|}} q_h(s|\pi_c,P)\cdot 1\right]\\
&+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\pi_c(a|s)\sum_{s\in S_h:p_s>\frac{1}{|S|}} p_s(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
&+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\pi_c(a|s)\sum_{s\in S_h:p_s\in[\frac{\epsilon}{6|S|},\frac{1}{|S|}]} p_s(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
\leq&\alpha_1 H+\frac{\epsilon}{6}\\
&+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s>\frac{1}{|S|}} p_s(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
&+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s\in[\frac{\epsilon}{6|S|},\frac{1}{|S|}]} p_s(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
=&\alpha_1 H+\frac{\epsilon}{6}\\
&+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s>\frac{1}{|S|}} p_s\mathbb{E}_{c\sim\mathcal{D}}\left[(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
&+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s\in[\frac{\epsilon}{6|S|},\frac{1}{|S|}]} p_s\mathbb{E}_{c\sim\mathcal{D}}\left[(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
\underbrace{\leq}_{\text{By ineq 4}}&\alpha_1 H+\frac{\epsilon}{6}+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s>\frac{1}{|S|}} p_s\epsilon^1_\star(p_s)+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s\in[\frac{\epsilon}{6|S|},\frac{1}{|S|}]} p_s\epsilon^1_\star(p_s)\\
=&\alpha_1 H+\frac{\epsilon}{6}+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s>\frac{1}{|S|}} p_s\frac{\epsilon}{6H|S||A|}+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:p_s\in[\frac{\epsilon}{6|S|},\frac{1}{|S|}]} p_s\frac{\epsilon}{6p_s|S||A|}\\
=&\alpha_1 H+\frac{\epsilon}{2}.
\end{aligned}
$$

■

Theorem 27 *With probability at least* $1-\delta$ *it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0)-V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)]\leq\epsilon+2\alpha_1 H,$$

where $\pi^\star=(\pi^\star_c)_{c\in\mathcal{C}}$ *is the optimal policy for the true model* $\mathcal{M}(c)$*, and* $\widehat{\pi}^\star=(\widehat{\pi}^\star_c)_{c\in\mathcal{C}}$ *is the optimal policy for* $\widehat{\mathcal{M}}(c)$*.*

Proof Assume the good events G_1 and G_2 hold. By Lemma 26 we have for $\pi^\star = (\pi_c^\star)_{c\in\mathcal{C}}$,that

$$\left|\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)]\right| \le \mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)|] \le \frac{1}{2}\epsilon + \alpha_1 H,$$

yielding,

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)] \le \frac{1}{2}\epsilon + \alpha_1 H.$$

Similarly, we have for $\widehat{\pi}^\star = (\widehat{\pi}_c^\star)_{c\in\mathcal{C}}$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \le \frac{1}{2}\epsilon + \alpha_1 H.$$

Also, for all $c \in \mathcal{C}$, since $\widehat{\pi}_c^\star$ is the optimal policy for $\widehat{\mathcal{M}}(c)$ we have $V^{\widehat{\pi}_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0) \ge V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)$, which implies that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)] \le 0.$$

Since by Lemma 23 we have that G_1 and G_2 hold with probability at least $1 - \delta/2$, the theorem follows by summing the above three inequalities. ■

B.3 Sample Complexity Bounds

Given standard sample complexity bounds for learning a function class using ERM, we can bound the required sample complexity of our Algorithm EXPLORE-KCFD. The following theorems state the sample complexity bounds.

Theorem 28 (Adaption of Theorem 19.2 in Anthony et al. (1999)) *Let $\mathcal{F}$ be a hypothesis space of real valued functions with a finite pseudo dimension, denoted $Pdim(\mathcal{F}) < \infty$. Then, $\mathcal{F}$ has a uniform convergence with*

$$m(\epsilon,\delta) = O\Big(\frac{1}{\epsilon^2}(Pdim(\mathcal{F})\ln\frac{1}{\epsilon} + \ln\frac{1}{\delta})\Big).$$

Theorem 29 (Adaption of Theorem 19.1 in Anthony et al. (1999)) *Let $\mathcal{F}$ be a hypothesis space of real valued functions with a finite fat-shattering dimension, denoted $fat_{\mathcal{F}}$. Then, $\mathcal{F}$ has a uniform convergence with*

$$m(\epsilon,\delta) = O\Big(\frac{1}{\epsilon^2}(fat_{\mathcal{F}}(\epsilon/256)\ln^2\frac{1}{\epsilon} + \ln\frac{1}{\delta})\Big).$$

B.3.1 Sample bounds for the ℓ_2 loss

We prove sample complexity bound of our algorithm for function classes with finite Pseudo dimension when using ℓ_2 loss.

Corollary 30 *Assume that for every $(s,a) \in S\times A$ we have that $Pdim(\mathcal{F}^R_{s,a}) < \infty$. Let $Pdim = \max_{(s,a)\in S\times A} Pdim(\mathcal{F}^R_{s,a})$. Then, after collecting*

$$O\Big(\frac{|S|^2|A|}{\epsilon}\ln\frac{|S||A|}{\delta} + \frac{H^4|S|^6|A|^5\,(Pdim\ln\frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta})}{\epsilon^4}\Big)$$

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

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

Proof Recall that for each state-action pair (s,a) such that s is $\frac{\epsilon}{6|S|}$-reachable, we run for $T_{s,a} = \lceil \frac{2}{p_s}(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}^R_{s,a}, \epsilon^2_\star(p_s), \delta_1))\rceil$ episodes. By Theorem 25, for $\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\geq\epsilon/6|S|}\sum_{a\in A} T_{s,a}$ samples we have with probability at least $1-\delta$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha_2 H.$$

Since for every $(s,a)\in S\times A$ we have that $Pdim(\mathcal{F}^R_{s,a}) < \infty$, and $Pdim = \max_{(s,a)\in S\times A} Pdim(\mathcal{F}^R_{s,a})$, by Theorem 28, for every $(s,a)\in S\times A$ we have

$$N_R(\mathcal{F}^R_{s,a}, \epsilon^2_\star(p_s), \delta_1) = O\Big(\frac{Pdim\ln\frac{1}{\epsilon^2_\star(p_s)} + \ln\frac{1}{\delta_1}}{\epsilon^4_\star(p_s)}\Big).$$

Using the accuracy-per-state function, we derive the overall sample complexity bound in the following computation.

$$\begin{aligned}
&\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\geq\epsilon/6|S|}\sum_{a\in A} T_{s,a} = \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\geq\epsilon/6|S|}\sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}^R_{s,a}, \epsilon^2_\star(p_s), \delta_1))\Big) \\
&= \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\geq\epsilon/6|S|}\sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{Pdim\ln\frac{1}{\epsilon^2_\star(p_s)} + \ln\frac{1}{\delta_1}}{\epsilon^4_\star(p_s)})\Big) \\
&= \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\in[\epsilon/6|S|,1/|S|]}\sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{Pdim\ln\frac{1}{\epsilon^2/36p_s^2|S|^2|A|^2} + \ln\frac{1}{\delta_1}}{\epsilon^4/36^2p_s^4|S|^4|A|^4})\Big) \\
&\quad + \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s>1/|S|}\sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{Pdim\ln\frac{1}{\epsilon^2/36H^2|S|^2|A|^2} + \ln\frac{1}{\delta_1}}{\epsilon^4/36^2H^4|S|^4|A|^4})\Big) \\
&= \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\in[\epsilon/6|S|,1/|S|]}\sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{p_s^4|S|^4|A|^4(Pdim\ln\frac{p_s^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{1}{\delta_1})}{\epsilon^4})\Big) \\
&\quad + \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s>1/|S|}\sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{H^4|S|^4|A|^4(Pdim\ln\frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{1}{\delta_1})}{\epsilon^4})\Big) \\
&\leq \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\in[\epsilon/6|S|,1/|S|]}\sum_{a\in A} O\Big(\frac{|S|}{\epsilon}\ln\frac{|S||A|}{\delta} + \frac{p_s^3|S|^4|A|^4(Pdim\ln\frac{|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta})}{\epsilon^4}\Big) \\
&\quad + \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s>1/|S|}\sum_{a\in A} O\Big(|S|\ln\frac{|S||A|}{\delta} + \frac{H^4|S|^5|A|^4(Pdim\ln\frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta})}{\epsilon^4}\Big) \\
&\underbrace{\leq}_{(\star)} \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\in[\epsilon/6|S|,1/|S|]}\sum_{a\in A} O\Big(\frac{|S|}{\epsilon}\ln\frac{|S||A|}{\delta} + \frac{|S||A|^4(Pdim\ln\frac{|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta})}{\epsilon^4}\Big) \\
&\quad + \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s>1/|S|}\sum_{a\in A} O\Big(|S|\ln\frac{|S||A|}{\delta} + \frac{H^4|S|^5|A|^4(Pdim\ln\frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta})}{\epsilon^4}\Big) \\
&= O\Big(\frac{|S|^2|A|}{\epsilon}\ln\frac{|S||A|}{\delta} + \frac{H^4|S|^6|A|^5\ (Pdim\ln\frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta})}{\epsilon^4}\Big)
\end{aligned}$$

Where $(\star)$ is since in that regime we have $p_s\in[\epsilon/6|S|, 1/|S|]$, hence $p_s^3\leq 1/|S|^3$. ■

We also show similar sample complexity for function classes with finite fat-shattering dimension when using ℓ_2 loss.

Remark 31 *The sample complexity for function classes with finite fat-shattering dimension with ℓ_2 loss, where in $Fdim$ below we also maximizes over $\epsilon_\star^2(p_s)$ and the maximum is bounded and independent of p_s.*

Corollary 32 *Assume that for every $(s,a) \in S \times A$ we have that $\mathcal{F}_{s,a}^R$ has finite fat-shattering dimention. Let $Fdim = \max_{(s,a)\in S\times A} fat_{\mathcal{F}_{s,a}^R}(\epsilon_\star^2(p_s)/256)$. Then, after collecting*

$$O\Big(\frac{|S|^2|A|}{\epsilon}\ln\frac{|S||A|}{\delta} + \frac{H^4|S|^6|A|^5\ (Fdim\ln^2\frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta})}{\epsilon^4}\Big)$$

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

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

Proof Recall that for each state-action pair (s,a) such that s is $\frac{\epsilon}{6|S|}$-reachable, we run for $T_{s,a} = \lceil \frac{2}{p_s}(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}_{s,a}^R, \epsilon_\star^2(p_s), \delta_1))\rceil$ episodes. By Theorem 25, for $\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\ge\epsilon/6|S|}\sum_{a\in A} T_{s,a}$ samples we have with probability at least $1-\delta$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

Since for every $(s,a)\in S\times A$ we have that $\mathcal{F}_{s,a}^R$ has finite fat-shattering dimension, and $Fdim = \max_{(s,a)\in S\times A} fat_{\mathcal{F}_{s,a}^R}(\epsilon_\star^2(p_s))$, by Theorem 29, for every $(s,a)\in S\times A$ we have

$$N_R(\mathcal{F}_{s,a}^R, \epsilon_\star^2(p_s), \delta_1) = O\Big(\frac{Fdim\ln^2\frac{1}{\epsilon_\star^2(p_s)} + \ln\frac{1}{\delta_1}}{\epsilon_\star^4(p_s)}\Big).$$

Using the accuracy-per-state function, we derive the overall sample complexity bound in the following computation.

$$\begin{aligned}
&\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\ge\epsilon/6|S|}\sum_{a\in A} T_{s,a}\\
&= \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\ge\epsilon/6|S|}\sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}_{s,a}^R, \epsilon_\star^2(p_s), \delta_1))\Big)\\
&= \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\ge\epsilon/6|S|}\sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{Fdim\ln^2\frac{1}{\epsilon_\star^2(p_s)} + \ln\frac{1}{\delta_1}}{\epsilon_\star^4(p_s)})\Big)\\
&= \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\in[\epsilon/6|S|,1/|S|]}\sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{Fdim\ln^2\frac{1}{\epsilon^2/36p_s^2|S|^2|A|^2} + \ln\frac{1}{\delta_1}}{\epsilon^4/36^2p_s^4|S|^4|A|^4})\Big)\\
&\quad + \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s>1/|S|}\sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{Fdim\ln^2\frac{1}{\epsilon^2/36H^2|S|^2|A|^2} + \ln\frac{1}{\delta_1}}{\epsilon^4/36^2H^4|S|^4|A|^4})\Big)\\
&= \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\in[\epsilon/6|S|,1/|S|]}\sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{p_s^4|S|^4|A|^4(Fdim\ln^2\frac{p_s^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{1}{\delta_1})}{\epsilon^4})\Big)\\
&\quad + \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s>1/|S|}\sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{H^4|S|^4|A|^4(Fdim\ln^2\frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{1}{\delta_1})}{\epsilon^4})\Big)\\
&\le \sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\in[\epsilon/6|S|,1/|S|]}\sum_{a\in A} O\Big(\frac{|S|}{\epsilon}\ln\frac{|S||A|}{\delta} + \frac{p_s^3|S|^4|A|^4(Fdim\ln^2\frac{|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta})}{\epsilon^4}\Big)
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s>1/|S|}\sum_{a\in A}O\Big(|S|\ln\frac{|S||A|}{\delta}+\frac{H^4|S|^5|A|^4(Fdim\ln^2\frac{H^2|S|^2|A|^2}{\epsilon^2}+\ln\frac{|S||A|}{\delta})}{\epsilon^4}\Big)\\
&\underbrace{\leq}_{(\star)}\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\in[\epsilon/6|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{|S|}{\epsilon}\ln\frac{|S||A|}{\delta}+\frac{|S||A|^4(Fdim\ln^2\frac{|S|^2|A|^2}{\epsilon^2}+\ln\frac{|S||A|}{\delta})}{\epsilon^4}\Big)\\
&+\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s>1/|S|}\sum_{a\in A}O\Big(|S|\ln\frac{|S||A|}{\delta}+\frac{H^4|S|^5|A|^4(Fdim\ln^2\frac{H^2|S|^2|A|^2}{\epsilon^2}+\ln\frac{|S||A|}{\delta})}{\epsilon^4}\Big)\\
&=O\Big(\frac{|S|^2|A|}{\epsilon}\ln\frac{|S||A|}{\delta}+\frac{H^4|S|^6|A|^5\ (Fdim\ln^2\frac{H^2|S|^2|A|^2}{\epsilon^2}+\ln\frac{|S||A|}{\delta})}{\epsilon^4}\Big)
\end{aligned}
$$

where $(\star)$ is since in that regime we have $p_s\in[\epsilon/6|S|,1/|S|]$, hence $p_s^3\leq 1/|S|^3$. ■

B.3.2 SAMPLE BOUNDS FOR THE ℓ_1 LOSS

We bound the sample complexity for function classes with finite Pseudo dimension with ℓ_1 loss.

Corollary 33 *Assume that for every $(s,a)\in S\times A$ we have that $Pdim(\mathcal{F}^R_{s,a})<\infty$. Let $Pdim=\max_{(s,a)\in S\times A}Pdim(\mathcal{F}^R_{s,a})$. Then, after collecting*

$$O\Big(\frac{|S|^2|A|}{\epsilon}\ln\frac{|S||A|}{\delta}+\frac{H^2|S|^4|A|^3\ (Pdim\ln\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|}{\delta})}{\epsilon^2}\Big)$$

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

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0)-V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)]\leq\epsilon+2\alpha_1 H.$$

Proof Recall that for each state-action pair (s,a) such that s is $\frac{\epsilon}{6|S|}$-reachable, we run for $T_{s,a}=\lceil\frac{2}{p_s}(\ln(\frac{1}{\delta_1})+N_R(\mathcal{F}^R_{s,a},\epsilon^1_\star(p_s),\delta_1))\rceil$ episodes. By Theorem 27, for $\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\geq\epsilon/6|S|}\sum_{a\in A}T_{s,a}$ samples we have with probability at least $1-\delta$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0)-V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)]\leq\epsilon+2\alpha_1 H.$$

Since for every $(s,a)\in S\times A$ we have that $Pdim(\mathcal{F}^R_{s,a})<\infty$, and $Pdim=\max_{(s,a)\in S\times A}Pdim(\mathcal{F}^R_{s,a})$, by Theorem 28, for every $(s,a)\in S\times A$ we have

$$N_R(\mathcal{F}^R_{s,a},\epsilon^1_\star(p_s),\delta_1)=O\Big(\frac{Pdim\ln\frac{1}{\epsilon^1_\star(p_s)}+\ln\frac{1}{\delta_1}}{\epsilon^2_\star(p_s)}\Big).$$

Using the accuracy-per-state function, we derive the overall sample complexity bound in the following computation.

$$
\begin{aligned}
&\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\geq\epsilon/6|S|}\sum_{a\in A}T_{s,a}\\
&=\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\geq\epsilon/6|S|}\sum_{a\in A}O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1})+N_R(\mathcal{F}^R_{s,a},\epsilon^1_\star(p_s),\delta_1))\Big)\\
&=\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\geq\epsilon/6|S|}\sum_{a\in A}O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1})+\frac{Pdim\ln\frac{1}{\epsilon^1_\star(p_s)}+\ln\frac{1}{\delta_1}}{\epsilon^2_\star(p_s)})\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{h=0}^{H-1} \sum_{s\in S_h: p_s \in [\epsilon/6|S|, 1/|S|]} \sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{Pdim \ln \frac{1}{\epsilon/6p_s|S||A|} + \ln\frac{1}{\delta_1}}{\epsilon^2/36p_s^2|S|^2|A|^2})\Big) \\
&\quad + \sum_{h=0}^{H-1} \sum_{s\in S_h: p_s > 1/|S|} \sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{Pdim \ln \frac{1}{\epsilon/6H|S||A|} + \ln\frac{1}{\delta_1}}{\epsilon^2/36H^2|S|^2|A|^2})\Big) \\
&= \sum_{h=0}^{H-1} \sum_{s\in S_h: p_s \in [\epsilon/6|S|, 1/|S|]} \sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{p_s^2|S|^2|A|^2(Pdim \ln \frac{p_s|S||A|}{\epsilon} + \ln\frac{1}{\delta_1})}{\epsilon^2})\Big) \\
&\quad + \sum_{h=0}^{H-1} \sum_{s\in S_h: p_s > 1/|S|} \sum_{a\in A} O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1}) + \frac{H^2|S|^2|A|^2(Pdim \ln \frac{H|S||A|}{\epsilon} + \ln\frac{1}{\delta_1})}{\epsilon^2})\Big) \\
&\leq \sum_{h=0}^{H-1} \sum_{s\in S_h: p_s \in [\epsilon/6|S|, 1/|S|]} \sum_{a\in A} O\Big(\frac{|S|}{\epsilon}\ln\frac{|S||A|}{\delta} + \frac{p_s|S|^2|A|^2(Pdim \ln \frac{|S||A|}{\epsilon} + \ln\frac{|S||A|}{\delta})}{\epsilon^2}\Big) \\
&\quad + \sum_{h=0}^{H-1} \sum_{s\in S_h: p_s > 1/|S|} \sum_{a\in A} O\Big(|S|\ln\frac{|S||A|}{\delta} + \frac{H^2|S|^3|A|^2(Pdim \ln \frac{H|S||A|}{\epsilon} + \ln\frac{|S||A|}{\delta})}{\epsilon^2}\Big) \\
&\leq \sum_{h=0}^{H-1} \sum_{s\in S_h: p_s \in [\epsilon/6|S|, 1/|S|]} \sum_{a\in A} O\Big(\frac{|S|}{\epsilon}\ln\frac{|S||A|}{\delta} + \frac{|S||A|^2(Pdim \ln \frac{|S||A|}{\epsilon} + \ln\frac{|S||A|}{\delta})}{\epsilon^2}\Big) \\
&\quad + \sum_{h=0}^{H-1} \sum_{s\in S_h: p_s > 1/|S|} \sum_{a\in A} O\Big(|S|\ln\frac{|S||A|}{\delta} + \frac{H^2|S|^3|A|^2(Pdim \ln \frac{H|S||A|}{\epsilon} + \ln\frac{|S||A|}{\delta})}{\epsilon^2}\Big) \\
&= O\Big(\frac{|S|^2|A|}{\epsilon}\ln\frac{|S||A|}{\delta} + \frac{H^2|S|^4|A|^3\ (Pdim \ln \frac{H|S||A|}{\epsilon} + \ln\frac{|S||A|}{\delta})}{\epsilon^2}\Big)
\end{aligned}
$$

where in the first inequality we used the upper bound on $1/p_s$ and in the second inequality we upper bounded p_s. ■

We also show similar sample complexity for function classes with finite fat-shattering dimension when using ℓ_1 loss.

Remark 34 *The sample complexity for function classes with finite fat-shattering dimension with ℓ_1 loss, where in $Fdim$ below we also maximizes over $\epsilon_\star(p_s)$ and the maximum is bounded and independent of p_s.*

Corollary 35 *Assume that for every $(s,a) \in S \times A$ we have that $\mathcal{F}^R_{s,a}$ has finite fat-shattering dimention. Let $Fdim = \max_{(s,a)\in S\times A} fat_{\mathcal{F}^R_{s,a}}(\epsilon_\star(p_s)/256)$. Then, for*

$$
O\Big(\frac{|S|^2|A|H^2}{\epsilon^2}\ (Fdim \ln^2\frac{\max\{|S|,H\}}{\epsilon} + \ln\frac{|S||A|}{\delta})\Big)
$$

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

$$
\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha_1 H.
$$

Proof Recall that for each state-action pair (s,a) such that s is $\frac{\epsilon}{6|S|}$ reachable, we run for $T_{s,a} = \lceil \frac{2}{p_s}(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}^R_{s,a}, \epsilon^1_\star(p_s), \delta_1))\rceil$ episodes. By Theorem 25, for $\sum_{h=0}^{H-1}\sum_{s\in S_h: p_s \geq \epsilon/6|S|}\sum_{a\in A} T_{s,a}$ samples we have with probability at least $1-\delta$ that

$$
\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)} - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha_1 H.
$$

Since for every $(s,a) \in S\times A$ we have that $\mathcal{F}^R_{s,a}$ has finite fat-shattering dimension, and $Fdim = \max_{(s,a)\in S\times A} fat_{\mathcal{F}^R_{s,a}}(\epsilon_\star(p_s))$, by Theorem 29, for every $(s,a) \in S \times A$ we have

$$
N_R(\mathcal{F}^R_{s,a}, \epsilon^2_\star(p_s), \delta_1) = O\Big(\frac{Fdim \ln^2 \frac{1}{\epsilon^1_\star(p_s)} + \ln\frac{1}{\delta_1}}{\epsilon^2_\star(p_s)}\Big).
$$

Using the accuracy-per-state function, we derive the overall sample complexity bound in the following computation.

$$
\begin{aligned}
&\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\geq\epsilon/6|S|}\sum_{a\in A}T_{s,a}=\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\geq\epsilon/6|S|}\sum_{a\in A}O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1})+N_R(\mathcal{F}^R_{s,a},\epsilon^1_\star(p_s),\delta_1))\Big)\\
&=\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\geq\epsilon/6|S|}\sum_{a\in A}O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1})+\frac{Fdim\ln^2\frac{1}{\epsilon^1_\star(p_s)}+\ln\frac{1}{\delta_1}}{\epsilon^2_\star(p_s)})\Big)\\
&=\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\in[\epsilon/6|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1})+\frac{Fdim\ln^2\frac{1}{\epsilon/6p_s|S||A|}+\ln\frac{1}{\delta_1}}{\epsilon^2/36p_s^2|S|^2|A|^2})\Big)\\
&\quad+\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s>1/|S|}\sum_{a\in A}O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1})+\frac{Fdim\ln^2\frac{1}{\epsilon/6H|S||A|}+\ln\frac{1}{\delta_1}}{\epsilon^2/36H^2|S|^2|A|^2})\Big)\\
&=\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\in[\epsilon/6|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1})+\frac{p_s^2|S|^2|A|^2(Fdim\ln^2\frac{p_s|S||A|}{\epsilon}+\ln\frac{1}{\delta_1})}{\epsilon^2})\Big)\\
&\quad+\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s>1/|S|}\sum_{a\in A}O\Big(\frac{1}{p_s}(\ln(\frac{1}{\delta_1})+\frac{H^2|S|^2|A|^2(Fdim\ln^2\frac{H|S||A|}{\epsilon}+\ln\frac{1}{\delta_1})}{\epsilon^2})\Big)\\
&\leq\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\in[\epsilon/6|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{|S|}{\epsilon}\ln\frac{|S||A|}{\delta}+\frac{p_s|S|^2|A|^2(Fdim\ln^2\frac{|S||A|}{\epsilon}+\ln\frac{|S||A|}{\delta})}{\epsilon^2}\Big)\\
&\quad+\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s>1/|S|}\sum_{a\in A}O\Big(|S|\ln\frac{|S||A|}{\delta}+\frac{H^2|S|^3|A|^2(Fdim\ln^2\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|}{\delta})}{\epsilon^2}\Big)\\
&\leq\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s\in[\epsilon/6|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{|S|}{\epsilon}\ln\frac{|S||A|}{\delta}+\frac{|S||A|^2(Fdim\ln^2\frac{|S||A|}{\epsilon}+\ln\frac{|S||A|}{\delta})}{\epsilon^2}\Big)\\
&\quad+\sum_{h=0}^{H-1}\sum_{s\in S_h:p_s>1/|S|}\sum_{a\in A}O\Big(|S|\ln\frac{|S||A|}{\delta}+\frac{H^2|S|^3|A|^2(Fdim\ln^2\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|}{\delta})}{\epsilon^2}\Big)\\
&=O\Big(\frac{|S|^2|A|}{\epsilon}\ln\frac{|S||A|}{\delta}+\frac{H^2|S|^4|A|^3\ (Fdim\ln^2\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|}{\delta})}{\epsilon^2}\Big)
\end{aligned}
$$

where in the first inequality we used the upper bound on $1/p_s$ and in the second inequality we upper bounded p_s. ■

Appendix C. Unknown and Context Free Dynamics.

When the dynamics is unknown, we have an additional hurdle which is the need to approximate it. To collect i.i.d samples efficiently for each state, we still need to find an exploration policy which (approximately) maximizes the probability to visit the target state.

We can compute approximate the true dynamics P by $\widehat{P}$, and use $\widehat{P}$ to compute a policy $\widehat{\pi}_s$ to reach state s. If $\widehat{P}\approx P$ this will result in a similar sample size to approximate the rewards, as in the known dynamics case. We will have a worse sample complexity due to the need to approximate the dynamics well.

C.1 Basic Lemmas

In this subsection, we present basic concentration-bounds based lemmas that used to compute the required sample complexity to obtain good dynamics approximation, with high probability.

Lemma 36 *Let $\delta' \in (0,1)$, layer $h \in [H]$ and a state $s_h \in S_h$. Let $\pi : S \to \Delta(A)$ be a policy that satisfies $q_h(s_h|\pi, P) \geq \beta$, for $\beta \in (0,1]$. Let m be the desired number of visits in s_h. Then, if running π for $T \geq \frac{2}{\beta}(\ln \frac{1}{\delta'} + m)$ episodes, the agent will visit state s_h at least m times, with probability at least $1 - \delta'$.*

Proof Follows form multiplicative Chernoff bound. ■

Lemma 37 *Let $\gamma, \delta_1 \in (0,1)$, $h \in [H-1]$ and $(s_h, a_h) \in S_h \times A$. For every $s_{h+1} \in S_{h+1}$, denote by $n(s_{h+1}|s_h, a_h)$ the number of times the agent observed a trajectory contains the triplet (s_h, a_h, s_{h+1}), out of m trajectories that contain the pair (s_h, a_h).*

Define for every $s_{h+1} \in S_{h+1}$: $\widehat{P}(s_{h+1}|s_h, a_h) = \frac{n(s_{h+1}|s_h, a_h)}{N_P(\gamma, \delta_1)}$.

Then, for $N_P(\gamma, \delta_1) \geq \frac{2}{\gamma^2}\Big(\ln\Big(\frac{1}{\delta_1} + (|S|+1)\ln 2\Big)\Big)$ we have with probability at least $1 - \delta_1$

$$\|P(\cdot|s_h, a_h) - \widehat{P}(\cdot|s_h, a_h)\|_1 \leq \gamma.$$

Proof By Bretagnolle Huber-Carol inequality we have

$$\begin{aligned}
\mathbb{P}[\|P(\cdot|s_h, a_h) - \widehat{P}(\cdot|s_h, a_h)\|_1 \geq \gamma] &= \mathbb{P}[\sum_{s_{h+1} \in S_{h+1}} |\frac{n(s_{h+1}|s_h, a_h)}{N_P(\gamma, \delta_1)} - P(s_{h+1}|s_h, a_h)|] \\
&\leq 2^{|S_{h+1}|+1} \exp\left(-\frac{N_P(\gamma, \delta_1)\gamma^2}{2}\right) \\
&\leq 2^{|S|+1} \exp\left(-\frac{N_P(\gamma, \delta_1)\gamma^2}{2}\right) \\
&\leq \delta_1 \iff N_P(\gamma, \delta_1) \geq \frac{2}{\gamma^2}\Big(\ln\frac{1}{\delta_1} + (|S|+1)\ln 2\Big).
\end{aligned}$$

■

Corollary 38 (Sample complexity for approximating the dynamics) *Let $\gamma = \frac{\epsilon}{48|S|H^2}$ and $\delta_1 = \frac{\delta}{6|S||A|H}$. Then*

$$N_P(\gamma, \delta_1) = O\Big(\frac{H^4|S|^2}{\epsilon^2}\Big(\ln\frac{|S||A|H}{\delta} + |S|\Big)\Big).$$

Lemma 39 (Dynamics distance bound implies occupancy measure distance bound) *Let $h \in [H]$ and fix a policy $\pi : S \to \Delta(A)$. Assume that for all $k < h$ and $(s_k, a_k) \in S_k \times A$ we have*

$$\|P(\cdot|s_k, a_k) - \widetilde{P}(\cdot|s_k, a_k)\|_1 \leq \gamma.$$

Then,

$$\|q_h(\cdot|\pi, P) - q_h(\cdot|\pi, \widetilde{P})\|_1 \leq \gamma h.$$

Proof We prove the lemma using induction on the horizon h. The base case is $h = 0$. As there exists unique start state s_0 the claim holds trivially.

We assume correctness for all $i < h$ and show for $i = h$. By definition we have

$$
\begin{aligned}
&\|q_h(\cdot|\pi, P) - q_h(\cdot|\pi, \widetilde{P})\|_1 = \\
&= \sum_{s_h \in S_h} |q_h(s_h|\pi, P) - q_h(s_h|\pi, \widetilde{P})| \\
&= \sum_{s_{h-1} \in S_{h-1}} \sum_{a_{h-1} \in A} \sum_{s_h \in S_h} \pi(a_{h-1}|s_{h-1}) |q_{h-1}(s_{h-1}|\pi, P) P(s_h|s_{h-1}, a_{h-1}) - q_{h-1}(s_{h-1}|\pi, \widetilde{P}) \widetilde{P}(s_h|s_{h-1}, a_{h-1})| \\
&\leq \sum_{s_{h-1} \in S_{h-1}} \sum_{a_{h-1} \in A} \pi(a_{h-1}|s_{h-1}) P(s_h|s_{h-1}, a_{h-1}) \sum_{s_h \in S_h} |q_{h-1}(s_{h-1}|\pi, P) - q_{h-1}(s_{h-1}|\pi, \widetilde{P})| \\
&\quad + \sum_{s_{h-1} \in S_{h-1}} q_{h-1}(s_{h-1}|\pi, \widetilde{P}) \sum_{a_{h-1} \in A} \pi(a_{h-1}|s_{h-1}) \sum_{s_h \in S_h} |P(s_h|s_{h-1}, a_{h-1}) - \widetilde{P}(s_h|s_{h-1}, a_{h-1})| \\
&\leq \|q_{h-1}(\cdot|\pi, P) - q_{h-1}(\cdot|\pi, \widetilde{P})\|_1 \\
&\quad + \sum_{s_{h-1} \in S_{h-1}} q_{h-1}(s_{h-1}|\pi, \widetilde{P}) \sum_{a_{h-1} \in A} \pi(a_{h-1}|s_{h-1}) \|P(\cdot|s_{h-1}, a_{h-1}) - \widetilde{P}(\cdot|s_{h-1}, a_{h-1})\|_1 \\
&\leq \gamma(h-1) + \gamma = \gamma h
\end{aligned}
$$

■

C.2 Algorithm

We start by an overview of our algorithms.

Algorithm EXPLORE-UCFD (Algorithm 7) works in phases, where in phase $h \in [H]$ we approximate layer h dynamics. We first collect samples for each (non-negligible) state in layer h and then use them to approximate the dynamics, using simple tabular estimation. Using the same sample we also estimate the rewards using ERM oracle. The required accuracy for each state-action pair (s_h, a_h) is determined by the accuracy-per-state function $\epsilon^i_\star(\cdot)$ (for $i \in \{1, 2\}$, with accordance to the used loss function) using $\widehat{p}_{s_h} := q_h(s_h|\widehat{p}_{s_h}, \widehat{P})$. We use $\epsilon^1_\star(p_s)$ for the ℓ_1 loss, which defines as

$$
\epsilon^1_\star(p_s) := \epsilon_\star(p_s) = \begin{cases} 1 & \text{, if } p_s < \frac{\epsilon}{24|S|} \\ \frac{\epsilon}{24H|S||A|} & \text{, if } p_s > \frac{1}{|S|} \\ \frac{\epsilon}{24p_s|S||A|} & \text{, if } p_s \in [\frac{\epsilon}{24|S|}, \frac{1}{|S|}] \end{cases}
$$

while for the ℓ_2 loss we use $\epsilon^2_\star(p_s) = (\epsilon_\star(p_s))^2$.

After collecting sufficient number of samples for every non-negligible state in layers up to $h-1$, we have a good approximation of the dynamics up to layer $h-1$. This yields a good approximation of the occupancy measure of layer h for any policy π (regardless of it being context-dependent or not) .

Hence, given a state $s_h \in S_h$ and approximate dynamics $\widehat{P}$ we compute $\widehat{\pi}_{s_h} = \arg\max_{\pi: S \to A} q_h(s_h|\pi, \widehat{P})$. We run $\widehat{\pi}_{s_h}$ to generate the sample of s_h. Since the dynamics is context-free, the policy is the same for all of the contexts.

In order to control the number episodes sampled we define non-negligible states as β-reachable w.r.t $\widehat{P}$, i.e., they have $q_h(s_h|\widehat{\pi}_{s_h}, \widehat{P}) \geq \beta$.

At the end of the sampling we have for each β-reachable state $s_h \in S_h$ with respect to $\widehat{P}$, and every action a_h a data set contains tuples of the form (s_h, a_h, s_{h+1}) to approximate the transition probability matrix via tabular mean estimation. (Recall that here the dynamics do no depend on the context, this will change in the context-dependent dynamics case.) For the rewards, we use tuples of the form $((c, s_h, a_h), r_h)$ and run the ERM to approximate the rewards.

Algorithm EXPLOIT-UCFD (Algorithm 8) get as inputs the MDP parameters, the dynamics approximation $\widehat{P}$ and the functions approximate the rewards (that computed using EXPLORE-UCFD algorithm). Given a context c it computed the approximated MDP $\widehat{\mathcal{M}}(c)$ and use it to compute a near optimal policy $\pi^\star_c$. Then, it run $\pi^\star_c$ to generate trajectory.

Recall that $\widehat{\mathcal{M}}(c) = (S, A, \widehat{P}, s_0, \widehat{r}^c, H)$ where we define $\forall s \in S, a \in A : \widehat{r}^c(s,a) = f_{s,a}(c)$, and $\widehat{P}$ is computed using tabular approximation.

Algorithm 7 Explore Unknown and Context-Free Dynamics CMDP(EXPLORE-UCFD)

inputs:

- MDP parameters: $S = \{S_0, S_1, \ldots, S_H\}$ - a layered states space, A - a finite actions space, s_0 - a unique start state, H - the horizon length.
- Accuracy and confidence parameters: ϵ,δ.
- $\forall s \in S, a \in A : \ \mathcal{F}^R_{s,a}$ - the function classes use to approximate the rewards function.
- $N_R(\mathcal{F}, \epsilon, \delta)$ - sample complexity function for the ERM oracle.
- γ - the required approximation error of the dynamics , β - the reachability parameter. (We have $\frac{\epsilon}{24|S|} \geq \beta \geq 2\gamma H$.)
- $N_P(\gamma, \delta_1)$ - sample complexity function for approximating the dynamics using tabular approximation.
- ℓ - loss function (assumed to be one of ℓ_1 or ℓ_2) and the appropriate accuracy-per-state function $\epsilon^i_\star$ (for $i \in \{1,2\}$ in accordance to ℓ).

set $\delta_1 = \frac{\delta}{6|S||A|H}$
set for all $(s,a) \in S \times A$: $n(s,a) = 0$ and for all $(s,a,s') \in S \times A \times S$: $n(s'|s,a) = 0$.
for $h \in [H-1]$ **do**
    let $s_{sink} \notin S$ be a new state which denotes a sink.
define the approximated dynamics for all $(s,a,s') \in S \times A \times S$: $\widehat{P}(s'|s,a) = \frac{n(s'|s,a)}{n(s,a)} \mathbb{I}[n(s,a) \geq N_P(\gamma, \delta_1)]$ and $\widehat{P}(s_{sink}|s,a) = \mathbb{I}[n(s,a) < N_P(\gamma, \delta_1)]$.
    for $s_h \in S_h$ **do**
        $(\widehat{\pi}_{s_h}, \widehat{p}_{s_h}) \leftarrow \texttt{FFP}(S \cup \{s_{sink}\}, A, \widehat{P}, s_0, H, s_h)$. $\triangleright$ $\widehat{p}_{s_h}$ is the highest probability to visit s_h under $\widehat{P}$ and $\widehat{\pi}_{s_h}$ is the policy that reach that probability.
        if $\widehat{p}_{s_h} \geq \beta$ **then**
            for $a_h \in A$ **do**
                compute $T_{s_h,a_h} = \lceil \frac{2}{\widehat{p}_{s_h} - \gamma h} (\ln(\frac{1}{\delta_1}) + \max\{N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon^i_\star(\widehat{p}_{s_h}), \delta_1), N_P(\gamma, \delta_1)\}) \rceil$
                initialize $Sample(s_h, a_h) = \emptyset$
                set $\widehat{\pi}_{s_h}(s_h) \leftarrow a_h$
                for $t = 1, 2, \ldots, T_{s_h,a_h}$ **do**
                    observe context c.
                    run $\widehat{\pi}_{s_h}$ to generate trajectory τ_t
                    if $(s_h, a_h, r_h, s_{h+1}) \in \tau_t$ for some $r_h \in [0,1]$ and $s_{h+1} \in S_{h+1}$ **then**
                        update sample: $Sample(s_h, a_h) = Sample(s_h, a_h) + \{((c, s_h, a_h), r_h)\}$
                        update counters: $n(s_h, a_h) \leftarrow n(s_h, a_h) + 1, \ \ n(s_{h+1}|s_h, a_h) \leftarrow n(s_{h+1}|s_h, a_h) + 1$
                if $|Sample(s_h, a_h)| \geq \max\{N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon_\star(\widehat{p}_{s_h}), \delta_1), N_P(\gamma, \delta_1)\}$ **then**
                    call Oracle: $f_{s_h,a_h} = \texttt{ERM}(\mathcal{F}^R_{s_h,a_h}, Sample(s_h, a_h), \ell)$
                else
                    return `FAIL`
        else
            set: $\forall a \in A : f_{s_h,a} = 0$
return $F = \{f_{s,a} : \ \forall s \in S, a \in A\}, \widehat{P}$

C.3 Analysis

C.3.1 ANALYSIS OUTLINE

We provide analysis for both ℓ_1 and ℓ_2 loss functions, in the agnostic case.

Algorithm 8 Exploit CMDP for Unknown Context-Free-Dynamics (EXPLOIT-UCFD)

inputs:

- The MDP parameters: $S = \{S_0, S_1, \ldots, S_H\}$,$A$,$s_0$,$H$.
- $\widehat{P}$ approximation of the context-free dynamics.
- $\{f_{s,a} | \forall (s,a) \in S \times A\}$ - the function use to approximate the reward for each state-action pair (as function of the context).

for $t = 1, 2, \ldots$ **do**
  observe context c_t
  define $\forall s \in S, a \in A : \widehat{r}^{c_t}(s,a) = f_{s,a}(c_t), \forall a \in A : \widehat{r}^{c_t}(s_{sink}, a) = 0$
  define the approximated MDP associated with c_t: $\widehat{\mathcal{M}}(c_t) = (S \cup \{s_{sink}\}, A, \widehat{P}, s_0, \widehat{r}^{c_t}, H)$
  $(\pi_t, V_t) \leftarrow \texttt{Planning}(\widehat{\mathcal{M}}(c_t))$
  run π_t in episode t.

In the analysis, we bound the error caused by the dynamics approximation (see Sub-subsection C.3.4) and the error caused by the rewards approximation for both the ℓ_2 loss (see Sub-subsection C.3.5) and the ℓ_1 loss (see Sub-subsection C.3.6).

We combine both errors to bound the expected value difference between the true model $\mathcal{M}(c)$ and $\widehat{\mathcal{M}}(c)$, for any context-dependent policy $\pi = (\pi_c)_{c \in \mathcal{C}}$, with high probability. (See Lemma 54 for the ℓ_2 loss and Lemma 58 for the ℓ_1).

Using that bound, we derive a bound on the expected value difference between the optimal context-dependent policy $\pi^\star = (\pi_c^\star)_{c \in \mathcal{C}}$ and our approximated optimal policy $\widehat{\pi}^\star = (\widehat{\pi}_c^\star)_{c \in \mathcal{C}}$, which holds with high probability. (See Theorem 25 and 27). .

Lastly, we derive sample complexity bound using known uniform convergence sample complexity bounds for the Pseudo dimension (See Theorem 28) and the fat-shattering dimension (See Theorem 29). For the sample complexity analysis, see Sub-subsection C.4.1 for the ℓ_2 loss, and C.4.2 for the ℓ_1 loss.

In the following analysis, we assume that $\frac{\epsilon}{24|S|} \geq \beta \geq 2\gamma H$, and later choose β and γ that satisfies that. We also choose $\delta_1 = \frac{\delta}{6|S||A|H}$. In addition, we use the following notation.

Definition 40 *For every state $s \in S$ we denote by $\widehat{p}_s$ the maximal (over the policies) probability to visit s under the approximated dynamics $\widehat{P}$.*

C.3.2 GOOD EVENTS

We analyze algorithms EXPLOR-UCFD (Algorithm 7) and EXPLOIT-UCFD (Algorithm 8) under the following good events, which we show that hold with high probability.

Event G_1. For every $h \in [H-1]$ let G_1^h denote the good event in which we have for every state and action pair $(s_h, a_h) \in S_h \times A$ such that s_h is β-reachable for $\widehat{P}$ that $n(s_h, a_h) \geq \max\{N_R(\mathcal{F}_{s,a}^R, \epsilon_\star^2(\widehat{p}_{s_h}), \delta_1), N_P(\gamma, \delta_1)\}$ for the ℓ_2 loss ($n(s_h, a_h) \geq \max\{N_R(\mathcal{F}_{s,a}^R, \epsilon_\star^1(\widehat{p}_{s_h}), \delta_1), N_P(\gamma, \delta_1)\}$ for the ℓ_1 loss) samples of were collected. We define $G_1 = \cap_{h \in [H-1]} G_1^h$.

Event G_2. For every $h \in [H-1]$ let G_2^h denote the good event in which we have for every state and action pair $(s_h, a_h) \in S_h \times A$ such that s_h is β-reachable for $\widehat{P}$ that $\|P(\cdot|s_h, a_h) - \widehat{P}(\cdot|s_h, a_h)\|_1 \leq \gamma$. (Here, we omit the entry $\widehat{P}(s_{sink}|s,a)$ of $\widehat{P}$). We define $G_2 = \cap_{h \in [H-1]} G_2^h$.

How are we about to use G_1 and G_2? Intuitively, given that G_1 holds, we have collected sufficient number of samples for each β-reachable state s and every action a. Hence, by $\widehat{P}$ definition we have that $\widehat{P}(s'|s,a) =$

$n(s,a,s')/n(s,a) \;\; \forall s' \neq s_{sink}$ and $\widehat{P}(s_{sink}|s,a) = 0$. Thus, we can ignore the entry related with the sink (which does not exist in the true dynamics P), and have that $\|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1 \leq \gamma$ with high probability, by Lemma 37.

Event G_3. For every $h \in [H-1]$ let G_3^h denote the good event in which we have for every state and action pair $(s_h, a_h) \in S_h \times A$ such that s_h is β-reachable for $\widehat{P}$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[(f_{s_h,a_h}(c) - r^c(s_h,a_h))^2] \leq \epsilon_\star^2(\widehat{p}_{s_h}) + \alpha_2^2(\mathcal{F}^R_{s_h,a_h})$$

for the ℓ_2 loss ($\mathbb{E}_{c\sim\mathcal{D}}[|f_{s_h,a_h}(c) - r^c(s_h,a_h)|] \leq \epsilon_\star(\widehat{p}_{s_h}) + \alpha_1(\mathcal{F}^R_{s_h,a_h})$ for the ℓ_1 loss). We define $G_3 = \cap_{h\in[H-1]} G_3^h$.

C.3.3 PROVING THE GOOD EVENTS HOLD WITH HIGH PROBABILITY

Lemma 41 $\mathbb{P}[G_3|G_1] \geq 1 - \frac{\delta}{6}$.

Proof By the ERM guarantees for each state-action pair when combined using union bound. ■

Lemma 42 (occupancy measure lower bound) *Let $h \in [H]$ and a policy $\pi : S \to \Delta(A)$. Under the good events G_1^k and G_2^k for every $k < h$, for every state $s_h \in S_h$ it holds that*

$$q_h(s_h|\pi, P) \geq q_h(s_h|\pi, \widehat{P}) - \gamma h.$$

Proof Define the dynamics $\widetilde{P}$ for all $k < h$ and $(s,a,s') \in S_k \times A \times S_{k+1}$ as follows:

$$\widetilde{P}(s'|s,a) = P(s'|s,a) \cdot \mathbb{I}[n(s,a) \geq N_P(\gamma,\delta_1)],$$

and

$$\widetilde{P}(s_{sink}|s,a) = \mathbb{I}[n(s,a) < N_P(\gamma,\delta_1)].$$

So, under the good events G_1^k and G_2^k for every $k < h$ and every $(s_k, a_k) \in S_k \times A_k$ we have that

$$\|\widetilde{P}(\cdot|s_k,a_k) - \widehat{P}(\cdot|s_k,a_k)\|_1 \leq \gamma.$$

(For states s_k which are β-reachable, it follows since G_1^k and G_2^k hold. For states s_k which are not β-reachable we have that $\widetilde{P}$ and $\widehat{P}$ are identical, i.e., they both transition to the sink with probability 1).

Hence, by Lemma 39 we have that $\|q_h(\cdot|\pi,\widetilde{P}) - q_h(\cdot|\pi,\widehat{P})\|_1 \leq \gamma h$, which implies that for all $s_h \in S_h$ we have

$$q_h(s_h|\pi,\widetilde{P}) \geq q_h(s_h|\pi,\widehat{P}) - \gamma h.$$

By $\widetilde{P}$ definition, we trivially have for all $h \in [H]$ and $s_h \in S_h$ that $q_h(s_h|\pi,P) \geq q_h(s_h|\pi,\widetilde{P})$. Hence, we obtained

$$q_h(s_h|\pi,P) \geq q_h(s_h|\pi,\widetilde{P}) \geq q_h(s_h|\pi,\widehat{P}) - \gamma h.$$

■

Lemma 43 *For every layer $h \in [H-1]$ it holds that $\mathbb{P}[G_2^h|G_1^h] \geq 1 - \frac{\delta}{6H}$.*

Proof Fix a layer $h \in [H]$. Since G_1^h holds, we have for every state-action pair $(s_h,a_h) \in S_h \times A$ such that s_h is β-reachable for $\widehat{P}$ that $n(s_h,a_h) \geq N_P(\gamma,\delta_1)$. Hence, by Lemma 37, for $N_P(\gamma,\delta_1) = O\Big(\frac{1}{\gamma^2}\Big(\ln\frac{1}{\delta_1} + |S|\Big)\Big)$ we have with probability at least $1-\delta_1$ that $\|P(\cdot|s_h,a_h) - \widehat{P}(\cdot|s_h,a_h)\|_1 \leq \gamma$. Since $\delta_1 = \frac{\delta}{6|S||A|H}$, using a union bound over all the pairs $(s_h,a_h) \in S_h \times A$ such that s_h is β-reachable for $\widehat{P}$ we obtain the lemma. ■

Lemma 44 *For every layer $h \in [H-1]$ it holds that $\mathbb{P}[G_1^h | G_1^k, G_2^k \ \forall k \in [h-1]] \geq 1 - \frac{\delta}{6H}$.*

Proof We prove the lemma using induction over the horizon h.

Base case: $h = 0$. As for state s_0 we collect at least $\max\{N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon^i_\star(\widehat{p}_s), \delta_1, N_P(\gamma, \delta_1\}$ samples in a deterministic manner, therefore we have $\mathbb{P}[G_1^0] = 1$.

Induction step: Assume the lemma holds for all $k \leq h$ and we show it holds for $h+1$. Given $G_1^k, G_2^k \ \forall k \in [h]$ hold, by Lemma 42 for every state $s_{h+1} \in S_{h+1}$ it holds that

$$q_{h+1}(s_{h+1}|\widehat{\pi}_{s_{h+1}}, P) \geq q_{h+1}(s_{h+1}|\widehat{\pi}_{s_{h+1}}, \widehat{P}) - \gamma(h+1) = \widehat{p}_{s_{h+1}} - \gamma(h+1).$$

Recall that for every action $a_{h+1} \in A$, the agent runs $\widehat{\pi}_{s_{h+1}}$ for $T_{s_{h+1},a_{h+1}}$ episodes, in which, when visiting s_{h+1} the agent plays action a_{h+1} (for $T_{s_{h+1},a_{h+1}}$ which defined in Algorithm 7).

Hence, by by Lemma 36, the agent collects at least $\max\{N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon^i_\star(\widehat{p}_s), \frac{\delta}{6|S||A|H}), N_P(\gamma, \frac{\delta}{6|S||A|H})\}$ examples of (s_{h+1}, a_{h+1}), with probability at lest $1 - \delta_1 = 1 - \frac{\delta}{6|S||A|H}$.

Using union bound over each pair $(s_{h+1}, a_{h+1}) \in S_{h+1} \times A$ such that s_{h+1} is β-reachable for $\widehat{P}$, we obtain that $\mathbb{P}[G_1^{h+1} | G_1^k, G_2^k \ \forall i \in [h]] \geq 1 - \delta_1|S||A| = 1 - \frac{\delta}{6H}$, which proves the induction step. ■

Lemma 45 $\mathbb{P}[G_1 \cap G_2] \geq 1 - \delta/3$.

Proof Recall that $G_1 = \cap_{h\in[H-1]} G_1^h$ and $G_2 = \cap_{h\in[H-1]} G_2^h$.

Let X be a random variable with support $[H-1]$ such that

$$X = \min_{k\in[H-1]} \{\overline{G}_1^k \cup \overline{G}_2^k \text{ holds }\}.$$

Meaning, X is the layer with the lowest index in which at least one of the good events G_1^h or G_2^h does not hold. If G_1^h and G_2^h hold for every layer $h \in [H-1]$ then $X = \perp$. By definition of X and Bayes rule (i.e., for two events A, B: $\mathbb{P}[A \cap B] = \mathbb{P}[A|B] \cdot \mathbb{P}[B]$) the following holds.

$$\begin{aligned}
\forall h \in [H-1]. \quad \mathbb{P}[X = h] &= \mathbb{P}[(\overline{G}_1^h \cup \overline{G}_2^h) \cap (\cap_{k\in[h-1]} G_1^k \cap G_2^k)] \\
&= \mathbb{P}[(\overline{G}_1^h \cup \overline{G}_2^h) | \cap_{k\in[h-1]} (G_1^k \cap G_2^k)] \cdot \underbrace{\mathbb{P}[\cap_{k\in[h-1]} (G_1^k \cap G_2^k)]}_{\leq 1} && \text{(By Bayes rule)} \\
&\leq \mathbb{P}[(\overline{G}_1^h \cup \overline{G}_2^h) | \cap_{k\in[h-1]} (G_1^k \cap G_2^k)] \\
&= \underbrace{\mathbb{P}[\overline{G}_1^h | \cap_{k\in[h-1]} (G_1^k \cap G_2^k)]}_{\leq \frac{\delta}{6H} \text{ by Lemma 44}} + \mathbb{P}[\overline{G}_2^h \cap G_1^h | \cap_{k\in[h-1]} (G_1^k \cap G_2^k)] \\
& && \text{(By union of disjoint events)} \\
&\leq \frac{\delta}{6H} + \mathbb{P}[\overline{G}_2^h \cap G_1^h | \cap_{k\in[h-1]} (G_1^k \cap G_2^k)] \\
&= \frac{\delta}{6H} + \mathbb{P}[\overline{G}_2^h | G_1^h, \cap_{k\in[h-1]} (G_1^k \cap G_2^k)] \cdot \underbrace{\mathbb{P}[G_1^h | \cap_{k\in[h-1]} (G_1^k \cap G_2^k)]]}_{\leq 1} \\
& && \text{(By Bayes rule)} \\
&\leq \frac{\delta}{6H} + \mathbb{P}[\overline{G}_2^h | G_1^h, \cap_{k\in[h-1]} (G_1^k \cap G_2^k)] \\
&= \frac{\delta}{6H} + \underbrace{\mathbb{P}[\overline{G}_2^h | G_1^h]}_{\leq \frac{\delta}{6H} \text{ by Lemma 43}} && (\overline{G}_2^h \text{ depended only on } G_1^h)
\end{aligned}$$

$$\leq 2\frac{\delta}{6H} = \frac{\delta}{3H}.$$

Lastly, by G_1 and G_2 definition we have

$$\begin{aligned}
\mathbb{P}[G_1 \cap G_2] &= 1 - \mathbb{P}[\overline{G}_1 \cup \overline{G}_2] \\
&= 1 - \mathbb{P}[\cup_{h\in[H-1]}(\overline{G}_1^h \cup \overline{G}_2^h)] \\
&= 1 - \mathbb{P}[\exists h \in [H-1].(\overline{G}_1^h \cup \overline{G}_2^h)] \\
&= 1 - \mathbb{P}[\exists h \in [H-1].X = h] \\
&= 1 - \mathbb{P}[\cup_{h\in[H-1]}\{X = h\}] \\
&\underbrace{\geq}_{\text{Union Bound}} 1 - \sum_{h\in[H-1]} \mathbb{P}[X = h] \\
&\geq 1 - \frac{\delta}{3}.
\end{aligned}$$

■

Lemma 46 *It holds that* $\mathbb{P}[G_1 \cap G_2 \cap G_3] \geq 1 - \delta/2$.

Proof By lemmas 45 and 41 when combined with a union bound. ■

C.3.4 BOUNDING THE ERROR CAUSED BY THE DYNAMICS APPROXIMATION

In the following, we consider an intermediate model $\widetilde{M}$, which defined as follows.

For any context $c \in \mathcal{C}$, we define $\widetilde{M}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}, r^c, s_0, H)$, where we extend the true rewards function r^c for the sink by defining $r^c(s_{sink}, a) := 0, \ \ \forall c \in \mathcal{C}, \ a \in A$, and $\widehat{P}$ is the approximated dynamics.

Recall the true MDP associated with the context c is $\mathcal{M}(c) = (S, A, P, r^c, s_0, H)$.

In the following lemma we bound the occupancy-measures differences under P and $\widehat{P}$, for every policy π under the good events.

Lemma 47 *Assume the good events G_1 and G_2 hold. Then, for every policy $\pi : s \to \Delta(A)$ and layer $h \in [H]$ it holds that*

$$\|q_h(\cdot|\pi, P) - q_h(\cdot|\pi, \widehat{P})\|_1 \leq \gamma h + \beta \sum_{k=0}^{h-1} |S_k|,$$

where

$$\forall h \in [H]. \ \ \|q_h(\cdot|\pi, P) - q_h(\cdot|\pi, \widehat{P})\|_1 := \sum_{s_h \in S_h} |q_h(s_h|\pi, P) - q_h(s_h|\pi, \widehat{P})|$$

(i.e., $q_h(s_{sink}|\pi, \widehat{P})$ is omitted, for all $h \in [H]$).

Remark 48 *Since $s_{sink} \notin S$, $q_h(s_{sink}|\pi, P)$ is not defined for the true dynamics P. In addition, by $\widehat{P}$ definition, from the sink there are no transitions to any other state and has zero reward. Hence, we can simply ignore it in the following analysis.*

We now prove Lemma 47.

Proof We show the lemma by induction on h.

The base case $h = 0$ holds trivially since there is a unique start state s_0. Hence $q_0(s_0|\pi, P) = q_0(s_0|\pi, \widehat{P}) = 1$.

For the induction step, we assume correctness for all $k < h$ and show for h.

For every $k \leq h$ we define $S_k^\beta = \{s_k \in S_k : s_k \text{ is } \beta\text{-reachable for } \widehat{P}\}$.

Since the good events G_1 and G_2 hold, we have for every $(s_k, a_k) \in S_k^\beta \times A$ that

$$\|P(\cdot|s_k, a_k) - \widehat{P}(\cdot|s_k, a_k)\|_1 \leq \gamma.$$

We remark that by definition $\widehat{P}(s_{sink}|s, a) = \mathbb{I}[n(s, a) < N_P(\gamma, \delta_1)] = 0$, under the good events and P is not defined for s_{sink}, hence we can ignore it when analysing the dynamics total variation distance for the β-reachable states.

Using the induction hypothesis we obtain,

$$\begin{aligned}
&\|q_h(\cdot|\pi, P) - q_h(\cdot|\pi, \widehat{P})\|_1 \\
=&\sum_{s_h \in S_h} |q_h(s_h|\pi, P) - q_h(s_h|\pi, \widehat{P})| \\
=&\sum_{s_{h-1} \in S_{h-1}} \sum_{a_{h-1} \in A} \sum_{s_h \in S_h} \pi(a_{h-1}|s_{h-1}) |q_{h-1}(s_{h-1}|\pi, P)P(s_h|s_{h-1}, a_{h-1}) - q_{h-1}(s_{h-1}|\pi, \widehat{P})\widehat{P}(s_h|s_{h-1}, a_{h-1})| \\
\leq&\sum_{s_{h-1} \in S_{h-1}} \sum_{a_{h-1} \in A} \pi(a_{h-1}|s_{h-1})P(s_h|s_{h-1}, a_{h-1}) \sum_{s_h \in S_h} |q_{h-1}(s_{h-1}|\pi, P) - q_{h-1}(s_{h-1}|\pi, \widehat{P})| \\
&+ \sum_{s_{h-1} \in S_{h-1}} q_{h-1}(s_{h-1}|\pi, \widehat{P}) \sum_{a_{h-1} \in A} \pi(a_{h-1}|s_{h-1}) \sum_{s_h \in S_h} |P(s_h|s_{h-1}, a_{h-1}) - \widehat{P}(s_h|s_{h-1}, a_{h-1})| \\
\leq&\|q_{h-1}(\cdot|\pi, P) - q_{h-1}(\cdot|\pi, \widehat{P})\|_1 \\
&+ \sum_{s_{h-1} \in S_{h-1}} q_{h-1}(s_{h-1}|\pi, \widehat{P}) \sum_{a_{h-1} \in A} \pi(a_{h-1}|s_{h-1}) \sum_{s_h \in S_h} |P(s_h|s_{h-1}, a_{h-1}) - \widehat{P}(s_h|s_{h-1}, a_{h-1})| \\
\leq&\gamma(h-1) + \beta \sum_{k=0}^{h-2} |S_k| + \sum_{s_{h-1} \in S_{h-1}^\beta} q_{h-1}(s_{h-1}|\pi, \widehat{P}) \sum_{a_{h-1} \in A} \pi(a_{h-1}|s_{h-1}) \underbrace{\|P(\cdot|s_{h-1}, a_{h-1}) - \widehat{P}(\cdot|s_{h-1}, a_{h-1})\|_1}_{\leq \gamma} \\
&+ \sum_{s_{h-1} \notin S_{h-1}^\beta} \underbrace{q_{h-1}(s_{h-1}|\pi, \widehat{P})}_{\leq \beta} \underbrace{\sum_{a_{h-1} \in A} \pi(a_{h-1}|s_{h-1})}_{=1} \underbrace{\sum_{s_h \in S_h} |P(s_h|s_{h-1}, a_{h-1}) - \widehat{P}(s_h|s_{h-1}, a_{h-1})|}_{\leq 1} \\
\leq&\gamma h + \beta \sum_{k=0}^{h-1} |S_k|.
\end{aligned}$$

■

Remark 49 *For $\beta = \frac{\epsilon}{24|S|H}$ and $\gamma = \frac{\epsilon}{48|S|H^2}$ we have $\beta - \gamma H \geq \frac{\epsilon}{48|S|H}$.*

Corollary 50 *Under the good events G_1 and G_2, for $\beta = \frac{\epsilon}{24|S|H}$ and $\gamma = \frac{\epsilon}{48|S|H^2}$ we have for all $h \in [H]$ that $\|q_h(\cdot|\pi, P) - q_h(\cdot|\pi, \widehat{P})\|_1 \leq \frac{3\epsilon}{48H} = \frac{\epsilon}{16H}$.*

Lemma 51 *Assume the good events G_1 and G_2 hold.*

Then, for the parameters choice of $\beta = \frac{\epsilon}{24|S|H}$ and $\gamma = \frac{\epsilon}{48|S|H^2}$,for every context-dependent policy $\pi = (\pi_c)_{c \in \mathcal{C}}$ it holds that,

$$|V_{\mathcal{M}(c)}^{\pi_c}(s_0) - V_{\widehat{M}(c)}^{\pi_c}(s_0)| \leq \frac{\epsilon}{16}.$$

Proof Recall that the true rewards function is not defined for s_{sink}, since $s_{sink} \notin S$. A natural extension of r^c to s_{sink} is by defining $\forall c \in \mathcal{C},\ \forall a \in A.\ \ r^c(s_{sink}, a) = 0$. Since P is also not defined for s_{sink}, we can simply ignore s_{sink}, as the second equality in the following calculation shows.

Fix a context $c \in \mathcal{C}$ and a context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$. Consider the following derivation.

$$
\begin{aligned}
&|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{M}(c)}(s_0)| \\
=&|\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,P)\cdot r^c(s_h,a_h) - \sum_{h=0}^{H-1}\sum_{s_h\in S_h\cup\{s_{sink}\}}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,\widehat{P})\cdot r^c(s_h,a_h)| \\
&= |\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,P)\cdot r^c(s_h,a_h) - \sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,\widehat{P})\cdot r^c(s_h,a_h)| \\
&\qquad\qquad \text{(Since we defined } r^c(s_{sink},a) := 0,\ \forall c, a) \\
&= |\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} (q_h(s_h,a_h|\pi_c,P) - q_h(s_h,a_h|\pi_c,\widehat{P}))r^c(s_h,a_h)| \\
&\leq \sum_{h=0}^{H}\sum_{s_h\in S_h}\sum_{a_h\in A} \pi(a_h|s_h)|r^c(s_h,a_h)||q_h(s_h|\pi_c,P) - q_h(s_h|\pi_c,\widehat{P})| \\
&\leq \sum_{h=0}^{H}\sum_{s_h\in S_h} |q_h(s_h|\pi_c,P) - q_h(s_h|\pi_c,\widehat{P})| \qquad (r^c \text{ is bounded in } [0,1], \text{ and } \textstyle\sum_{a\in A}\pi_c(a|s) = 1) \\
&\leq \frac{3\epsilon}{48H}H = \frac{\epsilon}{16}, \qquad \text{(By corollary 50)}
\end{aligned}
$$

which yields the lemma. ∎

Corollary 52 *Assume the good events G_1 and G_2 hold.*

Then, for the parameters choice $\beta = \frac{\epsilon}{24|S|H}$ and $\gamma = \frac{\epsilon}{48|S|H^2}$ for every context-dependent policy $\pi = (\pi_c)_{c\in \mathrm{C}}$ it holds that

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{M}(c)}(s_0)|] \leq \frac{\epsilon}{16}.$$

Proof Implied by taking expectation over both sided of the inequality stated in Lemma 51. ∎

C.3.5 BOUNDING THE ERROR CAUSED BY THE REWARDS APPROXIMATION FOR THE ℓ_2 LOSS.

In this sub-subsection, we bound the error caused by the rewards approximation, by bounding the expected value difference between the intermediate model $\widetilde{\mathcal{M}}$ and the approximated model $\widehat{\mathcal{M}}$. Here, we analyse the error for the ℓ_2 loss.

Recall the definition of $\widetilde{\mathcal{M}}$. For every context $c \in \mathcal{C}$ we define $\widetilde{\mathcal{M}}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}, r^c, s_0, H)$, where we extend the true rewards function r^c for the sink by defining $r^c(s_{sink}, a) := 0, \quad \forall c \in \mathcal{C},\ \ a \in A$, and $\widehat{P}$ is the approximated dynamics.

Also, recall the approximated MDP for the context c, $\widehat{\mathcal{M}}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}, \widehat{r}^c, s_0, H)$ which defined in Algorithm 8.

In the following analysis, Let $S^\beta(\widehat{P})$ be the set of β-reachable states for the dynamics $\widehat{P}$, and $\alpha_2^2 = \max_{(s,a\in S^\beta(\widehat{P})\times A)} \alpha_2^2(\mathcal{F}^R_{s,a})$ be the maximal agnostic approximation error.

Lemma 53 *Assume the good event G_3 holds.*

Then, for any context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$ it holds that

$$\mathbb{E}_{c\sim\mathcal{D}}\left[\left|V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)\right|\right] \leq \frac{\epsilon}{8} + \alpha_2 H.$$

Proof By construction of $\widehat{\mathcal{M}}(c)$, for any $s \notin S^{\beta}(\widehat{P})$, i.e., states which are not β-reachable for $\widehat{P}$, we set $f_{s,a}(c) = 0$ for any action a. Hence,

$$|r^c(s,a) - f_{s,a}(c)| \leq 1.$$

Since the good event G_3 holds, we have for every $(s,a) \in S \times A$ such that s is β-reachable for $\widehat{P}$ that

$$\epsilon^2_\star(\widehat{p}_s) + \alpha_2^2(\mathcal{F}^R_{s,a}) \underbrace{\geq}_{G_3} \mathbb{E}_{c\sim\mathcal{D}}[(f_{s,a}(c) - r^c(s,a))^2] \underbrace{\geq}_{\text{Jensen's inequality}} \mathbb{E}^2_{c\sim\mathcal{D}}[|f_{s,a}(c) - r^c(s,a)|].$$

Using that for all $a, b \in [0,\infty)$ it holds that $\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$, we obtain

$$\sqrt{\epsilon^2_\star(\widehat{p}_s)} + \alpha_2 \geq \sqrt{\epsilon^2_\star(\widehat{p}_s)} + \alpha_2(\mathcal{F}^R_{s,a}) \geq \sqrt{\epsilon^2_\star(\widehat{p}_s) + \alpha_2^2(\mathcal{F}^R_{s,a})} \geq \mathbb{E}_{c\sim\mathcal{D}}[|f_{s,a}(c) - r^c(s,a)|]. \tag{5}$$

The above implies that

$$\sqrt{\epsilon^2_\star(\widehat{p}_s)} \geq \mathbb{E}_{c\sim\mathcal{D}}[|f_{s,a}(c) - r^c(s,a)| - \alpha_2]. \tag{6}$$

Fix a context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$.

By definition we have:

$$\begin{aligned}
V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0) &= \sum_{h=0}^{H-1} \sum_{s\in S\cup\{s_{sink}\}} q_h(s|\pi_c,\widehat{P}) \sum_{a\in A} \pi_c(a|s) r^c(s,a) \\
&= \sum_{h=0}^{H-1} \sum_{s\in S} q_h(s|\pi_c,\widehat{P}) \sum_{a\in A} \pi_c(a|s) r^c(s,a) && (r^c(s_{sink},a) := 0, \ \ \forall c\in\mathcal{C}, a\in A.) \\
&= \sum_{h=0}^{H-1} \sum_{s\in S_h} q_h(s|\pi_c,\widehat{P}) \sum_{a\in A} \pi_c(a|s) r^c(s,a). && \text{(Since the MDP is layered and loop-free)}
\end{aligned}$$

Similarly,

$$\begin{aligned}
V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0) &= \sum_{h=0}^{H-1} \sum_{s\in S\cup\{s_{sink}\}} q_h(s|\pi_c,\widehat{P}) \sum_{a\in A} \pi_c(a|s) \widehat{r}^c(s,a) \\
&= \sum_{h=0}^{H-1} \sum_{s\in S} q_h(s|\pi_c,\widehat{P}) \sum_{a\in A} \pi_c(a|s) \widehat{r}^c(s,a) && (\widehat{r}^c(s_{sink},a) := 0, \ \ \forall c\in\mathcal{C}, a\in A.) \\
&= \sum_{h=0}^{H-1} \sum_{s\in S_h} q_h(s|\pi_c,\widehat{P}) \sum_{a\in A} \pi_c(a|s) \widehat{r}^c(s,a). && \text{(Since the MDP is layered and loop-free)} \\
&= \sum_{h=0}^{H-1} \sum_{s\in S_h} q_h(s|\pi_c,\widehat{P}) \sum_{a\in A} \pi_c(a|s) f_{s,a}(c). && \text{(By definition, } \widehat{r}^c(s,a) := f_{s,a}(c))
\end{aligned}$$

Recall that $\beta \leq \frac{\epsilon}{24|S|}$. Thus, if s is not β-reachable for $\widehat{P}$, then $\widehat{p}_s < \beta \leq \frac{\epsilon}{24|S|}$. Moreover, if $\widehat{p}_s \geq \frac{\epsilon}{24|S|}$ then s is β-reachable for $\widehat{P}$, and the good event G_3 guarantee is hold for s. Thus, when combining all the above, by linearity of expectation and triangle inequality we obtain,

$$\mathbb{E}_{c\sim\mathcal{D}}\left[\left|V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)\right|\right]$$

$$
\begin{aligned}
=&\mathbb{E}_{c\sim\mathcal{D}}\left[\left|\sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,\widehat{P})\sum_{a\in A}\pi_c(a|s)r^c(s,a) - \sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,\widehat{P})\sum_{a\in A}\pi_c(a|s)f_{s,a}(c)\right|\right]\\
\leq&\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,\widehat{P})\sum_{a\in A}\pi_c(a|s)|r^c(s,a) - f_{s,a}(c)|\right]\\
=&\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,\widehat{P})\sum_{a\in A}\pi_c(a|s)(|r^c(s,a) - f_{s,a}(c)| - \alpha_2 + \alpha_2)\right]
\end{aligned}
$$

$$
\begin{aligned}
=&\alpha_2 H + \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{s\in S_h} q_h(s|\pi_c,\widehat{P})\sum_{a\in A}\pi_c(a|s)(|r^c(s,a) - f_{s,a}(c)| - \alpha_2)\right]\\
=&\alpha_2 H + \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s<\frac{\epsilon}{24|S|}} q_h(s|\pi_c,\widehat{P})\pi_c(a|s)(|r^c(s,a) - f_{s,a}(c)| - \alpha_2)\right]\\
&+ \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s>\frac{1}{|S|}} q_h(s|\pi_c,\widehat{P})\pi_c(a|s)(|r^c(s,a) - f_{s,a}(c)| - \alpha_2)\right]\\
&+ \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s\in[\frac{\epsilon}{24|S|},\frac{1}{|S|}]} q_h(s|\pi_c,\widehat{P})\pi_c(a|s)(|r^c(s,a) - f_{s,a}(c)| - \alpha_2)\right]\\
\leq&\alpha_2 H + \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\pi_c(a|s)\sum_{s\in S_h:\widehat{p}_s<\frac{\epsilon}{24|S|}} q_h(s|\pi_c,\widehat{P})\cdot 1\right]\\
&+ \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\pi_c(a|s)\sum_{s\in S_h:\widehat{p}_s>\frac{1}{|S|}} \widehat{p}_s(|r^c(s,a) - f_{s,a}(c)| - \alpha_2)\right]\\
&+ \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\pi_c(a|s)\sum_{s\in S_h:\widehat{p}_s\in[\frac{\epsilon}{24|S|},\frac{1}{|S|}]} \widehat{p}_s(|r^c(s,a) - f_{s,a}(c)| - \alpha_2)\right]\\
&\leq \alpha_2 H + \frac{\epsilon}{24} + \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s>\frac{1}{|S|}} \widehat{p}_s(|r^c(s,a) - f_{s,a}(c)| - \alpha_2)\right]\\
&+ \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s\in[\frac{\epsilon}{24|S|},\frac{1}{|S|}]} \widehat{p}_s(|r^c(s,a) - f_{s,a}(c)| - \alpha_2)\right]\\
&= \alpha_2 H + \frac{\epsilon}{24} + \sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s>\frac{1}{|S|}} \widehat{p}_s\mathbb{E}_{c\sim\mathcal{D}}\left[(|r^c(s,a) - f_{s,a}(c)| - \alpha_2)\right]\\
&+ \sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s\in[\frac{\epsilon}{24|S|},\frac{1}{|S|}]} \widehat{p}_s\mathbb{E}_{c\sim\mathcal{D}}\left[(|r^c(s,a) - f_{s,a}(c)| - \alpha_2)\right]
\end{aligned}
$$

$$\leq \alpha_2 H + \frac{\epsilon}{24} + \sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s>\frac{1}{|S|}} \widehat{p}_s\sqrt{\epsilon^2_\star(\widehat{p}_s)} + \sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s\in[\frac{\epsilon}{24|S|},\frac{1}{|S|}]} \widehat{p}_s\sqrt{\epsilon^2_\star(\widehat{p}_s)} \quad \text{(By inequality (6))}$$

$$= \alpha_2 H + \frac{\epsilon}{24} + \sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s>\frac{1}{|S|}} \widehat{p}_s\frac{\epsilon}{24H|S||A|} + \sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s\in[\frac{\epsilon}{24|S|},\frac{1}{|S|}]} \widehat{p}_s\frac{\epsilon}{24\widehat{p}_s|S||A|}$$

$$= \alpha_2 H + \frac{\epsilon}{8},$$

as stated. ■

Combining both errors. In the following lemma, we combine the errors of both the dynamics and rewards approximation, to obtain an expected value-difference bound for the approximated and true models, which holds for every context-dependent policy. Using it, we drive our main result in Theorem 55.

Lemma 54 *Assume the good events G_1 ,G_2 and G_3 hold. Then, for every context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$ it holds that*

$$\mathbb{E}_c[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \frac{3}{16}\epsilon + \alpha_2 H.$$

Proof Fix a context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$. By triangle inequality and linearity of expectation we have,

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{M}(c)}(s_0)|] \leq \mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{M}(c)}(s_0)|] + \mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0) - V^{\pi}_{\widehat{M}(c)}(s_0)|]$$

By Corollary 52 we have

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{M}(c)}(s_0)|] \leq \frac{\epsilon}{16}.$$

By Lemma 53 we have

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0) - V^{\pi_c}_{\widehat{M}(c)}(s_0)|] \leq \frac{\epsilon}{8} + \alpha_2 H.$$

Hence,

$$\mathbb{E}_c[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \frac{3\epsilon}{16} + \alpha_2 H.$$

■

We have established the following theorem,

Theorem 55 *With probability at least $1-\delta$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \frac{3}{8}\epsilon + 2\alpha_2 H,$$

Where $\pi^\star = (\pi^\star_c)_{c\in\mathcal{C}}$ is the optimal context-dependent policy for $\mathcal{M}$ and $\widehat{\pi}^\star = (\widehat{\pi}^\star_c)_{c\in\mathcal{C}}$ is the optimal context-dependent policy for $\widehat{\mathcal{M}}$.

Proof Assume the good events G_1, G_2 and G_3 hold. By Lemma 54 we have for $\pi^\star$ that,

$$\left|\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)]\right| \leq \mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \frac{3\epsilon}{16} + \alpha_2 H,$$

yielding,

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] \leq \frac{3\epsilon}{16} + \alpha_2 H.$$

Similarly, we have for $\widehat{\pi}^\star$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \frac{3\epsilon}{16} + \alpha_2 H.$$

Since for all $c \in \mathcal{C}$, $\widehat{\pi}^\star_c$ is the optimal policy for $\widehat{\mathcal{M}}(c)$ we have $V^{\widehat{\pi}^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0) \geq V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)$, which implies that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] \leq 0.$$

By Lemma 46 we have that $\mathbb{P}[G_1 \cap G_2 \cap G_3] \geq 1 - \delta/2$. Hence the theorem follows by summing the above inequalities. ■

For the realizable case, i.e., where $\alpha_2 = 0$, we obtain the following corollary.

Corollary 56 *For $\alpha_2 = 0$, with probability at least $1 - \delta$ we have*

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon.$$

C.3.6 BOUNDING THE ERROR CAUSED BY THE REWARDS APPROXIMATION FOR THE ℓ_1 LOSS.

In this sub-subsection, we bound the error caused by the rewards approximation, by bounding the expected value difference between the intermediate model $\widetilde{\mathcal{M}}$ and the approximated model $\widehat{\mathcal{M}}$. Here e analyse the error for the ℓ_1 loss.

Recall the definition of $\widetilde{\mathcal{M}}$. For every context $c \in \mathcal{C}$ we define $\widetilde{\mathcal{M}}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}, r^c, s_0, H)$, where we extend the true rewards function r^c for the sink by defining $r^c(s_{sink}, a) := 0, \quad \forall c \in \mathcal{C},\ a \in A$, and $\widehat{P}$ is the approximated dynamics.

Also, recall the approximated MDP for the context c. $\widehat{\mathcal{M}}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}, \widehat{r}^c, s_0, H)$ which defined in Algorithm 8.

In the following analysis, let $S^\beta(\widehat{P})$ be the set of β-reachable states for the dynamics $\widehat{P}$, and $\alpha_1 = \max_{(s,a\in S^\beta(\widehat{P})\times A)} \alpha_1(\mathcal{F}^R_{s,a})$ be the maximal agnostic approximation error.

Lemma 57 *Assume the good event G_3 holds. Then, for every context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}\left[\left|V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)\right|\right] \leq \frac{\epsilon}{8} + \alpha_1 H.$$

Proof By construction of $\widehat{\mathcal{M}}$, for every state $s \notin S^\beta(\widehat{P})$, i.e., states which are not β-reachable for $\widehat{P}$, we set $f_{s,a}(c) = 0$ for any action a. Hence,

$$|r^c(s,a) - f_{s,a}(c)| \leq 1.$$

Since the good event G_3 holds, we have for every $(s,a) \in S \times A$ such that s is β-reachable for $\widehat{P}$ that

$$\epsilon_\star(\widehat{p}_s) + \alpha_1 \geq \epsilon_\star(\widehat{p}_s) + \alpha_1(\mathcal{F}^R_{s,a}) \underbrace{\geq}_{G_3} \mathbb{E}_{c\sim\mathcal{D}}[|f_{s,a}(c) - r^c(s,a)|]. \tag{7}$$

The above implies that

$$\epsilon_\star(\widehat{p}_s) \geq \mathbb{E}_{c\sim\mathcal{D}}[|f_{s,a}(c) - r^c(s,a)| - \alpha_1]. \tag{8}$$

Fix a context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$.

By definition we have:

$$V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0) = \sum_{h=0}^{H-1} \sum_{s\in S\cup\{s_{sink}\}} q_h(s|\pi_c, \widehat{P}) \sum_{a\in A} \pi_c(a|s) r^c(s,a)$$

$$= \sum_{h=0}^{H-1} \sum_{s \in S} q_h(s|\pi_c, \widehat{P}) \sum_{a \in A} \pi_c(a|s) r^c(s,a) \qquad (r^c(s_{sink}, a) := 0, \ \ \forall c \in \mathcal{C}, a \in A.)$$

$$= \sum_{h=0}^{H-1} \sum_{s \in S_h} q_h(s|\pi_c, \widehat{P}) \sum_{a \in A} \pi_c(a|s) r^c(s,a). \qquad \text{(Since the MDP is layered)}$$

Similarly,

$$V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0) = \sum_{h=0}^{H-1} \sum_{s \in S \cup \{s_{sink}\}} q_h(s|\pi_c, \widehat{P}) \sum_{a \in A} \pi_c(a|s) \widehat{r}^c(s,a)$$

$$= \sum_{h=0}^{H-1} \sum_{s \in S} q_h(s|\pi_c, \widehat{P}) \sum_{a \in A} \pi_c(a|s) \widehat{r}^c(s,a) \qquad (\widehat{r}^c(s_{sink}, a) := 0, \ \ \forall c \in \mathcal{C}, a \in A.)$$

$$= \sum_{h=0}^{H-1} \sum_{s \in S_h} q_h(s|\pi_c, \widehat{P}) \sum_{a \in A} \pi_c(a|s) \widehat{r}^c(s,a). \qquad \text{(Since the MDP is layered)}$$

$$= \sum_{h=0}^{H-1} \sum_{s \in S_h} q_h(s|\pi_c, \widehat{P}) \sum_{a \in A} \pi_c(a|s) f_{s,a}(c). \qquad \text{(By definition, } \widehat{r}^c(s,a) := f_{s,a}(c))$$

Recall that $\beta \leq \frac{\epsilon}{24|S|}$. Thus, if s is not β-reachable for $\widehat{P}$, then $\widehat{p}_s < \beta \leq \frac{\epsilon}{24|S|}$. Moreover, if $\widehat{p}_s \geq \frac{\epsilon}{24|S|}$ then s is β-reachable for $\widehat{P}$, and the good event G_3 guarantee is hold for s. Thus, when combining all the above, by linearity of expectation and triangle inequality we obtain,

$$\begin{aligned}
&\mathbb{E}_{c \sim \mathcal{D}} \left[\left| V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0) \right| \right] \\
=&\mathbb{E}_{c \sim \mathcal{D}} \left[\left| \sum_{h=0}^{H-1} \sum_{s \in S_h} q_h(s|\pi_c, \widehat{P}) \sum_{a \in A} \pi_c(a|s) r^c(s,a) - \sum_{h=0}^{H-1} \sum_{s \in S_h} q_h(s|\pi_c, \widehat{P}) \sum_{a \in A} \pi_c(a|s) f_{s,a}(c) \right| \right] \\
\leq&\mathbb{E}_{c \sim \mathcal{D}} \left[\sum_{h=0}^{H-1} \sum_{s \in S_h} q_h(s|\pi_c, \widehat{P}) \sum_{a \in A} \pi_c(a|s) |r^c(s,a) - f_{s,a}(c)| \right] \\
=&\mathbb{E}_{c \sim \mathcal{D}} \left[\sum_{h=0}^{H-1} \sum_{s \in S_h} q_h(s|\pi_c, \widehat{P}) \sum_{a \in A} \pi_c(a|s) (|r^c(s,a) - f_{s,a}(c)| - \alpha_1 + \alpha_1) \right] \\
=&\alpha_1 H + \mathbb{E}_{c \sim \mathcal{D}} \left[\sum_{h=0}^{H-1} \sum_{s \in S_h} q_h(s|\pi_c, \widehat{P}) \sum_{a \in A} \pi_c(a|s) (|r^c(s,a) - f_{s,a}(c)| - \alpha_1) \right] \\
=&\alpha_1 H + \mathbb{E}_{c \sim \mathcal{D}} \left[\sum_{h=0}^{H-1} \sum_{a \in A} \sum_{s \in S_h : \widehat{p}_s < \frac{\epsilon}{24|S|}} q_h(s|\pi_c, \widehat{P}) \pi_c(a|s) (|r^c(s,a) - f_{s,a}(c)| - \alpha_1) \right] \\
&+ \mathbb{E}_{c \sim \mathcal{D}} \left[\sum_{h=0}^{H-1} \sum_{a \in A} \sum_{s \in S_h : \widehat{p}_s > \frac{1}{|S|}} q_h(s|\pi_c, \widehat{P}) \pi_c(a|s) (|r^c(s,a) - f_{s,a}(c)| - \alpha_1) \right] \\
&+ \mathbb{E}_{c \sim \mathcal{D}} \left[\sum_{h=0}^{H-1} \sum_{a \in A} \sum_{s \in S_h : \widehat{p}_s \in [\frac{\epsilon}{24|S|}, \frac{1}{|S|}]} q_h(s|\pi_c, \widehat{P}) \pi_c(a|s) (|r^c(s,a) - f_{s,a}(c)| - \alpha_1) \right] \\
\leq&\alpha_1 H + \mathbb{E}_{c \sim \mathcal{D}} \left[\sum_{h=0}^{H-1} \sum_{a \in A} \pi_c(a|s) \sum_{s \in S_h : \widehat{p}_s < \frac{\epsilon}{24|S|}} q_h(s|\pi_c, \widehat{P}) \cdot 1 \right]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\pi_c(a|s)\sum_{s\in S_h:\widehat{p}_s>\frac{1}{|S|}}\widehat{p}_s(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
&+ \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\pi_c(a|s)\sum_{s\in S_h:\widehat{p}_s\in[\frac{\epsilon}{24|S|},\frac{1}{|S|}]}\widehat{p}_s(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
&\leq \alpha_1 H+\frac{\epsilon}{24}+\mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s>\frac{1}{|S|}}\widehat{p}_s(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
&+ \mathbb{E}_{c\sim\mathcal{D}}\left[\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s\in[\frac{\epsilon}{24|S|},\frac{1}{|S|}]}\widehat{p}_s(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
&= \alpha_1 H+\frac{\epsilon}{24}+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s>\frac{1}{|S|}}\widehat{p}_s\mathbb{E}_{c\sim\mathcal{D}}\left[(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
&+ \sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s\in[\frac{\epsilon}{24|S|},\frac{1}{|S|}]}\widehat{p}_s\mathbb{E}_{c\sim\mathcal{D}}\left[(|r^c(s,a)-f_{s,a}(c)|-\alpha_1)\right]\\
&\leq \alpha_1 H+\frac{\epsilon}{24}+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s>\frac{1}{|S|}}\widehat{p}_s\epsilon^1_\star(\widehat{p}_s)+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s\in[\frac{\epsilon}{24|S|},\frac{1}{|S|}]}\widehat{p}_s\epsilon^1_\star(\widehat{p}_s) \qquad \text{(By inequality (8))}\\
&= \alpha_1 H+\frac{\epsilon}{24}+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s>\frac{1}{|S|}}\widehat{p}_s\frac{\epsilon}{24H|S||A|}+\sum_{h=0}^{H-1}\sum_{a\in A}\sum_{s\in S_h:\widehat{p}_s\in[\frac{\epsilon}{24|S|},\frac{1}{|S|}]}\widehat{p}_s\frac{\epsilon}{24\widehat{p}_s|S||A|}\\
&= \alpha_1 H+\frac{\epsilon}{8},
\end{aligned}
$$

as stated. ■

Combining both errors. In the following lemma, we combine the errors of both the dynamics and rewards approximation, to obtain an expected value-difference bound for the approximated and true models, which holds for every context-dependent policy. Using it, we drive our main result in Theorem 59.

Lemma 58 *Assume the good events G_1 ,G_2 and G_3 hold. Then, for every policy context-dependent policy $\pi=(\pi_c)_{c\in\mathcal{C}}$ it holds that*

$$\mathbb{E}_c[|V^{\pi_c}_{\mathcal{M}(c)}(s_0)-V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|]\leq\frac{3}{16}\epsilon+\alpha_1 H.$$

Proof Fix a context-dependent policy $\pi=(\pi_c)_{c\in\mathcal{C}}$. Then,

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0)-V^{\pi_c}_{\widehat{M}(c)}(s_0)|]\leq\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0)-V^{\pi_c}_{\widetilde{M}(c)}(s_0)|]+\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)-V^{\pi_c}_{\widehat{M}(c)}(s_0)|]$$

By Corollary 52 we have

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0)-V^{\pi_c}_{\widetilde{M}(c)}(s_0)|]\leq\frac{\epsilon}{16}.$$

By Lemma 57 we have

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)-V^{\pi_c}_{\widehat{M}(c)}(s_0)|]\leq\alpha_1 H+\frac{\epsilon}{8}.$$

Hence,

$$\mathbb{E}_c[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \frac{3\epsilon}{16} + \alpha_1 H.$$

■

We have established the following theorem,

Theorem 59 *With probability at least $1-\delta$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \frac{3}{8}\epsilon + 2\alpha_1 H,$$

where $\pi^\star = (\pi^\star_c)_{c\in\mathcal{C}}$ is the optimal context-dependent policy for $\mathcal{M}$ and $\widehat{\pi}^\star = (\widehat{\pi}^\star_c)_{c\in\mathcal{C}}$ is the optimal context-dependent policy for $\widehat{\mathcal{M}}$.

Proof Assume the good events G_1, G_2 and G_3 hold. By Lemma 58 we have for $\pi^\star$ that,

$$\left|\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)]\right| \leq \mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \frac{3\epsilon}{16} + \alpha_1 H,$$

yielding,

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] \leq \frac{3\epsilon}{16} + \alpha_1 H.$$

Similarly, we have for $\widehat{\pi}^\star$ that,

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \frac{3\epsilon}{16} + \alpha_1 H.$$

Since for all $c\in\mathcal{C}$, $\widehat{\pi}^\star_c$ is the optimal policy for $\widehat{\mathcal{M}}(c)$ we have $V^{\widehat{\pi}^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0) \geq V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)$, which implies that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] \leq 0.$$

By Lemma 46 we have that $\mathbb{P}[G_1\cap G_2\cap G_3] \geq 1-\delta/2$. Hence the theorem follows by summing the above three inequalities. ■

For the realizable case, i.e., $\alpha_1 = 0$ we have the following corollary.

Corollary 60 *For $\alpha_1 = 0$, with probability at least $1-\delta$ we have*

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon.$$

C.4 Sample complexity bounds

We present sample complexity bounds based on dimension analysis. Recall Theorems 28 and 29,

Theorem 61 (Adaption of Theorem 19.2 in Anthony et al. (1999)) *Let $\mathcal{F}$ be a hypothesis space of real valued functions with a finite pseudo dimension, denoted $Pdim(\mathcal{F}) < \infty$. Then, $\mathcal{F}$ has a uniform convergence with*

$$m(\epsilon,\delta) = O\Big(\frac{1}{\epsilon^2}(Pdim(\mathcal{F})\ln\frac{1}{\epsilon} + \ln\frac{1}{\delta})\Big).$$

Theorem 62 (Adaption of Theorem 19.1 in Anthony et al. (1999)) *Let $\mathcal{F}$ be a hypothesis space of real valued functions with a finite fat-shattering dimension, denoted $fat_{\mathcal{F}}$. Then, $\mathcal{F}$ has a uniform convergence with*

$$m(\epsilon,\delta) = O\Big(\frac{1}{\epsilon^2}(fat_{\mathcal{F}}(\epsilon/256)\ln^2\frac{1}{\epsilon} + \ln\frac{1}{\delta})\Big).$$

Remark 63 *The calculations bellow hold for any set of weights $\{\widehat{p}_s \in [0,1]\}_{s\in S}$. Hence, although $\widehat{p}_s$ is a random variable that depends on the tabular approximation of the dynamics (which affected by the observations), we can use it to compute a general bound on the sample complexity of the algorithm.*

C.4.1 SAMPLE COMPLEXITY BOUNDS FOR THE ℓ_2 LOSS

We present sample complexity for function classes with finite Pseudo dimension with ℓ_2 loss.

Corollary 64 *Assume that for every $(s,a) \in S\times A$ we have that $Pdim(\mathcal{F}^R_{s,a}) < \infty$. Let $Pdim = \max_{(s,a)\in S\times A} Pdim(\mathcal{F}^R_{s,a})$. Then, after collecting*

$$O\Big(\frac{|S|^2|A|H}{\epsilon}\ln\frac{|S||A|H}{\delta} + \frac{H^4|S|^6|A|^5(\ Pdim\ln\frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big)$$

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

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

Proof Recall that for each state-action pair (s,a) such that s is β-reachable for $\widehat{P}$, Algorithm EXPLORE-UCFD runs for $T_{s,a} = \lceil \frac{2}{\widehat{p}_s-\gamma h}(\ln(\frac{1}{\delta_1}) + \max\{N_R(\mathcal{F}^R_{s,a},\epsilon^2_\star(\widehat{p}_s),\delta_1), N_P(\gamma,\delta_1)\})\rceil$ episodes. By Theorem 55, for $\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\ge\epsilon/24|S|h}\sum_{a\in A} T_{s,a}$ samples we have with probability at least $1-\delta$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

To simplify the analysis, assume that we first lean the dynamics (for each β-reachable state and every action) and then use it to approximate the rewards using an i.i.d sample of contexts and rewards for each non-negligible state and action. Note that in algorithm EXPLORE-UCFD we do not separate between the learning phases. By corollary 38, for $\gamma = \frac{\epsilon}{48|S|H^2}$ and $\delta_1 = \frac{\delta}{6|S||A|H}$ we have that

$$N_P(\gamma,\delta_1) = O\Big(\frac{H^4|S|^2}{\epsilon^2}\Big(\ln\Big(\frac{|S||A|H}{\delta} + |S|\Big)\Big)\Big).$$

Hence, to learn the dynamics for each β-reachable state s and action a for the approximate dynamics $\widehat{P}$, we have to collect

$$O\Big(|A||S|\frac{|S|H}{\epsilon}\frac{H^4|S|^2}{\epsilon^2}\Big(\ln\Big(\frac{|S||A|H}{\delta} + |S|\Big)\Big)\Big) = O\Big(\frac{H^5|S|^4|A|}{\epsilon^3}\Big(\ln\Big(\frac{|S||A|H}{\delta} + |S|\Big)\Big)\Big).$$

trajectories. (Since for every β-reachable state $s\in S_h$ and action $a\in A$ we have $\widehat{p}_s - \gamma h \ge \beta - \gamma h \ge \beta - \gamma H \ge \gamma H = O(\epsilon/|S|H)$).

To approximate the rewards, since for every $(s,a)\in S\times A$ we have that $Pdim(\mathcal{F}^R_{s,a}) < \infty$, and $Pdim = \max_{(s,a)\in S\times A} Pdim(\mathcal{F}^R_{s,a})$, by Theorem 28, for every $(s,a)\in S\times A$ we have

$$N_R(\mathcal{F}^R_{s,a},\epsilon^2_\star(\widehat{p}_s),\delta_1) = O\left(\frac{Pdim\ln\frac{1}{\epsilon^2_\star(\widehat{p}_s)} + \ln\frac{1}{\delta_1}}{\epsilon^4_\star(\widehat{p}_s)}\right)$$

By the accuracy-per-state function, for states s that satisfies $\widehat{p}_s < \frac{\epsilon}{24|S|}$ we have $\epsilon^2_\star(\widehat{p}_s) = 1$, hence for every action a, we have that $N_R(\mathcal{F}^R_{s,a},\epsilon^2_\star(\widehat{p}_s),\delta_1) = O(\ln(1/\delta))$. Thus they are negligible.

Overall, the sample complexity for learning the rewards is as follows.

$$\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\ge 24|S|}\sum_{a\in A} T_{s,a} = \sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\ge 24|S|}\sum_{a\in A} O\Big(\frac{1}{\widehat{p}_s-\gamma h}\left(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}^R_{s,a},\epsilon^2_\star(\widehat{p}_s),\delta_1)\right)\Big)$$

$$
\begin{aligned}
&= \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s \geq 24|S|} \sum_{a\in A} O\Big(\frac{1}{\widehat{p}_s - \gamma h}\Big(\ln\frac{1}{\delta_1} + \frac{Pdim \ln \frac{1}{\epsilon^2_\star(\widehat{p}_s)} + \ln\frac{1}{\delta_1}}{\epsilon^4_\star(\widehat{p}_s)}\Big)\Big) \\
&= \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s \in [\epsilon/24|S|, 1/|S|]} \sum_{a\in A} O\Big(\frac{1}{\widehat{p}_s - \gamma h}\Big(\ln\frac{1}{\delta_1} + \frac{Pdim \ln \frac{1}{\epsilon^2/576\widehat{p}^2_s|S|^2|A|^2} + \ln\frac{1}{\delta_1}}{\epsilon^4/576^2\widehat{p}^4_s|S|^4}|A|^4\Big)\Big) \\
&\quad + \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s > 1/|S|} \sum_{a\in A} O\Big(\frac{1}{\widehat{p}_s - \gamma h}\Big(\ln\frac{1}{\delta_1} + \frac{Pdim \ln \frac{1}{\epsilon^2/576H^2|S|^2|A|^2} + \ln\frac{1}{\delta_1}}{\epsilon^4/576^2H^4|S|^4|A|^4}\Big)\Big) \\
&= \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s \in [\epsilon/24|S|, 1/|S|]} \sum_{a\in A} O\Big(\frac{1}{\widehat{p}_s - \gamma h}\Big(\ln\frac{1}{\delta_1} + \frac{\widehat{p}^4_s|S|^4|A|^4(Pdim \ln \frac{\widehat{p}^2_s|S|^2|A|^2}{\epsilon^2} + \ln\frac{1}{\delta_1})}{\epsilon^4}\Big)\Big) \\
&\quad + \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s > 1/|S|} \sum_{a\in A} O\Big(\frac{1}{\widehat{p}_s - \gamma h}\Big(\ln(\frac{1}{\delta_1} + \frac{H^4|S|^4|A|^4(Pdim \ln \frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{1}{\delta_1})}{\epsilon^4}\Big)\Big) \\
&\underbrace{\leq}_{(\star)} \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s \in [\epsilon/24|S|, 1/|S|]} \sum_{a\in A} O\Big(\frac{|S|H}{\epsilon}\ln\frac{|S||A|H}{\delta} + \frac{\widehat{p_s}}{\widehat{p}_s - \gamma h}\frac{\widehat{p}^3_s|S|^4|A|^4(Pdim \ln \frac{|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big) \\
&\quad + \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s > 1/|S|} \sum_{a\in A} O\Big(|S|\ln\frac{|S||A|H}{\delta} + \frac{H^4|S|^5|A|^4(Pdim \ln \frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big) \\
&\underbrace{\leq}_{(\star\star)} \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s \in [\epsilon/24|S|, 1/|S|]} \sum_{a\in A} O\Big(\frac{|S|H}{\epsilon}\ln\frac{|S||A|H}{\delta} + \frac{|S||A|^4(Pdim \ln \frac{|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big) \\
&\quad + \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s > 1/|S|} \sum_{a\in A} O\Big(|S|\ln\frac{|S||A|H}{\delta} + \frac{H^4|S|^5|A|^4(Pdim \ln \frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big) \\
&= O\Big(\frac{|S|^2|A|H}{\epsilon}\ln\frac{|S||A|H}{\delta} + \frac{H^4|S|^6|A|^5(\ Pdim \ln \frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big)
\end{aligned}
$$

Where $(\star)$ is since for any $h \in [H]$ we have $\widehat{p}_s - \gamma h \geq \beta - \gamma h \geq \beta - \gamma H \geq \gamma H = O(\epsilon/|S|H)$, and if $\widehat{p}_s > \frac{1}{|S|}$ we have $\widehat{p}_s - \gamma h \geq \frac{1}{|S|} - \gamma h \geq \frac{1}{|S|} - \gamma H \geq \frac{1}{|S|} - \frac{1}{|S|H} = \frac{H-1}{|S|H} = O(1/|S|)$.

$(\star\star)$ is since $\widehat{p}^3_s \leq 1/|S|^3$. In addition, $\widehat{p}_s - \gamma h \geq \gamma H$ which implies that $\frac{\gamma H}{\widehat{p}_s - \gamma h} \leq 1$. Hence,

$$\frac{\widehat{p}_s}{\widehat{p}_s - \gamma h} = \frac{\widehat{p}_s - \gamma h + \gamma h}{\widehat{p}_s - \gamma h} = 1 + \frac{\gamma h}{\widehat{p}_s - \gamma h} \leq 1 + \frac{\gamma H}{\widehat{p}_s - \gamma h} \leq 2.$$

Since the MDP is layered $|S| \geq H$, hence, the overall sample complexity is

$$O\Big(\frac{|S|^2|A|H}{\epsilon}\ln\frac{|S||A|H}{\delta} + \frac{H^4|S|^6|A|^5(\ Pdim \ln \frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big).$$

■

We also show similar sample complexity for function classes with finite fat-shattering dimension when using ℓ_2 loss.

Remark 65 *The sample complexity for function classes with finite fat-shattering dimension with ℓ_2 loss, where in $Fdim$ below we also maximizes over $\epsilon^2_\star(\widehat{p}_s)$ and the maximum is bounded and independent of $\widehat{p}_s$.*

Corollary 66 *Assume that for every $(s,a) \in S \times A$ we have that $\mathcal{F}^R_{s,a}$ has a finite fat-shattering dimension. Let $Fdim = \max_{(s,a)\in S\times A} fat_{\mathcal{F}^R_{s,a}}(\epsilon^2_\star(\widehat{p}_s)/256)$. Then, after collecting*

$$O\Big(\frac{|S|^2|A|H}{\epsilon}\ln\frac{|S||A|H}{\delta} + \frac{H^4|S|^6|A|^5(\,Fdim\ln^2\frac{H^2|S|^2|A|^2}{\epsilon^2} + \ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big)$$

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

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

Proof Recall that for each state-action pair (s,a) such that s is $\frac{\epsilon}{24|S|}$-reachable for $\widehat{P}$, we run for $T_{s,a} = \lceil \frac{2}{\widehat{p}_s-\gamma h}(\ln(\frac{1}{\delta_1})+ \max\{N_R(\mathcal{F}^R_{s,a},\epsilon^2_\star(\widehat{p}_s),\delta_1), N_P(\gamma,\delta_1)\})\rceil$ episodes. By Theorem 55, for $\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\ge\epsilon/24|S|h}\sum_{a\in A}T_{s,a}$ samples we have with probability at least $1-\delta$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

To simplify the analysis, assume that we first lean the dynamics (for each β-reachable state and every action) and then use it to approximate the rewards using an i.i.d sample of contexts an rewards for each non-negligible state and action. Recall that in algorithm EXPLORE-UCFD we do not separate between the learning phases. By corollary 38, for $\gamma = \frac{\epsilon}{48|S|H^2}$ and $\delta_1 = \frac{\delta}{6|S||A|H}$ we have that

$$N_P(\gamma,\delta_1) = O\Big(\frac{H^4|S|^2}{\epsilon^2}\Big(\ln\Big(\frac{|S||A|H}{\delta}+|S|\Big)\Big)\Big).$$

Hence, to learn the dynamics for each β-reachable state a and action a for the approximate dynamics $\widehat{P}$, we have to collect

$$O\Big(|A||S|\frac{|S|H}{\epsilon}\frac{H^4|S|^2}{\epsilon^2}\Big(\ln\Big(\frac{|S||A|H}{\delta}+|S|\Big)\Big)\Big) = O\Big(\frac{H^5|S|^4|A|}{\epsilon^3}\Big(\ln\Big(\frac{|S||A|H}{\delta}+|S|\Big)\Big)\Big).$$

trajectories. (Since for every β-reachable state $s\in S_h$ and action $a\in A$ we have $\widehat{p}_s-\gamma h \ge \beta-\gamma h \ge \beta-\gamma H \ge \gamma H = O(\epsilon/|S|H)$).

To approximate the rewards, since for every $(s,a)\in S\times A$ we have that each state-action pair has a function class $\mathcal{F}_{s,a}$ with finite fat-shattering dimension. $Fdim = \max_{(s,a)\in S\times A} fat_{\mathcal{F}^R_{s,a}}(\epsilon^2_\star(\widehat{p}_s)/256)$, by Theorem 29, for every $(s,a)\in S\times A$ we have

$$N_R(\mathcal{F}^R_{s,a},\epsilon^2_\star(\widehat{p}_s),\delta_1) = O\Big(\frac{Fdim\ln^2\frac{1}{\epsilon^2_\star(\widehat{p}_s)}+\ln\frac{1}{\delta_1}}{\epsilon^4_\star(\widehat{p}_s)}\Big).$$

By the accuracy-per-state function, for states s that satisfies $\widehat{p}_s < \frac{\epsilon}{24|S|}$ we have $\epsilon^2_\star(\widehat{p}_s)=1$, hence for every action a, we have that $N_R(\mathcal{F}^R_{s,a},\epsilon^2_\star(\widehat{p}_s),\delta_1) = O(\ln(1/\delta))$. Thus they are negligible.

Overall, the sample complexity for learning the rewards is as follows.

$$\begin{aligned}
&\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\ge 24|S|}\sum_{a\in A}T_{s,a} = \sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\ge 24|S|}\sum_{a\in A}O\left(\frac{1}{\widehat{p}_s-\gamma h}\left(\ln\frac{1}{\delta_1}+N_R(\mathcal{F}^R_{s,a},\epsilon^2_\star(\widehat{p}_s),\delta_1)\right)\right)\\
&=\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\ge 24|S|}\sum_{a\in A}O\left(\frac{1}{\widehat{p}_s-\gamma h}\left(\ln\frac{1}{\delta_1}+\frac{Fdim\ln^2\frac{1}{\epsilon^2_\star(\widehat{p}_s)}+\ln\frac{1}{\delta_1}}{\epsilon^4_\star(\widehat{p}_s)}\right)\right)\\
&=\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\in[\epsilon/24|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{Fdim\ln^2\frac{1}{\epsilon^2/576\widehat{p}^2_s|S|^2|A|^2}+\ln\frac{1}{\delta_1}}{\epsilon^4/576^2\widehat{p}^4_s|S|^4|A|^4}\Big)\Big)\\
&\quad+\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s>1/|S|}\sum_{a\in A}O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{Fdim\ln^2\frac{1}{\epsilon^2/576H^2|S|^2|A|^2}+\ln\frac{1}{\delta_1}}{\epsilon^4/576^2H^4|S|^4|A|^4}\Big)\Big)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s\in[\epsilon/24|S|,1/|S|]} \sum_{a\in A} O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{\widehat{p}_s^4|S|^4|A|^4(Fdim\ln^2\frac{\widehat{p}_s^2|S|^2|A|^2}{\epsilon^2}+\ln\frac{1}{\delta_1})}{\epsilon^4}\Big)\Big)\\
&\quad+\sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s>1/|S|} \sum_{a\in A} O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{H^4|S|^4|A|^4(Fdim\ln^2\frac{H^2|S|^2|A|^2}{\epsilon^2}+\ln\frac{1}{\delta_1})}{\epsilon^4}\Big)\Big)\\
&\underbrace{\leq}_{(\star)} \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s\in[\epsilon/24|S|,1/|S|]} \sum_{a\in A} O\Big(\frac{|S|H}{\epsilon}\ln\frac{|S||A|H}{\delta}+\frac{\widehat{p}_s}{\widehat{p}_s-\gamma h}\frac{\widehat{p}_s^3|S|^4|A|^4(Fdim\ln^2\frac{|S|^2|A|^2}{\epsilon^2}+\ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big)\\
&\quad+\sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s>1/|S|} \sum_{a\in A} O\Big(|S|\ln\frac{|S||A|H}{\delta}+\frac{H^4|S|^5|A|^4(Fdim\ln^2\frac{H^2|S|^2|A|^2}{\epsilon^2}+\ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big)\\
&\underbrace{\leq}_{(\star\star)} \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s\in[\epsilon/24|S|,1/|S|]} \sum_{a\in A} O\Big(\frac{|S|H}{\epsilon}\ln\frac{|S||A|H}{\delta}+\frac{|S||A|^4(Fdim\ln^2\frac{|S|^2|A|^2}{\epsilon^2}+\ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big)\\
&\quad+\sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s>1/|S|} \sum_{a\in A} O\Big(|S|\ln\frac{|S||A|H}{\delta}+\frac{H^4|S|^5|A|^4(Fdim\ln^2\frac{H^2|S|^2|A|^2}{\epsilon^2}+\ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big)\\
&= O\Big(\frac{|S|^2|A|H}{\epsilon}\ln\frac{|S||A|H}{\delta}+\frac{H^4|S|^6|A|^5(\,Fdim\ln^2\frac{H^2|S|^2|A|^2}{\epsilon^2}+\ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big)
\end{aligned}
$$

Where $(\star)$ is since for any $h\in[H]$ we have $\widehat{p}_s-\gamma h\geq\beta-\gamma h\geq\beta-\gamma H\geq\gamma H=O(\epsilon/|S|H)$, and if $\widehat{p}_s>\frac{1}{|S|}$ we have $\widehat{p}_s-\gamma h\geq\frac{1}{|S|}-\gamma h\geq\frac{1}{|S|}-\gamma H\geq\frac{1}{|S|}-\frac{1}{|S|H}=\frac{H-1}{|S|H}=O(1/|S|)$.

$(\star\star)$ is since $\widehat{p}_s^3\leq 1/|S|^3$ in the appropriate regime. In addition, $\widehat{p}_s-\gamma h\geq\gamma H$ which implies that $\gamma H/\widehat{p}_s-\gamma h\leq 1$. Hence

$$\frac{\widehat{p}_s}{\widehat{p}_s-\gamma h}=\frac{\widehat{p}_s-\gamma h+\gamma h}{\widehat{p}_s-\gamma h}=1+\frac{\gamma h}{\widehat{p}_s-\gamma h}\leq 1+\frac{\gamma H}{\widehat{p}_s-\gamma h}\leq 2.$$

Hence, the overall sample complexity is

$$O\Big(\frac{|S|^2|A|H}{\epsilon}\ln\frac{|S||A|H}{\delta}+\frac{H^4|S|^6|A|^5(\,Fdim\ln^2\frac{H^2|S|^2|A|^2}{\epsilon^2}+\ln\frac{|S||A|H}{\delta})}{\epsilon^4}\Big).$$

■

C.4.2 SAMPLE COMPLEXITY BOUNDS FOR THE ℓ_1 LOSS

We present sample complexity bound for function classes with finite Pseudo dimension with ℓ_1 loss.

Corollary 67 *Assume that for every $(s,a)\in S\times A$ we have that $Pdim(\mathcal{F}_{s,a}^R)<\infty$. Let $Pdim=\max_{(s,a)\in S\times A}Pdim(\mathcal{F}_{s,a}^R)$. Then, after collecting*

$$O\Big(\frac{|S|^2|A|H}{\epsilon}\ln\frac{|S||A|H}{\delta}+\frac{H^5|S|^5|A|^3(\,Pdim\ln\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^3}\Big).$$

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

$$\mathbb{E}_{c\sim\mathcal{D}}[V_{\mathcal{M}(c)}^{\pi_c^\star}(s_0)-V_{\mathcal{M}(c)}^{\widehat{\pi}_c^\star}(s_0)]\leq\epsilon+2\alpha_1 H.$$

Proof Recall that for each state-action pair (s,a) such that s is $\frac{\epsilon}{24|S|}$-reachable for $\widehat{P}$, we run for $T_{s,a}=\lceil\frac{2}{\widehat{p}_s-\gamma h}(\ln(\frac{1}{\delta_1})+\max\{N_R(\mathcal{F}_{s,a}^R,\epsilon_\star(\widehat{p}_s),\delta_1),N_P(\gamma,\delta_1)\})\rceil$ episodes. By Theorem 59, for

$\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\geq\epsilon/24|S|h}\sum_{a\in A}T_{s,a}$ samples we have with probability at least $1-\delta$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0)-V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)]\leq\epsilon+2\alpha_1 H.$$

To simplify the analysis, assume that we first lean the dynamics (for each β reachable state and every action) and then use it to approximate the rewards using an i.i.d sample of contexts an rewards for each non-negligible state and action. Recall that in algorithm EXPLORE-UCFD we do not separate between the learning phases. By corollary 38, for $\gamma=\frac{\epsilon}{48|S|H^2}$ and $\delta_1=\frac{\delta}{6|S||A|H}$ we have that

$$N_P(\gamma,\delta_1)=O\Big(\frac{H^4|S|^2}{\epsilon^2}\Big(\ln\Big(\frac{|S||A|H}{\delta}+|S|\Big)\Big)\Big).$$

Hence, to learn the dynamics for each β-reachable state a and action a for the approximate dynamics $\widehat{P}$, we have to collect

$$O\Big(|A||S|\frac{|S|H}{\epsilon}\frac{H^4|S|^2}{\epsilon^2}\Big(\ln\Big(\frac{|S||A|H}{\delta}+|S|\Big)\Big)\Big)=O\Big(\frac{H^5|S|^4|A|}{\epsilon^3}\Big(\ln\Big(\frac{|S||A|H}{\delta}+|S|\Big)\Big)\Big).$$

trajectories. (Since for every β-reachable state $s\in S_h$ and action $a\in A$ we have $\widehat{p}_s-\gamma h\geq\beta-\gamma h\geq\beta-\gamma H\geq\gamma H=O(\epsilon/|S|H)$).

To approximate the rewards, Since for every $(s,a)\in S\times A$ we have that $Pdim(\mathcal{F}^R_{s,a})<\infty$, and $Pdim=\max_{(s,a)\in S\times A}Pdim(\mathcal{F}^R_{s,a})$, by Theorem 28, for every $(s,a)\in S\times A$ we have

$$N_R(\mathcal{F}^R_{s,a},\epsilon_\star(\widehat{p}_s),\delta_1)=O\Big(\frac{Pdim\ln\frac{1}{\epsilon_\star(\widehat{p}_s)}+\ln\frac{1}{\delta_1}}{\epsilon^2_\star(\widehat{p}_s)}\Big)$$

By the accuracy-per-state function, for states s that satisfies $\widehat{p}_s<\frac{\epsilon}{24|S|}$ we have $\epsilon_\star(\widehat{p}_s)=1$, hence for every action a, we have that $N_R(\mathcal{F}^R_{s,a},\epsilon_\star(\widehat{p}_s),\delta_1)=O(\ln(1/\delta))$. Thus they are negligible.

Overall, the sample complexity for learning the rewards is as follows.

$$\begin{aligned}
&\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\geq 24|S|}\sum_{a\in A}T_{s,a}=\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\geq 24|S|}\sum_{a\in A}O\left(\frac{1}{\widehat{p}_s-\gamma h}\left(\ln\frac{1}{\delta_1}+N_R(\mathcal{F}^R_{s,a},\epsilon_\star(\widehat{p}_s),\delta_1)\right)\right)\\
&=\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\geq 24|S|}\sum_{a\in A}O\left(\frac{1}{\widehat{p}_s-\gamma h}\left(\ln\frac{1}{\delta_1}+\frac{Pdim\ln\frac{1}{\epsilon_\star(\widehat{p}_s)}+\ln\frac{1}{\delta_1}}{\epsilon^2_\star(\widehat{p}_s)}\right)\right)\\
&=\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\in[\epsilon/24|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{Pdim\ln\frac{1}{\epsilon/24\widehat{p}_s|S||A|}+\ln\frac{1}{\delta_1}}{\epsilon^2/576\widehat{p}_s^2|S|^2|A|^2})\Big)\\
&\quad+\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s>1/|S|}\sum_{a\in A}O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{Pdim\ln\frac{1}{\epsilon/24H|S||A|}+\ln\frac{1}{\delta_1}}{\epsilon^2/576H^2|S|^2|A|^2}\Big)\Big)\\
&=\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\in[\epsilon/24|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{\widehat{p}_s^2|S|^2|A|^2(Pdim\ln\frac{\widehat{p}_s|S||A|}{\epsilon}+\ln\frac{1}{\delta_1})}{\epsilon^2}\Big)\Big)\\
&\quad+\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s>1/|S|}\sum_{a\in A}O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{H^2|S|^2|A|^2(Pdim\ln\frac{H|S||A|}{\epsilon}+\ln\frac{1}{\delta_1})}{\epsilon^2}\Big)\Big)\\
&\underbrace{\leq}_{(\star)}\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\in[\epsilon/24|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{|S|H}{\epsilon}\ln\frac{|S||A|H}{\delta}+\frac{\widehat{p}_s}{\widehat{p}_s-\gamma h}\frac{\widehat{p}_s|S|^2|A|^2(Pdim\ln\frac{|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^2}\Big)\\
&\quad+\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s>1/|S|}\sum_{a\in A}O\Big(|S|\ln\frac{|S||A|H}{\delta}+\frac{H^2|S|^3|A|^2(Pdim\ln\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^2}\Big)
\end{aligned}$$

$$\underbrace{\leq}_{(\star\star)} \sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s\in[\epsilon/24|S|,1/|S|]} \sum_{a\in A} O\Big(\frac{|S|H}{\epsilon}\ln\frac{|S||A|H}{\delta} + \frac{|S||A|^2(Pdim\ln\frac{|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^2}\Big)$$
$$+\sum_{h=0}^{H-1} \sum_{s\in S_h:\widehat{p}_s>1/|S|} \sum_{a\in A} O\Big(|S|\ln\frac{|S||A|H}{\delta} + \frac{H^2|S|^3|A|^2(Pdim\ln\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^2}\Big)$$
$$=O\Big(\frac{|S|^2|A|H}{\epsilon}\ln\frac{|S||A|H}{\delta} + \frac{H^2|S|^4|A|^3(\ Pdim\ln\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^2}\Big)$$

Where $(\star)$ is since for any $h\in[H]$ we have $\widehat{p}_s-\gamma h\geq\beta-\gamma h\geq\beta-\gamma H\geq\gamma H=O(\epsilon/|S|H)$, and if $\widehat{p}_s>\frac{1}{|S|}$ we have $\widehat{p}_s-\gamma h\geq\frac{1}{|S|}-\gamma h\geq\frac{1}{|S|}-\gamma H\geq\frac{1}{|S|}-\frac{1}{|S|H}=\frac{H-1}{|S|H}=O(1/|S|)$.

$(\star\star)$ is since $\widehat{p}_s\leq 1/|S|$ in the appropriate regime. In addition, $\widehat{p}_s-\gamma h\geq\gamma H$ which implies that $\gamma H/\widehat{p}_s-\gamma h\leq 1$. Hence

$$\frac{\widehat{p}_s}{\widehat{p}_s-\gamma h}=\frac{\widehat{p}_s-\gamma h+\gamma h}{\widehat{p}_s-\gamma h}=1+\frac{\gamma h}{\widehat{p}_s-\gamma h}\leq 1+\frac{\gamma H}{\widehat{p}_s-\gamma h}\leq 2.$$

Hence, the overall sample complexity is

$$O\Big(\frac{|S|^2|A|H}{\epsilon}\ln\frac{|S||A|H}{\delta} + \frac{H^5|S|^5|A|^3(\ Pdim\ln\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^3}\Big).$$

■

We also show similar sample complexity for function classes with finite fat-shattering dimension when using ℓ_1 loss.

Remark 68 *The sample complexity for function classes with finite fat-shattering dimension with ℓ_1 loss, where in $Fdim$ below we also maximizes over $\epsilon_\star(\widehat{p}_s)$ and the maximum is bounded and independent of $\widehat{p}_s$.*

Corollary 69 *Assume that for every $(s,a)\in S\times A$ we have that $\mathcal{F}^R_{s,a}$ has finite fat-shattering dimension. Let $Fdim=\max_{(s,a)\in S\times A} fat_{\mathcal{F}^R_{s,a}}(\epsilon_\star(\widehat{p}_s)/256)$. Then, after collecting*

$$O\Big(\frac{|S|^2|A|H}{\epsilon}\ln\frac{|S||A|H}{\delta} + \frac{H^5|S|^5|A|^3(\ Fdim\ln^2\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^3}\Big).$$

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

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0)-V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)]\leq\epsilon+2\alpha_1 H.$$

Proof Recall that for each state-action pair (s,a) such that s is $\frac{\epsilon}{24|S|}$ reachable for $\widehat{P}$, we run for $T_{s,a}=\lceil\frac{2}{\widehat{p}_s-\gamma h}(\ln(\frac{1}{\delta_1})+\max\{N_R(\mathcal{F}^R_{s,a},\epsilon_\star(\widehat{p}_s),\delta_1),N_P(\gamma,\delta_1)\})\rceil$ episodes. By Theorem 59, for $\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\geq\epsilon/24|S|h}\sum_{a\in A}T_{s,a}$ samples we have with probability at least $1-\delta$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0)-V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)]\leq\epsilon+2\alpha_1 H.$$

To simplify the analysis, assume that we first lean the dynamics (for each β-reachable state and every action) and then use it to approximate the rewards using an i.i.d sample of contexts an rewards for each non-negligible state and action. Recall that in algorithm EXPLORE-UCFD we do not separate between the learning phases. By corollary 38, for $\gamma=\frac{\epsilon}{48|S|H^2}$ and $\delta_1=\frac{\delta}{6|S||A|H}$ we have that

$$N_P(\gamma,\delta_1)=O\Big(\frac{H^4|S|^2}{\epsilon^2}\Big(\ln\Big(\frac{|S||A|H}{\delta}+|S|\Big)\Big)\Big).$$

Hence, to learn the dynamics for each β-reachable state a and action a for the approximate dynamics $\widehat{P}$, we have to collect

$$O\Big(|A||S|\frac{|S|H}{\epsilon}\frac{H^4|S|^2}{\epsilon^2}\Big(\ln\Big(\frac{|S||A|H}{\delta}+|S|\Big)\Big)\Big)=O\Big(\frac{H^5|S|^4|A|}{\epsilon^3}\Big(\ln\Big(\frac{|S||A|H}{\delta}+|S|\Big)\Big)\Big)$$

trajectories. (Since for every β-reachable state $s\in S_h$ and action $a\in A$ we have $\widehat{p}_s-\gamma h\geq\beta-\gamma h\geq\beta-\gamma H\geq\gamma H=O(\epsilon/|S|H)$).

To approximate the rewards, since or every $(s,a)\in S\times A$ we have that $\mathcal{F}^R_{s,a}$ has finite fat-shattering dimension, and $Fdim=\max_{(s,a)\in S\times A}fat_{\mathcal{F}^R_{s,a}}(\epsilon_\star(\widehat{p}_s)/256)$, by Theorem 29, for every $(s,a)\in S\times A$ we have

$$N_R(\mathcal{F}^R_{s,a},\epsilon_\star(\widehat{p}_s),\delta_1)=O\Big(\frac{Fdim\ln^2\frac{1}{\epsilon_\star(\widehat{p}_s)}+\ln\frac{1}{\delta_1}}{\epsilon^2_\star(\widehat{p}_s)}\Big)$$

By the accuracy-per-state function, for states s that satisfies $\widehat{p}_s<\frac{\epsilon}{24|S|}$ we have $\epsilon_\star(\widehat{p}_s)=1$, hence for every action a, we have that $N_R(\mathcal{F}^R_{s,a},\epsilon_\star(\widehat{p}_s),\delta_1)=O(\ln(1/\delta))$. Thus they are negligible.

Overall, the sample complexity for learning the rewards is as follows.

$$\begin{aligned}
&\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\geq 24|S|}\sum_{a\in A}T_{s,a}=\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\geq 24|S|}\sum_{a\in A}O\Big(\frac{1}{\widehat{p}_s-\gamma h}(\ln(\frac{1}{\delta_1})+N_R(\mathcal{F}^R_{s,a},\epsilon_\star(\widehat{p}_s),\delta_1))\Big)\\
=&\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\geq 24|S|}\sum_{a\in A}O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{Fdim\ln^2\frac{1}{\epsilon_\star(\widehat{p}_s)}+\ln\frac{1}{\delta_1}}{\epsilon^2_\star(\widehat{p}_s)}\Big)\Big)\\
=&\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\in[\epsilon/24|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{Fdim\ln^2\frac{1}{\epsilon/24\widehat{p}_s|S||A|}+\ln\frac{1}{\delta_1}}{\epsilon^2/576\widehat{p}^2_s|S|^2|A|^2}\Big)\Big)\\
&+\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s>1/|S|}\sum_{a\in A}O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{Fdim\ln^2\frac{1}{\epsilon/24H|S||A|}+\ln\frac{1}{\delta_1}}{\epsilon^2/576H^2|S|^2|A|^2}\Big)\Big)\\
=&\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\in[\epsilon/24|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{\widehat{p}^2_s|S|^2|A|^2(Fdim\ln^2\frac{\widehat{p}_s|S||A|}{\epsilon}+\ln\frac{1}{\delta_1})}{\epsilon^2}\Big)\Big)\\
&+\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s>1/|S|}\sum_{a\in A}O\Big(\frac{1}{\widehat{p}_s-\gamma h}\Big(\ln\frac{1}{\delta_1}+\frac{H^2|S|^2|A|^2(Fdim\ln^2\frac{H|S||A|}{\epsilon}+\ln\frac{1}{\delta_1})}{\epsilon^2}\Big)\Big)\\
\underbrace{\leq}_{(\star)}&\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\in[\epsilon/24|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{|S|H}{\epsilon}\ln\frac{|S||A|H}{\delta}+\frac{\widehat{p}_s}{\widehat{p}_s-\gamma h}\frac{\widehat{p}_s|S|^2|A|^2(Fdim\ln^2\frac{|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^2}\Big)\\
&+\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s>1/|S|}\sum_{a\in A}O\Big(|S|\ln\frac{|S||A|H}{\delta}+\frac{H^2|S|^3|A|^2(Fdim\ln^2\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^2}\Big)\\
\underbrace{\leq}_{(\star\star)}&\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s\in[\epsilon/24|S|,1/|S|]}\sum_{a\in A}O\Big(\frac{|S|H}{\epsilon}\ln\frac{|S||A|H}{\delta}+\frac{|S||A|^2(Fdim\ln^2\frac{|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^2}\Big)\\
&+\sum_{h=0}^{H-1}\sum_{s\in S_h:\widehat{p}_s>1/|S|}\sum_{a\in A}O\Big(|S|\ln\frac{|S||A|H}{\delta}+\frac{H^2|S|^3|A|^2(Fdim\ln^2\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^2}\Big)\\
=&O\Big(\frac{|S|^2|A|H}{\epsilon}\ln\frac{|S||A|H}{\delta}+\frac{H^2|S|^4|A|^3(\ Fdim\ln^2\frac{H|S||A|}{\epsilon}+\ln\frac{|S||A|H}{\delta})}{\epsilon^2}\Big)
\end{aligned}$$

Where $(\star)$ is since for any $h \in [H]$ we have $\widehat{p}_s - \gamma h \geq \beta - \gamma h \geq \beta - \gamma H \geq \gamma H = O(\epsilon/|S|H)$, and if $\widehat{p}_s > \frac{1}{|S|}$ we have $\widehat{p}_s - \gamma h \geq \frac{1}{|S|} - \gamma h \geq \frac{1}{|S|} - \gamma H \geq \frac{1}{|S|} - \frac{1}{|S|H} = \frac{H-1}{|S|H} = O(1/|S|)$.

$(\star\star)$ is since $\widehat{p}_s \leq 1/|S|$ in the appropriate regime. In addition, $\widehat{p}_s - \gamma h \geq \gamma H$ which implies that $\gamma H/\widehat{p}_s - \gamma h \leq 1$. Hence

$$\frac{\widehat{p}_s}{\widehat{p}_s - \gamma h} = \frac{\widehat{p}_s - \gamma h + \gamma h}{\widehat{p}_s - \gamma h} = 1 + \frac{\gamma h}{\widehat{p}_s - \gamma h} \leq 1 + \frac{\gamma H}{\widehat{p}_s - \gamma h} \leq 1 + \frac{\gamma H}{\gamma H} = 2.$$

Hence, the overall sample complexity is

$$O\Big(\frac{|S|^2|A|H}{\epsilon} \ln \frac{|S||A|H}{\delta} + \frac{H^5|S|^5|A|^3(\ Fdim \ln^2 \frac{H|S||A|}{\epsilon} + \ln \frac{|S||A|H}{\delta})}{\epsilon^3}\Big).$$

■

Appendix D. Known and Context Dependent Dynamics

In this section we address the challenging model of context dependent dynamics. Meaning, that each context induces a potentially different dynamics. Clearly, this implies that for any policy π (which can be either context-dependent or context-independent), the occupancy measure is determined by the context (due to the context-dependent dynamics). Hence, a state $s \in S$ that is highly-reachable for some context $c_1 \in \mathcal{C}$ might be poorly-reachable for a different context $c_2 \in \mathcal{C}$. (Something which is impossible in the context-free dynamics setting.)

For the known context-dependent dynamics setting we stay with a similar strategy as in the context-free dynamics, and do the approximation per state-action pair. In order to overcome the reachability issue, we define for each state s a subset of good contexts $\mathcal{C}^\beta(s)$ whose induced dynamics reaches s with non-negligible probability, i.e., β. A state s is (γ, β)-good if the probability of $\mathcal{C}^\beta(s)$ is at least γ. For each (γ, β)-good state s we build a sample in which the marginal distribution of the context is $\mathcal{D}$ restricted to $\mathcal{C}^\beta(s)$. We do this by using importance sampling. We can implement the importance sampling since the context-dependent dynamics are known, hence, the probability of reach state s under a good context c can be computed, say it is q. We accept a sample that reaches state s with probability $\beta/q \leq 1$. Given such that a sample we can use the ERM oracle and get a good approximation of rewards. Our approximate optimal policy is similar to the case of known context-free dynamics, with the modification that given a context c we use the dynamics P^c in the approximated MDP $\widehat{\mathcal{M}}(c)$.

D.1 Algorithm

We start with an overview of our algorithm EXPLORE-KCDD (Algorithm 10) which works in stages. Each stage learns a layer. When learning layer $h \in [H-1]$ we sample only the (γ, β)-good states of layer h.

Since the distribution over the contexts is unknown, we first need to approximate the probability $\mathbb{P}[c \in \mathcal{C}^\beta(s_h)]$ for each state $s_h \in S_h$, to approximate the set of (γ, β)-good states of layer h. We do it using mean estimation as described in algorithm AGC (i.e., Algorithm 9).

For every layer $h \in [H-1]$ we first approximate the set $S_h^{\gamma,\beta}$ of (γ, β)-good states. Then, for each $s_h \in S_h^{\gamma,\beta}$ and every action $a_h \in A$, we do the following for "sufficient" number of episodes:

(1) We observe the episode context c, compute $\pi^c_{s_h} = \arg\max_{\pi: S \to A} q_h(s_h|\pi, P^c)$ and set $\pi^c_{s_h}(s_h) \leftarrow a_h$, which guarantees that we perform action a_h in state s_h. (2) If $q_h(s_h|\pi^c_{s_h}, P^c) \geq \beta$, we run $\pi^c_{s_h}$ to generate a trajectory τ.

(3) If $(s_h, a_h, r_h) \in \tau$ we add $((c, s_h, a_h), r_h)$ to the sample of (s_h, a_h) with probability $\beta/q_h(s_h|\pi^c_{s_h}, P^c) \leq 1$. After collecting the samples, we approximate the rewards (as function of the context) using the ERM oracle, $f_{s_h,a_h} = \texttt{ERM}(\mathcal{F}^R_{s_h,a_h}, Sample(s_h, a_h), \ell)$.

(4) For states s which are not (γ, β)-good we set $f_{s,a} = 0$, for every action a.

Algorithm EXPLOIT-KCDD (Algorithm 11) get as inputs the MDP parameters (except for the context-dependent rewards function) and the functions approximate the rewards (that computed using algorithm EXPLORE-KCDD).

Given a context c it computes the approximated MDP $\widehat{\mathcal{M}}(c)$ and use it to compute a near optimal context-dependent policy $\hat{\pi}_c^\star$. Then, it run $\hat{\pi}_c^\star$ to generate trajectory. Recall that $\widehat{\mathcal{M}}(c) = (S, A, P^c, s_0, \widehat{r}^c, H)$ where we define $\forall s \in S, a \in A : \widehat{r}^c(s,a) = f_{s,a}(c)$.

Algorithm 9 Approximate Good Contexts (AGC)

inputs:
- MDP parameters: $S = \{S_0, S_1, \ldots, S_H\}$, A, H.
- P^c - The context-dependent dynamics.
- Reachability parameters: γ ,β
- Accuracy and confidence parameters ϵ_2, δ_2, where $\epsilon_2 \leq \gamma$.
- Current layer h and state s_h.

calculate $m(\epsilon_2, \delta_2) = \left\lceil \frac{\ln \frac{2}{\delta_2}}{2\epsilon_2^2} \right\rceil$
initialize $counter = 0$
for $t = 1, 2, ..., m(\epsilon_2, \delta_2)$ **do**
    observe context c_t
    if $c_t \in \mathcal{C}^\beta(s_h)$ **then**
        $Counter = Counter + 1$
$\widehat{p}_\beta(s_h) = \frac{Counter}{m(\epsilon_2, \delta_2)}$
return $\mathbb{I}[\widehat{p}_\beta(s_h) \geq \gamma - \epsilon_2]$ and $\widehat{p}_\beta(s_h)$

Remark 70 *The check whether $c \in \mathcal{C}^\beta(s)$ can be done in $poly(|S|, |A|, H)$ time by computing the maximal probability to visit s under the dynamics P^c, say it is p_s^c, and then check whether $p_s^c \geq \beta$.*

D.2 Analysis

D.2.1 Analysis Outline

In the following, we present analysis for both the ℓ_1 (see Sub-subsection D.2.4) and ℓ_2 (see Sub-subsection D.2.3) loss functions.

For both loss functions, our goal is to bound the expected value difference of the true and approximated models, i.e., $\mathcal{M}(c)$ and $\widehat{\mathcal{M}}(c)$, for any context-dependent policy $\pi = (\pi_c)_{c \in \mathcal{C}}$, with high probability. (See Lemmas 78 and 75).

Using that bound, we derive a bound on the expected value difference between the optimal context-dependent policy $\pi^\star = (\pi_c^\star)_{c \in \mathcal{C}}$ and our approximated optimal policy $\widehat{\pi}^\star = (\hat{\pi}_c^\star)_{c \in \mathcal{C}}$ on the true model, which holds with high probability. (See Theorems 79 and 76).

Lastly, we derive sample complexity bound using known uniform convergence sample complexity bounds for the Pseudo dimension (See Theorem 28) and the fat-shattering dimension (See Theorem 29). For the sample complexity analysis, see Sub-subsection D.3.2 for the ℓ_1 loss, and D.2.3 for the ℓ_2 loss.

D.2.2 Good Events

Event G_1. Let G_1 denote the good event in which for all $h \in [H-1]$ and $s_h \in S_h$ we have $|\widehat{p}_\beta(s_h) - \mathbb{P}_{c \sim \mathcal{D}}[c \in \mathcal{C}^\beta(s_h)]| \leq \epsilon_2$, for $\widehat{p}_\beta(s_h)$ that is defined in Algorithm AGC (i.e., Algorithm 9).

For $\epsilon_2 = \gamma/2$, event G_1 guarantees that for every layer $h \in [H-1]$ and state $s_h \in S_h$, if $s_h \in S_h^{\gamma,\beta}$, then Algorithm AGC will identity that s_h is (γ, β)-good. Hence, in Algorithm EXPLORE-KCDD we will collect samples for it.

The following lemma shows that event G_1 holds with high probability.

Algorithm 10 Explore Rewards for Known and Context-Dependent Dynamics (EXPLORE-KCDD)

inputs:

- CMDP parameters: $S = \{S_0, S_1, \ldots, S_H\}$ - a layered states space, A, P^c -a context-dependent transition probabilities matrix, s_0 - the unique start state, H - the horizon length.
- Accuracy and confidence parameters: ϵ, δ.
- $\forall s \in S, a \in A: \quad \mathcal{F}^R_{s,a}$ - the function classes use to approximate the rewards function.
- $N_R(\mathcal{F}, \epsilon, \delta)$ - sample complexity function for the ERM oracle.
- The extended readability parameters: β, γ.
- ℓ - a loss function (assumed to be ℓ_1 or ℓ_2).

set $\delta_1 = \frac{\delta}{6|S||A|}, \delta_2 = \frac{\delta}{6|S|}, \epsilon_2 = \gamma/2$

set $\epsilon_1 = \begin{cases} \frac{\epsilon^2}{64|S||A|H^2}, & \text{if } \ell = \ell_1 \\ \frac{\epsilon^3}{8^3|S||A|H^3}, & \text{if } \ell = \ell_2 \end{cases}$

for $h \in [H-1]$ **do**
  for $s_h \in S_h$ **do**
    $I(s_h), \widehat{p}_\beta(s_h) \leftarrow \texttt{AGC}(S, A, H, P^c, \delta_2, \epsilon_2, \gamma, \beta, h, s_h)$
    if $I(s_h) == 1$ **then**
      for $a_h \in A$ **do**
        initialize $Sample(s_h, a_h) = \emptyset$
        compute the required number of episodes

$$T_{s_h,a_h} = \lceil \frac{2}{\beta\gamma}(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon_1, \delta_1)) \rceil$$

        for $t = 1, 2, \ldots, T_{s_h,a_h}$ **do**
          observe context c_t
          $(\pi^{c_t}_{s_h}, p^{c_t}_{s_h}) \leftarrow \texttt{FFP}(S, A, P^c, s_0, H, s_h)$
          $\pi^{c_t}_{s_h}(s_h) \leftarrow a_h$
          if $p^{c_t}_{s_h} \geq \beta$ **then**
            run $\pi^{c_t}_{s_h}$ to generate trajectory τ_t
            if (s_h, a_h, r_h) is in τ_t, for a reward $r_h \in [0,1]$ **then**
              with probability $\frac{\beta}{p^{c_t}_{s_h}}$ add $((c_t, s_h, a_h), r_h)$ to $Sample(s_h, a_h)$
        if $|Sample(s_h, a_h)| \geq N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon_1, \delta_1)$ **then**
          $f_{s_h,a_h} = \texttt{ERM}(\mathcal{F}^R_{s_h,a_h}, Sample(s_h, a_h), \ell)$
        else
          set $f_{s_h,a_h} = 0$
    else
      set for all $a \in A$: $f_{s_h,a} = 0$

return $\{f_{s,a} : \forall (s,a) \in S \times A\}$

Algorithm 11 Exploit for Known and Context-Dependent Dynamics (EXPLOIT-KCDD)

inputs:

- MDP parameters: $S = \{S_0, S_1, \ldots, S_H\}$, A, s_0, H.
- P^c -A context-dependent transition probabilities matrix.
- Accuracy and confidence parameters: ϵ, δ.
- $\forall s \in S, a \in A: \quad \mathcal{F}^R_{s,a}$ - the function classes use to approximate the rewards function.
- $N_R(\mathcal{F}, \epsilon, \delta)$ - sample complexity function for the ERM oracle.
- Reachability parameters: γ,β and $S_h^{\gamma,\beta}$ for every $h \in [H-1]$
- Functions approximate the rewards for each state-action pair: $\{f_{s,a} : \forall (s,a) \in S \times A\}$.

for $t = 1, 2, \ldots$ **do**

observe context c_t

define $\widehat{\mathcal{M}}(c_t) = (S, A, P^{c_t}, \widehat{r}^{c_t}, s_0, H)$ where $\widehat{r}^{c_t}$ defined as:

$$\forall h \in [H-1], s_h \in S_h^{\gamma,\beta}, a_h \in A : \widehat{r}^{c_t}(s_h, a_h) = f_{s_h,a_h}(c_t)\mathbb{I}[c_t \in \mathcal{C}^\beta(s_h)]$$
$$\forall h \in [H-1], s_h \notin S_h^{\gamma,\beta}, a_h \in A : \widehat{r}^{c_t}(s_h, a_h) = 0$$

compute the optimal policy for $\widehat{\mathcal{M}}(c_t)$, $(\widehat{\pi}^{c_t}, \cdot) \leftarrow \texttt{Planning}(\widehat{\mathcal{M}}(c_t))$

run $\widehat{\pi}^{c_t}$ to generate trajectory.

Lemma 71 *For $\delta_2 = \delta/6|S|$ it holds that $\mathbb{P}[G_1] \geq 1 - \frac{\delta}{6}$.*

Proof For every state $s \in S$, by Hoeffding's inequality, for $m \geq \frac{\ln \frac{2}{\delta_2}}{2\epsilon_2^2}$ examples, we have with probability at least $1 - \delta_2$ that $|\widehat{p}_\beta(s) - \mathbb{P}_{c\sim\mathcal{D}}[c \in \mathcal{C}^\beta(s_h)]| \leq \epsilon_2$. Hence, using union bound over the states, we obtain the lemma. ■

Event G_2. Recall that for every $h \in [H-1]$ we define $S_h^{\gamma,\beta} = \{s_h \in S_h : \mathbb{P}[c \in \mathcal{C}^\beta(s_h)] \geq \gamma\}$ where $\mathcal{C}^\beta(s_h) = \{c \in \mathcal{C} : s_h \text{ is } \beta\text{-reachable for } P^c\}$.

Let G_2 denote the good event in which for every layer $h \in [H]$ and state $s_h \in S_h^{\gamma,\beta}$ for every action $a_h \in A$ we have that $|Sample(s_h, a_h)| \geq N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon_1, \delta_1)$.

The following lemma shows that event G_2 holds with high probability.

Lemma 72 *We have $\mathbb{P}_{c\sim\mathcal{D}}[G_2|G_1] \geq 1 - \delta/6$.*

Proof Fix a layer $h \in [H]$, a state $s_h \in S_h^{\gamma,\beta}$ and an action $a_h \in A$.

Let $p_\beta(s_h) := \mathbb{P}_{c\sim\mathcal{D}}[c \in \mathcal{C}^\beta(s_h)]$. Since $s_h \in S_h^{\gamma,\beta}$ it holds that $p_\beta(s_h) \geq \gamma$. Since G_1 holds we have that $p_\beta(s_h) - \epsilon_2 \leq \widehat{p}_\beta(s_h) \leq p_\beta(s_h) + \epsilon_2$ which yielding that $\widehat{p}_\beta(s_h) \geq p_\beta(s_h) - \epsilon_2 \geq \gamma - \epsilon_2$.

Hence, under G_1, the agent will identify that s_h is in $S_h^{\gamma,\beta}$ and try to collect at least $N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon_1, \frac{\delta}{6|S||A|})$ examples of it, for every the action a_h.

For a fixed context $c \in \mathcal{C}$, let $\pi^c_{s_h}$ denote the policy with the highest probability to visit s_h, which returned by algorithm `FFP`.

Since $s_h \in S_h^{\gamma,\beta}$ we have that $\mathbb{P}[c \in \mathcal{C}^\beta(s_h)] \geq \gamma$. Recall that we collect only examples of contexts $c \in \mathcal{C}^\beta(s_h)$.

Hence, the probability to observe a context $c \in \mathcal{C}^\beta(s_h)$ and then visit s_h when playing according to $\pi^c_{s_h}$ is at least $\gamma \cdot q_h(s_h|\pi^c_{s_h}, P^c) \geq \gamma\beta$, since for $c \in \mathcal{C}^\beta(s_h)$ we have that s_h is β-reachable for P^c, which implies that $q_h(s_h|\pi^c_{s_h}, P^c) \geq \beta$.

Since we use importance sampling, the probability that an observed example $((c, s_h, a_h), r_h)$ (for $c \in \mathcal{C}^\beta(s_h)$) will be added to $Sample(s_h, a_h)$ is $\frac{\beta}{q_h(s_h|\pi^c_{s_h}, P^c)}$. Overall, the probability of adding a sample of (c, s_h, a_h) to $Sample(s_h, a_h)$ is at least

$$\frac{\beta}{q_h(s_h|\pi^c_{s_h}, P^c)} \cdot q_h(s_h|\pi^c_{s_h}, P^c) \cdot \gamma = \beta\gamma.$$

Hence, in expectation, the agent needs to experience at most $\frac{1}{\beta\gamma}$ episodes to collect one such example of (s_h, a_h) for $s_h \in S^{\gamma,\beta}_h$.

Using Hoeffding's inequality, we obtain that with probability at least $1-\delta_1$, the agent will collect at least $N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon_1, \delta_1)$ examples after experiencing

$$T_{s_h,a_h} = \lceil \frac{2}{\beta\gamma}(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon_1, \delta_1))\rceil$$

episodes. For $\delta_1 = \frac{\delta}{6|S||A|}$, we obtain the lemma using union bound over $(s_h, a_h) \in S^{\gamma,\beta}_h \times A$ for every $h \in [H-1]$. ■

Event G_3. Let G_3 denote the good event in which for every layer $h \in [H-1]$ and state $s_h \in S^{\gamma,\beta}_h$ we have for every action $a_h \in A$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[(f_{s_h,a_h}(c) - r^c(s_h, a_h))^2 | c \in \mathcal{C}^\beta(s_h)] \leq \epsilon_1 + \alpha_2^2(\mathcal{F}^R_{s_h,a_h}).$$

for the ℓ_2 loss, (or $\mathbb{E}_{c\sim\mathcal{D}}[|f_{s_h,a_h}(c) - r^c(s_h, a_h)||c \in \mathcal{C}^\beta(s_h)] \leq \epsilon_1 + \alpha_1(\mathcal{F}^R_{s_h,a_h})$ for the ℓ_1 loss).

The following lemma shows that given events G_1 and G_2 hold, event G_3 holds with high probability.

Lemma 73 *We have $\mathbb{P}[G_3|G_1, G_2] \geq 1 - \delta/6$.*

Proof Since G_1 and G_2 hold, we have for every layer $h \in [H-1]$, state $s_h \in S^{\gamma,\beta}_h$ and action $a_h \in A$ that $|Sample(s_h, a_h)| \geq N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon_1, \delta_1)$. Hence, we compute f_{s_h,a_h} using the ERM oracle, and by the ERM guarantees (see 3), for every layer $h \in [H-1]$, state $s_h \in S^{\gamma,\beta}_h$ and an action $a_h \in A$ we have with probability at least $1 - \delta_1$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[(f_{s_h,a_h}(c) - r^c(s_h, a_h))^2 | c \in \mathcal{C}^\beta(s_h)] \leq \epsilon_1 + \alpha^2(\mathcal{F}^R_{s_h,a_h}).$$

for the ℓ_2 loss. ($\mathbb{E}_{c\sim\mathcal{D}}[|f_{s_h,a_h}(c) - r^c(s_h, a_h)||c \in \mathcal{C}^\beta(s_h)] \leq \epsilon_1 + \alpha_1(\mathcal{F}^R_{s_h,a_h})$ for the ℓ_1 loss.) For $\delta_1 = \frac{\delta}{6|S||A|}$ the lemma follows from union bound over each appropriate state-action pair. ■

By combining all the above, we obtain that all of the good events hold with high probability.

Lemma 74 *It holds that $\mathbb{P}[G_1 \cap G_2 \cap G_3] \geq 1 - \delta/2$.*

Proof By Lemmas 71, 72 and 73 when combined using an union bound. ■

D.2.3 ANALYSIS FOR THE ℓ_2 LOSS

Lemma 75 *Assume the good events G_1, G_2 and G_3 hold. Then for every context-dependent policy $\pi = (\pi_c)_{c\in\rfloor}$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \frac{\epsilon}{2} + \alpha_2 H,$$

where $\alpha_2^2 = \max_{(s_h,a_h)\in\cup_{h\in[H]} S^{\gamma,\beta}_h \times A} \alpha_2^2(\mathcal{F}^R_{s_h,a_h})$, for the following parameters choice: $\gamma = \frac{\epsilon}{8|S|H}$, $\beta = \frac{\epsilon}{8|S|}$ and $\epsilon_1 = \frac{\epsilon^3}{8^3|S||A|H^3}$.

Proof For all $h \in [H-1]$ and any context $c \in \mathcal{C}$, let us define the following subsets of S_h.

1. $B_1^{h,c} = \{s_h \in S_h : s_h \in S_h^{\beta,\gamma}, c \in \mathcal{C}^\beta(s_h)\}$.
2. $B_2^{h,c} = \{s_h \in S_h : s_h \in S_h^{\beta,\gamma}, c \notin \mathcal{C}^\beta(s_h)\}$.
3. $B_3^{h,c} = \{s_h \in S_h : s_h \notin S_h^{\beta,\gamma}, c \notin \mathcal{C}^\beta(s_h)\}$.
4. $B_4^{h,c} = \{s_h \in S_h : s_h \notin S_h^{\beta,\gamma}, c \in \mathcal{C}^\beta(s_h)\}$.

Clearly, $\cup_{i=1}^4 B_i^{h,c} = S_h$ for every $h \in [H-1]$ and $c \in \mathcal{C}$.

For $s_h \notin S_h^{\beta,\gamma}$ we have that $\mathbb{P}[c \in \mathcal{C}^\beta(s_h)] < \gamma$, hence,

$$\mathbb{P}_c[\exists h \in [H-1] : B_4^{h,c} \neq \emptyset] \;\; = \;\; \mathbb{P}_c[\exists h \in [H-1] \; \exists s_h \in S_h : s_h \notin S_h^{\beta,\gamma} \text{ and } c \in \mathcal{C}^\beta(s_h)] \;\; < \;\; \gamma|S|.$$

Fix a context $c \in \mathcal{C}$ and a context-dependent policy π (we will later take the expectation over the context). Consider the following derivation.

$$\begin{aligned}
|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)| &= |\sum_{h=0}^{H-1} \sum_{s_h \in S_h} \sum_{a_h \in A} q_h(s_h, a_h|\pi_c, P^c)(r^c(s_h,a_h) - \widehat{r}^c(s_h,a_h))| \\
&\leq \sum_{h=0}^{H-1} \sum_{s_h \in S_h} \sum_{a_h \in A} q_h(s_h, a_h|\pi_c, P^c)|r^c(s_h,a_h) - \widehat{r}^c(s_h,a_h)| \\
&= \underbrace{\sum_{h=0}^{H-1} \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h|\pi_c, P^c)|r^c(s_h,a_h) - \widehat{r}^c(s_h,a_h)|}_{(1)} \\
&+ \underbrace{\sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h|\pi_c, P^c)|r^c(s_h,a_h) - \widehat{r}^c(s_h,a_h)|}_{(2)} \\
&+ \underbrace{\sum_{h=0}^{H-1} \sum_{s_h \in B_4^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h|\pi_c, P^c)|r^c(s_h,a_h) - \widehat{r}^c(s_h,a_h)|}_{(3)}
\end{aligned}$$

We bound (1), (2) and (3) separately. For (1), under the good event G_3 we have for every layer $h \in [H-1]$, state $s_h \in S_h^{\gamma,\beta}$ and action $a_h \in A$ that

$$\mathbb{E}_{c\sim\mathcal{D}}\left[(f_{s_h,a_h}(c) - r^c(s_h,a_h))^2 - \alpha_2^2(\mathcal{F}^R_{s_h,a_h}) \,\middle|\, c \in \mathcal{C}^\beta(s_h)\right] \leq \epsilon_1.$$

Since $\mathbb{E}_{c\sim\mathcal{D}}[(f_{s_h,a_h}(c) - r^c(s_h,a_h))^2 | c \in \mathcal{C}^\beta(s_h)] \geq \alpha_2^2(\mathcal{F}^R_{s_h,a_h})$, for every layer $h \in [H-1]$, state $s_h \in S_h^{\gamma,\beta}$ and action $a_h \in A$, for a fixed constant $\rho \in [0,1]$ we obtain using Markov's inequality that

$$\begin{aligned}
&\mathbb{P}_c[|f_{s_h,a_h}(c) - r^c(s_h,a_h)| \geq \sqrt{\alpha_2^2(\mathcal{F}^R_{s_h,a_h}) + \rho} \,\middle|\, c \in \mathcal{C}^\beta(s_h)] = \\
=&\mathbb{P}_c[(f_{s_h,a_h}(c) - r^c(s_h,a_h))^2 - \alpha_2^2(\mathcal{F}^R_{s_h,a_h}) \geq \rho \,\middle|\, c \in \mathcal{C}^\beta(s_h)] \leq \frac{\epsilon_1}{\rho},
\end{aligned}$$

which using the following inequality (that holds since $\alpha_2(\mathcal{F}^R_{s_h,a_h}), \rho \in [0,1]$)

$$\sqrt{\alpha_2^2(\mathcal{F}^R_{s_h,a_h}) + \rho} \leq \alpha_2(\mathcal{F}^R_{s_h,a_h}) + \sqrt{\rho},$$

yielding that

$$\begin{aligned}
&\mathbb{P}_c[|f_{s_h,a_h}(c) - r^c(s_h,a_h)| \leq \alpha_2(\mathcal{F}^R_{s_h,a_h}) + \sqrt{\rho} | c \in \mathcal{C}^\beta(s_h)] \\
&\geq \mathbb{P}_c[|f_{s_h,a_h}(c) - r^c(s_h,a_h)| \geq \sqrt{\alpha_2^2(\mathcal{F}^R_{s_h,a_h}) + \rho} | c \in \mathcal{C}^\beta(s_h)] \\
&= \mathbb{P}_c[(f_{s_h,a_h}(c) - r^c(s_h,a_h))^2 - \alpha_2^2(\mathcal{F}^R_{s_h,a_h}) \geq \rho | c \in \mathcal{C}^\beta(s_h)] \\
&\geq 1 - \frac{\epsilon_1}{\rho}.
\end{aligned} \tag{9}$$

Let G_4 denote the following good event,

$$\forall h \in [H]\ \forall s_h \in B_1^{h,c}\ \forall a \in A : |f_{s_h,a_h}(c) - r^c(s_h,a_h)| \leq \alpha_2(\mathcal{F}^R_{s_h,a_h}) + \sqrt{\rho},$$

and denote by $\overline{G_4}$ the complementary event.

By $B_1^{h,c}$ definition, we have for all $h \in [H-1]$ and $s \in S_h$ that $c \in \mathcal{C}^\beta(s)$. Hence, when combining that with inequality 9 we obtain

$$\mathbb{P}_c[G_4] \geq 1 - \frac{\epsilon_1}{\rho}|S||A| \quad \text{and} \quad \mathbb{P}_c[\overline{G_4}] < \frac{\epsilon_1}{\rho}|S||A|.$$

When G_4 holds, then

$$\begin{aligned}
(1) &= \sum_{h=0}^{H-1} \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h)|r^c(s_h,a_h) - \widehat{r}^c(s_h,a_h)| \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h) \underbrace{|f_{s_h,a_h}(c) - r^c(s_h,a_h)|}_{\leq \alpha_2(\mathcal{F}^R_{s_h,a_h}) + \sqrt{\rho}} \\
&\leq \sum_{h=0}^{H-1} \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h) \underbrace{\alpha_2(\mathcal{F}^R_{s_h,a_h})}_{\leq \alpha_2} + \sqrt{\rho}H \leq \alpha_2 H + \sqrt{\rho}H.
\end{aligned}$$

Otherwise, when G_4 does not hold, then it is bounded by H.

Thus, by total expectation low we have

$$\mathbb{E}_{c\sim\mathcal{D}}[(1)] \leq \underbrace{\mathbb{E}_{c\sim\mathcal{D}}[(1)|G_4]}_{\leq \alpha_2 H + \sqrt{\rho}H} + \underbrace{\mathbb{P}[\overline{G_4}]}_{\leq \frac{\epsilon_1}{\rho}|S||A|} \cdot H \leq \alpha_2 H + \sqrt{\rho}H + \frac{\epsilon_1}{\rho}|S||A|H.$$

For (2), we have $c \notin C^\beta(s)$ for every $s \in \cup_{h\in[H-1]}(B_2^{h,c} \cup B_3^{h,c})$, which implies

$$\begin{aligned}
(2) &= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h) \underbrace{|r^c(s_h,a_h) - \widehat{r}^c(s_h,a_h)|}_{\leq 1} \\
&\leq \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h) \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} q_h(s_h|\pi_c, P^c) \underbrace{\sum_{a_h \in A} \pi(a_h|s_h)}_{=1}
\end{aligned}$$

$$= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \underbrace{q_h(s_h|\pi_c, P^c)}_{\leq \beta}$$
$$\leq \beta |S|.$$

Thus,

$$\mathbb{E}_{c\sim\mathcal{D}}[(2)] \leq \beta |S|.$$

For (3), when there exists $h \in [H-1]$ such that $B_4^{h,c} \neq \emptyset$, we have

$$(3) = \sum_{h=0}^{H-1} \sum_{s_h \in B_4^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h) \underbrace{|r^c(s_h,a_h) - \widehat{r}^c(s_h,a_h)|}_{\leq 1}$$
$$\leq \sum_{h=0}^{H-1} \underbrace{\sum_{s_h \in B_4^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h)}_{\leq 1} \leq H.$$

Let G_5 denote the good event in which $\forall h \in [H-1], B_4^{h,c} = \emptyset$. Denote by $\overline{G_5}$ the complement event of G_5. We showed that $\mathbb{P}_c[G_5] \geq 1 - \gamma|S|$ and $\mathbb{P}_c[\overline{G_5}] < \gamma|S|$.

Using total expectation we obtain

$$\mathbb{E}_{c\sim\mathcal{D}}[(3)] = \mathbb{P}_c[G_5] \cdot \mathbb{E}_{c\sim\mathcal{D}}[(3)|G_5] + \mathbb{P}_c[\overline{G_5}] \cdot \mathbb{E}_{c\sim\mathcal{D}}[(3)|\overline{G_5}]$$
$$\leq 1 \cdot 0 + \gamma|S|H = \gamma|S|H.$$

Overall, by linearity of expectation and the above we obtain

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \mathbb{E}_{c\sim\mathcal{D}}[(1)] + \mathbb{E}_{c\sim\mathcal{D}}[(2)] + \mathbb{E}_{c\sim\mathcal{D}}[(3)]$$
$$\leq \alpha_2 H + \sqrt{\rho}H + \frac{\epsilon_1}{\rho}|S||A|H + \beta|S| + \gamma|S|H.$$

Now, for $\gamma = \frac{\epsilon}{8|S|H}$, $\beta = \frac{\epsilon}{8|S|}$, $\rho = (\epsilon_1|S||A|)^{2/3}$ and $\epsilon_1 = \frac{\epsilon^3}{8^3|S||A|H^3}$ we have

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \alpha_2 H + (\epsilon_1|S||A|)^{1/3}H + \frac{\epsilon_1}{(\epsilon_1|S||A|)^{2/3}}|S||A|H + \frac{2}{8}\epsilon$$
$$= \alpha_2 H + (\frac{\epsilon^3}{8^3|S||A|H^3}|S||A|)^{1/3}H + \frac{\epsilon_1^{\frac{1}{3}}}{(|S||A|)^{2/3}}|S||A|H + \frac{2}{8}\epsilon$$
$$= \alpha_2 H + \frac{\epsilon}{8} + \frac{(\frac{\epsilon^3}{8^3|S||A|H^3})^{\frac{1}{3}}}{(|S||A|)^{2/3}}|S||A|H + \frac{2}{8}\epsilon$$
$$= \frac{\epsilon}{2} + \alpha_2 H.$$

The above inequality completes the proof of the lemma. ■

Theorem 76 *With probability at least $1-\delta$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha_2 H,$$

where $\pi^\star = (\pi_c^\star)_{c\in\mathcal{C}}$ is the optimal context-dependent policy for $\mathcal{M}$ and $\widehat{\pi}^\star = (\widehat{\pi}_c^\star)_{c\in\mathcal{C}}$ is the optimal context-dependent policy for $\widehat{\mathcal{M}}$.

Proof Assume the good events G_1, G_2 and G_3 hold. Then, by Lemma 75 we have for $\pi^\star$ that

$$\left|\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)]\right| \leq \mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \frac{\epsilon}{2} + \alpha_2 H,$$

yielding,

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] \leq \frac{\epsilon}{2} + \alpha_2 H.$$

Similarly we have for $\widehat{\pi}^\star$ that,

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \frac{\epsilon}{2} + \alpha_2 H.$$

Since for all $c \in \mathcal{C}$, $\widehat{\pi}^\star_c$ is an optimal policy for $\widehat{\mathcal{M}}(c)$ we have $V^{\widehat{\pi}^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0) \geq V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)$ which implies that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] \leq 0.$$

By Lemma 74 we have that $\mathbb{P}[G_1 \cap G_2 \cap G_3] \geq 1 - \delta/2$, hence the theorem implied by summing the above three inequalities . ■

Corollary 77 *When $\alpha_2 = 0$, with probability at least $1-\delta$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon.$$

where $\pi^\star_c$ is the optimal policy for $\mathcal{M}$ and $\widehat{\pi}^\star_c$ is the optimal policy for $\widehat{\mathcal{M}}$.

D.2.4 ANALYSIS FOR THE ℓ_1 LOSS.

Lemma 78 *Assume the good events G_1, G_2 and G_3 hold. Then we have for every context-dependent policy π that*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \frac{\epsilon}{2} + \alpha_1 H,$$

where $\alpha_1 = \max_{(s_h,a_h)\in\cup_{h\in[H]} S_h^{\gamma,\beta}\times A} \alpha_1(\mathcal{F}^R_{s_h,a_h})$, for the parameters choice $\gamma = \frac{\epsilon}{8|S|H}$, $\beta = \frac{\epsilon}{8|S|}$ and $\epsilon_1 = \frac{\epsilon^2}{64|S||A|H^2}$.

Proof

For all $h \in [H-1]$ and any context $c \in \mathcal{C}$, let us define the following subsets of S_h.

1. $B_1^{h,c} = \{s_h \in S_h : s_h \in S_h^{\beta,\gamma}, c \in \mathcal{C}^\beta(s_h)\}$.
2. $B_2^{h,c} = \{s_h \in S_h : s_h \in S_h^{\beta,\gamma}, c \notin \mathcal{C}^\beta(s_h)\}$.
3. $B_3^{h,c} = \{s_h \in S_h : s_h \notin S_h^{\beta,\gamma}, c \notin \mathcal{C}^\beta(s_h)\}$.
4. $B_4^{h,c} = \{s_h \in S_h : s_h \notin S_h^{\beta,\gamma}, c \in \mathcal{C}^\beta(s_h)\}$.

Clearly, $\cup_{i=1}^4 B_i^{h,c} = S_h$ for every $h \in [H-1]$ and $c \in \mathcal{C}$.

For $s_h \notin S_h^{\beta,\gamma}$ we have that $\mathbb{P}_c[c \in \mathcal{C}^\beta(s_h)] < \gamma$, hence,

$$\mathbb{P}_c[\exists h \in [H-1] : B_4^{h,c} \neq \emptyset] \;=\; \mathbb{P}_c[\exists h \in [H-1], s_h \in S_h : s_h \notin S_h^{\beta,\gamma}, \text{ and } c \in \mathcal{C}^\beta(s_h)] \;<\; \gamma|S|.$$

Fix a context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$. We have for any given context c (later we will take the expectation over c) the following

$$\begin{aligned}
|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)| &= |\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,P^c)(r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h))| \\
&\leq \sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,P^c)|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)| \\
&= \underbrace{\sum_{h=0}^{H-1}\sum_{s_h\in B_1^{h,c}}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,P^c)|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)|}_{(1)} \\
&+ \underbrace{\sum_{h=0}^{H-1}\sum_{s_h\in B_2^{h,c}\cup B_3^{h,c}}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,P^c)|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)|}_{(2)} \\
&+ \underbrace{\sum_{h=0}^{H-1}\sum_{s_h\in B_4^{h,c}}\sum_{a_h\in A} q^c_h(s_h,a_h|\pi)|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)|}_{(3)}
\end{aligned}$$

We bound (1), (2) and (3) separately.

For (1), under the good event G_3 for every layer $h\in[H-1]$, state $s_h\in S_h^{\gamma,\beta}$ and action $a_h\in A$ it holds that

$$\mathbb{E}_{c\sim\mathcal{D}}\Big[|f_{s_h,a_h}(c)-r^c(s_h,a_h)|-\alpha_1(\mathcal{F}^R_{s_h,a_h})\Big|\ c\in\mathcal{C}^\beta(s_h)\Big]\leq\epsilon_1.$$

Recall that $\mathbb{E}_{c\sim\mathcal{D}}\Big[|f_{s_h,a_h}(c)-r^c(s_h,a_h)|\ \Big|\ c\in\mathcal{C}^\beta(s_h)\Big]\geq\alpha_1(\mathcal{F}^R_{s_h,a_h})$. Hence, for every layer $h\in[H-1]$, state $s_h\in S_h^{\gamma,\beta}$ and an action $a_h\in A$, for a fixed constant $\rho\in[0,1]$ we obtain using Markov's inequality that

$$\begin{aligned}
&\mathbb{P}_c\Big[|f_{s_h,a_h}(c)-r^c(s_h,a_h)|\geq\alpha_1(\mathcal{F}^R_{s_h,a_h})+\rho\Big|\ c\in\mathcal{C}^\beta(s_h)\Big] \\
=&\mathbb{P}_c\Big[|f_{s_h,a_h}(c)-r^c(s_h,a_h)|-\alpha_1(\mathcal{F}^R_{s_h,a_h})\geq\rho\Big|\ c\in\mathcal{C}^\beta(s_h)\Big]\leq\frac{\epsilon_1}{\rho},
\end{aligned}$$

which implies that

$$\mathbb{P}_c\Big[|f_{s_h,a_h}(c)-r^c(s_h,a_h)|\leq\alpha_1(\mathcal{F}^R_{s_h,a_h})+\rho\Big|\ c\in\mathcal{C}^\beta(s_h)\Big]\geq 1-\frac{\epsilon_1}{\rho}. \tag{10}$$

Let G_4 denote the following good event,

$$\forall h\in[H-1]\ \forall s_h\in B_1^{h,c}\ \forall a\in A: |f_{s_h,a_h}(c)-r^c(s_h,a_h)|\leq\alpha_1(\mathcal{F}^R_{s_h,a_h})+\rho,$$

and denote by $\overline{G_4}$ the complementary event. By $B_1^{h,c}$ definition, we have for all $h\in[H-1]$ and $s\in S_h$ that $c\in\mathcal{C}^\beta(s)$. Hence, when combining that with inequality 10 we obtain

$$\mathbb{P}_c[G_4]\geq 1-\frac{\epsilon_1}{\rho}|S||A| \quad\text{and}\quad \mathbb{P}_c[\overline{G_4}]<\frac{\epsilon_1}{\rho}|S||A|.$$

When G_4 holds, then

$$(1)=\sum_{h=0}^{H-1}\sum_{s_h\in B_1^{h,c}}\sum_{a_h\in A} q_h(s_h|\pi_c,P^c)\pi_c(a_h|s_h)|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)|$$

$$= \sum_{h=0}^{H-1} \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h) \underbrace{|f_{s_h,a_h}(c) - r^c(s_h,a_h)|}_{\leq \alpha_1(\mathcal{F}^R_{s_h,a_h})+\rho}$$

$$\leq \sum_{h=0}^{H-1} \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h) \underbrace{\alpha_1(\mathcal{F}^R_{s_h,a_h})}_{\leq \alpha_1} + \rho H \leq \alpha_1 H + \rho H.$$

Otherwise, when G_4 does not hold, then it is bounded by H.

Thus,

$$\mathbb{E}_{c\sim\mathcal{D}}[(1)] \leq \underbrace{\mathbb{E}_{c\sim\mathcal{D}}[(1)|G_4]}_{\leq \alpha_1 H + \rho H} + \underbrace{\mathbb{P}[\overline{G_4}]}_{\leq \frac{\epsilon_1}{\rho}|S||A|} \cdot H$$

$$\leq \alpha_1 H + \rho H + \frac{\epsilon_1}{\rho}|S||A|H.$$

For (2), we have $c \notin C^\beta(s)$ for every $s \in \cup_{h\in[H-1]}(B_2^{h,c} \cup B_3^{h,c})$, which implies

$$(2) = \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c}\cup B_3^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h) \underbrace{|r^c(s_h,a_h) - \widehat{r}^c(s_h,a_h)|}_{\leq 1}$$

$$\leq \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c}\cup B_3^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h)$$

$$= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c}\cup B_3^{h,c}} q_h(s_h|\pi_c, P^c) \underbrace{\sum_{a_h \in A} \pi(a_h|s_h)}_{=1}$$

$$= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c}\cup B_3^{h,c}} \underbrace{q_h(s_h|\pi_c, P^c)}_{\leq \beta} \leq \beta|S|.$$

Thus,

$$\mathbb{E}_{c\sim\mathcal{D}}[(2)] \leq \beta|S|.$$

For (3), when there exists $h \in [H-1]$ such that $B_4^{h,c} \neq \emptyset$, we have

$$(3) = \sum_{h=0}^{H-1} \sum_{s_h \in B_4^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h) \underbrace{|r^c(s_h,a_h) - \widehat{r}^c(s_h,a_h)|}_{\leq 1}$$

$$\leq \sum_{h=0}^{H-1} \underbrace{\sum_{s_h \in B_4^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h)}_{\leq 1} \leq H.$$

Let G_5 denote the good event in which $\forall h \in [H-1], B_4^{h,c} = \emptyset$. Denote by $\overline{G_5}$ the complement event of G_5. We showed that $\mathbb{P}_c[G_5] \geq 1 - \gamma|S|$ and $\mathbb{P}_c[\overline{G_5}] < \gamma|S|$.

Using total expectation we obtain

$$\mathbb{E}_{c\sim\mathcal{D}}[(3)] = \mathbb{P}_c[G_5] \cdot \mathbb{E}_{c\sim\mathcal{D}}[(3)|G_5] + \mathbb{P}_c[\overline{G_5}] \cdot \mathbb{E}_{c\sim\mathcal{D}}[(3)|\overline{G_5}]$$

$$\leq 1 \cdot 0 + \gamma|S|H = \gamma|S|H.$$

Overall, by linearity of expectation and the above we obtain

$$\begin{aligned}\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] &\leq \mathbb{E}_{c\sim\mathcal{D}}[(1)] + \mathbb{E}_{c\sim\mathcal{D}}[(2)] + \mathbb{E}_{c\sim\mathcal{D}}[(3)] \\ &\leq \alpha_1 H + \rho H + \frac{\epsilon_1}{\rho}|S||A|H + \beta|S| + \gamma|S|H.\end{aligned}$$

Finally, for the parameters choice $\gamma = \frac{\epsilon}{8|S|H}$, $\beta = \frac{\epsilon}{8|S|}$, $\rho = (\epsilon_1|S||A|)^{1/2}$ and $\epsilon_1 = \frac{\epsilon^2}{64|S||A|H^2}$ it holds that

$$\begin{aligned}\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] &\leq \alpha_1 H + 2(\epsilon_1|S||A|)^{1/2}H + \frac{2}{8}\epsilon \\ &= \alpha_1 H + 2(\frac{\epsilon^2}{64|S||A|H^2}|S||A|)^{1/2}H + \frac{2}{8}\epsilon \\ &= \alpha_1 H + 2\frac{\epsilon}{4} \\ &= \frac{\epsilon}{2} + \alpha_1 H.\end{aligned}$$

The above inequality completes the proof of the lemma. ■

Theorem 79 *With probability at least $1-\delta$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha_1 H,$$

where $\pi^\star = (\pi^\star_c)_{c\in\mathcal{C}}$ is an optimal context-dependent policy for $\mathcal{M}$ and $\widehat{\pi}^\star = (\widehat{\pi}^\star_c)_{c\in\mathcal{C}}$ is an optimal context-dependent policy for $\widehat{\mathcal{M}}$.

Proof Assume the good events G_1, G_2 and G_3 hold. Then, by Lemma 78 we have for $\pi^\star$

$$\left|\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)]\right| \leq \mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \frac{\epsilon}{2} + \alpha_1 H,$$

yielding

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] \leq \frac{\epsilon}{2} + \alpha_1 H.$$

Similarly, we have for $\widehat{\pi}^\star$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \frac{\epsilon}{2} + \alpha_1 H.$$

Since for all $c\in\mathcal{C}$, $\widehat{\pi}^\star_c$ is the optimal policy for $\widehat{\mathcal{M}}(c)$ we have $V^{\widehat{\pi}^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0) \geq V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)$ which implies that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}^\star_c}_{\widehat{\mathcal{M}}(c)}] \leq 0.$$

By Lemma 74 we have that $\mathbb{P}[G_1\cap G_2\cap G_3] \geq 1-\delta/2$, hence the theorem implied by summing the above three inequalities. ■

Corollary 80 *When $\alpha_1 = 0$, with probability at least $1-\delta$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon,$$

where $\pi^\star_c$ is the optimal policy for $\mathcal{M}$ and $\widehat{\pi}^\star_c$ is the optimal policy for $\widehat{\mathcal{M}}$.

D.3 Sample complexity bounds.

We show dimension-based sample complexity bounds for both ℓ_1 and ℓ_2 loss functions.

Recall Theorems 28 and 29,

Theorem 81 (Adaption of Theorem 19.2 in Anthony et al. (1999)) *Let $\mathcal{F}$ be a hypothesis space of real valued functions with a finite pseudo dimension, denoted $Pdim(\mathcal{F}) < \infty$. Then, $\mathcal{F}$ has a uniform convergence with*

$$m(\epsilon,\delta) = O\Big(\frac{1}{\epsilon^2}(Pdim(\mathcal{F})\ln\frac{1}{\epsilon} + \ln\frac{1}{\delta})\Big).$$

Theorem 82 (Adaption of Theorem 19.1 in Anthony et al. (1999)) *Let $\mathcal{F}$ be a hypothesis space of real valued functions with a finite fat-shattering dimension, denoted $fat_{\mathcal{F}}$. Then, $\mathcal{F}$ has a uniform convergence with*

$$m(\epsilon,\delta) = O\Big(\frac{1}{\epsilon^2}(fat_{\mathcal{F}}(\epsilon/256)\ln^2\frac{1}{\epsilon} + \ln\frac{1}{\delta})\Big).$$

Remark 83 *In the following analysis we omit the sample complexity needed to approximate the faction of good contexts for every $s \in S$ as it is*

$$O\Big(\frac{|S|^3H^2\ln\frac{|S|}{\delta}}{\epsilon^2}\Big)$$

and is negligible additional term in the following analysis.

D.3.1 SAMPLE COMPLEXITY BOUNDS FOR THE ℓ_2 LOSS.

We show sample complexity bounds for function classes with finite Pseudo dimension with ℓ_2 loss.

Corollary 84 *Assume that for every $(s,a) \in S\times A$ we have that $Pdim(\mathcal{F}^R_{s,a}) < \infty$. Let $Pdim = \max_{(s,a)\in S\times A} Pdim(\mathcal{F}^R_{s,a})$. Then, after collecting*

$$O\Big(\frac{|S|^5|A|^3H^7}{\epsilon^8}\Big(Pdim\ln\frac{|S||A|H^3}{\epsilon^3} + \ln\frac{|S||A|}{\delta}\Big)\Big).$$

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

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

Proof In the worst-case for every layer $h \in [H-1]$ and every state-action pair (s,a) we collect

$$T_{s,a} = \lceil\frac{2}{\beta\gamma}(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}^R_{s_h,a_h},\epsilon_1,\delta_1))\rceil$$

trajectories. By Theorem 76, for $\gamma = \frac{\epsilon}{8|S|H}$, $\beta = \frac{\epsilon}{8|S|}$, $\epsilon_1 = \frac{\epsilon^3}{8^3|S||A|H^3}$, $\delta_1 = \frac{\delta}{6|S||A|}$ and $\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}T_{s,a}$ examples with probability at least $1-\delta$ it holds that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

Since for every $(s,a) \in S\times A$ we have that $Pdim(\mathcal{F}^R_{s,a}) < \infty$, and $Pdim = \max_{(s,a)\in S\times A} Pdim(\mathcal{F}^R_{s,a})$, by Theorem 28, for every $(s,a) \in S\times A$ we have

$$N_R(\mathcal{F}^R_{s,a},\epsilon_1,\delta_1) = O\Big(\frac{Pdim\ln\frac{1}{\epsilon_1} + \ln\frac{1}{\delta_1}}{\epsilon_1^2}\Big) = O\Big(\frac{|S|^2|A|^2H^6}{\epsilon^6}\Big(Pdim\ln\frac{|S||A|H^3}{\epsilon^3} + \ln\frac{|S||A|}{\delta}\Big)\Big)$$

Hence, for each state-action pair (s,a) we have

$$\begin{aligned}T_{s,a} =&O\Big(\frac{|S|}{\epsilon}\frac{|S|H}{\epsilon}\frac{|S|^2|A|^2H^6}{\epsilon^6}\Big(Pdim\ln\frac{|S||A|H^3}{\epsilon^3}+\ln\frac{|S||A|}{\delta}\Big)\Big)\\ =&O\Big(\frac{|S|^4|A|^2H^7}{\epsilon^8}\Big(Pdim\ln\frac{|S||A|H^3}{\epsilon^3}+\ln\frac{|S||A|}{\delta}\Big)\Big).\end{aligned}$$

When summing the above for every state-action pair we obtain that the overall sample complexity is

$$O\Big(\frac{|S|^5|A|^3H^7}{\epsilon^8}\Big(Pdim\ln\frac{|S||A|H^3}{\epsilon^3}+\ln\frac{|S||A|}{\delta}\Big)\Big).$$

■

We also show sample complexity bounds for function classes with finite fat-shattering dimension when using ℓ_2 loss.

Corollary 85 *Assume that for every $(s,a)\in S\times A$ we have that $\mathcal{F}^R_{s,a}$ has finite fat-shattering dimension. Let $Fdim=\max_{(s,a)\in S\times A} fat_{\mathcal{F}^R_{s,a}}(\epsilon_1/256)$ for $\epsilon_1=\frac{\epsilon^3}{8^3|S||A|H^3}$. Then, after collecting*

$$O\Big(\frac{|S|^5|A|^3H^7}{\epsilon^8}\Big(Fdim\ln^2\frac{|S||A|H^3}{\epsilon^3}+\ln\frac{|S||A|}{\delta}\Big)\Big).$$

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

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0)-V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)]\le\epsilon+2\alpha_2H.$$

Proof In the worst-case, for every layer $h\in[H-1]$ and every state-action pair $(s,a)\in S_h\times A$ we collect

$$T_{s,a}=\lceil\frac{2}{\beta\gamma}(\ln(\frac{1}{\delta_1})+N_R(\mathcal{F}^R_{s_h,a_h},\epsilon_1,\delta_1))\rceil$$

trajectories. By Theorem 76, for $\gamma=\frac{\epsilon}{8|S|H}$, $\beta=\frac{\epsilon}{8|S|}$, $\epsilon_1=\frac{\epsilon^3}{8^3|S||A|H^3}$, $\delta_1=\frac{\delta}{6|S||A|}$ and $\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A}T_{s,a}$ examples we have with probability at least $1-\delta$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0)-V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)]\le\epsilon+2\alpha_2H.$$

Since for every $(s,a)\in S\times A$ we have that $\mathcal{F}^R_{s,a}$ has finite fat-shattering dimension, and $Fdim=\max_{(s,a)\in S\times A} fat_{\mathcal{F}^R_{s,a}}(\epsilon_1/256)$ for $\epsilon_1=\frac{\epsilon^3}{8^3|S||A|H^3}$, by Theorem 29, for every $(s,a)\in S\times A$ we have

$$N_R(\mathcal{F}^R_{s,a},\epsilon_1,\delta_1)=O\Big(\frac{Fdim\ln^2\frac{1}{\epsilon_1}+\ln\frac{1}{\delta_1}}{\epsilon_1^2}\Big)=O\Big(\frac{|S|^2|A|^2H^6}{\epsilon^6}\Big(Fdim\ln^2\frac{|S||A|H^3}{\epsilon^3}+\ln\frac{|S||A|}{\delta}\Big)\Big)$$

Hence, for each state-action pair (s,a) we have

$$\begin{aligned}T_{s,a} =&O\Big(\frac{|S|}{\epsilon}\frac{|S|H}{\epsilon}\frac{|S|^2|A|^2H^6}{\epsilon^6}\Big(Fdim\ln^2\frac{|S||A|H^3}{\epsilon^3}+\ln\frac{|S||A|}{\delta}\Big)\Big)\\ =&O\Big(\frac{|S|^4|A|^2H^7}{\epsilon^8}\Big(Fdim\ln^2\frac{|S||A|H^3}{\epsilon^3}+\ln\frac{|S||A|}{\delta}\Big)\Big).\end{aligned}$$

When summing the above for every state-action pair we obtain that the overall sample complexity is

$$O\Big(\frac{|S|^5|A|^3H^7}{\epsilon^8}\Big(Fdim\ln^2\frac{|S||A|H^3}{\epsilon^3}+\ln\frac{|S||A|}{\delta}\Big)\Big).$$

■

D.3.2 SAMPLE COMPLEXITY BOUNDS FOR THE ℓ_1 LOSS.

We present sample complexity bounds for function classes with finite Pseudo dimension with ℓ_1 loss.

Corollary 86 *Assume that for every $(s,a) \in S \times A$ we have that $Pdim(\mathcal{F}^R_{s,a}) < \infty$. Let $Pdim = \max_{(s,a)\in S\times A} Pdim(\mathcal{F}^R_{s,a})$. Then, after collecting*

$$O\Big(\frac{|S|^5|A|^3H^5}{\epsilon^6}\Big(Pdim \ln\frac{|S||A|H^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta}\Big)\Big).$$

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

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha_1 H.$$

Proof In the worst-case, for every layer $h \in [H-1]$ and every state-action pair (s,a) we collect

$$T_{s,a} = \lceil \frac{2}{\beta\gamma}(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon_1, \delta_1))\rceil$$

trajectories. By Theorem 79, for $\gamma = \frac{\epsilon}{8|S|H}$, $\beta = \frac{\epsilon}{8|S|}$, $\epsilon_1 = \frac{\epsilon^2}{64|S||A|H^2}$, $\delta_1 = \frac{\delta}{6|S||A|}$ and $\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} T_{s,a}$ examples we have with probability at least $1-\delta$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha_1 H.$$

Since for every $(s,a) \in S \times A$ we have that $Pdim(\mathcal{F}^R_{s,a}) < \infty$, and $Pdim = \max_{(s,a)\in S\times A} Pdim(\mathcal{F}^R_{s,a})$, by Theorem 28, for every $(s,a) \in S \times A$ we have

$$N_R(\mathcal{F}^R_{s,a}, \epsilon_1, \delta_1) = O\Big(\frac{Pdim\ln\frac{1}{\epsilon_1} + \ln\frac{1}{\delta_1}}{\epsilon_1^2}\Big) = O\Big(\frac{|S|^2|A|^2H^4}{\epsilon^4}\Big(Pdim\ln\frac{|S||A|H^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta}\Big)\Big)$$

Hence, for each state-action pair (s,a) we have

$$\begin{aligned} T_{s,a} =& O\Big(\frac{|S|}{\epsilon}\frac{|S|H}{\epsilon}\frac{|S|^2|A|^2H^4}{\epsilon^4}\Big(Pdim\ln\frac{|S||A|H^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta}\Big)\Big) \\ =& O\Big(\frac{|S|^4|A|^2H^5}{\epsilon^6}\Big(Pdim\ln\frac{|S||A|H^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta}\Big).\Big) \end{aligned}$$

When summing the above for every state-action pair we obtain that the overall sample complexity is

$$O\Big(\frac{|S|^5|A|^3H^5}{\epsilon^6}\Big(Pdim\ln\frac{|S||A|H^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta}\Big)\Big).$$

■

We also show sample complexity bounds for function classes with finite fat-shattering dimension when using ℓ_1 loss.

Corollary 87 *Assume that for every $(s,a) \in S \times A$ we have that $\mathcal{F}^R_{s,a}$ has finite fat-shattering dimension. Let $Fdim = \max_{(s,a)\in S\times A} fat_{\mathcal{F}^R_{s,a}}(\epsilon_1/256)$ for $\epsilon_1 = \frac{\epsilon^2}{64|S||A|H^2}$. Then, after collecting*

$$O\Big(\frac{|S|^5|A|^3H^5}{\epsilon^6}\Big(Fdim\ln^2\frac{|S||A|H^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta}\Big)\Big).$$

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

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha_1 H.$$

Proof In the worst-case, for every layer $h \in [H-1]$ and every state-action pair (s,a) we collect

$$T_{s,a} = \lceil \frac{2}{\beta\gamma}(\ln(\frac{1}{\delta_1}) + N_R(\mathcal{F}^R_{s_h,a_h}, \epsilon_1, \delta_1))\rceil$$

trajectories. By Theorem 79, for $\gamma = \frac{\epsilon}{8|S|H}$, $\beta = \frac{\epsilon}{8|S|}$, $\epsilon_1 = \frac{\epsilon^2}{64|S||A|H^2}$, $\delta_1 = \frac{\delta}{6|S||A|}$ and $\sum_{h=0}^{H-1}\sum_{s\in S}\sum_{a\in A} T_{s,a}$ samples we have with probability at least $1-\delta$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha_1 H.$$

Since for every $(s,a) \in S\times A$ we have that $\mathcal{F}^R_{s,a}$ has finite fat-shattering dimension, and $Fdim = \max_{(s,a)\in S\times A} fat_{\mathcal{F}^R_{s,a}}(\epsilon_1/256)$ for $\epsilon_1 = \frac{\epsilon^2}{64|S||A|H^2}$, by Theorem 29, for every $(s,a) \in S \times A$ we have

$$N_R(\mathcal{F}^R_{s,a}, \epsilon_1, \delta_1) = O\Big(\frac{Fdim\ln^2\frac{1}{\epsilon_1} + \ln\frac{1}{\delta_1}}{\epsilon_1^2}\Big) = O\Big(\frac{|S|^2|A|^2H^4}{\epsilon^4}\Big(Fdim\ln^2\frac{|S||A|H^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta}\Big)\Big)$$

Hence, for each state-action pair (s,a) we have

$$\begin{aligned} T_{s,a} =& O\Big(\frac{|S|}{\epsilon}\frac{|S|H}{\epsilon}\frac{|S|^2|A|^2H^4}{\epsilon^4}\Big(Fdim\ln^2\frac{|S||A|H^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta}\Big)\Big) \\ =& O\Big(\frac{|S|^4|A|^2H^5}{\epsilon^6}\Big(Fdim\ln^2\frac{|S||A|H^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta}\Big)\Big). \end{aligned}$$

When summing the above for every state-action pair we obtain that the overall sample complexity is

$$O\Big(\frac{|S|^5|A|^3H^5}{\epsilon^6}\Big(Fdim\ln^2\frac{|S||A|H^2}{\epsilon^2} + \ln\frac{|S||A|}{\delta}\Big)\Big).$$

■

Appendix E. Unknown and Context Dependent Dynamics

In this section, we consider the most challenging case of unknown and context-dependent dynamics. Our approach requires a slight modification of our assumptions.

E.1 Modification of Assumptions

While we assume that the dynamics is context dependent, we will also assume that the partition to layers is the same for all contexts. As before, we are assuming the partition is known to the learner. (Note that the first layer is $S_0 = \{s_0\}$, namely there exist a single start state s_0 which is common for all the contexts.)

We also modify our assumption for the function approximation. We will have a function approximation per layer (and not per state-action pair). In addition we will assume that the dynamics are realizable by the function class, i.e., for each layer there is function in our class which models the dynamics correctly. For rewards we will also have a function approximation per layer, but it can be agnostic, i.e., even the best function in the class has a non-zero error.

In more details,

E.1.1 FUNCTION APPROXIMATION PER LAYER

We slightly modify our assumption for the function approximation class, which works per layer and not per state-action. For each layer $h \in [H-1]$ we have a function class for the dynamics $\mathcal{F}^P_h = \{f^P_h : \mathcal{C}\times S_h \times A \times S_{h+1} \to [0,1]\}$ and for the rewards $\mathcal{F}^R_h = \{f^R_h : \mathcal{C}\times S_h \times A \to [0,1]\}$. Intuitively, given that we are in state s perform action a and the context is c, functions $f^P_h \in \mathcal{F}^P_h$ and $f^R_h \in \mathcal{F}^R_h$, approximates the transition probability to state s', i.e., $P^c(s'|s,a)$, and the expected reward, i.e., $r^c(s,a)$, respectively.

Assumption 1 (layer dynamics realizability) *We assume the for every layer $h \in [H-1]$ there exist a function $f_h^\star \in \mathcal{F}_h^P$ for which,*

$$\forall (c, s, a, s') \in \mathcal{C} \times S_h \times A \times S_{h+1}.\ \ f_h^\star(c, s, a, s') = P^c(s'|s, a).$$

Namely, the true transition probability function of layer h is contained in $\mathcal{F}_h^P$.

Assumption 1 in particular implies that for every layer $h \in [H-1]$ it holds that $\alpha_1(\mathcal{F}_h^P) = \alpha_2^2(\mathcal{F}_h^P) = 0$ (for any distribution over (c, s, a, s')).

The functions $N_P(\mathcal{F}_h^P, \epsilon, \delta)$ and $N_R(\mathcal{F}_h^R, \epsilon, \delta)$ map a function class, required accuracy ϵ and confidence δ to the required number of samples for the ERM oracle guaranteed performance. For the dynamics the ERM guarantee is that with probability $1-\delta$, that $\mathbb{E}[\ell(f_h^P(x), y)] \leq \epsilon$. For the rewards the guarantee is that $\mathbb{E}[\ell(f_h^R(x), y)] \leq \epsilon + \alpha$, where α is the approximation error, and ℓ is the loss function.

E.1.2 REACHABILITY AND THE DOMAIN OF THE EXAMPLES

We redefine reachability with respect to the approximated context-dependent dynamics $\widehat{P}^c$.

For $\beta \in (0, 1]$ and layer $h \in [H-1]$ the *β-good contexts* of state $s_h \in S_h$ with respect to $\widehat{P}$ are

$$\widehat{\mathcal{C}}^\beta(s_h) = \{c \in \mathcal{C} : s_h \text{ is } \beta - \text{reachable for } \widehat{P}^c\}.$$

The *(γ, β)-good states* of layer $h \in [H-1]$ with respect to $\widehat{P}$ are

$$\widehat{S}_h^{\gamma,\beta} = \{s_h \in S_h : \mathbb{P}[c \in \widehat{\mathcal{C}}^\beta(s_h)] \geq \gamma\}.$$

The *target domain* we would like to collect sufficient number of examples from for each layer $h \in [H-1]$ is defined as

$$\mathcal{X}_h^{\gamma,\beta} = \{(c, s_h, a_h) : s_h \in \widehat{S}_h^{\gamma,\beta}, c \in \widehat{\mathcal{C}}^\beta(s_h), a_h \in A\}.$$

We remark that in the following algorithm, we approximate the set $\widehat{S}_h^{\gamma,\beta}$ for every layer $h \in [H-1]$. We denote the approximation by $\widetilde{S}_h^{\gamma,\beta}$. In the following analysis we show that with high probability, the set $\widetilde{S}_h^{\gamma,\beta}$ satisfies that

$$\widehat{S}_h^{\gamma,\beta} \subseteq \widetilde{S}_h^{\gamma,\beta} \subseteq \widehat{S}_h^{\gamma/2,\beta}.$$

Hence, for every layer $h \in [H-1]$, we have the *empirical domain* we collect examples from in practice, and is defined as $\widetilde{\mathcal{X}}_h^{\gamma,\beta} = \{(c, s_h, a_h) : s_h \in \widetilde{S}_h^{\gamma,\beta}, c \in \widehat{\mathcal{C}}^\beta(s_h), a_h \in A\}$. By the above, with high probability it holds that

$$\mathcal{X}_h^{\gamma,\beta} \subseteq \widetilde{\mathcal{X}}_h^{\gamma,\beta} \subseteq \mathcal{X}_h^{\gamma/2,\beta}.$$

We remark that the empirical domain $\widetilde{\mathcal{X}}_h^{\gamma,\beta}$ is determine before learning layer h, based on the approximation of the dynamics up to layer $h-1$ and the approximation of the set $\widehat{S}_h^{\gamma,\beta}$ (which is done before learning layer h).

E.2 Algorithm

Algorithm EXPLORE-UCDD (Algorithm 15) runs in H phases, one per layer. In phase $h \in [H-1]$ we maintain an approximate dynamics for all previous layers $k \leq h-1$, which we already learned. In phase h we run multiple iterations. In each iteration,
(1) we approximate the set of (γ, β)-good states for layer h. We denote the approximated set $\widetilde{S}_h^{\gamma,\beta}$.
(2) We select at random an approximately (γ, β)-good state $s_h \in \widetilde{S}_h^{\gamma,\beta}$ and an action $a_h \in A$.
(3) Given a context c and a state s_h we compute a policy $\widehat{\pi}_{s_h}^c$ which maximizes the probability of reaching state s_h under the approximated dynamics $\widehat{P}^c$.
(4) We run $\widehat{\pi}_{s_h}^c$. If it reaches s_h we play a_h, get a reward r_h and transits to s_{h+1}, we add: (a) to the dynamics data set $Sample^P(h)$: $((c, s_h, a_h, s'), \mathbb{I}[s_{h+1} = s'])$ for each $s' \in S_{h+1}$, (b) to the reward data set $Sample^R(h)$:

$((c, s_h, a_h), r_h)$.
(5) After collecting sufficient number of samples, we use the ERM oracle to (i) approximate the transition probabilities of layer h, i.e., $f_h^P = \texttt{ERM}(\mathcal{F}_h^P, Sample^P(h), \ell)$. (ii) approximate the rewards function of layer h, i.e., $f_h^R = \texttt{ERM}(\mathcal{F}_h^R, Sample^R(h), \ell)$.

Algorithm EXPLOIT-UCDD (Algorithm 16) gets as inputs the MDP parameters and the functions which approximate the rewards and the dynamics (that computed using EXPLORE-UCDD). Given a context c it computes the approximated MDP $\widehat{\mathcal{M}}(c)$ and use it to compute a the optimal policy for it, $\widehat{\pi}_c^\star$. Then, it run $\widehat{\pi}_c^\star$ to generate trajectory. Recall that $\widehat{\mathcal{M}}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}^c, s_0, \widehat{r}^c, H)$, where $\widehat{P}^c$, $\widehat{r}^c$ defined in algorithm EXPLOIT-UCDD.

Algorithm 12 Approximate Context-Dependent Dynamics (ACDD)

inputs:

- MDP parameters: $S = \{S_0, S_1, \ldots, S_H\}$, A, H.
- Layer $h \in [H-1]$ and approximation of the dynamics for every layer $l < h : f_l^P$.
- Reachability parameter β.
- The approximation of the sets of (γ, β)-good states $\widetilde{S}_k^{\gamma,\beta}$ for all $k \in [h-1]$

for a new state $s_{sink} \notin S$, define the approximated context-dependent dynamics as follows.

$$\forall (s,a) \in S \cup \{s_{sink}\} \times A : \widehat{P}^c(s|s_{sink}, a) = \mathbb{I}[s = s_{sink}]$$
$$\forall k \in [h-1], \quad (s_k, a_k, s_{k+1}) \in \widetilde{S}_k^{\gamma,\beta} \times A \times S_{k+1} :$$
$$\widehat{P}^c(s_{k+1}|s_k, a_k) = \mathbb{I}[c \in \widehat{C}^\beta(s_k)] \cdot \frac{f_k^P(c, s_k, a_k, s_{k+1})}{\sum_{s'_{k+1} \in S_{k+1}} f_k^P(c, s_k, a_k, s'_{k+1})}$$
$$\widehat{P}^c(s_{sink}|s_k, a_k) = \mathbb{I}[c \notin \widehat{C}^\beta(s_k)]$$
$$\forall k \in [h-1], (s_k, a_k, s_{k+1}) \in (S_k \setminus \widetilde{S}_k^{\gamma,\beta}) \times A \times S_{k+1} :$$
$$\widehat{P}^c(s_{k+1}|s_k, a_k) = 0, \widehat{P}^c(s_{sink}|s_k, a_k) = 1.$$

return $\widehat{P}^c$ ▷ Note that $\widehat{P}^c$ is a function of the context c.

Algorithm 13 Approximate Good States (AGS)

inputs:

- MDP parameters: $S = \{S_0, S_1, \ldots, S_H\}$, A, H.
- layer $h \in [H-1]$ and approximation of the dynamics for every layer $l < h : f_l^P$.
- Reachability parameters γ, β
- ϵ_2, δ_2 - accuracy and confidence.

set $\widetilde{S}_h^{\gamma,\beta} = \emptyset$
$\widehat{P}^c \leftarrow \texttt{ACDD}(S, A, H, h, \beta, \{f_k^P, \widetilde{S}_k^{\gamma,\beta} | k \in [h-1]\})$
for $s_h \in S_h$ **do**
  $I, p \leftarrow \texttt{AGC}(S, A, H, \widehat{P}^c, \delta_2, \epsilon_2, \gamma, \beta, h, s_h)$
  if $I == 1$ **then**
    $\widetilde{S}_h^{\gamma,\beta} \leftarrow \widetilde{S}_h^{\gamma,\beta} \cup \{s_h\}$
return $\widetilde{S}_h^{\gamma,\beta}$

Remark 88 *In the following algorithms, for a given context c and state $s \in S_h$, the check whether $c \in \widehat{C}^\beta(s)$ can be done in polynomial time in $|S|, |A|, H$ by computing the highest probability to visit s by any policy on the dynamics $\widehat{P}^c$.*

Algorithm 14 Approximate Good Contexts for UCDD (AGC)

inputs:
- MDP parameters: $S = \{S_0, S_1, \ldots, S_H\}$, A, H.
- P^c - The context-dependent dynamics.
- Reachability parameters: γ ,β
- Accuracy and confidence parameters ϵ_2, δ_2
- Current layer h and state s_h.

calculate $m(\epsilon_2, \delta_2) = \left\lceil \frac{\ln \frac{2}{\delta_2}}{2\epsilon_2^2} \right\rceil$
initialize $counter = 0$
for $t = 1, 2, ..., m(\epsilon_2, \delta_2)$ **do**
    observe context c_t
    if $c_t \in \mathcal{C}^{\beta}(s_h)$ **then**
        $Counter = Counter - 1$
$\widehat{p}_{\beta}(s_h) = \frac{Counter}{m(\epsilon_2, \delta_2)}$
return $\mathbb{I}[\widehat{p}_{\beta}(s_h) \geq \gamma - \epsilon_2]$ and $\widehat{p}_{\beta}(s_h)$

Since the CMDP is layered, to compute that, we need $\widehat{P}^c$ to be defined only on $(s, a) \in S_l \times A$ for all $\ell < h$. In the following algorithm, $\mathbb{I}[c \in \widehat{\mathcal{C}}^{\beta}(s)]$ is an indicator function that given a context c return 1 if and only if $c \in \widehat{\mathcal{C}}^{\beta}(s)$. By the above, the computation time of that function can be done in $poly(|S|, |A|, H)$ time.

E.3 Analysis Outline

We provide analysis for both ℓ_1 and ℓ_2 loss functions. For both of them, we first bound the expected value difference caused by the dynamics approximation for every context-dependent policy, with high probability. See Sub-subsection E.5.2, Lemma 120 for the ℓ_1 loss and Sub-subsection E.4.2, Lemma 100 for the ℓ_2 loss.

Then, we bound the expected value difference caused by the rewards approximation for every context-dependent policy, with high probability. See Sub-subsection E.5.3, Lemma 122 for the ℓ_1 loss and Sub-subsection E.4.3, Lemma 102 for the ℓ_2 loss.

The next step is to combine both bounds to obtain a bound the expected value difference between the true model $\mathcal{M}(c)$ and $\widehat{\mathcal{M}}(c)$, for any context-dependent policy $\pi = (\pi_c)_{c \in \mathcal{C}}$, with high probability. See Lemma 124 for the ℓ_1 loss and Lemma 104 for the ℓ_2.

Using the latter bound, we derive a bound on the expected value difference between the optimal context-dependent policy $\pi^{\star} = (\pi_c^{\star})_{c \in \mathcal{C}}$ and our approximated optimal policy $\widehat{\pi}^{\star} = (\widehat{\pi}_c^{\star})_{c \in \mathcal{C}}$ with respect to the true model $\mathcal{M}(c)$, which holds with high probability. This establish our main result. See Theorem 126 for the ℓ_1 loss and 106) for the ℓ_2 loss.

Lastly, we derive sample complexity bounds using known uniform convergence sample complexity bounds for the Pseudo dimension see Theorem 28) and the fat-shattering dimension (see Theorem 29). For the sample complexity analysis, see Sub-subsection E.6.2 for the ℓ_1 loss, and E.6.1 for the ℓ_2 loss.

E.4 Analysis for the ℓ_2 Loss

E.4.1 GOOD EVENTS

For the analysis of the algorithm, we define the following good events.

Event G_1. Intuitively, it states that the approximation of the probability that $c \in \widehat{\mathcal{C}}^{\beta}(s)$ is accurate for every state $s \in S$.

Algorithm 15 Explore Unknown and Context-Dependent Dynamics CMDP (EXPLORE-UCDD)

inputs:

- $S = \{S_0, S_1, \ldots, S_H\}$ - a layered states space, A - a finite actions space, s_0 - a unique start state, H - the horizon length.
- Accuracy and confidence parameters: ϵ, δ.
- $\forall h \in [H-1]: \quad \mathcal{F}_h^R, \mathcal{F}_h^P$ - the function classes use to approximate the expected reward and dynamics in layer h, respectively.
- $N_R(\mathcal{F}, \epsilon, \delta), N_P(\mathcal{F}, \epsilon, \delta)$ - sample complexity function for the ERM oracle, for the rewards and dynamics respectively.
- The reachability parameters $\gamma \in [0,1]$, $\beta \in [0,1]$.
- Loss function ℓ (assumed to be one of ℓ_1 or ℓ_2).

set $\delta_1 = \frac{\delta}{8H}, \delta_2 = \frac{\delta}{8|S|}$.

set $\epsilon_P = \begin{cases} \frac{\epsilon^3}{10 \cdot 2^8 \cdot 20^2 |A||S|^6 H^5}, & \text{if } \ell = \ell_2 \\ \frac{\epsilon^2}{10 \cdot 16 \cdot 20 |A||S|^4 H^3}, & \text{if } \ell = \ell_1 \end{cases}, \epsilon_R = \begin{cases} \frac{\epsilon^3}{20^3 |S||A| H^3}, & \text{if } \ell = \ell_2 \\ \frac{\epsilon^2}{20^2 |S||A| H^2}, & \text{if } \ell = \ell_1 \end{cases}, \epsilon_2 = \gamma/4.$

for $h \in [H-1]$ **do**

  initialize $Sample^R(h), Sample^P(h) = \emptyset$.

  compute the required number of episodes

$$T_h = \left\lceil \frac{8|S|}{\gamma \cdot \beta} \left(\ln \frac{1}{\delta_1} + 2 \max\{N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2), N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2)\} \right) \right\rceil$$

  $\widetilde{S}_h^{\gamma,\beta} \leftarrow \texttt{AGS}(S, A, H, h, \gamma, \beta, \{f_k^P, \widetilde{S}_k^{\gamma,\beta} : k \in [h-1]\}, \epsilon_2, \delta_2)$

  $\widehat{P}^c \leftarrow \texttt{ACDD}(S, A, H, h, \beta, \{f_k^P, \widetilde{S}_k^{\gamma,\beta} : k \in [h-1]\})$

  for $t = 1, 2, \ldots T_h$ **do**

    choose $(s_h, a_h) \in \widetilde{S}_h^{\gamma,\beta} \times A$ uniformly at random

    observe context c_t

    $(\widehat{\pi}_{s_h}^{c_t}, \widehat{p}_{s_h}^{c_t}) \leftarrow \texttt{FFP}(S, A, \widehat{P}^{c_t}, s_0, H, s_h)$.

    set $\widehat{\pi}_{s_h}^{c_t}(s_h) \leftarrow a_h$

    if $\widehat{p}_{s_h}^{c_t} \geq \beta$ **then**

      run $\widehat{\pi}_{s_h}^{c_t}$ and generate trajectory τ

      if (s_h, a_h, r_h, s_{h+1}) is in τ **then**

        update samples:

$$Sample^R(h) = Sample^R(h) + ((c_t, s_h, a_h), r_h)$$
$$Sample^P(h) = Sample^P(h) + \{((c_t, s_h, a_h, s'_{h+1}), \mathbb{I}[s_{h+1} = s'_{h+1}]) : s'_{h+1} \in S_{h+1}\}$$

  if $|Sample^R(h)| \geq 2 \cdot N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2)$ **then**

    $f_h^R = \texttt{ERM}(\mathcal{F}_h^R, Sample^R(h), \ell)$

  else

    return FAIL

  if $|Sample^P(h)| \geq 2 \cdot N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2)$ **then**

    $f_h^P = \texttt{ERM}(\mathcal{F}_h^P, Sample^P(h), \ell)$

  else

    return FAIL

return $\{f_h^R, f_h^P, \widetilde{S}_h^{\gamma,\beta} : \forall h \in [H-1]\}$

Algorithm 16 Exploit Unknown and Context-Dependent Dynamics CMDP (EXPLOIT-UCDD)

1: **inputs:**
- MDP parameters: $S = \{S_0, S_1, \ldots, S_H\}$,$A$,$s_0$,$H$ - the horizon length.
- Reachability parameter β.
- Function approximation for the rewards and dynamics for each layer $h \in [H-1]$ and the approximated set of (γ, β)-good contexts : $\{f_h^R, f_h^P, \widetilde{S}_h^{\gamma,\beta} : h \in [H-1]\}$

2: $\widehat{P}^c \leftarrow \texttt{ACDD}(S, A, H, H, \beta, \{f_k^P, \widetilde{S}_k^{\gamma,\beta} : k \in [H-1]\})$ ▷ $\widehat{P}^c$ is a function of the context c.
3: **for** $t = 1, 2, \ldots$ **do**
4: observe context c_t
5: define the reward approximation:

$$\forall h \in [H-1], s_h \in \widetilde{S}_h^{\gamma,\beta}, a_h \in A : \widehat{r}^c(s_h, a_h) = \mathbb{I}[c_t \in \widehat{\mathcal{C}}^\beta(s_h)] \cdot f_h^R(c_t, s_h, a_h)$$
$$\forall h \in [H-1], s_h \in S_h \setminus \widetilde{S}_h^{\gamma,\beta}, a_h \in A : \widehat{r}^{c_t}(s_h, a_h) = 0$$
$$\forall a \in A : \widehat{r}^{c_t}(s_{sink}, a) = 0$$

6: $\widehat{\mathcal{M}}(c_t) = (S \cup \{s_{sink}\}, A, \widehat{P}^{c_t}, \widehat{r}^{c_t}, s_0, H)$.
7: $\widehat{\pi}^{c_t} \leftarrow \texttt{Planning}(\widehat{\mathcal{M}}(c_t))$.
8: run $\widehat{\pi}^{c_t}$.

Formally, let $\widehat{p}_\beta(s)$ be the output of Algorithm AGC (see Algorithm 14) for the state $s \in S$, and denote $p_\beta(s) := \mathbb{P}[c \in \widehat{\mathcal{C}}^\beta(s)]$. For each layer $h \in [H-1]$ we define the event G_1^h as $G_1^h = \{|\widehat{p}_\beta(s_h) - p_\beta(s_h)| \leq \gamma/4 \quad \forall s_h \in S_h\}$ and define $G_1 = \cap_{h \in [H-1]} G_1^h$.

The good event G_1 guarantees that for every layer $h \in [H-1]$ and state $s_h \in S_h$, if $\widehat{p}_\beta(s_h) \geq \frac{3}{4}\gamma$ then $p_\beta(s_h) \geq \gamma/2$, which implies that $s_h \in \widehat{S}_h^{\gamma/2,\beta}$. This implies that for every layer h we sample only $(\gamma/2, \beta)$-good states for $\widehat{P}^c$.

More impotently, if $p_\beta(s_h) \geq \gamma$ then $\widehat{p}_\beta(s_h) \geq \frac{3}{4}\gamma$. Hence, we identify every (γ, β)-good state.

Thus, under the good event G_1, for every layer $h \in [H-1]$ the approximated set $\widetilde{S}_h^{\gamma,\beta}$ satisfies

$$\widehat{S}_h^{\gamma,\beta} \subseteq \widetilde{S}_h^{\gamma,\beta} \subseteq \widehat{S}_h^{\gamma/2,\beta}.$$

The following lemma shows that for our parameters choice, G_1 holds with high probability.

Lemma 89 *For $\epsilon_2 = \gamma/4$ and $\delta_2 = \frac{\delta}{8|S|}$, we have that $\mathbb{P}[G_1] \geq 1 - \delta/8$.*

Proof For each $s \in S$ we have that $\widehat{p}_\beta(s)$ is calculated over $m(\epsilon_2, \delta_2) = \left\lceil \frac{\ln \frac{2}{\delta_2}}{2\epsilon_2^2} \right\rceil$ examples. By Hoeffding's inequality combined with union bound, for $\epsilon_2 = \gamma/4$ and $\delta_2 = \frac{\delta}{8|S|}$, we obtain that $\mathbb{P}[G_1] \geq 1 - \delta/8$. ■

Sampling distributions. Recall that during the algorithm, for every layer $h \in [H-1]$ we collect examples of (c, s_h, a_h) for which $\widehat{p}_\beta(s_h) \geq \frac{3}{4}\gamma$, which under G_1 implies that $p_\beta(s_h) \geq \gamma/2$, context $c \in \widehat{C}^\beta(s_h)$ and actions $a_h \in A$.

For every layer $h \in [H-1]$ and reachability parameters γ and β we define the *target domain* we would like to collect examples from as

$$\mathcal{X}_h^{\gamma,\beta} = \{(c, s_h, a_h) : s_h \in \widehat{S}_h^{\gamma,\beta}, c \in \widehat{\mathcal{C}}^\beta(s_h), a_h \in A\},$$

recalling that

$$\widehat{C}^\beta(s_h) = \{c \in \mathcal{C} : s_h \text{ is } \beta - \text{reachable for } \widehat{P}^c\}$$

and

$$\widehat{S}_h^{\gamma,\beta} = \{s_h \in S_h : \mathbb{P}[c \in \widehat{\mathcal{C}}^{\beta}(s_h)] \geq \gamma\}.$$

Meaning, we would like to collect sufficient number of examples of (γ, β)-good states, appropriate good context and action for each layer.

In practice, we collect examples of states $s \in \widetilde{S}_h^{\gamma,\beta}$ which also contains states $s \in \widehat{S}_h^{\gamma/2,\beta}$. Under G_1 we have the guarantee that $\widehat{S}_h^{\gamma,\beta} \subseteq \widetilde{S}_h^{\gamma,\beta} \subseteq \widehat{S}_h^{\gamma/2,\beta}$.

Hence, we define the *empirical domain*

$$\widetilde{\mathcal{X}}_h^{\gamma,\beta} = \{(c, s_h, a_h) : s_h \in \widetilde{S}_h^{\gamma,\beta}, c \in \widehat{\mathcal{C}}^{\beta}(s_h), a_h \in A\},$$

We remark that before learning layer h, we compute $\widetilde{S}_h^{\gamma,\beta}$ based on the previous layers approximation for the dynamics which are fixed, hence $\widetilde{\mathcal{X}}_h^{\gamma,\beta}$ is fixed when learning layer h.

We also remark that under G_1 it holds, since $\widehat{S}_h^{\gamma,\beta} \subseteq \widetilde{S}_h^{\gamma,\beta} \subseteq \widehat{S}_h^{\gamma/2,\beta}$ it also holds that

$$\mathcal{X}_h^{\gamma,\beta} \subseteq \widetilde{\mathcal{X}}_h^{\gamma,\beta} \subseteq \mathcal{X}_h^{\gamma/2,\beta}.$$

We consider the marginal distributions of our observations, that sampled from $\widetilde{\mathcal{X}}_h^{\gamma,\beta}$.

For the rewards denote by $\widetilde{\mathcal{D}}_h^R$ the distribution over the collected examples $((c, s, a), r) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta} \times [0, 1]$, for each layer $h \in [H-1]$. It holds that

$$\begin{aligned}\widetilde{\mathcal{D}}_h^R((c, s_h, a_h), r_h) &= \mathbb{P}[((c, s_h, a_h), r_h) \in Sample^R(h) | (c, s_h, a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}] \\ &\propto \mathbb{P}[c | c \in \widehat{\mathcal{C}}^{\beta}(s_h)] \cdot q_h(s_h, a_h | \widehat{\pi}_{s_h}^c, P^c) \cdot \mathbb{P}[R^c(s_h, a_h) = r_h | c, s_h, a_h],\end{aligned}$$

where $\propto$ implies that we normalize to sum to 1.

Since under G_1 we have that $\mathcal{X}_h^{\gamma,\beta} \subseteq \widetilde{\mathcal{X}}_h^{\gamma,\beta}$, $\widetilde{\mathcal{D}}_h^R$ induces a marginal distribution over $\mathcal{X}_h^{\gamma,\beta} \times [0, 1]$, which we denote by $\mathcal{D}_h^R$. Clearly, it holds that

$$\begin{aligned}\mathcal{D}_h^R((c, s_h, a_h), r_h) &= \mathbb{P}[((c, s_h, a_h), r_h) \in Sample^R(h) | (c, s_h, a_h) \in \mathcal{X}_h^{\gamma,\beta}] \\ &\propto \mathbb{P}[c | c \in \widehat{\mathcal{C}}^{\beta}(s_h)] \cdot q_h(s_h, a_h | \widehat{\pi}_{s_h}^c, P^c) \cdot \mathbb{P}[R^c(s_h, a_h) = r_h | c, s_h, a_h],\end{aligned}$$

which is the desired marginal distribution over our target domain.

Similarly, for the next state we have,

$$\begin{aligned}&\widetilde{\mathcal{D}}_h^P((c, s_h, a_h, s'), \mathbb{I}[s_{h+1} = s']) \\ &\quad = \mathbb{P}[((c, s_h, a_h, s'), \mathbb{I}[s_{h+1} = s'])) \in Sample^P(h) | (c, s_h, a_h, s') \in (\widetilde{\mathcal{X}}_h^{\gamma,\beta} \times S_{h+1})] \\ &\quad \propto \mathbb{P}[c | c \in \widehat{\mathcal{C}}^{\beta}(s_h)] \cdot q_h(s_h, a_h | \widehat{\pi}_{s_h}^c, P^c) \cdot P^c(s' | s_h, a_h),\end{aligned}$$

and we denote $\mathcal{D}_h^P$ the induced marginal distribution over $(\mathcal{X}_h^{\gamma,\beta} \times S_{h+1}) \times [0, 1]$.

Remark 90 *When it is clear from the context, we use $\mathcal{D}_h^P$ and $\widetilde{\mathcal{D}}_h^P$ to also denote the induced distribution over $(c, s_h, a_h, s_{h+1}) \in \mathcal{X}_h^{\gamma,\beta} \times S_{h+1}$ and $(c, s_h, a_h, s_{h+1}) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta} \times S_{h+1}$, respectively, and drop the indicator bit. Similarly for $\mathcal{D}_h^R$ and $\widetilde{\mathcal{D}}_h^R$ we have $(c, s_h, a_h) \in \mathcal{X}_h^{\gamma,\beta}$ and $(c, s_h, a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$.*

Event G_2. Intuitively it states that sufficient number of examples have been collected for every layer $h \in [H-1]$.

Formally, let G_2^h be the event that

1. At least $\max\{N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2), N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2)\}$ examples of context, state and action from the target domain, i.e., $(c, s, a) \in \mathcal{X}_h^{\gamma,\beta}$, have been collected for layer $h \in [H-1]$.

2. At least $2\max\{N_R(\mathcal{F}_h^R,\epsilon_R,\delta_1/2), N_P(\mathcal{F}_h^P,\epsilon_P,\delta_1/2)\}$ examples of context, state and action from the empirical domain, i.e., $(c,s,a)\in\widetilde{\mathcal{X}}_h^{\gamma,\beta}$, have been collected for layer $h\in[H-1]$.

Let G_2 be the event $\cap_{h\in[H-1]}G_2^h$.

Event G_3. Intuitively states that the ERM guarantees for the approximation of the dynamics hold. Formally, let G_3^h denote the following event (for the ℓ_2 loss),

$$\mathbb{E}_{(c,s_h,a_h,s_{h+1})\sim\mathcal{D}_h^P}[(f_h^P(c,s_h,a_h,s_{h+1})-P^c(s_{h+1}|s_h,a_h))^2]\leq\epsilon_P,$$

and

$$\mathbb{E}_{(c,s_h,a_h,s_{h+1})\sim\widetilde{\mathcal{D}}_h^P}[(f_h^P(c,s_h,a_h,s_{h+1})-P^c(s_{h+1}|s_h,a_h))^2]\leq\epsilon_P.$$

Recall that we assume realizability for each layer. Define $G_3=\cap_{h\in[H-1]}G_3^h$.

The following lemma shows that if G_1 and G_2^h holds, then G_3^h holds with high probability. (We later show that G_2 holds with high probability.)

Lemma 91 *For any $h\in[H-1]$ it holds that $\mathbb{P}[G_3^h|G_1,G_2^h]\geq 1-\delta_1$.*

Proof Under G_1 and G_2^h we have collected sufficient number of examples from the domain $\mathcal{X}_h^{\gamma,\beta}\times S_{h+1}$ to approximate the transition probability function of layer h, for the accuracy parameter ϵ_P and confidence parameter $\delta_1/2$. By the ERM guarantees (see E.1.1), if sufficient number of examples have been collected, then the ERM output f_h^P satisfies that $\mathbb{E}_{\mathcal{D}_h^P}[(f_h^P(c,s_h,a_h,s_{h+1})-P^c(s_{h+1}|s_h,a_h))^2]\leq\epsilon_P$, with probability at least $1-\delta_1/2$. Similarly for $\widetilde{\mathcal{X}}_h^{\gamma,\beta}\times S_{h+1}$ it holds that $\mathbb{E}_{\widetilde{\mathcal{D}}_h^P}[(f_h^P(c,s_h,a_h,s_{h+1})-P^c(s_{h+1}|s_h,a_h))^2]\leq\epsilon_P$, with probability at least $1-\delta_1/2$. Hence the lemma follows by union bound. ■

The following lemma shows, inductively, that if for all the previous layers $i<h$ the good events G_1, G_2^i, G_3^i hold, then G_2^h holds with high probability, for the current layer h.

Lemma 92 *For every layer $h\in[H-1]$ it holds that*

$$\mathbb{P}[G_2^h|G_1,G_2^i,G_3^i\ \forall i\in[h-1]]\geq 1-(\delta_1+\frac{\epsilon_P}{\rho^2}|S|^2|A|).$$

Proof We prove the lemma using induction over the horizon h.

Base case. $h=0$.

By definition, the start state s_0 is $(1,1)$-good, which implies that for s_0 we collect samples in a deterministic manner. Thus, it holds that $\mathbb{P}[G_2^0]=1$.

Induction step. Assume the lemma holds for all $k<h$ and we show it holds for h.
Recall we collect examples of states $s_h\in S_h$ for which $\widehat{p}_\beta(s_h)\geq\frac{3}{4}\gamma$. Under G_1, if $\widehat{p}_\beta(s_h)\geq\frac{3}{4}\gamma$ then $\mathbb{P}[c\in\widehat{\mathcal{C}}^\beta(s_h)]\geq\gamma/2$. In addition, if $\mathbb{P}[c\in\widehat{\mathcal{C}}^\beta(s_h)]\geq\gamma$ then $\widehat{p}_\beta(s_h)\geq\frac{3}{4}\gamma$.

Thus, the set $\widetilde{S}_h^{\gamma,\beta}$ of approximately (γ,β)-good state for $\widehat{P}^c$ satisfies that $\widehat{S}_h^{\gamma,\beta}\subseteq\widetilde{S}_h^{\gamma,\beta}\subseteq\widehat{S}_h^{\gamma/2,\beta}$.

Given $G_1,G_2^k,G_3^k\ \forall k\in[h-1]$ hold, by Lemma 108, for β and ρ such that $\beta\geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$ the following holds.

$$\mathbb{P}_c\Big[\underbrace{\forall k\in[h],s_k\in S_k.\ \ q_k(s_k|\pi_c,P^c)\geq q_k(s_k|\pi_c,\widehat{P}^c)-\frac{4\rho|S|}{1-\rho^2|S|^2}k}_{(\star)}\Big]\geq 1-|A|\sum_{k=0}^{h-1}\frac{\epsilon_P}{\rho^2}|S_k||S_{k+1}|$$
$$\geq 1-|A||S|^2\frac{\epsilon_P}{\rho^2}.$$

Claim 1 *Assume inequality* $(\star)$ *holds. Then the probability to collect one example of* $(c, s_h, a_h) \in \mathcal{X}_h^{\gamma,\beta}$ *is at least* $\frac{1}{|S|}\gamma(\beta - \frac{4\rho|S|}{1-\rho^2|S|^2}h) \geq \frac{1}{|S|} \cdot \gamma \cdot \beta/2$.

Proof Consider the process of collecting a sample, as described in Algorithm 16:

1. The algorithm/agent chooses uniformly at random $(s, a) \in \widetilde{S}_h^{\gamma,\beta} \times A$. Under the good event G_1 we have that $\widehat{S}_h^{\gamma,\beta} \subseteq \widetilde{S}_h^{\gamma,\beta}$. Hence, the probability to choose $(s, a) \in \widehat{S}_h^{\gamma,\beta} \times A$ is at least $\frac{1}{|S|}$.

2. A context $c \sim \mathcal{D}$ is sampled. By $\widehat{S}_h^{\gamma,\beta}$ definition, the probability that $c \in \widehat{\mathcal{C}}^{\beta}(s)$ is at least γ.

   - If $c \in \widehat{\mathcal{C}}^{\beta}(s)$, the agent plays $\widehat{\pi}_s^c$ to generate a trajectory where the dynamics is P^c. By $(\star)$ and $\widehat{\mathcal{C}}^{\beta}(s)$ definition, the probability to observe (s, a) in a trajectory generated using $\widehat{\pi}_s^c$ where the dynamics is P^c is $q_h(s|\widehat{\pi}_s^c, P^c) \geq \beta - \frac{4\rho|S|}{1-\rho^2|S|^2}h \geq \beta/2$.
   - Otherwise quite iteration.

Overall, the probability to collect one example of a triplet $(c, s, a) \in \mathcal{X}_h^{\gamma,\beta}$ is at least $\frac{1}{|S|} \cdot \gamma \cdot (\beta - \frac{4\rho|S|}{1-\rho^2|S|^2}h) \geq \frac{1}{|S|} \cdot \gamma \cdot \frac{\beta}{2}$ (since $\beta \geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$). ■

Claim 2 *Assume inequality* $(\star)$ *holds. Then the probability to collect one example of* $(c, s_h, a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$ *is at least* $\frac{\gamma}{2}(\beta - \frac{4\rho|S|}{1-\rho^2|S|^2}h) \geq \gamma \cdot \beta/4$.

Proof Consider the process of collecting a sample, as described in Algorithm 16:

1. The algorithm/agent chooses uniformly at random $(s, a) \in \widetilde{S}_h^{\gamma,\beta} \times A$. Under the good event G_1 we have that $\widetilde{S}_h^{\gamma,\beta} \subseteq \widehat{S}_h^{\gamma/2,\beta}$.

2. A context $c \sim \mathcal{D}$ is sampled by the nature. By $\widehat{S}_h^{\gamma/2,\beta}$ definition, the probability to observe a context $c \in \widehat{\mathcal{C}}^{\beta}(s)$ is at least $\gamma/2$.

   - If $c \in \widehat{\mathcal{C}}^{\beta}(s)$, the agent plays $\widehat{\pi}_s^c$ to generate a trajectory where the dynamics is P^c. By $(\star)$ and $\widehat{\mathcal{C}}^{\beta}(s)$ definition, the probability to observe (s, a) in a trajectory generated using $\widehat{\pi}_s^c$ where the dynamics is P^c is $q_h(s|\widehat{\pi}_s^c, P^c) \geq \beta - \frac{4\rho|S|}{1-\rho^2|S|^2}h \geq \beta/2$.
   - Otherwise quite iteration.

Overall, the probability to collect one sample of some triplet $(c, s, a) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$ is at least $\gamma/2 \cdot (\beta - \frac{4\rho|S|}{1-\rho^2|S|^2}h) \geq \gamma \cdot \beta/4$ (since $\beta \geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$). ■

The above claims implies that if $(\star)$ holds, in expectation, the agent needs to experience at most $\frac{2|S|}{\gamma \cdot \beta}$ episodes to collect one example from $\mathcal{X}_h^{\gamma,\beta}$. In addition, in expectation, the agent needs to experience at most $\frac{4}{\gamma \cdot \beta}$ episodes to collect one example from $\widetilde{\mathcal{X}}_h^{\gamma,\beta}$.

Since under G_1 we have that $\mathcal{X}_h^{\gamma,\beta} \subseteq \widetilde{\mathcal{X}}_h^{\gamma,\beta} \subseteq \mathcal{X}_h^{\gamma/2,\beta}$, using multiplicative Chernoff bound we obtain that with probability at least $1 - \delta_1$ after experiencing

$$T_h = \left\lceil \frac{8|S|}{\gamma \cdot \beta} \left(\ln \frac{1}{\delta_1} + 2 \max\{N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2), N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2)\} \right) \right\rceil$$

episodes, the agent will collect at least $\max\{N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2)), N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2))\}$ examples from $\mathcal{X}_h^{\gamma,\beta}$ and $2\max\{N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2)), N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2))\}$ examples from $\widetilde{\mathcal{X}}_h^{\gamma,\beta}$.

Recall that T_h is exactly the number of episodes we run in Algorithm 16 when learning layer h. Hence, using union bound we obtain that

$$\mathbb{P}[G_2^h | G_1, G_2^i, G_3^i \ \forall i \in [h-1]] \geq 1 - (\delta_1 + |A||S|^2 \frac{\epsilon_P}{\rho^2}).$$

■

The following lemma shows that given G_1 holds, G_2 and G_3 holds with high probability.

Lemma 93 *The following holds.*

$$\mathbb{P}[G_2 \cap G_3 | G_1] \geq 1 - (2\delta_1 H + \frac{\epsilon_P}{\rho^2}|S|^2|A|H).$$

Proof Assume the good event G_1 holds. Recall that $G_2 = \cap_{h\in[H-1]} G_2^h$ and $G_3 = \cap_{h\in[H-1]} G_3^h$.

Let X be a random variable with support $[H-1]$ that satisfies

$$X = \min_{k\in[H-1]} \{\overline{G}_2^k \cup \overline{G}_3^k \text{ holds }\},$$

and otherwise $X = \perp$, meaning if G_2 and G_3 hold.

In words, X is the first layer in which at least one of the good events G_2^h or G_3^h does not hold.

By X definition and Bayes rule (i.e., $\mathbb{P}[A \cap B] = \mathbb{P}[A|B] \cdot \mathbb{P}[B]$) we have

$$\begin{aligned}
\forall h \in [H].\ \mathbb{P}[X = h | G_1] &= \mathbb{P}[(\overline{G}_2^h \cup \overline{G}_3^h) \cap (\cap_{k\in[h-1]} G_2^k \cap G_3^k) | G_1] \\
&= \mathbb{P}[(\overline{G}_2^h \cup \overline{G}_3^h) | G_1, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)] \cdot \underbrace{\mathbb{P}[(\cap_{k\in[h-1]} G_2^k \cap G_3^k) | G_1]}_{\leq 1} && \text{(Base rule)} \\
&\leq \mathbb{P}[(\overline{G}_2^h \cup \overline{G}_3^h) | G_1, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)] \\
&= \underbrace{\mathbb{P}[\overline{G}_2^h | G_1, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)]}_{\leq \delta_1 + \frac{\epsilon_P}{\rho^2}|S|^2|A| \text{ by Lemma 92}} + \mathbb{P}[\overline{G}_3^h \cap G_2^h | G_1, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)] \\
& && \text{(Union of disjoint events)} \\
&\leq \delta_1 + \frac{\epsilon_P}{\rho^2}|S||A| + \mathbb{P}[\overline{G}_3^h \cap G_2^h | G_1, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)] \\
&= \delta_1 + \frac{\epsilon_P}{\rho^2}|S||A| + \mathbb{P}[\overline{G}_3^h | G_1, G_2^h, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)] \cdot \underbrace{\mathbb{P}[G_2^h | G_1 \cap (\cap_{k\in[h-1]} G_2^k \cap G_3^k)]]}_{\leq 1} \\
& && \text{(caused by rule)} \\
&\leq \delta_1 + \frac{\epsilon_P}{\rho^2}|S||A| + \mathbb{P}[\overline{G}_3^h | G_1, G_2^h, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)] \\
&= \delta_1 + \frac{\epsilon_P}{\rho^2}|S|^2|A| + \underbrace{\mathbb{P}[\overline{G}_3^k | G_1, G_2^h]}_{\leq \delta_1 \text{ by Lemma 91}} \\
&\leq 2\delta_1 + \frac{\epsilon_P}{\rho^2}|S|^2|A|.
\end{aligned}$$

Now, by G_2 and G_3 definitions we have

$$\mathbb{P}[G_2 \cap G_3 | G_1] = 1 - \mathbb{P}[\overline{G}_2 \cup \overline{G}_3 | G_1]$$

$$
\begin{aligned}
&= 1 - \mathbb{P}[\cup_{h\in[H-1]}(\overline{G}_2^h \cup \overline{G}_3^h)|G_1] \\
&= 1 - \mathbb{P}[\exists h \in [H-1].(\overline{G}_2^h \cup \overline{G}_3^h)|G_1] \\
&= 1 - \mathbb{P}[\exists h \in [H-1].X = h|G_1] \\
&= 1 - \mathbb{P}[\cup_{h\in[H-1]}\{X = h\}|G_1] \\
&= 1 - \sum_{h=0}^{H-1} \mathbb{P}[X = h|G_1] && \text{(Union bound)} \\
&\geq 1 - (2\delta_1 H + \frac{\epsilon_P}{\rho^2}|S|^2|A|H),
\end{aligned}
$$

as stated. ■

Event G_4. Intuitively states that the ERM guarantees for the approximation of the rewards function hold (for the ℓ_2 loss). Let G_4 denote the good event

$$\forall h \in [H-1].\ \mathbb{E}_{(c,s_h,a_h)\sim\mathcal{D}_h^R}[(f_h^R(c,s_h,a_h) - r^c(s_h,a_h))^2] \leq \epsilon_R + \alpha_2^2(\mathcal{F}_h^R).$$

The following lemma shows that given G_1 and G_2 hold, we have that G_4 holds with high probability.

Lemma 94 *It holds that* $\mathbb{P}[G_4|G_1, G_2] \geq 1 - \delta_1 H$.

Proof Since G_1 and G_2 hold, for every layer $h \in [H-1]$ sufficient number of examples $((c, s_h, a_h), r_h) \in \mathcal{X}_h^{\gamma,\beta} \times [0,1]$ have been collected for the ERM to output a function f_h^R the satisfies

$$\mathbb{E}_{(c,s_h,a_h)\sim\mathcal{D}_h^R}[(f_h^R(c,s_h,a_h) - r^c(s_h,a_h))^2] \leq \epsilon_R + \alpha_2^2(\mathcal{F}_h^R)$$

with probability at least $1-\delta_1$. Hence, the lemma follows by the ERM guarantees (see E.1.1) and an union bound over every layer $h \in [H-1]$. ■

Overall, all the good events hols with high probability.

Corollary 95 *The following holds.*

$$\mathbb{P}[G_1, G_2, G_3, G_4] \geq 1 - \left(\frac{\delta}{8} + 3\delta_1 H + \frac{\epsilon_P}{\rho^2}|S|^2|A|H\right).$$

Proof Followed from union bound over the results of Lemmas 89, 94 and 93. ■

E.4.2 ANALYSIS OF THE ERROR CAUSED BY THE DYNAMICS APPROXIMATION UNDER THE GOOD EVENTS

In the following analysis, for any context $c \in \mathcal{C}$ we consider an intermediate MDP associated with it: $\widetilde{\mathcal{M}}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}^c, r^c, s_0, H)$, where $\widehat{P}^c$ is the approximation of the dynamics P^c and r^c is the true rewards function extended to s_{sink} by defining that $r^c(s_{sink}, a) := 0,\ \ \forall c \in \mathcal{C},\ a \in A$. Recall the true MDP associated with this context is $\mathcal{M}(c) = (S, A, P^c, r^c, s_0, H)$.

Lemma 96 *Let* $\rho \in [0, \frac{1}{|S|})$ *and* $h \in [H-1]$. *Assume the good events* $G_1, G_2^k, G_3^k,\ \forall k \in [h]$ *hold, then it holds that*

$$\mathop{\mathbb{P}}_{(c,s_h,a_h)}\left[\|\widehat{P}^c(\cdot|s_h,a_h) - P^c(\cdot|s_h,a_h)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}\Big|(c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}\right] \geq 1 - \frac{\epsilon_P}{\rho^2}|S_{h+1}|,$$

where $\widehat{P}^c$ is the dynamics defined in Algorithm 12 and

$$\|\widehat{P}^c(\cdot|s_h,a_h) - P^c(\cdot|s_h,a_h)\|_1 := \sum_{s_{h+1}\in S_{h+1}} |\widehat{P}^c(s_{h+1}|s_h,a_h) - P^c(s_{h+1}|s_h,a_h)|$$

(i.e., the entry of s_{sink} in $\widehat{P}^c$ is ignored).

Proof Under G_1 it holds that $\mathcal{X}_h^{\gamma,\beta} \subseteq \widetilde{\mathcal{X}}_h^{\gamma,\beta}$. Recall that for all $(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$ we have that $\widehat{P}^c(s_{sink}|s_h,a_h) = 0$ by $\widehat{P}^c$ definition. Hence, for all $(c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}$ we have that $\widehat{P}^c(s_{sink}|s_h,a_h) = 0$.

In addition, the true dynamics P^c is not defined for s_{sink} since $s_{sink} \notin S$. A natural extension of P^c to s_{sink} is by defining that $\forall(s,a) \in S \times A.\ \ P^c(s_{sink}|s,a) := 0$. By that extension, we have for all $(c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}$ that $P^c(s_{sink}|s_h,a_h) = \widehat{P}^c(s_{sink}|s_h,a_h) = 0$. Hence, we can simply ignore s_{sink} in the following analysis.

Under the good event G_3^h, we have

$$\begin{aligned}
&\mathop{\mathbb{P}}_{(c,s_h,a_h,s_{h+1})}\left[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)| \geq \rho \middle| (c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}\right] = \\
&= \mathbb{P}_{\mathcal{D}_h^P}[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)| \geq \rho] \\
&= \mathbb{P}_{\mathcal{D}_h^P}[(f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h))^2 \geq \rho^2] \\
&\leq \frac{\mathbb{E}_{\mathcal{D}_h^P}[(f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h))^2]}{\rho^2} && \text{(By Markov's inequality)} \\
&\leq \frac{\epsilon_P}{\rho^2}. && \text{(Under } G_h^3\text{)}
\end{aligned}$$

Hence,

$$\mathop{\mathbb{P}}_{(c,s_h,a_h,s_{h+1})}\left[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)| \leq \rho \middle| (c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}\right] \geq 1 - \frac{\epsilon_P}{\rho^2}.$$

As $P^c(\cdot|s_h,a_h)$ is a distribution, we have for every context c that $\sum_{s_{h+1}\in S_{h+1}} P^c(s_{h+1}|s_h,a_h) = 1$.

Thus, by union bound over $s_{h+1} \in S_{h+1}$ we obtain

$$\mathbb{P}_{(c,s_h,a_h)}\left[1-\rho|S| \leq \sum_{s_{h+1}\in S_{h+1}} f_h^P(c,s_h,a_h,s_{h+1}) \leq 1+\rho|S| \middle| (c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}\right] \geq 1 - \frac{\epsilon_P}{\rho^2}|S_{h+1}|.$$

Hence, we further conclude

$$\mathop{\mathbb{P}}_{(c,s_h,a_h)}\left[\forall s_{h+1} \in S_{h+1}. \frac{P^c(s_{h+1}|s_h,a_h) - \rho}{1+\rho|S|} \leq \underbrace{\frac{f_h^P(c,s_h,a_h,s_{h+1})}{\sum_{s'\in S_{h+1}} f_h^P(c,s_h,a_h,s')}}_{=\widehat{P}^c(s_{h+1}|s_h,a_h)} \leq \frac{P^c(s_{h+1}|s_h,a_h) + \rho}{1-\rho|S|} \ \middle| (c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}\right]$$

$$\geq 1 - \frac{\epsilon_P}{\rho^2}|S_{h+1}|. \tag{11}$$

Fix a tuple $(c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}$ and assume the event of inequality (11) holds.
Denote $S_{h+1}^+ = \{s_{h+1} \in S_{h+1} : \widehat{P}^c(s_{h+1}|s_h,a_h) \geq P^c(s_{h+1}|s_h,a_h)\}$ and consider the following derivation.

$$\begin{aligned}
\|\widehat{P}^c(\cdot|s_h,a_h) - P^c(\cdot|s_h,a_h)\|_1 &= \sum_{s_{h+1}\in S_{h+1}} |\widehat{P}^c(s_{h+1}|s_h,a_h) - P^c(s_{h+1}|s_h,a_h)| \\
&= \sum_{s_{h+1}\in S_{h+1}^+} (\widehat{P}^c(s_{h+1}|s_h,a_h) - P^c(s_{h+1}|s_h,a_h))
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{s_{h+1}\in S_{h+1}\setminus S_{h+1}^{+}}\left(P^c(s_{h+1}|s_h,a_h)-\widehat{P}^c(s_{h+1}|s_h,a_h)\right)\\
\leq&\sum_{s_{h+1}\in S_{h+1}^{+}}\left(\frac{P^c(s_{h+1}|s_h,a_h)+\rho}{1-\rho|S|}-P^c(s_{h+1}|s_h,a_h)\right)\\
&+\sum_{s_{h+1}\in S_{h+1}\setminus S_{h+1}^{+}}\left(P^c(s_{h+1}|s_h,a_h)-\frac{P^c(s_{h+1}|s_h,a_h)-\rho}{1+\rho|S|}\right)\\
=&\sum_{s_{h+1}\in S_{h+1}^{+}}\frac{P^c(s_{h+1}|s_h,a_h)+\rho-(1-\rho|S|)P^c(s_{h+1}|s_h,a_h)}{1-\rho|S|}\\
&+\sum_{s_{h+1}\in S_{h+1}\setminus S_{h+1}^{+}}\frac{-P^c(s_{h+1}|s_h,a_h)+\rho+(1+\rho|S|)P^c(s_{h+1}|s_h,a_h)}{1+\rho|S|}\\
=&\frac{1}{1-\rho|S|}\sum_{s_{h+1}\in S_{h+1}^{+}}(\rho+\rho|S|P^c(s_{h+1}|s_h,a_h))\\
&+\frac{1}{1+\rho|S|}\sum_{s_{h+1}\in S_{h+1}\setminus S_{h+1}^{+}}(\rho+\rho|S|P^c(s_{h+1}|s_h,a_h))\\
\leq&\frac{2\rho|S|}{1-\rho|S|}+\frac{2\rho|S|}{1+\rho|S|}\\
=&\frac{4\rho|S|}{1-\rho^2|S|^2}.
\end{aligned}
$$

By inequality (11), the above holds with probability at least $1-\frac{\epsilon_P}{\rho^2}|S_{h+1}|$ over $(c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}$. Hence the lemma follows.

■

Lemma 97 *For the parameters choice $\beta=\frac{\epsilon}{20|S|H}\in(0,1)$, $\rho=\frac{\beta}{16|S|H}\in(0,\frac{1}{|S|})$, and $\epsilon_P=\frac{\epsilon^3}{10\cdot 2^8 20^2|A||S|^6H^5}$, we have $\beta\geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$. In addition, under the good events G_1, G_2^k and G_3^k for all $k\in[h]$, we have that*

$$
\mathbb{P}\left[\|\widehat{P}^c(\cdot|s_h,a_h)-P^c(\cdot|s_h,a_h)\|_1\leq\frac{\epsilon}{40|S|H^2}\middle|(c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}\right]\geq 1-\frac{\epsilon}{10|S||A|H}.
$$

Proof An immediate implication of lemma 96. ■

Lemma 98 (occupancy measures difference) *Let $\rho\in[0,\frac{1}{|S|})$ and $\beta\in(0,1]$ for which $\beta\geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$. Under the good events G_1,G_2,G_3, for every context-dependent policy π it holds that*

$$
\mathbb{P}_c\left[\forall h\in[H].\ \|q_h(\cdot|\pi_c,P^c)-q_h(\cdot|\pi_c,\widehat{P}^c)\|_1\leq\frac{4\rho|S|}{1-\rho^2|S|^2}h+\beta\sum_{k=1}^{h-1}|S_k|\right]\geq 1-\left(\frac{\epsilon_P}{\rho^2}|A|\sum_{i=0}^{H-1}|S_i||S_{i+1}|+\gamma\sum_{i=1}^{H-1}|S_i|\right)
$$

for a fixed $\rho\in[0,\frac{1}{|S|})$ and $\beta\geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$, where we define

$$
\forall h\in[H].\ \ \|q_h(\cdot|\pi_c,P^c)-q_h(\cdot|\pi_c,\widehat{P}^c)\|_1:=\sum_{s_h\in S_h}|q_h(s_h|\pi_c,P^c)-q_h(s_h|\pi_c,\widehat{P}^c)|
$$

(i.e., $q_h(s_{sink}|\pi_c,\widehat{P}^c)$ is ignored for all $h\in[H]$).

Remark 99 *Since $s_{sink} \notin S$, $q_h(s_{sink}|\pi_c, P^c)$ is not defined for the true dynamics P^c. In addition, by $\widehat{P}^c$ definition, from the sink there are no transitions to any other state, hence, we can simply ignore it in the following analysis.*

Proof We will show the lemma by induction over the horizon, h.

For the base case $h = 0$ the claim holds trivially (with probability 1) since the start state s_0 is unique.

For the induction step, assume that it holds for h, namely,

$$\mathbb{P}_c\left[\forall k \in [h]. \quad \|q_k(\cdot|\pi_c, P^c) - q_k(\cdot|\pi_c, \widehat{P}^c)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}k + \beta\sum_{i=1}^{k-1}|S_i|\right] \geq 1 - \left(\frac{\epsilon_P}{\rho^2}|A|\sum_{i=0}^{h-1}|S_i||S_{i+1}| + \gamma\sum_{i=1}^{h-1}|S_i|\right)$$

and prove for $h+1$.

Under the good events G_1 , G_2 and G_3 by Lemma 96 it holds that

$$\mathbb{P}_{(c,s_h,a_h)}\left[\|\widehat{P}^c(\cdot|s_h,a_h) - P^c(\cdot|s_h,a_h)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}\Big|(c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}\right] \geq 1 - \frac{\epsilon_P}{\rho^2}|S_{h+1}|.$$

Consider the following derivation for any fixed context $c \in \mathcal{C}$. (Later we will take the probability over $c \sim \mathcal{D}$.)

$$\begin{aligned}
&\|q_{h+1}(\cdot|\pi_c, P^c) - q_{h+1}(\cdot|\pi_c, \widehat{P}^c)\|_1 \\
&= \sum_{s_{h+1}\in S_{h+1}} |q_{h+1}(s_{h+1}|\pi_c, P^c) - q_{h+1}(s_{h+1}|\pi_c, \widehat{P}^c)| \\
&= \sum_{s_{h+1}\in S_{h+1}} |\sum_{s_h\in S_h}\sum_{a_h\in A}(q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h) - q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)\widehat{P}^c(s_{h+1}|s_h,a_h))| \\
&\leq \underbrace{\sum_{s_h\in S_h}\sum_{a_h\in A}\sum_{s_{h+1}\in S_{h+1}} |q_h(s_h|\pi_c, P^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h) - q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h)|}_{(1)} \\
&\quad + \underbrace{\sum_{s_h\in S_h}\sum_{a_h\in A}\sum_{s_{h+1}\in S_{h+1}} |q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h) - q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)\widehat{P}^c(s_{h+1}|s_h,a_h)|}_{(2)}
\end{aligned}$$

We bound (1) and (2) separately.

For (1), since $P^c(\cdot|s_h,a_h)$, $\pi_c(\cdot|s_h)$ are distributions, the following holds with probability 1.

$$\begin{aligned}
(1) &= \sum_{s_h\in S_h} |q_h(\cdot|\pi_c, P^c) - q_h(s_h|\pi_c, \widehat{P}^c)| \sum_{a_h\in A}\pi_c(a_h|s_h) \sum_{s_{h+1}\in S_{h+1}} P^c(s_{h+1}|s_h,a_h) \\
&= \|q_h(\cdot|\pi_c, P^c) - q_h(\cdot|\pi_c, \widehat{P}^c)\|_1 \sum_{a_h\in A}\pi_c(a_h|s_h) \sum_{s_{h+1}\in S_{h+1}} P^c(s_{h+1}|s_h,a_h) \\
&= \|q_h(\cdot|\pi_c, P^c) - q_h(\cdot|\pi_c, \widehat{P}^c)\|_1.
\end{aligned}$$

For (2), we define for any given context c and every layer $h \in [H-1]$ the following subsets of S_h .

1. $B_1^{h,c} = \{s_h \in S_h : s_h \in \widehat{S}_h^{\gamma,\beta}$ and $c \in \widehat{C}^\beta(s_h)\}$.
2. $B_2^{h,c} = \{s_h \in S_h : s_h \in \widehat{S}_h^{\gamma,\beta}$ and $c \notin \widehat{C}^\beta(s_h)\}$.
3. $B_3^{h,c} = \{s_h \in S_h : s_h \notin \widehat{S}_h^{\gamma,\beta}$ and $c \notin \widehat{C}^\beta(s_h)\}$.

4. $B_4^{h,c} = \{s_h \in S_h : s_h \notin \widehat{S}_h^{\gamma,\beta} \text{ and } c \in \widehat{C}^{\beta}(s_h)\}$.

Clearly, $\cup_{i=1}^4 B_i^{h,c} = S_h$ for all $h \in [H-1]$ and $c \in \mathcal{C}$.

By definition of $B_1^{h,c}$, for every layer $h \in [H-1]$ we have that $s_h \in B_1^{h,c}$ if and only if for every action $a_h \in A$ it holds that $(c, s_h, a_h) \in \mathcal{X}_h^{\gamma,\beta}$.

By definition of $B_4^{h,c}$, for every layer $h \in [H-1]$ it holds that

$$\mathbb{P}_c[B_4^{h,c} \neq \emptyset] = \mathbb{P}_c[\exists s_h \in S_h : s_h \notin \widehat{S}_h^{\gamma,\beta} \text{ and } c \in \widehat{C}^{\beta}(s_h)] \leq \gamma |S_h|.$$

Thus, for every $h \in [H-1]$ we have $\mathbb{P}_c[B_4^{h,c} = \emptyset] \geq 1 - \gamma|S_h|$.

In the following, we assume that $B_4^{h,c} = \emptyset$. Since $\mathbb{P}_c[B_4^{h,c} = \emptyset] \geq 1 - \gamma|S_h|$, it will only add $\gamma|S_h|$ to the probability of the error.

Consider the following derivation

$$\begin{aligned}
(2) &= \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} \sum_{s_{h+1} \in S_{h+1}} |q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h) - q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)\widehat{P}^c(s_{h+1}|s_h,a_h)| \\
&+ \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} \sum_{s_{h+1} \in S_{h+1}} |q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h) - q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)\widehat{P}^c(s_{h+1}|s_h,a_h)| \\
&+ \sum_{s_h \in B_4^{h,c}} \sum_{a_h \in A} \sum_{s_{h+1} \in S_{h+1}} |q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h) - q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)\widehat{P}^c(s_{h+1}|s_h,a_h)| \\
&= \sum_{s_h \in B_1^{h,c}} q_h(s_h|\pi_c, \widehat{P}^c) \sum_{a_h \in A} \pi_c(a_h|s_h) \sum_{s_{h+1} \in S_{h+1}} |P^c(s_{h+1}|s_h,a_h) - \widehat{P}^c(s_{h+1}|s_h,a_h)| \\
&+ \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h) \underbrace{\sum_{s_{h+1} \in S_{h+1}} |P^c(s_{h+1}|s_h,a_h) - \widehat{P}^c(s_{h+1}|s_h,a_h)|}_{\leq 1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (B_4^{h,c} = \emptyset \text{ w.p. at least } 1 - \gamma|S_h|) \\
&\leq \sum_{s_h \in B_1^{h,c}} q_h(s_h|\pi_c, \widehat{P}^c) \sum_{a_h \in A} \pi_c(a_h|s_h) \|P^c(\cdot|s_h,a_h) - \widehat{P}^c(\cdot|s_h,a_h)\|_1 + \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} q_h(s_h|\pi_c, \widehat{P}^c) \underbrace{\sum_{a_h \in A} \pi_c(a_h|s_h)}_{=1} \\
&\leq \underbrace{\sum_{s_h \in B_1^{h,c}} q_h(s_h|\pi_c, \widehat{P}^c) \sum_{a_h \in A} \pi_c(a_h|s_h)}_{\leq 1} \frac{4\rho|S|}{1-\rho^2|S|^2} + \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \underbrace{q_h(s_h|\pi_c, \widehat{P}^c)}_{\leq q_h(s_h|\widehat{\pi}^c_{s_h}, \widehat{P}^c) < \beta} \underbrace{\sum_{a_h \in A} \pi_c(a_h|s_h)}_{=1} \\
&\qquad\qquad \text{(By Lemma 96 and union bound over } (s_h, a_h) \in B_1^{h,c} \times A, \text{ holds w.p. at least } 1 - |A||S_h|\tfrac{\epsilon_P}{\rho^2}|S_{h+1}|) \\
&= \frac{4\rho|S|}{1-\rho^2|S|^2} + \beta|S_h|.
\end{aligned}$$

Hence,

$$\mathbb{P}_c\left[(2) \leq \frac{4\rho|S|}{1-\rho^2|S|^2} + \beta|S_h|\right] \geq 1 - \left(|A||S_h|\frac{\epsilon_P}{\rho^2}|S_{h+1}| + \gamma|S_h|\right). \tag{12}$$

In addition, we showed above that

$$\mathbb{P}_c\left[\|q_{h+1}(\cdot|\pi_c, P^c) - q_{h+1}(\cdot|\pi_c, \widehat{P}^c)\|_1 \leq (1) + (2)\right] = 1,$$

and

$$\mathbb{P}_c\left[(1) = \|q_h(\cdot|\pi_c, P^c) - q_h(\cdot|\pi_c, \widehat{P}^c)\|_1\right] = 1.$$

Thus, by combining all the above inequalities with the induction hypothesis we obtain

$$\mathbb{P}_c\left[\forall k\in[h+1].\quad \|q_k(\cdot|\pi_c,P^c)-q_k(\cdot|\pi_c,\widehat{P}^c)\|_1\le\frac{4\rho|S|}{1-\rho^2|S|^2}k+\beta\sum_{i=1}^{k-1}|S_i|\right]$$

$$=\mathbb{P}_c\Bigg[\forall k\in[h].\quad \|q_k(\cdot|\pi_c,P^c)-q_k(\cdot|\pi_c,\widehat{P}^c)\|_1\le\frac{4\rho|S|}{1-\rho^2|S|^2}k+\beta\sum_{i=1}^{k-1}|S_i|\ \text{ and }$$

$$\|q_{h+1}(\cdot|\pi_c,P^c)-q_{h+1}(\cdot|\pi_c,\widehat{P}^c)\|_1\le\frac{4\rho|S|}{1-\rho^2|S|^2}(h+1)+\beta\sum_{i=1}^{h}|S_i|\Bigg]$$

$$\ge\mathbb{P}_c\Bigg[\forall k\in[h].\quad \|q_k(\cdot|\pi_c,P^c)-q_k(\cdot|\pi_c,\widehat{P}^c)\|_1\le\frac{4\rho|S|}{1-\rho^2|S|^2}k+\beta\sum_{i=1}^{k-1}|S_i|\ \text{ and }$$

$$(1)+(2)\le\frac{4\rho|S|}{1-\rho^2|S|^2}(h+1)+\beta\sum_{i=1}^{h}|S_i|\Bigg]\quad\left(\text{Since }\mathbb{P}_c\left[\|q_{h+1}(\cdot|\pi_c,P^c)-q_{h+1}(\cdot|\pi_c,\widehat{P}^c)\|_1\le(1)+(2)\right]=1.\right)$$

$$\ge\mathbb{P}_c\left[\forall k\in[h].\quad \|q_k(\cdot|\pi_c,P^c)-q_k(\cdot|\pi_c,\widehat{P}^c)\|_1\le\frac{4\rho|S|}{1-\rho^2|S|^2}k+\beta\sum_{i=1}^{k-1}|S_i|\ \text{ and }(2)\le\frac{4\rho|S|}{1-\rho^2|S|^2}+\beta|S_h|\right]$$

$$\left(\text{Since }\mathbb{P}_c\left[(1)=\|q_h(\cdot|\pi_c,P^c)-q_h(\cdot|\pi_c,\widehat{P}^c)\|_1\right]=1.\right)$$

$$\ge1-\left(\frac{\epsilon_P}{\rho^2}|A|\sum_{i=0}^{h-1}|S_i||S_{i+1}|+\gamma\sum_{i=1}^{h-1}|S_i|+\frac{\epsilon_P}{\rho^2}|A||S_h||S_{h+1}|+\gamma|S_h|\right)=1-\left(\frac{\epsilon_P}{\rho^2}|A|\sum_{i=0}^{h}|S_i||S_{i+1}|+\gamma\sum_{i=1}^{h}|S_i|\right),$$

(By the induction hypothesis and equation (12))

as stated. ■

Lemma 100 (expected value difference caused by dynamics approximation) *Under the good events G_1, G_2 and G_3, for every context-dependent policy $\pi=(\pi_c)_{c\in\mathcal{C}}$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0)-V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|]\le\frac{4\rho|S|}{1-\rho^2|S|^2}H^2+\beta|S|H+H|S|^2|A|\frac{\epsilon_P}{\rho^2}+\gamma H|S|,$$

for $\rho\in[0,\frac{1}{|S|})$ and $\beta\ge2H\frac{4\rho|S|}{1-\rho^2|S|^2}$.

Proof Recall that the true rewards function is not defined for s_{sink}, since $s_{sink}\notin S$. For the intermediate MDP $\widetilde{\mathcal{M}}(c)$ we extended r^c to s_{sink} by defining $\forall a\in A.\ \ r^c(s_{sink},a)=0$ for every context $c\in\mathcal{C}$. Since P^c is also not defined for s_{sink}, we can simply omit s_{sink}, as the second equality in the following derivation shows.

Consider the following derivation for any fixed $c\in\mathcal{C}$. (Later we will take the expectation over c.)

$$|V^{\pi_c}_{\mathcal{M}(c)}(s_0)-V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|$$

$$=\left|\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A}q_h(s_h,a_h|\pi_c,P^c)\cdot r^c(s_h,a_h)-\sum_{h=0}^{H-1}\sum_{s_h\in S_h\cup\{s_{sink}\}}\sum_{a_h\in A}q_h(s_h,a_h|\pi_c,\widehat{P}^c)\cdot r^c(s_h,a_h)\right|$$

$$=\left|\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A}q_h(s_h,a_h|\pi_c,P^c)\cdot r^c(s_h,a_h)-\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A}q_h(s_h,a_h|\pi_c,\widehat{P}^c)\cdot r^c(s_h,a_h)\right|$$

$$(r^c(s_{sink},a):=0,\ \forall c,a)$$

$$=\left|\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A}(q_h(s_h,a_h|\pi_c,P^c)-q_h(s_h,a_h|\pi_c,\widehat{P}^c))r^c(s_h,a_h)\right|$$

$$
\begin{aligned}
&\leq \sum_{h=0}^{H-1} \sum_{s_h \in S_h} \sum_{a_h \in A} \left| q_h(s_h, a_h | \pi_c, P^c) - q_h(s_h, a_h | \pi_c, \widehat{P}^c) \right| \underbrace{|r^c(s_h, a_h)|}_{\leq 1} \\
&\leq \sum_{h=0}^{H-1} \sum_{s_h \in S_h} \sum_{a_h \in A} \left| q_h(s_h, a_h | \pi_c, P^c) - q_h(s_h, a_h | \pi_c, \widehat{P}^c) \right| \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in S_h} \sum_{a_h \in A} \pi_c(a_h|s_h) |q_h(s_h | \pi_c, P^c) - q_h(s_h | \pi_c, \widehat{P}^c)| \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in S_h} |q_h(s_h | \pi_c, P^c) - q_h(s_h | \pi_c, \widehat{P}^c)| \underbrace{\sum_{a_h \in A} \pi_c(a_h|s_h)}_{=1} \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in S_h} |q_h(s_h | \pi_c, P^c) - q_h(s_h | \pi_c, \widehat{P}^c)| \\
&= \sum_{h=0}^{H-1} \|q_h(\cdot | \pi_c, P^c) - q_h(\cdot | \pi_c, \widehat{P}^c)\|_1.
\end{aligned}
$$

Denote by G_8 the good event

$$
\forall h \in [H] : \|q_h(\cdot | \pi_c, P^c) - q_h(\cdot | \pi_c, \widehat{P}^c)\|_1 \leq \frac{4\rho|S|}{1 - \rho^2 |S|^2} h + \beta \sum_{i=1}^{h-1} |S_i|,
$$

and denote by $\overline{G_8}$ its complementary event.

By Lemma 98 we have

$$
\mathbb{P}[G_8] \geq 1 - \left(\frac{\epsilon_P}{\rho^2} |A| \sum_{i=0}^{H-1} |S_i||S_{i+1}| + \gamma \sum_{i=1}^{H-1} |S_i| \right) \geq 1 - \left(|S|^2 |A| \frac{\epsilon_P}{\rho^2} + |S|\gamma \right).
$$

If G_8 holds, then

$$
\begin{aligned}
&\sum_{h=0}^{H-1} \|q_h(\cdot | \pi_c, P^c) - q_h(\cdot | \pi_c, \widehat{P}^c)\|_1 \leq \sum_{h=0}^{H-1} \left(\frac{4\rho|S|}{1 - \rho^2 |S|^2} h + \beta \sum_{i=1}^{h-1} |S_i| \right) \\
&\leq \sum_{h=0}^{H-1} \left(\frac{4\rho|S|}{1 - \rho^2 |S|^2} H + \beta |S| \right) \leq \frac{4\rho|S|}{1 - \rho^2 |S|^2} H^2 + \beta |S| H.
\end{aligned}
$$

Otherwise,

$$
\sum_{h=0}^{H-1} \|q_h(\cdot | \pi_c, P^c) - q_h(\cdot | \pi_c, \widehat{P}^c)\|_1 \leq \sum_{h=0}^{H-1} 1 \leq H.
$$

Using total expectation low we obtain

$$
\begin{aligned}
&\mathbb{E}_{c \sim \mathcal{D}} \left[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)| \right] \\
\leq &\mathbb{P}\left[G_8\right] \mathbb{E}_{c \sim \mathcal{D}} \left[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)| \, | G_8 \right] + \mathbb{P}\left[\overline{G_8}\right] \mathbb{E}_{c \sim \mathcal{D}} \left[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)| \mid \overline{G_8} \right] \\
\leq &\mathbb{E}_{c \sim \mathcal{D}} \left[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)| \mid G_8 \right] + \mathbb{P}\left[\overline{G_8}\right] \mathbb{E}_{c \sim \mathcal{D}} \left[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)| \mid \overline{G_8} \right] \\
\leq &\frac{4\rho|S|}{1 - \rho^2 |S|^2} H^2 + \beta |S| H + \mathbb{P}\left[\overline{G_8}\right] H
\end{aligned}
$$

$$\leq \frac{4\rho|S|}{1-\rho^2|S|^2}H^2 + \beta|S|H + H|S|^2|A|\frac{\epsilon_P}{\rho^2} + \gamma H|S|$$

which proves the lemma. ■

For out parameters choice, we obtain,

Corollary 101 *Under the good events G_1, G_2 and G_3, For $\epsilon_P = \frac{\epsilon^3}{10\cdot 2^8\cdot 20^2|A||S|^6H^5}$, $\gamma = \frac{\epsilon}{20|S|H}$ $\beta = \frac{\epsilon}{20|S|H}$, $\rho = \frac{\beta}{16|S|H}$, we have that $\rho \in [0, \frac{1}{|S|})$, $\beta \geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$ and*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|] \leq \frac{\epsilon}{40|S|} + \frac{2\epsilon}{10}.$$

Proof Implied by assigning the detailed parameters to the results of Lemma 100. ■

E.4.3 ANALYSIS OF THE ERROR CAUSED BY THE REWARDS APPROXIMATION UNDER THE GOOD EVENTS

Recall that for every context $c \in \mathcal{C}$, define the following two MDPs. The intermediate MDP $\widetilde{M}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}^c, r^c)$ and the approximated MDP $\widehat{M}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}^c, \widehat{r}^c)$, where r^c is the true rewards function extended to s_{sink} by defining $r^c(s_{sink}, a) := 0, \quad \forall c \in \mathcal{C}, a \in A$. $\widehat{P}^c, \widehat{r}^c$ are the approximation of the dynamics and the rewards as defined in Algorithm 16.

Lemma 102 (expected value difference caused by rewards approximation) *Under the good events G_1, G_2 and G_4, for any context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$ it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\widetilde{M}(c)}(s_0) - V^{\pi_c}_{\widehat{M}(c)}(s_0)|] = \alpha_2 H + 2(\epsilon_R|S||A|)^{\frac{1}{3}}H + \beta|S| + \gamma|S|H,$$

where $\alpha_2^2 := \max_{h\in[H-1]} \alpha_2^2(\mathcal{F}_h^R)$.

Proof

Recall that $\widehat{r}^c(s_{sink}, a) := 0, \forall c \in \mathcal{C}, a \in A$ by definition. In addition, since r^c is the true rewards function and $s_{sink} \notin S$, r^c is not defined for s_{sink}. We naturally extended it to s_{sink} by defining $r^c(s_{sink}, a) := 0, \forall c \in \mathcal{C}, a \in A$. Hence, we can simply ignore s_{sink} as the following computation shows.

Let us recall the definition of the following subsets of S_h for every $h \in [H-1]$ given any context $c \in \mathcal{C}$.

1. $B_1^{h,c} = \{s_h \in S_h : s_h \in \widehat{S}_h^{\gamma,\beta}$ and $c \in \widehat{C}^\beta(s_h)\}$
2. $B_2^{h,c} = \{s_h \in S_h : s_h \in \widehat{S}_h^{\gamma,\beta}$ and $c \notin \widehat{C}^\beta(s_h)\}$
3. $B_3^{h,c} = \{s_h \in S_h : s_h \notin \widehat{S}_h^{\gamma,\beta}$ and $c \notin \widehat{C}^\beta(s_h)\}$
4. $B_4^{h,c} = \{s_h \in S_h : s_h \notin \widehat{S}_h^{\gamma,\beta}$ and $c \in \widehat{C}^\beta(s_h)\}$

Clearly, $\cup_{i=1}^4 B_i^h = S_h$.

By definition $s_h \in B_1^{h,c}$ if and only if for every action $a_h \in A$ we have that $(c, s_h, a_h) \in \mathcal{X}_h^{\gamma,\beta}$.

For $s_h \notin \widehat{S}_h^{\gamma,\beta}$ we have that $\mathbb{P}[c \in \widehat{\mathcal{C}}^\beta(s_h)] < \gamma$, hence,

$$\mathbb{P}_c[\exists h \in [H-1] : B_4^{h,c} \neq \emptyset] \quad = \quad \mathbb{P}_c[\exists h \in [H-1], s_h \in S_h : s_h \notin \widehat{S}_h^{\gamma,\beta}, \text{ and } c \in \widehat{\mathcal{C}}^\beta(s_h)] \quad < \quad \gamma|S|.$$

Fix a context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$. The following holds for any given context c. (Later we will take the expectation over c).

$$
\begin{aligned}
|V^{\pi_c}_{\widehat{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{M}(c)}(s_0)| &= \left|\sum_{h=0}^{H-1}\sum_{s_h\in S_h\cup\{s_{sink}\}}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,\widehat{P}^c)(r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h))\right| \\
&= \left|\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,\widehat{P}^c)(r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h))\right| \\
&\qquad \text{(By definition, } r^c(s_{sink},a)=\widehat{r}^c(s_{sink},a)=0,\ \ \forall c\in\mathcal{C}, a\in A.\text{)} \\
&\le \sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,\widehat{P}^c)|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)| \\
&= \underbrace{\sum_{h=0}^{H-1}\sum_{s_h\in B_1^{h,c}}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,\widehat{P}^c)|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)|}_{(1)} \\
&\quad + \underbrace{\sum_{h=0}^{H-1}\sum_{s_h\in B_2^{h,c}\cup B_3^{h,c}}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,\widehat{P}^c)|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)|}_{(2)} \\
&\quad + \underbrace{\sum_{h=0}^{H-1}\sum_{s_h\in B_4^{h,c}}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,\widehat{P}^c)|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)|}_{(3)}.
\end{aligned}
$$

We bound (1), (2) and (3) separately.

For (1), under the good events G_1, G_2 , G_3 and G_4, we have for all $h\in[H-1]$ that

$$\mathbb{E}_{(c,s_h,a_h)\sim\mathcal{D}_h^R}[(f_h^R(c,s_h,a_h)-r^c(s_h,a_h))^2-\alpha_2^2(\mathcal{F}_h^R)]\le\epsilon_R.$$

Since $\mathbb{E}_{\mathcal{D}_h^R}[(f_h^R(c,s_h,a_h)-r^c(s_h,a_h))^2]\ge\alpha_2^2(\mathcal{F}_h^R)$, for all $h\in[H-1]$ and $\xi\in(0,1]$ we obtain using Markov's inequality that

$$
\begin{aligned}
&\mathbb{P}_{(c,s_h,a_h)}[|f_h^R(c,s_h,a_h)-r^c(s_h,a_h)|\ge\sqrt{\alpha_2^2(\mathcal{F}^R_{s_h,a_h})+\xi^2}\ \Big|(c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}] = \\
&= \mathbb{P}_{(c,s_h,a_h)}[(f_h^R(c,s_h,a_h)-r^c(s_h,a_h))^2-\alpha_2^2(\mathcal{F}^R_{s_h,a_h})\ge\xi^2\ \Big|(c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}] \\
&\le \frac{\mathbb{E}_{\mathcal{D}_h^R}[(f_h^R(c,s_h,a_h)-r^c(s_h,a_h))^2-\alpha_2^2(\mathcal{F}_h^R)]}{\xi^2}\le\frac{\epsilon_R}{\xi^2}.
\end{aligned}
$$

Since $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$, for $a,b\in[0,1]$, it holds that

$$
\begin{aligned}
&\mathbb{P}\left[\left|f_h^R(c,s_h,a_h)-r^c(s_h,a_h)\ \right|\ \ \le\alpha_2(\mathcal{F}^R_{s_h,a_h})+\xi\Big|(c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}\right] \\
&\ge\mathbb{P}\left[\left|f_h^R(c,s_h,a_h)-r^c(s_h,a_h)\right|\ \ \le\sqrt{\alpha_2^2(\mathcal{F}^R_{s_h,a_h})+\xi^2}\ \ \Big|\ \ (c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}\right]\ge 1-\frac{\epsilon_R}{\xi^2}.
\end{aligned}
$$

Let G_5 denote the following good event.

$$\forall h\in[H-1]\ \forall s_h\in B_1^{h,c}\ \forall a\in A: |f_h^R(c,s_h,a_h)-r^c(s_h,a_h)|\le\alpha_2(\mathcal{F}^R_{s_h,a_h})+\xi$$

and denote by $\overline{G_5}$ the complementary event. By the above and union bound over $h \in [H-1]$ and $(s_h, a_h) \in B_1^{h,c} \times A$ it holds that $\mathbb{P}_c[G_5] \geq 1 - \frac{\epsilon_R}{\xi^2}|S||A|$ and $\mathbb{P}_c[\overline{G_5}] \leq \frac{\epsilon_R}{\xi^2}|S||A|$.

If G_5 holds then,

$$\begin{aligned}
(1) &= \sum_{h=0}^{H-1} \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c) \left| r^c(s_h, a_h) - \widehat{r}^c(s_h, a_h) \right| \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} \pi_c(a_h|s_h) q_h(s_h | \pi_c, \widehat{P}^c) \left| f_h^R(c, s_h, a_h) - r^c(s_h, a_h) \right| \\
&\leq \sum_{h=0}^{H-1} \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} \pi_c(a_h|s_h) q_h(s_h | \pi_c, \widehat{P}^c) (\alpha_2(\mathcal{F}_h^R) + \xi) \\
&\leq \sum_{h=0}^{H-1} \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} \pi_c(a_h|s_h) q_h(s_h | \pi_c, \widehat{P}^c) (\alpha_2 + \xi) \leq \alpha_2 H + \xi H.
\end{aligned}$$

Otherwise,

$$(1) = \sum_{h=0}^{H-1} \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c) \underbrace{\left| r^c(s_h, a_h) - \widehat{r}^c(s_h, a_h) \right|}_{\leq 1} \leq H.$$

Thus,

$$\mathbb{E}_{c \sim \mathcal{D}}[(1)] \leq \alpha_2 H + \xi H + \frac{\epsilon_R}{\xi^2}|S||A|H.$$

For (2), consider the following derivation:

$$\begin{aligned}
(2) &= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c) \underbrace{\left| r^c(s_h, a_h) - \widehat{r}^c(s_h, a_h) \right|}_{\leq 1} \\
&\leq \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c) \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} \pi_c(a_h|s_h) q_h(s_h | \pi_c, \widehat{P}^c) \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} q_h(s_h | \pi_c, \widehat{P}^c) \underbrace{\sum_{a_h \in A} \pi_c(a_h|s_h)}_{=1} \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \underbrace{q_h(s_h | \pi_c, \widehat{P}^c))}_{\leq q_h(s_h | \widehat{\pi}_{s_h}^c, \widehat{P}^c) < \beta} \leq \beta |S|.
\end{aligned}$$

Thus,

$$\mathbb{E}_{c \sim \mathcal{D}}[(2)] \leq \beta |S|.$$

For (3), let G_6 denote the good event in which $\forall h \in [H-1], B_4^{h,c} = \emptyset$. Denote by $\overline{G_6}$ the complement event of G_6. We showed that $\mathbb{P}_c[G_6] \geq 1 - \gamma|S|$ thus $\mathbb{P}_c[\overline{G_6}] \leq \gamma|S|$.

If G_6 holds, then $(3) = 0$. Otherwise,

$$\begin{aligned}(3) &= \sum_{h=0}^{H-1} \sum_{s_h \in B_4^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c) \underbrace{|r^c(s_h, a_h) - \widehat{r}^c(s_h, a_h)|}_{\leq 1} \\ &\leq \sum_{h=0}^{H-1} \underbrace{\sum_{s_h \in B_4^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c)}_{\leq 1} \leq H.\end{aligned}$$

Using total expectation we obtain

$$\begin{aligned}\mathbb{E}_{c\sim\mathcal{D}}[(3)] &= \mathbb{P}[G_6]\mathbb{E}_{c\sim\mathcal{D}}[(3)|G_6] + \mathbb{P}[\overline{G_6}]\mathbb{E}_{c\sim\mathcal{D}}[(3)|\overline{G_5}] \\ &\leq 1 \cdot 0 + \gamma|S| \cdot H \\ &= \gamma|S|H.\end{aligned}$$

Overall, by linearity of expectation and the above, we obtain for $\xi = (\epsilon_R|S|\ |A|)^{\frac{1}{3}}$ that

$$\begin{aligned}&\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\widetilde{M}(c)}(s_0) - V^{\pi_c}_{\widehat{M}(c)}(s_0)|] \\ &\leq \mathbb{E}_{c\sim\mathcal{D}}[(1)] + \mathbb{E}_{c\sim\mathcal{D}}[(2)] + \mathbb{E}_{c\sim\mathcal{D}}[(3)] \\ &\leq \alpha_2 H + \xi H + \frac{\epsilon_R}{\xi^2}|S||A|H + \beta|S| + \gamma|S|H \\ &= \alpha_2 H + 2(\epsilon_R|S||A|)^{\frac{1}{3}}H + \beta|S| + \gamma|S|H,\end{aligned}$$

as stated. ■

Corollary 103 *Under the good events G_1, G_2 and G_4, for $\gamma = \frac{\epsilon}{20|S|H}$, $\beta = \frac{\epsilon}{20|S|H}$ and $\epsilon_R = \frac{\epsilon^3}{20^3|S||A|H^3}$, we have for any context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$ that*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\widetilde{M}(c)}(s_0) - V^{\pi_c}_{\widehat{M}(c)}(s_0)|] \leq \alpha_2 H + \frac{3\epsilon}{20} + \frac{\epsilon}{20H}$$

Proof Implied by assigning the detailed parameters to the results of Lemma 102. ■

E.4.4 COMBINING VALUE DIFFERENCES CAUSED BY DYNAMICS AND REWARDS APPROXIMATION TO A SUB-OPTIMALITY BOUND

Let SP2 denote the following parameters set.

- $\gamma = \frac{\epsilon}{20|S|H} \in (0, 1)$.
- $\beta = \frac{\epsilon}{20|S|H} \in (0, 1)$.
- $\rho = \frac{\beta}{16|S|H} \in (0, \frac{1}{|S|})$.
- $\epsilon_P = \frac{\epsilon^3}{10\cdot 2^8 20^2|A||S|^6H^5}$.
- $\epsilon_R = \frac{\epsilon_3}{20^3H^4}$.

We remark that for our choice of ρ and β it holds that $\rho \in [0, \frac{1}{|S|})$ and $\beta \geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$.

Lemma 104 (expected value difference) *Under the good events G_1, G_2,G_3 and G_4, for every context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$ it holds that,*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \alpha_2 H + \frac{1}{2}\epsilon,$$

where $\mathcal{M}(c)$ is the true MDP associated with the context c and $\widehat{\mathcal{M}}(c)$ is it's the approximated model, for the parameters set SP2.

Proof For any context $c \in \mathcal{C}$, consider the intermediate MDP $\widetilde{\mathcal{M}}(c) = (S, A, \widehat{P}^c, r^c, H, s_0)$. Using triangle inequality and linearity of expectation we obtain

$$\begin{aligned}\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] &= \mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0) + V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \\ &\leq \mathbb{E}_{c\sim\mathcal{D}}[\underbrace{|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|}_{(1)}] + \mathbb{E}_{c\sim\mathcal{D}}[\underbrace{|V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|}_{(2)}]\end{aligned}$$

By Lemma 100 we have

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|] \leq \frac{4\rho|S|}{1-\rho^2|S|^2}H^2 + |S|^2|A|H\frac{\epsilon_P}{\rho^2} + \gamma|S|H + \beta|S|H.$$

By Lemma 102 we have

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\widetilde{M}(c)}(s_0) - V^{\pi_c}_{\widehat{M}(c)}(s_0)|] \leq \alpha_2 H + 2(\epsilon_R|S||A|)^{\frac{1}{3}}H + \beta|S| + \gamma|S|H.$$

Overall,

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] = \frac{4\rho|S|}{1-\rho^2|S|^2}H^2 + |S|^2|A|H\frac{\epsilon_P}{\rho^2} + 2\gamma|S|H + 2\beta|S|H + \alpha_2 H + 2(\epsilon_R|S||A|)^{\frac{1}{3}}H$$

For the parameters set SP2 we have, $\gamma = \frac{\epsilon}{20|S|H} \in (0,1)$, $\beta = \frac{\epsilon}{20|S|H} \in (0,1)$, $\rho = \frac{\beta}{16|S|H} \in (0, \frac{1}{|S|})$, $\epsilon_P = \frac{\epsilon^3}{10\cdot 2^8 20^2|A||S|^6H^5}$, $\epsilon_R = \frac{\epsilon^3}{20^3|S||A|H^3}$.

In addition it holds that $\beta < \frac{1}{2|S|}$, which implies that $0 < \rho < \frac{1}{|S|}$.

We also have that

$$2H\frac{4\rho|S|}{1-\rho^2|S|^2} = \frac{8H|S|\frac{\beta}{16|S|H}}{1-\frac{\beta^2|S|^2}{2^8|S|^2H^2}} = \frac{\frac{\beta}{2}}{1-\underbrace{\frac{\beta^2}{2^8H^2}}_{\leq 1/2}} \leq 2\frac{\beta}{2} = \beta.$$

Hence, the constrains on ρ and β are satisfied.

Finally,

$$\begin{aligned}\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] &= \frac{4\rho|S|}{1-\rho^2|S|^2}H^2 + |S|^2|A|H\frac{\epsilon_P}{\rho^2} + 2\gamma|S|H + 2\beta|S|H + \alpha_2 H + 2(\epsilon_R|S||A|)^{\frac{1}{3}}H \\ &\leq \frac{\frac{1}{4}\beta H}{1-\underbrace{\frac{\beta^2}{2^8H^2}}_{\leq 1/2}} + 2^8|S|^4|A|H^3\frac{\epsilon_P}{\beta^2} + 2\frac{\epsilon}{10} + \alpha_2 H + \frac{\epsilon}{10}\end{aligned}$$

$$\leq \frac{1}{2}\beta H + 2^8 20^2 |S|^6 |A| H^5 \frac{\epsilon_P}{\epsilon^2} + 3\frac{\epsilon}{10} + \alpha_2 H$$
$$= \frac{1}{2}\frac{\epsilon}{20|S|} + 4\frac{\epsilon}{10} + \alpha_2 H$$
$$\leq \frac{1}{2}\epsilon + \alpha_2 H,$$

as stated. ■

The following corollary shows that for our choice of parameters, all good events holds with high probability.

Corollary 105 *For the parameters set SP2 it holds that* $\mathbb{P}[G_1, G_2, G_3, G_4] \geq 1 - (\frac{\delta}{2} + \frac{\epsilon}{10})$.

Proof By Corollary 95 it holds that $\mathbb{P}[G_1, G_2, G_3, G_4] \geq 1 - (\frac{\delta}{4} + 3\delta_1 H + \frac{\epsilon_P}{\rho^2}|S|^2|A|H)$. Hence by ρ, β, ϵ_P and δ_1 choice we obtain

$$\mathbb{P}[\cap_{i\in[4]} G_i] \geq 1 - \left(\frac{\delta}{8} + 3\delta_1 H + \frac{\epsilon_P}{\rho^2}|S|^2|A|H\right)$$
$$= 1 - \frac{\delta}{2} - \frac{\epsilon_P}{\beta^2} 2^8 |S|^4 |A| H^3$$
$$= 1 - \frac{\delta}{2} - 2^8 20^2 |S|^6 |A| H^5 \frac{\epsilon_P}{\epsilon^2}$$
$$= 1 - \frac{\delta}{2} - \frac{\epsilon}{10}.$$

■

Finally, the following theorem bound the expected sub-optimality of our approximated optimal policy $\widehat{\pi}^\star$.

Theorem 106 (expected sub-optimality bound) *With probability at least* $1 - (\delta + \frac{\epsilon}{5})$ *it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha_2 H,$$

where $\pi^\star = (\pi_c^\star)_{c\in\mathcal{C}}$ *is the optimal context-dependent policy for* $\mathcal{M}$ *and* $\widehat{\pi}^\star = (\widehat{\pi}_c^\star)_{c\in\mathcal{C}}$ *is the optimal context-dependent policy for* $\widehat{\mathcal{M}}$.

Proof Assume the good events G_1, G_2, G_3 and G_4 hold.

Then, by Lemma 104, we have for $\pi^\star$

$$\left|\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)]\right| \leq \mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \frac{1}{2}\epsilon + \alpha_2 H,$$

yielding,

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)] \leq \frac{1}{2}\epsilon + \alpha_2 H.$$

Similarly, we obtain for $\widehat{\pi}_c^\star$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \leq \frac{1}{2}\epsilon + \alpha_2 H.$$

Since for all $c \in \mathcal{C}$, $\widehat{\pi}_c^\star$ is the optimal policy for $\widehat{\mathcal{M}}(c)$ we have $V^{\widehat{\pi}_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0) \geq V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)$ which implies that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)] \leq 0.$$

Since by Corollary 105 we have that $\mathbb{P}[G_1, G_2, G_3, G_4] \geq 1 - (\frac{\delta}{2} + \frac{\epsilon}{10})$, the theorem implied by summing the above three inequalities. ■

E.4.5 ADDITIONAL LEMMAS FOR BOUNDING THE SAMPLE COMPLEXITY FOR THE ℓ_2 LOSS

Lemma 107 *Let $\rho \in [0, \frac{1}{|S|})$ and $h \in [H-1]$. Assume the good events $G_1, G_2^k, G_3^k,\ \forall k \in [h]$ hold, then we have*

$$\mathbb{P}\left[\|\widehat{P}^c(\cdot|s_h,a_h) - P^c(\cdot|s_h,a_h)\|_1 \le \frac{4\rho|S|}{1-\rho^2|S|^2}\Big|(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}\right] \ge 1 - \frac{\epsilon_P}{\rho^2}|S_{h+1}|,$$

where $\widehat{P}^c$ is the dynamics defined in Algorithm 12 and

$$\|\widehat{P}^c(\cdot|s_h,a_h) - P^c(\cdot|s_h,a_h)\|_1 := \sum_{s_{h+1}\in S_{h+1}} |\widehat{P}^c(s_{h+2}|s_h,a_h) - P^c(s_{h+1}|s_h,a_h)|$$

(i.e., the entry of s_{sink} is in $\widehat{P}$ is ignored).

Proof We prove similarly to shown for Lemma 96, when using the good events G_3^k for all $k \in [h]$ guarantees for the distribution $\widetilde{D}_h^{\gamma,\beta}$ over $\widetilde{\mathcal{X}}_h^{\gamma,\beta} \times S_{h+1}$.

Recall that for $(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$ we have that $\widehat{P}^c(s_{sink}|s_h,a_h) = 0$ by $\widehat{P}^c$ definition. In addition, the true dynamics P^c is not defined for s_{sink} since $s_{sink} \notin S$. A natural extension of P^c to s_{sink} is by defining that $\forall(s,a) \in S \times A.\ \ P^c(s_{sink}|s,a) := 0$. By that extension, we have for all $(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$ that $P^c(s_{sink}|s_h,a_h) = \widehat{P}^c(s_{sink}|s_h,a_h) = 0$. Hence, we can simply ignore s_{sink} in the following analysis.

Under the good event G_3^h, it holds that

$$\begin{aligned}
&\mathbb{P}_{(c,s_h,a_h,s_{h+1})}\left[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)| \ge \rho\Big|(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}\right] = \\
&= \mathbb{P}_{\widetilde{\mathcal{D}}_h^P}[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)| \ge \rho] \\
&= \mathbb{P}_{\widetilde{\mathcal{D}}_h^P}[(f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h))^2 \ge \rho^2] \\
&\le \frac{\mathbb{E}_{\widetilde{\mathcal{D}}_h^P}[(f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h))^2]}{\rho^2} && \text{(By Markov's inequality)} \\
&\le \frac{\epsilon_P}{\rho^2}. && \text{(Under } G_h^3\text{)}
\end{aligned}$$

Hence,

$$\mathbb{P}_{(c,s_h,a_h,s_{h+1})}[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)| \le \rho] \ge 1 - \frac{\epsilon_P}{\rho^2}.$$

As $P^c(\cdot|s_h,a_h)$ is a distribution, we have for every context c that $\sum_{s_{h+1}\in S_{h+1}} P^c(s_{h+1}|s_h,a_h) = 1$.

Thus, by union bound over $s_{h+1} \in S_{h+1}$ we obtain

$$\mathbb{P}_{(c,s_h,a_h)}\left[1-\rho|S| \le \sum_{s_{h+1}\in S_{h+1}} f_h^P(c,s_h,a_h,s_{h+1}) \le 1+\rho|S|\Big|(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}\right] \ge 1 - \frac{\epsilon_P}{\rho^2}|S_{h+1}|.$$

Hence, we further conclude that

$$\begin{aligned}
&\mathbb{P}_{(c,s_h,a_h)}\left[\forall s_{h+1} \in S_{h+1}.\ \frac{P^c(s_{h+1}|s_h,a_h)-\rho}{1+\rho|S|} \le \underbrace{\frac{f_h^P(c,s_h,a_h,s_{h+1})}{\sum_{s'\in S_{h+1}} f_h^P(c,s_h,a_h,s')}}_{=\widehat{P}^c(s_{h+1}|s_h,a_h)} \le \frac{P^c(s_{h+1}|s_h,a_h)+\rho}{1-\rho|S|}\ \Big|(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}\right] \\
&\ge 1 - \frac{\epsilon_P}{\rho^2}|S_{h+1}|.
\end{aligned} \tag{13}$$

For any fixed tuple $(c, s_h, a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$ denote $S_{h+1}^+ = \{s_{h+1} \in S_{h+1} : \widehat{P}^c(s_{h+1}|s_h, a_h) \geq P^c(s_{h+1}|s_h, a_h)\}$ and consider the following derivation:

$$\begin{aligned}
\|\widehat{P}^c(\cdot|s_h, a_h) - P^c(\cdot|s_h, a_h)\|_1 &= \sum_{s_{h+1} \in S_{h+1}} |\widehat{P}^c(s_{h+1}|s_h, a_h) - P^c(s_{h+1}|s_h, a_h)| \\
&= \sum_{s_{h+1} \in S_{h+1}^+} (\widehat{P}^c(s_{h+1}|s_h, a_h) - P^c(s_{h+1}|s_h, a_h)) \\
&\quad + \sum_{s_{h+1} \in S_{h+1} \setminus S_{h+1}^+} (P^c(s_{h+1}|s_h, a_h) - \widehat{P}^c(s_{h+1}|s_h, a_h)) \\
&\leq \sum_{s_{h+1} \in S_{h+1}^+} \left(\frac{P^c(s_{h+1}|s_h, a_h) + \rho}{1 - \rho|S|} - P^c(s_{h+1}|s_h, a_h) \right) \\
&\quad + \sum_{s_{h+1} \in S_{h+1} \setminus S_{h+1}^+} \left(P^c(s_{h+1}|s_h, a_h) - \frac{P^c(s_{h+1}|s_h, a_h) - \rho}{1 + \rho|S|} \right) \\
&= \sum_{s_{h+1} \in S_{h+1}^+} \frac{P^c(s_{h+1}|s_h, a_h) + \rho - (1 - \rho|S|)P^c(s_{h+1}|s_h, a_h)}{1 - \rho|S|} \\
&\quad + \sum_{s_{h+1} \in S_{h+1} \setminus S_{h+1}^+} \frac{-P^c(s_{h+1}|s_h, a_h) + \rho + (1 + \rho|S|)P^c(s_{h+1}|s_h, a_h)}{1 + \rho|S|} \\
&= \frac{1}{1 - \rho|S|} \sum_{s_{h+1} \in S_{h+1}^+} (\rho + \rho|S|P^c(s_{h+1}|s_h, a_h)) \\
&\quad + \frac{1}{1 + \rho|S|} \sum_{s_{h+1} \in S_{h+1} \setminus S_{h+1}^+} (\rho + \rho|S|P^c(s_{h+1}|s_h, a_h)) \\
&\leq \frac{2\rho|S|}{1 - \rho|S|} + \frac{2\rho|S|}{1 + \rho|S|} \\
&= \frac{4\rho|S|}{1 - \rho^2|S|^2}.
\end{aligned}$$

By inequality 13, the above holds with probability at least $1 - \frac{\epsilon_P}{\rho^2}|S_{h+1}|$ over $(c, s_h, a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$. Hence the lemma follows.

■

Lemma 108 *Fix $\beta \in (0, 1]$ and $\rho \in [0, \frac{1}{|S|})$ such that $\beta \geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$.*

Then, for every (context-dependent) policy $\pi = (\pi_c)_{c \in \mathcal{C}}$ and a layer $h \in [H-1]$, under the good events $G_1, G_2^i, G_3^i, \forall i \in [h-1]$ the following holds.

$$\mathbb{P}_c \left[\forall k \in [h], s_k \in S_k.\ q_k(s_k|\pi_c, P^c) \geq q_k(s_k|\pi_c, \widehat{P}^c) - \frac{4\rho|S|}{1 - \rho^2|S|^2}k \right] \geq 1 - |A| \sum_{k=0}^{h-1} \frac{\epsilon_P}{\rho^2}|S_k||S_{k+1}|.$$

Proof For every context $c \in \mathcal{C}$ define the dynamics $\widetilde{P}^c$ over $S \cup \{s_{sink}\} \times A$:

$$\forall (s, a) \in S \cup \{s_{sink}\} \times A : \widetilde{P}^c(s|s_{sink}, a) = \begin{cases} 1 & \text{, if } s = s_{sink} \\ 0 & \text{, otherwise} \end{cases}.$$

In addition we define

$$\forall k \in [h-1],\ \ \forall (s_k, a_k, s_{k+1}) \in \widetilde{S}_k^{\gamma,\beta} \times A \times S_{k+1}:$$
$$\widetilde{P}^c(s_{k+1}|s_k,a_k) = \begin{cases} P^c(s_{k+1}|s_k,a_k) & \text{, if } c \in \widehat{C}^\beta(s_k) \\ 0 & \text{, otherwise} \end{cases}$$
$$\widetilde{P}^c(s_{sink}|s_k,a_k) = \begin{cases} 0 & \text{, if } c \in \widehat{\mathcal{C}}^\beta(s_k) \\ 1 & \text{, otherwise} \end{cases}$$
$$\forall k \in [h-1],\ \ \forall (s_k, a_k, s_{k+1}) \in (S_k \setminus \widetilde{S}_k^{\gamma,\beta}) \times A \times S_{k+1}:$$
$$\widetilde{P}^c(s_{k+1}|s_k,a_k) = 0, \quad \widetilde{P}^c(s_{sink}|s_k,a_k) = 1.$$

Clearly, by definition of $\widetilde{P}^c$, we have for every (context-dependent) policy π that

$$\mathbb{P}_c\left[\forall k \in [h], s_k \in S_k.\ \ q_k(s_k|\pi_c, P^c) \geq q_k(s_k|\pi_c, \widetilde{P}^c)\right] = 1. \tag{14}$$

By Lemma 107 under the good events $G_1, G_2^k, G_3^k\ \ \forall k \in [h-1]$ for any $k \in [h-1]$ it holds that

$$\mathbb{P}_{(c,s_k,a_k)}\left[\|\widehat{P}^c(\cdot|s_k,a_k) - \widetilde{P}^c(\cdot|s_k,a_k)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}\middle|(c,s_k,a_k) \in \widetilde{\mathcal{X}}_k^{\gamma,\beta}\right] \geq 1 - \frac{\epsilon_P}{\rho^2}|S_{k+1}|, \tag{15}$$

We now show that

$$\mathbb{P}_{(c,s_k,a_k)}\left[\|\widehat{P}^c(\cdot|s_k,a_k) - \widetilde{P}^c(\cdot|s_k,a_k)\|_1 = 0\middle|(c,s_k,a_k) \notin \widetilde{\mathcal{X}}_k^{\gamma,\beta}\right] = 1. \tag{16}$$

For every layer $k \in [h-1]$ we have by definition that $(c,s_k,a_k) \in \widetilde{\mathcal{X}}_k^{\gamma,\beta}$ if and only if $s_k \in \widetilde{S}_k^{\gamma,\beta}$ and $c \in \widehat{\mathcal{C}}^\beta(s_k)$.

By the definition of $\widetilde{P}^c$ and $\widehat{P}^c$ we have for every layer $k \in [h-1]$ and context $c \in \mathcal{C}$ that

$$\forall (s_k,a_k) \in (S_k \setminus \widetilde{S}_k^{\gamma,\beta}) \times A.\ \ \|\widehat{P}^c(\cdot|s_k,a_k) - \widetilde{P}^c(\cdot|s_k,a_k)\|_1 = 0$$

In addition, by definition of $\widehat{P}^c$ and $\widetilde{P}^c$, for every layer $k \in [h-1]$ if $(s_k,a_k) \in \widetilde{S}_k^{\gamma,\beta} \times A$ but $c \notin \widehat{\mathcal{C}}^\beta(s_k)$, then

$$\|\widehat{P}^c(\cdot|s_k,a_k) - \widetilde{P}^c(\cdot|s_k,a_k)\|_1 = 0.$$

Thus, equation (16) follows.

Using total probability low, equations (15) and (16) yield that

$$\mathbb{P}_{(c,s_k,a_k)}\left[\|\widehat{P}^c(\cdot|s_k,a_k) - \widetilde{P}^c(\cdot|s_k,a_k)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}\right] \geq 1 - \frac{\epsilon_P}{\rho^2}|S_{k+1}|,$$

which by union bound over $(s_k,a_k) \in S_k \times A$ for every layer $k \in [h-1]$ implies that

$$\mathbb{P}_c\left[\forall k \in [h-1], (s_k,a_k) \in S_k \times A.\ \ \|\widehat{P}^c(\cdot|s_k,a_k) - \widetilde{P}^c(\cdot|s_k,a_k)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}\right] \geq 1 - |A|\sum_{k=0}^{h-1}\frac{\epsilon_P}{\rho^2}|S_k||S_{k+1}|. \tag{17}$$

By Theorem 137 the above yields that

$$\mathbb{P}_c\left[\forall k \in [h].\ \ \|q_k(\cdot|\pi_c,\widetilde{P}^c) - q_k(\cdot|\pi_c,\widehat{P}^c)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}k\right] \geq 1 - |A|\sum_{k=0}^{h-1}\frac{\epsilon_P}{\rho^2}|S_k||S_{k+1}|,$$

which in particularly implies that

$$\mathbb{P}_c\left[\forall k \in [h], s_k \in S_k.\ \ q_k(s_k|\pi_c,\widetilde{P}^c) \geq q_k(s_k|\pi_c,\widehat{P}^c) - \frac{4\rho|S|}{1-\rho^2|S|^2}k\right] \geq 1 - |A|\sum_{k=0}^{h-1}\frac{\epsilon_P}{\rho^2}|S_k||S_{k+1}|. \tag{18}$$

Finally, the lemma follows by combining inequalities 14 and 18. ■

E.5 Analysis for the ℓ_1 Loss

E.5.1 GOOD EVENTS

For the analysis of the algorithm, we define the following good events.

Event G_1. Intuitively, it states that the approximation of the probability that $c \in \widehat{\mathcal{C}}^{\beta}(s)$ is accurate for every state $s \in S$.

Formally, let $\widehat{p}_{\beta}(s)$ be the output of Algorithm AGC (see Algorithm 14) for the state $s \in S$, and denote $p_{\beta}(s) := \mathbb{P}[c \in \widehat{\mathcal{C}}^{\beta}(s)]$. For each layer $h \in [H-1]$ we define the event G_1^h as $G_1^h = \{|\widehat{p}_{\beta}(s_h) - p_{\beta}(s_h)| \leq \gamma/4 \quad \forall s_h \in S_h\}$ and define $G_1 = \cap_{h \in [H-1]} G_1^h$.

The good event G_1 guarantees that for every layer $h \in [H-1]$ and state $s_h \in S_h$, if $\widehat{p}_{\beta}(s_h) \geq \frac{3}{4}\gamma$ then $p_{\beta}(s_h) \geq \gamma/2$, which implies that $s_h \in \widehat{S}_h^{\gamma/2,\beta}$. This implies that for every layer h we sample only $(\gamma/2, \beta)$-good states for $\widehat{P}^c$.

More impotently, if $p_{\beta}(s_h) \geq \gamma$ then $\widehat{p}_{\beta}(s_h) \geq \frac{3}{4}\gamma$. Hence, we identify every (γ, β)-good state.

Thus, under the good event G_1, for every layer $h \in [H-1]$ the approximated set $\widetilde{S}_h^{\gamma,\beta}$ satisfies

$$\widehat{S}_h^{\gamma,\beta} \subseteq \widetilde{S}_h^{\gamma,\beta} \subseteq \widehat{S}_h^{\gamma/2,\beta}.$$

The following lemma shows that for our parameters choice, G_1 holds with high probability.

Lemma 109 *For $\epsilon_2 = \gamma/4$ and $\delta_2 = \frac{\delta}{8|S|}$, we have that $\mathbb{P}[G_1] \geq 1 - \delta/8$.*

Proof For each $s \in S$ we have that $\widehat{p}_{\beta}(s)$ is calculated over $m(\epsilon_2, \delta_2) = \left\lceil \frac{\ln \frac{2}{\delta_2}}{2\epsilon_2^2} \right\rceil$ examples. By Hoeffding's inequality combined with union bound, for $\epsilon_2 = \gamma/4$ and $\delta_2 = \frac{\delta}{8|S|}$, we obtain that $\mathbb{P}[G_1] \geq 1 - \delta/8$. ■

Sampling distributions. Recall that during the algorithm, for every layer $h \in [H-1]$ we collect examples of (c, s_h, a_h) for which $\widehat{p}_{\beta}(s_h) \geq \frac{3}{4}\gamma$, which under G_1 implies that $p_{\beta}(s_h) \geq \gamma/2$, context $c \in \widehat{C}^{\beta}(s_h)$ and actions $a_h \in A$.

For every layer $h \in [H-1]$ and reachability parameters γ and β we define the *target domain* we would like to collect examples from as

$$\mathcal{X}_h^{\gamma,\beta} = \{(c, s_h, a_h) : s_h \in \widehat{S}_h^{\gamma,\beta}, c \in \widehat{\mathcal{C}}^{\beta}(s_h), a_h \in A\},$$

recalling that

$$\widehat{C}^{\beta}(s_h) = \{c \in \mathcal{C} : s_h \text{ is } \beta - \text{reachable for } \widehat{P}^c\}$$

and

$$\widehat{S}_h^{\gamma,\beta} = \{s_h \in S_h : \mathbb{P}[c \in \widehat{\mathcal{C}}^{\beta}(s_h)] \geq \gamma\}.$$

Meaning, we would like to collect sufficient number of examples of (γ, β)-good states, appropriate good context and action for each layer.

In practice, we collect examples of states $s \in \widetilde{S}_h^{\gamma,\beta}$ which also contains states $s \in \widehat{S}_h^{\gamma/2,\beta}$. Under G_1 we have the guarantee that $\widehat{S}_h^{\gamma,\beta} \subseteq \widetilde{S}_h^{\gamma,\beta} \subseteq \widehat{S}_h^{\gamma/2,\beta}$.

Hence, we define the *empirical domain*

$$\widetilde{\mathcal{X}}_h^{\gamma,\beta} = \{(c, s_h, a_h) : s_h \in \widetilde{S}_h^{\gamma,\beta}, c \in \widehat{\mathcal{C}}^{\beta}(s_h), a_h \in A\},$$

We remark that before learning layer h, we compute $\widetilde{S}_h^{\gamma,\beta}$ based on the previous layers approximation for the dynamics which are fixed, hence $\widetilde{\mathcal{X}}_h^{\gamma,\beta}$ is fixed when learning layer h.

We also remark that under G_1 it holds, since $\widehat{S}_h^{\gamma,\beta} \subseteq \widetilde{S}_h^{\gamma,\beta} \subseteq \widehat{S}_h^{\gamma/2,\beta}$ it also holds that

$$\mathcal{X}_h^{\gamma,\beta} \subseteq \widetilde{\mathcal{X}}_h^{\gamma,\beta} \subseteq \mathcal{X}_h^{\gamma/2,\beta}.$$

We consider the marginal distributions of our observations, that sampled from $\widetilde{\mathcal{X}}_h^{\gamma,\beta}$.

For the rewards denote by $\widetilde{\mathcal{D}}_h^R$ the distribution over the collected examples $((c,s,a),r) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta} \times [0,1]$, for each layer $h \in [H-1]$. It holds that

$$\begin{aligned}\widetilde{\mathcal{D}}_h^R((c,s_h,a_h),r_h) &= \mathbb{P}[((c,s_h,a_h),r_h) \in Sample^R(h) | (c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}] \\ &\propto \mathbb{P}[c | c \in \widehat{\mathcal{C}}^{\beta}(s_h)] \cdot q_h(s_h,a_h|\widehat{\pi}^c_{s_h}, P^c) \cdot \mathbb{P}[R^c(s_h,a_h) = r_h | c, s_h, a_h],\end{aligned}$$

where $\propto$ implies that we normalize to sum to 1.

Since under G_1 we have that $\mathcal{X}_h^{\gamma,\beta} \subseteq \widetilde{\mathcal{X}}_h^{\gamma,\beta}$, $\widetilde{\mathcal{D}}_h^R$ induces a marginal distribution over $\mathcal{X}_h^{\gamma,\beta} \times [0,1]$, which we denote by $\mathcal{D}_h^R$. Clearly, it holds that

$$\begin{aligned}\mathcal{D}_h^R((c,s_h,a_h),r_h) &= \mathbb{P}[((c,s_h,a_h),r_h) \in Sample^R(h) | (c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}] \\ &\propto \mathbb{P}[c | c \in \widehat{\mathcal{C}}^{\beta}(s_h)] \cdot q_h(s_h,a_h|\widehat{\pi}^c_{s_h}, P^c) \cdot \mathbb{P}[R^c(s_h,a_h) = r_h | c, s_h, a_h],\end{aligned}$$

which is the desired marginal distribution over our target domain.

Similarly, for the next state we have,

$$\begin{aligned}&\widetilde{\mathcal{D}}_h^P((c,s_h,a_h,s'), \mathbb{I}[s_{h+1} = s']) \\ &\quad = \mathbb{P}[((c,s_h,a_h,s'), \mathbb{I}[s_{h+1} = s'])) \in Sample^P(h) | (c,s_h,a_h,s') \in (\widetilde{\mathcal{X}}_h^{\gamma,\beta} \times S_{h+1})] \\ &\quad \propto \mathbb{P}[c | c \in \widehat{\mathcal{C}}^{\beta}(s_h)] \cdot q_h(s_h,a_h|\widehat{\pi}^c_{s_h}, P^c) \cdot P^c(s'|s_h,a_h),\end{aligned}$$

and we denote $\mathcal{D}_h^P$ the induced marginal distribution over $(\mathcal{X}_h^{\gamma,\beta} \times S_{h+1}) \times [0,1]$.

Remark 110 *When it is clear from the context, we use $\mathcal{D}_h^P$ and $\widetilde{\mathcal{D}}_h^P$ to also denote the induced distribution over $(c,s_h,a_h,s_{h+1}) \in \mathcal{X}_h^{\gamma,\beta} \times S_{h+1}$ and $(c,s_h,a_h,s_{h+1}) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta} \times S_{h+1}$, respectively, and drop the indicator bit. Similarly for $\mathcal{D}_h^R$ and $\widetilde{\mathcal{D}}_h^R$ we have $(c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}$ and $(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$.*

Event G_2. Intuitively it states that sufficient number of examples have been collected for every layer $h \in [H-1]$.

Formally, let G_2^h be the event that

1. At least $\max\{N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2), N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2)\}$ examples of context, state and action from the target domain, i.e., $(c,s,a) \in \mathcal{X}_h^{\gamma,\beta}$, have been collected for layer $h \in [H-1]$.
2. At least $2\max\{N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2), N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2)\}$ examples of context, state and action from the empirical domain, i.e., $(c,s,a) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$, have been collected for layer $h \in [H-1]$.

Let G_2 be the event $\cap_{h \in [H-1]} G_2^h$.

Event G_3. Intuitively states that the ERM guarantees for the approximation of the dynamics hold. Let G_3^h denote the following event (for the ℓ_1 loss) that

$$\mathbb{E}_{(c,s_h,a_h,s_{h+1}) \sim \mathcal{D}_h^P}[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)|] \leq \epsilon_P$$

and

$$\mathbb{E}_{(c,s_h,a_h,s_{h+1}) \sim \widetilde{\mathcal{D}}_h^P}[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)|] \leq \epsilon_P.$$

Recall that we assume realizability for each layer. Define $G_3 = \cap_{h \in [H-1]} G_3^h$.

The following lemma shows that if G_1 and G_2^h holds, then G_3^h holds with high probability. (We later show that G_2 holds with high probability.)

Lemma 111 *For any $h \in [H-1]$ we have $\mathbb{P}[G_3^h | G_1, G_2^h] \geq 1 - \delta_1$.*

Proof Under G_1 and G_2^h we have collected sufficient number of examples from the domain $\mathcal{X}_h^{\gamma,\beta} \times S_{h+1}$ to approximate the transition probability function of layer h, for the accuracy parameter ϵ_P and confidence parameter $\delta_1/2$. By the ERM guarantees (see E.1.1), if sufficient number of examples have been collected, then the ERM output f_h^P satisfies that $\mathbb{E}_{\mathcal{D}_h^P}[|f_h^P(c, s_h, a_h, s_{h+1}) - P^c(s_{h+1}|s_h, a_h)|] \leq \epsilon_P$, with probability at least $1 - \delta_1/2$. Similarly for $\widetilde{\mathcal{X}}_h^{\gamma,\beta} \times S_{h+1}$ it holds that $\mathbb{E}_{\widetilde{\mathcal{D}}_h^P}[|f_h^P(c, s_h, a_h, s_{h+1}) - P^c(s_{h+1}|s_h, a_h)|] \leq \epsilon_P$, with probability at least $1 - \delta_1/2$. Hence the lemma follows by union bound. ■

The following lemma shows, inductively, that if in all the previous layers $i < h$ we have that G_1, G_2^i, G_3^i hold, then G_2^h holds with high probability in the current layer h.

Lemma 112 *For each layer $h \in [H-1]$ it holds that $\mathbb{P}[G_2^h | G_1, G_2^i, G_3^i \ \forall i \in [h-1]] \geq 1 - (\delta_1 + \frac{\epsilon_P}{\rho}|S|^2|A|)$.*

Proof We prove the lemma using induction over the horizon h.

Base case. $h = 0$.

By definition, the start state s_0 is $(1,1)$-good, which implies that for s_0 we collect samples in a deterministic manner. Thus, it holds that $\mathbb{P}[G_2^0] = 1$.

Induction step. Assume the lemma holds for all $k < h$ and we show it holds for h.
Recall we collect examples of states $s_h \in S_h$ for which $\widehat{p}_\beta(s_h) \geq \frac{3}{4}\gamma$. Under G_1, if $\widehat{p}_\beta(s_h) \geq \frac{3}{4}\gamma$ then $\mathbb{P}[c \in \widehat{\mathcal{C}}^\beta(s_h)] \geq \gamma/2$. In addition, if $\mathbb{P}[c \in \widehat{\mathcal{C}}^\beta(s_h)] \geq \gamma$ then $\widehat{p}_\beta(s_h) \geq \frac{3}{4}\gamma$.

Thus, the set $\widetilde{S}_h^{\gamma,\beta}$ of approximately (γ,β)-good state for $\widehat{P}^c$ satisfies that $\widehat{S}_h^{\gamma,\beta} \subseteq \widetilde{S}_h^{\gamma,\beta} \subseteq \widehat{S}_h^{\gamma/2,\beta}$.

Given $G_1, G_2^k, G_3^k \ \forall k \in [h-1]$ hold, by Lemma 128, for β and ρ such that $\beta \geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$ it holds that

$$\mathbb{P}_c\left[\underbrace{\forall k \in [h], s_k \in S_k.\ \ q_k(s_k|\pi_c, P^c) \geq q_k(s_k|\pi_c, \widehat{P}^c) - \frac{4\rho|S|}{1-\rho^2|S|^2}k}_{(\star)}\right] \geq 1 - |A|\sum_{k=0}^{h-1}\frac{\epsilon_P}{\rho}|S_k||S_{k+1}|$$
$$\geq 1 - |A||S|^2\frac{\epsilon_P}{\rho}.$$

Claim 3 *Assume inequality $(\star)$ holds. Then the probability to collect one example of $(c, s_h, a_h) \in \mathcal{X}_h^{\gamma,\beta}$ is at least $\frac{1}{|S|}\gamma(\beta - \frac{4\rho|S|}{1-\rho^2|S|^2}h) \geq \frac{1}{|S|} \cdot \gamma \cdot \beta/2$.*

Proof Consider the process of collecting a sample, as described in Algorithm 16:

1. The algorithm/agent chooses uniformly at random $(s,a) \in \widetilde{S}_h^{\gamma,\beta} \times A$. Under the good event G_1 we have that $\widehat{S}_h^{\gamma,\beta} \subseteq \widetilde{S}_h^{\gamma,\beta}$. Hence, the probability to choose $(s,a) \in \widehat{S}_h^{\gamma,\beta} \times A$ is at least $\frac{1}{|S|}$.

2. A context $c \sim \mathcal{D}$ is sampled. By $\widehat{S}_h^{\gamma,\beta}$ definition, the probability that $c \in \widehat{\mathcal{C}}^\beta(s)$ is at least γ.

   - If $c \in \widehat{\mathcal{C}}^\beta(s)$, the agent plays $\widehat{\pi}_s^c$ to generate a trajectory where the dynamics is P^c. By $(\star)$ and $\widehat{\mathcal{C}}^\beta(s)$ definition, the probability to observe (s,a) in a trajectory generated using $\widehat{\pi}_s^c$ where the dynamics is P^c is $q_h(s|\widehat{\pi}_s^c, P^c) \geq \beta - \frac{4\rho|S|}{1-\rho^2|S|^2}h \geq \beta/2$.
   - Otherwise quite iteration.

Overall, the probability to collect one example of a triplet $(c,s,a) \in \mathcal{X}_h^{\gamma,\beta}$ is at least $\frac{1}{|S|} \cdot \gamma \cdot (\beta - \frac{4\rho|S|}{1-\rho^2|S|^2}h) \geq \frac{1}{|S|} \cdot \gamma \cdot \frac{\beta}{2}$ (since $\beta \geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$). ■

Claim 4 *Assume inequality $(\star)$ holds. Then the probability to collect one example of $(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$ is at least $\frac{\gamma}{2}(\beta - \frac{4\rho|S|}{1-\rho^2|S|^2}h) \geq \gamma \cdot \beta/4$.*

Proof Consider the process of collecting a sample, as described in Algorithm 16:

1. The algorithm/agent chooses uniformly at random $(s,a) \in \widetilde{S}_h^{\gamma,\beta} \times A$. Under the good event G_1 we have that $\widetilde{S}_h^{\gamma,\beta} \subseteq \widehat{S}_h^{\gamma/2,\beta}$.

2. A context $c \sim \mathcal{D}$ is sampled by the nature. By $\widehat{S}_h^{\gamma/2,\beta}$ definition, the probability to observe a context $c \in \widehat{\mathcal{C}}^{\beta}(s)$ is at least $\gamma/2$.

   - If $c \in \widehat{\mathcal{C}}^{\beta}(s)$, the agent plays $\widehat{\pi}_s^c$ to generate a trajectory where the dynamics is P^c. By $(\star)$ and $\widehat{\mathcal{C}}^{\beta}(s)$ definition, the probability to observe (s,a) in a trajectory generated using $\widehat{\pi}_s^c$ where the dynamics is P^c is $q_h(s|\widehat{\pi}_s^c, P^c) \geq \beta - \frac{4\rho|S|}{1-\rho^2|S|^2}h \geq \beta/2$.
   - Otherwise quite iteration.

Overall, the probability to collect one sample of some triplet $(c,s,a) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$ is at least $\gamma/2 \cdot (\beta - \frac{4\rho|S|}{1-\rho^2|S|^2}h) \geq \gamma \cdot \beta/4$ (since $\beta \geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$). ■

The above claims implies that if $(\star)$ holds, in expectation, the agent needs to experience at most $\frac{2|S|}{\gamma \cdot \beta}$ episodes to collect one example from $\mathcal{X}_h^{\gamma,\beta}$. In addition, in expectation, the agent needs to experience at most $\frac{4}{\gamma \cdot \beta}$ episodes to collect one example from $\widetilde{\mathcal{X}}_h^{\gamma,\beta}$.

Since under G_1 we have that $\mathcal{X}_h^{\gamma,\beta} \subseteq \widetilde{\mathcal{X}}_h^{\gamma,\beta} \subseteq \mathcal{X}_h^{\gamma/2,\beta}$, using multiplicative Chernoff bound we obtain that with probability at least $1-\delta_1$ after experiencing

$$T_h = \left\lceil \frac{8|S|}{\gamma \cdot \beta} \left(\ln \frac{1}{\delta_1} + 2\max\{N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2), N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2)\} \right) \right\rceil$$

episodes, the agent will collect at least $\max\{N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2)), N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2))\}$ examples from $\mathcal{X}_h^{\gamma,\beta}$ and $2\max\{N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2)), N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2))\}$ examples from $\widetilde{\mathcal{X}}_h^{\gamma,\beta}$. Recall that T_h is exactly the number of episodes we run in Algorithm 16 when learning layer h. Hence, using union bound we obtain that

$$\mathbb{P}[G_2^h | G_1, G_2^i, G_3^i \; \forall i \in [h-1]] \geq 1 - (\delta_1 + |A||S|^2\frac{\epsilon_P}{\rho}).$$

■

The following lemma shows that, given G_1, G_2 and G_3 hold with high probability.

Lemma 113 *The following holds.*

$$\mathbb{P}[G_2 \cap G_3 | G_1] \geq 1 - (2\delta_1 H + \frac{\epsilon_P}{\rho}|S|^2|A|H).$$

Proof Assume the good event G_1 holds. Recall that $G_2 = \cap_{h\in[H-1]} G_2^h$ and $G_3 = \cap_{h\in[H-1]} G_3^h$.

Let X be a random variable with support $[H-1]$ that satisfies

$$X = \min_{k\in[H-1]} \{\overline{G}_2^k \cup \overline{G}_3^k \text{ holds }\},$$

and otherwise $X = \perp$, (i.e., if G_2 and G_3 hold). In words, X is the first layer in which at least one of the good events G_2^h or G_3^h does not hold.

By X definition and Bayes rule (i.e., $\mathbb{P}[A \cap B] = \mathbb{P}[A|B] \cdot \mathbb{P}[B]$) we have

$$\begin{aligned}
\forall h \in [H].\ \mathbb{P}[X = h|G_1] &= \mathbb{P}[(\overline{G}_2^h \cup \overline{G}_3^h) \cap (\cap_{k\in[h-1]} G_2^k \cap G_3^k)|G_1] \\
&= \mathbb{P}[(\overline{G}_2^h \cup \overline{G}_3^h)|G_1, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)] \cdot \underbrace{\mathbb{P}[(\cap_{k\in[h-1]} G_2^k \cap G_3^k)|G_1]}_{\leq 1} && \text{(Bayes rule)} \\
&\leq \mathbb{P}[(\overline{G}_2^h \cup \overline{G}_3^h)|G_1, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)] \\
&= \underbrace{\mathbb{P}[\overline{G}_2^h|G_1, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)]}_{\leq \delta_1 + \frac{\epsilon_P}{\rho}|S|^2|A| \text{ by Lemma 112}} + \mathbb{P}[\overline{G}_3^h \cap G_2^h|G_1, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)] \\
& && \text{(Union of disjoint events)} \\
&\leq \delta_1 + \frac{\epsilon_P}{\rho^2}|S||A| + \mathbb{P}[\overline{G}_3^h \cap G_2^h|G_1, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)] \\
&= \delta_1 + \frac{\epsilon_P}{\rho^2}|S||A| + \mathbb{P}[\overline{G}_3^h|G_1, G_2^h, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)] \cdot \underbrace{\mathbb{P}[G_2^h|G_1 \cap (\cap_{k\in[h-1]} G_2^k \cap G_3^k)]]}_{\leq 1} \\
& && \text{(caused by rule)} \\
&\leq \delta_1 + \frac{\epsilon_P}{\rho}|S||A| + \mathbb{P}[\overline{G}_3^h|G_1, G_2^h, (\cap_{k\in[h-1]} G_2^k \cap G_3^k)] \\
&= \delta_1 + \frac{\epsilon_P}{\rho^2}|S|^2|A| + \underbrace{\mathbb{P}[\overline{G}_3^k|G_1, G_2^h]}_{\leq \delta_1 \text{ by Lemma 111}} \\
&\leq 2\delta_1 + \frac{\epsilon_P}{\rho}|S|^2|A|
\end{aligned}$$

Lastly, by G_2 and G_3 definitions we have

$$\begin{aligned}
\mathbb{P}[G_2 \cap G_3|G_1] &= 1 - \mathbb{P}[\overline{G}_2 \cup \overline{G}_3|G_1] \\
&= 1 - \mathbb{P}[\cup_{h\in[H-1]} (\overline{G}_2^h \cup \overline{G}_3^h)|G_1] \\
&= 1 - \mathbb{P}[\exists h \in [H-1].(\overline{G}_2^h \cup \overline{G}_3^h)|G_1] \\
&= 1 - \mathbb{P}[\exists h \in [H-1].X = h|G_1] \\
&= 1 - \mathbb{P}[\cup_{h\in[H-1]} \{X = h\}|G_1] \\
&\geq 1 - \sum_{h=0}^{H-1} \mathbb{P}[X = h|G_1] && \text{(Union bound.)} \\
&\geq 1 - (2\delta_1 H + \frac{\epsilon_P}{\rho}|S|^2|A|H),
\end{aligned}$$

as stated. ■

Event G_4. Intuitively states that the ERM guarantees for the approximation of the rewards function hold (for the ℓ_1 loss). Let G_4 denote the good event

$$\forall h \in [H-1]\ \ \mathbb{E}_{(c,s_h,a_h)\sim\mathcal{D}_h^R}[|f_h^R(c, s_h, a_h) - r^c(s_h, a_h)|] \leq \epsilon_R + \alpha_1(\mathcal{F}_h^R).$$

The following lemma shows that if G_1 and G_2 hold, the G_4 hold with high probability.

Lemma 114 *It holds that* $\mathbb{P}[G_4|G_1,G_2] \geq 1-\delta_1 H$.

Proof Since G_1 and G_2 hold, for every layer $h \in [H-1]$ sufficient number of examples $((c,s_h,a_h),r_h) \in \mathcal{X}_h^{\gamma,\beta} \times [0,1]$ have been collected for the ERM to output a function f_h^R the satisfies

$$\mathbb{E}_{(c,s_h,a_h)\sim\mathcal{D}_h^R}[|f_h^R(c,s_h,a_h) - r^c(s_h,a_h)|] \leq \epsilon_R + \alpha_2^2(\mathcal{F}_h^R)$$

with probability at least $1-\delta_1$. Hence, the lemma follows by the ERM guarantees (see E.1.1) and an union bound over every layer $h \in [H-1]$. ■

The following corollary shows that all four good events hold with high probability.

Corollary 115 *It holds that*

$$\mathbb{P}[G_1,G_2,G_3,G_4] \geq 1-\left(\frac{\delta}{8}+3\delta_1 H+\frac{\epsilon_P}{\rho}|S|^2|A|H\right).$$

Proof Followed from union bound over the results of Lemmas 109, 113 and 114. ■

E.5.2 ANALYSIS OF THE ERROR CAUSED BY THE DYNAMICS APPROXIMATION UNDER THE GOOD EVENTS

In the following analysis, for any context $c \in \mathcal{C}$ we consider an intermediate MDP associated with it: $\widetilde{\mathcal{M}}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}^c, r^c, s_0, H)$, where $\widehat{P}^c$ is the approximation of the dynamics P^c and r^c is the true rewards function extended to s_{sink} by defining that $r^c(s_{sink},a) := 0, \ \ \forall c \in \mathcal{C}, \ a \in A$. Recall the true MDP associated with this context is

$$\mathcal{M}(c) = (S, A, P^c, r^c, s_0, H).$$

Lemma 116 *Let* $\rho \in [0, \frac{1}{|S|})$ *and* $h \in [H-1]$. *Assume the good events* $G_1, G_2^k, G_3^k, \ \forall k \in [h]$ *hold, then it holds that*

$$\mathbb{P}\left[\|\widehat{P}^c(\cdot|s_h,a_h) - P^c(\cdot|s_h,a_h)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}\middle|(c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}\right] \geq 1-\frac{\epsilon_P}{\rho}|S_{h+1}|,$$

where $\widehat{P}^c$ *is the dynamics defined in Algorithm 12 and*

$$\|\widehat{P}^c(\cdot|s_h,a_h) - P^c(\cdot|s_h,a_h)\|_1 := \sum_{s_{h+1}\in S_{h+1}} |\widehat{P}^c(s_{h+1}|s_h,a_h) - P^c(s_{h+1}|s_h,a_h)|$$

(i.e., the entry of s_{sink} *in* $\widehat{P}^c$ *is ignored).*

Proof

Under G_1 it holds that $\mathcal{X}_h^{\gamma,\beta} \subseteq \widetilde{\mathcal{X}}_h^{\gamma,\beta}$. Recall that for all $(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$ we have that $\widehat{P}^c(s_{sink}|s_h,a_h) = 0$ by $\widehat{P}^c$ definition. Hence, for all $(c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}$ we have that $\widehat{P}^c(s_{sink}|s_h,a_h) = 0$.

In addition, the true dynamics P^c is not defined for s_{sink} since $s_{sink} \notin S$. A natural extension of P^c to s_{sink} is by defining that $\forall (s,a) \in S \times A. \ \ P^c(s_{sink}|s,a) := 0$. By that extension, we have for all $(c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}$ that $P^c(s_{sink}|s_h,a_h) = \widehat{P}^c(s_{sink}|s_h,a_h) = 0$. Hence, we can simply ignore s_{sink} in the following analysis.

Under the good event G_3^h, by Markov's inequality we have

$$\begin{aligned}
&\mathbb{P}_{(c,s_h,a_h,s_{h+1})}\left[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)| \geq \rho \Big| (c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}\right] = \\
&= \mathbb{P}_{\mathcal{D}_h^P}[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)| \geq \rho] \\
&\leq \frac{\mathbb{E}_{\mathcal{D}_h^P}[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)|]}{\rho} && \text{(By Markov's inequality)} \\
&\leq \frac{\epsilon_P}{\rho}. && \text{(Under } G_h^3\text{)}
\end{aligned}$$

Hence,

$$\mathbb{P}_{(c,s_h,a_h,s_{h+1})}\left[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)| \leq \rho \Big| (c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}\right] \geq 1 - \frac{\epsilon_P}{\rho}.$$

Since $P^c(\cdot|s_h,a_h)$ is a distribution over $s_{h+1} \in S_{h+1}$, we have that $\sum_{s_{h+1}\in S_{h+1}} P^c(s_{h+1}|s_h,a_h) = 1$. Thus, by union bound applied on $s_{h+1} \in S_{h+1}$, we obtain

$$\mathbb{P}_{(c,s_h,a_h)}\left[1-\rho|S| \leq \sum_{s_{h+1}\in S_{h+1}} f_h^P(c,s_h,a_h,s_{h+1}) \leq 1+\rho|S| \Big| (c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}\right] \geq 1 - \frac{\epsilon_P}{\rho}|S_{h+1}|.$$

Hence, we further conclude that

$$\begin{aligned}
&\mathbb{P}_{(c,s_h,a_h)}\left[\forall s_{h+1} \in S_{h+1}. \frac{P^c(s_{h+1}|s_h,a_h)-\rho}{1+\rho|S|} \leq \underbrace{\frac{f_h^P(c,s_h,a_h,s_{h+1})}{\sum_{s'\in S_{h+1}} f_h^P(c,s_h,a_h,s')}}_{=\widehat{P}^c(s_{h+1}|s_h,a_h)} \leq \frac{P^c(s_{h+1}|s_h,a_h)+\rho}{1-\rho|S|} \;\Big|\; (c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}\right] \\
&\geq 1 - \frac{\epsilon_P}{\rho}|S_{h+1}|.
\end{aligned} \tag{19}$$

Fix a tuple $(c,s_h,a_h) \in \mathcal{X}_h^{\gamma,\beta}$ and assume the event of inequality (19) holds.
Denote $S_{h+1}^+ = \{s_{h+1} \in S_{h+1} : \widehat{P}^c(s_{h+1}|s_h,a_h) \geq P^c(s_{h+1}|s_h,a_h)\}$ and consider the following derivation.

$$\begin{aligned}
\|\widehat{P}^c(\cdot|s_h,a_h) - P^c(\cdot|s_h,a_h)\|_1 &= \sum_{s_{h+1}\in S_{h+1}} |\widehat{P}^c(s_{h+1}|s_h,a_h) - P^c(s_{h+1}|s_h,a_h)| \\
&= \sum_{s_{h+1}\in S_{h+1}^+} (\widehat{P}^c(s_{h+1}|s_h,a_h) - P^c(s_{h+1}|s_h,a_h)) \\
&\quad + \sum_{s_{h+1}\in S_{h+1}\setminus S_{h+1}^+} (P^c(s_{h+1}|s_h,a_h) - \widehat{P}^c(s_{h+1}|s_h,a_h)) \\
&\leq \sum_{s_{h+1}\in S_{h+1}^+} \left(\frac{P^c(s_{h+1}|s_h,a_h)+\rho}{1-\rho|S|} - P^c(s_{h+1}|s_h,a_h)\right) \\
&\quad + \sum_{s_{h+1}\in S_{h+1}\setminus S_{h+1}^+} \left(P^c(s_{h+1}|s_h,a_h) - \frac{P^c(s_{h+1}|s_h,a_h)-\rho}{1+\rho|S|}\right) \\
&= \sum_{s_{h+1}\in S_{h+1}^+} \frac{P^c(s_{h+1}|s_h,a_h)+\rho-(1-\rho|S|)P^c(s_{h+1}|s_h,a_h)}{1-\rho|S|}
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{s_{h+1}\in S_{h+1}\setminus S^+_{h+1}}\frac{-P^c(s_{h+1}|s_h,a_h)+\rho+(1+\rho|S|)P^c(s_{h+1}|s_h,a_h)}{1+\rho|S|}\\
=&\frac{1}{1-\rho|S|}\sum_{s_{h+1}\in S^+_{h+1}}(\rho+\rho|S|P^c(s_{h+1}|s_h,a_h))\\
&+\frac{1}{1+\rho|S|}\sum_{s_{h+1}\in S_{h+1}\setminus S^+_{h+1}}(\rho+\rho|S|P^c(s_{h+1}|s_h,a_h))\\
\le&\frac{2\rho|S|}{1-\rho|S|}+\frac{2\rho|S|}{1+\rho|S|}\\
=&\frac{4\rho|S|}{1-\rho^2|S|^2}.
\end{aligned}
$$

By inequality (19), the above holds with probability at least $1-\frac{\epsilon_P}{\rho}|S_{h+1}|$ over $(c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}$. Hence the lemma follows.

■

Lemma 117 *For the parameters choice $\beta=\frac{\epsilon}{20|S|H}\in(0,1)$, $\rho=\frac{\beta}{16|S|H}\in(0,\frac{1}{|S|})$, and $\epsilon_P=\frac{\epsilon^2}{10\cdot16\cdot20|A||S|^4H^3}$ we have $\beta\ge 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$. In addition, under the good events G_1, G_2^k and G_3^k for all $k\in[h]$ it holds that*

$$
\mathbb{P}\left[\|\widehat{P}^c(\cdot|s_h,a_h)-P^c(\cdot|s_h,a_h)\|_1\le\frac{\epsilon}{40|S|H^2}\middle|(c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}\right]\ge 1-\frac{\epsilon}{10|S||A|H}.
$$

Proof An immediate implication of lemma 96. ■

Lemma 118 (occupancy measure difference) *Under the good events G_1, G_2 and G_3, we have for any (context-dependent) policy $\pi=(\pi_c)_{c\in\mathcal{C}}$ that*

$$
\mathbb{P}_c\left[\forall h\in[H].\quad\|q_h(\cdot|\pi_c,P^c)-q_h(\cdot|\pi_c,\widehat{P}^c)\|_1\le\frac{4\rho|S|}{1-\rho^2|S|^2}h+\beta\sum_{k=1}^{h-1}|S_k|\right]\ge 1-\left(\frac{\epsilon_P}{\rho}|A|\sum_{i=0}^{H-1}|S_i||S_{i+1}|+\gamma\sum_{i=1}^{H-1}|S_i|\right)
$$

for a fixed $\rho\in[0,\frac{1}{|S|})$ and $\beta\in(0,1]$ for which $\beta\ge 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$, where

$$
\forall h\in[H].\ \|q_h(\cdot|\pi_c,P^c)-q_h(\cdot|\pi_c,\widehat{P}^c)\|_1:=\sum_{s_h\in S_h}|q_h(s_h|\pi_c,P^c)-q_h(s_h|\pi_c,\widehat{P}^c)|
$$

(i.e., $q_h(s_{sink}|\pi_c,\widehat{P}^c)$ is ignored for all $h\in[H]$).

Remark 119 *Since $s_{sink}\notin S$, $q_h(s_{sink}|\pi_c,P^c)$ is not defined for the true dynamics P^c. In addition, by $\widehat{P}^c$ definition, from the sink there are no transitions to any other state, hence, we can simply ignore it in the following analysis.*

Proof We will show the lemma by induction over the horizon, h.

For the base case $h=0$ the claim holds trivially (with probability 1) since the start state s_0 is unique.

For the induction step, assume that it holds up to layer h, namely

$$
\mathbb{P}\Big[\forall k\in[h].\quad\|q_k(\cdot|\pi_c,P^c)-q_k(\cdot|\pi_c,\widehat{P}^c)\|_1\le\frac{4\rho|S|}{1-\rho^2|S|^2}k+\beta\sum_{i=1}^{k-1}|S_i|\Big]\ge 1-\left(\frac{\epsilon_P}{\rho}|A|\sum_{i=0}^{h-1}|S_i||S_{i+1}|+\gamma\sum_{i=1}^{h-1}|S_i|\right)
$$

and prove for layer $h+1$.

by Lemma 116 it holds that

$$\mathbb{P}_{(c,s_h,a_h)}\left[\|\widehat{P}^{c'}(\cdot|s_h,a_h)-P^{c'}(\cdot|s_h,a_h)\|_1 \le \frac{4\rho|S|}{1-\rho^2|S|^2}\Big|(c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}\right] \ge 1-\frac{\epsilon_P}{\rho}|S_{h+1}|.$$

Consider the following derivation for any (fixed) context c. (Later we will take the probability over $c\sim\mathcal{D}$.)

$$\begin{aligned}
&\|q_{h+1}(\cdot|\pi_c,P^c)-q_{h+1}(\cdot|\pi_c,\widehat{P}^c)\|_1\\
&=\sum_{s_{h+1}\in S_{h+1}}|q_{h+1}(s_{h+1}|\pi_c,P^c)-q_{h+1}(s_{h+1}|\pi_c,\widehat{P}^c)|\\
&=\sum_{s_{h+1}\in S_{h+1}}|\sum_{s_h\in S_h}\sum_{a_h\in A}(q_h(s_h|\pi_c,P^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h)-q_h(s_h|\pi_c,\widehat{P}^c)\pi_c(a_h|s_h)\widehat{P}^c(s_{h+1}|s_h,a_h))|\\
&\le\underbrace{\sum_{s_h\in S_h}\sum_{a_h\in A}\sum_{s_{h+1}\in S_{h+1}}|q_h(s_h|\pi_c,P^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h)-q_h(s_h|\pi_c,\widehat{P}^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h)|}_{(1)}\\
&\quad+\underbrace{\sum_{s_h\in S_h}\sum_{a_h\in A}\sum_{s_{h+1}\in S_{h+1}}|q_h(s_h|\pi_c,\widehat{P}^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h)-q_h(s_h|\pi_c,\widehat{P}^c)\pi_c(a_h|s_h)\widehat{P}^c(s_{h+1}|s_h,a_h)|}_{(2)}.
\end{aligned}$$

We bound (1) and (2) separately.

For (1), since $P^c(\cdot|s_h,a_h)$ and $\pi_c(\cdot|s_h)$ are distributions, we have

$$\begin{aligned}
(1)&=\sum_{s_h\in S_h}|q_h(\cdot|\pi_c,P^c)-q_h(s_h|\pi_c,\widehat{P}^c)|\sum_{a_h\in A}\pi_c(a_h|s_h)\sum_{s_{h+1}\in S_{h+1}}P^c(s_{h+1}|s_h,a_h)\\
&=\|q_h(\cdot|\pi_c,P^c)-q_h(\cdot|\pi_c,\widehat{P}^c)\|_1\sum_{a_h\in A}\pi_c(a_h|s_h)\sum_{s_{h+1}\in S_{h+1}}P^c(s_{h+1}|s_h,a_h)\\
&=\|q_h(\cdot|\pi_c,P^c)-q_h(\cdot|\pi_c,\widehat{P}^c)\|_1,
\end{aligned}$$

which holds with probability 1.

To bound (2), for all $h\in[H-1]$, let us define the following subsets of S_h for any given context c.

1. $B_1^{h,c}=\{s_h\in S_h: s_h\in\widehat{S}^{\gamma,\beta}\text{ and }c\in\widehat{C}^{\beta}(s_h)\}$.
2. $B_2^{h,c}=\{s_h\in S_h: s_h\in\widehat{S}^{\gamma,\beta}\text{ and }c\notin\widehat{C}^{\beta}(s_h)\}$.
3. $B_3^{h,c}=\{s_h\in S_h: s_h\notin\widehat{S}^{\gamma,\beta}\text{ and }c\notin\widehat{C}^{\beta}(s_h)\}$.
4. $B_4^{h,c}=\{s_h\in S_h: s_h\notin\widehat{S}^{\gamma,\beta}\text{ and }c\in\widehat{C}^{\beta}(s_h)\}$.

Clearly, for every layer $h\in[H-1]$ and context $c\in\mathcal{C}$ it holds that $\cup_{i=1}^4 B_i^{h,c}=S_h$.

By definition of $B_1^{h,c}$, for every layer $h\in[H-1]$ we have that $s_h\in B_1^{h,c}$ if and only if for every action $a_h\in A$ it holds that $(c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}$.

By definition of $B_4^{h,c}$, for every layer $h\in[H-1]$ we have that

$$\mathbb{P}_c[B_4^{h,c}\neq\emptyset]=\mathbb{P}_c[\exists s_h\in S_h: s_h\notin\widehat{S}_h^{\gamma,\beta}\text{ and }c\in\widehat{C}^{\beta}(s_h)]\le\gamma|S_h|.$$

Thus, for every $h\in[H-1]$ we have $\mathbb{P}_c[B_4^{h,c}=\emptyset]\ge 1-\gamma|S_h|$.

In the following, we assume that $B_4^{h,c} = \emptyset$, since $\mathbb{P}_c[B_4^{h,c} = \emptyset] \geq 1 - \gamma|S_h|$, it will only add $\gamma|S_h|$ to the probability of the error.

Consider the following derivation

$$
\begin{aligned}
(2) &= \sum_{s_h \in B_1} \sum_{a_h \in A} \sum_{s_{h+1} \in S_{h+1}} |q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h) - q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)\widehat{P}^c(s_{h+1}|s_h,a_h)| \\
&+ \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} \sum_{s_{h+1} \in S_{h+1}} |q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h) - q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)\widehat{P}^c(s_{h+1}|s_h,a_h)| \\
&+ \sum_{s_h \in B_4^{h,c}} \sum_{a_h \in A} \sum_{s_{h+1} \in S_{h+1}} |q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)P^c(s_{h+1}|s_h,a_h) - q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h)\widehat{P}^c(s_{h+1}|s_h,a_h)| \\
&= \sum_{s_h \in B_1^{h,c}} q_h(s_h|\pi_c, \widehat{P}^c) \sum_{a_h \in A} \pi_c(a_h|s_h) \sum_{s_{h+1} \in S_{h+1}} |P^c(s_{h+1}|s_h,a_h) - \widehat{P}^c(s_{h+1}|s_h,a_h)| \\
&+ \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} q_h(s_h|\pi_c, \widehat{P}^c)\pi_c(a_h|s_h) \underbrace{\sum_{s_{h+1} \in S_{h+1}} |P^c(s_{h+1}|s_h,a_h) - \widehat{P}^c(s_{h+1}|s_h,a_h)|}_{\leq 1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (B_4^{h,c} = \emptyset \text{ w.p. at least } 1 - \gamma|S_h|) \\
&\leq \sum_{s_h \in B_1^{h,c}} q_h(s_h|\pi_c, \widehat{P}^c) \sum_{a_h \in A} \pi_c(a_h|s_h) \|P^c(\cdot|s_h,a_h) - \widehat{P}^c(\cdot|s_h,a_h)\|_1 + \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} q_h(s_h|\pi_c, \widehat{P}^c) \underbrace{\sum_{a_h \in A} \pi_c(a_h|s_h)}_{=1} \\
&\leq \underbrace{\sum_{s_h \in B_1^{h,c}} q_h(s_h|\pi_c, \widehat{P}^c) \sum_{a_h \in A} \pi_c(a_h|s_h)}_{\leq 1} \frac{4\rho|S|}{1-\rho^2|S|^2} + \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \underbrace{q_h(s_h|\pi_c, \widehat{P}^c)}_{\leq q_h(s_h|\widehat{\pi}^c_{s_h}, \widehat{P}^c) < \beta} \underbrace{\sum_{a_h \in A} \pi_c(a_h|s_h)}_{=1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By Lemma 96 and union bound over } (s_h,a_h) \in B_1^{h,c} \times A \text{, holds w.p. at least } 1 - |A||S_h|\tfrac{\epsilon_P}{\rho}|S_{h+1}|) \\
&= \frac{4\rho|S|}{1-\rho^2|S|^2} + \beta|S_h|.
\end{aligned}
$$

Hence,

$$\mathbb{P}_c\left[(2) \leq \frac{4\rho|S|}{1-\rho^2|S|^2} + \beta|S_h|\right] \geq 1 - \left(|A||S_h|\frac{\epsilon_P}{\rho}|S_{h+1}| + \gamma|S_h|\right).$$

In addition, we proved above that

$$\mathbb{P}_c\left[\|q_{h+1}(\cdot|\pi_c, P^c) - q_{h+1}(\cdot|\pi_c, \widehat{P}^c)\|_1 \leq (1) + (2)\right] = 1,$$

and

$$\mathbb{P}_c\left[(1) = \|q_h(\cdot|\pi_c, P^c) - q_h(\cdot|\pi_c, \widehat{P}^c)\|_1\right] = 1.$$

Thus, by combining all the above inequalities with the induction hypothesis we obtain

$$
\begin{aligned}
&\mathbb{P}_c\left[\forall k \in [h+1].\ \ \|q_k(\cdot|\pi_c, P^c) - q_k(\cdot|\pi_c, \widehat{P}^c)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}k + \beta\sum_{i=1}^{k-1}|S_i|\right] \\
&\geq \mathbb{P}_c\Bigg[\forall k \in [h].\ \ \|q_k(\cdot|\pi_c, P^c) - q_k(\cdot|\pi_c, \widehat{P}^c)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}k + \beta\sum_{i=1}^{k-1}|S_i| \text{ and} \\
&(1) + (2) \leq \frac{4\rho|S|}{1-\rho^2|S|^2}(h+1) + \beta\sum_{i=1}^{h}|S_i|\Bigg]
\end{aligned}
$$

$$\geq \mathbb{P}_c\left[\forall k \in [h].\quad \|q_k(\cdot|\pi_c, P^c) - q_k(\cdot|\pi_c, \widehat{P}^c)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}k + \beta\sum_{i=1}^{k-1}|S_i| \text{ and } (2) \leq \frac{4\rho|S|}{1-\rho^2|S|^2} + \beta|S_h|\right]$$

$$\geq 1 - \left(\frac{\epsilon_P}{\rho}|A|\sum_{i=0}^{h-1}|S_i||S_{i+1}| + \gamma\sum_{i=1}^{h-1}|S_i| + \frac{\epsilon_P}{\rho}|A||S_h||S_{h+1}| + \gamma|S_h|\right) = 1 - \left(\frac{\epsilon_P}{\rho}|A|\sum_{i=0}^{h}|S_i||S_{i+1}| + \gamma\sum_{i=1}^{h}|S_i|\right),$$

as stated.

■

Lemma 120 (expected value difference caused by dynamics approximation) *Then, under the good events G_1, G_2 and G_3, for every context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$ we have*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|] \leq \frac{4\rho|S|}{1-\rho^2|S|^2}H^2 + \beta|S|H + H|S|^2|A|\frac{\epsilon_P}{\rho} + \gamma H|S|,$$

where $\rho \in [0, \frac{1}{|S|})$ and $\beta \in (0, 1]$ for which $\beta \geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$.

Proof Recall that the true rewards function is not defined for s_{sink}, since $s_{sink} \notin S$. For the intermediate MDP $\widetilde{\mathcal{M}}(c)$ we extended r^c to s_{sink} by defining $\forall a \in A.\ r^c(s_{sink}, a) = 0$ for every context $c \in \mathcal{C}$. Since P^c is also not defined for s_{sink}, we can simply omit s_{sink}, as the second equality in the following derivation shows.

Consider the following derivation for any fixed $c \in \mathcal{C}$. (Later we will take the expectation over c.)

$$\begin{aligned}
&|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)| \\
&= \left|\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} q_h(s_h, a_h|\pi_c, P^c)\cdot r^c(s_h, a_h) - \sum_{h=0}^{H-1}\sum_{s_h\in S_h\cup\{s_{sink}\}}\sum_{a_h\in A} q_h(s_h, a_h|\pi_c, \widehat{P}^c)\cdot r^c(s_h, a_h)\right| \\
&= \left|\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} q_h(s_h, a_h|\pi_c, P^c)\cdot r^c(s_h, a_h) - \sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} q_h(s_h, a_h|\pi_c, \widehat{P}^c)\cdot r^c(s_h, a_h)\right| \\
&\qquad\qquad (r^c(s_{sink}, a) := 0,\ \forall c, a) \\
&= \left|\sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} (q_h(s_h, a_h|\pi_c, P^c) - q_h(s_h, a_h|\pi_c, \widehat{P}^c))r^c(s_h, a_h)\right| \\
&\leq \sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} |q_h(s_h, a_h|\pi_c, P^c) - q_h(s_h, a_h|\pi_c, \widehat{P}^c)|\underbrace{|r^c(s_h, a_h)|}_{\leq 1} \\
&\leq \sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} |q_h(s_h, a_h|\pi_c, P^c) - q_h(s_h, a_h|\pi_c, \widehat{P}^c)| \\
&= \sum_{h=0}^{H-1}\sum_{s_h\in S_h}\sum_{a_h\in A} \pi_c(a_h|s_h)|q_h(s_h|\pi_c, P^c) - q_h(s_h|\pi_c, \widehat{P}^c)| \\
&= \sum_{h=0}^{H-1}\sum_{s_h\in S_h} |q_h(s_h|\pi_c, P^c) - q_h(s_h|\pi_c, \widehat{P}^c)|\underbrace{\sum_{a_h\in A}\pi_c(a_h|s_h)}_{=1} \\
&= \sum_{h=0}^{H-1}\sum_{s_h\in S_h} |q_h(s_h|\pi_c, P^c) - q_h(s_h|\pi_c, \widehat{P}^c)|
\end{aligned}$$

$$= \sum_{h=0}^{H-1} \|q_h(\cdot|\pi_c, P^c) - q_h(\cdot|\pi_c, \widehat{P}^c)\|_1.$$

Denote by G_8 the good event

$$\forall h \in [H] : \|q_h(\cdot|\pi_c, P^c) - q_h(\cdot|\pi_c, \widehat{P}^c)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}h + \beta\sum_{i=1}^{h-1}|S_i|,$$

and denote by $\overline{G_8}$ its complementary event.

By Lemma 118 we have

$$\mathbb{P}_c[G_8] \geq 1 - \left(\frac{\epsilon_P}{\rho}|A|\sum_{i=0}^{H-1}|S_i||S_{i+1}| + \gamma\sum_{i=1}^{H-1}|S_i|\right) \geq 1 - \left(|S|^2|A|\frac{\epsilon_P}{\rho} + |S|\gamma\right).$$

If G_8 holds, then

$$\begin{aligned}\sum_{h=0}^{H-1}\|q_h(\cdot|\pi_c, P^c) - q_h(\cdot|\pi_c, \widehat{P}^c)\|_1 &\leq \sum_{h=0}^{H-1}\frac{4\rho|S|}{1-\rho^2|S|^2}h + \beta\sum_{i=1}^{h-1}|S_i| \\ &\leq \sum_{h=0}^{H-1}\frac{4\rho|S|}{1-\rho^2|S|^2}H + \beta|S| \leq \frac{4\rho|S|}{1-\rho^2|S|^2}H^2 + \beta|S|H.\end{aligned}$$

Otherwise,

$$\sum_{h=0}^{H-1}\|q_h(\cdot|\pi_c, P^c) - q_h(\cdot|\pi_c, \widehat{P}^c)\|_1 \leq \sum_{h=0}^{H-1}1 \leq H.$$

Using total expectation low we obtain

$$\begin{aligned}&\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|] \\ \leq&\mathbb{P}[G_8]\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)||G_8] + \mathbb{P}[\overline{G_8}]\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)||\overline{G_8}] \\ \leq&\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)||G_8] + \mathbb{P}[\overline{G_8}]\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)||\overline{G_8}] \\ \leq&\frac{4\rho|S|}{1-\rho^2|S|^2}H^2 + \beta|S|H + \mathbb{P}[\overline{G_8}]H \\ \leq&\frac{4\rho|S|}{1-\rho^2|S|^2}H^2 + \beta|S|H + H|S|^2|A|\frac{\epsilon_P}{\rho} + \gamma H|S|.\end{aligned}$$

■

Corollary 121 *Under the good events G_1, G_2 and G_3, for the parameter choice $\gamma = \frac{\epsilon}{20|S|H}$, $\beta = \frac{\epsilon}{20|S|H}$, $\rho = \frac{\beta}{16|S|H}$, $\epsilon_P = \frac{\epsilon^2}{10\cdot16\cdot20|A||S|^4H^3}$, we have for every policy π that*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|] \leq \frac{\epsilon}{40|S|} + \frac{2\epsilon}{10} \leq 0.225\epsilon.$$

Proof Implied by assigning the detailed parameters to the results of Lemma 120. ■

E.5.3 ANALYSIS OF THE ERROR CAUSED BY THE REWARDS APPROXIMATION UNDER THE GOOD EVENTS

Recall that for every $c \in \mathcal{C}$, define the following two MDPs. The intermediate MDP $\widetilde{M}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}^c, r^c)$, and the approximated MDP $\widehat{M}(c) = (S \cup \{s_{sink}\}, A, \widehat{P}^c, \widehat{r}^c)$ where r^c is the true rewards function extended to s_{sink} by defining $r^c(s_{sink}, a) := 0, \quad \forall c \in \mathcal{C}, a \in A$. $\widehat{P}^c$, $\widehat{r}^c$ are the approximation of the dynamics and the rewards as defined algorithm 16.

Lemma 122 (expected value difference caused by rewards approximation) *Then, under the good events G_1, G_2 and G_4, for every context-dependent policy $\pi = (\pi_c)_{c \in \mathcal{C}}$ we have*

$$\mathbb{E}_{c \sim \mathcal{D}}[|V^{\pi_c}_{\widetilde{M}(c)}(s_0) - V^{\pi_c}_{\widehat{M}(c)}(s_0)|] \leq \alpha_1 H + 2(\epsilon_R |S||A|)^{\frac{1}{2}} H + \beta |S| + \gamma |S| H$$

where $\alpha_1 := \max_{h \in [H-1]} \alpha_1(\mathcal{F}_h^R)$.

Proof Recall that $\widehat{r}^c(s_{sink}, a) := 0, \forall c \in \mathcal{C}, a \in A$ by definition. In addition, since r^c is the true rewards function and $s_{sink} \notin S$, r^c is not defined for s_{sink}. We naturally extended it to s_{sink} by defining $r^c(s_{sink}, a) := 0, \forall c \in \mathcal{C}, a \in A$. Hence, we can simply ignore s_{sink} as the following analysis shows.

Let us recall the definition of the following subsets of S_h for every $h \in [H-1]$ and a given context $c \in \mathcal{C}$.

1. $B_1^{h,c} = \{s_h \in S_h : s_h \in \widehat{S}_h^{\gamma,\beta}$ and $c \in \widehat{C}^{\beta}(s_h)\}$.
2. $B_2^{h,c} = \{s_h \in S_h : s_h \in \widehat{S}_h^{\gamma,\beta}$ and $c \notin \widehat{C}^{\beta}(s_h)\}$.
3. $B_3^{h,c} = \{s_h \in S_h : s_h \notin \widehat{S}_h^{\gamma,\beta}$ and $c \notin \widehat{C}^{\beta}(s_h)\}$.
4. $B_4^{h,c} = \{s_h \in S_h : s_h \notin \widehat{S}_h^{\gamma,\beta}$ and $c \in \widehat{C}^{\beta}(s_h)\}$.

Clearly, $\cup_{i=1}^4 B_i^h = S_h$.

By definition $s_h \in B_1^{h,c}$ if and only if for every action $a_h \in A$ we have that $(c, s_h, a_h) \in \mathcal{X}_h^{\gamma,\beta}$.

For $s_h \notin \widetilde{S}_h^{\gamma,\beta}$ we have that $\mathbb{P}_c[c \in \widehat{\mathcal{C}}^{\beta}(s_h)] < \gamma$, hence,

$$\mathbb{P}_c[\exists h \in [H-1] : B_4^{h,c} \neq \emptyset] \quad = \quad \mathbb{P}_c[\exists h \in [H-1], s_h \in S_h : s_h \notin \widehat{S}_h^{\gamma,\beta}, \text{ and } c \in \widehat{\mathcal{C}}^{\beta}(s_h)] \quad < \quad \gamma |S|$$

Fix a context-dependent policy π. The following holds for any given context c. (Later we will take the expectation over c).

$$\begin{aligned}
|V^{\pi_c}_{\widetilde{M}(c)}(s_0) - V^{\pi_c}_{\widehat{M}(c)}(s_0)| &= |\sum_{h=0}^{H-1} \sum_{s_h \in S_h \cup \{s_{sink}\}} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c)(r^c(s_h, a_h) - \widehat{r}^c(s_h, a_h))| \\
&= |\sum_{h=0}^{H-1} \sum_{s_h \in S_h} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c)(r^c(s_h, a_h) - \widehat{r}^c(s_h, a_h))| \\
&\qquad \text{(By definition, } r^c(s_{sink}, a) = \widehat{r}^c(s_{sink}, a) = 0, \quad \forall c \in \mathcal{C}, a \in A.\text{)} \\
&\leq \sum_{h=0}^{H-1} \sum_{s_h \in S_h} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c) |r^c(s_h, a_h) - \widehat{r}^c(s_h, a_h)| \\
&= \underbrace{\sum_{h=0}^{H-1} \sum_{s_h \in B_1^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c) |r^c(s_h, a_h) - \widehat{r}^c(s_h, a_h)|}_{(1)}
\end{aligned}$$

$$+\underbrace{\sum_{h=0}^{H-1}\sum_{s_h\in B_2^{h,c}\cup B_3^{h,c}}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,\widehat{P}^c)|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)|}_{(2)}$$

$$+\underbrace{\sum_{h=0}^{H-1}\sum_{s_h\in B_4^{h,c}}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,\widehat{P}^c)|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)|}_{(3)}.$$

We bound (1), (2) and (3) separately.

For (1), under the good events G_1, G_2 , G_3 and G_4, we have for all $h\in[H-1]$ that

$$\mathbb{E}_{\mathcal{D}_h^R}[|f_h^R(c,s_h,a_h)-r^c(s_h,a_h)|-\alpha_1(\mathcal{F}_h^R)]\le\epsilon_R.$$

Since $\mathbb{E}_{\mathcal{D}_h^R}[|f_h^R(c,s_h,a_h)-r^c(s_h,a_h)|]\ge\alpha_1(\mathcal{F}_h^R)$, for all $h\in[H-1]$ and $\xi\in(0,1]$ we obtain using Markov's inequality that

$$\mathbb{P}[|f_h^R(c,s_h,a_h)-r^c(s_h,a_h)|\ge\alpha_1(\mathcal{F}^R_{s_h,a_h})+\xi\ \Big|(c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}]\le\frac{\epsilon_R}{\xi}.$$

Hence,

$$\mathbb{P}[|f_h^R(c,s_h,a_h)-r^c(s_h,a_h)|\ \le\alpha_1(\mathcal{F}^R_{s_h,a_h})+\xi\Big|(c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}]\ge1-\frac{\epsilon_R}{\xi}.$$

Let G_5 denote the following good event.

$$\forall h\in[H-1]\ \forall s_h\in B_1^{h,c}\ \forall a\in A.\ |f_h^R(c,s_h,a_h)-r^c(s_h,a_h)|\le\alpha_1(\mathcal{F}^R_{s_h,a_h})+\xi$$

and denote by $\overline{G_5}$ the complementary event. By the above and an union bound over $(s_h,a_h)\in B_1^{h,c}\times A$ for all $h\in[H-1]$ we have, $\mathbb{P}_c[G_5]\ge1-\frac{\epsilon_R}{\xi}|S||A|$ and $\mathbb{P}_c[\overline{G_5}]\le\frac{\epsilon_R}{\xi}|S||A|$.

If G_5 holds then,

$$\begin{aligned}
(1)&=\sum_{h=0}^{H-1}\sum_{s_h\in B_1^{h,c}}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,\widehat{P}^c)|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)|\\
&=\sum_{h=0}^{H-1}\sum_{s_h\in B_1^{h,c}}\sum_{a_h\in A}\pi_c(a_h|s_h)q_h(s_h|\pi_c,\widehat{P}^c)|f_h^R(c,s_h,a_h)-r^c(s_h,a_h)|\\
&\le\sum_{h=0}^{H-1}\sum_{s_h\in B_1^{h,c}}\sum_{a_h\in A}\pi_c(a_h|s_h)q_h(s_h|\pi_c,\widehat{P}^c)(\alpha_1(\mathcal{F}_h^R)+\xi)\\
&\le\sum_{h=0}^{H-1}\sum_{s_h\in B_1^{h,c}}\sum_{a_h\in A}\pi_c(a_h|s_h)q_h(s_h|\pi_c,\widehat{P}^c)(\alpha_1+\xi)\le\alpha_1H+\xi H.
\end{aligned}$$

Otherwise,

$$(1)=\sum_{h=0}^{H-1}\sum_{s_h\in B_1^{h,c}}\sum_{a_h\in A} q_h(s_h,a_h|\pi_c,\widehat{P}^c)\underbrace{|r^c(s_h,a_h)-\widehat{r}^c(s_h,a_h)|}_{\le1}\le H.$$

Thus,

$$\mathbb{E}_{c\sim\mathcal{D}}[(1)]\le\alpha_1H+\xi H+\frac{\epsilon_R}{\xi}|S||A|H.$$

For (2), consider the following derivation:

$$\begin{aligned}
(2) &= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c) \underbrace{|r^c(s_h, a_h) - \widehat{r}^c(s_h, a_h)|}_{\leq 1} \\
&\leq \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c) \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \sum_{a_h \in A} \pi_c(a_h|s_h) q_h(s_h | \pi_c, \widehat{P}^c) \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} q_h(s_h | \pi_c, \widehat{P}^c) \underbrace{\sum_{a_h \in A} \pi_c(a_h|s_h)}_{=1} \\
&= \sum_{h=0}^{H-1} \sum_{s_h \in B_2^{h,c} \cup B_3^{h,c}} \underbrace{q_h(s_h | \pi_c, \widehat{P}^c)}_{\leq q_h(s_h | \widehat{\pi}^c_{s_h}, \widehat{P}^c) < \beta} \beta |S|.
\end{aligned}$$

Thus,

$$\mathbb{E}_{c \sim \mathcal{D}}[(2)] \leq \beta |S|.$$

For (3), let G_6 denote the good event in which $\forall h \in [H-1], B_4^{h,c} = \emptyset$. Denote by $\overline{G_6}$ the complement event of G_5. We showed that $\mathbb{P}_c[G_6] \geq 1 - \gamma|S|$ thus $\mathbb{P}_c[\overline{G_6}] \leq \gamma|S|$.
If G_6 holds, then $(3) = 0$. Othwewise,

$$\begin{aligned}
(3) &= \sum_{h=0}^{H-1} \sum_{s_h \in B_4^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c) \underbrace{|r^c(s_h, a_h) - \widehat{r}^c(s_h, a_h)|}_{\leq 1} \\
&\leq \sum_{h=0}^{H-1} \underbrace{\sum_{s_h \in B_4^{h,c}} \sum_{a_h \in A} q_h(s_h, a_h | \pi_c, \widehat{P}^c)}_{\leq 1} \leq H.
\end{aligned}$$

Using total expectation we obtain

$$\mathbb{E}_{c \sim \mathcal{D}}[(3)] \leq \gamma |S| H.$$

Overall, by linearity of expectation and the above, we obtain for $\xi = (\epsilon_R |S||A|)^{\frac{1}{2}}$ that

$$\begin{aligned}
\mathbb{E}_{c \sim \mathcal{D}}[|V^{\pi_c}_{\widehat{M}(c)}(s_0) - V^{\pi_c}_{\widehat{M}(c)}(s_0)|] &\leq \mathbb{E}_{c \sim \mathcal{D}}[(1)] + \mathbb{E}_{c \sim \mathcal{D}}[(2)] + \mathbb{E}_{c \sim \mathcal{D}}[(3)] \\
&\leq \alpha_1 H + \xi H + \frac{\epsilon_R}{\xi} |S||A| H + \beta |S| + \gamma |S| H \\
&= \alpha_1 H + 2(\epsilon_R |S||A|)^{\frac{1}{2}} H + \beta |S| + \gamma |S| H,
\end{aligned}$$

as stated. ■

Corollary 123 *Under the good events G_1, G_2 and G_4, for $\gamma = \frac{\epsilon}{20|S|H}$, $\beta = \frac{\epsilon}{20|S|H}$. and $\epsilon_R = \frac{\epsilon^2}{20^2 |S||A| H^2}$, we have for any context-dependent policy $\pi = (\pi_c)_{c \in \mathcal{C}}$ that*

$$\mathbb{E}_{c \sim \mathcal{D}}[|V^{\pi_c}_{\widehat{M}(c)}(s_0) - V^{\pi_c}_{\widehat{M}(c)}(s_0)|] \leq \alpha_1 H + \frac{3\epsilon}{20} + \frac{\epsilon}{20H}$$

Proof Implied by assigning the detailed parameters to the results of Lemma 122. ■

E.5.4 COMBINING VALUE DIFFERENCES CAUSED BY DYNAMICS AND REWARDS APPROXIMATION TO SUB-OPTIMALITY BOUND

Let the following set of selected parameters, called SP1, be

- $\gamma = \frac{\epsilon}{20|S|H} \in (0,1)$.
- $\beta = \frac{\epsilon}{20|S|H} \in (0,1)$.
- $\rho = \frac{\beta}{16|S|H} \in (0, \frac{1}{|S|})$.
- $\epsilon_P = \frac{\epsilon^2}{10 \cdot 16 \cdot 20|A||S|^4 H^3}$.
- $\epsilon_R = \frac{\epsilon^2}{20^2|S||A|H^2}$.

We remark that for our choice of ρ and β it holds that $\rho \in [0, \frac{1}{|S|})$ and $\beta \geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$.

Lemma 124 (expected value difference) *Under the good events G_1, G_2,G_3 and G_4, we have for every policy context-dependent policy $\pi = (\pi_c)_{c\in\mathcal{C}}$ that*

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \alpha_1 H + \frac{1}{2}\epsilon$$

where $\mathcal{M}(c)$ is the true MDP associated with the context c and $\widehat{\mathcal{M}}(c)$ is it's the approximated model, using parameters SP1.

Proof For a fixed $c \in \mathcal{C}$, consider the intermediate MDP $\widetilde{\mathcal{M}}(c) = (S, A, \widehat{P}^c, r^c, H, s_0)$. Using triangle inequality and linearity of expectation we obtain

$$\begin{aligned}\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] &= \mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0) + V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \\ &\leq \mathbb{E}_{c\sim\mathcal{D}}[\underbrace{|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|}_{(1)}] + \mathbb{E}_{c\sim\mathcal{D}}[\underbrace{|V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|}_{(2)}]\end{aligned}$$

By Lemma 120 we have

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0)|] \leq \frac{4\rho|S|}{1-\rho^2|S|^2}H^2 + \beta|S|H + H|S|^2|A|\frac{\epsilon_P}{\rho} + \gamma H|S|.$$

By Lemma 122 we have

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\widetilde{\mathcal{M}}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] \leq \alpha_1 H + 2(\epsilon_R|S||A|)^{\frac{1}{2}}H + \beta|S| + \gamma|S|H.$$

Overall,

$$\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] = \frac{4\rho|S|}{1-\rho^2|S|^2}H^2 + |S|^2|A|H\frac{\epsilon_P}{\rho} + 2\gamma|S|H + 2\beta|S|H + \alpha_1 H + 2(\epsilon_R|S||A|)^{\frac{1}{2}}H$$

For the parameters set SP1 we have that $\beta < \frac{1}{2|S|}$, which implies that $0 < \rho < \frac{1}{|S|}$. We also have that

$$2H\frac{4\rho|S|}{1-\rho^2|S|^2} = \frac{8H|S|\frac{\beta}{16|S|H}}{1-\frac{\beta^2|S|^2}{2^8|S|^2H^2}} = \frac{\frac{\beta}{2}}{1-\underbrace{\frac{\beta^2}{2^8H^2}}_{\leq 1/2}} \leq 2\frac{\beta}{2} = \beta.$$

Hence, the constrains on ρ and β are both hold.

Finally,

$$\begin{aligned}
\mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c}_{\mathcal{M}(c)}(s_0) - V^{\pi_c}_{\widehat{\mathcal{M}}(c)}(s_0)|] &= \frac{4\rho|S|}{1-\rho^2|S|^2}H^2 + |S|^2|A|H\frac{\epsilon_P}{\rho} + 2\gamma|S|H + 2\beta|S|H + \alpha_1 H + 2(\epsilon_R|S||A|)^{\frac{1}{2}}H \\
&\leq \frac{\frac{1}{4}\beta H}{1-\underbrace{\frac{\beta^2}{2^8H^2}}_{\leq 1/2}} + 16|S|^3|A|H^2\frac{\epsilon_P}{\beta} + 2\frac{\epsilon}{10} + \alpha_1 H + \frac{\epsilon}{10} \\
&\leq \frac{1}{2}\beta H + 16|S|^3|A|H^2\frac{\epsilon_P}{\beta} + 2\frac{\epsilon}{10} + \alpha_1 H + \frac{\epsilon}{10} \\
&\leq \frac{1}{2}\beta H + 16\cdot 20|S|^4|A|H^3\frac{\epsilon_P}{\epsilon} + 3\frac{\epsilon}{10} + \alpha_1 H \\
&= \frac{1}{2}\frac{\epsilon}{20|S|} + 4\frac{\epsilon}{10} + \alpha_1 H \\
&\leq \frac{1}{2}\epsilon + \alpha_1 H,
\end{aligned}$$

as stated. ■

The following corollary shows that for our choice of parameters, all good events holds with high probability.

Corollary 125 *Using parameters SP1 we have* $\mathbb{P}[G_1, G_2, G_3, G_4] \geq 1 - (\frac{\delta}{2} + \frac{\epsilon}{10})$.

Proof By Corollary 115 we have that $\mathbb{P}[G_1, G_2, G_3, G_4] \geq 1 - (\frac{\delta}{8} + 3\delta_1 H + \frac{\epsilon_P}{\rho}|S|^2|A|H)$. Hence by ρ, β, ϵ_P and δ_1 choice we obtain

$$\begin{aligned}
\mathbb{P}[G_1, G_2, G_3, G_4] &\geq 1 - (\frac{\delta}{8} + 3\delta_1 H + \frac{\epsilon_P}{\rho}|S|^2|A|H) \\
&= 1 - \frac{\delta}{2} - 16|S|^3|A|H^2\frac{\epsilon_P}{\beta} \\
&= 1 - \frac{\delta}{2} - 16\cdot 20|S|^4|A|H^3\frac{\epsilon_P}{\epsilon} \\
&= 1 - \frac{\delta}{2} - \frac{\epsilon}{10},
\end{aligned}$$

as stated. ■

Finally, the following theorem bound the expected sub-optimality of our approximated optimal policy $\widehat{\pi}^\star$.

Theorem 126 (expected suboptimality bound) *With probability at least* $1 - (\delta + \frac{\epsilon}{5})$ *it holds that*

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi^\star_c}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}^\star_c}_{\mathcal{M}(c)}(s_0)] \leq \epsilon + 2\alpha_1 H,$$

where $\pi^\star = (\pi^\star_c)_{c\in\mathcal{C}}$ *is the optimal policy for* $\mathcal{M}$ *and* $\widehat{\pi}^\star = (\widehat{\pi}^\star_c)_{c\in\mathcal{C}}$ *is the optimal policy for* $\widehat{\mathcal{M}}$.

Proof Assume the good events G_1, G_2, G_3 and G_4 hold.

By Lemma 124, we have for $\pi^\star$

$$\left|\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)]\right| \le \mathbb{E}_{c\sim\mathcal{D}}[|V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)|] \le \frac{1}{2}\epsilon + \alpha_1 H,$$

yielding,

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)] \le \frac{1}{2}\epsilon + \alpha_1 H.$$

Similarly, we obtain for $\widehat{\pi}^\star$,

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \le \frac{1}{2}\epsilon + \alpha_1 H.$$

Since for all $c \in \mathcal{C}$, $\widehat{\pi}_c^\star$ is the optimal policy for $\widehat{\mathcal{M}}(c)$ we have $V^{\widehat{\pi}_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0) \ge V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)$ which implies that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)] - \mathbb{E}_{c\sim\mathcal{D}}[V^{\widehat{\pi}_c^\star}_{\widehat{\mathcal{M}}(c)}(s_0)] \le 0.$$

Since by Corollary 125 we have that $\mathbb{P}[G_1, G_2, G_3, G_4] \ge 1 - (\frac{\delta}{2} + \frac{\epsilon}{10})$, the theorem implied by summing the above three inequalities. ■

E.5.5 ADDITIONAL LEMMAS FOR BOUNDING THE SAMPLE COMPLEXITY FOR THE ℓ_1 LOSS

Lemma 127 *Let $\rho \in [0, \frac{1}{|S|})$ and $h \in [H-1]$. Assume the good events $G_1, G_2^k, G_3^k,\ \forall k \in [h]$ hold, then it holds that*

$$\mathbb{P}\left[\|\widehat{P}^c(\cdot|s_h,a_h) - P^c(\cdot|s_h,a_h)\|_1 \le \frac{4\rho|S|}{1-\rho^2|S|^2}\middle|(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}\right] \ge 1 - \frac{\epsilon_P}{\rho}|S_{h+1}|,$$

where $\widehat{P}^c$ is the dynamics defined in Algorithm 12 and

$$\|\widehat{P}^c(\cdot|s_h,a_h) - P^c(\cdot|s_h,a_h)\|_1 := \sum_{s_{h+1}\in S_{h+1}} |\widehat{P}^c(s_{h+1}|s_h,a_h) - P^c(s_{h+1}|s_h,a_h)|$$

(i.e., the entry of s_{sink} in $\widehat{P}^c$ is ignored).

Proof

We prove similarly to shown for Lemma 116, when using the good events G_3^k for all $k \in [h]$ guarantees for the distribution $\widetilde{D}_h^{\gamma,\beta}$ over $\widetilde{\mathcal{X}}_h^{\gamma,\beta} \times S_{h+1}$.

Recall that for $(c, s_h, a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$ we have that $\widehat{P}^c(s_{sink}|s_h,a_h) = 0$ by $\widehat{P}^c$ definition. In addition, the true dynamics P^c is not defined for s_{sink} since $s_{sink} \notin S$. A natural extension of P^c to s_{sink} is by defining that $\forall(s,a) \in S \times A$. $P^c(s_{sink}|s,a) := 0$. By that extension, we have for all $(c, s_h, a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$ that $P^c(s_{sink}|s_h,a_h) = \widehat{P}^c(s_{sink}|s_h,a_h) = 0$. Hence, we can simply ignore s_{sink} in the following analysis.

Under the good event G_3^h, by Markov's inequality we have

$$\begin{aligned}
&\mathbb{P}_{(c,s_h,a_h,s_{h+1})}\left[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)| \ge \rho\middle|(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}\right] = \\
&= \mathbb{P}_{\mathcal{D}_h^P}[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)| \ge \rho] \\
&\le \frac{\mathbb{E}_{\mathcal{D}_h^P}[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)|]}{\rho} && \text{(By Markov's inequality)} \\
&\le \frac{\epsilon_P}{\rho}. && \text{(Under } G_h^3\text{)}
\end{aligned}$$

Hence,

$$\mathbb{P}_{(c,s_h,a_h,s_{h+1})}\left[|f_h^P(c,s_h,a_h,s_{h+1}) - P^c(s_{h+1}|s_h,a_h)| \le \rho \Big| (c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}\right] \ge 1 - \frac{\epsilon_P}{\rho}.$$

Since $P^c(\cdot|s_h,a_h)$ is a distribution over $s_{h+1} \in S_{h+1}$, we have that $\sum_{s_{h+1}\in S_{h+1}} P^c(s_{h+1}|s_h,a_h) = 1$. Thus, by union bound applied on $s_{h+1} \in S_{h+1}$, we obtain

$$\mathbb{P}_{(c,s_h,a_h)}\left[1-\rho|S| \le \sum_{s_{h+1}\in S_{h+1}} f_h^P(c,s_h,a_h,s_{h+1}) \le 1+\rho|S| \Big| (c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}\right] \ge 1 - \frac{\epsilon_P}{\rho}|S_{h+1}|.$$

Hence, we further conclude that

$$\mathbb{P}_{(c,s_h,a_h)}\left[\forall s_{h+1}\in S_{h+1}. \frac{P^c(s_{h+1}|s_h,a_h)-\rho}{1+\rho|S|} \le \underbrace{\frac{f_h^P(c,s_h,a_h,s_{h+1})}{\sum_{s'\in S_{h+1}} f_h^P(c,s_h,a_h,s')}}_{=\widehat{P}^c(s_{h+1}|s_h,a_h)} \le \frac{P^c(s_{h+1}|s_h,a_h)+\rho}{1-\rho|S|} \Big| (c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}\right]$$
$$\ge 1 - \frac{\epsilon_P}{\rho}|S_{h+1}|. \tag{20}$$

Fix a tuple $(c,s_h,a_h) \in \widetilde{\mathcal{X}}_h^{\gamma,\beta}$ and assume the event of inequality (20) holds.
Denote $S_{h+1}^+ = \{s_{h+1} \in S_{h+1} : \widehat{P}^c(s_{h+1}|s_h,a_h) \ge P^c(s_{h+1}|s_h,a_h)\}$ and consider the following derivation.

$$\begin{aligned}
\|\widehat{P}^c(\cdot|s_h,a_h) - P^c(\cdot|s_h,a_h)\|_1 &= \sum_{s_{h+1}\in S_{h+1}} |\widehat{P}^c(s_{h+1}|s_h,a_h) - P^c(s_{h+1}|s_h,a_h)| \\
&= \sum_{s_{h+1}\in S_{h+1}^+} (\widehat{P}^c(s_{h+1}|s_h,a_h) - P^c(s_{h+1}|s_h,a_h)) \\
&\quad + \sum_{s_{h+1}\in S_{h+1}\setminus S_{h+1}^+} (P^c(s_{h+1}|s_h,a_h) - \widehat{P}^c(s_{h+1}|s_h,a_h)) \\
&\le \sum_{s_{h+1}\in S_{h+1}^+} \left(\frac{P^c(s_{h+1}|s_h,a_h)+\rho}{1-\rho|S|} - P^c(s_{h+1}|s_h,a_h)\right) \\
&\quad + \sum_{s_{h+1}\in S_{h+1}\setminus S_{h+1}^+} \left(P^c(s_{h+1}|s_h,a_h) - \frac{P^c(s_{h+1}|s_h,a_h)-\rho}{1+\rho|S|}\right) \\
&= \sum_{s_{h+1}\in S_{h+1}^+} \frac{P^c(s_{h+1}|s_h,a_h)+\rho-(1-\rho|S|)P^c(s_{h+1}|s_h,a_h)}{1-\rho|S|} \\
&\quad + \sum_{s_{h+1}\in S_{h+1}\setminus S_{h+1}^+} \frac{-P^c(s_{h+1}|s_h,a_h)+\rho+(1+\rho|S|)P^c(s_{h+1}|s_h,a_h)}{1+\rho|S|} \\
&= \frac{1}{1-\rho|S|} \sum_{s_{h+1}\in S_{h+1}^+} (\rho+\rho|S|P^c(s_{h+1}|s_h,a_h)) \\
&\quad + \frac{1}{1+\rho|S|} \sum_{s_{h+1}\in S_{h+1}\setminus S_{h+1}^+} (\rho+\rho|S|P^c(s_{h+1}|s_h,a_h)) \\
&\le \frac{2\rho|S|}{1-\rho|S|} + \frac{2\rho|S|}{1+\rho|S|}
\end{aligned}$$

$$=\frac{4\rho|S|}{1-\rho^2|S|^2}.$$

By inequality (20), the above holds with probability at least $1-\frac{\epsilon_P}{\rho}|S_{h+1}|$ over $(c,s_h,a_h)\in\mathcal{X}_h^{\gamma,\beta}$. Hence the lemma follows.

■

Lemma 128 *Fix $\beta\in(0,1]$ and $\rho\in[0,\frac{1}{|S|})$ such that $\beta\geq 2H\frac{4\rho|S|}{1-\rho^2|S|^2}$.*

Then, for every (context-dependent) policy $\pi=(\pi_c)_{c\in\mathcal{C}}$, and a layer $h\in[H-1]$, under the good events $G_1,G_2^i,G_3^i,\forall i\in[h-1]$ the following holds.

$$\mathbb{P}_c\left[\forall k\in[h],s_k\in S_k.\ \ q_k(s_k|\pi_c,P^c)\geq q_k(s_k|\pi_c,\widehat{P}^c)-\frac{4\rho|S|}{1-\rho^2|S|^2}k\right]\geq 1-|A|\sum_{k=0}^{h-1}\frac{\epsilon_P}{\rho}|S_k||S_{k+1}|$$

Proof For every context $c\in\mathcal{C}$ and define the dynamics $\widetilde{P}^c$ over $S\cup\{s_{sink}\}\times A$:

$$\forall(s,a)\in S\cup\{s_{sink}\}\times A:\widetilde{P}^c(s|s_{sink},a)=\begin{cases}1 & ,\text{if } s=s_{sink}\\ 0 & ,\text{otherwise}\end{cases}$$

In addition we define

$$\begin{aligned}
&\forall k\in[h-1],\ \ \forall(s_k,a_k,s_{k+1})\in\widetilde{S}_k^{\gamma,\beta}\times A\times S_{k+1}:\\
&\widetilde{P}^c(s_{k+1}|s_k,a_k)=\begin{cases}P^c(s_{k+1}|s_k,a_k) & ,\text{if } c\in\widehat{C}^\beta(s_k)\\ 0 & ,\text{otherwise}\end{cases}\\
&\widetilde{P}^c(s_{sink}|s_k,a_k)=\begin{cases}0 & ,\text{if } c\in\widehat{\mathcal{C}}^\beta(s_k)\\ 1 & ,\text{otherwise}\end{cases}\\
&\forall k\in[h-1],\forall(s_k,a_k,s_{k+1})\in(S_k\setminus\widetilde{S}_k^{\gamma,\beta})\times A\times S_{k+1}:\\
&\widetilde{P}^c(s_{k+1}|s_k,a_k)=0,\quad\widetilde{P}^c(s_{sink}|s_k,a_k)=1.
\end{aligned}$$

Clearly, by definition of $\widetilde{P}^c$, we have for every (context-dependent) policy π that

$$\mathbb{P}_c\left[\forall k\in[h],s_k\in S_k.\ \ q_k(s_k|\pi_c,P^c)\geq q_k(s_k|\pi_c,\widetilde{P}^c)\right]=1. \tag{21}$$

By Lemma 127 under the good events $G_1,G_2^k,G_3^k\ \ \forall k\in[h-1]$ we have for any $k\in[h-1]$ that

$$\mathop{\mathbb{P}}_{(c,s_k,a_k)}\left[\|\widetilde{P}^c(\cdot|s_k,a_k)-\widehat{P}^c(\cdot|s_k,a_k)\|_1\leq\frac{4\rho|S|}{1-\rho^2|S|^2}\middle|(c,s_k,a_k)\in\widetilde{\mathcal{X}}_k^{\gamma,\beta}\right]\geq 1-\frac{\epsilon_P}{\rho}|S_{k+1}|. \tag{22}$$

We now show that

$$\mathop{\mathbb{P}}_{(c,s_k,a_k)}\left[\|\widehat{P}^c(\cdot|s_k,a_k)-\widetilde{P}^c(\cdot|s_k,a_k)\|_1=0\middle|(c,s_k,a_k)\notin\widetilde{\mathcal{X}}_k^{\gamma,\beta}\right]=1. \tag{23}$$

For every layer $k\in[h-1]$ we have by definition that $(c,s_k,a_k)\in\widetilde{\mathcal{X}}_k^{\gamma,\beta}$ if and only if $s_k\in\widetilde{S}_k^{\gamma,\beta}$ and $c\in\widehat{\mathcal{C}}^\beta(s_k)$.

By the definition of $\widetilde{P}^c$ and $\widehat{P}^c$ we have for every layer $k\in[h-1]$ and context $c\in\mathcal{C}$ that

$$\forall(s_k,a_k)\in(S_k\setminus\widetilde{S}_k^{\gamma,\beta})\times A.\ \ \|\widehat{P}^c(\cdot|s_k,a_k)-\widetilde{P}^c(\cdot|s_k,a_k)\|_1=0$$

In addition, by definition of $\widehat{P}^c$ and $\widetilde{P}^c$, for every layer $k \in [h-1]$ if $(s_k, a_k) \in \widetilde{S}_k^{\gamma,\beta} \times A$ but $c \notin \widehat{\mathcal{C}}^{\beta}(s_k)$, then

$$\|\widehat{P}^c(\cdot|s_k, a_k) - \widetilde{P}^c(\cdot|s_k, a_k)\|_1 = 0.$$

Thus, equation (23) follows.

Using total probability low, equations (22) and (23) yield that

$$\mathop{\mathbb{P}}_{(c,s_k,a_k)}\left[\|\widehat{P}^c(\cdot|s_k, a_k) - \widetilde{P}^c(\cdot|s_k, a_k)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}\right] \geq 1 - \frac{\epsilon_P}{\rho^2}|S_{k+1}|,$$

which by union bound over $(s_k, a_k) \in S_k \times A$ for every layer $k \in [h-1]$ implies that

$$\mathbb{P}_c\left[\forall k \in [h-1], (s_k, a_k) \in S_k \times A.\ \|\widehat{P}^c(\cdot|s_k, a_k) - \widetilde{P}^c(\cdot|s_k, a_k)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}\right] \geq 1 - |A| \sum_{k=0}^{h-1} \frac{\epsilon_P}{\rho}|S_k||S_{k+1}|. \tag{24}$$

By Theorem 137 the above yields that

$$\mathbb{P}_c\left[\forall k \in [h].\ \|q_k(\cdot|\pi_c, \widetilde{P}^c) - q_k(\cdot|\pi_c, \widehat{P}^c)\|_1 \leq \frac{4\rho|S|}{1-\rho^2|S|^2}k\right] \geq 1 - |A| \sum_{k=0}^{h-1} \frac{\epsilon_P}{\rho}|S_k||S_{k+1}|,$$

which in particularly implies that

$$\mathbb{P}_c\left[\forall k \in [h],\ s_k \in s_k.\ q_k(s_k|\pi_c, \widetilde{P}^c) \geq q_k(s_k|\pi_c, \widehat{P}^c) - \frac{4\rho|S|}{1-\rho^2|S|^2}\right] \geq 1 - |A| \sum_{k=0}^{h-1} \frac{\epsilon_P}{\rho}|S_k||S_{k+1}|. \tag{25}$$

Finally, the lemma follows by combining inequalities (21) and (25).

■

E.6 Sample Complexity Bounds

We show sample complexity bounds for both ℓ_1 and ℓ_2 loss functions. Recall Theorems 28 and 29,

Theorem 129 (Adaption of Theorem 19.2 in Anthony et al. (1999)) *Let $\mathcal{F}$ be a hypothesis space of real valued functions with a finite pseudo dimension, denoted $Pdim(\mathcal{F}) < \infty$. Then, $\mathcal{F}$ has a uniform convergence with*

$$m(\epsilon, \delta) = O\Big(\frac{1}{\epsilon^2}(Pdim(\mathcal{F}) \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})\Big).$$

Theorem 130 (Adaption of Theorem 19.1 in Anthony et al. (1999)) *Let $\mathcal{F}$ be a hypothesis space of real valued functions with a finite fat-shattering dimension, denoted $fat_{\mathcal{F}}$. Then, $\mathcal{F}$ has a uniform convergence with*

$$m(\epsilon, \delta) = O\Big(\frac{1}{\epsilon^2}(fat_{\mathcal{F}}(\epsilon/256) \ln^2 \frac{1}{\epsilon} + \ln \frac{1}{\delta})\Big).$$

Remark 131 *In the following analysis we omit the sample complexity needed to approximate the faction of good contexts for every $s \in S$ as it is*

$$O\Big(\frac{|S|^3 H^2 \ln \frac{|S|}{\delta}}{\epsilon^2}\Big)$$

which is negligible additional term in the following analysis.

E.6.1 SAMPLE COMPLEXITY BOUNDS FOR THE ℓ_2 LOSS.

We show sample complexity for function classes with finite Pseudo dimension with ℓ_2 loss.

Corollary 132 *Assume that for every $h \in [H-1]$ we have that $Pdim(\mathcal{F}_h^R), Pdim(\mathcal{F}_h^P) < \infty$. Let $Pdim = \max_{h\in[H-1]} \max\{Pdim(\mathcal{F}_h^R), Pdim(\mathcal{F}_h^P)\}$. Then, after collecting*

$$O\Big(\frac{|A|^2|S|^{15}H^{13}}{\epsilon^8}\Big(Pdim \ln \frac{|A||S|^6H^5}{\epsilon^3} + \ln\frac{H}{\delta}\Big)\Big).$$

trajectories, with probability at least $1-(\delta+\frac{\epsilon}{5})$ it holds that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

Proof Recall that for every layer $h \in [H-1]$ our algorithm run for

$$T_h = \left\lceil \frac{8|S|}{\gamma\cdot\beta}\left(\ln\frac{1}{\delta_1} + 2\max\{N_P(\mathcal{F}_h^P,\epsilon_P,\delta_1/2)), N_R(\mathcal{F}_h^R,\epsilon_R,\delta_1/2))\}\right)\right\rceil.$$

episodes. Recall our choice of parameters is $\gamma = \frac{\epsilon}{20|S|H}$, $\beta = \frac{\epsilon}{20|S|H}$, $\rho = \frac{\beta}{16|S|H}$, $\epsilon_P = \frac{\epsilon^3}{10\cdot 2^8\cdot 20^2|A||S|^6H^5}$, $\epsilon_R = \frac{\epsilon^3}{20^3|S||A|H^3}$, $\delta_1 = \frac{\delta}{8H}$.

By Theorem 106, for this choice of parameters and $\sum_{h=0}^{H-1} T_h$ examples we have with probability at least $1-(\delta+\frac{\epsilon}{5})$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

Since for every $h\in[H-1]$ we have that $Pdim(\mathcal{F}_h^R), Pdim(\mathcal{F}_h^P) < \infty$ and $Pdim = \max_{h\in[H-1]}\max\{Pdim(\mathcal{F}_h^R), Pdim(\mathcal{F}_h^P)\}$, by Theorem 28, for every $h\in[H-1]$ we have

$$N_P(\mathcal{F}_h^P,\epsilon_P,\delta_1) = O\Big(\frac{Pdim\ln\frac{1}{\epsilon_P}+\ln\frac{1}{\delta_1}}{\epsilon_P^2}\Big) = O\Big(\frac{|A|^2|S|^{12}H^{10}}{\epsilon^6}\Big(Pdim\ln\frac{|A||S|^6H^5}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big),$$

and

$$N_R(\mathcal{F}_h^R,\epsilon_R,\delta_1) = O\Big(\frac{Pdim\ln\frac{1}{\epsilon_R}+\ln\frac{1}{\delta_1}}{\epsilon_R^2}\Big) = O\Big(\frac{|S|^2|A|^2H^6}{\epsilon^6}\Big(Pdim\ln\frac{|S||A|H^3}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big).$$

Hence for every $h\in[H-1]$ we have that

$$\max\{N_P(\mathcal{F}_h^P,\epsilon_P,\delta_1), N_R(\mathcal{F}_h^R,\epsilon_R,\delta_1)\} = O\Big(\frac{|A|^2|S|^{12}H^{10}}{\epsilon^6}\Big(Pdim\ln\frac{|A||S|^6H^5}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big).$$

By our choice of β and γ it holds that

$$T_h = O\Big(|S|\frac{|S|^2H^2}{\epsilon^2}\frac{|A|^2|S|^{12}H^{10}}{\epsilon^6}\Big(Pdim\ln\frac{|A||S|^6H^5}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big) = O\Big(\frac{|A|^2|S|^{15}H^{12}}{\epsilon^8}\Big(Pdim\ln\frac{|A||S|^6H^5}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big).$$

By summing the above for every layer $h\in[H-1]$ we obtain that the sample complexity is

$$O\Big(\frac{|A|^2|S|^{15}H^{13}}{\epsilon^8}\Big(Pdim\ln\frac{|A||S|^6H^5}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big).$$

■

We show sample complexity for function classes with finite fat-shattering dimension with ℓ_2 loss:

Corollary 133 *Assume that for every $h \in [H-1]$ we have that $\mathcal{F}_h^R$ and $\mathcal{F}_h^R$ has finite fat-shattering dimension. Let $Fdim = \max_{h\in[H-1]} \max\{fat_{\mathcal{F}_h^R}(\epsilon_R), fat_{\mathcal{F}_h^P}(\epsilon_R)\}$. Then, after collecting*

$$O\Big(\frac{|A|^2|S|^{15}H^{13}}{\epsilon^8}\Big(Fdim\ln^2\frac{|A||S|^6H^5}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big).$$

trajectories, with probability at least $1-(\delta+\frac{\epsilon}{5})$ it holds that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

Proof Recall that for every layer $h \in [H-1]$ our algorithm run for

$$T_h = \left\lceil \frac{8|S|}{\gamma\cdot\beta}\left(\ln\frac{1}{\delta_1} + 2\max\{N_P(\mathcal{F}_h^P,\epsilon_P,\delta_1/2), N_R(\mathcal{F}_h^R,\epsilon_R,\delta_1/2)\}\right)\right\rceil$$

episodes. Recall our choice of parameters is $\gamma = \frac{\epsilon}{20|S|H}$, $\beta = \frac{\epsilon}{20|S|H}$, $\rho = \frac{\beta}{16|S|H}$, $\epsilon_P = \frac{\epsilon^3}{10\cdot 2^8 20^2|A||S|^6H^5}$, $\epsilon_R = \frac{\epsilon^3}{20^3|S||A|H^3}$, $\delta_1 = \frac{\delta}{8H}$.

By Theorem 106, for this choice of parameters and $\sum_{h=0}^{H-1} T_h$ samples we have with probability at least $1-(\delta+\frac{\epsilon}{5})$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \le \epsilon + 2\alpha_2 H.$$

Since every $h \in [H-1]$ we have that $\mathcal{F}_h^R$ and
$\mathcal{F}_h^R$ has finite fat-shattering dimension, and $Fdim = \max_{h\in[H-1]} \max\{fat_{\mathcal{F}_h^R}(\epsilon_R), fat_{\mathcal{F}_h^P}(\epsilon_R)\}$, by Theorem 29, for every $h \in [H-1]$ we have

$$N_P(\mathcal{F}_h^P,\epsilon_P,\delta_1) = O\Big(\frac{Pdim\ln\frac{1}{\epsilon_P}+\ln\frac{1}{\delta_1}}{\epsilon_P^2}\Big) = O\Big(\frac{|A|^2|S|^{12}H^{10}}{\epsilon^6}\Big(Fdim\ln^2\frac{|A||S|^6H^5}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big),$$

and

$$N_R(\mathcal{F}_h^R,\epsilon_R,\delta_1) = O\Big(\frac{Pdim\ln\frac{1}{\epsilon_R}+\ln\frac{1}{\delta_1}}{\epsilon_R^2}\Big) = O\Big(\frac{|S|^2|A|^2H^6}{\epsilon^6}\Big(Fdim\ln^2\frac{|S||A|H^3}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big).$$

Hence for every $h \in [H-1]$ we have that

$$\max\{N_P(\mathcal{F}_h^P,\epsilon_P,\delta_1), N_R(\mathcal{F}_h^R,\epsilon_R,\delta_1)\} = O\Big(\frac{|A|^2|S|^{12}H^{10}}{\epsilon^6}\Big(Fdim\ln^2\frac{|A||S|^6H^5}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big).$$

By our choice of β and γ it holds that

$$T_h = O\Big(|S|\frac{|S|^2H^2}{\epsilon^2}\frac{|A|^2|S|^{12}H^{10}}{\epsilon^6}\Big(Fdim\ln^2\frac{|A||S|^6H^5}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big) = O\Big(\frac{|A|^2|S|^{15}H^{12}}{\epsilon^8}\Big(Fdim\ln^2\frac{|A||S|^6H^5}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big).$$

By summing the above for every layer $h \in [H-1]$ we obtain that the sample complexity is

$$O\Big(\frac{|A|^2|S|^{15}H^{13}}{\epsilon^8}\Big(Fdim\ln^2\frac{|A||S|^6H^5}{\epsilon^3}+\ln\frac{H}{\delta}\Big)\Big).$$

■

E.6.2 SAMPLE COMPLEXITY BOUNDS FOR THE ℓ_1 LOSS

We show sample complexity for function classes with finite Pseudo dimension with ℓ_1 loss using the set of parameters SP1.

Corollary 134 *Assume that for every $h \in [H-1]$ we have that $Pdim(\mathcal{F}_h^R), Pdim(\mathcal{F}_h^P) < \infty$. Let $Pdim = \max_{h\in[H-1]} \max\{Pdim(\mathcal{F}_h^R), Pdim(\mathcal{F}_h^P)\}$. Then, after collecting*

$$O\Big(\frac{|A|^2|S|^{11}H^9}{\epsilon^6}\Big(Pdim \ln \frac{|A||S|^4H^3}{\epsilon^2} + \ln \frac{H}{\delta}\Big)\Big)$$

trajectories, with probability at least $1-(\delta+\frac{\epsilon}{5})$ it holds that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \leq 2\alpha_1 H + \epsilon.$$

Proof Recall that for every layer $h \in [H-1]$ our algorithm run for

$$T_h = \left\lceil \frac{8|S|}{\gamma\cdot\beta}\left(\ln\frac{1}{\delta_1} + 2\max\{N_P(\mathcal{F}_h^P,\epsilon_P,\delta_1/2), N_R(\mathcal{F}_h^R,\epsilon_R,\delta_1/2)\}\right)\right\rceil$$

episodes. Recall SP1 parameters are $\gamma = \frac{\epsilon}{20|S|H}$, $\beta = \frac{\epsilon}{20|S|H}$, $\rho = \frac{\beta}{16|S|H}$, $\epsilon_P = \frac{\epsilon^2}{10\cdot 16\cdot 20|A||S|^4H^3}$, $\epsilon_R = \frac{\epsilon^2}{20^2|S||A|H^2}$ $\delta_1 = \frac{\delta}{8H}$.

By Theorem 126, for this choice of parameters and $\sum_{h=0}^{H-1} T_h$ examples we have with probability at least $1-(\delta+\frac{\epsilon}{5})$ that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \leq 2\alpha_1 H + \epsilon.$$

Since every $h\in[H-1]$ we have that $Pdim(\mathcal{F}_h^R), Pdim(\mathcal{F}_h^P) < \infty$. and $Pdim = \max_{h\in[H-1]}\max\{Pdim(\mathcal{F}_h^R), Pdim(\mathcal{F}_h^P)\}$, by Theorem 28, for every $h\in[H-1]$ we have

$$N_P(\mathcal{F}_h^P,\epsilon_P,\delta_1) = O\Big(\frac{Pdim\ln\frac{1}{\epsilon_P} + \ln\frac{1}{\delta_1}}{\epsilon_P^2}\Big) = O\Big(\frac{|A|^2|S|^8H^6}{\epsilon^4}\Big(Pdim\ln\frac{|A||S|^4H^3}{\epsilon^2} + \ln\frac{H}{\delta}\Big)\Big),$$

and

$$N_R(\mathcal{F}_h^R,\epsilon_R,\delta_1) = O\Big(\frac{Pdim\ln\frac{1}{\epsilon_R} + \ln\frac{1}{\delta_1}}{\epsilon_R^2}\Big) = O\Big(\frac{|S|^2|A|^2H^4}{\epsilon^4}\Big(Pdim\ln\frac{|S||A|H^2}{\epsilon^2} + \ln\frac{H}{\delta}\Big)\Big).$$

Hence for every $h\in[H-1]$ we have that

$$\max\{N_P(\mathcal{F}_h^P,\epsilon_P,\delta_1), N_R(\mathcal{F}_h^R,\epsilon_R,\delta_1)\} = O\Big(\frac{|A|^2|S|^8H^6}{\epsilon^4}\Big(Pdim\ln\frac{|A||S|^4H^3}{\epsilon^2} + \ln\frac{H}{\delta}\Big)\Big).$$

By our choice of β and γ it holds that

$$T_h = O\Big(|S|\frac{|S|^2H^2}{\epsilon^2}\frac{|A|^2|S|^8H^6}{\epsilon^4}\Big(Pdim\ln\frac{|A||S|^4H^3}{\epsilon^2} + \ln\frac{H}{\delta}\Big)\Big) = O\Big(\frac{|A|^2|S|^{11}H^8}{\epsilon^6}\Big(Pdim\ln\frac{|A||S|^4H^3}{\epsilon^2} + \ln\frac{H}{\delta}\Big)\Big).$$

By summing the above for every layer $h\in[H-1]$ we obtain that the sample complexity is

$$O\Big(\frac{|A|^2|S|^{11}H^9}{\epsilon^6}\Big(Pdim\ln\frac{|A||S|^4H^3}{\epsilon^2} + \ln\frac{H}{\delta}\Big)\Big).$$

■

We show sample complexity for function classes with finite fat-shattering dimension with ℓ_1 loss:

Corollary 135 *Assume that for every $h\in[H-1]$ we have that $\mathcal{F}_h^R$ and $\mathcal{F}_h^R$ has finite fat-shattering dimension. Let $Fdim = \max_{h\in[H-1]}\max\{fat_{\mathcal{F}_h^R}(\epsilon_R), fat_{\mathcal{F}_h^P}(\epsilon_R)\}$. Then, after collecting*

$$O\Big(\frac{|A|^2|S|^{11}H^9}{\epsilon^6}\Big(Fdim\ln^2\frac{|A||S|^4H^3}{\epsilon^2} + \ln\frac{H}{\delta}\Big)\Big).$$

examples, with probability at least $1-(\delta+\frac{\epsilon}{5})$ it holds that

$$\mathbb{E}_{c\sim\mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \leq 2\alpha_1 H + \epsilon.$$

Proof Recall that for every layer $h \in [H-1]$ our algorithm run for

$$T_h = \left\lceil \frac{8|S|}{\gamma \cdot \beta} \left(\ln \frac{1}{\delta_1} + 2 \max\{N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1/2), N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1/2)\} \right) \right\rceil$$

episodes. Recall our choice of parameters is $\gamma = \frac{\epsilon}{20|S|H}$, $\beta = \frac{\epsilon}{20|S|H}$, $\rho = \frac{\beta}{16|S|H}$, $\epsilon_P = \frac{\epsilon^2}{10 \cdot 16 \cdot 20|A||S|^4H^3}$, $\epsilon_R = \frac{\epsilon^2}{20^2|S||A|H^2}$, $\delta_1 = \frac{\delta}{8H}$.

By Theorem 126, for this choice of parameters and $\sum_{h=0}^{H-1} T_h$ samples we have with probability at least $1 - (\delta + \frac{\epsilon}{5})$ that

$$\mathbb{E}_{c \sim \mathcal{D}}[V^{\pi_c^\star}_{\mathcal{M}(c)}(s_0) - V^{\widehat{\pi}_c^\star}_{\mathcal{M}(c)}(s_0)] \leq 2\alpha_1 H + \epsilon.$$

Since every $h \in [H-1]$ we have that $\mathcal{F}_h^R$ and $\mathcal{F}_h^R$ has finite fat-shattering dimension, and $Fdim = \max_{h \in [H-1]} \max\{fat_{\mathcal{F}_h^R}(\epsilon_R), fat_{\mathcal{F}_h^P}(\epsilon_R)\}$, by Theorem 29, for every $h \in [H-1]$ we have

$$N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1) = O\Big(\frac{Fdim \ln^2 \frac{1}{\epsilon_P} + \ln \frac{1}{\delta_1}}{\epsilon_P^2}\Big) = O\Big(\frac{|A|^2|S|^8H^6}{\epsilon^4}\Big(Fdim \ln^2 \frac{|A||S|^4H^3}{\epsilon^2} + \ln \frac{H}{\delta}\Big)\Big),$$

and

$$N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1) = O\Big(\frac{Fdim \ln^2 \frac{1}{\epsilon_R} + \ln \frac{1}{\delta_1}}{\epsilon_R^2}\Big) = O\Big(\frac{|S|^2|A|^2H^4}{\epsilon^4}\Big(Fdim \ln^2 \frac{|S||A|H^2}{\epsilon^2} + \ln \frac{H}{\delta}\Big)\Big).$$

Hence for every $h \in [H-1]$ we have that

$$\max\{N_P(\mathcal{F}_h^P, \epsilon_P, \delta_1), N_R(\mathcal{F}_h^R, \epsilon_R, \delta_1)\} = O\Big(\frac{|A|^2|S|^8H^6}{\epsilon^4}\Big(Fdim \ln^2 \frac{|A||S|^4H^3}{\epsilon^2} + \ln \frac{H}{\delta}\Big)\Big).$$

By our choice of β and γ it holds that

$$T_h = O\Big(|S|\frac{|S|^2H^2}{\epsilon^2}\frac{|A|^2|S|^8H^6}{\epsilon^4}\Big(Fdim \ln^2 \frac{|A||S|^4H^3}{\epsilon^2} + \ln \frac{H}{\delta}\Big)\Big) = O\Big(\frac{|A|^2|S|^{11}H^8}{\epsilon^6}\Big(Fdim \ln^2 \frac{|A||S|^4H^3}{\epsilon^2} + \ln \frac{H}{\delta}\Big)\Big).$$

By summing the above for every layer $h \in [H-1]$ we obtain that the sample complexity is

$$O\Big(\frac{|A|^2|S|^{11}H^9}{\epsilon^6}\Big(Fdim \ln^2 \frac{|A||S|^4H^3}{\epsilon^2} + \ln \frac{H}{\delta}\Big)\Big).$$

■

E.7 Useful Lemmas and Theorems

Definition 136 *Let M be $n \times m$ matrix. Let us denote*

$$\|M\|_{1,\infty} = \max_{i \in [n]} \sum_{j=0}^{m-1} |M_{ij}|$$

The following is a known theorem from Markov Chains theory.

Theorem 137 *Let $P_1(\cdot|\cdot,\cdot) : S \times (S \times A) \to [0,1]$ and $P_2(\cdot|\cdot,\cdot) : S \times (S \times A) \to [0,1]$ denote two transition probabilities functions of two MDPs. Let $\pi(\cdot|\cdot) : A \times S \to [0,1]$ be a policy such that the induced Markov chains $P_1^\pi(s'|s) = \langle \pi(\cdot|s), P_1(s'|s,\cdot)\rangle$ and $P_2^\pi(s'|s) = \langle \pi(\cdot|s), P_2(s'|s,\cdot)\rangle$ satisfies for some $\alpha \geq 0$ that*

$$\|P_1^\pi - P_2^\pi\|_{1,\infty} \leq \alpha.$$

Let $q_1^h(\cdot|\pi)$ and $q_2^h(\cdot|\pi)$ be the distribution over states after trajectories of length h of P_1^π and P_2^π respectively. Then,

$$\|q_1^h - q_2^h\|_1 \leq \alpha h.$$