Sample Complexity Characterization for Linear Contextual MDPs

Junze Deng
Department of ECE,
The Ohio State University

Yuan Cheng
National University of Singapore

Shaofeng Zou
Department of EE & CSE,
University at Buffalo

Yingbin Liang
Department of ECE,
The Ohio State University

Abstract

Contextual Markov decision processes (CMDPs) describe a class of reinforcement learning problems in which the transition kernels and reward functions can change over time with different MDPs indexed by a context variable. While CMDPs serve as an important framework to model many real-world applications with time-varying environments, they are largely unexplored from a theoretical perspective. In this paper, we study CMDPs under two linear function approximation models: Model I with context-varying representations and common linear weights for all contexts; and Model II with common representations for all contexts and context-varying linear weights. For both models, we propose novel model-based algorithms and show that they enjoy guaranteed ϵ-suboptimality gap with desired polynomial sample complexity. In particular, instantiating our result for the first model to the tabular CMDP improves the existing result by removing the reachability assumption. Our result for the second model is the first-known result for such a type of function approximation models. Comparison between our results for the two models further indicates that having context-varying features leads to much better sample efficiency than having common representations for all contexts under linear CMDPs.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 Introduction

Reinforcement learning (RL) (Sutton and Barto, 2018) aims to optimize the interaction process between agents and the environment, and has succeeded in many practical applications, e.g., games (Silver et al., 2016), robotics (Levine et al., 2016; Gu et al., 2017), and recommendation systems (Li et al., 2010). Typically, Markov Decision Processes (MDPs) are employed to model these interaction processes (Bertsekas, 2011). In a single MDP, the transition kernel and reward function remain invariant across different episodes. The objective of the agent is to learn a policy that maximizes the cumulative reward under the *same* MDP throughout the interaction.

However, many real-world applications involve *time-varying* transition kernels and reward functions. These scenarios are influenced by an additional variable known as a *context*. To capture such complexities, Contextual Markov Decision Processes (CMDPs) are utilized (Modi et al., 2018; Levy and Mansour, 2022a; Hallak et al., 2015). For instance, consider a multi-user recommendation system where transition probabilities and reward functions may differ significantly among users. Describing the diverse user behaviors using a single MDP becomes challenging. In contrast, a CMDP enables the modeling of a multi-user recommendation system by introducing a context-dependent transition kernel and a context-dependent reward function (Kabra et al., 2021).

Although CMDPs have the capability to model various real-world applications with time-varying environments, study of their theoretical performance remains limited. Recently, Sodhani et al. (2022); Modi et al. (2018) have investigated a particular class of CMDPs, known as Lipschitz CMDPs (or smooth CMDPs). Additionally, Dong et al. (2020); Jiang et al. (2017) studied the class of CMDPs with low Bellman rank. Moreover, Levy and Mansour (2022b) explored the tabular CMDPs with a minimal reachability assumption and a finite state space. To further generalize such study to a large or even infinite-state space, Amani et al. (2022)

studied CMDPs with linear function approximation, where the reward function is context-dependent but the transition kernel is context-independent.

In this paper, we further explore more general CMDP models with linear function approximation, where both the transition kernel and the reward function are context-dependent and are linear functions of features and weights, and the state space is large or even continuous. In particular, we study the following two models for CMDPs, both of which are well justified in practical applications (see Section 3.2 for motivating examples).

- In Model I, the transition and the reward can be decomposed into a linear function of known context-varying representation and unknown common linear weights that are shared across all contexts. *This model generalizes the CMDP in Amani et al. (2022) with a fixed transition to context-varying transitions.*
- In Model II, the transition kernel and the reward function can be decomposed into a linear function of known common representation and unknown context-varying linear weights. *Such a CMDP model has not been studied in the past.*

We summarize our main contributions below.

Novel model-based algorithm. For both models, we design model-based algorithms to enable the use of all historical data (generated under *different* environments) simultaneously for learning the transition model and the reward function. This is in contrast to the model-free algorithms designed for single linear MDPs (Jin et al., 2020) and linear CMDP with fixed transition kernel (Amani et al., 2022), where all all historical data (in the past episodes) are collected under a fixed transition kernel and can hence be used to directly estimate the value function.

Novel bonus term design. For Model II, the bonus term for promoting optimism is designed to be the *squared* norm of features, which is based on a novel decomposition of the value function uncertainty into the context-dependent and context-independent components, so that context-independent components can be bounded by the squared norm of features. This is in contrast to the standard UCB bonus design (that adopts the norm itself) in the previous studies of function approximation for *fixed* transition kernels (Amani et al., 2022; Jin et al., 2020; Hu et al., 2022).

Theoretical guarantee. For both models, we provide provable upper bounds on the average suboptimality gap and the corresponding sample complexity to achieve such a near-optimal performance. Specifically, for Model I, Our result improves that for tabular CMDP in (Levy and Mansour, 2022b) by removing their reachability assumption, and thus enjoys a better sample efficiency if the reachability lower bound is small, which is often the case in practice (Agarwal et al., 2020). Our result for context-varying transitions has the same sample complexity as that in (Amani et al., 2022) for CMDP with fixed transitions, indicating that our handling of context-varying transition does not incur additional sample complexity. For Model II, our result is the first in the literature established for such a type of function approximation models.

2 Related Work

Contextual MDPs (CMDPs). The study of CMDPs was originated by the work of (Hallak et al., 2015). Since then, several specialized classes of CMDPs have been explored, incorporating additional structural assumptions. One notable class is that of smooth CMDPs, which was proposed in (Modi et al., 2018). The authors developed a framework for designing Probably Approximately Correct (PAC) algorithms specifically tailored for smooth CMDPs with a finite context space. This work focused on achieving efficient and accurate decision-making in CMDPs by leveraging the smoothness assumptions. Sodhani et al. (2022) further studied the properties of Lipschitz block CMDPs. Another important class of CMDPs with low Bellman rank was introduced in (Jiang et al., 2017). Further, function approximation techniques were used in (Levy and Mansour, 2022a) to obtain the sample complexity, where they assumed the access to an empirical risk minimization oracle. Then CMDPs with linear function approximation were studied in (Amani et al., 2022), where only the reward is context-dependent. More recently, tabular CMDPs with both context-dependent transition and context-dependent reward were studied in (Levy and Mansour, 2022b). In this paper, we focus on more general CMDPs with both context-dependent transition and context-dependent reward, where the transition kernel can be modeled with linear function approximation. Thus, our models include the CMDPs studied in (Levy and Mansour, 2022b; Amani et al., 2022) as special cases.

Contextual Bandits. Contextual bandits can be viewed as a natural extension of the classical multi-armed bandit problem (Slivkins et al., 2019). In contextual bandit settings, additional information, known as context, is provided to the decision-making agent. This context influences the reward associated with each action. Further, contextual bandits can be viewed as special cases of CMDPs with horizon one. The major challenge here lies in estimating the reward func-

tion, which is commonly solved using regression-based methods, e.g., (Chu et al., 2011; Foster et al., 2018; Agarwal et al., 2014; Xu and Zeevi, 2020).

Adversarial RL and Nonstationary RL. There have been two other lines of research that model time-varying transition kernels and reward functions. The first line is on adversarial RL (Neu et al., 2012; Zimin and Neu, 2013; Lykouris et al., 2021; Rosenberg and Mansour, 2019; Jin et al., 2019; Xiong et al., 2021), which allows time variations in reward functions but assumes an identical transition kernel over time. The second line is on nonstationary RL Zhong et al. (2021); Mao et al. (2021); Touati and Vincent (2020); Zhou et al. (2020); Cheng et al. (2023b); Xiong et al. (2020); Feng et al. (2023), where both transition kernels and reward functions can be time-varying. Note that both adversarial RL and nonstationary RL assume that rewards and/or samples taken under current transitions can be used only in the *next* episode, whereas the agent in contextual MDPs can access and exploit the rewards and samples taken under the *current* MDP (i.e., the context) to achieve better performance.

Linear MDPs. Linear function approximation in Markov Decision Processes (MDPs) has been widely explored in the literature, e.g.,Jin et al. (2020); Du et al. (2019); Wang et al. (2019); He et al. (2021); Zhang et al. (2021); Chu et al. (2011). The use of linear function approximation allows for efficient and scalable representation of value functions or policies in MDPs, facilitating the handling of high-dimensional state spaces. For liner MDPs, (Jin et al., 2020) developed a standard framework of Upper Confidence Bound (UCB) based model-free algorithms, and (He et al., 2022; Hu et al., 2022) designed algorithms that are nearly minimax optimal. Most previous works on linear MDPs (Jin et al., 2020; Amani et al., 2022) adopt model-free approaches to approximate the value function directly. However, these approaches are not directly applicable to settings where the transition kernel is changing over time, i.e., context-varying. In this paper, we develop model-based approaches that effectively take advantage of historical data collected under time-varying transition kernels for better model estimation.

3 Preliminaries and Problem Formulation

Notation. For a positive integer H, let $[H] := \{1, 2, ..., H\}$. For a vector x, define the vector norm of x w.r.t. a positive symmetric matrix A by $\|x\|_A := \sqrt{xA^\top x}$. For a finite set $\mathcal{A}$, we use $\mathcal{U}(\mathcal{A})$ to denote the uniform distribution over the set $\mathcal{A}$.

3.1 Contextual MDPs

A Contextual Markov Desicion Process (CMDP) can be described by a tuple $(\mathcal{W}, \mathcal{S}, \mathcal{A}, \mathcal{M})$, where $\mathcal{W}$ is the context space, which can be continuous or infinite, $\mathcal{S}$ is the state space, which can be continuous and infinite, $\mathcal{A}$ is the finite action space with the cardinality K, and the mapping $\mathcal{M}$ maps a context $w \in \mathcal{W}$ to a Markov Decision Process (MDP): $\mathcal{M}(w) = (\mathcal{S}, \mathcal{A}, P_w, r_w, H)$. Specifically, H is the horizon length, and $P_w = \{P_{w,h}\}_{h=1}^H$ denotes the time-dependent and context-dependent transition kernel, i.e., $P_{w,h}(s'|s, a)$ is the probability of reaching state s' in the next step given the state-action pair (s, a) at step h when the context is w. For convenience, we denote $P_h(s'|s, a, w) := P_{w,h}(s'|s, a)$. Furthermore, $r_w = \{r_{w,h}\}_{h=1}^H$ denotes the deterministic reward function where $r_{w,h} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function at step h given the context is w. We also write $r_h(s, a, w) := r_{w,h}(s, a)$.

At the beginning of each episode, a context w is drawn randomly from a distribution q, and the agent then experiences the MDP $\mathcal{M}(w)$ for the current episode. Each episode starts with a fixed initial state s_1 independent of the context. At each step h, the agent observes a state s_h, takes an action $a_h \in \mathcal{A}$ under a possibly context-dependent policy π_w, receives a reward $r_h(s_h, a_h, w)$, and the system transits to the next state s_{h+1} following the probability $P_h(s_{h+1}|s_h, a_h, w)$.

For a given MDP, we use $a_h \sim \pi$ to denote that an action a_h is selected according to a policy π. We use $s_h \sim (P, \pi)$ to denote the distribution of s_h applying policy π under the transition kernel P for $h - 1$ steps. Then we use $\mathbb{E}_{(s_h,a_h)\sim(P,\pi)}$ to denote the expectation over states $s_h \sim (P, \pi)$ and actions $a_h \sim \pi$. Given an MDP: $M = (\mathcal{S}, \mathcal{A}, P, r, H)$ and a policy π, we denote by $V_{h,M}^\pi$ the value function under the MDP M and the policy π starting from step h and state s_h, i.e.,

$$V_{h,M}^\pi = \mathbb{E}_{(s_{h'},a_{h'})\sim(P,\pi)} \left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'})|s_h\right].$$

For simplicity, we use V_M^π to denote $V_{1,M}^\pi$. For the MDP $\mathcal{M}(w) = (\mathcal{S}, \mathcal{A}, P_w, r_w, H)$, we use π_w^* to denote an optimal policy that maximizes the value function w.r.t. the corresponding context w, i.e.,

$$\pi_w^* = \arg\max_\pi V_{\mathcal{M}(w)}^\pi.$$

For the CMDP problem, our goal is to obtain a series of policies $\{\pi_w^n\}_{n=1}^N$ over time steps $n = 1, \ldots, N$ that minimize the following *average sub-optimality gap*, which takes the expectation over context and the average over time:

$$\frac{1}{N} \sum_{n=1}^N \mathop{\mathbb{E}}_{w\sim q} \left[V_{\mathcal{M}(w)}^{\pi_w^*} - V_{\mathcal{M}(w)}^{\pi_w^n}\right].$$

A sequence of policy $\{\pi_w^n\}_{n=1}^N$ is ϵ-optimal if the average suboptimality gap is less than ϵ.

3.2 Two Linear Function Approximation Models for CMDPs

In this paper, we study two different linear function approximation models with context-varying transition kernels and reward functions.

The first model defined below has context-varying representations and the same linear weights for all contexts.

Definition 1 (Model I: CMDPs with varying representation)**.** *The environment transition kernel P_w admits a linear decomposition of a known representation $\phi_h : \mathcal{S} \times \mathcal{A} \times \mathcal{W} \to \mathbb{R}^d$ and an unknown linear weights function $\mu_h : \mathcal{S} \to \mathbb{R}^d$ as follows:*

$$P_h(s'|s, a, w) = \langle \phi_h(s, a, w), \mu_h(s') \rangle. \tag{1}$$

Moreover, the reward function r_w admits a similar linear decomposition of a known feature function $\psi_h : \mathcal{S} \times \mathcal{A} \times \mathcal{W} \to \mathbb{R}^d$ and an unknown linear weights $\eta_h \in \mathbb{R}^d$:

$$r_h(s, a, w) = \langle \psi_h(s, a, w), \eta_h \rangle. \tag{2}$$

For normalization, we assume $\|\phi_h(s, a, w)\|_2 \leq 1$ and $\|\psi_h(s, a, w)\|_2 \leq 1$ for any $(s, a, w) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W}$. For any function $g \to [0, 1]$, we assume $\left\|\int \mu_h(s) g(s) ds\right\|_2 \leq \sqrt{d}$. Furthermore, we assume that $\|\eta_h\|_2 \leq \sqrt{d}$.

Model I captures read-world applications in which the environment transition depends on a few dominating features, where the features can vary rather quickly but the weights are relatively steady. For example, in multi-user recommendation systems, features are user-dependent and change across uses. But the importance of individual features can remain the same regardless of user identity, meaning that the roles that these features play in the system dynamics are the same for all users.

In this model, we take the following assumption widely used in the RL literature (Cheng et al., 2023a; Levy and Mansour, 2022b; Sun et al., 2019).

Assumption 1. *The learning agent has access to a finite model class Ψ_1, where the true model $\mu_h(s) \in \Psi_1$ for any $h \in [H]$.*

In this paper, we assume that the cardinality of the function class is finite. It can be extended to an infinite function class with bounded statistical complexity such as a bounded covering number (Sun et al., 2019; Uehara et al., 2021).

The second model has the same features for all contexts but context-varying linear weights.

Definition 2 (Model II: CMDPs with varying linear weights)**.** *The environment transition kernel P_w admits a linear decomposition with a known representation $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and an unknown linear weight function $\mu_h : \mathcal{S} \times \mathcal{W} \to \mathbb{R}^d$:*

$$P_h(s'|s, a, w) = \langle \phi_h(s, a), \mu_h(s', w) \rangle. \tag{3}$$

Moreover, the reward function r_w admits a similar linear decomposition of a known feature function $\psi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and an unknown linear weights function: $\eta_h : \mathcal{W} \to \mathbb{R}^d$:

$$r_h(s, a, w) = \langle \psi_h(s, a), \eta_h(w) \rangle. \tag{4}$$

For normalization, we assume $\|\phi_h(s, a)\|_2 \leq 1$ and $\|\psi_h(s, a)\|_2 \leq 1$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. For any $w \in \mathcal{W}$ and function $g \to [0, 1]$, we assume $\left\|\int \mu_h(s, w) g(s) ds\right\|_2 \leq \sqrt{d}$. Furthermore, we assume that $\|\eta_h(w)\|_2 \leq \sqrt{d}$ for any $w \in \mathcal{W}$.

Model II captures many read-world applications where environmental transitions rely on relatively stable features but with varying weights depending on contexts. For example, suppose an autonomous car drives over several traffic patterns. These traffic patterns can share common features, but the role that these features play in the system dynamics can change with traffic patterns, i.e., the linear weights are different across contexts.

For the second model, we take the following assumption similar to the previous one.

Assumption 2. *The learning agent has access to two finite model classes Ψ_2 and Ψ_3, where the true model $\mu_h(s, a) \in \Psi_2$ and $\eta_h(s, w) \in \Psi_3$ for any $h \in [H]$.*

The above two linear contextual MDP models take the linear function approximation structure in the standard linear MDP (Jin et al., 2020) for each context. However, solving a CMDP is significantly harder than solving a single MDP due to the context-varying transition kernel. The key challenges lie in how to use historical data taken over different MDPs to benefit the learning in future MDPs; and the context and space can be infinite.

4 Model I: Varying Representation

In this section, we consider the CMDPs with varying representations as defined in Definition 1.

Technical Challenge: Although there exists a line of works in single linear MDPs (Jin et al., 2020) and

linear CMDP with fixed transition kernel and context-dependent reward (Amani et al., 2022), directly extending their model-free algorithms to our model is challenging. This is because in those studies, the transition kernel is fixed, and hence all historical data (in the past episodes) can be used to directly estimate the value function, because they are generated via the same environment. This is not feasible for our model, because the transition kernel is context-varying. Therefore, we will design a model-based algorithm, which makes it convenient to use all historical data (generated under different environments) to learn the transition model and the reward function first, and then conduct planning for each context.

4.1 Algorithm

We propose a novel model-based algorithm, particularly designed to exploit data generated by context-varying environments. As we mention above, the previous model-free algorithms for fixed transition kernel (Jin et al., 2020; Amani et al., 2022) are not applicable here. Our algorithm is presented in Algorithm 1.

Estimation of transition kernel and reward. At the beginning of each episode n, the agent observes a context w_n. For each step $h \in [H]$, the agent uses historical data $\{(s_h^\tau, a_h^\tau, w_\tau, r_h^\tau)\}_{\tau=1}^{n-1}$ to estimate the reward function as follows:

$$\hat{\eta}_h^n = \underset{\eta_h \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{\tau=1}^{n-1} \left(\langle \eta_h, \psi_h(s_h^\tau, a_h^\tau, w_\tau)\rangle - r_h^\tau\right)^2 + \xi_n \left\|\eta_h\right\|_2^2,$$

where $\xi_n > 0$ is some constant to be specified later. Moreover, for each step $h \in [H]$, the agent uses an maximum likelihood estimation (MLE) oracle on the collected data set $\mathcal{D}_h^n = \{(s_h^\tau, a_h^\tau, s_{h+1}^\tau, w_\tau)\}_{\tau=1}^{n-1}$ to estimate the weights of transition kernel as follows:

$$\begin{aligned}\hat{\mu}_h^n &= \mathrm{MLE}(\mathcal{D}_h^n) \\ &= \underset{\mu_h \in \Psi_1}{\operatorname{argmax}} \sum_{(s,a,s',w) \in \mathcal{D}_h^n} \log\langle \phi_h(s,a,w), \mu_h(s')\rangle.\end{aligned} \tag{5}$$

Design of UCB bonus terms. We define the following two matrices:

$$\begin{aligned}\hat{\Sigma}_h^n &= \sum_{\tau=1}^{n-1} \phi_h(s_h^\tau, a_h^\tau, w_\tau)\phi_h(s_h^\tau, a_h^\tau, w_\tau)^\top + \lambda_n I, \\ \hat{\Lambda}_h^n &= \sum_{\tau=1}^{n-1} \psi_h(s_h^\tau, a_h^\tau, w_\tau)\psi_h(s_h^\tau, a_h^\tau, w_\tau)^\top + \xi_n I,\end{aligned} \tag{6}$$

where $\lambda_n = \gamma_1 d\log(2nH/\delta)$, $\xi_n = \gamma_2 d\log(2nH/\delta)$, $\gamma_1, \gamma_2 = \mathcal{O}(1)$ and I denotes the identity matrix. The UCB bonus terms for the transition kernel and the reward are then defined as follows:

$$\hat{b}_h^n(s,a,w) = \min\{\alpha_n \left\|\phi_h(s,a,w)\right\|_{(\hat{\Sigma}_h^n)^{-1}}, H\}, \tag{7}$$

$$\hat{c}_h^n(s,a,w) = \min\{\beta_n \left\|\psi_h(s,a,w)\right\|_{(\hat{\Lambda}_h^n)^{-1}}, 1\}, \tag{8}$$

where $\alpha_n = 5H\sqrt{2\lambda_n d + 4\log(2nH|\Psi_1|/\delta)}$ and $\beta_n = \sqrt{d\xi_n}$. Note that $\lambda_n, \xi_n = \mathcal{O}(d)$, and therefore, $\alpha_n = \widetilde{\mathcal{O}}(dH)$ and $\beta_n = \mathcal{O}(d)$. Moreover, the parameter α_n depends on the size of the model class Ψ_1. The design of α_n is to bound the gap of value functions due to the estimation error of the transition kernel. The design of β_n is to bound the gap of value functions due to the estimation error of the reward function. See further details in Remark 1.

Then the agent uses the estimated function $\hat{\mu}_h^n(s)$ to update the transition as

$$\hat{P}_h^n(s'|s,a,w) = \langle \phi_h(s,a,w), \hat{\mu}_h^n(s')\rangle,$$

and uses the estimated reward weights $\hat{\eta}_h^n$ with defined bonus terms to update the optimistic reward function as

$$\hat{r}_h^n(s,a,w) = \hat{f}_h^n(s,a,w) + \hat{b}_h^n(s,a,w) + \hat{c}_h^n(s,a,w), \tag{9}$$

where $\hat{b}_h^n$ and $\hat{c}_h^n$ are defined in eqs. (7) and (8) and

$$\hat{f}_h^n(s,a,w) = \begin{cases} \langle \hat{\eta}_h^n, \psi_h(s,a,w)\rangle & \text{if} \langle \hat{\eta}_h^n, \psi_h(s,a,w)\rangle \in [0,1] \\ 1 & \text{if} \langle \hat{\eta}_h^n, \psi_h(s,a,w)\rangle > 1 \\ 0 & \text{if} \langle \hat{\eta}_h^n, \psi_h(s,a,w)\rangle < 0 \end{cases}. \tag{10}$$

Remark 1. *It can be shown that with high probability, we have:*

$$|(P_{w,h} - \hat{P}_{w,h}^n) V_{h+1, \mathcal{M}^{(r,\hat{P}^n)}(w)}(s,a)| \le \hat{b}_h^n(s,a,w)$$

and

$$|\langle \eta_h - \hat{\eta}_h^n, \psi_h(s,a,w)\rangle| \le \hat{c}_h^n(s,a,w)$$

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

Planning and exploration. For the estimated MDP $\hat{\mathcal{M}}(w) = (\mathcal{S}, \mathcal{A}, H, \hat{P}_w^n, \hat{r}_w^n)$, the agent defines a truncated value function iteratively using the optimistic reward function and the estimated transition kernel:

$$\begin{aligned}&\bar{\bar{Q}}_{h,\hat{\mathcal{M}}(w)}^{\pi_w}(s_h, a_h) \\ &\quad = \min\left\{3H, \hat{r}_h^n(s,a,w) + \hat{P}_{h,w}^n \bar{\bar{V}}_{h+1,\hat{\mathcal{M}}(w)}^{\pi_w}(s_h, a_h)\right\}, \\ &\bar{\bar{V}}_{h,\hat{\mathcal{M}}(w)}^{\pi_w}(s_h) = \mathop{\mathbb{E}}_{\pi_w}\left[\bar{\bar{Q}}_{h,\hat{\mathcal{M}}(w)}^{\pi_w}(s_h, a_h)\right].\end{aligned} \tag{11}$$

Note that the truncation threshold of $3H$ is specially designed to provide a valid bound for the optimistic reward, which consists of three elements including the

bonus term for the kernel estimation error, the bonus term for the reward estimation error and the estimated reward, and each of their corresponding value functions can be truncated to H.

In the first episode, the agent starts with a random policy and executes such a policy to collect data. Then in the following episodes, the agent finds an optimal context-varying policy π_w^n of the truncated value function:

$$\pi_w^n = \arg\max_{\pi} \bar{\bar{V}}^{\pi}_{\hat{\mathcal{M}}(w)}. \tag{12}$$

Then the agent executes the policy π_w^n in the current episode and collects data.

Algorithm 1 CMDPs with varying representation

Initialization: regularizers λ_n, ξ_n, model class Ψ_1, MLE data set $\mathcal{D}_h^n = \emptyset$

for $n = 1, ..., N$ **do**
  observe context w_n.
  if $n = 1$ **then**
    set $\pi^1_{w_1}$ as a random policy
  else
    for $h = 1, ..., H$ **do**
      $\hat{\eta}_h^n = \left(\hat{\Lambda}_h^n\right)^{-1} \cdot \sum_{\tau=1}^{n-1} \psi_h(s_h^\tau, a_h^\tau, w_\tau) r_h^\tau$,
      where $\hat{\Lambda}_h^n$ is defined in eq. (6).
      $\hat{\mu}_h^n = \text{MLE}(\mathcal{D}_h^n)$.
      $\hat{P}_h^n(s'|s,a,w) = \langle \phi_h(s,a,w), \hat{\mu}_h^n(s') \rangle$.
      $\hat{r}_h^n(s,a,w) = \hat{f}_h^n(s,a,w) + \hat{b}_h^n(s,a,w)$
        $+\hat{c}_h^n(s,a,w)$,
      where $\hat{r}_h^n$ is defined in eq. (9).
    end for
    $\hat{\mathcal{M}}(w) = (\mathcal{S}, \mathcal{A}, H, \hat{P}_w^n, \hat{r}_w^n)$.
    $\pi^n_{w_n} = \text{argmax}_\pi \ \bar{\bar{V}}^{\pi}_{\hat{\mathcal{M}}(w_n)}$,
    where $\bar{\bar{V}}^{\pi}_{\hat{\mathcal{M}}(w_n)}$ is defined in eq. (11)
  end if
  Execute policy $\pi^n_{w_n}$, collect the trajectory:
  $s_1^n, a_1^n, r_1^n, ..., s_H^n, a_H^n, r_H^n$.
  $\mathcal{D}_h^{n+1} = \mathcal{D}_h^n \cup \{(s_h^n, a_h^n, s_{h+1}^n, w_n)\}$ for $h \in [H]$.
end for

4.2 Theoretical Analysis

In this section, we develop an upper bound on the average sub-optimality gap and characterize the required sample complexity for finding a near-optimal policy in Theorem 1. The detailed proof of Theorem 1 is deferred to Appendix A.

Theorem 1. *Consider a CMDP with varying representations as defined in Definition 1. Under Assumption 1, for any $\delta \in (0,1)$, with probability at least $1 - 3\delta/2$, the sequence of policies $\{\pi^n_{w_n}\}_{n=1}^N$ generated by Algorithm 1 satisfies that*

$$\frac{1}{N}\sum_{n=1}^{N} \mathop{\mathbb{E}}_{w\sim q}\left[V^{\pi_w^*}_{\mathcal{M}(w)} - V^{\pi_w^n}_{\mathcal{M}(w)}\right] \leq \left(42H\sqrt{2d\lambda_N + 4\log(2HN|\Psi_1|/\delta)} + 6\sqrt{d\xi_N}\right) \cdot H\sqrt{\frac{2d}{N}}\sqrt{\log\left(1 + \frac{N}{d\lambda}\right)},$$

where $\lambda = \min\{\lambda_1, \xi_1\}$. To achieve an ϵ average sub-optimality gap, at most $\mathcal{O}\left(\frac{H^4 d^3 \log(|\Psi_1|/\delta)}{\epsilon^2}\right)$ episodes are needed.

We highlight the significance of Theorem 1 via its comparisons with highly relevant existing studies as follows. **(i)** Our result generalizes that in (Amani et al., 2022) with only context-dependent reward to both context-dependent transition kernel and context-dependent reward. Our result has the same sample complexity as that in (Amani et al., 2022) indicating that our handling of context-varying transition does not incur additional sample complexity compared to the setting with a fixed transition kernel. **(ii)** Our result improves that in (Levy and Mansour, 2022b) by removing the reachability assumption required for tabular MDP, and thus enjoys a better sample efficiency if $q_{\min}$ is small, which is often the case in practice (Agarwal et al., 2020). **(iii)** Our sample complexity for CMDPs is $\tilde{\mathcal{O}}(\frac{H^4 d^3}{\epsilon^2})$, which is the same as that of LSVI-UCB for a single MDP (Jin et al., 2020). Thus, with the same sample efficiency, our approach can also solve CMDPs, as long as the contexts share the common linear weights.

5 Model II: Varying Linear Weights

In this section, we consider Model II with unknown context-varying linear weights and the same known features for all contexts as defined in Definition 2.

Technical Challenge: The bonus term design to promote optimism for Model II cannot apply the standard UCB bonus term in the previous studies of function approximation for fixed transition kernels (Amani et al., 2022; Jin et al., 2020; Hu et al., 2022). The reason is that the features (i.e., representations) in Model II are fixed, and hence cannot fully upper-bound the context-varying value function gap due to the estimation error of the transition kernel and the reward function. Therefore, we will develop a novel decomposition of the value function gap into the context-dependent and context-independent components, so that context-independent components can be bounded by the squared norm of features.

5.1 Algorithm

We adopt a model-based design to learn the transition model first, and conduct planning for each new context. The detailed algorithm is presented in Algorithm 2.

Estimation of transition kernel and reward. At the beginning of episode n, the agent observes a context w_n. The agent uses collected data $\mathcal{D}_h^n = \{(s_h^\tau, a_h^\tau, s_{h+1}^\tau, r_h^\tau, w_\tau)\}_{\tau=1}^{n-1}$ to estimate the transition by an MLE oracle:

$$\begin{aligned}\widetilde{\mu}_h^n &= \mathrm{MLE}(\mathcal{D}_h^n) \\ &= \operatorname*{argmax}_{\mu_h \in \Psi_2} \sum_{(s,a,s',w)\in\mathcal{D}_h^n} \log\langle \phi_h(s,a), \mu_h(s',w)\rangle\end{aligned}$$

and the reward function by a least square regression (LSR) oracle:

$$\begin{aligned}\widetilde{\eta}_h^n &= \mathrm{LSR}(\mathcal{D}_h^n) \\ &= \operatorname*{argmin}_{\eta_h \in \Psi_3} \sum_{(s,a,r,w)\in\mathcal{D}_h^n} \left(\langle \psi_h(s,a), \eta_h(w)\rangle - r\right)^2.\end{aligned}$$

Novel design of UCB bonus terms. We first define the following matrices:

$$\begin{aligned}\widetilde{\Sigma}_h^n &= \sum_{\tau=1}^{n-1} \phi_h(s_h^\tau, a_h^\tau)\phi_h(s_h^\tau, a_h^\tau)^\top + \widetilde{\lambda}_n I, \\ \widetilde{\Lambda}_h^n &= \sum_{\tau=1}^{n-1} \psi_h(s_h^\tau, a_h^\tau)\psi_h(s_h^\tau, a_h^\tau)^\top + \widetilde{\xi}_n I, \end{aligned} \tag{13}$$

where $\widetilde{\lambda}_n = \widetilde{\gamma}_1 d\log(2nH/\delta)$, $\widetilde{\xi}_n = \widetilde{\gamma}_2 d\log(2nH/\delta)$ and $\widetilde{\gamma}_1, \widetilde{\gamma}_2 = \mathcal{O}(1)$.

To design a UCB bonus term, we first note that the following important remark.

Remark 2. *It can be shown (see Lemma 12 in the Appendix) that the value function gap $|(P_{w,h} - \widetilde{P}_{w,h}^n)V_{h+1,\mathcal{M}^{(r,\tilde{P}^n)}}(s,a)|$ for any $(s,a,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$ can be decomposed (via an upper bound) into the context-independent and context-dependent components, where the features $\phi_h(s,a)$ can play a role only in the context-independent term because the features do not change over contexts in Model II.*

Lemma 12 in the Appendix also indicates that only the *squared* norm of the features $\phi_h(s,a)$ will serve as valid bonus terms to upper-bound the context-independent component of the value function gap. This is very different from the standard UCB bonus terms based on the norm of features in the previous studies for fixed transition kernels (Amani et al., 2022; Jin et al., 2020; Hu et al., 2022). Thus, we design the UCB bonus term $\widetilde{b}_h^n(s,a)$ as follows:

$$\widetilde{b}_h^n(s,a) = \min\left\{\widetilde{\alpha}_n \left\|\phi_h(s,a)\right\|^2_{(\widetilde{\Sigma}_h^n)^{-1}}, H\right\}, \tag{14}$$

where $\widetilde{\alpha}_n = \frac{25}{2\sqrt{K}} CH\sqrt{dN}$ with $C = \frac{p_{\max}}{p_{\min}}$, and $p_{\max}$ and $p_{\min}$ are defined in Assumption 3. Similarly, we define the bonus term for the estimation error of the reward function as:

$$\widetilde{c}_h^n(s,a) = \min\left\{\widetilde{\beta}_n \left\|\psi_h(s,a)\right\|^2_{(\widetilde{\Lambda}_h^n)^{-1}}, 1\right\}, \tag{15}$$

where $\widetilde{\alpha}_n = \frac{25}{2\sqrt{K}} C\sqrt{dN}$. Then the agent uses the estimated function $\widetilde{\mu}_h^n(s,w)$ to update the estimate of the transition kernel as

$$\widetilde{P}_h^n(s'|s,a,w) = \langle \phi_h(s,a), \widetilde{\mu}_h^n(s',w)\rangle,$$

and further updates the optimistic reward function using the estimated function $\widetilde{\eta}_h^n(w)$ and the above defined bonus terms as

$$\widetilde{r}_h^n(s,a,w) = \widetilde{f}_h^n(s,a,w) + \widetilde{b}_h^n(s,a) + \widetilde{c}_h^n(s,a), \tag{16}$$

where $\widetilde{b}_h^n$ and $\widetilde{c}_h^n$ are defined in eqs. (14) and (15) and

$$\widetilde{f}_h^n(s,a,w) = \begin{cases} \langle \widetilde{\eta}_h^n(w), \psi_h(s,a)\rangle & \text{if} \langle \widetilde{\eta}_h^n(w), \psi_h(s,a)\rangle \in [0,1] \\ 1 & \text{if} \langle \widetilde{\eta}_h^n(w), \psi_h(s,a)\rangle > 1 \\ 0 & \text{if} \langle \widetilde{\eta}_h^n(w), \psi_h(s,a)\rangle < 0 \end{cases}. \tag{17}$$

Planning and exploration. For the estimated MDP $\widetilde{\mathcal{M}}(w) = (\mathcal{S}, \mathcal{A}, H, \widetilde{P}_w^n, \widetilde{r}_w^n)$, the agent defines a truncated value function iteratively using the optimistic reward function and the estimated transition kernel:

$$\begin{aligned}&\bar{\bar{Q}}^{\pi_w}_{h,\widetilde{\mathcal{M}}(w)}(s_h,a_h) \\ &\quad = \min\left\{3H, \widetilde{r}_h^n(s,a,w) + \widetilde{P}_{h,w}^n \bar{\bar{V}}^{\pi_w}_{h+1,\widetilde{\mathcal{M}}(w)}(s_h,a_h)\right\}, \\ &\bar{\bar{V}}^{\pi_w}_{h,\widetilde{\mathcal{M}}(w)}(s_h) = \mathop{\mathbb{E}}_{\pi_w}\left[\bar{\bar{Q}}^{\pi_w}_{h,\widetilde{\mathcal{M}}(w)}(s_h,a_h)\right]. \end{aligned} \tag{18}$$

In the first episode, the agent starts with a random policy and executes such a policy to collect data. Then in the following episodes, with the estimated transition kernel $\widetilde{P}_w^n$ and the optimistic reward $\widetilde{r}_w^n$, the agent finds the optimal context-dependent policy π_w^n for the MDP $\widetilde{\mathcal{M}}(w) = (\mathcal{S}, \mathcal{A}, H, \widetilde{P}_w^n, \widetilde{r}_w^n)$ as

$$\pi_{w_n}^n = \operatorname{argmax} \bar{\bar{V}}^{\pi}_{\widetilde{\mathcal{M}}(w_n)}.$$

To collect data, the agent does not simply execute the policy $\pi_{w_n}^n$ in the entire episode. Instead, for each $h \in [H]$, the agent first executes $\pi_{w_n}^n$ for h steps, and then chooses the next action according to a uniform distribution $\mathcal{U}(\mathcal{A})$ and observes the next state. In this way, s_h^n follows the distribution of $(P, \pi_{w_n}^n)$, and a_h^n follows the uniform distribution. Such a uniform choice of actions provides a context-independent distribution of the action, which helps the agent to use history data from all previous contexts to estimate the transition and the reward function of the current context.

Algorithm 2 CMDPs with varying linear weights

Initialization: regularizers $\widetilde{\lambda}_n, \widetilde{\xi}_n$, model classes Ψ_2 and Ψ_3, MLE data set $\mathcal{D}_h^n = \emptyset$.

for $n = 1, ..., N$ **do**
  observe context w_n.
  if $n = 1$ **then**
    set $\pi_{w_1}^1$ as a random policy.
  else if $n \geq 2$ **then**
    for $h = 1, ..., H$ **do**
      $\widetilde{\mu}_h^n = \text{MLE}(\mathcal{D}_h^n)$, $\widetilde{\eta}_h^n = \text{LSR}(\mathcal{D}_h^n)$.
      $\widetilde{P}_h^n(s'|s,a,w) = \langle \phi_h(s,a), \widetilde{\mu}_h^n(s', w) \rangle$.
    end for
    $\widetilde{r}_h^n(s,a,w) = \widetilde{f}_h^n(s,a,w) + \widetilde{b}_h^n(s,a) + \widetilde{c}_h^n(s,a)$,
    where $\widetilde{r}_h^n$ is defined in eq. (16)
    $\widetilde{\mathcal{M}}(w) = (S, A, H, \widetilde{P}_w^n, \widetilde{r}_w^n)$.
    $\pi_{w_n}^n = \text{argmax}\ \bar{V}^{\pi}_{\widetilde{\mathcal{M}}(w_n)}$.
  end if
  for $h = 1, ..., H$ **do**
    use $\pi_{w_n}^n$: roll into s_h, take an action uniformly $a_h \sim \mathcal{U}(A)$, reach next state s_{h+1}.
    collect trajectory $s_1^n, a_1^n, ..., s_h^n, a_h^n, s_{h+1}^n$.
    $\mathcal{D}_h^{n+1} = \mathcal{D}_h^n \cup \{(s_h^n, a_h^n, s_{h+1}^n, r_h^n, w_n)\}$.
  end for
end for

5.2 Theoretical Analysis

In this subsection, we develop an upper bound on the average sub-optimality gap and characterize the sample complexity for finding a near-optimal policy.

Suppose that Assumption 2 holds. We further adopt a standard assumption also taken by the study of CMDP (Levy and Mansour, 2022b). Specifically, given the context w and policy π_w, let $s_h \sim (P_w, \pi_w)$, and let $p(s_h|\pi_w, P_w)$ denote the density function of s_h.

Assumption 3. *For any context $w \in \mathcal{W}$, any step $h \in [H]$ and any context-dependent policy π_w, there exists constants $0 < p_{\min} \leq p_{\max} < \infty$ such that $p(s_h|\pi_w, P_w) \in [p_{\min}, p_{\max}]$ for any $s_h \in \mathcal{S}$.*

As discussed in Levy and Mansour (2022b), the existence of $p_{\min}$ can be removed by mixing each transition kernel with a uniform distribution, while still keeping the sub-optimality gap sublinear without $p_{\min}$. This assumption helps to bound the maximal uncertainty of reaching a state at step h under any context and policy. In this way, the maximal difference between the probability of reaching the current state and reaching any previous state in the history data can be estimated, which allows to guarantee the accuracy of learning the current MDP based on history data.

We present the following theorem and defer the detailed proof of Theorem 2 to Appendix B.

Theorem 2. *Consider Model II of CMDP with varying linear weights as defined in definition 2. Under Assumptions 2 and 3, for any $\delta \in (0,1)$, with probability at least $1 - 3\delta/2$, the sequence of policies $\pi_{w_n}^n$ generated by Algorithm 2 satisfies that*

$$\begin{aligned}
&\tfrac{1}{N}\textstyle\sum_{n=1}^{N} \mathop{\mathbb{E}}_{w\sim q}\left[V_{\mathcal{M}(w)}^{\pi_w^*} - V_{\mathcal{M}(w)}^{\pi_w^n}\right] \\
&\leq 912CH^2\sqrt{\tfrac{d^3K}{N}}\log\left(1 + \tfrac{N}{\widetilde{\lambda}d}\right) \\
&\quad + \tfrac{H^2}{C}\sqrt{\tfrac{K}{dN}}\left(2\widetilde{\xi}_N d + C^2\log\left(\tfrac{2HN|\Psi_3|}{\delta}\right) + 4\widetilde{\lambda}_N d\right. \\
&\quad \left. + 8C^2\log\left(\tfrac{2HN|\Psi_2|}{\delta^2}\right)\right),
\end{aligned}$$

where $\widetilde{\lambda} = \min\{\widetilde{\lambda}_1, \widetilde{\xi}_1\}$ and $C = \sqrt{\frac{p_{\max}}{p_{\min}}}$.
To achieve an ϵ average sub-optimality gap, at most $\mathcal{O}\left(\frac{H^4 d^3 K \log^2(|\Psi_2||\Psi_3|/\delta^2)}{\epsilon^2} \cdot \frac{p_{\max}}{p_{\min}}\right)$ episodes are needed.

Compared to Theorem 1 on Model I, Theorem 2 on Model II requires an additional factor of $\mathcal{O}\left(\frac{Kp_{\max}}{p_{\min}}\right)$ in the sample complexity. This is mainly because that the unknown weights are context-varying, and a uniform choice of actions was adopted to facilitate the learning of the varying weights. Such a uniform choice of actions causes an additional factor of $\mathcal{O}(K)$ in the sample complexity. Moreover, in order to use data collected in the previous contexts to estimate the current MDP, handling the distribution shift among different contexts and policies further introduces an additional sample complexity of order $\mathcal{O}\left(\frac{p_{\max}}{p_{\min}}\right)$.

The result in Theorem 2 improves existing studies as follows. **(i)** Our result generalizes those results in (Amani et al., 2022) to cases where the transition kernel can also be context-varying. **(ii)** Our Model II includes the tabular CMDP in (Levy and Mansour, 2022b) as a special case, whereas our model is more general allowing the state space to be infinite.

6 Conclusion

In this paper, we investigated CMDPs whose transition kernel and reward are both context-varying. More specifically, we considered two different linear function approximation models, which are defined in Model I and Model II respectively. We designed model-based methods for both models, where the design for each model features novel elements to deal with the unique challenge of the model. We further provided provable upper bounds on the average sub-optimality gap for both models and the corresponding sample complexity to achieve ϵ average sub-optimality gap.

Acknowledgments

The work of J. Deng and Y. Liang was supported in part by the U.S. National Science Foundation under the grants CCF-1761506, RINGS- 2148253, ECCS-2113860, and DMS-2134145. The work of S. Zou was supported by the National Science Foundation under Grant 2007783.

References

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. (2014). Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning (ICML)*.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. (2020). Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems (NeurIPS)*, 33:20095–20107.

Amani, S., Yang, L. F., and Cheng, C.-A. (2022). Provably efficient lifelong reinforcement learning with linear function approximation. *arXiv preprint arXiv:2206.00270*.

Bertsekas, D. P. (2011). Dynamic Programming and Optimal Control 3rd edition, volume II. *Belmont, MA: Athena Scientific*.

Cheng, Y., Huang, R., Yang, J., and Liang, Y. (2023a). Improved sample complexity for reward-free reinforcement learning under low-rank MDPs. *International Conference on Learning Representations (ICLR)*.

Cheng, Y., Yang, J., and Liang, Y. (2023b). Provably efficient algorithm for nonstationary low-rank mdps. *Advances in Neural Information Processing Systems (NeurIPS)*.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Dong, K., Peng, J., Wang, Y., and Zhou, Y. (2020). Root-n-regret for learning in markov decision processes with function approximation and low bellman rank. In *Conference on Learning Theory (COLT)*.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. (2019). Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*.

Feng, S., Yin, M., Huang, R., Wang, Y.-X., Yang, J., and Liang, Y. (2023). Non-stationary reinforcement learning under general function approximation. *International Conference on Machine Learning (ICML)*.

Foster, D., Agarwal, A., Dudík, M., Luo, H., and Schapire, R. (2018). Practical contextual bandits with regression oracles. In *International Conference on Machine Learning (ICML)*.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *IEEE international conference on robotics and automation*.

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

He, J., Zhao, H., Zhou, D., and Gu, Q. (2022). Nearly minimax optimal reinforcement learning for linear markov decision processes. *arXiv preprint arXiv:2212.06132*.

He, J., Zhou, D., and Gu, Q. (2021). Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning (ICML)*.

Hu, P., Chen, Y., and Huang, L. (2022). Nearly minimax optimal reinforcement learning with linear function approximation. In *International Conference on Machine Learning (ICML)*.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning (ICML)*.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. (2019). Learning adversarial MDPs with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192*.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory (COLT)*.

Kabra, A., Agarwal, A., and Parihar, A. S. (2021). Potent real-time recommendations using multimodel contextual reinforcement learning. *IEEE Transactions on Computational Social Systems*.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373.

Levy, O. and Mansour, Y. (2022a). Learning efficiently function approximation for contextual MDP. *arXiv preprint arXiv:2203.00995*.

Levy, O. and Mansour, Y. (2022b). Optimism in face of a context: Regret guarantees for stochastic contextual mdp. *arXiv preprint arXiv:2207.11126*.

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

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. (2021). Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory (COLT)*.

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., and Basar, T. (2021). Near-optimal model-free reinforcement learning in non-stationary episodic mdps. In *International Conference on Machine Learning (ICML)*.

Modi, A., Jiang, N., Singh, S., and Tewari, A. (2018). Markov decision processes with continuous side information. In *Algorithmic Learning Theory (ALT)*.

Neu, G., György, A., and Szepesvári, C. (2012). The adversarial stochastic shortest path problem with unknown transition probabilities. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Rosenberg, A. and Mansour, Y. (2019). Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning (ICML)*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., and Lanctot, M. (2016). Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484.

Slivkins, A. et al. (2019). Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286.

Sodhani, S., Meier, F., Pineau, J., and Zhang, A. (2022). Block contextual mdps for continual learning. In *Learning for Dynamics and Control Conference*, pages 608–623. PMLR.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. (2019). Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory (COLT)*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, Massachusetts.

Touati, A. and Vincent, P. (2020). Efficient learning in non-stationary linear markov decision processes. *CoRR*, abs/2010.12870.

Uehara, M., Zhang, X., and Sun, W. (2021). Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*.

Wang, Y., Wang, R., Du, S. S., and Krishnamurthy, A. (2019). Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*.

Xiong, H., Xu, T., Liang, Y., and Zhang, W. (2021). Non-asymptotic convergence of adam-type reinforcement learning algorithms under markovian sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10460–10468.

Xiong, H., Zhao, L., Liang, Y., and Zhang, W. (2020). Finite-time analysis for double q-learning. *Advances in neural information processing systems (NeurIPS)*, 33:16628–16638.

Xu, Y. and Zeevi, A. (2020). Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*.

Zanette, A., Cheng, C.-A., and Agarwal, A. (2021). Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory (COLT)*.

Zhang, Z., Yang, J., Ji, X., and Du, S. S. (2021). Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:4342–4355.

Zhong, H., Yang, Z., Wang, Z., and Szepesvári, C. (2021). Optimistic policy optimization is provably efficient in non-stationary mdps. *CoRR*, abs/2110.08984.

Zhou, H., Chen, J., Varshney, L. R., and Jagmohan, A. (2020). Nonstationary reinforcement learning with linear function approximation. *CoRR*.

Zimin, A. and Neu, G. (2013). Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials for Paper Submission: Sample Complexity Characterization for Linear Contextual MDPs

Notation. Recall that we use $\mathcal{M}(w) = (\mathcal{S}, \mathcal{A}, P_w, r_w, H)$ to denote the environment MDP w.r.t. the context w. We use $\mathcal{M}^{(r',P')}(w)$ to denote an MDP with a transition kernel P'_w and a reward function r'_w, i.e., $\mathcal{M}^{(r',P')}(w) = (\mathcal{S}, \mathcal{A}, P'_w, r'_w, H)$. Here we define a truncated value function for any MDP $\mathcal{M}$ under a policy π and at step h as:

$$\bar{Q}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}(s_h, a_h) = \min\left\{H, r'_h(s_h, a_h, w) + P'_{h,w}\bar{V}^{\pi}_{h+1,\mathcal{M}^{(r',P')}(w)}(s_h, a_h)\right\},$$
$$\bar{V}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}(s_h) = \mathop{\mathbb{E}}_{\pi}\left[\bar{Q}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}(s_h, a_h)\right]. \tag{19}$$

For any two probability measures P and Q, we use $\|P - Q\|_{TV}$ to denote the total variation distance between them.

A Proof of Theorem 1

In this section, we first prove some useful lemmas and then prove Theorem 1.

A.1 Supporting Lemmas

We consider the model defined in Definition 1. We first introduce the following MLE guarantee on the estimation error established in Agarwal et al. (2020).

Lemma 1. *(MLE guarantee). Suppose Assumption 1 holds. Given $\delta \in (0,1)$, we have the following inequality holds with probability at least $1 - \delta/2$ for all $h \in [H]$ and $n \in [N]$:*

$$\sum_{\tau=1}^{n-1} \mathop{\mathbb{E}}_{\substack{w\sim q \\ (s_h,a_h)\sim(P_w,\pi_w^\tau)}} \|\langle\hat{\mu}_h^n(\cdot) - \mu_h(\cdot), \phi_h(s_h, a_h, w)\rangle\|_{TV}^2 \le \zeta_n,$$

where $\zeta_n := \log(2|\Psi_1|nH/\delta)$.

The following lemma can be obtained from Lemma 39 in Zanette et al. (2021) and Lemma 11 in Uehara et al. (2021).

Lemma 2. *For $\hat{\Sigma}_h^n$ and $\hat{\Lambda}_h^n$ defined in eq. (6), we define their corresponding expected value as follows:*

$$\Sigma_h^n = \sum_{\tau=1}^{n-1} \mathop{\mathbb{E}}_{\substack{w\sim q \\ (s_h,a_h)\sim(P_w,\pi_w^\tau)}} \phi_h(s_h, a_h, w)\phi_h(s_h, a_h, w)^\top + \lambda_n I,$$

$$\Lambda_h^n = \sum_{\tau=1}^{n-1} \mathop{\mathbb{E}}_{\substack{w\sim q \\ (s_h,a_h)\sim(P_w,\pi_w^\tau)}} \phi_h(s_h, a_h, w)\phi_h(s_h, a_h, w)^\top + \xi_n I,$$

where $\lambda_n = \gamma_1 d\log(2nH/\delta)$ and $\xi_n = \gamma_2 d\log(2nH/\delta)$. We define the following two events:

$$\mathcal{E}_1 = \Big\{\forall n \in [N], h \in [H], s \in \mathcal{S}, a \in \mathcal{A}, w \in \mathcal{W}$$
$$\frac{1}{5}\|\phi_h(s,a,w)\|_{(\Sigma_h^n)^{-1}} \le \|\phi_h(s,a,w)\|_{(\hat{\Sigma}_h^n)^{-1}} \le 3\|\phi_h(s,a,w)\|_{(\Sigma_h^n)^{-1}}\Big\};$$

$$\mathcal{E}_2 = \Big\{ \forall n \in [N], h \in [H], s \in \mathcal{S}, a \in \mathcal{A}, w \in \mathcal{W} \\ \frac{1}{5} \left\|\psi_h(s,a,w)\right\|_{\left(\Lambda_h^n\right)^{-1}} \le \left\|\psi_h(s,a,w)\right\|_{\left(\hat{\Lambda}_h^n\right)^{-1}} \le 3 \left\|\psi_h(s,a,w)\right\|_{\left(\Lambda_h^n\right)^{-1}} \Big\}.$$

Let $\mathcal{E}_0 := \mathcal{E}_1 \cup \mathcal{E}_2$ denote the intersection of the two events. Then $\mathbb{P}(\mathcal{E}_0) \ge 1-\delta$.

We further prove a number of supporting lemmas.

Lemma 3. *Suppose $\hat{P}_h^n(\cdot|s,a,w) = \langle \hat{\mu}_h^n(\cdot), \phi_h(s,a,w)\rangle$ is the estimated context-varying linear transition kernel at step $h \in [H]$ in episode $n \in [N]$. Consider a generic non-negative function $f : \mathcal{S} \to \mathbb{R}$, which is bounded by B, i.e., $f(s) \in [0,B]$ for any $s \in \mathcal{S}$. Then for any $(s,a,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$, with probability at least $1-\delta/2$, we have:*

$$\left| \int_{\mathcal{S}} f\left(s'\right) \left(\hat{\mu}_h^n\left(s'\right) - \mu_h\left(s'\right)\right)^\top \phi_h(s,a,w) ds' \right| \le \min\left\{ \hat{\alpha}_n \left\|\phi_h(s,a,w)\right\|_{\left(\Sigma_h^n\right)^{-1}}, B \right\},$$

where $\hat{\alpha}_n = B\sqrt{2\lambda_n d + 4\zeta_n}$.

Proof. First, we have:

$$\begin{aligned} &\left| \int_{\mathcal{S}} f\left(s'\right) \left(\hat{\mu}_h^n\left(s'\right) - \mu_h\left(s'\right)\right)^\top \phi_h(s,a,w) ds' \right| \\ &\quad \overset{(i)}{\le} \left\|\phi_h(s,a,w)\right\|_{\left(\Sigma_h^n\right)^{-1}} \cdot \left\| \int_{\mathcal{S}} f(s') \left(\hat{\mu}_h^n\left(s'\right) - \mu_h(s')\right) ds' \right\|_{\left(\Sigma_h^n\right)}, \end{aligned} \tag{20}$$

where (i) follows from the Cauchy-Schwarz inequality. Then, we further derive

$$\begin{aligned} &\left\| \int_{\mathcal{S}} f(s) \left(\hat{\mu}_h^n(s) - \mu_h(s)\right) ds \right\|_{\left(\Sigma_h^n\right)}^2 \\ &\quad = \lambda_n \cdot \left\| \int_{\mathcal{S}} f(s) (\hat{\mu}_h^n(s) - \mu_h(s)) ds \right\|^2 \\ &\qquad + \sum_{\tau=1}^{n-1} \mathop{\mathbb{E}}_{\substack{w\sim q \\ (s_h,a_h)\sim(P_w,\pi_w^\tau)}} \left(\int_{\mathcal{S}} f(s) \left(\hat{\mu}_h^n(s) - \mu_h(s)\right)^\top \phi_h\left(s_h,a_h,w\right) ds \right)^2 \\ &\quad \overset{(i)}{\le} 2\lambda_n d B^2 + 4B^2 \sum_{\tau=1}^{n-1} \mathop{\mathbb{E}}_{\substack{w\sim q \\ (s_h,a_h)\sim(P_w,\pi_w^\tau)}} \left\| \langle \hat{\mu}_h^n(\cdot) - \mu_h(\cdot), \phi_h\left(s_h,a_h,w\right)\rangle \right\|_{TV}^2 \\ &\quad \overset{(ii)}{\le} 2\lambda_n d B^2 + 4B^2 \zeta_n, \end{aligned}$$

where (i) follows from Definition 1 and from the definition of total variation distance, and (ii) follows from Lemma 1. By substituting the above equation into eq. (20) and setting $\hat{\alpha}_n = B\sqrt{2\lambda_n d + 4\zeta_n}$, we have:

$$\left| \int_{\mathcal{S}} f\left(s'\right) \left(\hat{\mu}_h^n\left(s'\right) - \mu_h\left(s'\right)\right)^\top \phi_h(s,a,w) ds' \right| \le \hat{\alpha}_n \left\|\phi_h(s,a,w)\right\|_{\left(\Sigma_h^n\right)^{-1}}. \tag{21}$$

Also, since $f(s) \in [0,B]$ for any $s \in \mathcal{S}$, we have:

$$\begin{aligned} &\left| \int_{\mathcal{S}} f\left(s'\right) \left(\hat{\mu}_h^n\left(s'\right) - \mu_h\left(s'\right)\right)^\top \phi_h(s,a,w) ds' \right| \\ &\quad = \left| \mathop{\mathbb{E}}_{s'\sim\hat{P}_h^n(\cdot|s,a,w)} f(s') - \mathop{\mathbb{E}}_{s'\sim P_h(\cdot|s,a,w)} f(s') \right| \\ &\quad \le B. \end{aligned} \tag{22}$$

By combining eq. (21) and eq. (22), we complete the proof. □

Lemma 4. *Given the event $\mathcal{E}_0$ defined in Lemma 2 occurs, for any $(s,a,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$, define the function $\hat{b}_h^n(s,a,w) = \min\{\alpha_n \|\phi_h(s,a,w)\|_{(\hat{\Sigma}_h^n)^{-1}}, H\}$, where $\alpha_n = 5H\sqrt{2\lambda_n d + 4\zeta_n}$. Then for any context w, with probability at least $1-\delta/2$, we have:*

$$\left|V^{\pi_w}_{\mathcal{M}(w)} - V^{\pi_w}_{\mathcal{M}^{(r,\hat{P}^n)}(w)}\right| \le \bar{V}^{\pi_w}_{\mathcal{M}^{(\hat{b}^n,\hat{P}^n)}(w)}.$$

Proof. Recall the definitions of the truncated value functions $\bar{V}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}$ and $\bar{Q}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}$ for a generic MDP $\mathcal{M}^{(r',P')}(w) = (\mathcal{S},\mathcal{A},P'_w,r'_w,H)$ are given by

$$\bar{Q}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}(s_h,a_h) = \min\{H, r'(s_h,a_h,w) + P'_{h,w}\bar{V}^{\pi}_{h+1,\mathcal{M}^{(r',P')}(w)}(s_h,a_h)\},$$
$$\bar{V}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}(s_h) = \mathop{\mathbb{E}}_{\pi}\left[\bar{Q}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}(s_h,a_h)\right].$$

We complete the proof by induction. For the base case $h = H+1$, we have $\left|V^{\pi_w}_{H+1,\mathcal{M}(w)}(s_{H+1}) - V^{\pi_w}_{H+1,\mathcal{M}^{(r,\hat{P}^n)}(w)}(s_{H+1})\right| = 0 = \bar{V}^{\pi_w}_{H+1,\mathcal{M}^{(\hat{b}^n,\hat{P}^n)}(w)}(s_{H+1})$.
Now we assume that $\left|V^{\pi_w}_{h+1,\mathcal{M}(w)}(s_{h+1}) - V^{\pi_w}_{h+1,\mathcal{M}^{(r,\hat{P}^n)}(w)}(s_{h+1})\right| \le \bar{V}^{\pi_w}_{h+1,\mathcal{M}^{(\hat{b}^n,\hat{P}^n)}(w)}(s_{h+1})$ holds for all $s_{h+1}\in\mathcal{S}$. Then according to Bellman equation, for all s_h, a_h, we have:

$$\begin{aligned}
&\left|Q^{\pi_w}_{h,\mathcal{M}^{(r,\hat{P}^n)}(w)}(s_h,a_h) - Q^{\pi_w}_{h,\mathcal{M}(w)}(s_h,a_h)\right| \\
&\quad= \left|\hat{P}^n_{h,w}V^{\pi_w}_{h+1,\mathcal{M}^{(r,\hat{P}^n)}(w)}(s_h,a_h) - P_{h,w}V^{\pi_w}_{h+1,\mathcal{M}(w)}(s_h,a_h)\right| \\
&\quad= \left|\hat{P}^n_{h,w}\left(V^{\pi_w}_{h+1,\mathcal{M}^{(r,\hat{P}^n)}(w)} - V^{\pi_w}_{h+1,\mathcal{M}(w)}\right)(s_h,a_h)\right. \\
&\qquad\qquad\left. + \left(\hat{P}^n_{h,w} - P_{h,w}\right)V^{\pi_w}_{h+1,\mathcal{M}(w)}(s_h,a_h)\right| \\
&\quad\overset{(i)}{\le} \min\left\{H, \hat{b}^n_h(s_h,a_h,w) + \hat{P}^n_{h,w}\left|V^{\pi_w}_{h+1,\mathcal{M}^{(r,\hat{P}^n)}(w)} - V^{\pi_w}_{h+1,\mathcal{M}(w)}\right|(s_h,a_h)\right\} \\
&\quad\overset{(ii)}{\le} \min\left\{H, \hat{b}^n_h(s_h,a_h,w) + \hat{P}^n_{h,w}\bar{V}^{\pi_w}_{h+1,\mathcal{M}^{(\hat{b}^n,\hat{P}^n)}(w)}(s_h,a_h)\right\} \\
&\quad= \bar{Q}^{\pi_w}_{h,\mathcal{M}^{(\hat{b}^n,\hat{P}^n)}(w)}(s_h,a_h),
\end{aligned} \tag{23}$$

where (i) follows from the fact that $\left|Q^{\pi_w}_{h,\mathcal{M}^{(r,\hat{P}^n)}(w)}(s_h,a_h) - Q^{\pi_w}_{h,\mathcal{M}(w)}(s_h,a_h)\right|$ is bounded by H and from Lemma 3, and (ii) follows from the recursive hypothesis. Then, we have:

$$\begin{aligned}
&\left|V^{\pi_w}_{h,\mathcal{M}(w)}(s_h) - V^{\pi_w}_{h,\mathcal{M}^{(r,\hat{P}^n)}(w)}(s_h)\right| \\
&\quad= \left|\mathop{\mathbb{E}}_{\pi_w}\left[Q^{\pi_w}_{h,\mathcal{M}(w)}(s_h,a_h)\right] - \mathop{\mathbb{E}}_{\pi_w}\left[Q^{\pi_w}_{h,\mathcal{M}^{(r,\hat{P}^n)}(w)}(s_h,a_h)\right]\right| \\
&\quad\le \mathop{\mathbb{E}}_{\pi_w}\left[\left|Q^{\pi_w}_{h,\mathcal{M}(w)}(s_h,a_h) - Q^{\pi_w}_{h,\mathcal{M}^{(r,\hat{P}^n)}(w)}(s_h,a_h)\right|\right] \\
&\quad\overset{(i)}{\le} \mathop{\mathbb{E}}_{\pi_w}\left[\bar{Q}^{\pi_w}_{h,\mathcal{M}^{(\hat{b}^n,\hat{P}^n)}(w)}(s_h,a_h)\right] \\
&\quad= \bar{V}^{\pi_w}_{h,\mathcal{M}^{(\hat{b}^n,\hat{P}^n)}(w)}(s_h),
\end{aligned}$$

where (i) follows from eq. (23). Therefore, by induction, we conclude that:

$$\left|V^{\pi_w}_{\mathcal{M}(w)} - V^{\pi_w}_{\mathcal{M}^{(r,\hat{P}^n)}(w)}\right| \le \bar{V}^{\pi_w}_{\mathcal{M}^{(\hat{b}^n,\hat{P}^n)}(w)}.$$

□

Lemma 5. *Suppose $\hat{f}_h^n(s,a,w)$ is the estimated reward defined in eq. (10), then for any $(s,a,w)\in\mathcal{S}\times\mathcal{A}\times\mathcal{W}$, we have:*

$$\left|\hat{f}^n_h(s,a,w) - r_h(s,a,w)\right| \le \min\left\{\beta_n \|\psi_h(s,a,w)\|_{(\hat{\Lambda}^n_h)^{-1}}, 1\right\},$$

where $\beta_n = \sqrt{\xi_n d}$.

Proof. Since the values of both $r_h(s,a,w)$ and $\hat{f}_h^n(s,a,w)$ are restricted to $[0,1]$ for any (s,a,w) and any $h \in [H], n \in [N]$, we conclude that $\left|\hat{f}_h^n(s,a,w) - r_h(s,a,w)\right| \leq 1$. Also, according to the definition of $\hat{f}_h^n(s,a,w)$ defined in eq. (10), we have:

$$\left|\hat{f}_h^n(s,a,w) - r_h(s,a,w)\right| \leq \left|\langle \hat{\eta}_h^n, \psi_h(s,a,w)\rangle - r_h(s,a,w)\right|. \tag{24}$$

Now consider:

$$\begin{aligned}
\hat{\eta}_h^n - \eta_h &= \left(\hat{\Lambda}_h^n\right)^{-1} \sum_{\tau=1}^{n-1} \psi_h\left(s_h^\tau, a_h^\tau, w_\tau\right) r_h^\tau - \eta_h \\
&= \left(\hat{\Lambda}_h^n\right)^{-1} \left(\sum_{\tau=1}^{n-1} \psi_h\left(s_h^\tau, a_h^\tau, w_\tau\right) r_h^\tau - \hat{\Lambda}_h^n \cdot \eta_h\right) \\
&\overset{(i)}{=} -\xi_n \left(\hat{\Lambda}_h^n\right)^{-1} \eta_h,
\end{aligned}$$

where (i) follows from the definition of the matrix $\hat{\Lambda}_h^n$ and the reward function $r_h^\tau = \psi_h\left(s_h^\tau, a_h^\tau, w_\tau\right)^\top \eta_h$. Due to the linear structure, for any $(s,a,w) \in \mathcal{S} \times \mathcal{A} \times \mathcal{W}$, we have:

$$\begin{aligned}
\left|\langle \hat{\eta}_h^n, \psi_h(s,a,w)\rangle - r_h(s,a,w)\right| &= \left|\langle \hat{\eta}_h^n - \eta_h, \psi_h(s,a,w)\rangle\right| \\
&= \left\langle \left(I - \xi_n \left(\hat{\Lambda}_h^n\right)^{-1}\right) \eta_h - \eta_h, \psi_h(s,a,w) \right\rangle \\
&= \left| \left\langle -\xi_n \left(\hat{\Lambda}_h^n\right)^{-1} \eta_h, \psi_h(s,a,w) \right\rangle \right| \\
&\leq \sqrt{\xi_n} \left\|\eta_h\right\| \cdot \left\|\psi_h(s,a,w)\right\|_{\left(\hat{\Lambda}_h^n\right)^{-1}} \\
&\overset{(i)}{\leq} \sqrt{\xi_n d} \left\|\psi_h(s,a,w)\right\|_{\left(\hat{\Lambda}_h^n\right)^{-1}},
\end{aligned} \tag{25}$$

where (i) follows from the normalization in Definition 1. By combining eqs. (24) and (25), and the fact that $\left|\hat{f}_h^n(s,a,w) - r_h(s,a,w)\right| \leq 1$. Then we complete the proof. □

Lemma 6. *Define function* $\hat{c}_h^n(s,a,w) = \min\left\{\beta_n \left\|\psi_h(s,a,w)\right\|_{\left(\hat{\Lambda}_h^n\right)^{-1}}, 1\right\}$. *Assume that the event* $\mathcal{E}_0$ *defined in Lemma 2 occurs. Then for any context* w*, we have:*

$$\left|\bar{V}^{\pi_w}_{\mathcal{M}(\hat{f}^n, \hat{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(r, \hat{P}^n)(w)}\right| \leq \bar{V}^{\pi_w}_{\mathcal{M}(\hat{c}^n, \hat{P}^n)(w)}.$$

Proof. Note that the values of both $\hat{f}_h^n(s,a,w)$ and $\hat{c}_h^n(s,a,w)$ are restricted to $[0,1]$ for any (s,a,w) and any $h \in [H], n \in [N]$. Following from Lemma 19, it is equivalent to prove:

$$\left|V^{\pi_w}_{\mathcal{M}(\hat{f}^n, \hat{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(r, \hat{P}^n)(w)}\right| \leq V^{\pi_w}_{\mathcal{M}(\hat{c}^n, \hat{P}^n)(w)}.$$

Then we have:

$$\begin{aligned}
\left|V^{\pi_w}_{\mathcal{M}(\hat{f}^n, \hat{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(r, \hat{P}^n)(w)}\right| &\overset{(i)}{=} \left|\sum_{h=1}^{H} \mathop{\mathbb{E}}_{(s_h,a_h)\sim\left(\hat{P}_w^n, \pi_w\right)} \left(\hat{f}_h^n\left(s_h, a_h, w\right) - r_h\left(s_h, a_h, w\right)\right)\right| \\
&\leq \sum_{h=1}^{H} \mathop{\mathbb{E}}_{(s_h,a_h)\sim\left(\hat{P}_w^n, \pi_w\right)} \left|\hat{f}_h^n\left(s_h, a_h, w\right) - r_h\left(s_h, a_h, w\right)\right| \\
&\overset{(ii)}{\leq} \sum_{h=1}^{H} \mathop{\mathbb{E}}_{(s_h,a_h)\sim\left(\hat{P}_w^n, \pi_w\right)} \hat{c}_h^n(s_h, a_h, w) \\
&= V^{\pi_w}_{\mathcal{M}(\hat{c}^n, \hat{P}^n)(w)},
\end{aligned} \tag{26}$$

where (i) follows from Lemma 23 and (ii) follows from Lemma 5. □

Lemma 7. *Given the event $\mathcal{E}_0$ occurs, then for any context-dependent policy π_w with probability at least $1-\delta/2$, we have*

$$\left|\bar{V}^{\pi_w}_{\mathcal{M}(\hat{f}^n,\hat{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(w)}\right| \leq \bar{V}^{\pi_w}_{\mathcal{M}(\hat{b}^n,\hat{P}^n)(w)} + \bar{V}^{\pi_w}_{\mathcal{M}(\hat{c}^n,\hat{P}^n)(w)}$$

Proof. By combining the bounds on the estimation error of both the reward and the transition kernel, characterized respectively in Lemma 6 and Lemma 4, we have:

$$\begin{aligned}\left|\bar{V}^{\pi_w}_{\mathcal{M}(\hat{f}^n,\hat{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(w)}\right| &= \left|\bar{V}^{\pi_w}_{\mathcal{M}(\hat{f}^n,\hat{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(r,\hat{P}^n)(w)} + V^{\pi_w}_{\mathcal{M}(r,\hat{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(w)}\right| \\ &\leq \left|\bar{V}^{\pi_w}_{\mathcal{M}(\hat{f}^n,\hat{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(r,\hat{P}^n)(w)}\right| + \left|V^{\pi_w}_{\mathcal{M}(r,\hat{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(w)}\right| \\ &\leq \bar{V}^{\pi_w}_{\mathcal{M}(\hat{b}^n,\hat{P}^n)(w)} + \bar{V}^{\pi_w}_{\mathcal{M}(\hat{c}^n,\hat{P}^n)(w)}.\end{aligned}$$

□

A.2 Proof of Theorem 1

We first restate Theorem 1 below.

Theorem 3 (Restatement of Theorem 1)**.** *Consider a CMDP with varying representations as defined in Definition 1. Under Assumption 1, for any $\delta \in (0,1)$, with probability at least $1-3\delta/2$, the sequence of policies $\{\pi^n_{w_n}\}_{n=1}^N$ generated by Algorithm 1 satisfies that*

$$\begin{aligned}&\frac{1}{N}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\left[V^{\pi^*_w}_{\mathcal{M}(w)} - V^{\pi^n_w}_{\mathcal{M}(w)}\right] \\ &\quad\leq \left(42H\sqrt{2d\lambda_N + 4\log(2HN|\Psi_1|/\delta)} + 6\sqrt{d\xi_N}\right) H\sqrt{\frac{2d}{N}}\cdot\sqrt{\log\left(1+\frac{N}{d\lambda}\right)},\end{aligned}$$

where $\lambda_n = \gamma_1 d\log(2nH/\delta)$, $\xi_n = \gamma_2 d\log(2nH/\delta)$, $\gamma_1,\gamma_2 = \mathcal{O}(1)$ and $\lambda = \min\{\lambda_1,\xi_1\}$. To achieve an ϵ average sub-optimality gap, at most $\mathcal{O}\left(\frac{H^4d^3\log(|\Psi_1|/\delta)}{\epsilon^2}\right)$ episodes are needed.

Proof. First, we derive an optimistic estimation of the optimal value function.

$$\begin{aligned}V^{\pi^*_w}_{\mathcal{M}(w)} &\overset{(i)}{\leq} \bar{V}^{\pi^*_w}_{\mathcal{M}(\hat{f}^n,\hat{P}^n)(w)} + \bar{V}^{\pi^*_w}_{\mathcal{M}(\hat{b}^n,\hat{P}^n)(w)} + \bar{V}^{\pi^*_w}_{\mathcal{M}(\hat{c}^n,\hat{P}^n)(w)} \\ &\overset{(ii)}{\leq} \bar{\bar{V}}^{\pi^*_w}_{\mathcal{M}(\hat{f}^n+\hat{b}^n+\hat{c}^n,\hat{P}^n)(w)} \\ &\overset{(iii)}{\leq} \bar{\bar{V}}^{\pi^n_w}_{\mathcal{M}(\hat{f}^n+\hat{b}^n+\hat{c}^n,\hat{P}^n)(w)} \\ &\overset{(iv)}{\leq} \bar{\bar{V}}^{\pi^n_w}_{\mathcal{M}(\hat{f}^n,\hat{P}^n)(w)} + \bar{\bar{V}}^{\pi^n_w}_{\mathcal{M}(\hat{b}^n,\hat{P}^n)(w)} + \bar{\bar{V}}^{\pi^n_w}_{\mathcal{M}(\hat{c}^n,\hat{P}^n)(w)},\end{aligned} \tag{27}$$

where (i) follows from Lemma 7, (ii) follows from Lemma 20, (iii) follows from the definition of the greed policy $\pi^n_w := \text{argmax}_\pi \bar{\bar{V}}^{\pi}_{\mathcal{M}(\hat{f}^n+\hat{b}^n+\hat{c}^n,\hat{P}^n)(w)}$ and (iv) follows from Lemma 21, Then the average suboptimality gap can be

bounded by:

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\left[V^{\pi_w^*}_{\mathcal{M}(w)}-V^{\pi_w^n}_{\mathcal{M}(w)}\right]\\
\leq&\frac{1}{N}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\left[\bar{\bar{V}}^{\pi_w^n}_{\mathcal{M}(\hat{f}^n,\hat{P}^n)(w)}+\bar{\bar{V}}^{\pi_w^n}_{\mathcal{M}(\hat{b}^n,\hat{P}^n)(w)}+\bar{\bar{V}}^{\pi_w^n}_{\mathcal{M}(\hat{c}^n,\hat{P}^n)(w)}-V^{\pi_w^n}_{\mathcal{M}(w)}\right]\\
\leq&\frac{1}{N}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\left[\left|\bar{\bar{V}}^{\pi_w^n}_{\mathcal{M}(\hat{f}^n,\hat{P}^n)(w)}-V^{\pi_w^n}_{\mathcal{M}(w)}\right|+\bar{\bar{V}}^{\pi_w^n}_{\mathcal{M}(\hat{b}^n,\hat{P}^n)(w)}+\bar{\bar{V}}^{\pi_w^n}_{\mathcal{M}(\hat{c}^n,\hat{P}^n)(w)}\right]\\
\overset{(i)}{\leq}&\frac{1}{N}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\left[2\bar{\bar{V}}^{\pi_w^n}_{\mathcal{M}(\hat{b}^n,\hat{P}^n)(w)}+2\bar{\bar{V}}^{\pi_w^n}_{\mathcal{M}(\hat{c}^n,\hat{P}^n)(w)}\right],
\end{aligned}\tag{28}$$

where (i) follows from Lemma 7.

We next provide an upper bound on $\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\bar{\bar{V}}^{\pi_w^n}_{\mathcal{M}(\hat{b}^n,\hat{P}^n)(w)}$. Define $g_h^n(s,a,w):=(\hat{P}^n_{h,w}-P_{h,w})\bar{\bar{V}}^{\pi_w^n}_{h+1,\mathcal{M}(\hat{b}^n,\hat{P}^n)(w)}(s,a)$. Then due to Lemma 22 in Appendix D, we have:

$$\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\bar{\bar{V}}^{\pi_w^n}_{\mathcal{M}(\hat{b}^n,\hat{P}^n)(w)}\leq\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}V^{\pi_w^n}_{\mathcal{M}(\hat{b}^n,P)(w)}+\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}V^{\pi_w^n}_{\mathcal{M}(g^n,P)(w)}.\tag{29}$$

First, we bound the first term in the right-hand-side of eq. (29). By applying Lemma 24 in Appendix D, we can obtain a bound on the summation of the expected value function $V^{\pi_w}_{\mathcal{M}(\hat{b}^n,P)(w)}$ as follows:

$$\begin{aligned}
\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}V^{\pi_w^n}_{\mathcal{M}(\hat{b}^n,P)(w)}=&\sum_{n=1}^{N}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{\substack{w\sim q\\(s_h,a_h)\sim(P_w,\pi_w^n)}}\left[\alpha_n\left\|\phi_h(s_h,a_h,w)\right\|_{(\hat{\Sigma}_h^n)^{-1}}\right]\\
\overset{(i)}{\leq}&3\alpha_N\sqrt{N}\sum_{h=1}^{H}\sqrt{\sum_{n=1}^{N}\mathop{\mathbb{E}}_{\substack{w\sim q\\(s_h,a_h)\sim(P_w,\pi_w^n)}}\left[\left\|\phi_h(s_h,a_h,w)\right\|^2_{(\Sigma_h^n)^{-1}}\right]}\\
\overset{(ii)}{\leq}&3\alpha_N\sqrt{N}H\cdot\sqrt{2d\log\left(1+\frac{N}{d\lambda}\right)},
\end{aligned}\tag{30}$$

where (i) follows from the Cauchy-Schwarz inequality and because the event $\mathcal{E}_0$ occurs, and (ii) follows from Lemma 24.

Next, we bound the second term in the right-hand-side of eq. (29). We obtain the bound on the summation of the expected value function $V^{\pi_w}_{\mathcal{M}(g^n,P)(w)}$ as follows:

$$\begin{aligned}
\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}V^{\pi_w^n}_{\mathcal{M}(g^n,P)(w)}\leq&\sum_{n=1}^{N}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{\substack{w\sim q\\(s_h,a_h)\sim(P_w,\pi_w^n)}}\left|g_h^n(s_h,a_h,w)\right|\\
\overset{(i)}{\leq}&\sum_{n=1}^{N}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{\substack{w\sim q\\(s_h,a_h)\sim(P_w,\pi_w^n)}}\left[\frac{3}{5}\alpha_n\left\|\phi_h(s_h,a_h,w)\right\|_{(\Sigma_h^n)^{-1}}\right]\\
\overset{(ii)}{\leq}&\frac{3}{5}\alpha_N\sqrt{N}\sum_{h=1}^{H}\sqrt{\sum_{n=1}^{N}\mathop{\mathbb{E}}_{\substack{w\sim q\\(s_h,a_h)\sim(P_w,\pi_w^n)}}\left[\left\|\phi_h(s_h,a_h,w)\right\|^2_{(\Sigma_h)^{-1}}\right]}\\
\overset{(iii)}{\leq}&\frac{3}{5}\alpha_N\sqrt{N}H\cdot\sqrt{2d\log\left(1+\frac{N}{d\lambda}\right)},
\end{aligned}\tag{31}$$

where (i) follows from Lemma 3 and the fact that $\bar{\bar{V}}^{\pi_w}_{h,\mathcal{M}(\hat{b}^n,\hat{P}^n)(w)}(s_h,a_h)\leq 3H$ for any $h\in[H]$, (ii) follows from the Cauchy-Schwarz inequality, and (iii) follows from Lemma 24. Then by combining eqs. (29) to (31) together

with the definition of α_n we obtain:

$$\sum_{n=1}^{N} \mathop{\mathbb{E}}_{w\sim q} \bar{V}^{\pi_w^n}_{\mathcal{M}(\hat{b}^n,\hat{P}^n)(w)} \leq \frac{18}{5}\alpha_N\sqrt{N}H \cdot \sqrt{2d\log\left(1+\frac{N}{d\lambda}\right)}. \tag{32}$$

Now, we provide an upper bound on $\sum_{n=1}^{N} \mathop{\mathbb{E}}_{w\sim q} \bar{V}^{\pi_w^n}_{\mathcal{M}(\hat{c}^n,\hat{P}^n)(w)}$. Due to Lemma 22 in Appendix D, we can show that

$$\sum_{n=1}^{N} \mathop{\mathbb{E}}_{w\sim q} \bar{V}^{\pi_w^n}_{\mathcal{M}(\hat{c}^n,\hat{P}^n)(w)} \leq \sum_{n=1}^{N} \mathop{\mathbb{E}}_{w\sim q} V^{\pi_w^n}_{\mathcal{M}(\hat{c}^n,P)(w)} + \sum_{n=1}^{N} \mathop{\mathbb{E}}_{w\sim q} V^{\pi_w^n}_{\mathcal{M}(l^n,P)(w)}, \tag{33}$$

where $l_h^n(s,a,w) := (\hat{P}^n_{h,w} - P_{h,w})\bar{V}^{\pi_w^n}_{h+1,\mathcal{M}(\hat{c}^n,\hat{P}^n)(w)}(s,a)$. We first provide an upper bound on the first term in the right-hand-side of eq. (33):

$$\begin{aligned}
\sum_{n=1}^{N} \mathop{\mathbb{E}}_{w\sim q} V^{\pi_w^n}_{\mathcal{M}(\hat{c}^n,P)(w)} &= \sum_{n=1}^{N}\sum_{h=1}^{H} \mathop{\mathbb{E}}_{\substack{w\sim q\\ (s_h,a_h)\sim(P_w,\pi_w^n)}} \left[\beta_n \left\|\psi_h(s_h,a_h,w)\right\|_{(\hat{\Lambda}_h^n)^{-1}}\right] \\
&\overset{(i)}{\leq} 3\beta_N\sqrt{N}\sum_{h=1}^{H}\sqrt{\sum_{n=1}^{N} \mathop{\mathbb{E}}_{\substack{w\sim q\\ (s_h,a_h)\sim(P_w,\pi_w^n)}} \left[\left\|\psi_h(s_h,a_h,w)\right\|^2_{(\Lambda_h^n)^{-1}}\right]} \\
&\overset{(ii)}{\leq} 3\beta_N\sqrt{N}H \cdot \sqrt{2d\log\left(1+\frac{N}{d\lambda}\right)},
\end{aligned} \tag{34}$$

where (i) follows from the Cauchy-Schwarz inequality and the event $\mathcal{E}_0$ occurs, and (ii) follows from Lemma 24. Then, since $\bar{V}^{\pi_w}_{h,\mathcal{M}(\hat{c}^n,\hat{P}^n)(w)}(s,a) \leq 3H$ for any $h \in [H]$, we bound the second term in the right-hand-side of eq. (33) similarly to eq. (31) and obtain:

$$\sum_{n=1}^{N} \mathop{\mathbb{E}}_{w\sim q} V^{\pi_w^n}_{\mathcal{M}(l^n,P)(w)} \leq \frac{3}{5}\alpha_N\sqrt{N}H \cdot \sqrt{2d\log\left(1+\frac{N}{d\lambda}\right)}. \tag{35}$$

Combining eqs. (33) to (35) together with the definitions of α_n and β_n, we obtain:

$$\sum_{n=1}^{N} \mathop{\mathbb{E}}_{w\sim q} \bar{V}^{\pi_w^n}_{\mathcal{M}(\hat{c}^n,\hat{P}^n)(w)} \leq \left(\frac{3}{5}\alpha_N + 3\beta_N\right)\sqrt{N}H \cdot \sqrt{2d\log\left(1+\frac{N}{d\lambda}\right)}. \tag{36}$$

By substituting eqs. (32) and (36) into eq. (28), we can bound the average suboptimality gap as follows:

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N} \mathop{\mathbb{E}}_{w\sim q}\left[V^{\pi_w^*}_{\mathcal{M}(w)} - V^{\pi_w^n}_{\mathcal{M}(w)}\right] \\
&\quad\leq \frac{1}{N}\left(\frac{42}{5}\alpha_N + 6\beta_N\right)\sqrt{N}H \cdot \sqrt{2d\log\left(1+\frac{N}{d\lambda}\right)} \\
&\quad= \left(42H\sqrt{2d\lambda_N + 4\log(2HN|\Psi_1|/\delta)} + 6\sqrt{d\xi_N}\right)H\sqrt{\frac{2d}{N}} \cdot \sqrt{\log\left(1+\frac{N}{d\lambda}\right)},
\end{aligned}$$

where $\lambda_n = \gamma_1 d\log(2nH/\delta)$, $\xi_n = \gamma_2 d\log(2nH/\delta)$ and $\lambda = \min\{\lambda_1, \xi_1\}$.

Note that $\lambda_N, \xi_N = \mathcal{O}(d)$, and it can be seen that the upper bound of the average suboptimality gap is of the order $\mathcal{O}\left(\sqrt{\frac{H^4 d^3 \log(|\Psi_1|/\delta)}{N}}\right)$. Then to achieve an ϵ average sub-optimality gap, at most $\mathcal{O}\left(\frac{H^4 d^3 \log(|\Psi_1|/\delta)}{\epsilon^2}\right)$ episodes are needed. This completes the proof. □

B Proof of Theorem 2.

In this section, we first prove some useful lemmas and then prove Theorem 2.

B.1 Supporting Lemmas

We consider the model defined in Definition 2. We first introduce the following MLE guarantee on the estimation error established in Agarwal et al. (2020). Note that the form of the MLE guarantee is different from that in Lemma 1 because the model is different.

Lemma 8. *(MLE guarantee). Suppose Assumption 2 holds. Given $\delta \in (0,1)$, the following inequality holds with probability at least $1-\delta/2$ for all $h \in [H]$ and $n \in [N]$:*

$$\sum_{\tau=1}^{n-1} \mathop{\mathbb{E}}_{\substack{w\sim q \\ s_h\sim(P_w,\pi_w^\tau) \\ a_h\sim\mathcal{U}(\mathcal{A})}} \|\langle \widetilde{\mu}_h(\cdot,w)-\mu_h(\cdot,w), \phi_h(s_h,a_h)\rangle\|_{TV}^2 \le \zeta_n,$$

where $\zeta_n := \log(2|\Psi_2|nH/\delta)$.

Lemma 9. *(LSR guarantee). Suppose Assumption 2 holds. Given $\delta \in (0,1)$, the following inequality holds with probability at least $1-\delta/2$ for all $h \in [H]$ and $n \in [N]$:*

$$\sum_{\tau=1}^{n-1} \mathop{\mathbb{E}}_{\substack{w\sim q \\ s_h\sim(P_w,\pi_w^\tau) \\ a_h\sim\mathcal{U}(\mathcal{A})}} \|\langle \widetilde{\eta}_h(w)-\eta_h(w), \psi_h(s_h,a_h)\rangle\|_2^2 \le \zeta_n',$$

where $\zeta_n' := \log(2|\Psi_3|nH/\delta)$.

Proof. We present the detailed proof in Appendix C. □

The following lemma can be obtained from Lemma 39 in Zanette et al. (2021) and Lemma 11 in Uehara et al. (2021).

Lemma 10. *For $\widetilde{\Sigma}_h^n$ defined in eq. (13), we define its expected value as follows:*

$$\Sigma_h^n = \sum_{\tau=1}^{n-1} \mathop{\mathbb{E}}_{\substack{w\sim q \\ s_h\sim(P_w,\pi_w^\tau) \\ a_h\sim\mathcal{U}(\mathcal{A})}} \phi_h(s_h,a_h)\phi_h(s_h,a_h)^\top + \widetilde{\lambda}_n I,$$

$$\Lambda_h^n = \sum_{\tau=1}^{n-1} \mathop{\mathbb{E}}_{\substack{w\sim q \\ s_h\sim(P_w,\pi_w^\tau) \\ a_h\sim\mathcal{U}(\mathcal{A})}} \psi_h(s_h,a_h)\psi_h(s_h,a_h)^\top + \widetilde{\xi}_n I$$

where $\widetilde{\lambda}_n = \widetilde{\gamma}_1 d\log(2nH/\delta)$ and $\widetilde{\xi}_n = \widetilde{\gamma}_2 d\log(2nH/\delta)$. We further define $\widetilde{\mathcal{E}}_0 = \widetilde{\mathcal{E}}_1 \cup \widetilde{\mathcal{E}}_2$ where:

$$\widetilde{\mathcal{E}}_1 = \Bigg\{ \forall n\in[N], h\in[H], s\in\mathcal{S}, a\in\mathcal{A}, \\ \frac{1}{5}\|\phi_h(s,a)\|_{(\Sigma_h^n)^{-1}} \le \|\phi_h(s,a)\|_{(\widetilde{\Sigma}_h^n)^{-1}} \le 3\|\phi_h(s,a)\|_{(\Sigma_h^n)^{-1}} \Bigg\},$$

$$\widetilde{\mathcal{E}}_2 = \Bigg\{ \forall n\in[N], h\in[H], s\in\mathcal{S}, a\in\mathcal{A}, \\ \frac{1}{5}\|\psi_h(s,a)\|_{(\Lambda_h^n)^{-1}} \le \|\psi_h(s,a)\|_{(\widetilde{\Lambda}_h^n)^{-1}} \le 3\|\psi_h(s,a)\|_{(\Lambda_h^n)^{-1}} \Bigg\}.$$

Then we have $\mathbb{P}(\widetilde{\mathcal{E}}_0) \ge 1-\delta$.

We next prove a number of supporting lemmas that are useful for our proof.

Lemma 11. *Suppose Assumption 3 holds. Given any function $f : \mathcal{S}\times\mathcal{A}\times\mathcal{W}\to\mathbb{R}$, for any given context $w\in\mathcal{W}$, we have:*

$$\mathop{\mathbb{E}}_{\substack{w_\tau\sim q \\ s_h\sim(P_{w_\tau},\pi_{w_\tau}^\tau) \\ a_h\sim\mathcal{U}(\mathcal{A})}} f(s_h,a_h,w) \le C^2 \mathop{\mathbb{E}}_{\substack{s_h\sim(P_w,\pi_w^\tau) \\ a_h\sim\mathcal{U}(\mathcal{A})}} f(s_h,a_h,w),$$

where $C = \sqrt{\frac{p_{\max}}{p_{\min}}}$.

Proof. Recall that we use $p(s_h|\pi_w, P_w)$ to denote the probability density of s_h when $s_h \sim (P_w, \pi_w)$. For any given $w_\tau \in \mathcal{W}$, we have:

$$\begin{aligned}\mathop{\mathbb{E}}_{\substack{s_h\sim(P_{w_\tau},\pi^\tau_{w_\tau})\\ a_h\sim\mathcal{U}(\mathcal{A})}} f(s_h,a_h,w) &= \int_{\mathcal{S}} \sum_{a_h} f(s_h,a_h,w)\cdot p(s_h|\pi_{w_\tau},P_{w_\tau})\cdot\frac{1}{K}ds_h \\ &\overset{(i)}{\leq} \frac{p_{\max}}{p_{\min}}\int_{\mathcal{S}}\sum_{a_h} f(s_h,a_h,w)\cdot p(s_h|\pi_w,P_w)\cdot\frac{1}{K}ds_h \\ &= C^2 \mathop{\mathbb{E}}_{\substack{s_h\sim(P_w,\pi^\tau_w)\\ a_h\sim\mathcal{U}(\mathcal{A})}} f(s_h,a_h,w),\end{aligned}$$

where (i) follows from Assumption 3. Note that the right-hand-side of the above equation is independent of w_τ. By taking the expectation over w_τ on both sides, we obtain the desired result. □

To simplify the notation, we define

$$\widetilde{\zeta}^n_h(w) = \sum_{\tau=1}^{n-1} \mathop{\mathbb{E}}_{\substack{s_h\sim(P_w,\pi^\tau_w)\\ a_h\sim\mathcal{U}(\mathcal{A})}} \left\|\langle\widetilde{\mu}^n_h(\cdot,w)-\mu_h(\cdot,w),\phi_h(s_h,a_h)\rangle\right\|^2_{TV}. \tag{37}$$

Now we present a lemma to bound the value function gap $|(P_{w,h}-\widetilde{P}^n_{w,h})V_{w,h+1}(s,a)|$. Differently from Lemma 3 in Appendix A, we cannot bound the context-varying gap by the norms of context-independent representations as bonus terms for any (s,a,w). The following lemma helps to decompose the context-varying value function gap into context-dependent and context-independent components.

Lemma 12. *Suppose $\widetilde{P}^n_h(\cdot|s_h,a_h,w) = \langle\widetilde{\mu}^n_h(\cdot,w),\phi_h(s_h,a_h)\rangle$ is the estimated context-varying linear transition kernel at step $h\in[H]$ in episode $n\in[N]$. Consider a generic non-negative function $f:\mathcal{S}\to\mathbb{R}$ which is bounded by B, i.e., $f(s)\in[0,B]$ for any $s\in\mathcal{S}$. Then any $(s_h,a_h,w)\in\mathcal{S}\times\mathcal{A}\times\mathcal{W}$, with probability at least $1-\delta/2$, we have:*

$$\begin{aligned}&\left|\int_{\mathcal{S}} f(s')\left(\widetilde{\mu}^n_h(s',w)-\mu_h(s',w)\right)^\top\phi_h(s_h,a_h)ds'\right| \\ &\quad\leq \min\left\{\frac{CB\sqrt{dN}}{2\sqrt{K}}\left\|\phi_h(s_h,a_h)\right\|^2_{(\Sigma^n_h)^{-1}},B\right\}+\frac{B\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d+2C^2\widetilde{\zeta}^n_h(w)).\end{aligned}$$

Proof. Consider any given $(s_h,a_h,w)\in\mathcal{S}\times\mathcal{A}\times\mathcal{W}$. We first obtain:

$$\begin{aligned}&\left|\int_{\mathcal{S}} f(s')(\widetilde{\mu}^n_h(s',w)-\mu_h(s',w))^\top\phi_h(s_h,a_h)ds'\right| \\ &\quad\overset{(i)}{\leq} \left\|\phi_h(s_h,a_h)\right\|_{(\Sigma^n_h)^{-1}}\cdot\left\|\int_{\mathcal{S}} f(s')\left(\widetilde{\mu}^n_h(s',w)-\mu_h(s',w)\right)ds'\right\|_{(\Sigma^n_h)},\end{aligned} \tag{38}$$

where (i) follows from the Cauchy-Schwarz inequality. Then, we further derive that

$$\begin{aligned}
&\left\|\int_{\mathcal{S}} f(s)\left(\widetilde{\mu}_h^n(s,w)-\mu_h(s,w)\right)ds\right\|^2_{(\Sigma_h^n)}\\
&\quad=\widetilde{\lambda}_n\cdot\left\|\int_{\mathcal{S}} f(s)(\widetilde{\mu}_h^n(s,w)-\mu_h(s,w))ds\right\|^2\\
&\qquad+\sum_{\tau=1}^{n-1}\mathop{\mathbb{E}}_{\substack{w_\tau\sim q\\ s'_h\sim(P_{w_\tau},\pi^\tau_{w_\tau})\\ a'_h\sim\mathcal{U}(\mathcal{A})}}\left(\int_{\mathcal{S}} f(s)\left(\widetilde{\mu}_h^n(s,w)-\mu_h(s,w)\right)^\top\phi_h\left(s'_h,a'_h\right)ds\right)^2\\
&\quad\overset{(i)}{\leq} 2\lambda_n dB^2+4B^2\sum_{\tau=1}^{n-1}\mathop{\mathbb{E}}_{\substack{w_\tau\sim q\\ s'_h\sim(P_{w_\tau},\pi^\tau_{w_\tau})\\ a'_h\sim\mathcal{U}(\mathcal{A})}}\|\langle\widetilde{\mu}_h^n(\cdot,w)-\mu_h(\cdot,w),\phi_h\left(s'_h,a'_h\right)\rangle\|^2_{TV}\\
&\quad\overset{(ii)}{\leq} B^2\left(2\lambda_n d+4C^2\widetilde{\zeta}_h^n(w)\right),
\end{aligned}\tag{39}$$

where (i) follow from Definition 2 and from the definition of the total variation distance, and (ii) follows from Lemma 11 and the notation eq. (37). Notice that since $f(s)\in[0,B]$ for any $s\in\mathcal{S}$, we have:

$$\begin{aligned}
&\left|\int_{\mathcal{S}} f\left(s'\right)\left(\widetilde{\mu}_h^n\left(s'\right)-\mu_h\left(s'\right)\right)^\top\phi_h(s,a,w)ds'\right|\\
&\qquad=\left|\mathop{\mathbb{E}}_{s'\sim\widetilde{P}_h^n(\cdot|s,a,w)}f(s')-\mathop{\mathbb{E}}_{s'\sim P_h(\cdot|s,a,w)}f(s')\right|\leq B.
\end{aligned}\tag{40}$$

Then, we have:

$$\begin{aligned}
&\left|\int_{\mathcal{S}} f\left(s'\right)\left(\widetilde{\mu}_h^n\left(s',w\right)-\mu_h\left(s',w\right)\right)^\top\phi_h(s_h,a_h)ds'\right|\\
&\quad\overset{(i)}{\leq}\min\left\{\|\phi_h(s_h,a_h)\|_{(\Sigma_h^n)^{-1}}\cdot\sqrt{B^2\left(2\lambda_n d+4C^2\widetilde{\zeta}_h^n(w)\right)},B\right\}\\
&\quad=\min\left\{\sqrt{\frac{CB\sqrt{dN}}{\sqrt{K}}\|\phi_h(s_h,a_h)\|^2_{(\Sigma_h^n)^{-1}}}\cdot\sqrt{\frac{B\sqrt{K}}{C\sqrt{dN}}(2\widetilde{\lambda}_n d+4C^2\widetilde{\zeta}_h^n(w))},B\right\}\\
&\quad\overset{(ii)}{\leq}\min\left\{\frac{CB\sqrt{dN}}{2\sqrt{K}}\|\phi_h(s_h,a_h)\|^2_{(\Sigma_h^n)^{-1}}+\frac{B\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d+2C^2\widetilde{\zeta}_h^n(w)),B\right\}\\
&\quad\overset{(iii)}{\leq}\min\left\{\frac{CB\sqrt{dN}}{2\sqrt{K}}\|\phi_h(s_h,a_h)\|^2_{(\Sigma_h^n)^{-1}},B\right\}+\frac{B\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d+2C^2\widetilde{\zeta}_h^n(w)),
\end{aligned}$$

where (i) follows from eqs. (38) to (40), (ii) follows from the fact that $ab\leq\frac{1}{2}a^2+\frac{1}{2}b^2$ and (iii) follows from the fact that $\min\{a,b+c\}\leq\min\{a,b\}+\min\{a,c\}$ for $a,b,c\geq 0$. □

Lemma 13. *Given that the event $\widetilde{\mathcal{E}}_0$ occurs, for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, define the function $\widetilde{b}_h^n(s,a):=\min\left\{\widetilde{\alpha}_n\|\phi_h(s,a)\|^2_{\left(\widetilde{\Sigma}_h^n\right)^{-1}},H\right\}$, where $\widetilde{\alpha}_n=\frac{25CH\sqrt{dN}}{2\sqrt{K}}$. Then for any context-dependent policy π_w, with probability at least $1-\delta/2$, we have:*

$$\mathop{\mathbb{E}}_{w\sim q}\left|V^{\pi_w}_{\mathcal{M}(w)}-V^{\pi_w}_{\mathcal{M}^{(r,\widetilde{P}^n)}(w)}\right|\leq\mathop{\mathbb{E}}_{w\sim q}\bar{V}^{\pi_w}_{\mathcal{M}^{(\widetilde{b}^n,\widetilde{P}^n)}(w)}+\frac{H^2\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d+2C^2\zeta_n).$$

Proof. Recall the truncated value functions $\bar{V}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}$ and $\bar{Q}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}$ for a generic MDP $\mathcal{M}^{(r',P')}(w)=$

$(\mathcal{S}, \mathcal{A}, P'_w, r'_w, H)$ are defined as:

$$\bar{Q}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}(s_h, a_h) = \min\{H, r'(s_h, a_h, w) + P'_{h,w}\bar{V}^{\pi}_{h+1,\mathcal{M}^{(r',P')}(w)}(s_h, a_h)\},$$
$$\bar{V}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}(s_h) = \mathop{\mathbb{E}}_{\pi}\left[\bar{Q}^{\pi}_{h,\mathcal{M}^{(r',P')}(w)}(s_h, a_h)\right].$$

We first prove that

$$\left|V^{\pi_w}_{\mathcal{M}(w)} - V^{\pi_w}_{\mathcal{M}^{(r,\tilde{P}^n)}(w)}\right| \le \bar{V}^{\pi_w}_{\mathcal{M}^{(\tilde{b}^n,\tilde{P}^n)}(w)} + \frac{H\sqrt{K}}{C\sqrt{dN}}\sum_{h=1}^{H}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_h(w)),$$

holds by induction. For the base case $h = H$, we have

$$\begin{aligned}
&\left|V^{\pi_w}_{H,\mathcal{M}(w)}(s_H) - V^{\pi_w}_{H,\mathcal{M}^{(r,\tilde{P}^n)}(w)}(s_H)\right| \\
&\overset{(i)}{\le} \mathop{\mathbb{E}}_{a_H\sim\pi_w}\left|\left(P_{H,w} - \widetilde{P}^n_{H,w}\right)V^{\pi_w}_{H+1,\mathcal{M}(w)}(s_H, a_H)\right| \\
&\overset{(ii)}{\le} \mathop{\mathbb{E}}_{a_H\sim\pi_w}\left[\min\left\{H, \frac{CH\sqrt{dN}}{2\sqrt{K}}\left\|\phi_H(s_H, a_H)\right\|^2_{(\Sigma^n_H)^{-1}}\right\}\right] + \frac{H\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_H(w)) \\
&= \bar{V}^{\pi_w}_{H,\mathcal{M}^{(\tilde{b}^n,\tilde{P}^n)}(w)}(s_H) + \frac{H\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_H(w)),
\end{aligned}$$

where (i) follows from Lemma 23 and (ii) follows from Lemma 12.

Now we assume that

$$\begin{aligned}
&\left|V^{\pi_w}_{h+1,\mathcal{M}(w)}(s_{h+1}) - V^{\pi_w}_{h+1,\mathcal{M}^{(r,\tilde{P}^n)}(w)}(s_{h+1})\right| \\
&\qquad \le \bar{V}^{\pi_w}_{h+1,\mathcal{M}^{(\tilde{b}^n,\tilde{P}^n)}(w)}(s_{h+1}) + \sum_{k=h+1}^{H}\frac{H\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_k(w))
\end{aligned}$$

holds for all $s_{h+1} \in \mathcal{S}$. Then following from the Bellman equation, for all s_h, a_h, we have:

$$\begin{aligned}
&\left|Q^{\pi_w}_{h,\mathcal{M}^{(r,\tilde{P}^n)}(w)}(s_h, a_h) - Q^{\pi_w}_{h,\mathcal{M}(w)}(s_h, a_h)\right| \\
&\qquad = \left|\widetilde{P}^n_{h,w}V^{\pi_w}_{h+1,\mathcal{M}^{(r,\tilde{P}^n)}(w)}(s_h, a_h) - P_{h,w}V^{\pi_w}_{h+1,\mathcal{M}(w)}(s_h, a_h)\right| \\
&\qquad = \left|\widetilde{P}^n_{h,w}\left(V^{\pi_w}_{h+1,\mathcal{M}^{(r,\tilde{P}^n)}(w)} - V^{\pi_w}_{h+1,\mathcal{M}(w)}\right)(s_h, a_h)\right. \\
&\qquad\qquad\qquad \left. + \left(\widetilde{P}^n_{h,w} - P_{h,w}\right)V^{\pi_w}_{h+1,\mathcal{M}(w)}(s_h, a_h)\right| \\
&\qquad \le \left|\widetilde{P}^n_{h,w}\left(V^{\pi_w}_{h+1,\mathcal{M}^{(r,\tilde{P}^n)}(w)} - V^{\pi_w}_{h+1,\mathcal{M}(w)}\right)(s_h, a_h)\right| \\
&\qquad\qquad\qquad + \left|\left(\widetilde{P}^n_{h,w} - P_{h,w}\right)V^{\pi_w}_{h+1,\mathcal{M}(w)}(s_h, a_h)\right|. \qquad (41)
\end{aligned}$$

Now we consider the first term in eq. (41):

$$\begin{aligned}
&\left|\widetilde{P}_{h,w}\left(V^{\pi_w}_{h+1,\mathcal{M}^{(r,\tilde{P}^n)}(w)} - V^{\pi_w}_{h+1,\mathcal{M}(w)}\right)(s_h, a_h)\right| \\
&\qquad \le \widetilde{P}_{h,w}\left|V^{\pi_w}_{h+1,\mathcal{M}^{(r,\tilde{P}^n)}(w)} - V^{\pi_w}_{h+1,\mathcal{M}(w)}\right|(s_h, a_h) \\
&\qquad \overset{(i)}{\le} \widetilde{P}_{h,w}\bar{V}^{\pi_w}_{h+1,\mathcal{M}^{(\tilde{b}^n,\tilde{P}^n)}(w)}(s_h, a_h) + \sum_{k=h+1}^{H}\frac{H\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_k(w)), \qquad (42)
\end{aligned}$$

where (i) follows from the induction hypothesis. Then we upper-bound the second term in eq. (41) as follows:

$$\begin{aligned}
&\left|\left(\widetilde{P}^n_{h,w} - P_{h,w}\right) V^{\pi_w}_{h+1,\mathcal{M}(w)}(s_h,a_h)\right| \\
&\quad\overset{(i)}{\leq} \frac{CH\sqrt{dN}}{2\sqrt{K}} \left\|\phi_h(s_h,a_h)\right\|^2_{(\Sigma^n_h)^{-1}} + \frac{H\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_h(w)) \\
&\quad\overset{(ii)}{\leq} \widetilde{b}^n_h(s_h,a_h) + \frac{H\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_h(w)),
\end{aligned} \tag{43}$$

where (i) follows from Lemma 13 and (ii) follows from that the event $\widetilde{\mathcal{E}}_0$ occurs. Then we have:

$$\begin{aligned}
&\left|Q^{\pi_w}_{h,\mathcal{M}^{(r,\tilde{P}^n)}(w)}(s_h,a_h) - Q^{\pi_w}_{h,\mathcal{M}(w)}(s_h,a_h)\right| \\
&\quad\overset{(i)}{\leq} \min\left\{H, \widetilde{b}^n_h(s_h,a_h) + \widetilde{P}_{h,w}\bar{V}^{\pi_w}_{h+1,\mathcal{M}^{(\tilde{b}^n,\tilde{P}^n)}(w)}(s_h,a_h) + \sum_{k=h}^{H}\frac{H\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_k(w))\right\} \\
&\quad\overset{(ii)}{\leq} \min\left\{H, \widetilde{b}^n_h(s_h,a_h) + \widetilde{P}_{h,w}\bar{V}^{\pi_w}_{h+1,\mathcal{M}^{(\tilde{b}^n,\tilde{P}^n)}(w)}(s_h,a_h)\right\} + \sum_{k=h}^{H}\frac{H\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_k(w)) \\
&\quad= \bar{Q}^{\pi}_{h,\mathcal{M}^{(\tilde{b}^n,\tilde{P}^n)}(w)}(s_h,a_h) + \sum_{k=h}^{H}\frac{H\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_k(w))
\end{aligned} \tag{44}$$

where (i) follows from eqs. (41) to (43) and the fact that $\left|Q^{\pi_w}_{h,\mathcal{M}^{(r,\tilde{P}^n)}(w)}(s_h,a_h) - Q^{\pi_w}_{h,\mathcal{M}(w)}(s_h,a_h)\right|$ is bounded by H, and (ii) follows from that fact that $\min\{a, b+c\} \leq \min\{a,b\} + \min\{a,c\}$ if $a,b,c \geq 0$. Then, by the definition of $\bar{V}^{\pi_w}_{h,\mathcal{M}^{(\tilde{b}^n,\tilde{P}^n)}(w)}(s_h)$, we have:

$$\begin{aligned}
&\left|V^{\pi_w}_{h,\mathcal{M}(w)}(s_h) - V^{\pi_w}_{h,\mathcal{M}^{(r,\tilde{P}^n)}(w)}(s_h)\right| \\
&\quad= \left|\mathop{\mathbb{E}}_{\pi_w}\left[Q^{\pi_w}_{h,\mathcal{M}(w)}(s_h,a_h)\right] - \mathop{\mathbb{E}}_{\pi_w}\left[Q^{\pi_w}_{h,\mathcal{M}^{(r,\tilde{P}^n)}(w)}(s_h,a_h)\right]\right| \\
&\quad\leq \mathop{\mathbb{E}}_{\pi_w}\left[\left|Q^{\pi_w}_{h,\mathcal{M}(w)}(s_h,a_h) - Q^{\pi_w}_{h,\mathcal{M}^{(r,\tilde{P}^n)}(w)}(s_h,a_h)\right|\right] \\
&\quad\overset{(i)}{\leq} \mathop{\mathbb{E}}_{\pi_w}\left[\bar{Q}^{\pi_w}_{h,\mathcal{M}^{(\tilde{b}^n,\tilde{P}^n)}(w)}(s_h,a_h)\right] + \sum_{k=h}^{H}\frac{H\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_k(w)) \\
&\quad= \bar{V}^{\pi_w}_{h,\mathcal{M}^{(\tilde{b}^n,\tilde{P}^n)}(w)}(s_h) + \sum_{k=h}^{H}\frac{H\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_k(w)),
\end{aligned}$$

where (i) follows from eq. (23). Therefore, by induction, we can conclude that:

$$\left|V^{\pi_w}_{\mathcal{M}(w)} - V^{\pi_w}_{\mathcal{M}^{(r,\tilde{P}^n)}(w)}\right| \leq \bar{V}^{\pi_w}_{\mathcal{M}^{(\tilde{b}^n,\tilde{P}^n)}(w)} + \sum_{h=1}^{H}\frac{H\sqrt{K}}{C\sqrt{dN}}(\widetilde{\lambda}_n d + 2C^2\widetilde{\zeta}^n_h(w)).$$

Hence, taking the expectations over the context w on the both sides of the above equation and applying Lemma 8 complete the proof. □

Next we use a similar idea to bound the estimation error of the reward function. To simplify the notation, we define

$$\widetilde{\zeta}^{n\prime}_h(w) = \sum_{\tau=1}^{n-1}\mathop{\mathbb{E}}_{\substack{s_h\sim(P_w,\pi^\tau_w)\\ a_h\sim\mathcal{U}(\mathcal{A})}}\left\|\langle\widetilde{\eta}^n_h(\cdot,w) - \eta_h(\cdot,w), \psi_h(s_h,a_h)\rangle\right\|^2_2. \tag{45}$$

Lemma 14. *Suppose $\widetilde{\eta}^n_h$ is obtained by $\widetilde{\eta}^n_h = \mathrm{LSR}(\mathcal{G}^n_h)$ in Algorithm 2 at step h in episode n and $\widetilde{f}^n_h(s,a,w)$ is the estimated reward function defined in eq. (17). For any $(s_h,a_h,w) \in \mathcal{S}\times\mathcal{A}\times\mathcal{W}$, with probability at least*

$1-\delta/2$, *we have:*

$$\left|\widetilde{f}_h^n(s_h,a_h,w)-r_h(s_h,a_h,w)\right| \leq \min\left\{\frac{C\sqrt{dN}}{2\sqrt{K}}\left\|\psi_h(s_h,a_h)\right\|^2_{(\Lambda_h^n)^{-1}},1\right\}+\frac{\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d+C^2\widetilde{\zeta}_h^{n\prime}(w)).$$

Proof. Following from the fact that $r_h(s_h,a_h,w)\in[0,1]$ for any (s_h,a_h,w) any $h\in[H],n\in[N]$ and the definition in eq. (17), we have:

$$\left|\widetilde{f}_h^n(s_h,a_h,w)-r_h(s_h,a_h,w)\right| \leq \left|\langle\widetilde{\eta}_h^n,\psi_h(s_h,a_h,w)\rangle-r_h(s_h,a_h,w)\right|. \tag{46}$$

Then for any given $(s_h,a_h,w)\in\mathcal{S}\times\mathcal{A}\times\mathcal{W}$, we obtain:

$$\begin{aligned}&|\langle\widetilde{\eta}_h^n(w)-\eta_h(w),\psi_h(s_h,a_h)\rangle|\\ &\quad\overset{(i)}{\leq}\left\|\widetilde{\eta}_h^n(w)-\eta_h(w)\right\|_{\left(\Lambda_h^n\right)}\cdot\left\|\psi_h(s_h,a_h)\right\|_{\left(\Lambda_h^n\right)^{-1}},\end{aligned} \tag{47}$$

where (i) follows from the Cauchy-Schwarz inequality. Then, we further derive that

$$\begin{aligned}&\left\|\widetilde{\eta}_h^n(w)-\eta_h(w)\right\|^2_{\left(\Lambda_h^n\right)}\\ &\quad=\widetilde{\xi}_n\cdot\left\|\widetilde{\eta}_h^n(w)-\eta_h(w)\right\|^2_{\left(\Lambda_h^n\right)}\\ &\qquad+\sum_{\tau=1}^{n-1}\mathop{\mathbb{E}}_{\substack{w_\tau\sim q\\ s_h'\sim(P_{w_\tau},\pi^\tau_{w_\tau})\\ a_h'\sim\mathcal{U}(\mathcal{A})}}\left(\left(\widetilde{\eta}_h^n(w)-\eta_h(w)\right)^\top\phi_h\left(s_h',a_h'\right)\right)^2\\ &\quad\overset{(i)}{\leq}2\widetilde{\xi}_n d+C^2\widetilde{\zeta}_h^{n\prime}(w),\end{aligned} \tag{48}$$

where (i) follows from Definition 2 and from Lemma 11. Then we have:

$$\begin{aligned}&|\langle\widetilde{\eta}_h^n(w)-\eta_h(w),\psi_h(s_h,a_h)\rangle|\\ &\quad\overset{(i)}{\leq}\left\|\psi_h(s_h,a_h)\right\|_{\left(\Lambda_h^n\right)^{-1}}\cdot\sqrt{2\widetilde{\xi}_n d+C^2\widetilde{\zeta}_h^{n\prime}(w)}\\ &\quad=\sqrt{\frac{C\sqrt{dN}}{\sqrt{K}}\left\|\psi_h(s_h,a_h)\right\|^2_{(\Lambda_h^n)^{-1}}}\cdot\sqrt{\frac{\sqrt{K}}{C\sqrt{dN}}(2\widetilde{\xi}_n d+C^2\widetilde{\zeta}_h^{n\prime}(w))}\\ &\quad\overset{(ii)}{\leq}\frac{C\sqrt{dN}}{2\sqrt{K}}\left\|\psi_h(s_h,a_h)\right\|^2_{(\Lambda_h^n)^{-1}}+\frac{\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d+C^2\widetilde{\zeta}_h^{n\prime}(w)),\end{aligned} \tag{49}$$

where (i) follows from eqs. (47) and (48) and (ii) follows from the fact that $ab\leq\frac{1}{2}a^2+\frac{1}{2}b^2$. By combining eqs. (46) and (49), and the fact that $\left|\widetilde{f}_h^n(s_h,a_h,w)-r_h(s_h,a_h,w)\right|\leq 1$, we have:

$$\begin{aligned}&\left|\widetilde{f}_h^n(s_h,a_h,w)-r_h(s_h,a_h,w)\right|\\ &\quad\leq\min\left\{\frac{C\sqrt{dN}}{2\sqrt{K}}\left\|\psi_h(s_h,a_h)\right\|^2_{(\Lambda_h^n)^{-1}}+\frac{\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d+C^2\widetilde{\zeta}_h^{n\prime}(w)),1\right\}\\ &\quad\leq\min\left\{\frac{C\sqrt{dN}}{2\sqrt{K}}\left\|\psi_h(s_h,a_h)\right\|^2_{(\Lambda_h^n)^{-1}},1\right\}+\frac{\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d+C^2\widetilde{\zeta}_h^{n\prime}(w)),\end{aligned}$$

which completes the proof. □

Lemma 15. *Define the function* $\widetilde{c}_h^n(s,a) := \min\left\{\widetilde{\beta}_n \left\|\psi_h(s,a)\right\|^2_{\left(\widetilde{\Lambda}_h^n\right)^{-1}}, 1\right\}$, *where* $\widetilde{\beta}_n = \frac{25C\sqrt{dN}}{2\sqrt{K}}$. *Assume that the event* $\widetilde{\mathcal{E}}_0$ *occurs. Then for any context-dependent policy* π_w, *with probability at least* $1-\delta/2$, *we have:*

$$\mathop{\mathbb{E}}_{w\sim q}\left|\bar{V}^{\pi_w}_{\mathcal{M}(\tilde{f}^n,\tilde{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(r,\tilde{P}^n)(w)}\right| \le \mathop{\mathbb{E}}_{w\sim q}\bar{V}^{\pi_w}_{\mathcal{M}(\tilde{c}^n,\tilde{P}^n)(w)} + \frac{H\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d + C^2\zeta_n').$$

Proof. Note that the values of both $\widetilde{f}_h^n(s,a,w)$ and $\widetilde{c}_h^n(s,a,w)$ are restricted to $[0,1]$ for any (s,a,w) and any $h\in[H], n\in[N]$. Following from Lemma 19, it is equivalent to prove:

$$\mathop{\mathbb{E}}_{w\sim q}\left|V^{\pi_w}_{\mathcal{M}(\tilde{f}^n,\tilde{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(r,\tilde{P}^n)(w)}\right| \le \mathop{\mathbb{E}}_{w\sim q}V^{\pi_w}_{\mathcal{M}(\tilde{c}^n,\tilde{P}^n)(w)} + \frac{H\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d + C^2\zeta_n').$$

Then, we have:

$$\begin{aligned}
&\left|V^{\pi_w}_{\mathcal{M}(r,\tilde{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(\tilde{f}^n,\tilde{P}^n)(w)}\right| \\
&\quad\overset{(i)}{=} \left|\sum_{h=1}^H \mathop{\mathbb{E}}_{(s_h,a_h)\sim(\widetilde{P}_w^n,\pi_w)} \widetilde{f}_h^n(s_h,a_h,w) - r_h(s_h,a_h,w)\right| \\
&\quad\le \sum_{h=1}^H \mathop{\mathbb{E}}_{(s_h,a_h)\sim(\widetilde{P}_w^n,\pi_w)} \left|\widetilde{f}_h^n(s_h,a_h,w) - r_h(s_h,a_h,w)\right| \\
&\quad\overset{(ii)}{\le} \sum_{h=1}^H \left[\mathop{\mathbb{E}}_{(s_h,a_h)\sim(\widetilde{P}_w^n,\pi_w)} \beta_n \left\|\psi_h(s_h,a_h)\right\|^2_{\left(\widetilde{\Lambda}_h^n\right)^{-1}} + \frac{\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d + C^2\widetilde{\zeta}_h^{n\prime}(w))\right] \\
&\quad= V^{\pi_w}_{\mathcal{M}(\tilde{c}^n,\tilde{P}^n)(w)} + \sum_{h=1}^H \frac{\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d + C^2\widetilde{\zeta}_h^{n\prime}(w)),
\end{aligned} \tag{50}$$

where (i) follows from Lemma 23 and (ii) follows from Lemma 14 and from the occurrence of event $\widetilde{\mathcal{E}}_0$. Then, taking expectations over the context w on the both sides of the above equation and applying Lemma 9 complete the proof. □

Lemma 16. *Suppose the event* $\widetilde{\mathcal{E}}_0$ *occurs. Then for any context-dependent policy* π_w, *with probability at least* $1-\delta/2$, *we have*

$$\begin{aligned}
&\mathop{\mathbb{E}}_{w\sim q}\left|\bar{V}^{\pi_w}_{\mathcal{M}(\tilde{f}^n,\tilde{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(w)}\right| \\
&\qquad\le \mathop{\mathbb{E}}_{w\sim q}\bar{V}^{\pi_w}_{\mathcal{M}(\tilde{b}^n,\tilde{P}^n)(w)} + \mathop{\mathbb{E}}_{w\sim q}\bar{V}^{\pi_w}_{\mathcal{M}(\tilde{c}^n,\tilde{P}^n)(w)} + \frac{H^2\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d + C^2\zeta_n' + 2\widetilde{\lambda}_n d + 4C^2\zeta_n).
\end{aligned}$$

Proof. By combining the bounds on the estimation error of both the reward and the transition kernel, characterized respectively in Lemma 13 and Lemma 15, and the fact that $H\ge 1$, we have:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{w\sim q}\left|\bar{V}^{\pi_w}_{\mathcal{M}(\tilde{f}^n,\tilde{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(w)}\right| \\
&\quad= \mathop{\mathbb{E}}_{w\sim q}\left|\bar{V}^{\pi_w}_{\mathcal{M}(\tilde{f}^n,\tilde{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(r,\tilde{P}^n)(w)} + V^{\pi_w}_{\mathcal{M}(r,\tilde{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(w)}\right| \\
&\quad\le \mathop{\mathbb{E}}_{w\sim q}\left|\bar{V}^{\pi_w}_{\mathcal{M}(\tilde{f}^n,\tilde{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(r,\tilde{P}^n)(w)}\right| + \mathop{\mathbb{E}}_{w\sim q}\left|V^{\pi_w}_{\mathcal{M}(r,\tilde{P}^n)(w)} - V^{\pi_w}_{\mathcal{M}(w)}\right| \\
&\quad\le \mathop{\mathbb{E}}_{w\sim q}\bar{V}^{\pi_w}_{\mathcal{M}(\tilde{b}^n,\tilde{P}^n)(w)} + \mathop{\mathbb{E}}_{w\sim q}\bar{V}^{\pi_w}_{\mathcal{M}(\tilde{c}^n,\tilde{P}^n)(w)} + \frac{H^2\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d + C^2\zeta_n' + 2\widetilde{\lambda}_n d + 4C^2\zeta_n),
\end{aligned}$$

which completes the proof. □

B.2 Proof of Theorem 2.

We first restate Theorem 2 below.

Theorem 4 (Restatement of Theorem 2)**.** *Consider a CMDP with varying linear weights as defined in Definition 2. Under Assumptions 2 and 3, for any $\delta \in (0,1)$, with probability at least $1-3\delta/2$, the sequence of policies $\pi^n_{w_n}$ generated by Algorithm 2 satisfies that*

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\left[V^{\pi^*_w}_{\mathcal{M}(w)}-V^{\pi^n_w}_{\mathcal{M}(w)}\right]\\
&\quad\leq 912CH^2\sqrt{\frac{d^3K}{N}}\log\left(1+\frac{N}{\widetilde{\lambda}d}\right)\\
&\qquad+\frac{H^2}{C}\sqrt{\frac{K}{dN}}\left(2\widetilde{\xi}_N d+C^2\log\left(\frac{2HN|\Psi_3|}{\delta}\right)+4\widetilde{\lambda}_N d+8C^2\log\left(\frac{2HN|\Psi_2|}{\delta}\right)\right),
\end{aligned}$$

where $C=\sqrt{\frac{p_{\max}}{p_{\min}}}$, $\widetilde{\lambda}_n=\widetilde{\gamma}_1 d\log(2nH/\delta)$, $\widetilde{\xi}_n=\widetilde{\gamma}_1 d\log(2nH/\delta)$, $\widetilde{\gamma}_1,\widetilde{\gamma}_2=\mathcal{O}(1)$ and $\widetilde{\lambda}=\min\{\widetilde{\lambda}_1,\widetilde{\xi}_1\}$. To achieve an ϵ average sub-optimality gap, at most $\mathcal{O}\left(\frac{H^4d^3K\log^2(|\Psi_2||\Psi_3|/\delta^2)}{\epsilon^2}\cdot\frac{p_{\max}}{p_{\min}}\right)$ episodes are needed.

Proof. First, we derive an optimistic upper bound for the expectation of optimal value function.

$$\begin{aligned}
\mathop{\mathbb{E}}_{w\sim q}V^{\pi^*_w}_{\mathcal{M}(w)}&\overset{(i)}{\leq}\mathop{\mathbb{E}}_{w\sim q}\left[\bar{V}^{\pi^*_w}_{\mathcal{M}(\tilde{f}^n,\tilde{P}^n)(w)}+\bar{V}^{\pi^*_w}_{\mathcal{M}(\tilde{b}^n,\tilde{P}^n)(w)}+\bar{V}^{\pi^*_w}_{\mathcal{M}(\tilde{c}^n,\tilde{P}^n)(w)}\right]\\
&\qquad+\frac{H^2\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d+C^2\zeta'_n+2\widetilde{\lambda}_n d+4C^2\zeta_n)\\
&\overset{(ii)}{\leq}\mathop{\mathbb{E}}_{w\sim q}\left[\bar{V}^{\pi^*_w}_{\mathcal{M}(\tilde{f}^n+\tilde{b}^n+\tilde{c}^n,\tilde{P}^n)(w)}\right]+\frac{H^2\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d+C^2\zeta'_n+2\widetilde{\lambda}_n d+4C^2\zeta_n)\\
&\overset{(iii)}{\leq}\mathop{\mathbb{E}}_{w\sim q}\left[\bar{V}^{\pi^n_w}_{\mathcal{M}(\tilde{f}^n+\tilde{b}^n+\tilde{c}^n,\tilde{P}^n)(w)}\right]+\frac{H^2\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d+C^2\zeta'_n+2\widetilde{\lambda}_n d+4C^2\zeta_n)\\
&\overset{(iv)}{\leq}\mathop{\mathbb{E}}_{w\sim q}\left[\bar{V}^{\pi^n_w}_{\mathcal{M}(\tilde{f}^n,\tilde{P}^n)(w)}+\bar{V}^{\pi^n_w}_{\mathcal{M}(\tilde{b}^n,\tilde{P}^n)(w)}+\bar{V}^{\pi^n_w}_{\mathcal{M}(\tilde{c}^n,\tilde{P}^n)(w)}\right]\\
&\qquad+\frac{H^2\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d+C^2\zeta'_n+2\widetilde{\lambda}_n d+4C^2\zeta_n),
\end{aligned}\tag{51}$$

where (i) follows from Lemma 16, (ii) follows from Lemma 20, (iii) follows from the greedy policy $\pi^n_w=\operatorname{argmax}_\pi \bar{V}^{\pi}_{\mathcal{M}(\tilde{f}^n+\tilde{b}^n+\tilde{c}^n,\tilde{P}^n)(w)}$ and (iv) follows from Lemma 21. Then the average suboptimality gap can be bounded as

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\left[V^{\pi^*_w}_{\mathcal{M}(w)}-V^{\pi^n_w}_{\mathcal{M}(w)}\right]\\
&\overset{(i)}{\leq}\frac{1}{N}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\left[\bar{V}^{\pi^n_w}_{\mathcal{M}(\tilde{f}^n,\tilde{P}^n)(w)}+\bar{V}^{\pi^n_w}_{\mathcal{M}(\tilde{b}^n,\tilde{P}^n)(w)}+\bar{V}^{\pi^n_w}_{\mathcal{M}(\tilde{c}^n,\tilde{P}^n)(w)}-V^{\pi^n_w}_{\mathcal{M}(w)}\right]\\
&\qquad+\frac{H^2\sqrt{K}}{2C\sqrt{dN}}(2\widetilde{\xi}_n d+C^2\zeta'_n+2\widetilde{\lambda}_n d+4C^2\zeta_n)\\
&\overset{(ii)}{\leq}\frac{1}{N}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\left[2\bar{V}^{\pi^n_w}_{\mathcal{M}(\tilde{b}^n,\tilde{P}^n)(w)}+2\bar{V}^{\pi^n_w}_{\mathcal{M}(\tilde{c}^n,\tilde{P}^n)(w)}\right]+\frac{H^2\sqrt{K}}{C\sqrt{dN}}(2\widetilde{\xi}_n d+C^2\zeta'_n+2\widetilde{\lambda}_n d+4C^2\zeta_n),
\end{aligned}\tag{52}$$

where (i) follows from eq. (51), and (ii) follows from Lemma 16.

We next provide an upper bound on $\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\bar{V}^{\pi^n_w}_{\mathcal{M}(\tilde{b}^n,\tilde{P}^n)(w)}$. Define $g^n_h(s,a,w)=(\widetilde{P}^n_{h,w}-$

$P_{h,w})\bar{\bar{V}}^{\pi_w^n}_{h+1,\mathcal{M}(\tilde{b}^n,\tilde{P}^n)(w)}(s_h,a_h)$. Following from Lemma 22 in Appendix D, we have:

$$\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\bar{V}^{\pi_w^n}_{\mathcal{M}(\tilde{b}^n,\tilde{P}^n)(w)}\leq\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}V^{\pi_w^n}_{\mathcal{M}(\tilde{b}^n,P)(w)}+\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}V^{\pi_w^n}_{\mathcal{M}(g^n,P)(w)}. \tag{53}$$

For the first term in the right-hand-side of eq. (53), we obtain an upper bound on the summation of the expected value functions $V^{\pi_w}_{\mathcal{M}(\tilde{b}^n,P)(w)}$ as follows:

$$\begin{aligned}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}V^{\pi_w^n}_{\mathcal{M}(\tilde{b}^n,P)(w)}&=\sum_{n=1}^{N}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{\substack{w\sim q\\(s_h,a_h)\sim(P_w,\pi_w^n)}}\left[\widetilde{\alpha}_n\left\|\phi_h(s_h,a_h)\right\|^2_{(\widetilde{\Sigma}_h^n)^{-1}}\right]\\&\overset{(i)}{\leq}9K\widetilde{\alpha}_N\sum_{h=1}^{H}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{\substack{w\sim q\\s_h\sim(P_w,\pi_w^n)\\a_h\sim\mathcal{U}(\mathcal{A})}}\left[\left\|\phi_h(s_h,a_h)\right\|^2_{(\Sigma_h^n)^{-1}}\right]\\&\overset{(ii)}{\leq}18dHK\widetilde{\alpha}_N\cdot\log\left(1+\frac{N}{d\widetilde{\lambda}}\right),\end{aligned} \tag{54}$$

where (i) follows because the event $\widetilde{\mathcal{E}}_0$ occurs and from the importance sampling, and (ii) follows from Lemma 24. For the second term in the right-hand-side of eq. (53), we derive

$$\begin{aligned}&\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}V^{\pi_w^n}_{\mathcal{M}(g^n,P)(w)}\\&\quad\leq\sum_{n=1}^{N}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{\substack{w\sim q\\(s_h,a_h)\sim(P_w,\pi_w^n)}}\left|g_h^n(s_h,a_h,w)\right|\\&\quad\overset{(i)}{\leq}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{3\widetilde{\alpha}_n}{25}\mathop{\mathbb{E}}_{\substack{w\sim q\\(s_h,a_h)\sim(P_w,\pi_w)}}\left\|\phi_h(s_h,a_h)\right\|^2_{(\Sigma_h^n)^{-1}}+\frac{H^2\sqrt{KN}}{C\sqrt{d}}(\widetilde{\lambda}_Nd+2C^2\zeta_N)\\&\quad\overset{(ii)}{\leq}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{3K\widetilde{\alpha}_n}{25}\mathop{\mathbb{E}}_{\substack{w\sim q\\s_h\sim(P_w,\pi_w)\\a_h\sim\mathcal{U}(\mathcal{A})}}\left\|\phi_h(s_h,a_h)\right\|^2_{(\Sigma_h^n)^{-1}}+\frac{H^2\sqrt{KN}}{C\sqrt{d}}(\widetilde{\lambda}_Nd+2C^2\zeta_N)\\&\quad\overset{(iii)}{\leq}\frac{3HK\widetilde{\alpha}_N}{25}\cdot2d\log\left(1+\frac{N}{d\widetilde{\lambda}}\right)+\frac{H^2\sqrt{KN}}{C\sqrt{d}}(\widetilde{\lambda}_Nd+2C^2\zeta_N),\end{aligned} \tag{55}$$

where (i) follows from Lemma 12 and the fact that $\bar{\bar{V}}^{\pi_w}_{h,\mathcal{M}(\tilde{b}^n,\tilde{P}^n)(w)}(s_h,a_h)\leq 3H$ for any $h\in[H]$, (ii) follows from the importance sampling, and (iii) follows from Lemma 24. Then by combining eqs. (53) to (55), we have:

$$\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\bar{V}^{\pi_w}_{\mathcal{M}(\tilde{b}^n,\tilde{P}^n)(w)}\leq\frac{456dHK\widetilde{\alpha}_n}{25}\cdot\log\left(1+\frac{N}{d\widetilde{\lambda}}\right)+\frac{H^2\sqrt{KN}}{C\sqrt{d}}(\widetilde{\lambda}_Nd+2C^2\zeta_N). \tag{56}$$

Next we provide an upper bound on $\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\bar{V}^{\pi_w^n}_{\mathcal{M}(\tilde{c}^n,\tilde{P}^n)(w)}$. Define $l_h^n(s,a,w)=(\widetilde{P}^n_{h,w}-P_{h,w})\bar{V}^{\pi_w^n}_{h+1,\mathcal{M}(\tilde{c}^n,\tilde{P}^n)(w)}(s_h,a_h)$. Following from Lemma 22 in Appendix D, we have:

$$\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\bar{V}^{\pi_w^n}_{\mathcal{M}(\tilde{c}^n,\tilde{P}^n)(w)}\leq\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}V^{\pi_w^n}_{\mathcal{M}(\tilde{c}^n,P)(w)}+\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}V^{\pi_w^n}_{\mathcal{M}(l^n,P)(w)}. \tag{57}$$

For the first term in eq. (57), we obtain an upper bound on the summation of the expected value function

$V^{\pi_w^n}_{\mathcal{M}^{(\bar{c}^n,P)}(w)}$ as follows:

$$\begin{aligned}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}V^{\pi_w^n}_{\mathcal{M}^{(\bar{c}^n,P)}(w)} &= \sum_{n=1}^{N}\sum_{h=1}^{H}\mathop{\mathbb{E}}_{\substack{w\sim q\\(s_h,a_h)\sim(P_w,\pi_w^n)}}\left[\widetilde{\beta}_n\left\|\psi_h(s_h,a_h)\right\|^2_{(\widetilde{\Lambda}_h^n)^{-1}}\right]\\ &\overset{(i)}{\leq} 9K\widetilde{\beta}_N\sum_{h=1}^{H}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{\substack{w\sim q\\ s_h\sim(P_w,\pi_w^n)\\ a_h\sim\mathcal{U}(\mathcal{A})}}\left[\left\|\psi_h(s_h,a_h)\right\|^2_{(\Lambda_h^n)^{-1}}\right]\\ &\overset{(ii)}{\leq} 18dHK\widetilde{\beta}_N\cdot\log\left(1+\frac{N}{d\widetilde{\lambda}}\right),\end{aligned} \tag{58}$$

where (i) follows because the event $\widetilde{\mathcal{E}}_0$ occurs and from the importance sampling, and (ii) follows from Lemma 24. Then, since $\bar{V}^{\pi_w}_{h,\mathcal{M}^{(\bar{c}^n,\bar{P}^n)}(w)}(s,a)\leq 3H$ for any $h\in[H]$, we bound the second term in the right-hand-side of eq. (57) similarly to eq. (55) and obtain:

$$\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}V^{\pi_w^n}_{\mathcal{M}^{(l^n,P)}(w)}\leq\frac{3HK\widetilde{\alpha}_N}{25}\cdot 2d\log\left(1+\frac{N}{d\widetilde{\lambda}}\right)+\frac{H^2\sqrt{KN}}{C\sqrt{d}}(\widetilde{\lambda}_N d+2C^2\zeta_N). \tag{59}$$

Then by combining eqs. (57) to (59), we obtain:

$$\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\bar{V}^{\pi_w^n}_{\mathcal{M}^{(\bar{c}^n,\bar{P}^n)}(w)}\leq\left(\frac{6HK\widetilde{\alpha}_N}{25}+18HK\widetilde{\beta}_N\right)\cdot d\log\left(1+\frac{N}{d\widetilde{\lambda}}\right)+\frac{H^2\sqrt{KN}}{C\sqrt{d}}(\widetilde{\lambda}_N d+2C^2\zeta_N). \tag{60}$$

By substituting eqs. (56) and (60) into eq. (52), we have

$$\begin{aligned}&\frac{1}{N}\sum_{n=1}^{N}\mathop{\mathbb{E}}_{w\sim q}\left[V^{\pi_w^*}_{\mathcal{M}(w)}-V^{\pi_w^n}_{\mathcal{M}(w)}\right]\\ &\quad\leq\left(\frac{912dHK\widetilde{\alpha}_N}{25N}+18HK\widetilde{\beta}_N\right)\cdot\log\left(1+\frac{N}{d\widetilde{\lambda}}\right)+\frac{H^2\sqrt{K}}{C\sqrt{dN}}(2\widetilde{\xi}_N d+C^2\zeta'_N+4\widetilde{\lambda}_N d+8C^2\zeta_N)\\ &\quad\leq 912CH^2\sqrt{\frac{d^3K}{N}}\log\left(1+\frac{N}{\widetilde{\lambda}d}\right)\\ &\qquad+\frac{H^2}{C}\sqrt{\frac{K}{dN}}\left(2\widetilde{\xi}_N d+C^2\log\left(\frac{2HN|\Psi_3|}{\delta}\right)+4\widetilde{\lambda}_N d+8C^2\log\left(\frac{2HN|\Psi_2|}{\delta}\right)\right),\end{aligned}$$

where $C=\sqrt{\frac{p_{\max}}{p_{\min}}}$, $\widetilde{\lambda}_n=\widetilde{\gamma}_1 d\log(2nH/\delta)$, $\widetilde{\xi}_n=\widetilde{\gamma}_1 d\log(2nH/\delta)$, $\widetilde{\gamma}_1,\widetilde{\gamma}_2=\mathcal{O}(1)$ and $\widetilde{\lambda}=\min\{\widetilde{\lambda}_1,\widetilde{\xi}_1\}$.

Note that since $\widetilde{\lambda}_N=\widetilde{\mathcal{O}}(d)$, the upper bound on the average suboptimality gap is of the order $\mathcal{O}\left(\sqrt{\frac{H^4d^3KC^2\log^2(|\Psi_2||\Psi_3|/\delta^2)}{N}}\right)$. Then it can be seen that to achieve an ϵ average sub-optimality gap, at most $\mathcal{O}\left(\frac{H^4d^3K\log^2(|\Psi_2||\Psi_3|/\delta^2)}{\epsilon^2}\cdot\frac{p_{\max}}{p_{\min}}\right)$ episodes are needed. This completes the proof. □

C Least Square Regression (LSR) Guarantee

In this section, we derive an upper bound on a LSR estimation of a generic deterministic model. To simplify the notation, we denote the instance space as $\mathcal{X}$. Our goal is to estimate a deterministic function $f^*(x)$ that belongs to a function class $\mathcal{F}:\mathcal{X}\to\mathbb{R}$. Our estimation is based on a dataset $D:=\{x_i,y_i\}_{i=1}^n$, where $x_i\sim\mathcal{D}_i=\mathcal{D}_i(x_{1:i-1},y_{1:i-1})$ and $y_i=f^*(x_i)$, where $\mathcal{D}_i$ depends on the previous samples. We further define a tangent sequence $\mathcal{D}':=\{x'_i,y'_i\}_{i=1}^n$, where $x'_i\sim\mathcal{D}_i=\mathcal{D}_i(x_{1:i-1},y_{1:i-1})$ and $y'_i=f^*(x'_i)$. We obtain the estimator via the following minimization problem:

$$\hat{f}=\operatorname*{argmin}_{f\in\mathcal{F}}\sum_{i=1}^{n}\|f(x_i)-f^*(x_i)\|_2^2.$$

We first prove the following decoupling inequality, which is inspired by Lemma 24 of (Agarwal et al., 2020).

Lemma 17. *Suppose D is a dataset of n samples and D' is a tangent sequence. Let $L(f, D) = \sum_{i=1}^{n} l(f, (x_i, y_i))$ be any function that decomposes additively where l is any function. We denote $\hat{f}(D)$ as an estimator taking input random variable D. Then:*

$$\mathbb{E}_D \left[\exp\left(\mathbb{E}_{D'} \left(L\left(\hat{f}(D), D' \right) \right) - L(\hat{f}(D), D) - \log |\mathcal{F}| \right) \right] \leq 1.$$

Proof. Let π be the uniform distribution over $\mathcal{F}$ and let $g : \mathcal{F} \to \mathbb{R}$ be any function. Define $\mu(f) := \frac{\exp(g(f))}{\sum_f \exp(g(f))}$, which is clearly a probability distribution. Now consider any other probability distribution $\hat{\pi}$ over $\mathcal{F}$, and we have

$$\begin{aligned}
0 \leq \mathrm{KL}(\hat{\pi} \| \mu) &= \sum_f \hat{\pi}(f) \log(\hat{\pi}(f)) + \sum_f \hat{\pi}(f) \log \left(\sum_{f'} \exp\left(g\left(f' \right) \right) \right) - \sum_f \hat{\pi}(f) g(f) \\
&= \mathrm{KL}(\hat{\pi} || \pi) - \sum_f \hat{\pi}(f) g(f) + \log \mathbb{E}_{f \sim \pi} \exp(g(f)) \\
&\leq \log |\mathcal{F}| - \sum_f \hat{\pi}(f) g(f) + \log \mathbb{E}_{f \sim \pi} \exp(g(f)).
\end{aligned}$$

Re-arranging the above equation, we obtain that

$$\sum_f \hat{\pi}(f) g(f) - \log |\mathcal{F}| \leq \log \mathbb{E}_{f \sim \pi} \exp(g(f)).$$

Now we let $\hat{\pi}(f) = 1\{\hat{f}(D)\}$ and $g(f) = \mathbb{E}_{D'} L(f, D') - L(f, D)$, and have:

$$\begin{aligned}
\mathbb{E}_{D'} L(f(D), D') - L(f(D), D) - \log |\mathcal{F}| &\leq \log \mathbb{E}_{f \sim \pi} \frac{\exp \mathbb{E}_{\mathcal{D}'} \left(L\left(\widehat{f}(\mathcal{D}), \mathcal{D}' \right) \right)}{\exp(L(\widehat{f}(\mathcal{D}), \mathcal{D}))} \\
&\leq \log \mathbb{E}_{f \sim \pi} \frac{\mathbb{E}_{\mathcal{D}'} \exp \left(L\left(\widehat{f}(\mathcal{D}), \mathcal{D}' \right) \right)}{\exp(L(\widehat{f}(\mathcal{D}), \mathcal{D}))}.
\end{aligned}$$

We exponentiate both sides of the above equation and then take expectation over D on both sides, and obtain

$$\mathbb{E}_D \left[\exp(\mathbb{E}_{D'} L(f(D), D') - L(f(D), D) - \log |\mathcal{F}|) \right] \leq \mathbb{E}_D \left[\mathbb{E}_{f \sim \pi} \frac{\mathbb{E}_{\mathcal{D}'} \exp \left(L\left(\widehat{f}(\mathcal{D}), \mathcal{D}' \right) \right)}{\exp(L(\widehat{f}(\mathcal{D}), \mathcal{D}))} \right].$$

Note that, conditioned on $\mathcal{D}$, the samples in the tangent sequence $\mathcal{D}'$ are independent, which yields

$$\mathbb{E}_{\mathcal{D}'} \exp \left[L\left(\widehat{f}(\mathcal{D}), \mathcal{D}' \right) \mid \mathcal{D} \right] = \prod_{i=1}^{n} \exp \left(\mathbb{E}_{(x_i, y_i) \sim \mathcal{D}_i} \left[l\left(f, (x_i, y_i) \right) \right] \right).$$

Then we can conclude that

$$\mathbb{E}_D \left[\exp\left(\mathbb{E}_{D'} \left(L\left(\hat{f}(D), D' \right) \right) - L(\hat{f}(D), D) - \log |\mathcal{F}| \right) \right] \leq 1.$$

□

Now we present the LSR guarantee as follows.

Lemma 18. *Assume $|\mathcal{F}| \leq \infty$ and $f^* \in \mathcal{F}$. Then with probability at least $1 - \delta$, we have:*

$$\sum_{i=1}^{n} \mathbb{E}_{x_i \sim \mathcal{D}_i} \left\| f^*(x_i) - f(x_i) \right\|_2^2 \leq \log |\mathcal{F}| / \delta.$$

Proof. We first apply the Chernoff bound to Lemma 17. With probability $1-\delta$, we have:

$$\mathbb{E}_{D'} L(\hat{f}(D), D') \leq L(\hat{f}(D), D) + \log \frac{|\mathcal{F}|}{\delta}.$$

Since $f^* \in \mathcal{F}$, we have $L(\hat{f}(D), D) \leq L(f^*, D) = 0$. Then we derive

$$\begin{aligned}
\mathbb{E}_{D'} L(\hat{f}(D), D') &= \mathbb{E}_{D'} \left[\sum_{i=1}^{n} \|\hat{f}(x_i') - f^*(x_i')\|^2 \middle| D \right] \\
&= \mathop{\mathbb{E}}_{x_i' \sim D_i} \left[\sum_{i=1}^{n} \|\hat{f}(x_i') - f^*(x_i')\|^2 \right] \\
&= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{x_i \sim D_i} \|\hat{f}(x_i) - f^*(x_i)\|^2,
\end{aligned}$$

which completes the proof. □

D Auxiliary Lemmas

The following lemma proves that the truncated value function is equal to the value function if the reward function is bounded by one.

Lemma 19. *For a generic reward function r' such that $r_h'(s,a,w) \in [0,1]$ for any (s,a,w) and any $h \in [H]$, a generic transition kernel P', and any context w, we have:*

$$\bar{V}^{\pi_w}_{\mathcal{M}(r',P')(w)} = V^{\pi_w}_{\mathcal{M}(r',P')(w)},$$

where $\bar{V}^{\pi_w}_{\mathcal{M}(r',P')(w)}$ is defined by eq. (19).

Proof. We develop the proof by induction. For the case $h = H+1$, we have $\bar{V}^{\pi_w}_{H+1,\mathcal{M}(r',P')(w)}(s_{H+1}) = 0 = V^{\pi_w}_{H+1,\mathcal{M}(r',P')(w)}(s_{H+1})$. Assume that $\bar{V}^{\pi_w}_{h+1,\mathcal{M}(r',P')(w)}(s_{h+1}) = V^{\pi_w}_{h+1,\mathcal{M}(r',P')(w)}(s_{h+1})$ holds for any s_{h+1}. Then we have:

$$\begin{aligned}
\bar{Q}^{\pi_w}_{h,\mathcal{M}(r',P')(w)}(s_h,a_h) &= \min\left\{H, r_h'(s_h,a_h,w) + P'_{h,w}\bar{V}^{\pi}_{h+1,\mathcal{M}(r',P')(w)}(s_h,a_h)\right\} \\
&\overset{(i)}{=} \min\left\{H, r_h'(s_h,a_h,w) + P'_{h,w}V^{\pi}_{h+1,\mathcal{M}(r',P')(w)}(s_h,a_h)\right\} \\
&= \min\left\{H, Q^{\pi_w}_{h,\mathcal{M}(r',P')(w)}(s_h,a_h)\right\} \\
&\overset{(ii)}{=} Q^{\pi_w}_{h,\mathcal{M}(r',P')(w)}(s_h,a_h),
\end{aligned}$$

where (i) follows from the induction hypothesis and (ii) follows from the fact that the reward function r' is always bounded by 1. By taking expectations on the both sides of the above equation, we conclude that for all s_h,

$$\bar{V}^{\pi_w}_{h,\mathcal{M}(r',P')(w)}(s_h) = V^{\pi_w}_{h,\mathcal{M}(r',P')(w)}(s_h),$$

which completes the proof. □

Next, we present two lemmas to prove the relationship between the value function of a sum of reward functions and the sum of value functions with each corresponding to a reward function.

Lemma 20. *Consider three MDPs denoted as $(\mathcal{S}, \mathcal{A}, P', r^{(i)}, H)$ for $i = 1,2,3$, where P' is a generic transition kernel and $r^{(i)}$ are generic reward functions. Then for any context w and context-dependent policy π_w, we have:*

$$\sum_{i=1}^{3} \bar{V}^{\pi_w}_{\mathcal{M}(r^{(i)},P')(w)} \leq \bar{\bar{V}}^{\pi_w}_{\mathcal{M}(r^{(1)}+r^{(2)}+r^{(3)},P'')(w)}.$$

Proof. We develop the proof by induction. For the case $h = H + 1$, we have

$$\sum_{i=1}^{3} \bar{V}^{\pi_w}_{H+1,\mathcal{M}^{(r^{(i)},P')}(w)}(s_{H+1}) = 0 = \bar{\bar{V}}^{\pi_w}_{H+1,\mathcal{M}^{(r^{(1)}+r^{(2)}+r^{(3)},P')}(w)}(s_{H+1}).$$

We assume that $\sum_{i=1}^{3} \bar{V}^{\pi_w}_{H+1,\mathcal{M}^{(r^{(i)},P')}(w)}(s_{h+1}) \le \bar{\bar{V}}^{\pi_w}_{H+1,\mathcal{M}^{(r^{(1)}+r^{(2)}+r^{(3)},P')}(w)}(s_{H+1})$ holds for any s_{h+1}. Then by the definition in eq. (19), we have:

$$\begin{aligned}
&\sum_{i=1}^{3} \bar{Q}^{\pi_w}_{h,\mathcal{M}^{(r^{(i)},P')}(w)}(s_h, a_h) \\
&\quad\le \min\left\{3H, \sum_{i=1}^{3}\left[r_h^{(i)}(s_h, a_h, w) + P'_{h,w}\bar{V}^{\pi_w}_{h+1,\mathcal{M}^{(r^{(i)},P')}(w)}(s_h, a_h)\right]\right\} \\
&\quad\overset{(i)}{\le} \min\left\{3H, \sum_{i=1}^{3}\left[r_h^{(i)}(s_h, a_h, w)\right] + P'_{h,w}\bar{\bar{V}}^{\pi_w}_{h+1,\mathcal{M}^{(r^{(1)}+r^{(2)}+r^{(3)},P')}(w)}(s_h, a_h)\right\} \\
&\quad= \bar{\bar{Q}}^{\pi_w}_{h,\mathcal{M}^{(r^{(1)}+r^{(2)}+r^{(3)},P')}(w)}(s_h, a_h),
\end{aligned}$$

where (i) follows from the induction hypothesis. By taking expectations on the both sides of the above equation, we conclude that for any s_h,

$$\sum_{i=1}^{3} \bar{V}^{\pi_w}_{h,\mathcal{M}^{(r^{(i)},P')}(w)}(s_h) \le \bar{\bar{V}}^{\pi_w}_{h,\mathcal{M}^{(r^{(1)}+r^{(2)}+r^{(3)},P')}(w)}(s_h),$$

which completes the proof. □

Lemma 21. *Suppose there are three MDPs denoted as: $(\mathcal{S}, \mathcal{A}, P', r^{(i)}, H)$ for $i = 1, 2, 3$ where P' is a generic transition kernel and $r^{(i)}$ are generic reward functions. Then for any context w and context-dependent policy π_w, we have:*

$$\bar{\bar{V}}^{\pi_w}_{\mathcal{M}^{(r^{(1)}+r^{(2)}+r^{(3)},P')}(w)} \le \sum_{i=1}^{3} \bar{\bar{V}}^{\pi_w}_{\mathcal{M}^{(r^{(i)},P')}(w)}.$$

Proof. We develop the proof by induction. For the case $h = H + 1$, we have $\bar{\bar{V}}^{\pi_w}_{H+1,\mathcal{M}^{(r^{(1)}+r^{(2)}+r^{(3)},P')}(w)}(s_{H+1}) = 0 = \sum_{i=1}^{3} \bar{\bar{V}}^{\pi_w}_{H+1,\mathcal{M}^{(r^{(i)},P')}(w)}(s_{H+1})$.
We assume that

$$\bar{\bar{V}}^{\pi_w}_{h+1,\mathcal{M}^{(r^{(1)}+r^{(2)}+r^{(3)},P')}(w)}(s_{h+1}) \le \sum_{i=1}^{3} \bar{\bar{V}}^{\pi_w}_{h+1,\mathcal{M}^{(r^{(i)},P')}(w)}(s_{h+1})$$

holds for any s_{h+1}. Then by the definition in eq. (11), we have:

$$\begin{aligned}
&\bar{\bar{Q}}^{\pi_w}_{h,\mathcal{M}^{(r^{(1)}+r^{(2)}+r^{(3)},P')}(w)}(s_h, a_h) \\
&\quad= \min\left\{3H, \sum_{i=1}^{3} r_h^{(i)}(s_h, a_h, w) + P'_{h,w}\bar{\bar{V}}^{\pi_w}_{h+1,\mathcal{M}^{(r^{(1)}+r^{(2)}+r^{(3)},P')}(w)}(s_h, a_h)\right\} \\
&\quad\overset{(i)}{\le} \min\left\{3H, \sum_{i=1}^{3}\left[r_h^{(i)}(s_h, a_h, w)\right] + \sum_{i=1}^{3}\left[P'_{h,w}\bar{\bar{V}}^{\pi_w}_{h+1,\mathcal{M}^{(r^{(i)},P')}(w)}(s_h, a_h)\right]\right\} \\
&\quad\overset{(ii)}{\le} \sum_{i=1}^{3} \min\left\{3H, r_h^{(i)}(s_h, a_h, w) + P'_{h,w}\bar{\bar{V}}^{\pi_w}_{h+1,\mathcal{M}^{(r^{(i)},P')}(w)}(s_h, a_h)\right\} \\
&\quad= \sum_{i=1}^{3} \bar{\bar{Q}}^{\pi_w}_{h,\mathcal{M}^{(r^{(i)},P')}(w)}(s_h, a_h),
\end{aligned}$$

where (i) follows from the induction hypothesis and (ii) follows from the fact that $\min\{a, b+c\} \leq \min\{a, b\} + \min\{a, c\}$ if $a, b, c \geq 0$. By taking expectations on the both sides of the above equation, we conclude that for any s_h, we have:

$$\bar{\bar{V}}^{\pi_w}_{h,\mathcal{M}^{(r^{(1)}+r^{(2)}+r^{(3)},P')}(w)}(s_h) \leq \sum_{i=1}^{3} \bar{\bar{V}}^{\pi_w}_{h,\mathcal{M}^{(r^{(i)},P')}(w)}(s_h).$$

By induction, we complete the proof. □

We next present a lemma to upper-bound the difference between a truncated value function and a value function with the same generic reward function but different transition kernels by a constructed bounded value function. The detailed lemma is presented as follows:

Lemma 22. *For a generic reward function r', any two transition kernels P' and P'' and any context w, we can obtain:*

$$\bar{\bar{V}}^{\pi_w}_{\mathcal{M}^{(r',P')}(w)} - V^{\pi_w}_{\mathcal{M}^{(r',P'')}(w)} \leq V^{\pi_w}_{\mathcal{M}^{(g^n,P'')}(w)},$$

where $g^n_h(s, a, w) := (P'_{h,w} - P''_{h,w})\bar{\bar{V}}^{\pi_w}_{h+1,\mathcal{M}^{(r',P')}}(s, a)$.

Proof. For any context w, we have:

$$\begin{aligned}
&\bar{\bar{V}}^{\pi_w}_{\mathcal{M}^{(r',P')}(w)} - V^{\pi_w}_{\mathcal{M}^{(r',P'')}(w)} \\
&\overset{(i)}{\leq} \mathop{\mathbb{E}}_{\pi_w}\left[P'_{1,w}\bar{\bar{V}}^{\pi_w}_{2,\mathcal{M}^{(r',P')}(w)}(s_1,a_1) - P''_{1,w}V^{\pi_w}_{2,\mathcal{M}^{(r',P'')}(w)}(s_1,a_1)\right] \\
&= \mathop{\mathbb{E}}_{\pi_w}\left[\left(P'_{1,w} - P''_{1,w}\right)\bar{\bar{V}}^{\pi_w}_{2,\mathcal{M}^{(r',P')}(w)}(s_1,a_1) + P''_{1,w}\left(\bar{\bar{V}}^{\pi_w}_{2,\mathcal{M}^{(r',P')}(w)} - V^{\pi_w}_{2,\mathcal{M}^{(r',P'')}(w)}\right)(s_1,a_1)\right] \\
&= \mathop{\mathbb{E}}_{\pi_w}\left[g^n_1(s_1,a_1,w) + P''_{1,w}\left(\bar{\bar{V}}^{\pi_w}_{2,\mathcal{M}^{(r',P')}(w)} - V^{\pi_w}_{2,\mathcal{M}^{(r',P'')}(w)}\right)(s_1,a_1)\right] \\
&\overset{(ii)}{\leq} \mathop{\mathbb{E}}_{(s_h,a_h)\sim(P''_w,\pi_w)}\left[\sum_{h=1}^{H} g^n_h(s_h,a_h,w)\right] = V^{\pi_w}_{\mathcal{M}^{(g^n,P'')}(w)},
\end{aligned}$$

where (i) follows from the definition in eq. (11) and (ii) follows by iteratively extracting the terms $g^n_h(s_h, a_h, w)$ from the value function gaps. □

The following lemma (Dann et al., 2017) provides an expression on the difference of two value functions under different MDPs.

Lemma 23. *(Simulation Lemma). Suppose P' and P'' are transition kernels of two MDPs, and r', r'' are the corresponding reward functions. Given any policy π, we have :*

$$\begin{aligned}
&V^{\pi}_{h,P',r'}(s_h) - V^{\pi}_{h,P'',r''}(s_h) \\
&\quad= \sum_{h'=h}^{H} \mathop{\mathbb{E}}_{\substack{s_{h'}\sim(P'',\pi)\\ a_{h'}\sim\pi}} \left[r'\left(s_{h'},a_{h'}\right) - r''\left(s_{h'},a_{h'}\right) + \left(P'_{h'} - P''_{h'}\right)V^{\pi}_{h'+1,P',r'}\left(s_{h'},a_{h'}\right) \mid s_h\right] \\
&\quad= \sum_{h'=h}^{H} \mathop{\mathbb{E}}_{\substack{s_{h'}\sim(P',\pi)\\ a_{h'}\sim\pi}} \left[r'\left(s_{h'},a_{h'}\right) - r''\left(s_{h'},a_{h'}\right) + \left(P'_{h'} - P''_{h'}\right)V^{\pi}_{h'+1,P'',r''}\left(s_{h'},a_{h'}\right) \mid s_h\right].
\end{aligned}$$

We next present a widely-used lemma for linear MDPs here, which is Lemma G.2 in Agarwal et al. (2020) and Lemma 10 in Uehara et al. (2021).

Lemma 24. *(Elliptical Potential Lemma) Consider a sequence of $d \times d$ positive semidefinite matrices $X_1, ..., X_N$ satisfying $\mathrm{tr}(X_n) \leq 1$ for any $n \in [N]$. Define $M_0 = \lambda_0$ and $M_n = M_{n-1} + X_n$. Then we have:*

$$\sum_{n=1}^{N} \mathrm{tr}\left(X_n M_{n-1}^{-1}\right) \leq 2\log\det\left(M_N\right) - 2\log\det\left(M_0\right) \leq 2d\log\left(1 + \frac{N}{d\lambda_0}\right).$$