Sample and Oracle Efficient Reinforcement Learning for MDPs with Linearly-Realizable Value Functions

Zakaria Mhammedi MHAMMEDI@GOOGLE.COM
Google Research, NYC

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Designing sample-efficient and computationally feasible reinforcement learning (RL) algorithms is particularly challenging in environments with large or infinite state and action spaces. In this paper, we advance this effort by presenting an efficient algorithm for Markov Decision Processes (MDPs) where the state-action value function of any policy is linear in a given feature map. This challenging setting can model environments with infinite states and actions, strictly generalizes classic linear MDPs, and currently lacks a computationally efficient algorithm under online access to the MDP. Specifically, we introduce a new RL algorithm that efficiently finds a near-optimal policy in this setting, using a number of episodes and calls to a cost-sensitive classification (CSC) oracle that are both polynomial in the problem parameters. Notably, our CSC oracle can be efficiently implemented when the feature dimension is constant, representing a clear improvement over state-of-the-art methods, which require solving non-convex problems with horizon-many variables and can incur computational costs that are exponential in the horizon.

Keywords: Reinforcement learning, linear function approximation, linear Q^π.

1. Introduction

The field of reinforcement learning (RL) is advancing rapidly, making the development of sample-efficient algorithms increasingly important as the dimensionality of modern problems continues to grow. Traditional RL methods used in practice often lack sample complexity guarantees, meaning there are no bounds on the number of interactions with the environment needed to find a near-optimal policy. Designing RL algorithms with provable guarantees is particularly challenging, especially in applications involving large state and action spaces. While theoretical algorithms that offer such guarantees exist, they often rely on strong assumptions about the environment and are typically computationally infeasible for practical use. In this paper, we take a step toward bridging this gap by focusing on a classical environment assumption that has been widely studied in recent years in the RL context, yet has lacked computationally efficient algorithms.

Specifically, we consider MDPs where the state-action value functions of any policy are linear in some known feature map, which we refer to as linearly Q^π-realizable MDPs (Lattimore et al., 2020; Du et al., 2020). This assumption is particularly interesting because it does not impose direct constraints on the dynamics of the MDP, unlike classic linear MDPs (Jin et al., 2020; Yang and Wang, 2019, 2020), where the transition operator is assumed to have a low-rank structure. However, with less structure to exploit, learning in linearly Q^π-realizable MDPs is significantly more challenging. Consequently, these MDPs have primarily been studied in online RL with state resetting (Hao et al., 2022; Weisz et al., 2021a; Li et al., 2021; Yin et al., 2022; Weisz et al., 2022; Mhammedi et al., 2024), which is a setting where the agent has access to a local simulator allowing it to revisit previously encountered state-action pairs. While sample- and computationally efficient algorithms

are known in the local simulator setting, it was not until recently that Weisz et al. (2024) developed a sample-efficient algorithm for linearly Q^π-realizable MDPs without relying on a local simulator, though this approach is not computationally efficient. Weisz et al. (2024) left open the question of whether a computationally efficient algorithm could be developed for this setting.

Contributions. In this paper, we present an RL algorithm that finds a near-optimal policy in linearly Q^π-realizable MDPs using a number of episodes and calls to a cost-sensitive classification (CSC) oracle over policies that are both polynomial in the problem parameters. Additionally, we show that, due to the nature of the policy class involved, our CSC oracle can be efficiently implemented when the feature dimension is *constant*.[1] This represents a clear improvement over the algorithm by Weisz et al. (2024), which relies on *global optimism* and requires solving non-convex optimization problems with horizon-many variables that can lead to a computational cost exponential in the horizon.

Although cost-sensitive classification is NP-hard in the worst case, it can be reduced to binary classification (Beygelzimer et al., 2009; Langford and Beygelzimer, 2005), a problem for which many practical algorithms are available and which forms the foundation of empirical machine learning. It is also worth noting that certain algorithms based on global optimism such as `OLIVE` (Jiang et al., 2017) (which is similar to the algorithm in (Weisz et al., 2024)) are known to be incompatible with oracle-efficient implementations for various common RL oracles, including the CSC oracle considered in this paper. This highlights a separation between these computationally intractable algorithms and our approach, which can be implemented in an oracle-efficient manner.

This paper extends the ongoing exploration of structural assumptions that allow for both sample- and computationally-efficient algorithms in RL. We believe our techniques will inspire more efficient algorithms for planning and RL with general function approximation, beyond the paper's setting.

Paper organization. The remainder of this paper is organized as follows. In Section 2, we formally introduce the setup considered in this paper, along with preliminary results and a high-level overview of our approach. Section 3 presents the main result, focusing on the sample and computational complexity of our algorithm, `Optimistic-PSDP`. In Section 4, we explore the algorithm in detail, providing both high-level descriptions and key insights into its design. An extended discussion of the algorithm design and challenges is provided in Appendix C. Appendix D provides a proof sketch of the main guarantee of `Optimistic-PSDP`. Related work and proofs are deferred to the appendix.

2. Setup and Preliminaries

In Section 2.1, we outline the reinforcement learning (RL) setting considered in this paper and introduce the notation used throughout. In Section 2.2, we present key structural results for linearly Q^π-realizable MDPs, which are essential to our analysis, along with a high-level overview of our approach. In Section 2.3, we describe our computational oracle and its properties. Finally, Section 5 discusses limitations of our approach and outlines directions for future work.

2.1. Online Reinforcement Learning in Linearly Realizable MDPs

A Markov Decision Process (MDP) is a tuple $\mathcal{M} = (\mathcal{X}, \mathcal{A}, P, \rho_{\text{init}}, R, H)$, where $\mathcal{X}$ is a large or potentially infinite state space, $\mathcal{A}$ is the action space, which we assume is finite and abbreviate $A = |\mathcal{A}|$, $H \in \mathbb{N}$ is the horizon, $R : \mathcal{X} \times \mathcal{A} \to [0,1]$ is the reward function, $P : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$

1. This means that the computational complexity is polynomial in all parameters except for the feature dimension, where it can be exponential.

is the transition distribution, and $\rho_{\text{init}} \in \Delta(\mathcal{X})$ is the initial state distribution. A (Markovian) policy π is a map $\pi : \mathcal{X} \to \Delta(\mathcal{A})$; we use Π to denote the set of all such maps. When a policy is executed, it generates a trajectory $(\boldsymbol{x}_1, \boldsymbol{a}_1, \boldsymbol{r}_1), \ldots, (\boldsymbol{x}_H, \boldsymbol{a}_H, \boldsymbol{r}_h)$ via the process $\boldsymbol{a}_h \sim \pi(\boldsymbol{x}_h), \boldsymbol{r}_h \sim R(\boldsymbol{x}_h, \boldsymbol{a}_h), \boldsymbol{x}_{h+1} \sim P(\cdot \mid \boldsymbol{x}_h, \boldsymbol{a}_h)$, initialized from $\boldsymbol{x}_1 \sim \rho_{\text{init}}$ (we use $\boldsymbol{x}_{H+1}$ to denote a terminal state with zero reward). We write $\mathbb{P}^\pi[\cdot]$ and $\mathbb{E}^\pi[\cdot]$ to denote the law and expectation under this process. We assume (following, e.g., (Jiang et al., 2017; Mhammedi et al., 2023; Weisz et al., 2024)) that the MDP $\mathcal{M}$ is *layered* so that $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_H$ for $\mathcal{X}_i \cap \mathcal{X}_j = \varnothing$ for all $i \neq j$, where $\mathcal{X}_h \subseteq \mathcal{X}$ is the subset of states in $\mathcal{X}$ that are reachable at layer h.[2]

The goal in an MDP is to find a policy $\widehat{\pi} \in \Pi$ such that

$$J(\pi^\star) - \mathbb{E}[J(\widehat{\pi})] \leq \varepsilon, \tag{1}$$

where $J(\pi) \coloneqq \mathbb{E}^\pi\left[\sum_{h=1}^H \boldsymbol{r}_h\right]$ is expected reward of policy π, and $\pi^\star \in \arg\max_{\pi \in \Pi} J(\pi)$ is the optimal policy. Since the state space $\mathcal{X}$ can be very large or even infinite, achieving the goal in (1) efficiently can be very challenging in general without any additional assumptions on the MDP. In this paper, we assume that the MDP is *linear Q^π-realizable.*

Linearly Q^π-realizable MDPs. To present our main assumption, we need to define the state-action value functions (or Q-functions); the Q-function for a policy $\pi \in \Pi$ at layer $h \in [H]$ is defined as

$$Q_h^\pi(x, a) \coloneqq \mathbb{E}^\pi\left[\sum_{\ell=h}^H \boldsymbol{r}_\ell \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right], \tag{2}$$

for all $(x, a) \in \mathcal{X}_h \times \mathcal{A}$. Our main assumption in this paper asserts that the Q-functions are *linear* with respect to some *known* feature map $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$.

Assumption 2.1 (Linear-Q^π)**.** *$\exists c > 0$: $\forall h \in [H]$, $\forall \pi \in \Pi$, $\exists \theta_h^\pi \in \mathbb{B}_d(cH) \subseteq \mathbb{R}^d$ such that*

$$\forall (x, a) \in \mathcal{X}_h \times \mathcal{A}, \quad Q_h^\pi(x, a) = \phi(x, a)^\top \theta_h^\pi. \tag{3}$$

Furthermore, the feature map ϕ is known to the agent and satisfies $\|\phi(x, a)\| \leq 1$, for all $(x, a) \in \mathcal{X} \times \mathcal{A}$. We take $c = 1$ for simplicity and to avoid carrying around an extra parameter.

Results of this paper can easily be extended to the setting where the linear Q^π assumption holds approximately see, e.g., Weisz et al. (2024). However, to simplify the presentation, we focus on the exact linear Q^π assumption in Assumption 2.1.

We refer to an MDP that satisfies Assumption 2.1 as *linearly Q^π-realizable*. Linearly Q^π-realizable MDPs are particularly interesting because, unlike many other common assumptions in the RL context, the assumption in (3) (or its approximate version) does not directly set any constraints on the dynamics of the MDP (which are captured by the transition operator P). For instance, in classic linear MDPs, the transition operator P is typically assumed to have a low-rank structure.

2. On its own, the layered assumption comes with no loss of generality as one can always augment the state by adding information about the current layer. However, for linearly Q^π-realizable MDPs (the main assumption we make in this paper), layering *does* comes with some loss generality. Tackling linearly Q^π-realizable MDPs without the layering assumption is an interesting open problem.

Online Reinforcement Learning. To achieve the goal in (1), we consider the standard *online* reinforcement learning framework, where the agent/algorithm repeatedly interacts with an *unknown* MDP, where the transition operator P and the reward function R are unknown, by executing a policy and observing the resulting trajectory, with the goal of maximizing the total reward. Formally, for each episode $t \in [N_{\mathtt{episodes}}]$, the agent selects a policy $\pi^{(t)} = \{\pi_h^{(t)}\}_{h=1}^H$, executes it in the underlying MDP $\mathcal{M}$ and observes the trajectory $\{(\boldsymbol{x}_h^{(t)}, \boldsymbol{a}_h^{(t)}, \boldsymbol{r}_h^{(t)})\}_{h=1}^H$. The goal is to achieve (1) with as few episodes of interaction with the MDP as possible; we say that an algorithm is *sample-efficient* if the number of trajectories it requires to find $\widehat{\pi}$ satisfying (1) under Assumption 2.1 is $N_{\mathtt{episodes}} = \mathrm{poly}(d, A, H, \varepsilon^{-1})$ (where d is the dimension of the feature ϕ) with no dependence on the state space $\mathcal{X}$, which can be very large or infinite.

As mentioned earlier, Weisz et al. (2024) has already developed a sample-efficient, though not computationally efficient, algorithm for linearly Q^π-realizable MDPs. In this paper, we introduce a sample-efficient algorithm for this setting that makes a polynomial number of calls to a cost-sensitive classification oracle. Moreover, we show that this oracle can be implemented efficiently when the feature dimension is constant. The results of this paper bring us one step closer to understanding the computational complexity of sample-efficient online RL in linearly Q^π-realizable MDPs.

Notation. In what follows, we let $V_h^\pi(x) \coloneqq \mathbb{E}^\pi\big[\sum_{\ell=h}^H \boldsymbol{r}_\ell \mid \boldsymbol{x}_h = x\big]$ denote the state value function at layer $h \in [H]$. We denote by $\pi^\star$ the optimal deterministic policy that maximizes $Q^{\pi^\star}$, and write $Q^\star \coloneqq Q^{\pi^\star}$ and $V^\star \coloneqq V^{\pi^\star}$. We denote by

$$\Theta_h \coloneqq \{\theta_h^\pi \in \mathbb{R}^d \mid \pi \in \Pi\}, \tag{4}$$

the set of all parameter vectors corresponding to state-action value functions, where θ_h^π is as in Assumption 2.1. To simplify notation, we let $\varphi(\cdot, a, a') \coloneqq \phi(\cdot, a) - \phi(\cdot, a')$, for $a, a' \in \mathcal{A}$, and for any matrix $W \in \mathbb{R}^{d\times d}$ define:

$$\varphi(x; W) = W(\phi(x, a_x) - \phi(x, a'_x)), \quad \text{with} \quad (a_x, a'_x) \in \operatorname*{arg\,max}_{(a,a')\in\mathcal{A}^2} \|W(\phi(x, a) - \phi(x, a'))\|. \tag{5}$$

Argmax tie-breaking. Whenever we write $\pi(\cdot) = \arg\max_{a\in\mathcal{A}} \phi(\cdot, a)^\top\theta$, for some $\theta \in \mathbb{R}^d$, ties in the argmax are broken by choosing the action with the smallest index.

Additional notation. For any $m, n \in \mathbb{N}$, we denote by $[m\,..\,n]$ the integer interval $\{m, \dots, n\}$. For any sequence of objects $o_1, o_2, \dots$, we define $o_{m:n} \coloneqq (o_i)_{i\in[m..n]}$. We also let $[n] \coloneqq [1\,..\,n]$. We use $\widetilde{O}(\cdot)$ to denote a bound up to factors polylogarithmic in parameters appearing in the expression. We use the notation $a \propto b$ to mean that there are absolute constants $c, C > 0$ such that $ca \le b \le Ca$. We define $\pi_{\mathtt{unif}} \in \Pi$ as the random policy that selects actions in $\mathcal{A}$ uniformly. For $1 \le t \le h \le H$ and any pair of policies $\pi, \pi' \in \Pi$, we define $\pi \circ_t \pi' \in \Pi$ as the policy given by $(\pi \circ_t \pi')(x_\ell) = \pi(x_\ell)$ for all $\ell < t$ and $(\pi \circ_t \pi')(x_\ell) = \pi'(x_\ell)$ for all $\ell \in [t\,..\,H]$. For the analysis only, we let $\boldsymbol{x}_0$ denote a fictitious state such that $\phi(\boldsymbol{x}_0, a) = 0$, for all $a \in \mathcal{A}$.

We use $\|\cdot\|$ to denote the Euclidean norm in $\mathbb{R}^d$, and let $\mathbb{B}_d(r) \subseteq \mathbb{R}^d$ denote the Euclidean ball of radius r; we drop the d subscript when the dimension is clear from the context. We let $\|\cdot\|_{\mathrm{op}}$ denote the matrix Operator norm. For a positive semi-definite matrix $A \in \mathbb{R}^{d\times d}$, we write $\|x\|_A \coloneqq \sqrt{x^\top Ax}$ for $x \in \mathbb{R}^d$. We denote by $\mathbb{S}_{++}^{d\times d}$ the set of positive definite matrices in $\mathbb{R}^{d\times d}$. For a symmetric matrix $A \in \mathbb{R}^{d\times d}$ and $c > 0$, we denote by $\mathcal{S}(A, c)$ the linear subspace spanned by the eigenvectors of A corresponding to eigenvalues that are at least c. Given a subspace $S \subseteq \mathbb{R}^d$, we denote by

$\mathrm{Proj}_S : \mathbb{R}^d \to S$ the orthogonal projection operator onto S. We use $\dagger$ to denote the Moore-Penrose pseudo-inverse. Finally, we use $\otimes$ to denote the tensor product of vectors/matrices; for a vector $v \in \mathbb{R}^d$ and matrix $M \in \mathbb{R}^{d\times d}$, we let $v \otimes M \in \mathbb{R}^{d\times d\times d}$ be the tensor that satisfies $(v \otimes M)_{i,j,k} = v_i M_{j,k}$.

2.2. Background and High-Level Overview of the Approach

In this section, we introduce key concepts and structural results for linearly Q^π-realizable MDPs. Most of the results presented here are based on the works of (Weisz et al., 2024; Tkachuk et al., 2024). The proofs for these statements can be found in Appendix G. We also provide a high-level overview of our approach along the way. Central to our analysis is the notation of *range of states*.

Definition 2.1 (Range of state). *For $h \in [H]$, the range of a state $x \in \mathcal{X}_h$ is defined as*

$$\mathtt{Rg}(x) \coloneqq \sup_{a,a'\in\mathcal{A}} \sup_{\theta_h\in\Theta_h} \langle \phi(x,a) - \phi(x,a'), \theta_h \rangle, \tag{6}$$

where we recall that Θ_h is the set of all parameter vectors corresponding to Q-functions; see (4).

The key insight from Weisz et al. (2024) is that states with low range, such as those where $\mathtt{Rg}(x) \le O(\varepsilon)$, are not particularly significant when it comes to learning an $O(\varepsilon)$-optimal policy in a linearly Q^π-realizable MDP. More formally, for any two policies π and π', if we define $\tilde{\pi}(\cdot) = \mathbb{I}\{\mathtt{Rg}(\cdot) \ge \varepsilon\} \cdot \pi(\cdot) + \mathbb{I}\{\mathtt{Rg}(\cdot) < \varepsilon\} \cdot \pi'(\cdot)$, then $J(\pi) \le J(\tilde{\pi}) + O(\varepsilon)$. This means that the actions a policy takes in low-range states have minimal impact. What is more, Weisz et al. (2024); Tkachuk et al. (2024) show that if one has access to the range function $\mathtt{Rg}$, there is a way in which one can "skip" over low-range states to essentially reduce a linearly Q^π-realizable MDP to a linear MDP.

The challenge, of course, is that the range function $\mathtt{Rg}$ is *not* known to the agent, and we need to use some proxy for it. To introduce the proxy we use in this paper, we must first define a few key concepts. We begin with the notion of the *design range* of a state, which restricts the supremum over θ in (6) to an approximate design for the set Θ_h.

Definition 2.2 (Approximate design). *Let $\mathcal{C} = \{c^z\}_{z\in\mathcal{Z}} \subseteq \mathbb{R}^d$ be a set indexed by an abstract set $\mathcal{Z}$. A distribution $\rho \in \Delta(\mathcal{Z})$ such that $|\mathrm{supp}\,\rho| < \infty$ is an approximate optimal design for $\mathcal{C}$ if*

$$\sup_{c\in\mathcal{C}} \|c\|^2_{\mathrm{G}(\rho)^\dagger} \le 2d, \quad \textit{where} \quad \mathrm{G}(\rho) \coloneqq \sum_{z\in\mathrm{supp}\,\rho} \rho(z) c^z (c^z)^\top.$$

Approximate design for Θ_h. For $h \in [H]$, let $\rho_h \in \Delta(\Pi)$ be an approximate optimal design for $\Theta_h = \{\theta^\pi_h \mid \pi \in \Pi\}$ with $|\mathrm{supp}\,\rho_h| \le \tilde{d} \coloneqq 4d \log\log d + 16$; such a distribution ρ_h is guaranteed to exist by (Todd, 2016, Part (ii) of Lemma 3.9). For the rest of the paper, we fix such an approximate design ρ_h. With this, we now define the notation of design range.

Definition 2.3 (Design range). *Let $h \in [H]$ and $\rho_h \in \Delta(\Pi)$ be the approximate optimal design just defined. The design range of $x \in \mathcal{X}_h$ is defined as*

$$\mathrm{Rg}^{\mathtt{D}}(x) \coloneqq \sup_{a,a'\in\mathcal{A}} \max_{\pi\in\mathrm{supp}\,\rho_h} \langle \phi(x,a) - \phi(x,a'), \theta^\pi_h \rangle.$$

The design range $\mathrm{Rg}^{\mathtt{D}}(\cdot)$ provides a good approximation of the range $\mathtt{Rg}(\cdot)$ in the sense that $\mathrm{Rg}^{\mathtt{D}}(\cdot) \propto \mathtt{Rg}(\cdot)$. In fact, we trivially have that $\mathrm{Rg}^{\mathtt{D}}(\cdot) \le \mathtt{Rg}(\cdot)$. And, as the next lemma shows, we also have $\mathtt{Rg}(x) \le \sqrt{2d} \cdot \mathrm{Rg}^{\mathtt{D}}(x)$.

Lemma 2.1 (Restatement of Proposition 4.5 in (Weisz et al., 2024))**.** *For $h \in [H]$ and $x \in \mathcal{X}_h$, $\mathsf{Rg}(x) \le \sqrt{2d} \cdot \mathrm{Rg}^{\mathsf{D}}(x)$.*

We say that a function is *admissible* if it is dominated by the design range.

Definition 2.4 (Admissibility)**.** *For $h \in [H]$ and $\alpha > 0$, a function $F : \mathcal{X}_h \to \mathbb{R}$ is α-admissible if for all $x \in \mathcal{X}_h$, $F(x) \le \mathrm{Rg}^{\mathsf{D}}(x)/\alpha$.*

Admissibility plays a crucial role in the analysis of our algorithm. The following lemma, which restates (Weisz et al., 2024, Lemma 4.1), shows that the conditional expectation of an admissible function is *linear* in the feature map ϕ. While in a linear MDP the conditional expectation of any function is linear in ϕ, it is not always the case for linearly Q^π-realizable MDPs (Weisz et al., 2024).

Lemma 2.2 (Admissible realizability)**.** *For $h \in [H]$ and $\alpha > 0$, if $f : \mathcal{X}_h \to \mathbb{R}$ is α-admissible (Definition 2.4), then for all $\ell \in [h-1]$ and $\tilde{\pi} \in \Pi$, there exists $\theta \in \mathbb{B}(4\tilde{d}H/\alpha)$, where $\tilde{d} \coloneqq 4d \log\log d + 16$, such that for all $(x, a) \in \mathcal{X}_\ell \times \mathcal{A}$,*

$$\mathbb{E}^{\tilde{\pi}}[f(\boldsymbol{x}_h) \mid \boldsymbol{x}_\ell = x, \boldsymbol{a}_\ell = a] = \phi(x, a)^\top \theta.$$

Algorithms for the classical MDP setting crucially rely on the linearity of conditional expectations to efficiently learn a near-optimal policy; this linearity enables effective exploration and planning. Our algorithm will also need to leverage this linearity. The challenge in our setting is that only admissible functions are guaranteed to have linear conditional expectations as reflected in Lemma 2.2. The next lemma shows how any non-admissible function can essentially be turned into an admissible one by "ignoring" low-range states, when one has access to the design range function $\mathrm{Rg}^{\mathsf{D}}(\cdot)$.

Lemma 2.3. *Let $\ell \in [H]$, $L > 0$, $\gamma > 0$, and $f : \mathcal{X}_\ell \to [-L, L]$ be given, and let F be the function*

$$F(x) = \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(x) \ge \gamma\} \cdot f(x),$$

for all $x \in \mathcal{X}_\ell$. Then, F is α-admissible with $\alpha = \gamma/L$; that is, for all $x \in \mathcal{X}_\ell$, $F(x) \le L\mathrm{Rg}^{\mathsf{D}}(x)/\gamma$.

Fortunately, as discussed right after Definition 2.1 in the prequel, the cost of "ignoring" low-range states is minimal when it comes to finding a near-optimal policy. The remaining challenge is that we do not have direct access to the range function Rg^{D} and need to rely on a proxy.

Proxy for the range. To introduce our proxy for the range, we need one more concept.

Definition 2.5 (Valid preconditioning)**.** *For $\nu > 0$, a matrix W_h is a valid ν-preconditioning for layer $h \in [H]$ if there exists $k \in \mathbb{N}$ and vectors $(w_i)_{i \in [k]} \subset \mathbb{R}^d$ such that $W_h = \left(H^{-2} I + \sum_{i=1}^k w_i w_i^\top\right)^{-1/2}$ and for all $i \in [k]$*

$$\sup_{\theta \in \Theta_h} |\theta^\top w_i| \le 1, \qquad \left\| \left(H^{-2} I + \sum_{j \in [i-1]} w_j w_j^\top \right)^{-1/2} w_i \right\|^2 \ge \frac{1}{2}, \qquad \text{and} \qquad \|w_i\| \le \nu^{-1}.$$

As reflected by the statement of the next lemma, any valid preconditioning can be seen as parameterizing an ellipsoid that encapsulates the set Θ_h; the more accurately this ellipsoid approximates Θ_h, the better we can approximate $\mathrm{Rg}^{\mathsf{D}}(\cdot)$.

Lemma 2.4. *Let $h \in [H]$ be given. For $\nu > 0$, let $W_h \in \mathbb{R}^{d\times d}$ be a valid ν-preconditioning for layer $h \in [H]$ (Definition 2.5). Then, we have*

$$\sup_{\theta\in\Theta_h} \|\theta\|^2_{W_h^{-2}} = \sup_{\theta\in\Theta_h} \|W_h^{-1}\theta\|^2 \le 5d\log(1+16H^4\nu^{-4}).$$

In light of this, given a valid precondining matrix W_h and some small parameter $\mu > 0$, we will essentially use the function $\|\varphi(\cdot; W_h)\|$, where φ is as in (5), as a proxy for the design range Rg^{D}. The next lemma shows that a one-sided inequality always holds between the two.

Lemma 2.5. *Let $\ell \in [H]$ be given. For $\nu > 0$, let $W_\ell \in \mathbb{R}^{d\times d}$ be a valid ν-preconditioning for layer $\ell \in [H]$ (see Definition 2.5). Then, we have*

$$\forall x \in \mathcal{X}_\ell, \quad \mathrm{Rg}^{\mathsf{D}}(x) \le \sqrt{d_\nu}\cdot\|\varphi(x; W_\ell)\|,$$

where $d_\nu \coloneqq 5d\log(1+16H^4\nu^{-4})$ and $\varphi(\cdot; W_\ell)$ is as in (5).

Unfortunately, the reverse side of the inequality does not necessarily hold; if it did, meaning that $\|\varphi(\cdot; W_\ell)\| \propto \mathrm{Rg}^{\mathsf{D}}(\cdot)$, then for any bounded function f over $\mathcal{X}_\ell$, the map $\mathbb{I}\{\|\varphi(\cdot; W_\ell)\| \ge \mu\}\cdot f(\cdot)$ would be admissible (as can be shown through a similar proof as that of Lemma 2.3). However, even if $\|\varphi(\cdot; W_\ell)\|$ is not proportional to $\mathrm{Rg}^{\mathsf{D}}(\cdot)$, we can still use $\|\varphi(\cdot; W_\ell)\|$ as an effective proxy for the design range by adopting the following strategy, which we detail in Section 4.1:

1. We use $\|\varphi(\cdot; W_\ell)\|$ as a proxy for the design range until we encounter a function f such that the map $g_f : x \mapsto \mathbb{I}\{\|\varphi(x; W_\ell)\| \ge \mu\}\cdot f(x)$ is not admissible. We can "witness" the non-admissibility of g_f if, for example, we identify a layer $h \in [\ell-1]$ and a policy π for which $\inf_\theta |\mathbb{E}^\pi[g_f(\boldsymbol{x}_\ell) - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top\theta]|$ is too large—this quantity should be small for an admissible function (see Lemma 2.2).
2. When we witness non-admissibility, we can compute a non-zero preconditioning vector w_ℓ (as detailed in Section 4.1) such that the new matrix $(W_\ell^{-2} + w_\ell w_\ell^\top)^{-1/2}$ is also a valid preconditioning matrix. We then use this new matrix in Step 1 and repeat the process.

The next lemma shows that we can only update the preconditioning matrix at most $\widetilde{O}(d)$ times; in other words, non-admissibility can only be witnessed $\widetilde{O}(d)$ times. And, as long as non-admissibility is not observed for certain functions of interest (such as bonus functions in our case), $\|\varphi(\cdot; W_h)\|$ will serve as a reliable proxy for the design range $\mathrm{Rg}^{\mathsf{D}}(\cdot)$.

Lemma 2.6 (Length of a preconditioning)**.** *For $\nu > 0$, let $W_h \in \mathbb{R}^{d\times d}$ be a valid ν-preconditioning for layer $h \in [H]$, and let $k \in \mathbb{N}$ be the length of the corresponding sequence $(w_i)_{i\in[k]}$; see Definition 2.5. Then, $k \le 4d\log(1+16\nu^{-4}H^4)$.*

2.3. Benchmark Policies and Computational Oracle

In this subsection, we describe the computational oracle our algorithm requires. For this, we need to introduce a class of benchmark policies.

Benchmark policies. We consider the set Π_{Bench} of benchmark policies such that $\pi \in \Pi_{\mathsf{Bench}}$ if and only if there exist $\gamma > 0$, $\pi', \pi'' \in \Pi_{\mathsf{Base}} \coloneqq \{x \mapsto \pi(x;\theta) = \arg\max_{a \in \mathcal{A}} \theta^\top \phi(x,a) \mid \theta \in \mathbb{B}(H)\}$ and $\theta_1, \dots, \theta_{\tilde{d}} \in \mathbb{B}(H)$ with $\tilde{d} = 4d \log\log d + 16$ such that

$$\pi(\cdot) = \mathbb{I}\left\{\max_{a,a' \in \mathcal{A}, i \in [\tilde{d}]} \varphi(\cdot, a, a')^\top \theta_i \geq \gamma\right\} \cdot \pi'(\cdot) + \mathbb{I}\left\{\max_{a,a' \in \mathcal{A}, i \in [\tilde{d}]} \varphi(\cdot, a, a')^\top \theta_i < \gamma\right\} \cdot \pi''(\cdot), \quad (7)$$

where $\varphi(\cdot, a, a') \coloneqq \phi(\cdot, a) - \phi(\cdot, a')$. Note that from the definition of the design range in Definition 2.3, there exist $\theta_1, \dots, \theta_{\tilde{d}} \in \mathbb{B}(H)$ such that $\mathrm{Rg}^{\mathsf{D}}(\cdot) = \max_{a,a' \in \mathcal{A}, i \in [\tilde{d}]} \langle \phi(\cdot, a) - \phi(\cdot, a'), \theta_i \rangle$, and so

$$\forall \pi', \pi'' \in \Pi_{\mathsf{Base}}, \forall \gamma > 0, \quad \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\cdot) < \gamma\} \cdot \pi'(\cdot) + \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\cdot) \geq \gamma\} \cdot \pi''(\cdot) \in \Pi_{\mathsf{Bench}}. \quad (8)$$

The *growth function* of the policy class Π_{Bench} at layer $h \in [H]$ is defined as

$$\mathcal{G}_h(\Pi_{\mathsf{Bench}}, n) \coloneqq \sup_{(x^{(1)}, \dots, x^{(n)}) \in \mathcal{X}_h^n} \left|\left\{(\pi(x^{(1)}), \dots, \pi(x^{(n)})) \mid \pi \in \Pi_{\mathsf{Bench}}\right\}\right|.$$

The growth function of Π_{Bench} is defined as $\mathcal{G}(\Pi_{\mathsf{Bench}}, n) \coloneqq \max_{h \in [H]} \mathcal{G}_h(\Pi_{\mathsf{Bench}}, n)$. The growth function is a key concept in the study of statistical generalization (Mohri et al., 2012). Since we will be using the benchmark class Π_{Bench} to learn a good policy, it is essential to bound the growth function of Π_{Bench} to ensure that our algorithm is sample efficient. The next lemma shows that the logarithm of the growth function of Π_{Bench} is polynomial in the problem parameters, which is sufficient to meet our sample efficiency requirements.

Lemma 2.7. *For any $n \in \mathbb{N}$, the growth function $\mathcal{G}(\Pi_{\mathsf{Bench}}, n)$ is at most $(9^2 n A^2/d)^{(d+1)^2}$.*

We now describe our computational oracle and how it uses the class Π_{Bench}.

Computational oracle. Our algorithm requires a *Cost-Sensitive Classification* (CSC) oracle over policies in Π_{Bench} such that given any $n \in \mathbb{N}$, $h \in [H]$, and $(c^{(1)}, x^{(1)}, a^{(1)}), \dots, (c^{(n)}, x^{(n)}, a^{(n)}) \in \mathbb{R} \times \mathcal{X}_h \times \mathcal{A}$, the oracle returns

$$\pi' \in \arg\min_{\pi \in \Pi_{\mathsf{Bench}}} \sum_{i=1}^n c^{(i)} \cdot \mathbb{I}\{\pi(x^{(i)}) = a^{(i)}\}.$$

Although cost-sensitive classification is NP-hard in the worst case (Dann et al., 2018), it can be simplified to binary classification (Beygelzimer et al., 2009; Langford and Beygelzimer, 2005), a problem for which numerous practical algorithms are available, forming a foundation of empirical machine learning. Moreover, the CSC oracle has been successfully used in practical algorithms for contextual bandits, imitation learning, and structured prediction (Langford and Zhang, 2007; Agarwal et al., 2014; Ross and Bagnell, 2014; Chang et al., 2015). Further, when the feature dimension is constant, we show that our CSC oracle can be efficiently implemented.

Lemma 2.8 (Restates Lemma M.1)**.** *Let $n \in \mathbb{N}$, $h \in [H]$, and $(c^{(1)}, x^{(1)}, a^{(1)}), \dots, (c^{(n)}, x^{(n)}, a^{(n)}) \in \mathbb{R} \times \mathcal{X}_h \times \mathcal{A}$ be given. Then, for the benchmark policy class Π_{Bench} in Section 2.3, it is possible to find $\pi' \in \arg\min_{\pi \in \Pi_{\mathsf{Bench}}} \sum_{i=1}^n c^{(i)} \cdot \mathbb{I}\{\pi(x^{(i)}) = a^{(i)}\}$, in $O(\mathrm{poly}(n, d, A) \cdot (9n^2A^2/d)^{(d+1)^2})$ time.*

3. Main Result: Guarantee of `Optimistic-PSDP`

The following is the main guarantee of `Optimistic-PSDP` (the proof is in Appendix L).

Theorem 3.1. *Let $\varepsilon, \delta \in (0,1)$ be given and consider a call to* `Optimistic-PSDP`$(\Pi_{\mathsf{Bench}}, \varepsilon, \delta)$ *(Algorithm 1) with Π_{Bench} as in Section 2.3. Then, with probability at least $1-2\delta$, we have*

$$J(\pi^\star) - J(\widehat{\pi}_{1:H}) \le \varepsilon,$$

where $\widehat{\pi}_{1:H}$ is the policy returned by Algorithm 1. The # of episodes Algorithm 1 uses is at most $\mathrm{poly}(A, H, d, 1/\varepsilon, \log(1/\delta))$ and the # of calls to the CSC oracle in Section 2.3 is at most $\widetilde{O}(d^2H^6)$.

As desired, the sample complexity of `Optimistic-PSDP` is polynomial in the problem parameters. However, we note that our approach incurs a $\mathrm{poly}(A)$ factor in the sample complexity compared to the non-computationally efficient method of Weisz et al. (2024), which is based on global optimism. It is unclear if this factor can be eliminated when using a local optimism-based approach like ours.

Oracle complexity. As mentioned earlier, our algorithm requires calls to a CSC oracle (see Section 2.3 for the oracle's definition). As stated in Theorem 3.1, the number of oracle calls made by `Optimistic-PSDP` does not depend on the desired suboptimality ε; it does not grow with $O(1/\varepsilon)$.

Computational complexity and practicality. We now revisit a few key points regarding the complexity and practicality of the CSC oracle (see Section 2.3):

- The policy optimization oracle we require can be implemented efficiently when the feature dimension is constant (see Lemma 2.8).
- While implementing the oracle is NP-hard in general (for non-constant feature dimension), it can be reduced to binary classification, allowing the use of well-established machine learning algorithms (see Section 2.3).

Comparison to previous algorithms. Note that our result strictly improves on those of Weisz et al. (2024) in terms of computational complexity. The algorithm of Weisz et al. (2024) relies on global optimism, which involves solving non-convex optimization problems in $\mathbb{R}^{dH}$, leading to a computational complexity exponential in the horizon in the worst case. In contrast, the computational complexity of `Optimistic-PSDP` is polynomial in the horizon (see Theorem 3.1 and Lemma 2.8).

It is also worth noting that certain algorithms based on global optimism such as `OLIVE` (Jiang et al., 2017) (which is similar to the algorithm in (Weisz et al., 2024)) are known to be incompatible with oracle-efficient implementation for various common RL oracles, including the CSC oracle considered in this paper (see Dann et al. (2018)). Therefore, Theorem 3.1 separates these computationally intractable algorithms from our algorithm. We leave open the question of whether an algorithm can be developed that is computationally efficient without relying on a computational oracle.

4. Algorithm and Intuition

In this section, we describe our algorithm, `Optimistic-PSDP`, and provide some intuition behind its design. An extended version of this discussion, including the key challenges addressed by our solution, can be found in Appendix C.1.

Our main algorithm, `Optimistic-PSDP` (Algorithm 1), builds on the classical Policy Search by Dynamic Programming (PSDP) algorithm (see, e.g., Bagnell et al. (2003)) by incorporating bonuses

Algorithm 1 `Optimistic-PSDP`: Optimistic Policy Search by Dynamic Programming.

input: Policy class Π', suboptimality $\varepsilon \in (0,1)$, confidence $\delta \in (0,1)$.

initialize: $U_h^{(1)} \leftarrow 0$, $W_h^{(1)} \leftarrow H^{-2} I$, $\Psi_h^{(1)} \leftarrow (\pi_{\mathtt{unif}}, 0)$ for all $h \in [0\,..\,H]$, $\tau' \leftarrow 1$, $J \leftarrow 0$.

Set parameters T, $n_{\mathtt{traj}}$, μ, ν, λ, and β as in (63) in Appendix F.

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

`/* Fit optimistic values in a dynamic programming fashion. */`

  for $h = H, \ldots, 1$ **do**

    Update $(\hat{\theta}_h^{(t)}, w_{h+1:H}^{(t,h)}) \leftarrow \mathtt{FitValue}_h(\Psi_{h-1}^{(t)}, \widehat{\pi}_{h+1:H}^{(t)}, U_{h+1:H}^{(t)}, W_{h+1:H}^{(t)}; \Pi', \mu, \nu, \lambda, \beta, \varepsilon, \delta, n_{\mathtt{traj}})$.

    Update $W_\ell^{(t)} \leftarrow ((W_\ell^{(t)})^{-2} + w_\ell^{(t,h)} (w_\ell^{(t,h)})^\top)^{-1/2}$, for all $\ell \in [h+1\,..\,H]$.

    Set $\widehat{\pi}_h^{(t)}(\cdot) \leftarrow \arg\max_{a \in \mathcal{A}} \phi(\cdot, a)^\top \hat{\theta}_h^{(t)}$.

`/* Update the preconditioning matrices for the next iteration. */`

  Update $W_{1:H}^{(t+1)} \leftarrow W_{1:H}^{(t)}$.

`/* Identify policy that leads to "uncertain" states-actions and update design matrix. */`

  for $h = 1, \ldots, H$ **do**

    $(u_h^{(t)}, \tilde{\pi}_{1:h}^{(t)}, v_h^{(t)}) \leftarrow \mathtt{UncertainPolicy}_h(\Psi_{0:h-1}^{(t)}, \widehat{\pi}_{1:h}^{(t)}, U_h^{(t)}; \beta, n_{\mathtt{traj}})$.

    Set $U_h^{(t+1)} \leftarrow U_h^{(t)} + u_h^{(t)} (u_h^{(t)})^\top$ and $\Psi_h^{(t+1)} \leftarrow \Psi_h^{(t)} \cup \{(\tilde{\pi}_{1:h}^{(t)}, v_h^{(t)})\}$.

`/* Evalutating the policy` $\widehat{\pi}^{(t)}$ `. */`

  Compute $J^{(t)} \leftarrow \mathtt{Evaluate}(\widehat{\pi}_{1:H}^{(t)}, n_{\mathtt{traj}})$.

  if $J < J^{(t)}$ **then** set $J \leftarrow J^{(t)}$ and $\tau' \leftarrow t$.

return $\widehat{\pi}_{1:H} = \widehat{\pi}_{1:H}^{(\tau')}$.

into the rewards. In a nutshell, the algorithm learns a policy in a dynamic programming fashion by fitting optimistic value functions for each layer $h = H, \ldots, 1$. It is well known in RL, that adding the right bonuses helps in driving exploration, which is what we use them for in our algorithm. `Optimistic-PSDP` consists of three subroutines: `FitValue` (Algorithm 2), `UncertainPolicy` (Algorithm 4), and `Evaluate` (Algorithm 3). We present these subroutines in Appendix C.1 for space considerations, but a high-level description of their functionality is provided below.

Before diving into the specifics of `Optimistic-PSDP`, we provide an overview of its key variables.

Key Variables in `Optimistic-PSDP` (Algorithm 1). `Optimistic-PSDP` runs for $T = \widetilde{O}(d)$ iterations, where at each iteration $t \in [T]$, the algorithm maintains the following variables:

- $\Psi_h^{(t)} \subset \Pi \times \mathbb{R}^d$ consists of t policy-vector pairs. The subroutines $\mathtt{FitValue}_h$ and $\mathtt{UncertainPolicy}_h$ within `Optimistic-PSDP` use the policies in $\Psi_h^{(t)}$ to generate trajectories up to layer h. Ideally, $\Psi_h^{(t)}$ should contain policies that provide good coverage over the state-action space at layer h. Intuitively, the set $\Psi_h^{(t)}$ plays the role of a core set of policies (Agarwal et al., 2020; Zanette et al., 2021; Du et al., 2019; Wang et al., 2021).
- $U_h^{(t)}$ is a "design matrix" for layer h consisting of the sum of outer products of (truncated) expected features vectors $\bar{\phi}_h^{\pi,v} := \mathbb{E}^\pi[\mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \geq 0\} \cdot \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)]$, for $(\pi, v) \in \Psi_h^{(t)}$; that is, $U_h^{(t)} \approx \sum_{(\pi,v) \in \Psi_h^{(t)}} \bar{\phi}_h^{\pi,v} (\bar{\phi}_h^{\pi,v})^\top$. In the sequel (see Remark C.2), we will elucidate the role of the

indicator term $\mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \geq 0\}$ in the definition of $\bar{\phi}_h^{\pi,v}$. The matrix $U_h^{(t)}$ is used to define bonus functions that are added to the rewards (as in the next bullet point).

- $W_h^{(t)}$ is a valid preconditioning (Definition 2.5) for all t (with high probability). The matrix $W_h^{(t)}$ and the design matrix $U_h^{(t)}$ are used to define the bonus function $b_h^{(t)} : \mathcal{X}_h \to [0, H]$ as

$$b_h^{(t)}(\cdot) = \min\left(H, \frac{\varepsilon}{4H} \max_{a \in \mathcal{A}} \|\phi(\cdot, a)\|_{(\beta I + U_h^{(t)})^{-1}}\right) \cdot \mathbb{I}\{\|\varphi(\cdot; W_h^{(t)})\| \geq \mu\}, \tag{9}$$

  for some parameter μ and suboptimality ε.
- $\hat{\theta}_h^{(t)}$ is a parameter vector at layer h used to approximate the optimistic value function at layer h:

$$Q_h^{(t)}(x, a) \coloneqq Q_h^{\widehat{\pi}^{(t)}}(x, a) + \mathbb{E}^{\widehat{\pi}^{(t)}}\left[\sum_{\ell=h}^{H} b_\ell^{(t)}(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right]. \tag{10}$$

  Essentially, `Optimistic-PSDP` ($\texttt{FitValue}_h$, specifically) ensures that $\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}_h^{(t)} + b_h^{(t)}(\boldsymbol{x}_h) \approx Q_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h)$ in expectation under trajectories generated with policies in $\Psi_{h-1}^{(t)}$ (see (11) below).
- $\widehat{\pi}_h^{(t)}(\cdot) = \arg\max_{a \in \mathcal{A}} \phi(\cdot, a)^\top \hat{\theta}_h^{(t)}$ represents the policy at layer h in iteration t.

In each iteration $t \in [T]$, `Optimistic-PSDP` computes the policies $\widehat{\pi}_{1:H}^{(t)}$ in a dynamic programming fashion by fitting the optimistic value functions $(Q_h^{(t)})$ at each layer h. The subroutine `FitValue` Algorithm 2, which we describe next, is responsible for fitting these value functions.

4.1. `FitValue` (Algorithm 2)

In each iteration $t \in [T]$, starting from $h = H$ and progressing down to $h = 1$, `Optimistic-PSDP` invokes $\texttt{FitValue}_h$ with the input $(\Psi_{h-1}^{(t)}, \widehat{\pi}_{h+1:H}^{(t)}, U_{h+1:H}^{(t)}, W_{h+1:H}^{(t)})$, returning the pair $(\hat{\theta}_h^{(t)}, w_{h+1:H}^{(t,h)})$. The vectors $w_{h+1:H}^{(t,h)}$ are then used in Line 5 of Algorithm 1 to update the preconditioning matrices $(W_{h+1:H}^{(t)})$. The $\texttt{FitValue}_h$ subroutine ensures, with high probability, that $(W_h^{(t)})$ are valid preconditionings (see Lemma J.5 and recall the definition of a valid preconditioning in Definition 2.5). Consequently, using Lemma 2.6—which bounds the maximum length of a sequence of non-zero preconditioning vectors—it can be shown that $w_\ell^{(t,h)}$ is non-zero only on $\widetilde{O}(d)$ iterations; these are the iterations where the preconditioning matrix $W_\ell^{(t)}$ actually changes (see Line 5). Intuitively, $W_\ell^{(t)}$ does not update too frequently.

On iterations t where the preconditioning matrix is not updated (which is the case for most iterations as long as $T = \Omega(d \log d)$ is sufficiently large), $\texttt{FitValue}_h$ ensures that $\phi(\cdot, \cdot)^\top \hat{\theta}_h^{(t)}$ is a good approximation of the optimistic value function $Q_h^{(t)}$ in (10) in expectation under trajectories generated using policies in $\Psi_{h-1}^{(t)}$. More specifically, the subroutine $\texttt{FitValue}_h$ guarantees that on most iterations $t \in [T]$: for all $(\pi, v) \in \Psi_{h-1}^{(t)}$ and $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$ (see Lemma J.5),

$$\left|\mathbb{E}^{\pi \circ_h \tilde{\pi}}\left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\} \cdot (Q_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h) - \widehat{Q}_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h))\right]\right| \leq O\left(\frac{\varepsilon}{8TH\sqrt{d}}\right), \tag{11}$$

where $\widehat{Q}_h^{(t)}(x, a) = \phi(x, a)^\top \hat{\theta}_h^{(t)} + b_h^{(t)}(x)$, for all $(x, a) \in \mathcal{X}_h \times \mathcal{A}$, and $b_h^{(t)}$ is as in (9). The bound in (11) is a core result for the analysis of `Optimistic-PSDP`. Note that this bound is weaker than the typical least-squares error bounds in reinforcement learning, which bound the squared approximation error of the (optimistic) value function $Q_h^{(t)}$. However, this bound suffices for our purposes. In our setting, bounding the squared approximation error is not feasible because the bonus term in

the definition of $Q_h^{(t)}$ in (10) is not necessarily linear in the feature vector $\phi(x,a)$; though the term $Q_h^{\widehat{\pi}^{(t)}}(x,a)$ in the definition of $Q_h^{(t)}$ is linear in $\phi(x,a)$ thanks to Assumption 2.1, the second "bonus" term $\mathbb{E}^{\widehat{\pi}^{(t)}}\left[\sum_{\ell=h}^H b_\ell^{(t)}(\boldsymbol{x}_\ell) \mid h = x, \boldsymbol{a}_h = a\right]$ is not necessarily linear. As will become clearer in the sequel, it is precisely due to this non-linearity that we need (11) to hold for all $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$; this is also why we require a CSC oracle over policies in `FitValue` (see Line 17 in Algorithm 5—the full version of `FitValue`). We comment on the CSC oracle calls within `FitValue` in Remark C.1. In Remark C.2, we also comment on the presence of the peculiar term $\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\}$ in (11).

The reason (11) is possible at all is because we have multiplied the bonuses by $\mathbb{I}\{\|\varphi(\cdot; W_\ell)\| \geq \mu\}$ (see (9)), where, as noted in Section 2.2, we use $\|\varphi(\cdot; W_\ell)\|$ as a proxy for the design range $\mathrm{Rg}^{\mathsf{D}}(\cdot)$. From Lemma 2.3 in Section 2.2, we know that for any $\gamma > 0$ and function f, the map $g_f : x \mapsto \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(x) \geq \gamma\} \cdot f(\cdot)$ is admissible, which in turn means that the conditional expectation $\mathbb{E}^\pi[g_f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a]$ is linear in $\phi(x,a)$ for any policy π. And, as already mentioned in Section 2.2, even though $\|\varphi(\cdot; W_\ell)\|$ may not be a good proxy for the range function at the start of the algorithm, the algorithm updates the preconditioning matrix W_ℓ each time we "witness" non-admissibility of $\mathbb{I}\{\|\varphi(\cdot; W_\ell)\| \geq \mu\} \cdot f(\cdot)$ for some f. And, after each update, $\|\varphi(\cdot; W_\ell)\|$ becomes a better proxy for $\mathrm{Rg}^{\mathsf{D}}(\cdot)$.

More specifically, we show (see Appendix H) that for any $\ell \in [h+1\,..\,H]$, any function $f : \mathcal{X}_\ell \to [-L, L]$ for $L > 0$, and any policy $\pi \in \Pi$, if W_ℓ is a valid preconditioning for layer ℓ, then for some small parameters $\mu, \lambda > 0$, one of the following holds:

1. The discrepancy $\left|\mathbb{E}^\pi\left[\mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} \cdot f(\boldsymbol{x}_\ell) - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}^W\right]\right|$ is small for the standard λ-regularized least-squares parameter
$$\hat{\theta}^W \approx \Sigma_\lambda^{-1} \mathbb{E}^\pi[\mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} \cdot f(\boldsymbol{x}_\ell) \cdot \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)] \in \mathbb{R}^d,$$
where $\Sigma_\lambda := \lambda I + \mathbb{E}^\pi[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top]$; or
2. It is possible to compute a non-zero vector $w_\ell \in \mathbb{R}^d$ such that $(W_\ell^{-2} + w_\ell w_\ell^\top)^{-1/2}$ remains a valid preconditioning matrix for layer ℓ. Specifically, by letting
$$\hat{\vartheta}^W := \Sigma_\lambda^{-1} \mathbb{E}^\pi\left[\mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} \cdot f(\boldsymbol{x}_\ell) \cdot \phi(\boldsymbol{x}_h, \boldsymbol{a}_h) \otimes \tfrac{\varphi(\boldsymbol{x}_\ell; W_\ell)\varphi(\boldsymbol{x}_\ell; W_\ell)^\top}{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\|^2}\right] \in \mathbb{R}^{d \times d \times d},$$
the preconditioning vector w_ℓ can be computed by first solving the eigenvalue problems
$$z^\pm \in \operatorname*{arg\,max}_{z \in \mathbb{B}(1)} \pm \begin{pmatrix} z^\top \mathbb{E}^\pi\left[\mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} \cdot f(\boldsymbol{x}_\ell) \cdot \frac{\varphi(\boldsymbol{x}_\ell; W_\ell)\varphi(\boldsymbol{x}_\ell; W_\ell)^\top}{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\|^2}\right] z \\ -\hat{\vartheta}^W[\mathbb{E}^\pi[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)], z, z] \end{pmatrix},$$
then setting $w_\ell = \mathrm{Proj}_{\mathcal{S}(W_\ell, \nu)}(\bar{z})$ where $\bar{z}$ is the vector in $\{z^-, z^+\}$ with the highest absolute objective value in the previous display.

In light of Lemma 2.6, by repeatedly testing if the absolute value of the expected difference $\mathbb{E}^\pi\left[f(\boldsymbol{x}_\ell) \cdot \mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}^W\right]$ is small and updating the preconditioning matrix W_ℓ if it is not, it is possible to achieve a good linear fit of $f(\cdot) \cdot \mathbb{I}\{\|\varphi(\cdot; W_\ell)\| \geq \mu\}$ (in expectation under π) after at most $\widetilde{O}(d)$ updates to the matrix W_ℓ.

In the context of $\texttt{FitValue}_h$, f plays the role of the bonus $H \wedge \frac{\varepsilon}{4H} \max_{a \in \mathcal{A}} \|\phi(\cdot, a)\|_{(\beta I + U_\ell)^{-1}}$, with $U_\ell = U_\ell^{(t)}$. By calling $\texttt{FitValue}_h$ and updating the preconditioning matrix $\widetilde{O}(d)$ times, `Optimistic-PSDP` ensures that on *most iterations* (as long as $T = \Omega(d \log d)$ is sufficiently large), a good linear fit of the truncated bonuses $(b_h^{(t)})$ is achieved, thereby ensuring that (11) holds. However, the caveat is that due to the truncation term $\mathbb{I}\{\|\varphi(\cdot; W_\ell)\| \geq \mu\}$ the bonuses may become less effective at driving exploration. This is a significant issue and we address how it is managed in Appendix C.2.

4.2. `UncertainPolicy` (Algorithm 4) and `Evaluate` (Algorithm 3)

`Optimistic-PSDP` iteratively updates the policy sets $(\Psi_h^{(t)})$ and the design matrices $(U_h^{(t)})$ using the subroutine `UncertainPolicy`. The policies in $\Psi_{h-1}^{(t)}$ are constructed to ensure good coverage over the state space, enabling effective transfer of guarantees from (11) to other policies (i.e., change-of-measure). Inspired by classical linear MDPs, coverage quality is measured via the "diversity" of expected feature vectors induced by policies, akin to a G-optimal design. While such a design exists for $m = \widetilde{O}(d)$, constructing it in our setting is challenging without access to additional oracles.

To address this challenges, `Optimistic-PSDP` employs a greedy construction via the subroutine `UncertainPolicy`, which ensures that for some large enough $t = \Omega(d)$: for all $\ell \in [0\,..\,h-1]$ and $(\pi, v) \in \Psi_\ell^{(t)}$, $\mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}^{(t)}}\left[\max_{a\in\mathcal{A}} \|\phi(\boldsymbol{x}_h, a)\|_{(\beta I + U_h^{(t)})^{-1}}\right] \lesssim 2\sqrt{d}$. This "on-policy" version of G-optimality is sufficient to perform the required change-of-measures in our analysis; see §C.1.4 for more detail.

Finally, the `Evaluate` subroutine evaluates the performance of the policy $\widehat{\pi}_{1:H}^{(t)}$ at iteration t by computing the average sum of rewards across $n_{\texttt{traj}}$ trajectories. It is invoked at the end of each iteration to assess the policy's performance, and `Optimistic-PSDP` ultimately returns the best-performing policy after T rounds. Our analysis shows that for $T = \Omega(d\log(d))$, there exists at least one $O(\varepsilon)$-optimal policy among $(\widehat{\pi}_{1:H}^{(t)})_{t\in[T]}$.

5. Conclusion, Limitations and Future Work

In this paper, we presented an algorithm for the linear Q^π-realizable setting that is both sample-efficient and requires only a polynomial number of calls to a cost-sensitive classification oracle over policies. We further showed that this oracle can be implemented efficiently when the feature dimension is constant. The techniques developed in this work may be of independent interest and could potentially extend to other reinforcement learning settings beyond linear function approximation. In particular, the core ideas behind `Optimistic-PSDP`—notably, the combination of PSDP with a distribution shift detection mechanism (via the `UncertainPolicy` subroutine)—may prove broadly useful. For instance, it can be shown that the `Optimistic-PSDP` framework generalizes the `RVFS` algorithm (Mhammedi et al., 2024) while removing its recursive structure, and our analysis offers a significantly simpler proof strategy. This template has also been recently applied to the linear $Q_\beta^\star$ setting in regularized MDPs (Foster et al., 2025).

That said, several important questions remain open. First, it is unclear whether a computationally efficient algorithm can be designed for non-constant feature dimensions without relying on a classification oracle. Second, we do not yet know whether the $\mathrm{poly}(A)$ dependence in the sample complexity can be removed under a local optimism-based approach like ours. Addressing these challenges could lead to more broadly applicable and computationally tractable algorithms for RL.

Finally, as noted in Footnote 2, the layered MDP assumption does restrict generality in the context of linearly Q^π-realizable MDPs. Removing this assumption could open the door to algorithms that extend to the infinite-horizon setting.

References

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

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 33: 13399–13412, 2020.

Philip Amortila, Nan Jiang, Dhruv Madeka, and Dean P Foster. A few expert queries suffices for sample-efficient rl with resets and linear value approximation. *Advances in Neural Information Processing Systems*, 35:29637–29648, 2022.

Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Zhaohan Daniel Guo, and Charles Blundell. Agent57: Outperforming the atari human benchmark. In *International conference on machine learning*, pages 507–517. PMLR, 2020.

James Bagnell, Sham M Kakade, Jeff Schneider, and Andrew Ng. Policy search by dynamic programming. *Advances in neural information processing systems*, 16, 2003.

Alina Beygelzimer, John Langford, and Pradeep Ravikumar. Error-correcting tournaments. In *International Conference on Algorithmic Learning Theory*, pages 247–262. Springer, 2009.

Kai-Wei Chang, Akshay Krishnamurthy, Alekh Agarwal, Hal Daumé III, and John Langford. Learning to search better than your teacher. In *International Conference on Machine Learning*, pages 2058–2066. PMLR, 2015.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient PAC RL with rich observations. In *Advances in neural information processing systems*, pages 1422–1432, 2018.

Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient Q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, pages 8060–8070, 2019.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. Go-explore: a new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. First return, then explore. *Nature*, 590(7847):580–586, 2021.

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

Dylan J. Foster, Zakaria Mhammedi, and Dhruv Rohatgi. Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration, 2025. URL https://arxiv.org/abs/2503.07453.

Paul Goldberg and Mark Jerrum. Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 361–369, 1993.

Noah Golowich and Ankur Moitra. Linear bellman completeness suffices for efficient online reinforcement learning with few actions. *arXiv preprint arXiv:2406.11640*, 2024.

Zhaohan Guo, Shantanu Thakoor, Miruna Pîslar, Bernardo Avila Pires, Florent Altché, Corentin Tallec, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, et al. Byol-explore: Exploration by bootstrapped prediction. *Advances in neural information processing systems*, 35: 31855–31870, 2022.

Botao Hao, Nevena Lazic, Dong Yin, Yasin Abbasi-Yadkori, and Csaba Szepesvari. Confident least square value iteration with local access to a simulator. In *International Conference on Artificial Intelligence and Statistics*, pages 2420–2435. PMLR, 2022.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

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

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

Ying Jin. Upper bounds on the natarajan dimensions of some function classes. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 1020–1025. IEEE, 2023.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998.

John Langford and Alina Beygelzimer. Sensitive error correcting output codes. In *Learning Theory: 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005. Proceedings 18*, pages 158–172. Springer, 2005.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.

Gen Li, Yuxin Chen, Yuejie Chi, Yuantao Gu, and Yuting Wei. Sample-efficient reinforcement learning is feasible for linearly realizable mdps with limited revisiting. *Advances in Neural Information Processing Systems*, 34:16671–16685, 2021.

Zakaria Mhammedi, Adam Block, Dylan J Foster, and Alexander Rakhlin. Efficient model-free exploration in low-rank mdps. *arXiv preprint arXiv:2307.03997*, 2023.

Zakaria Mhammedi, Dylan J Foster, and Alexander Rakhlin. The power of resets in online reinforcement learning. *arXiv preprint arXiv:2404.15417*, 2024.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, 2012.

Stephane Ross and J Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014.

Tim Salimans and Richard Chen. Learning montezuma's revenge from a single demonstration. *arXiv preprint arXiv:1812.03381*, 2018.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, 31, 2018.

Arash Tavakoli, Vitaly Levdik, Riashat Islam, Christopher M Smith, and Petar Kormushev. Exploring restart distributions. *arXiv preprint arXiv:1811.11298*, 2018.

Volodymyr Tkachuk, Gellért Weisz, and Csaba Szepesvári. Trajectory data suffices for statistically efficient learning in offline rl with linear qpi-realizability and concentrability. *arXiv preprint arXiv:2405.16809*, 2024.

Michael J Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

Yuanhao Wang, Ruosong Wang, and Sham Kakade. An exponential lower bound for linearly realizable mdp with constant suboptimality gap. *Advances in Neural Information Processing Systems*, 34:9521–9533, 2021.

Gellert Weisz, Philip Amortila, Barnabás Janzer, Yasin Abbasi-Yadkori, Nan Jiang, and Csaba Szepesvári. On query-efficient planning in mdps under linear realizability of the optimal state-value function. In *Conference on Learning Theory*, pages 4355–4385. PMLR, 2021a.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021b.

Gellért Weisz, András György, Tadashi Kozuno, and Csaba Szepesvári. Confident approximate policy iteration for efficient local planning in q^π-realizable mdps. *Advances in Neural Information Processing Systems*, 35:25547–25559, 2022.

Gellért Weisz, András György, and Csaba Szepesvári. Online rl in linearly q^π-realizable mdps is as easy as in linear mdps if you learn what to ignore. *arXiv preprint arXiv:2310.07811*, 2023.

Gellért Weisz, András György, and Csaba Szepesvári. Online rl in linearly qpi-realizable mdps is as easy as in linear mdps if you learn what to ignore. *Advances in Neural Information Processing Systems*, 36, 2024.

Lin Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Dong Yin, Botao Hao, Yasin Abbasi-Yadkori, Nevena Lazić, and Csaba Szepesvári. Efficient local planning with linear function approximation. In *International Conference on Algorithmic Learning Theory*, pages 1165–1192. PMLR, 2022.

Dong Yin, Sridhar Thiagarajan, Nevena Lazic, Nived Rajaraman, Botao Hao, and Csaba Szepesvari. Sample efficient deep reinforcement learning via local planning. *arXiv preprint arXiv:2301.12579*, 2023.

Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pages 4473–4525. PMLR, 2021.

Contents of Appendix

Appendix A. Organization of the Appendix

This appendix is organized as follows:

- In Appendix B, we present related work in the context of linear function approximation.
- In Appendix C, we provide an extended discussion of the algorithm design and challenges.
- In Appendix D, we provide a proof sketch of the main guarantee of `Optimistic-PSDP`.
- In Appendix E, we present the full versions of the `FitValue` and `UncertainPolicy` algorithms.
- Appendix F details the choice of parameters for the `Optimistic-PSDP` algorithm.
- Appendix G contains the proofs for structural results related to linearly Q^π-realizable MDPs.
- Appendix H showcases how a non-zero preconditioning vector can be computed when a linear fit fails (see discussion in § C.1.2).
- Appendix I provides the guarantees of the `UncertainPolicy` algorithm, including both statements and proofs of guarantees.
- Appendix J outlines the guarantees of `FitValue`, along with proofs and additional structural results.
- Appendix K presents the guarantee for the `Evaluate` procedure.
- Appendix L contains the proof of Theorem 3.1.
- Appendix M discusses the implementation of the CSC oracle over Π_{Bench}.
- Appendix N provides the upper bound on the growth function of Π_{Bench} and includes the proof of Lemma 2.7.
- Appendix O contains a collection of helper lemmas used throughout the analysis.

Appendix B. Related Work

Our work builds on a substantial body of literature in reinforcement learning (RL) with linear function approximation, a foundational framework for understanding which structural assumptions enable the design of sample- and computationally-efficient RL algorithms. A prominent example is the linear MDP framework (Jin et al., 2020), where the transition operator is assumed to have a low-rank structure. Linear MDPs have been studied extensively, with sample-efficient algorithms developed for both offline and online settings. In contrast, as mentioned earlier, the linear Q^π-assumption, which generalizes the linear MDP assumption, has mostly been explored in the presence of a local simulator (Hao et al., 2022; Weisz et al., 2021a; Li et al., 2021; Yin et al., 2022; Weisz et al., 2022) due to the inherent challenges of the setting.

In the offline RL setting, where the learner is provided with a fixed dataset and cannot interact with the MDP to generate new trajectories, Wang et al. (2021) established an exponential-in-horizon

sample complexity lower bound for learning in linearly Q^π-realizable MDPs, even under a feature coverage assumption. Later, Tkachuk et al. (2024) showed that with a stronger concentrability assumption (Foster et al., 2021), linearly Q^π-realizable MDPs become statistically tractable in the offline setting, though no computationally efficient algorithm was provided. In the online RL setting, which is the focus of this paper, Weisz et al. (2021b) developed a sample-efficient algorithm, but it, too, lacks computational efficiency.

Local simulators: Theoretical research. RL with local simulators has received extensive interest in the context of linear function approximation. Most notably, Weisz et al. (2021a) show that reinforcement learning with linear $V^\star$ is tractable with local simulator access, and Li et al. (2021) show that RL with linear $Q^\star$ and a state-action gap is tractable; online RL is known to be intractable under the same assumptions (Weisz et al., 2021a; Wang et al., 2021). Amortila et al. (2022) show that the gap assumption can be removed if a small number of expert queries are available. Also of note are the works of Yin et al. (2022); Weisz et al. (2022), which give computationally-efficient algorithms under linear Q^π-realizability for all π; this setting is known to be tractable in the online RL model (Weisz et al., 2023), but computationally-efficient algorithms are currently only known for RLLS.

Global simulators. Global simulators—in which the agent can query arbitrary state-action pairs and observe next state transitions—have also received theoretical investigation, but like local simulators, results are largely restricted to tabular reinforcement learning and linear models (Kearns and Singh, 1998; Kakade, 2003; Sidford et al., 2018; Du et al., 2020; Yang and Wang, 2019; Lattimore et al., 2020).

Local simulators: Empirical research. The Go-Explore algorithm (Ecoffet et al., 2019, 2021) uses local simulator access to achieve state-of-the-art performance for the Atari games Montezuma's Revenge and Pitfall—both notoriously difficult games that require systematic exploration. To the best of our knowledge, the performance of Go-Explore on these tasks has yet to be matched by online reinforcement learning; the performing agents (Badia et al., 2020; Guo et al., 2022) are roughly a factor of four worse in terms of cumulative reward. Interestingly, like `Optimistic-PSDP`, Go-Explore makes use of core sets of informative state-action pairs to guide exploration. However, Go-Explore uses an ad-hoc, domain specific approach to designing the core set, and does not use function approximation to drive exploration.

Recent work of Yin et al. (2023) provides an empirical framework for online RL with local planning that can take advantage of deep neural function approximation, and is inspired by the theoretical works in Weisz et al. (2021a); Li et al. (2021); Yin et al. (2022); Weisz et al. (2022). This approach does not have provable guarantees, but achieves super-human performance at Montezuma's Revenge.

Other notable empirical works that incorporate local simulator access, as highlighted by Yin et al. (2023), include Schulman et al. (2015); Salimans and Chen (2018); Tavakoli et al. (2018).

Appendix C. Extended Algorithm Overview and Key Challenges

In this section, we describe our algorithm, `Optimistic-PSDP`, offer some intuition behind its design (Appendix C.1), and outline the challenges that motivated our approach (Appendix C.2). This section expands on Section 4, providing additional details on the `UncertainPolicy` and `Evaluate` subroutines and outlining the challenges that shaped our specific algorithmic solution.

C.1. High-Level Algorithm Description and Intuition

Our main algorithm, `Optimistic-PSDP` (Algorithm 1), builds on the classical Policy Search by Dynamic Programming (PSDP) algorithm (see, e.g., Bagnell et al. (2003)) by incorporating bonuses into the rewards. In a nutshell, the algorithm learns a policy in a dynamic programming fashion by fitting optimistic value functions for each layer $h = H, \dots, 1$.[3] It is well known in RL, that adding the right bonuses helps in driving exploration, which is what we use them for in our algorithm. `Optimistic-PSDP` consists of three subroutines: `FitValue` (Algorithm 2), `UncertainPolicy` (Algorithm 4), and `Evaluate` (Algorithm 3). Note that Algorithm 2 and Algorithm 4 are simplified (asymptotic) versions of the full algorithms in Appendix E; Algorithm 5 and Algorithm 6, respectively.

Before delving into the specifics of `Optimistic-PSDP`, we first provide an overview of the key variables in Algorithm 1.

C.1.1. KEY VARIABLES IN `Optimistic-PSDP` (ALGORITHM 1)

`Optimistic-PSDP` runs for $T = \widetilde{O}(d)$ iterations, where at each iteration $t \in [T]$, the algorithm maintains the following variables:

- $\Psi_h^{(t)} \subset \Pi \times \mathbb{R}^d$ consists of t policy-vector pairs. The subroutines $\texttt{FitValue}_h$ and $\texttt{UncertainPolicy}_h$ within `Optimistic-PSDP` use the policies in $\Psi_h^{(t)}$ to generate trajectories up to layer h. Ideally, $\Psi_h^{(t)}$ should contain policies that provide good coverage over the state-action space at layer h. Intuitively, the set $\Psi_h^{(t)}$ plays the role of a core set of policies (Agarwal et al., 2020; Zanette et al., 2021; Du et al., 2019; Wang et al., 2021).

- $U_h^{(t)}$ is a "design matrix" for layer h consisting of the sum of outer products of (truncated) expected features vectors $\bar{\phi}_h^{\pi,v} \coloneqq \mathbb{E}^\pi[\mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \geq 0\} \cdot \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)]$, for $(\pi, v) \in \Psi_h^{(t)}$; that is,

$$U_h^{(t)} \approx \sum_{(\pi,v)\in\Psi_h^{(t)}} \bar{\phi}_h^{\pi,v} (\bar{\phi}_h^{\pi,v})^\top.$$

  In the sequel (see Remark C.2), we will elucidate the role of the indicator term $\mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \geq 0\}$ in the definition of $\bar{\phi}_h^{\pi,v}$. The matrix $U_h^{(t)}$ is used to define bonus functions that are added to the rewards (as in the next bullet point).

- $W_h^{(t)}$ is a valid preconditioning (Definition 2.5) for all t (with high probability). The matrix $W_h^{(t)}$ and the design matrix $U_h^{(t)}$ are used to define the bonus function $b_h^{(t)} : \mathcal{X}_h \to [0, H]$ as

$$b_h^{(t)}(\cdot) = \min\left(H, \frac{\varepsilon}{4H} \max_{a\in\mathcal{A}} \|\phi(\cdot, a)\|_{(\beta I + U_h^{(t)})^{-1}}\right) \cdot \mathbb{I}\{\|\varphi(\cdot; W_h^{(t)})\| \geq \mu\}, \tag{12}$$

3. The approach of incorporating local optimism into `PSDP` has also been recently employed by Golowich and Moitra (2024) for the linear Bellman complete setting.

for some parameter μ and suboptimality ε.

- $\hat{\theta}_h^{(t)}$ is a parameter vector at layer h used to approximate the optimistic value function at layer h:

$$Q_h^{(t)}(x,a) \coloneqq Q_h^{\widehat{\pi}^{(t)}}(x,a) + \mathbb{E}^{\widehat{\pi}^{(t)}}\left[\sum_{\ell=h}^{H} b_\ell^{(t)}(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right]. \tag{13}$$

  Essentially, `Optimistic-PSDP` (or more specifically $\texttt{FitValue}_h$) ensures that $\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}_h^{(t)} + b_h^{(t)}(\boldsymbol{x}_h) \approx Q_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h)$ in expectation under trajectories generated with policies in $\Psi_{h-1}^{(t)}$ (see (14) below).

- $\widehat{\pi}_h^{(t)}(\cdot) = \arg\max_{a \in \mathcal{A}} \phi(\cdot, a)^\top \hat{\theta}_h^{(t)}$ represents the policy at layer h in iteration t.

In each iteration $t \in [T]$, `Optimistic-PSDP` computes the policies $\widehat{\pi}_{1:H}^{(t)}$ in a dynamic programming fashion by fitting the optimistic value functions $(Q_h^{(t)})$ at each layer h. The subroutine `FitValue` (informal version in Algorithm 2), which we describe next, is responsible for fitting these value functions.

Algorithm 2 Informal (asymptotic) version of $\texttt{FitValue}_h$ (Algorithm 5); for the formal version, all expectations are replaced by finite sample estimates, where the input n represents the sample size.

input: h, Ψ_{h-1}, $\widehat{\pi}_{h+1:H}$, $U_{h+1:H}$, $W_{h+1:H}$, Π', $\mu, \nu, \lambda, \beta, \varepsilon, \delta, n$.

initialize: For all $\ell \in [h+1\,..\,H]$, $w_\ell \leftarrow 0$.

1: Define $\varepsilon' = 2dc\varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\frac{\delta}{6HT}, \Pi', |\Psi_{h-1}|) + \frac{8cd\nu HA}{\mu\lambda}$, with $\varepsilon^{\mathsf{b}}_{\mathsf{reg}}$ be as in (87) and $c \coloneqq 20d\log(1 + 16H^4\nu^{-4})$.

2: Define bonuses $b_\ell(\cdot) = \min\big(H, \frac{\varepsilon}{4H}\max_{a\in\mathcal{A}}\|\phi(\cdot,a)\|_{(\beta I+U_\ell)^{-1}}\big)\cdot\mathbb{I}\{\|\varphi(\cdot;W_\ell)\| \geq \mu\}$.

`/* Fitting the rewards and bonuses. */`

3: Set $\Sigma_h \leftarrow \lambda I + \sum_{(\pi,v)\in\Psi_{h-1}} \mathbb{E}^{\pi\circ_h\pi_{\mathsf{unif}}\circ_{h+1}\widehat{\pi}_{h+1:H}}[\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top]$.

4: Set $\hat{\theta}^{\mathsf{r}}_h \leftarrow \Sigma_h^{-1}\sum_{(\pi,v)\in\Psi_{h-1}} \mathbb{E}^{\pi\circ_h\pi_{\mathsf{unif}}\circ_{h+1}\widehat{\pi}_{h+1:H}}[\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)\cdot\sum_{\ell=h}^H \boldsymbol{r}_\ell]$.

5: Set $\hat{\theta}^{\mathsf{b}}_{h,\ell} \leftarrow \Sigma_h^{-1}\sum_{(\pi,v)\in\Psi_{h-1}} \mathbb{E}^{\pi\circ_h\pi_{\mathsf{unif}}\circ_{h+1}\widehat{\pi}_{h+1:H}}[\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)\cdot b_\ell(\boldsymbol{x}_\ell)]$.

6: Set $\hat{\theta}_h \leftarrow \hat{\theta}^{\mathsf{r}}_h + \sum_{\ell=h+1}^H \hat{\theta}^{\mathsf{b}}_{h,\ell} \in \mathbb{R}^d$.

`/* Check the quality of the linear fit for the bonuses and compute new preconditioning vectors. */`

7: Define $\Delta_{h,\ell}(\pi, v, \tilde{\pi}) \leftarrow \mathbb{E}^{\pi\circ_h\tilde{\pi}\circ_{h+1}\widehat{\pi}_{h+1:H}}\Big[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1},\boldsymbol{a}_{h-1})^\top v \geq 0\}\Big(b_\ell(\boldsymbol{x}_\ell) - \phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top\hat{\theta}^{\mathsf{b}}_{h,\ell}\Big)\Big]$.

8: Define $\mathcal{H} \leftarrow \{\ell \in [h+1\,..\,H] : \max_{(\pi,v)\in\Psi_{h-1}}\max_{\tilde{\pi}\in\Pi'}|\Delta_{h,\ell}(\pi,v,\tilde{\pi})| > \varepsilon'\}$. `// $\varepsilon'$ as in Line 1`

9: **for** $\ell \in \mathcal{H}$ **do**

10: Define $B_\ell(\cdot) = b_\ell(\cdot)\cdot\|\varphi(\cdot;W_\ell)\|^{-2}\cdot\varphi(\cdot;W_\ell)\varphi(\cdot;W_\ell)^\top \in \mathbb{R}^{d\times d}$.

11: Set $\hat{\vartheta}^{\mathsf{b}}_{h,\ell} \leftarrow \Sigma_h^{-1}\sum_{(\pi,v)\in\Psi_{h-1}} \mathbb{E}^{\pi\circ_h\pi_{\mathsf{unif}}\circ_{h+1}\widehat{\pi}_{h+1:H}}[\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)\otimes B_\ell(\boldsymbol{x}_\ell)] \in \mathbb{R}^{d\times d\times d}$.

12: Compute $((\pi_{h,\ell}, v_{h,\ell}), \tilde{\pi}_{h,\ell}) \in \arg\max_{((\pi,v),\tilde{\pi})\in\Psi_{h-1}\times\Pi'}|\Delta_{h,\ell}(\pi,v,\tilde{\pi})|$.

13: For $(\pi, v, \tilde{\pi}) = (\pi_{h,\ell}, v_{h,\ell}, \tilde{\pi}_{h,\ell})$, compute

$$z_\ell \leftarrow \arg\max_{z\in\mathbb{B}(1)}\Big|\mathbb{E}^{\pi\circ_h\tilde{\pi}\circ_{h+1}\widehat{\pi}_{h+1:H}}\Big[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1},\boldsymbol{a}_{h-1})^\top v \geq 0\}\cdot\Big(z^\top B_\ell(\boldsymbol{x}_\ell)z - \hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(\boldsymbol{x}_h,\boldsymbol{a}_h), z, z]\Big)\Big]\Big|.$$

14: Set $\tilde{z}_\ell \leftarrow \mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(z_\ell)$.

15: $w_\ell \leftarrow W_\ell^{-1}\tilde{z}_\ell$.

16: **return** $(\hat{\theta}_h, w_{h+1:H})$.

C.1.2. FitValue (ALGORITHM 2)

In each iteration $t \in [T]$, starting from $h = H$ and progressing down to $h = 1$, `Optimistic-PSDP` invokes $\texttt{FitValue}_h$ with the input $(\Psi^{(t)}_{h-1}, \widehat{\pi}^{(t)}_{h+1:H}, U^{(t)}_{h+1:H}, W^{(t)}_{h+1:H})$, returning the pair $(\hat{\theta}^{(t)}_h, w^{(t,h)}_{h+1:H})$. The vectors $w^{(t,h)}_{h+1:H}$ are then used in Line 5 of Algorithm 1 to update the preconditioning matrices $(W^{(t)}_{h+1:H})$. The $\texttt{FitValue}_h$ subroutine ensures, with high probability, that $(W^{(t)}_h)$ are valid preconditionings (see Lemma J.5 and recall the definition of a valid preconditioning in Definition 2.5). Consequently, using Lemma 2.6—which bounds the maximum length of a sequence of non-zero preconditioning vectors—it can be shown that $w^{(t,h)}_\ell$ is non-zero only on $\widetilde{O}(d)$ iterations; these are the iterations where the preconditioning matrix $W^{(t)}_\ell$ actually changes (see Line 5). Intuitively, $W^{(t)}_\ell$ does not update too frequently.

On iterations t where the preconditioning matrix is not updated (which is the case for most iterations as just argued), $\texttt{FitValue}_h$ ensures that $\phi(\cdot,\cdot)^\top \hat{\theta}_h^{(t)}$ is a good approximation of the optimistic value function $Q_h^{(t)}$ in (13) in expectation under trajectories generated using policies in $\Psi_{h-1}^{(t)}$. More specifically, the subroutine $\texttt{FitValue}_h$ guarantees that on most iterations $t \in [T]$: for all $(\pi, v) \in \Psi_{h-1}^{(t)}$ and $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$ (see Lemma J.5),

$$\left|\mathbb{E}^{\pi \circ_h \tilde{\pi}}\left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\} \cdot (Q_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h) - \widehat{Q}_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h))\right]\right| \leq O\left(\frac{\varepsilon}{8TH\sqrt{d}}\right), \tag{14}$$

where $\widehat{Q}_h^{(t)}(x,a) = \phi(x,a)^\top \hat{\theta}_h^{(t)} + b_h^{(t)}(x)$, for all $(x,a) \in \mathcal{X}_h \times \mathcal{A}$, and $b_h^{(t)}$ is as in (12). The bound in (14) is a core result for the analysis of `Optimistic-PSDP`. Note that this bound is weaker than the typical least-squares error bounds in reinforcement learning, which bound the squared approximation error of the (optimistic) value function $Q_h^{(t)}$. However, this bound suffices for our purposes. In our setting, bounding the squared approximation error is not feasible because the bonus term in the definition of $Q_h^{(t)}$ in (13) is not necessarily linear in the feature vector $\phi(x,a)$; though the term $Q_h^{\widehat{\pi}^{(t)}}(x,a)$ in the definition of $Q_h^{(t)}$ is linear in $\phi(x,a)$ thanks to Assumption 2.1, the second "bonus" term $\mathbb{E}^{\widehat{\pi}^{(t)}}\left[\sum_{\ell=h}^H b_\ell^{(t)}(\boldsymbol{x}_\ell) \mid h = x, \boldsymbol{a}_h = a\right]$ is not necessarily linear. As will become clearer in the sequel, it is precisely due to this non-linearity that we need (14) to hold for all $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$; this is also why we require a CSC oracle over policies in `FitValue` (see Line 17 in Algorithm 5—the full version of `FitValue`). We comment on the CSC oracle calls within `FitValue` in Remark C.1 below. In Remark C.2, we also comment on the presence of the peculiar term $\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\}$ in (14).

The reason (14) is possible at all is because we have multiplied the bonuses by $\mathbb{I}\{\|\varphi(\cdot; W_\ell)\| \geq \mu\}$ (see (12)), where, as noted in Section 2.2, we use $\|\varphi(\cdot; W_\ell)\|$ as a proxy for the design range $\mathrm{Rg}^{\mathsf{D}}(\cdot)$. From Lemma 2.3 in Section 2.2, we know that for any $\gamma > 0$ and function f, the map $g_f : x \mapsto \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(x) \geq \gamma\} \cdot f(\cdot)$ is admissible, which in turn means that the conditional expectation $\mathbb{E}^\pi[g_f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a]$ is linear in $\phi(x,a)$ for any policy π. And, as already mentioned in Section 2.2, even though $\|\varphi(\cdot; W_\ell)\|$ may not be a good proxy for the range function at the start of the algorithm, the algorithm updates the preconditioning matrix W_ℓ each time we "witness" non-admissibility of $\mathbb{I}\{\|\varphi(\cdot; W_\ell)\| \geq \mu\} \cdot f(\cdot)$ for some f. And, after each update, $\|\varphi(\cdot; W_\ell)\|$ becomes a better proxy for $\mathrm{Rg}^{\mathsf{D}}(\cdot)$.

More specifically, we show (see Appendix H or the proof of Lemma J.4 for the guarantee of `FitValue`) that for any $\ell \in [h+1\,..\,H]$, any function $f : \mathcal{X}_\ell \to [-L, L]$ for $L > 0$, and any policy $\pi \in \Pi$, if W_ℓ is a valid preconditioning for layer ℓ, then for some small parameters $\mu, \lambda > 0$, one of the following holds:

1. The discrepancy $\left|\mathbb{E}^\pi\left[\mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} \cdot f(\boldsymbol{x}_\ell) - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}^W\right]\right|$ is small for the standard λ-regularized least-squares parameter
$$\hat{\theta}^W \approx \Sigma_\lambda^{-1} \mathbb{E}^\pi[\mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} \cdot f(\boldsymbol{x}_\ell) \cdot \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)] \in \mathbb{R}^d,$$
where $\Sigma_\lambda \coloneqq \lambda I + \mathbb{E}^\pi[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top]$; or
2. It is possible to compute a non-zero vector $w_\ell \in \mathbb{R}^d$ such that $(W_\ell^{-2} + w_\ell w_\ell^\top)^{-1/2}$ remains a valid preconditioning matrix for layer ℓ. Specifically, by letting
$$\hat{\vartheta}^W \coloneqq \Sigma_\lambda^{-1} \mathbb{E}^\pi\left[\mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} \cdot f(\boldsymbol{x}_\ell) \cdot \phi(\boldsymbol{x}_h, \boldsymbol{a}_h) \otimes \frac{\varphi(\boldsymbol{x}_\ell; W_\ell)\varphi(\boldsymbol{x}_\ell; W_\ell)^\top}{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\|^2}\right] \in \mathbb{R}^{d \times d \times d},$$

the preconditioning vector w_ℓ can be computed by first solving the eigenvalue problems

$$z^\pm \in \arg\max_{z\in\mathbb{B}(1)} \pm \begin{pmatrix} z^\top \mathbb{E}^\pi\left[\mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} \cdot f(\boldsymbol{x}_\ell) \cdot \frac{\varphi(\boldsymbol{x}_\ell;W_\ell)\varphi(\boldsymbol{x}_\ell;W_\ell)^\top}{\|\varphi(\boldsymbol{x}_\ell;W_\ell)\|^2}\right] z \\ -\hat{\vartheta}^W[\mathbb{E}^\pi[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)], z, z] \end{pmatrix},$$

then setting $w_\ell = \mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(\bar{z})$ where $\bar{z}$ is the vector in $\{z^-, z^+\}$ with the highest absolute objective value in the previous display.

In light of Lemma 2.6, by repeatedly testing if the discrepancy

$$\left|\mathbb{E}^\pi\left[f(\boldsymbol{x}_\ell) \cdot \mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}^W\right]\right|$$

is small and updating the preconditioning matrix W_ℓ if it is not, it is possible to achieve a good linear fit of $f(\cdot) \cdot \mathbb{I}\{\|\varphi(\cdot; W_\ell)\| \geq \mu\}$ (in expectation under π) after at most $\widetilde{O}(d)$ updates to the matrix W_ℓ.

In the context of $\texttt{FitValue}_h$, f plays the role of the "untruncated" bonus

$$H \wedge \frac{\varepsilon}{4H} \max_{a\in\mathcal{A}} \|\phi(\cdot, a)\|_{(\beta I + U_\ell)^{-1}}.$$

By calling $\texttt{FitValue}_h$ and updating the preconditioning matrix $\widetilde{O}(d)$ times, `Optimistic-PSDP` ensures that on *most iterations*, a good linear fit of the truncated bonuses $(b_h^{(t)})$ is achieved, thereby ensuring that (14) holds. However, the caveat is that due to the truncation term $\mathbb{I}\{\|\varphi(\cdot; W_\ell)\| \geq \mu\}$ the bonuses may become less effective at driving exploration. We address how this issue is managed in the following section. Essentially, states $x \in \mathcal{X}_\ell$ for which $\|\varphi(x; W_\ell)\| < \mu$ are low-range states (as demonstrated by Lemma 2.5 and Lemma 2.1) and can, to a certain extent, be ignored without affecting the algorithm's ability to explore; as discussed in Section 2.2, low-range states do not significantly impact the process of finding a near-optimal policy.

Remark C.1 (Calls to CSC oracle within `FitValue`)**.** *In the informal version of* `FitValue` *(Algorithm 2) it is not clear how the policy optimization steps in Line 8 and Line 12 reduce to the cost-sensitive classification problem over* Π_{Bench} *described in Section 2.2. However, the reduction can immediately be seen in the full version of* `FitValue` *in Algorithm 5, where the discrepancy* $\Delta_{h,\ell}(\pi, v, \tilde{\pi})$ *(see Line 12) is given by*

$$\Delta_{h,\ell}(\pi, v, \tilde{\pi}) = \sum_{(x_{1:H}, a_{1:H}, r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \frac{A \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \geq 0\}}{n} \cdot \left(b_\ell(x_\ell) - \phi(x_h, a_h)^\top \hat{\theta}^{\mathfrak{b}}_{h,\ell}\right) \cdot \mathbb{I}\{\tilde{\pi}(x_h) = a_h\}.$$

Thus, the optimization steps in Line 13 and Line 17 of Algorithm 5 can be reduced to solving

$$\arg\max_{\tilde{\pi}\in\Pi'} \ \mathfrak{s} \cdot \sum_{(x_{1:H}, a_{1:H}, r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} c(x_{1:H}, a_{1:H}, r_{1:H}; v) \cdot \mathbb{I}\{\tilde{\pi}(x_h) = a_h\},$$

for all $(\pi, v) \in \Psi_{h-1}$ *and* $\mathfrak{s} \in \{-1, 1\}$*, where*

$$c(x_{1:H}, a_{1:H}, r_{1:H}; v) := \frac{A \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \geq 0\}\left(b_\ell(x_\ell) - \phi(x_h, a_h)^\top \hat{\theta}^{\mathfrak{b}}_{h,\ell}\right)}{n}.$$

This problem is now clearly of the form that our CSC oracle in Section 2.3 solves.

Remark C.2. *The reader may have notice the peculiar term* $\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\}$ *in the guarantee we presented in (14) for the optimistic value function fit. As we will see shortly when discussing* `UncertainPolicy`*, this term is needed for a change of measure argument in the analysis. We would have not needed it if we could get the guarantee in (14) in a least-square sense, which we cannot under Assumption 2.1 alone.*[4]

C.1.3. `Evaluate` (ALGORITHM 3)

Before introducing the second main component of `Optimistic-PSDP`, the `UncertainPolicy` subroutine, we first describe the `Evaluate` subroutine in Algorithm 1. The `Evaluate` subroutine is used to evaluate the performance of the policy $\widehat{\pi}_{1:H}^{(t)}$ at iteration t by calculating the average sum of rewards across $n_{\texttt{traj}}$ trajectories. This subroutine is invoked at the end of each iteration to evaluate the policy's performance. `Optimistic-PSDP` ultimately returns the best-performing policy after T rounds. Our analysis of `Optimistic-PSDP` relies on showing that for $T = \Omega(d)$, there will be at least one $O(\varepsilon)$-optimal policy among $(\widehat{\pi}_{1:H}^{(t)})_{t\in[T]}$.

Algorithm 3 `Evaluate`$_h$: Evaluate a policy.

input: $\widehat{\pi}_{1:H}$, n.

`/* Gather trajectory data. */`

Set $\mathcal{D} \leftarrow \varnothing$.

for $n = 1, \ldots, n$ **do**

  Sample trajectory $(\boldsymbol{x}_1, \boldsymbol{a}_1, \boldsymbol{r}_1, \ldots, \boldsymbol{x}_H, \boldsymbol{a}_H, \boldsymbol{r}_H) \sim \mathbb{P}^{\widehat{\pi}_{1:H}}$.

  Update $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\boldsymbol{x}_{1:H}, \boldsymbol{a}_{1:H}, \boldsymbol{r}_{1:H})\}$.

Set $J \leftarrow \frac{1}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H})\in\mathcal{D}} \sum_{h\in[H]} r_h$.

return J.

4. The term $\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\}$ is reminiscent of a term introduced by Mhammedi et al. (2023) in their "reward-free" objective to get rid of the requirement for reachability in low-rank MDPs.

Algorithm 4 Informal (asymptotic) version of $\texttt{UncertainPolicy}_h$ (Algorithm 6); for the formal version, all expectations are replaced by finite sample estimates, where the input n represents the sample size.

input: h, $\Psi_{0:h-1}$, $\widehat{\pi}_{1:h}$, U_h, β, n.
Set $\kappa_h \leftarrow 0$.
for $i \in [d]$ **do**
  Set $v_{h,i} = (\beta I + U_h)^{-1/2} e_i \in \mathbb{R}^d$.
  Set $\pi_{h,i,-}(\cdot) = \arg\max_{a \in \mathcal{A}} -\phi(\cdot, a)^\top v_{h,i}$.
  Set $\pi_{h,i,+}(\cdot) = \arg\max_{a \in \mathcal{A}} \phi(\cdot, a)^\top v_{h,i}$.
  for $\ell = 0, \dots, h-1$ **do**
    for $(\pi, v) \in \Psi_\ell$ **do**
      Set $u_{h,i,\ell,\pi,-} \leftarrow \mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:h}} \left[\phi(\boldsymbol{x}_h, \pi_{h,i,-}(\boldsymbol{x}_h)) \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \pi_{h,i,-}(\boldsymbol{x}_h))^\top v_{h,i} \le 0\}\right]$.
      Set $u_{h,i,\ell,\pi,+} \leftarrow \mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:h}} \left[\phi(\boldsymbol{x}_h, \pi_{h,i,+}(\boldsymbol{x}_h)) \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \pi_{h,i,+}(\boldsymbol{x}_h))^\top v_{h,i} \ge 0\}\right]$.
      Set $(\mathfrak{s}, u_{h,i,\ell,\pi}) \in \arg\max_{(s,u) \in \{(-, u_{h,i,\ell,\pi,-}), (+, u_{h,i,\ell,\pi,+})\}} |\langle u, v_{h,i} \rangle|$.
      if $|\langle u_{h,i,\ell,\pi}, v_{h,i} \rangle| \ge \kappa_h$ **then**
        Set $\kappa_h \leftarrow |\langle u_{h,i,\ell,\pi}, v_{h,i} \rangle|$.
        Set $u_h \leftarrow u_{h,i,\ell,\pi}$ and $v_h \leftarrow \mathfrak{s} \cdot v_{h,i}$.
        Set $\tilde{\pi}_{1:h} \leftarrow \pi' \circ_{\ell+1} \widehat{\pi}_{\ell+1:h-1} \circ_h \pi_{h,i,\mathfrak{s}}$.
return $(u_h, \tilde{\pi}_{1:h}, v_h)$.

C.1.4. UncertainPolicy (ALGORITHM 4)

Optimistic-PSDP uses UncertainPolicy to update the policy sets $(\Psi_h^{(t)})$ and the design matrices $(U_h^{(t)})$. Notice that in (14), we bound the difference between $Q_h^{(t)}$ and $\widehat{Q}_h^{(t)}$ in expectation under trajectories generated using policies from $\Psi_{h-1}^{(t)}$. Thus, it is desirable for the policies in $\Psi_{h-1}^{(t)}$ to have good coverage over the state space; informally, policies with good coverage would allow us to transfer the guarantee in (14) to any other policy (i.e., perform a change of measure) with minimal cost. Taking inspiration from the classical linear MDP setting, one way we can measure the quality of the coverage of a set of policies $\{\pi^{(1)}, \dots, \pi^{(m)}\}$ is by looking at the "diversity" of the expected feature vectors they induce. For example, suppose the policies $\pi^{(1)}, \dots, \pi^{(m)}$ induce a G-optimal design in the space of expected features $\{\mathbb{E}^\pi[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)] \mid \pi \in \Pi\}$; that is, suppose that

$$\forall \pi \in \Pi, \ \ \mathbb{E}^\pi \left[\max_{a \in \mathcal{A}} \|\phi(\boldsymbol{x}_h, a)\|_{U^\dagger}\right] \le \sqrt{2d}, \ \ \text{where} \ \ U \coloneqq \sum_{i=1}^m \mathbb{E}^{\pi^{(i)}}[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)] \mathbb{E}^{\pi^{(i)}}[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top]. \tag{15}$$

Such a set is guaranteed to exist for $m = \widetilde{O}(d)$ (see e.g. Todd (2016)). In this case, for any vector $\theta \in \mathbb{R}^d$ and any policy $\pi \in \Pi$, we can perform the following change of measure (see proof of

Lemma L.1):

$$|\mathbb{E}^{\pi}[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \theta]| \leq \sqrt{d} \cdot \mathbb{E}^{\pi}\left[\max_{a \in \mathcal{A}} \|\phi(\boldsymbol{x}_h, a)\|_{U^\dagger}\right] \cdot \sum_{i \in [m]} \left|\mathbb{E}^{\pi^{(i)}}\left[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \theta\right]\right|, \tag{16}$$

$$\leq \sqrt{2}d \cdot \sum_{i \in [m]} \left|\mathbb{E}^{\pi^{(i)}}\left[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \theta\right]\right|. \quad \text{(by (15))} \tag{17}$$

This implies that if $|\mathbb{E}^{\pi^{(i)}}\left[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \theta\right]|$ is small for all $i \in [m]$, it will also be small for

$$|\mathbb{E}^{\pi}\left[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \theta\right]|,$$

for *any* $\pi \in \Pi$. Thus, one might hope to construct a set $\Psi_{h-1}^{(t)}$ with policies $\pi^{(1)}, \ldots, \pi^{(m)}$ satisfying (15), which would allow us to transfer the guarantee in (14) from the policies in $\Psi_{h-1}^{(t)}$ to any other policy with minimal overhead. However, there are two challenges to achieving this:

1. Although a set of policies satisfying (15) always exists with $m = \widetilde{O}(d)$, there is no straightforward way to compute such a set in our setting. Even in the much simpler linear MDP setting, finding such a set would require solving a non-convex optimization problem.

2. Even if $\Psi_{h-1}^{(t)}$ consisted of policies satisfying (15), we would not necessarily be able to perform a change of measure in (14) as we did in (17). This is because $(Q_h^{(t)} - \widehat{Q}_h^{(t)})(x, a)$ is not necessarily linear in the feature map $\phi(x, a)$, as it would be in the standard linear MDP setting.

To address the first challenge, we will follow a "greedy" approach to construct the set $\Psi_{h-1}^{(t)}$ and the corresponding design matrix $U_{h-1}^{(t)}$ using the $\texttt{UncertainPolicy}_h$ subroutine. The call to $\texttt{UncertainPolicy}_h$ in Optimistic-PSDP at iteration t returns a tuple $(u_h^{(t)}, \tilde{\pi}_{1:h}^{(t)}, v_h^{(t)})$ that is used to update $\Psi_h^{(t)}$ and $U_h^{(t)}$ as (see Line 10):

$$U_h^{(t+1)} \leftarrow U_h^{(t)} + u_h^{(t)}(u_h^{(t)})^\top \quad \text{and} \quad \Psi_h^{(t+1)} \leftarrow \Psi_h^{(t)} \cup \{(\tilde{\pi}_{1:h}^{(t)}, v_h^{(t)})\}. \tag{18}$$

Furthermore, the tuple $(u_h^{(t)}, \tilde{\pi}_{1:h}^{(t)}, v_h^{(t)})$ satisfies, with high probability (see Lemma I.3),

$$u_h^{(t)} \approx \mathbb{E}^{\tilde{\pi}_{1:h}^{(t)}}\left[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h) \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v_h^{(t)} \geq 0\}\right], \tag{19}$$

and for all $\ell \in [0 \,..\, h-1]$ and $(\pi, v) \in \Psi_\ell^{(t)}$:

$$\mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}^{(t)}}\left[\max_{a \in \mathcal{A}} \|\phi(\boldsymbol{x}_h, a)\|_{(\beta I + U_h^{(t)})^{-1}}\right] \lesssim 2\sqrt{d}\|u_h^{(t)}\|_{(\beta I + U_h^{(t)})^{-1}}. \tag{20}$$

Now, thanks to the update rule in (18) for $U_h^{(t+1)}$, a standard elliptical potential argument (see proof of Lemma L.2) implies that for a large enough iteration $t = \Omega(d)$, we have $\|u_h^{(t)}\|_{(\beta I + U_h^{(t)})^{-1}} \leq O(1)$. And so, plugging this into (20) implies that for such a t, we have that for all $\ell \in [0 \,..\, h-1]$ and $(\pi, v) \in \Psi_\ell^{(t)}$,

$$\mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}^{(t)}}\left[\max_{a \in \mathcal{A}} \|\phi(\boldsymbol{x}_h, a)\|_{(\beta I + U_h^{(t)})^{-1}}\right] \lesssim 2\sqrt{d}. \tag{21}$$

This inequality, which suffices for our analysis, can be seen as a weaker "on-policy" version of the G-optimal design inequality in (17), where "on-policy" refers to the fact that the expectation in (21) is taken under the algorithm's own policies $\widehat{\pi}_{1:H}$, in addition to the policies in $(\Psi_\ell^{(t)})$.[5]

Before addressing how we tackle the second challenge (Item 2), there are two additional points worth discussing:

- First, the reason we include indicators of the form $\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\}$ in the guarantee in (14) for $\texttt{FitValue}_h$ and in the expression of $u_h^{(t)}$ in (19) is precisely because we want an inequality like (20) to hold. In general, we cannot find a policy π such that

$$\mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}^{(t)}_{\ell+1:H}} \left[\max_{a \in \mathcal{A}} \|\phi(\boldsymbol{x}_h, a)\|_{(\beta I + U_h^{(t)})^{-1}} \right] \lesssim 2\sqrt{d} \|\mathbb{E}^\pi[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)]\|_{(\beta I + U_h^{(t)})^{-1}}, \tag{22}$$

  for all $\ell \in [0\,..\,h-1]$ and $(\pi, v) \in \Psi_\ell^{(t)}$; Jensen's inequality essentially goes the wrong way (notice the expectation being inside the norm on the right-hand side of (22)).

- Second, one may wonder why we require (20) to hold for all $\ell \in [0\,..\,h-1]$ instead of just $\ell = 0$; that is, bounding $\mathbb{E}^{\widehat{\pi}^{(t)}_{1:H}} \left[\max_{a \in \mathcal{A}} \|\phi(\boldsymbol{x}_h, a)\|_{(\beta I + U_h^{(t)})^{-1}} \right]$. The reason for this is related to the second challenge in Item 2, which we will discuss next.

C.2. Challenge: Non-Linearity of Optimistic Value Functions

The second challenge described in Item 2—the non-linearity of $Q_h^{(t)} - \widehat{Q}_h^{(t)}$—is much more serious. We cannot perform a change of measure (as described in Appendix C.1) unless the error in (14) is linear in ϕ; the analysis of $\texttt{Optimistic-PSDP}$ (and essentially any other RL algorithm) requires performing a change of measure at some step. To understand how we can resolve this issue, we need to closely examine the step in the analysis that requires a change of measure. Similar to typical analyses of algorithms that employ local optimism, our approach involves showing via backward induction that for all $\ell = H+1, \ldots, 1$, $(x, a) \in \mathcal{X}_\ell \times \mathcal{A}$ and $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$:

$$Q_\ell^{\tilde{\pi}}(x, a) \leq \widetilde{Q}_\ell^{(t)}(x, a), \tag{23}$$

$$V_\ell^{\tilde{\pi}}(x) \leq V_\ell^{(t)}(x), \tag{24}$$

where $\widetilde{Q}_\ell^{(t)} \coloneqq Q_\ell^{(t)} - b_\ell^{(t)}$ and $V_\ell^{(t)}(\cdot) = Q_\ell^{(t)}(\cdot, \widehat{\pi}_\ell^{(t)}(\cdot))$. Without going into too much detail, instantiating (24) with $\tilde{\pi} = \pi^\star$ and $\ell = 1$, and using an elliptical potential argument to relate $Q_h^{\widehat{\pi}^{(t)}}$ to $Q_h^{(t)}$ leads to a suboptimality guarantee for $\widehat{\pi}^{(t)}_{1:H}$. Now, let's try to perform one step of backward induction to see where the non-linearity of $Q_h^{(t)}$ is problematic. Fix $h \in [H]$ and assume that (23) and (24) hold for all $\ell \in [h+1\,..\,H]$, and we want to show that they hold for $\ell = h$. First, (23) follows easily from (24) with $\ell = h+1$; in fact, we have that for all $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$ and $(x, a) \in \mathcal{X}_h \times \mathcal{A}$:

$$\begin{aligned} Q_h^{\tilde{\pi}}(x, a) = R(x, a) + \mathbb{E}[V_{h+1}^{\tilde{\pi}}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] &\leq R(x, a) + \mathbb{E}[V_{h+1}^{(t)}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a], \\ &= \widetilde{Q}_h^{(t)}(x, a). \end{aligned}$$

Now, let's see how we can use this to show (24) for $\ell = h$. Using (23) and assuming that $\widehat{Q}_h^{(t)}$ is such that $\widehat{\pi}_h^{(t)}(\cdot) = \arg\max_{a \in \mathcal{A}} \widehat{Q}_h^{(t)}(\cdot, a)$, a standard decomposition (see proof of Theorem 3.1) implies

5. In the RL literature, it is known that on-policy guarantees are sufficient when employing optimism.

that for all $x \in \mathcal{X}_h$:

$$V_h^{\tilde{\pi}}(x) \leq \widetilde{V}_h^{(t)}(x) + (Q_h^{(t)} - \widehat{Q}_h^{(t)})(x, \tilde{\pi}(x)) + (\widehat{Q}_h^{(t)} - Q_h^{(t)})(x, \widehat{\pi}_h^{(t)}(x)), \tag{25}$$

where $\widetilde{V}_h^{(t)} \coloneqq V_h^{(t)} - b_h^{(t)}$. Now, if $Q_h^{(t)} - \widehat{Q}_h^{(t)}$ was linear (which is not the case in our setting); that is, if there exists $\theta_h^{(t)}$ such that $(Q_h^{(t)} - \widehat{Q}_h^{(t)})(\cdot, \cdot) = \phi(\cdot, \cdot)^\top \theta_h^{(t)}$, then by some standard algebra (similar to the steps in (16)), we could perform the following change of measure: for all $(x, a) \in \mathcal{X}_h \times \mathcal{A}$,

$$\begin{aligned}
&(Q_h^{(t)} - \widehat{Q}_h^{(t)})(x, a) \\
&\lesssim \sqrt{d} \|\phi(x, a)\|_{(\beta I + U_h^{(t)})^{-1}} \sum_{\tau \in [t-1]} |(\theta_h^{(t)})^\top u_h^{(\tau)}|, \\
&\lesssim \sqrt{d} \|\phi(x, a)\|_{(\beta I + U_h^{(t)})^{-1}} \sum_{(\pi, v) \in \Psi_h^{(t)}} \left| \mathbb{E}^\pi \left[(\theta_h^{(t)})^\top \phi(\boldsymbol{x}_h, \boldsymbol{a}_h) \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \geq 0\} \right] \right|, \quad \text{(by (19))}
\end{aligned}$$

and by definition of $\theta_h^{(t)}$:

$$\begin{aligned}
&= \sqrt{d} \|\phi(x, a)\|_{(\beta I + U_h^{(t)})^{-1}} \sum_{(\pi, v) \in \Psi_h^{(t)}} \left| \mathbb{E}^\pi \left[(Q_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h) - \widehat{Q}_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h)) \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \geq 0\} \right] \right|, \\
&\lesssim \frac{\varepsilon}{8H} \|\phi(x, a)\|_{(\beta I + U_h^{(t)})^{-1}},
\end{aligned} \tag{26}$$

where the last step follows from (14). Technically, the right-hand side of (14) involves a sum over elements in $\Psi_{h-1}^{(t)}$ rather than $\Psi_h^{(t)}$; the intention here is not to be overly precise, but rather to illustrate how the point-wise error $(Q_h^{(t)} - \widehat{Q}_h^{(t)})(x, a)$ can be bounded in terms of expected errors under rollouts when $(Q_h^{(t)} - \widehat{Q}_h^{(t)})(x, a)$ is linear. Plugging (26) into the decomposition in (25) and setting $b_h^{(t)}(\cdot) \coloneqq \frac{\varepsilon}{4H} \max_{a \in \mathcal{A}} \|\phi(\cdot, a)\|_{(\beta I + U_h^{(t)})^{-1}}$ shows (24) for $\ell = h$ and completes the induction.

Unfortunately, the argument just presented relies on the linearity of $Q_h^{(t)} - \widehat{Q}_h^{(t)}$, which we do not have; the linear-Q^π assumption does not imply the linearity of $Q_h^{(t)} - \widehat{Q}_h^{(t)}$ due to the presence of bonus terms in the definition of $Q_h^{(t)}$. In the next sectiom, we explain how we can overcome this issue.

Remark C.3 (Rollouts versus value-iteration)**.** *Our main algorithm,* `Optimistic-PSDP`*, uses rollouts to estimate the parameters of the value function within the dynamic programming loop. This differs from the standard value-iteration approach, where the value function is updated by solving the Bellman equation. We use rollouts because, unlike value iteration, they allow for the value decomposition in (25), which is crucial to our analysis. A similar decomposition was also key in Golowich and Moitra (2024) to tackle the linear Bellman complete setting.*

C.2.1. First Key Idea: Regaining Linearity by Going One Layer Back

The key to resolving this issue is to perform the change of measure in a slightly different way (which ultimately comes at the cost of requiring the CSC oracle in the call to `FitValue` in Line 12). First, for the backward induction, we are now going to aim at showing that for all $\ell = H + 1, \ldots, 1$, $(x, a) \in \mathcal{X}_\ell \times \mathcal{A}$, and $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$:

$$Q_\ell^{\tilde{\pi}}(x, a) \leq Q_\ell^{(t)}(x, a), \tag{27}$$

$$V_\ell^{\tilde{\pi}}(x) \leq V_\ell^{(t)}(x) + (Q_\ell^{(t)} - \widehat{Q}_\ell^{(t)})(x, \tilde{\pi}(x)) + (\widehat{Q}_\ell^{(t)} - Q_\ell^{(t)})(x, \widehat{\pi}_\ell^{(t)}(x)). \tag{28}$$

Suppose that (27) and (28) hold for all $\ell \in [h+1\,..\,H]$, and we want to show that they hold for $\ell = h$. Once we show (27), (28) will follow easily by a standard decomposition as we did in (25) (see also the proof of Theorem 3.1). Let's examine how we can prove (27) for $\ell = h$ by using the fact that (28) holds for $\ell = h+1$. To do this, we will focus on the comparator policy $\tilde{\pi} = \widehat{\pi}^\star$ that satisfies:

$$\widehat{\pi}^\star_h(\cdot) = \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\cdot) \geq \gamma\} \cdot \pi^\star(\cdot) + \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\cdot) < \gamma\} \cdot \widehat{\pi}^{(t)}_h(\cdot), \tag{29}$$

for some small $\gamma > 0$, where Rg^{D} is as defined in Section 2.3. First, note that $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$ by (8). Second, we will see that the policy $\widehat{\pi}^\star$, which is a mixture of the optimal policy $\pi^\star$ and $\widehat{\pi}^{(t)}$, serves as a sufficiently strong benchmark to allow us to derive a good bound on the suboptimality of $\widehat{\pi}^{(t)}$ relative to $\pi^\star$ by applying (28) with $\ell = 1$. The strength of $\widehat{\pi}^\star$ lies in the fact that the definition of Rg^{D} ensures that the actions taken by a policy on low-range states—those where $\mathrm{Rg}^{\mathsf{D}}(\cdot) < \gamma$—have minimal impact; see the discussion in Section 2.2 following Definition 2.1.

By definition of $\widehat{\pi}^\star$ in (29), we have that if $\tilde{\pi} = \widehat{\pi}^\star_h$, then for all $x \in \mathcal{X}_{h+1}$:

$$\mathrm{Rg}^{\mathsf{D}}(x) < \gamma \implies (Q^{(t)}_{h+1} - \widehat{Q}^{(t)}_{h+1})(x, \tilde{\pi}(x)) + (\widehat{Q}^{(t)}_{h+1} - Q^{(t)}_{h+1})(x, \widehat{\pi}^{(t)}_{h+1}(x)) = 0.$$

This implies that $x \mapsto (Q^{(t)}_{h+1} - \widehat{Q}^{(t)}_{h+1})(x, \tilde{\pi}(x)) + (\widehat{Q}^{(t)}_{h+1} - Q^{(t)}_{h+1})(x, \widehat{\pi}^{(t)}_{h+1}(x))$ is α-admissible with $\alpha = \gamma/(2H)$ (see Lemma 2.3). Thus, by Lemma 2.2, there exists $\theta^{(t)}_h$ such that for all $(x,a) \in \mathcal{X}_h \times \mathcal{A}$:

$$\begin{aligned}
&\phi(x,a)^\top \theta^{(t)}_h \\
&= \mathbb{E}\left[(Q^{(t)}_{h+1} - \widehat{Q}^{(t)}_{h+1})(\boldsymbol{x}_{h+1}, \tilde{\pi}(\boldsymbol{x}_{h+1})) + (\widehat{Q}^{(t)}_{h+1} - Q^{(t)}_{h+1})(\boldsymbol{x}_{h+1}, \widehat{\pi}^{(t)}_{h+1}(\boldsymbol{x}_{h+1})) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right].
\end{aligned}$$

With this, we can now show (27) for $\ell = h$ by performing a change of measure as follows: for all $(x,a) \in \mathcal{X}_h \times \mathcal{A}$,

$$\begin{aligned}
&Q^{\widehat{\pi}^\star}_h(x,a) \\
&= R(x,a) + \mathbb{E}[V^{\widehat{\pi}^\star}_{h+1}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a], \\
&\leq R(x,a) + \mathbb{E}[V^{(t)}_{h+1}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] \\
&\quad + \mathbb{E}\left[(Q^{(t)}_{h+1} - \widehat{Q}^{(t)}_{h+1})(\boldsymbol{x}_{h+1}, \widehat{\pi}^\star_{h+1}(\boldsymbol{x}_{h+1})) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] \\
&\quad + \mathbb{E}\left[(\widehat{Q}^{(t)}_{h+1} - Q^{(t)}_{h+1})(\boldsymbol{x}_{h+1}, \widehat{\pi}^{(t)}_{h+1}(\boldsymbol{x}_{h+1})) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right], \quad \text{(by (28))} \\
&\leq R(x,a) + \mathbb{E}[V^{(t)}_{h+1}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] + \phi(x,a)^\top \theta^{(t)}_h, \\
&= \widetilde{Q}^{(t)}_h(x,a) + \phi(x,a)^\top \theta^{(t)}_h.
\end{aligned} \tag{30}$$

Now, by similar steps as in (26), we have that for all $(x,a) \in \mathcal{X}_h \times \mathcal{A}$:

$$\begin{aligned}
&\phi(x,a)^\top \theta^{(t)}_h \\
&\lesssim \sqrt{d}\|\phi(x,a)\|_{(\beta I + U^{(t)}_h)^{-1}} \sum_{\tau \in [t-1]} |(\theta^{(t)}_h)^\top u^{(\tau)}_h|, \\
&\lesssim \sqrt{d}\|\phi(x,a)\|_{(\beta I + U^{(t)}_h)^{-1}} \sum_{(\pi,v) \in \Psi^{(t)}_h} \left|\mathbb{E}^\pi\left[(\theta^{(t)}_h)^\top \phi(\boldsymbol{x}_h, \boldsymbol{a}_h) \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \geq 0\}\right]\right|, \quad \text{(by (20))},
\end{aligned}$$

and by definition of $\theta_h^{(t)}$ and the triangle inequality:

$$\begin{aligned}
&\leq \sqrt{d}\|\phi(x,a)\|_{(\beta I+U_h^{(t)})^{-1}} \\
&\quad \times \sum_{(\pi,v)\in\Psi_h^{(t)}} \left|\mathbb{E}^{\pi\circ_{h+1}\widehat{\pi}^\star}\left[(Q_{h+1}^{(t)}(\boldsymbol{x}_{h+1},\boldsymbol{a}_{h+1}) - \widehat{Q}_{h+1}^{(t)}(\boldsymbol{x}_{h+1},\boldsymbol{a}_{h+1}))\cdot\mathbb{I}\{\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top v\geq 0\}\right]\right| \\
&+ \sqrt{d}\|\phi(x,a)\|_{(\beta I+U_h^{(t)})^{-1}} \\
&\quad \times \sum_{(\pi,v)\in\Psi_h^{(t)}} \left|\mathbb{E}^{\pi\circ_{h+1}\widehat{\pi}^{(t)}}\left[(Q_{h+1}^{(t)}(\boldsymbol{x}_{h+1},\boldsymbol{a}_{h+1}) - \widehat{Q}_{h+1}^{(t)}(\boldsymbol{x}_{h+1},\boldsymbol{a}_{h+1}))\cdot\mathbb{I}\{\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top v\geq 0\}\right]\right|, \\
&\lesssim \frac{\varepsilon}{4H}\|\phi(x,a)\|_{(\beta I+U_h^{(t)})^{-1}}, \qquad (31)
\end{aligned}$$

where the last inequality follows from (14). Thus, by combining (31) and (30), we get that $Q_h^{\widehat{\pi}^\star}(x,a)\leq \widetilde{Q}_h^{(t)}(x,a)+\frac{\varepsilon}{4H}\max_{a'\in\mathcal{A}}\|\phi(x,a')\|_{(\beta I+U_h^{(t)})^{-1}}$. On the other hand, we also know that $Q_h^{\widehat{\pi}^\star}(x,a)\leq H$, and so using that $\min(c+b,H)\leq\min(c,H)+b$ for all $c,b\geq 0$, we get that

$$Q_h^{\widehat{\pi}^\star}(x,a)\leq \widetilde{Q}_h^{(t)}(x,a)+\bar{b}_h^{(t)}(x), \quad \text{where} \quad \bar{b}_h^{(t)}(x)\coloneqq \min\left(H,\frac{\varepsilon}{4H}\max_{a\in\mathcal{A}}\|\phi(x,a)\|_{(\beta I+U_h^{(t)})^{-1}}\right). \qquad (32)$$

It seems that we are almost done with proving (27) for $\ell=h$. However, the term $\bar{b}_h^{(t)}$ is not exactly the same as $b_h^{(t)}$ in (12); it is missing the indicator $\mathbb{I}\{\varphi(\cdot;W_h)\geq\mu\}$. Recall that we need this indicator to ensure the linear fit in (14).

C.2.2. Second Key Idea: Skipping Over Low Range States

To deal with the issue that the bonus in (32) is missing the indicator $\mathbb{I}\{\varphi(\cdot;W_h)\geq\mu\}$, we will use a special value decomposition (generalizing (25)) which essentially involves the value functions over different layers and treats low-range states in a special way. Again, we need to slightly modify the target inequalities for our induction; We modify (27) and (28) so that the goal is to show via backward induction over $\ell=H+1,\ldots,1$ that for all $(x,a)\in\mathcal{X}_\ell\times\mathcal{A}$:

$$Q_\ell^{\widehat{\pi}^\star}(x,a)\leq Q_\ell^{(t)}(x,a)+\bar{b}_\ell^{(t)}(x)\cdot\mathbb{I}\{\|\varphi(x;W_\ell^{(t)})\|<\mu\}, \qquad (33)$$

$$V_\ell^{\widehat{\pi}^\star}(x)\leq V_\ell^{(t)}(x)+\xi_\ell^{(t)}(x,\widehat{\pi}_\ell^\star(x))-\xi_\ell^{(t)}(x,\widehat{\pi}_\ell^{(t)}(x))+\bar{b}_\ell^{(t)}(x)\cdot\mathbb{I}\{\|\varphi(x;W_\ell^{(t)})\|<\mu\}, \qquad (34)$$

where for $(\tilde{x},\tilde{a})\in\mathcal{X}_\ell\times\mathcal{A}$, $\xi_\ell^{(t)}(\tilde{x},\tilde{a})\coloneqq Q_\ell^{(t)}(\tilde{x},\tilde{a})-\widehat{Q}_\ell^{(t)}(\tilde{x},\tilde{a})$. We note that because of the term $\bar{b}_\ell^{(t)}(x)\cdot\mathbb{I}\{\|\varphi(x;W_\ell^{(t)})\|<\mu\}$, which is not necessarily admissible (Definition 2.4), we cannot easily recover (33) for $\ell=h$ from (34) with $\ell=h+1$. Instead, we will show (33) for $\ell=h$ by leveraging (34) for *all* $\ell\in[h+1\,..\,H]$ and the following "skip-step" decomposition (aimed at replace the steps in e.g. (30)) which we prove in Lemma O.8: for all $(x,a)\in\mathcal{X}_h\times\mathcal{A}$:

$$\begin{aligned}
&\mathbb{E}[V_{h+1}^{\widehat{\pi}^\star}(\boldsymbol{x}_{h+1})-V_{h+1}^{(t)}(\boldsymbol{x}_{h+1})\mid \boldsymbol{x}_h=x,\boldsymbol{a}_h=a] \\
&\leq \sum_{\ell=h+1}^{H}\mathbb{E}^{\widehat{\pi}^{(t)}}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell)\geq\gamma\}\prod_{k=h+1}^{\ell-1}\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k)<\gamma\}\left(\xi_\ell^{(t)}(\boldsymbol{x}_\ell,\widehat{\pi}_\ell^\star(\boldsymbol{x}_\ell))-\xi_\ell^{(t)}(\boldsymbol{x}_\ell,\widehat{\pi}_\ell^{(t)}(\boldsymbol{x}_\ell))\right)\mid \boldsymbol{x}_h=x,\boldsymbol{a}_h=a\right] \\
&\quad+\sum_{\ell=h+1}^{H}\mathbb{E}^{\widehat{\pi}^{(t)}}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell)\geq\gamma\}\prod_{k=h+1}^{\ell-1}\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k)<\gamma\}\cdot\mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell;W_\ell^{(t)})\|<\mu\}\cdot\bar{b}_\ell^{(t)}(\boldsymbol{x}_\ell)\mid \boldsymbol{x}_h=x,\boldsymbol{a}_h=a\right]. \qquad (35)
\end{aligned}$$

The advantage of this decomposition, compared to say (25), is that it eliminates any inadmissible terms on the right-hand side. In fact, by Lemma G.1 in Section 2.2 we have that all the terms on the right-hand side of (35) are linear in $\phi(x,a)$, allowing us to perform a change of measure as in (31), which in turn enables us to prove (33) for $\ell = h$.

Finally, coming back to some earlier points from Appendix C.1, the reason we needed (22) for all $\ell \in [0\,..\,h-1]$ instead of just $\ell = 0$ is precisely because we are using the skip-step decomposition in (35) (see the proof sketch in Appendix D for more detail). Additionally, the policy optimization step in `FitOptVal` is needed because we are bounding the conditional expectation:

$$\mathbb{E}\left[(Q^{(t)}_{h+1} - \widehat{Q}^{(t)}_{h+1})(\boldsymbol{x}_{h+1}, \widehat{\pi}^\star_{h+1}(\boldsymbol{x}_{h+1})) + (\widehat{Q}^{(t)}_{h+1} - Q^{(t)}_{h+1})(\boldsymbol{x}_{h+1}, \widehat{\pi}^{(t)}_{h+1}(\boldsymbol{x}_{h+1})) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right];$$

that is, we are bounding the regression error "one layer back" (as reflected by the title of §C.2.1), and so we need to ensure that we measure the error $(Q^{(t)}_{h+1} - \widehat{Q}^{(t)}_{h+1})(\cdot, \tilde{\pi}(\cdot))$ for all policies $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$.

Appendix D. Proof Sketch of the Main Theorem (Theorem 3.1)

We now provide a proof sketch of Theorem 3.1, with the full proof deferred to Appendix L. We recommend that readers first review Appendix C.1.

Let $\tau \in [T]$ be an iteration such that for all $h \in [H]$, $\ell \in [0\,..\,h-1]$, and $(\pi, v) \in \Psi^{(\tau)}_\ell$,

$$\left|\mathbb{E}^{\pi \circ_h \tilde{\pi}}\left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\} \cdot (Q^{(\tau)}_h(\boldsymbol{x}_h, \boldsymbol{a}_h) - \widehat{Q}^{(\tau)}_h(\boldsymbol{x}_h, \boldsymbol{a}_h))\right]\right| \lesssim \frac{\varepsilon}{16H^2T^2d^2}, \tag{36}$$

and

$$\mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}^{(\tau)}_{\ell+1:H}}\left[\max_{a \in \mathcal{A}} \|\phi(\boldsymbol{x}_h, a)\|_{(\beta I + U^{(\tau)}_h)^{-1}}\right] \lesssim 2\sqrt{d}. \tag{37}$$

We gave a high-level explanation of why such a τ exists in Appendix C.1; see also the formal statements for these bounds in Lemma L.2 (the bound we display in (36) is merely a slightly tighter version of the one we presented earlier in (14)). Further, define

$$\widehat{\pi}^\star_h(\cdot) \coloneqq \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\cdot) < \gamma\} \cdot \widehat{\pi}^{(\tau)}_h(\cdot) + \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\cdot) \geq \gamma\} \cdot \pi^\star(\cdot),$$

for $h \in [H]$, $\gamma = \mu/\sqrt{d\nu}$ (μ, ν are as in Algorithm 1), and Rg^{D} as in Definition 2.3. Note that by (8) and the fact that $\widehat{\pi}^{(\tau)}, \pi^\star \in \Pi_{\mathsf{Base}}$, where $\Pi_{\mathsf{Base}} = \{x \mapsto \arg\max_{a \in \mathcal{A}} \theta^\top \phi(x,a) \mid \theta \in \mathbb{B}(H)\}$, we have

$$\widehat{\pi}^\star_h \in \Pi_{\mathsf{Bench}}. \tag{38}$$

At a high-level, our strategy will be to show the sequence of inequalities:

$$\begin{aligned} J(\pi^\star) &\leq J(\widehat{\pi}^\star_{1:H}) + O(\varepsilon), \\ J(\widehat{\pi}^\star_{1:H}) &\leq J(\widehat{\pi}^{(\tau)}_{1:H}) + O(\varepsilon), \\ J(\widehat{\pi}^{(\tau)}_{1:H}) &\leq J(\widehat{\pi}_{1:H}) + O(\varepsilon). \end{aligned} \tag{39}$$

Summing up these inequalities and telescoping would imply the desired result. The first inequality reflects the fact that it does not matter too much what actions a policy chooses on low range states (see also discussion in Section 2.2 right after Definition 2.1).

Suboptimality of $\widehat{\pi}^\star_{1:H}$. First, by the performance difference lemma (see Lemma O.7), we have

$$\begin{aligned}
&J(\pi^\star) - J(\widehat{\pi}^\star_{1:H}) \\
&= \sum_{h=1}^H \mathbb{E}^{\widehat{\pi}^\star}\Big[Q_h^{\pi^\star}(\boldsymbol{x}_h, \pi^\star(\boldsymbol{x}_h)) - Q_h^{\pi^\star}(\boldsymbol{x}_h, \widehat{\pi}^\star_h(\boldsymbol{x}_h))\Big], \\
&= \sum_{h=1}^H \mathbb{E}^{\widehat{\pi}^\star}\Big[\langle \phi(\boldsymbol{x}_h, \pi^\star(\boldsymbol{x}_h)) - \phi(\boldsymbol{x}_h, \widehat{\pi}^\star_h(\boldsymbol{x}_h)), \theta_h^{\pi^\star}\rangle\Big], \\
&= \sum_{h=1}^H \mathbb{E}^{\widehat{\pi}^\star}\Big[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_h) < \gamma\} \cdot \langle \phi(\boldsymbol{x}_h, \pi^\star(\boldsymbol{x}_h)) - \phi(\boldsymbol{x}_h, \widehat{\pi}^{(\tau)}_h(\boldsymbol{x}_h)), \theta_h^{\pi^\star}\rangle\Big], \quad \text{(by definition of } \widehat{\pi}^\star) \\
&\le \sum_{h=1}^H \mathbb{E}^{\widehat{\pi}^\star}\Big[\mathbb{I}\{\mathrm{Rg}(\boldsymbol{x}_h) < \sqrt{2d}\gamma\} \cdot \langle \phi(\boldsymbol{x}_h, \pi^\star(\boldsymbol{x}_h)) - \phi(\boldsymbol{x}_h, \widehat{\pi}^{(\tau)}_h(\boldsymbol{x}_h)), \theta_h^{\pi^\star}\rangle\Big], \quad \text{(by Lemma 2.1)} \\
&\le \sum_{h=1}^H \mathbb{E}^{\widehat{\pi}^\star}\Big[\mathbb{I}\{\mathrm{Rg}(\boldsymbol{x}_h) < \sqrt{2d}\gamma\} \cdot \mathrm{Rg}(\boldsymbol{x}_h)\Big], \\
&\le H\sqrt{2d}\gamma = H\sqrt{2d/d_\nu}\mu. \qquad (40)
\end{aligned}$$

This implies (39) for the choices of parameters in Algorithm 1. We now bound the suboptimality of $\widehat{\pi}^{(\tau)}_{1:H}$ relative to $\widehat{\pi}^\star_{1:H}$, where we recall that $\tau \in [T]$ is such that (36) and (37) hold.

Suboptimality of $\widehat{\pi}^{(\tau)}_{1:H}$. For this part of the proof sketch, we recall some definitions from Appendix C.1; for $\ell \in [H]$, $(x,a) \in \mathcal{X}_\ell \times \mathcal{A}$, and $t \in [T]$, we define

$$\bar{b}_\ell^{(t)}(x) := H \wedge \frac{\varepsilon \cdot \max_{a' \in \mathcal{A}} \|\phi(x,a')\|_{(\beta I + U_\ell^{(t)})^{-1}}}{4H}, \quad b_\ell^{(t)}(x) := \bar{b}_\ell^{(t)}(x)\mathbb{I}\{\|\varphi(x; W_\ell^{(t)})\| \ge \mu\}, \quad (41)$$

$$Q_\ell^{(t)}(x,a) = Q_\ell^{\widehat{\pi}^{(t)}}(x,a) + \mathbb{E}^{\widehat{\pi}^{(t)}}\left[\sum_{k=\ell}^H b_k^{(t)}(\boldsymbol{x}_k) \mid \boldsymbol{x}_\ell = x, \boldsymbol{a}_\ell = a\right], \quad \text{and} \quad V_\ell^{(t)}(x) := Q_\ell^{(t)}(x, \widehat{\pi}_\ell^{(t)}(x)). \quad (42)$$

Here, $\bar{b}_\ell^{(t)}$ represents the untruncated bonus and $Q_\ell^{(t)}$ represents the optimistic value function.

With this, we proceed via backward induction to show that for all $\ell = H+1, \ldots, 1$ and $(x,a) \in \mathcal{X}_\ell \times \mathcal{A}$:

$$Q_\ell^{\widehat{\pi}^\star}(x,a) \le Q_\ell^{(\tau)}(x,a) + \bar{b}_\ell^{(\tau)}(x) \cdot \mathbb{I}\{\|\varphi(x; W_\ell^{(\tau)})\| < \mu\}, \quad (43)$$

$$V_\ell^{\widehat{\pi}^\star}(x) \le V_\ell^{(\tau)}(x) + \xi_\ell^{(\tau)}(x, \widehat{\pi}^\star_\ell(x)) - \xi_\ell^{(\tau)}(x, \widehat{\pi}^{(\tau)}_\ell(x)) + \bar{b}_\ell^{(\tau)}(x) \cdot \mathbb{I}\{\|\varphi(x; W_\ell^{(\tau)})\| < \mu\}, \quad (44)$$

where for $(\tilde{x}, \tilde{a}) \in \mathcal{X}_\ell \times \mathcal{A}$,

$$V_\ell^{\widehat{\pi}^\star}(x) := Q_\ell^{\widehat{\pi}^\star}(x, \widehat{\pi}^\star_\ell(x)), \quad \xi_\ell^{(\tau)}(\tilde{x}, \tilde{a}) := Q_\ell^{(\tau)}(\tilde{x}, \tilde{a}) - \widehat{Q}_\ell^{(\tau)}(\tilde{x}, \tilde{a}),$$

and

$$\widehat{Q}_\ell^{(t)}(\tilde{x}, \tilde{a}) = \phi(\tilde{x}, \tilde{a})^\top \hat{\theta}_\ell^{(t)} + b_\ell^{(t)}(\tilde{x}),$$

with the convention that $Q_{H+1}^{\widehat{\pi}^\star} \equiv Q_{H+1}^{(\tau)} \equiv \widehat{Q}_{H+1}^{(\tau)} \equiv 0$.

The base case $\ell = H+1$ follows trivially by the convention that $Q_{H+1}^{\widehat{\pi}^\star} \equiv Q_{H+1}^{(\tau)} \equiv \widehat{Q}_{H+1}^{(\tau)} \equiv 0$. Now, let $h \in [H]$ and suppose that the induction hypothesis holds for all $\ell = [h+1 \,..\, H+1]$. We show that it holds for $\ell = h$.

We show (43) for $\ell = h$. Fix $(x, a) \in \mathcal{X}_h \times \mathcal{A}$. By the skip-step decomposition in Lemma O.8 (as alluded to in Appendix C.2) we get that

$$\begin{aligned}
&\mathbb{E}[V^{\widehat{\pi}^\star}_{h+1}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] \\
&= \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \cdot \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot V^{\widehat{\pi}^\star}_\ell(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] \\
&\quad + \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\prod_{k=h+1}^{\ell} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot R(\boldsymbol{x}_\ell, \boldsymbol{a}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right].
\end{aligned} \tag{45}$$

We instantiate Lemma O.8 again, this time with the optimistic value function, to get that (full details are in Appendix L)

$$\begin{aligned}
&\mathbb{E}[V^{(\tau)}_{h+1}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] \\
&= \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \cdot \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot V^{(\tau)}_\ell(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] \\
&\quad + \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\prod_{k=h+1}^{\ell} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot \tilde{r}_\ell(\boldsymbol{x}_\ell, \boldsymbol{a}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right],
\end{aligned} \tag{46}$$

where $\tilde{r}_\ell(\tilde{x}, \tilde{a}) \coloneqq R(\tilde{x}, \tilde{a}) + b^{(\tau)}_\ell(\tilde{x})$ and $b^{(\tau)}_\ell$ is as in (41). Thus, combining (45) and (46) and using that $\tilde{r}_\ell(\cdot) \geq R(\cdot)$, we get

$$\begin{aligned}
&\mathbb{E}[V^{\widehat{\pi}^\star}_{h+1}(\boldsymbol{x}_{h+1}) - V^{(\tau)}_{h+1}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] \\
&\leq \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \cdot \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot \left(V^{\widehat{\pi}^\star}_\ell(\boldsymbol{x}_\ell) - V^{(\tau)}_\ell(\boldsymbol{x}_\ell)\right) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right],
\end{aligned}$$

and so by the induction hypothesis; in particular (44) for $\ell \in [h+1 \,..\, H+1]$, we have

$$\begin{aligned}
&\mathbb{E}[V^{\widehat{\pi}^\star}_{h+1}(\boldsymbol{x}_{h+1}) - V^{(\tau)}_{h+1}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] \\
&\leq \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \left(\xi^{(\tau)}_\ell(\boldsymbol{x}_\ell, \widehat{\pi}^\star_\ell(\boldsymbol{x}_\ell)) - \xi^{(\tau)}_\ell(\boldsymbol{x}_\ell, \widehat{\pi}^{(\tau)}_\ell(\boldsymbol{x}_\ell))\right) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] \\
&\quad + \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot \mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W^{(\tau)}_\ell)\| < \mu\} \cdot \bar{b}^{(\tau)}_\ell(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right].
\end{aligned} \tag{47}$$

Note that this is the inequality we presented in (35). Now, by Lemma 2.5, we have that $\mathrm{Rg}^{\mathsf{D}}(\tilde{x}) \leq \sqrt{d_\nu} \cdot \|\varphi(\tilde{x}; W^{(\tau)}_\ell)\|$ for all $\tilde{x} \in \mathcal{X}_\ell$. Thus, since $\mu = \sqrt{d_\nu} \cdot \gamma$, we have that for all $\tilde{x} \in \mathcal{X}_\ell$, $\mathbb{I}\{\|\varphi(\tilde{x}; W^{(\tau)}_\ell)\| < \mu\} = 1$ only if $\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\tilde{x}) < \gamma\} = 1$. This implies that the second sum in (47) is zero, and so

$$\mathbb{E}[V^{\widehat{\pi}^\star}_{h+1}(\boldsymbol{x}_{h+1}) - V^{(\tau)}_{h+1}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] \leq \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[g^{(\tau)}_\ell(\boldsymbol{x}_{h+1:\ell}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right], \tag{48}$$

where

$$g^{(\tau)}_\ell(\boldsymbol{x}_{h+1:\ell}) \coloneqq \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \cdot \left(\xi^{(\tau)}_\ell(\boldsymbol{x}_\ell, \widehat{\pi}^\star_\ell(\boldsymbol{x}_\ell)) - \xi^{(\tau)}_\ell(\boldsymbol{x}_\ell, \widehat{\pi}^{(\tau)}_\ell(\boldsymbol{x}_\ell))\right). \tag{49}$$

Now, in Lemma G.1, we show that for any $\pi \in \Pi$, $f : \mathcal{X}_\ell \to [-L, L]$ for $L > 0$, there exists $\theta \in \mathbb{R}^d$ such that $\|\theta\| \le \mathrm{poly}(L, d, H, \gamma^{-1})$ and

$$\mathbb{E}^\pi \left[\prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \ge \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = \tilde{x}, \boldsymbol{a}_h = \tilde{a} \right] = \phi(\tilde{x}, \tilde{a})^\top \theta,$$

for $(\tilde{x}, \tilde{a}) \in \mathcal{X}_h \times \mathcal{A}$. Instantiating this with $f(\cdot) = \xi_\ell^{(\tau)}(\cdot, \widehat{\pi}_\ell^\star(\cdot)) - \xi_\ell^{(\tau)}(\cdot, \widehat{\pi}_\ell^{(\tau)}(\cdot))$, we get that there exists $\theta_{h,\ell}^{(\tau)} \in \mathbb{R}^d$ such that $\|\theta_{h,\ell}^{(\tau)}\| \le \mathrm{poly}(L, d, H, \gamma^{-1}, \lambda^{-1})$ and

$$\forall (\tilde{x}, \tilde{a}) \in \mathcal{X}_h \times \mathcal{A}, \quad \mathbb{E}^{\widehat{\pi}^{(\tau)}} \left[g_\ell^{(\tau)}(\boldsymbol{x}_{h+1:\ell}) \mid \boldsymbol{x}_h = \tilde{x}, \boldsymbol{a}_h = \tilde{a} \right] = \phi(\tilde{x}, \tilde{a})^\top \theta_{h,\ell}^{(\tau)}. \tag{50}$$

Thus, by a change of measure argument (similar to the steps taken in (26); see also formal statement in Lemma L.1), we have that (ignoring some low-order terms to simplify the presentation)

$$\begin{aligned} &\left| \mathbb{E}^{\widehat{\pi}^{(\tau)}} \left[g_\ell^{(\tau)}(\boldsymbol{x}_{h+1:\ell}) \mid \boldsymbol{x}_h = x, a_h = a \right] \right| \\ &\lesssim \sqrt{d} \|\phi(x, a)\|_{(\beta I + U_h^{(\tau)})^{-1}} \sum_{(\pi, v) \in \Psi_h^{(\tau)}} \left| \mathbb{E}^\pi [\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \theta_{h,\ell}^{(\tau)} \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \ge 0\}] \right|. \end{aligned} \tag{51}$$

Now, using (50) again and the law of total expectation, we get that for any $(\pi, v) \in \Psi_h^{(\tau)}$:

$$\left| \mathbb{E}^\pi [\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \theta_{h,\ell}^{(\tau)} \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \ge 0\}] \right| = \left| \mathbb{E}^{\pi \circ_{h+1} \widehat{\pi}^{(\tau)}} [g_\ell^{(\tau)}(\boldsymbol{x}_{h+1:\ell}) \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \ge 0\}] \right|,$$

and so by letting $\mathbf{I}_{h,\ell,v} \coloneqq \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \ge 0\} \cdot \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\}$ and using the definition of $g_\ell^{(\tau)}$ in (49):

$$\begin{aligned} &\left| \mathbb{E}^\pi [\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \theta_{h,\ell}^{(\tau)} \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \ge 0\}] \right| \\ &= \left| \mathbb{E}^{\pi \circ_{h+1} \widehat{\pi}^{(\tau)}} \left[\mathbf{I}_{h,\ell,v} \cdot \mathbb{E} \left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \ge \gamma\} \cdot \left(\xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^\star(\boldsymbol{x}_\ell)) - \xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^{(\tau)}(\boldsymbol{x}_\ell)) \right) \mid \boldsymbol{x}_{\ell-1}, \boldsymbol{a}_{\ell-1} \right] \right] \right|. \end{aligned} \tag{52}$$

Now, by Lemma 2.3, the function $x \mapsto \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(x) \ge \gamma\} \cdot \left(\xi_\ell^{(\tau)}(x, \widehat{\pi}_\ell^\star(x)) - \xi_\ell^{(\tau)}(x, \widehat{\pi}_\ell^{(\tau)}(x)) \right)$ is α-admissible with $\alpha = \mathrm{poly}(\lambda, 1/H)$, and so by Lemma 2.2, there exists $\tilde{\theta}_{\ell-1}^{(\tau)} \in \mathbb{R}^d$ such that $\|\tilde{\theta}_{\ell-1}^{(\tau)}\| \le \mathrm{poly}(d, H, \gamma^{-1}, \lambda^{-1})$ and for all $(\tilde{x}, \tilde{a}) \in \mathcal{X}_{\ell-1} \times \mathcal{A}$:

$$\begin{aligned} &\phi(\tilde{x}, \tilde{a})^\top \tilde{\theta}_{\ell-1}^{(\tau)} \\ &= \mathbb{E} \left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \ge \gamma\} \cdot \left(\xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^\star(\boldsymbol{x}_\ell)) - \xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^{(\tau)}(\boldsymbol{x}_\ell)) \right) \mid \boldsymbol{x}_{\ell-1} = \tilde{x}, \boldsymbol{a}_{\ell-1} = \tilde{a} \right], \quad (53) \\ &= \mathbb{E} \left[\xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^\star(\boldsymbol{x}_\ell)) - \xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^{(\tau)}(\boldsymbol{x}_\ell)) \mid \boldsymbol{x}_{\ell-1} = \tilde{x}, \boldsymbol{a}_{\ell-1} = \tilde{a} \right], \quad (54) \end{aligned}$$

where the last equality follows by the fact that $\widehat{\pi}_\ell^\star(\cdot) = \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\cdot) < \gamma\} \cdot \widehat{\pi}_\ell^{(\tau)}(\cdot) + \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\cdot) \ge \gamma\} \cdot \pi^\star(\cdot)$, by definition. Using (53) and a change of measure argument again (see Lemma L.1), we get that for all $(\tilde{x}, \tilde{a}) \in \mathcal{X}_{\ell-1} \times \mathcal{A}$:

$$\begin{aligned} &\left| \mathbb{E} \left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \ge \gamma\} \cdot \left(\xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^\star(\boldsymbol{x}_\ell)) - \xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^{(\tau)}(\boldsymbol{x}_\ell)) \right) \mid \boldsymbol{x}_{\ell-1} = \tilde{x}, \boldsymbol{a}_{\ell-1} = \tilde{a} \right] \right| \\ &\lesssim \sqrt{d} \|\phi(\tilde{x}, \tilde{a})\|_{(\beta I + U_{\ell-1}^{(\tau)})^{-1}} \sum_{(\pi', v') \in \Psi_{\ell-1}^{(\tau)}} \left| \mathbb{E}^{\pi'} [\phi(\boldsymbol{x}_{\ell-1}, \boldsymbol{a}_{\ell-1})^\top \tilde{\theta}_{\ell-1}^{(\tau)} \cdot \mathbb{I}\{\phi(\boldsymbol{x}_{\ell-1}, \boldsymbol{a}_{\ell-1})^\top v' \ge 0\}] \right|, \end{aligned}$$

and so using the property of $\tilde{\theta}^{(\tau)}_{\ell-1}$ in (54), the law of total expectation, and the triangle inequality, we get

$$\begin{aligned}
&\lesssim \sqrt{d}\|\phi(\tilde{x},\tilde{a})\|_{(\beta I+U^{(\tau)}_{\ell-1})^{-1}} \sum_{(\pi',v')\in\Psi^{(\tau)}_{\ell-1}} \left|\mathbb{E}^{\pi'}[\xi^{(\tau)}_\ell(\boldsymbol{x}_\ell,\widehat{\pi}^{(\tau)}_\ell(\boldsymbol{x}_\ell))\cdot\mathbb{I}\{\phi(\boldsymbol{x}_{\ell-1},\boldsymbol{a}_{\ell-1})^\top v'\ge 0\}]\right| \\
&\quad + \sqrt{d}\|\phi(\tilde{x},\tilde{a})\|_{(\beta I+U^{(\tau)}_{\ell-1})^{-1}} \sum_{(\pi',v')\in\Psi^{(\tau)}_{\ell-1}} \left|\mathbb{E}^{\pi'}[\xi^{(\tau)}_\ell(\boldsymbol{x}_\ell,\widehat{\pi}^{\star}_\ell(\boldsymbol{x}_\ell))\cdot\mathbb{I}\{\phi(\boldsymbol{x}_{\ell-1},\boldsymbol{a}_{\ell-1})^\top v'\ge 0\}]\right|, \\
&\lesssim \frac{\varepsilon}{8H^2Td^{3/2}}\|\phi(\tilde{x},\tilde{a})\|_{(\beta I+U^{(\tau)}_{\ell-1})^{-1}}, \qquad (55)
\end{aligned}$$

where the last inequality follows by (36) (see also Lemma L.2 for a formal statement) and the fact that $\widehat{\pi}^\star \in \Pi_{\mathsf{Bench}}$ (see (38)). Plugging (55) into (52), we get that for all $(\pi,v)\in\Psi^{(\tau)}_h$:

$$\begin{aligned}
&\left|\mathbb{E}^{\pi}[\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top\theta^{(\tau)}_{h,\ell}\cdot\mathbb{I}\{\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top v\ge 0\}]\right| \\
&\lesssim \frac{\varepsilon}{8H^2Td^{3/2}}\mathbb{E}^{\pi\circ_{h+1}\widehat{\pi}^{(\tau)}}\left[\mathbf{I}_{h,\ell,v}\cdot\|\phi(\boldsymbol{x}_{\ell-1},\boldsymbol{a}_{\ell-1})\|_{(\beta I+U^{(\tau)}_{\ell-1})^{-1}}\right],
\end{aligned}$$

where we recall that $\mathbf{I}_{h,\ell,v}=\mathbb{I}\{\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top v\ge 0\}\cdot\prod_{k=h+1}^{\ell-1}\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k)<\gamma\}\le 1$. Thus, we have for all $(\pi,v)\in\Psi^{(\tau)}_h$:

$$\begin{aligned}
&\left|\mathbb{E}^{\pi}[\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top\theta^{(\tau)}_{h,\ell}\cdot\mathbb{I}\{\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top v\ge 0\}]\right| \\
&\le \frac{\varepsilon}{8H^2Td^{3/2}}\mathbb{E}^{\pi\circ_{h+1}\widehat{\pi}^{(\tau)}}\left[\mathbf{I}_{h,\ell,v}\cdot\|\phi(\boldsymbol{x}_{\ell-1},\boldsymbol{a}_{\ell-1})\|_{(\beta I+U^{(\tau)}_{\ell-1})^{-1}}\right], \\
&\lesssim \frac{\varepsilon}{4H^2Td^{1/2}},
\end{aligned}$$

where the last inequality follows by (20) and (21); see also Lemma L.2 for a formal statement. Combining this with (51) and (48), we get that

$$\begin{aligned}
&Q^{\widehat{\pi}^\star}_h(x,a)-\big(R(x,a)+\mathbb{E}[V^{(\tau)}_{h+1}(\boldsymbol{x}_{h+1})\mid\boldsymbol{x}_h=x,\boldsymbol{a}_h=a]\big) \\
&=\mathbb{E}[V^{\widehat{\pi}^\star}_{h+1}(\boldsymbol{x}_{h+1})-V^{(\tau)}_{h+1}(\boldsymbol{x}_{h+1})\mid\boldsymbol{x}_h=x,\boldsymbol{a}_h=a], \\
&\le\frac{\varepsilon}{4H}\cdot\|\phi(x,a)\|_{(\beta I+U^{(\tau)}_h)^{-1}}\cdot\varepsilon. \qquad (56)
\end{aligned}$$

On the other hand, we have that

$$Q^{\widehat{\pi}^\star}_h(x,a)\le H \quad\text{and}\quad R(x,a)+\mathbb{E}[V^{(\tau)}_{h+1}(\boldsymbol{x}_{h+1})\mid\boldsymbol{x}_h=x,\boldsymbol{a}_h=a]\ge 0.$$

Combining this with (56) and the fact that $\min(H,c+b)\le c+\min(H,b)$ for $c,b\ge 0$, we get

$$\begin{aligned}
&Q^{\widehat{\pi}^\star}_h(x,a) \\
&\lesssim \min\left(H, R(x,a)+\mathbb{E}[V^{(\tau)}_{h+1}(\boldsymbol{x}_{h+1})\mid\boldsymbol{x}_h=x,\boldsymbol{a}_h=a]+\frac{\varepsilon}{4H}\cdot\|\phi(x,a)\|_{(\beta I+U^{(\tau)}_h)^{-1}}\right), \\
&\le R(x,a)+\mathbb{E}[V^{(\tau)}_{h+1}(\boldsymbol{x}_{h+1})\mid\boldsymbol{x}_h=x,\boldsymbol{a}_h=a]+\min\left(H,\frac{\varepsilon}{4H}\cdot\|\phi(x,a)\|_{(\beta I+U^{(\tau)}_h)^{-1}}\right), \\
&\le R(x,a)+\mathbb{E}[V^{(\tau)}_{h+1}(\boldsymbol{x}_{h+1})\mid\boldsymbol{x}_h=x,\boldsymbol{a}_h=a]+\min\left(H,\frac{\varepsilon}{4H}\cdot\max_{\tilde{a}\in\mathcal{A}}\|\phi(x,\tilde{a})\|_{(\beta I+U^{(\tau)}_h)^{-1}}\right), \\
&= R(x,a)+\mathbb{E}[V^{(\tau)}_{h+1}(\boldsymbol{x}_{h+1})\mid\boldsymbol{x}_h=x,\boldsymbol{a}_h=a]+b^{(\tau)}_h(x)+\bar{b}^{(\tau)}_h(x)\cdot\mathbb{I}\{\|\varphi(x;W^{(\tau)}_h)\|<\mu\}, \qquad (57) \\
&= Q^{(\tau)}_h(x,a)+\bar{b}^{(\tau)}_h(x)\cdot\mathbb{I}\{\|\varphi(x;W^{(\tau)}_h)\|<\mu\},
\end{aligned}$$

where (57) follows by the definitions of $b_h^{(\tau)}$ and $\bar{b}_h^{(\tau)}$ in (41), and the last inequality follows by the definition of $Q_h^{(\tau)}$ in (42). This shows (43) for $\ell = h$.
We show (44) for $\ell = h$. We have

$$
\begin{aligned}
&V_h^{\widehat{\pi}^\star}(x) - V_h^{(\tau)}(x) \\
&= Q_h^{\widehat{\pi}^\star}(x, \widehat{\pi}_h^\star(x)) - Q_h^{(\tau)}(x, \widehat{\pi}_h^{(\tau)}(x)), \\
&\leq Q_h^{\widehat{\pi}^\star}(x, \widehat{\pi}_h^\star(x)) - Q_h^{(\tau)}(x, \widehat{\pi}_h^{(\tau)}(x)) + \widehat{Q}_h^{(\tau)}(x, \widehat{\pi}_h^{(\tau)}(x)) - \widehat{Q}_h^{(\tau)}(x, \widehat{\pi}_h^\star(x)), \quad \text{(by definition of } \widehat{\pi}_h^{(\tau)}) \\
&= Q_h^{\widehat{\pi}^\star}(x, \widehat{\pi}_h^\star(x)) - Q_h^{(\tau)}(x, \widehat{\pi}_h^\star(x)) + Q_h^{(\tau)}(x, \widehat{\pi}_h^\star(x)) - Q_h^{(\tau)}(x, \widehat{\pi}_h^{(\tau)}(x)) \\
&\quad + \widehat{Q}_h^{(\tau)}(x, \widehat{\pi}_h^{(\tau)}(x)) - \widehat{Q}_h^{(\tau)}(x, \widehat{\pi}_h^\star(x)), \\
&\leq \xi_h^{(\tau)}(x, \widehat{\pi}_h^\star(x)) - \xi_h^{(\tau)}(x, \widehat{\pi}_h^{(\tau)}(x)) + \bar{b}_h^{(\tau)}(x) \cdot \mathbb{I}\{\|\varphi(x; W_h^{(\tau)})\| < \mu\},
\end{aligned}
$$

where the last inequality follows by (43) with $\ell = h$. This shows (44) for $\ell = h$ and completes the induction. Instantiation (44) with $\ell = 1$ and using the definition of $V_1^{(\tau)}$ in (42), we get that

$$
\begin{aligned}
&J(\widehat{\pi}_{1:H}^\star) - J(\widehat{\pi}_{1:H}^{(\tau)}) \\
&= \mathbb{E}\left[V_1^{\widehat{\pi}^\star}(\boldsymbol{x}_1)\right] - \mathbb{E}\left[V_1^{\widehat{\pi}^{(\tau)}}(\boldsymbol{x}_1)\right], \\
&= \mathbb{E}\left[V_1^{\widehat{\pi}^\star}(\boldsymbol{x}_1)\right] - \mathbb{E}\left[V_1^{(\tau)}(\boldsymbol{x}_1)\right] + \mathbb{E}\left[\sum_{h=1}^H b_h^{(\tau)}(\boldsymbol{x}_h)\right], \\
&\leq \mathbb{E}\left[\xi_1^{(\tau)}(\boldsymbol{x}_1, \widehat{\pi}_1^\star(\boldsymbol{x}_1)) - \xi_1^{(\tau)}(\boldsymbol{x}_1, \widehat{\pi}_1^{(\tau)}(\boldsymbol{x}_1)) + \bar{b}_1^{(\tau)}(\boldsymbol{x}_1) \cdot \mathbb{I}\{\|\varphi(\boldsymbol{x}_1; W_1^{(\tau)})\| < \mu\}\right] + \mathbb{E}\left[\sum_{h=1}^H b_h^{(\tau)}(\boldsymbol{x}_h)\right], \\
&\leq 2\varepsilon + \mathbb{E}\left[\bar{b}_1^{(\tau)}(\boldsymbol{x}_1) \cdot \mathbb{I}\{\|\varphi(\boldsymbol{x}_1; W_1^{(\tau)})\| < \mu\}\right] + \mathbb{E}\left[\sum_{h=1}^H b_h^{(\tau)}(\boldsymbol{x}_h)\right], \quad \text{(see below)} \qquad (58) \\
&\leq 2\varepsilon + 2\mathbb{E}\left[\sum_{h=1}^H b_h^{(\tau)}(\boldsymbol{x}_h)\right], \qquad (59)
\end{aligned}
$$

where (58) follows by (36) (see Lemma L.2 for a formal statement) and the fact that $\widehat{\pi}^\star \in \Pi_{\mathsf{Bench}}$ (see (38)). Now, by the definition of the bonuses $(b_h^{(\tau)})$ in (41), and the bounds in (20) and (21), we get that (see also Lemma L.2 for a formal statement)

$$
\mathbb{E}\left[\sum_{h=1}^H b_h^{(\tau)}(\boldsymbol{x}_h)\right] \leq d\varepsilon/2,
$$

and so plugging this into (59), we get

$$
J(\widehat{\pi}_{1:H}^\star) - J(\widehat{\pi}_{1:H}^{(\tau)}) \lesssim 2\varepsilon + d\varepsilon. \qquad (60)
$$

Suboptimality of $\widehat{\pi}_{1:H}$. Let τ' be as in Algorithm 1 just before the algorithm returns, and note that the policy $\widehat{\pi}_{1:H}$ satisfies $\widehat{\pi}_{1:H} = \widehat{\pi}_{1:H}^{(\tau')}$. Now, by Lemma L.3 (guarantee of `Evaluate`), we have that

$$
J(\widehat{\pi}_{1:H}^{(\tau)}) \lesssim \max_{t \in [T]} J(\widehat{\pi}_{1:H}^{(t)}) \leq J(\widehat{\pi}_{1:H}^{(\tau')}) + H\sqrt{\frac{2\log T}{n_{\mathtt{traj}}}}. \qquad (61)
$$

Combining (40), (60), and (61), we get

$$J(\pi^\star) - J(\widehat{\pi}_{1:H}) \le H\sqrt{\frac{2\log(T)}{n_{\mathtt{traj}}}} + 2\varepsilon + d\varepsilon + H\sqrt{2d/d_\nu}\cdot\mu. \tag{62}$$

Using the choices of β, μ, ν, T, and $n_{\mathtt{traj}}$, in Algorithm 1, we get that the right-hand of side of (62) is at most $O(\varepsilon)$, which completes the proof sketch.

Appendix E. Full Versions of `FitValue` and `UncertainPolicy`

Algorithm 6 $\mathtt{UncertainPolicy}_h$: Compute design direction and corresponding partial policy.

input: h, $\Psi_{0:h-1}$, $\widehat{\pi}_{1:h}$, U_h, β, n.

/* Gathering trajectories for updating the design matrices. */

for $\ell = 0, \dots, h-1$ **do**
  for $(\pi, v) \in \Psi_\ell$ **do**
    $\mathcal{D}_{\ell,\pi} \leftarrow \varnothing$.
    for $n = 1, \dots, n$ **do**
      Sample trajectory $(\boldsymbol{x}_1, \boldsymbol{a}_1, \boldsymbol{r}_1, \dots, \boldsymbol{x}_h, \boldsymbol{a}_h, \boldsymbol{r}_h) \sim \mathbb{P}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:h}}$.
      Update $\mathcal{D}_{\ell,\pi} \leftarrow \mathcal{D}_{\ell,\pi} \cup \{(\boldsymbol{x}_{1:h}, \boldsymbol{a}_{1:h})\}$.

/* Compute new design direction. */

Set $\kappa_h \leftarrow 0$.
for $i \in [d]$ **do**
  Set $v_{h,i} = (\beta I + U_h)^{-1/2} e_i \in \mathbb{R}^d$.
  Set $\pi_{h,i,-}(\cdot) = \arg\max_{a\in\mathcal{A}} -\phi(\cdot, a)^\top v_{h,i}$.
  Set $\pi_{h,i,+}(\cdot) = \arg\max_{a\in\mathcal{A}} \phi(\cdot, a)^\top v_{h,i}$.
  for $\ell = 0, \dots, h-1$ **do**
    for $(\pi, v) \in \Psi_\ell$ **do**
      Set $u_{h,i,\ell,\pi,-} \leftarrow \frac{1}{n}\sum_{(x_{1:h},a_{1:h})\in\mathcal{D}_{\ell,\pi}} \phi(x_h, \pi_{h,i,-}(x_h)) \cdot \mathbb{I}\{\phi(x_h, \pi_{h,i,-}(x_h))^\top v_{h,i} \le 0\}$.
      Set $u_{h,i,\ell,\pi,+} \leftarrow \frac{1}{n}\sum_{(x_{1:h},a_{1:h})\in\mathcal{D}_{\ell,\pi}} \phi(x_h, \pi_{h,i,+}(x_h)) \cdot \mathbb{I}\{\phi(x_h, \pi_{h,i,+}(x_h))^\top v_{h,i} \ge 0\}$.
      Set $(\mathfrak{s}, u_{h,i,\ell,\pi}) \in \arg\max_{(s,u)\in\{(-,u_{h,i,\ell,\pi,-}),(+,u_{h,i,\ell,\pi,+})\}} |\langle u, v_{h,i}\rangle|$.
      if $|\langle u_{h,i,\ell,\pi}, v_{h,i}\rangle| \ge \kappa_h$ **then**
        Set $\kappa_h \leftarrow |\langle u_{h,i,\ell,\pi}, v_{h,i}\rangle|$.
        Set $u_h \leftarrow u_{h,i,\ell,\pi}$ and $v_h \leftarrow \mathfrak{s}\cdot v_{h,i}$.
        Set $\tilde{\pi}_{1:h} \leftarrow \pi' \circ_{\ell+1} \widehat{\pi}_{\ell+1:h-1} \circ_h \pi_{h,i,\mathfrak{s}}$.
return $(u_h, \tilde{\pi}_{1:h}, v_h)$.

Algorithm 5 `FitValue`$_h$: Fit bonuses and update preconditioning matrices if necessary.

input: h, Ψ_{h-1}, $\widehat{\pi}_{h+1:H}$, $U_{h+1:H}$, $W_{h+1:H}$, Π', $\mu, \nu, \lambda, \beta, \varepsilon, \delta, n$.

initialize: For all $\ell \in [h+1\,..\,H]$, $w_\ell \leftarrow 0$.

Define $\varepsilon' = 2dc\varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\frac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n) + \frac{8cd\nu HA}{\mu\lambda}$, with $\varepsilon^{\mathsf{b}}_{\mathsf{reg}}$ be as in (87) and $c := 20d\log(1 + 16H^4\nu^{-4})$.

`/* Gather trajectory data. */`

for $(\pi, v) \in \Psi_{h-1}$ **do**

  Set $\widehat{\mathcal{D}}_{h,\pi} \leftarrow \varnothing$.

  for $n = 1, \ldots, n$ **do**

    Sample trajectory $(\boldsymbol{x}_1, \boldsymbol{a}_1, \boldsymbol{r}_1, \ldots, \boldsymbol{x}_H, \boldsymbol{a}_H, \boldsymbol{r}_H) \sim \mathbb{P}^{\pi \circ_h \pi_{\mathsf{unif}} \circ_{h+1} \widehat{\pi}_{h+1:H}}$.

    Update $\widehat{\mathcal{D}}_{h,\pi} \leftarrow \widehat{\mathcal{D}}_{h,\pi} \cup \{(\boldsymbol{x}_{1:H}, \boldsymbol{a}_{1:H}, \boldsymbol{r}_{1:H})\}$.

Define bonuses $b_\ell(\cdot) = \min\big(H, \frac{\varepsilon}{4H}\max_{a\in\mathcal{A}} \|\phi(\cdot, a)\|_{(\beta I + U_\ell)^{-1}}\big) \cdot \mathbb{I}\{\|\varphi(\cdot; W_\ell)\| \geq \mu\}$.

`/* Fitting the reward and bonuses. */`

Set $\Sigma_h \leftarrow \lambda n|\Psi_{h-1}|I + \sum_{(\pi,v)\in\Psi_{h-1}} \sum_{(x_{1:H}, a_{1:H}, r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \phi(x_h, a_h)\phi(x_h, a_h)^\top$.

Set $\hat{\theta}^{\mathsf{r}}_h \leftarrow \Sigma_h^{-1} \sum_{(\pi,v)\in\Psi_{h-1}} \sum_{(x_{1:H}, a_{1:H}, r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \phi(x_h, a_h) \cdot \sum_{\ell=h}^H r_\ell \in \mathbb{R}^d$.

Set $\hat{\theta}^{\mathsf{b}}_{h,\ell} \leftarrow \Sigma_h^{-1} \sum_{(\pi,v)\in\Psi_{h-1}} \sum_{(x_{1:H}, a_{1:H}, r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \phi(x_h, a_h) \cdot b_\ell(x_\ell) \in \mathbb{R}^d$.

Set $\hat{\theta}_h \leftarrow \hat{\theta}^{\mathsf{r}}_h + \sum_{\ell=h+1}^H \hat{\theta}^{\mathsf{b}}_{h,\ell}$.

`/* Check the quality of the linear fit for the bonuses and compute new preconditioning vectors. */`

Define $\Delta_{h,\ell}(\pi, v, \tilde{\pi}) \leftarrow \sum_{(x_{1:H}, a_{1:H}, r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \frac{A\cdot\mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \geq 0\}}{n} \cdot \mathbb{I}\{\tilde{\pi}(x_h) = a_h\}\Big(b_\ell(x_\ell) - \phi(x_h, a_h)^\top \hat{\theta}^{\mathsf{b}}_{h,\ell}\Big)$.

Define $\mathcal{H} \leftarrow \{\ell \in [h+1\,..\,H] : \max_{(\pi,v)\in\Psi_{h-1}} \max_{\tilde{\pi}\in\Pi'} |\Delta_{h,\ell}(\pi, v, \tilde{\pi})| > \varepsilon'\}$. `// ε′ defined in Line 1`

for $\ell \in \mathcal{H}$ **do**

  Define $B_\ell(\cdot) = b_\ell(\cdot) \cdot \|\varphi(\cdot; W_\ell)\|^{-2} \cdot \varphi(\cdot; W_\ell)\varphi(\cdot; W_\ell)^\top \in \mathbb{R}^{d\times d}$.

  Set $\hat{\vartheta}^{\mathsf{b}}_{h,\ell} \leftarrow \Sigma_h^{-1} \sum_{(\pi,v)\in\Psi_{h-1}} \sum_{(x_{1:H}, a_{1:H}, r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \phi(x_h, a_h) \otimes B_\ell(x_\ell) \in \mathbb{R}^{d\times d\times d}$.

  Compute $((\pi_{h,\ell}, v_{h,\ell}), \tilde{\pi}_{h,\ell}) \in \arg\max_{((\pi,v),\tilde{\pi})\in\Psi_{h-1}\times\Pi'} |\Delta_{h,\ell}(\pi, v, \tilde{\pi})|$.

  For $(\pi, v, \tilde{\pi}) = (\pi_{h,\ell}, v_{h,\ell}, \tilde{\pi}_{h,\ell})$, compute $z_\ell \in \arg\max_{z\in\mathbb{B}(1)} |\mathfrak{L}(z)|$, where

$$\mathfrak{L}(z) := \sum_{(x_{1:H}, a_{1:H}, r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \geq 0\} \cdot \mathbb{I}\{\tilde{\pi}(x_h) = a_h\} \cdot \Big(z^\top B_\ell(x_\ell) z - \hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(x_h, a_h), z, z]\Big).$$

  Set $\tilde{z}_\ell \leftarrow \mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(z_\ell)$.

  $w_\ell \leftarrow W_\ell^{-1}\tilde{z}_\ell$.

return $(\hat{\theta}_h, w_{h+1:H})$.

Appendix F. Choice of Parameters for `Optimistic-PSDP`

For $\mathfrak{c} = \mathrm{polylog}(d, A, 1/\delta, 1/\varepsilon)$ sufficiently large, we set the parameters as:

$$T = 32dH^2 \cdot \mathfrak{c}, \quad n_{\mathtt{traj}} = \frac{A^8 d^{70} H^{80} \cdot \mathfrak{c}}{\varepsilon^{24}}, \quad \mu = \frac{\varepsilon}{H \cdot \mathfrak{c}}, \quad \nu = \frac{\varepsilon^6}{A^3 H^{20} d^{19} \mathfrak{c}}, \quad \lambda = \frac{\varepsilon^4}{A^2 d^{14} H^{11}}, \quad \beta = \frac{\varepsilon^{12}}{A^4 d^{24} H^{40}}. \tag{63}$$

Appendix G. Proofs for Structural Results for Linearly Q^π-Realizable MDPs

In this section, we present proofs for the structural results for linearly Q^π-realizable MDPs we presented in Section 2.2. Many of the results are modification of existing ones in (Weisz et al., 2024).

Proof of Lemma 2.1. Fix $h \in [H]$ and $x \in \mathcal{X}$. Let ρ_h be the approximate optimal design for Θ_h in Section 2.2. Further, let $\mathcal{C}_h := \mathrm{supp}\, \rho_h$ and $\mathrm{G}(\rho_h) := \sum_{\pi \in \mathcal{C}_h} \rho_h(\pi) \theta_h^\pi (\theta_h^\pi)^\top$. Then,

$$\begin{aligned}
\mathtt{Rg}(x) &= \sup_{a,a' \in \mathcal{A}} \sup_{\theta_h \in \Theta_h} \langle \phi(x,a) - \phi(x,a'), \theta_h \rangle, \\
&= \sup_{a,a' \in \mathcal{A}} \sup_{\theta_h \in \Theta_h} \langle \phi(x,a) - \phi(x,a'), \theta_h \rangle, \\
&\le \sup_{a,a' \in \mathcal{A}} \sup_{\theta_h \in \Theta_h} \|\phi(x,a) - \phi(x,a')\|_{\mathrm{G}(\rho_h)} \cdot \|\theta_h\|_{\mathrm{G}(\rho_h)^\dagger}, \\
&\le \sqrt{\sup_{a,a' \in \mathcal{A}} \sum_{\pi \in \mathcal{C}_h} \rho_h(\pi) \cdot \langle \phi(x,a) - \phi(x,a'), \theta_h^\pi \rangle^2} \cdot \sup_{\theta_h \in \Theta_h} \|\theta_h\|_{\mathrm{G}(\rho_h)^\dagger}, \\
&\le \sqrt{\sup_{a,a' \in \mathcal{A}} \max_{\pi \in \mathcal{C}_h} \langle \phi(x,a) - \phi(x,a'), \theta_h^\pi \rangle^2} \cdot \sqrt{2d}, \quad \text{(see below)} \qquad (64) \\
&= \sqrt{2d} \cdot \mathrm{Rg}^{\mathsf{D}}(x),
\end{aligned}$$

where (64) follows by the fact that $\rho_h \in \Delta(\Pi)$ is an approximate optimal design for $\Theta_h = \{\theta_h^\pi \mid \pi \in \Pi\}$; see Definition 2.2, and the last equality follows by the definition of the design range. This completes the proof. □

Proof of Lemma 2.3. Fix $x \in \mathcal{X}_\ell$. If $\mathrm{Rg}^{\mathsf{D}}(x) < \gamma$, then $F(x) = 0 \le \mathrm{Rg}^{\mathsf{D}}(x)/\alpha$. Now, if $\mathrm{Rg}^{\mathsf{D}}(x) \ge \gamma$, then $F(x) = f(x) \le L \le L \cdot \mathrm{Rg}^{\mathsf{D}}(x)/\gamma$, which completes the proof. □

We now state and prove a generalization of Lemma 2.2 that involves "skipping" over intermediate low-range states, which will be crucial to our analysis.[6]

Lemma G.1. *Let $\ell \in [H]$, $L > 0$, $\gamma > 0$, and $f : \mathcal{X}_\ell \to [-L, L]$ be given. Then, for any $h \in [\ell - 1]$ and $\pi \in \Pi$, there exists $\theta \in \mathbb{B}(4\tilde{d}H^2 L/\gamma)$, where $\tilde{d} := 4d \log\log d + 16$, such that for all $(x,a) \in \mathcal{X}_h \times \mathcal{A}$:*

$$\mathbb{E}^\pi \left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \ge \gamma\} \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a \right] = \phi(x,a)^\top \theta.$$

6. We thank Gellért Weisz for a discussion that led to the realization of the lemma's result.

Proof of Lemma G.1. We show via induction that for all $i = 1, \dots, \ell - h$, there exists $\theta_i \in (i-1)\mathbb{B}(4\tilde{d}HL/\gamma)$ such that for all $(x, a) \in \mathcal{X}_h \times \mathcal{A}$:

$$\begin{aligned}
&\mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] \\
&= \mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+i}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] + \phi(x, a)^\top \theta_i, \qquad (65)
\end{aligned}$$

where we use the convention that $\prod_{k=\ell}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} = 1$. The base case $i = 1$ follows trivially. Now, suppose that the result holds for $i \in [\ell - h - 1]$, and we show that it holds for $i + 1$. We have for all $(x, a) \in \mathcal{X}_h \times \mathcal{A}$:

$$\begin{aligned}
&\mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+i}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] \\
&= \mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+i+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] \\
&\quad - \mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_{h+i}) \geq \gamma\} \cdot \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+i+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right]. \qquad (66)
\end{aligned}$$

Thus, it suffices to show that there exists $\tilde{\theta}_i \in \mathbb{B}(4\tilde{d}HL/\gamma)$ such that $(x, a) \in \mathcal{X}_h \times \mathcal{A}$:

$$\phi(x, a)^\top \tilde{\theta}_i = \mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_{h+i}) \geq \gamma\} \cdot \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+i+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right].$$

Combined with (66), this would imply (65) with i replaced by $i + 1$ and $\theta_{i+1} = \theta_i - \tilde{\theta}_i$.

Now, define

$$g(x') \coloneqq \mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+i+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_{h+i} = x', \boldsymbol{a}_{h+i} = \pi(x')\right],$$

for all $x' \in \mathcal{X}_{h+i}$. By the law of total expectation, we have for all $(x, a) \in \mathcal{X}_h \times \mathcal{A}$:

$$\begin{aligned}
&\mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_{h+i}) \geq \gamma\} \cdot \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+i+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] \\
&= \mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_{h+i}) \geq \gamma\} \cdot g(\boldsymbol{x}_{h+i}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right].
\end{aligned}$$

Now, by Lemma 2.3, $F : x' \mapsto \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(x') \geq \gamma\} \cdot g(x')$ is α-admissible (Definition 2.4) with $\alpha \coloneqq L/\gamma$, and so by Lemma 2.2, there exists $\tilde{\theta}_i \in \mathbb{B}(4\tilde{d}HL/\gamma)$ such that for all $(x, a) \in \mathcal{X}_h \times \mathcal{A}$:

$$\begin{aligned}
&\mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_{h+i}) \geq \gamma\} \cdot \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+i+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] \\
&= \phi(x, a)^\top \tilde{\theta}_i.
\end{aligned}$$

Thus, by (66) and the induction hypothesis (i.e. Eq. (65)), we have that for all $(x,a) \in \mathcal{X}_h \times \mathcal{A}$:

$$\mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \ge \gamma\} \prod_{k=h+i+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right]$$
$$= \phi(x,a)^\top \theta_i - \phi(x,a)^\top \tilde{\theta}_i.$$

This combined with (66) implies (65) with i replaced by $i+1$ and $\theta_{i+1} = \theta_i - \tilde{\theta}_i$, which completes the induction. Instantiating (65) with $i = \ell - h$ (recalling that we are using the convention that $\prod_{k=\ell}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} = 1$) implies that there is some $\theta_{h-\ell} \in (h-\ell-1)\mathbb{B}(4\tilde{d}HL/\gamma)$ such that for all $(x,a) \in \mathcal{X}_h \times \mathcal{A}$:

$$\mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \ge \gamma\} \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right]$$
$$= \mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \ge \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] + \phi(x,a)^\top \theta_{\ell-h}. \tag{67}$$

By Lemma 2.3 and Lemma 2.2 again, there exists $\tilde{\theta}_{\ell-h} \in \mathbb{B}(2\tilde{d}HL/\gamma)$ such that for all $(x,a) \in \mathcal{X}_h \times \mathcal{A}$:

$$\mathbb{E}^\pi\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \ge \gamma\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] = \phi(x,a)^\top \tilde{\theta}_{\ell-h}.$$

Combining this with (67) implies the desired result with $\theta \coloneqq \tilde{\theta}_{\ell-h} + \theta_{\ell-h} \in (\ell-h)\mathbb{B}(4\tilde{d}HL/\gamma) \subseteq \mathbb{B}(4\tilde{d}H^2L/\gamma)$. □

Proof of Lemma 2.4. Let $(w_i)_{i\in[k]}$ be the sequence of vectors corresponding to W_h in the definition of a valid preconditioning; see Definition 2.5. With this, we have $W_h = \left(H^{-2}I + \sum_{i\in[k]} w_i w_i^\top\right)^{-1/2}$. Then, for all $\theta \in \Theta_h$:

$$\|W_h^{-1}\theta\|^2 = \theta^\top\left(H^{-2}I + \sum_{i\in[k]} w_i w_i^\top\right)\theta \le \|\theta\|^2 H^{-2} + \sum_{i\in[k]} \langle \theta, w_i\rangle^2 \overset{(a)}{\le} H^{-2}\|\theta\|^2 + k \le 1 + 4d\log(1 + 16H^4\nu^{-4}),$$

where (a) follows by the definition of a valid preconditioning (Definition 2.5); i.e. that $\sup_{\theta\in\Theta_h} |\langle \theta, w_i\rangle| \le 1$, for all $i \in [k]$, and the last inequality follows by Lemma 2.6 and the fact that $\sup_{\theta\in\Theta_h} \|\theta\| \le H$ (see Assumption 2.1). □

Proof of Lemma 2.5. Aspects of this proof are inspired by the proof of (Weisz et al., 2024, Proposition 4.5).

Fix $x \in \mathcal{X}_\ell$. Let ρ_ℓ be the approximate optimal design for Θ_ℓ in Section 2.2. By definition of the design range in Definition 2.3, we have

$$\mathrm{Rg}^{\mathsf{D}}(x) = \max_{\pi\in\mathrm{supp}(\rho_\ell)} \sup_{a,a'\in\mathcal{A}} \langle \phi(x,a) - \phi(x,a'), \theta_\ell^\pi\rangle,$$
$$= \max_{\pi\in\mathrm{supp}(\rho_\ell)} \sup_{a,a'\in\mathcal{A}} \langle W_\ell\phi(x,a) - W_\ell\phi(x,a'), W_\ell^{-1}\theta_\ell^\pi\rangle,$$

and by Cauchy Schwarz and the definition of $\varphi(\cdot;\cdot)$ in Lemma J.6, we have

$$\le \max_{\pi\in\mathrm{supp}(\rho_\ell)} \|\varphi(x;W_\ell)\| \cdot \|W_\ell^{-1}\theta_\ell^\pi\|,$$
$$\le \sqrt{d_\nu} \cdot \|\varphi(x;W_\ell)\|,$$

where the last inequality follows by the fact that W_ℓ is a valid ν-preconditioning and Lemma 2.4. □

Appendix H. Fit or Precondition

Algorithm 7 `LinearFit`$_h$: Given layers $h, \ell \in [H]$ such that $h < \ell$, a function f, and a valid preconditioning matrix W_ℓ, either return $\hat{\theta}_h^f$ such that $\mathbb{E}^\pi[f(\boldsymbol{x}_\ell) \cdot \mathbb{I}\{\varphi(\boldsymbol{x}_\ell; W_\ell)\} - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}_h^f]$ is small, or compute a new, non-zero vector $w_\ell \in \mathbb{R}^d$ such that $(W_\ell^{-2} + w_\ell w_\ell^\top)^{-1/2}$ remains a valid preconditioning. For practical implementation, expectations would be replaced with empirical averages; we focus on the asymptotic version of the algorithm for simplicity.

input: h, ℓ, f, Ψ, $\widehat{\pi}_{h+1:H}$, $W_{h+1:H}$, μ, ν, λ.

initialize: $w_\ell \leftarrow 0$.

Define $\varepsilon' = \frac{8c\sqrt{\lambda}d\tilde{d}H}{\sqrt{\zeta}\mu} + \frac{8cd\nu L}{\mu\lambda}$, with $c \coloneqq 20d\log(1 + 16H^4\nu^{-4})$, $\tilde{d} \coloneqq 5d\log\log(1 + 16H^4\nu^{-4})$, and $\zeta = \frac{1}{8d}$.

`/* Fitting the truncated version of the function` f. `*/`

Set $\Sigma_h \leftarrow \lambda I + \sum_{\pi\in\Psi} \mathbb{E}^\pi[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top]$.

Set $\hat{\theta}_h^f \leftarrow \Sigma_h^{-1} \sum_{\pi\in\Psi} \mathbb{E}^{\pi\circ_{h+1}\widehat{\pi}_{h+1:H}}[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h) \cdot f(\boldsymbol{x}_\ell) \cdot \mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\}] \in \mathbb{R}^d$.

`/* If the quality of the linear fit is poor, compute new non-zero preconditioning vector. */`

Define $\Delta_{h,\ell}(\pi) \leftarrow \mathbb{E}^{\pi\circ_{h+1}\widehat{\pi}_{h+1:H}}\left[f(\boldsymbol{x}_\ell) \cdot \mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}_h^f\right]$.

for $\max_{\pi\in\Psi} |\Delta_{h,\ell}(\pi)| > \varepsilon'$ **do** `// `ε' `as in Line 1`

  Define $F(\cdot) = f(\cdot) \cdot \mathbb{I}\{\|\varphi(\cdot; W_\ell)\| \geq \mu\} \cdot \|\varphi(\cdot; W_\ell)\|^{-2} \cdot \varphi(\cdot; W_\ell)\varphi(\cdot; W_\ell)^\top \in \mathbb{R}^{d\times d}$.

  Set $\hat{\vartheta}_{h,\ell}^f \leftarrow \Sigma_h^{-1} \sum_{\pi\in\Psi} \mathbb{E}^{\pi\circ_{h+1}\widehat{\pi}_{h+1:H}}[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h) \otimes F(\boldsymbol{x}_\ell)] \in \mathbb{R}^{d\times d\times d}$.

  Compute $\pi_{h,\ell} \in \arg\max_{\pi\in\Psi} |\Delta_{h,\ell}(\pi)|$.

  For $\pi = \pi_{h,\ell}$, compute

$$z_\ell \leftarrow \arg\max_{z\in\mathbb{B}(1)} \left|\mathbb{E}^{\pi\circ_{h+1}\widehat{\pi}_{h+1:H}}\left[z^\top F(\boldsymbol{x}_\ell) z - \hat{\vartheta}_{h,\ell}^f[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h), z, z]\right]\right|.$$

  Set $\tilde{z}_\ell \leftarrow \mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(z_\ell)$.

  $w_\ell \leftarrow W_\ell^{-1}\tilde{z}_\ell$.

return $(\hat{\theta}_h^f, w_\ell)$.

In this section, we show that for any $h \in [H]$, $\ell \in [h+1\,..\,H]$, any function $f : \mathcal{X}_\ell \to [-L, L]$ for $L > 0$, and any policy $\pi \in \Pi$, if W_ℓ is a valid preconditioning for layer ℓ, then for some small parameters $\mu, \lambda > 0$, one of the following holds:

1. The discrepancy $\inf_{\theta\in\mathbb{R}^d} |\mathbb{E}^\pi[\mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} \cdot f(\boldsymbol{x}_\ell) - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top\theta]|$ is small; intuitively, one can think of this as corresponding to case where $\mathbb{E}^\pi[\mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell)\| \geq \mu\} \cdot f(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a]$ is approximately linear in $\phi(x, a)$ (which would be the case if $\varphi(\cdot; W_\ell) \propto \mathrm{Rg}^{\mathsf{D}}(\cdot)$); or
2. It is possible to compute a non-zero vector $w_\ell \in \mathbb{R}^d$ such that $(W_\ell^{-2} + w_\ell w_\ell^\top)^{-1/2}$ remains a valid preconditioning matrix for layer ℓ; the procedure for finding such a vector is displayed in

Algorithm 7; note that many components of Algorithm 7 are very similar to those of `FitValue` (Algorithm 2), which is used in our main algorithm, `Optimistic-PSDP`.

Next, we state this result formally and provide a proof.

Lemma H.1. *Let $h \in [H]$, $\ell \in [h+1\,..\,H]$, $L > 0$, $f : \mathcal{X}_\ell \to [-L, L]$, $\Psi, \widehat{\pi}_{h+1:H}, W_{h+1:H}, \mu, \nu, \lambda$ such that $W_{h+1}, \dots, W_H \in \mathbb{S}^{d\times d}_{++}$ be given and consider a call to $\texttt{LinearFit}_h(\ell, f, \widehat{\pi}_{h+1:H}, W_{h+1:H}, \mu, \nu, \lambda)$. Further, let $d_\nu \coloneqq 5d\log(1+16H^4\nu^{-4})$. Then, the variables in Algorithm 7 are such that either I) $w_\ell = 0$ and*

$$\forall \pi \in \Psi, \quad \left|\mathbb{E}^{\pi \circ_{h+1} \widehat{\pi}_{h+1:H}}\left[f(\boldsymbol{x}_\ell)\cdot \mathbb{I}\{\varphi(\boldsymbol{x}_\ell; W_\ell)\} - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}^f_h\right]\right| \le \frac{8c\sqrt{\lambda}d\tilde{d}H}{\sqrt{\zeta}\mu} + \frac{8cd\nu L}{\mu\lambda}, \tag{68}$$

where $c \coloneqq 20d\log(1+16H^4\nu^{-4})$, $\tilde{d} \coloneqq 5d\log\log(1+16H^4\nu^{-4})$, and $\zeta = \frac{1}{8d}$; or II) the following points hold:

1. $\|w_\ell\| \le \nu^{-1}$;
2. $\|W_\ell w_\ell\| \ge 1/2$; *and*
3. $\sup_{\theta\in\Theta_\ell} |\langle \theta, w_\ell\rangle| \le 1$.

Lemma H.1 implies that if W_ℓ is a valid ν-preconditioning (see Definition 2.5), then in the second case of Lemma H.1, $(W_\ell^{-2} + w_\ell w_\ell^\top)^{-1/2}$ remains a valid ν-preconditioning.

Proof of Lemma H.1. Let ρ_ℓ be the approximate design for Θ_ℓ in Section 2.2 and define

$$G_\ell \coloneqq W_\ell^{-1}\left(\sum_{\pi\in\operatorname{supp}\rho_\ell} \rho_\ell(\pi)\theta^\pi_\ell(\theta^\pi_\ell)^\top\right)W_\ell^{-1} \in \mathbb{R}^{d\times d}.$$

To simplify notation in this proof, we let $c \coloneqq 20d\log(1+16H^4\nu^{-4})$ and for any $z \in \mathbb{R}^d$, we write $z_\parallel \coloneqq \mathrm{Proj}_{\mathcal{S}(G_\ell,\zeta)}(z)$ and $z_\perp \coloneqq \mathrm{Proj}_{\mathcal{S}(G_\ell,\zeta)^\perp}(z)$, where $\zeta \coloneqq \frac{1}{8d}$.

Case where $w_\ell \neq 0$. First, assume that w_ℓ in Algorithm 7 is non-zero. Let

$$\pi_{h,\ell} \in \arg\max_{\pi\in\Psi} \Delta_{h,\ell}(\pi),$$

where $\Delta_{h,\ell}$ is as in Algorithm 5. From Line 5 of Algorithm 7, we have

$$|\Delta_{h,\ell}(\pi_{h,\ell})| \ge \frac{8c\sqrt{\lambda}d\tilde{d}H}{\sqrt{\zeta}\mu} + \frac{8cd\nu L}{\mu\lambda}. \tag{69}$$

Moving forward, we let

$$M \coloneqq \mathbb{E}^{\pi_{h,\ell}\circ_{h+1}\widehat{\pi}_{h+1:H}}\left[F(\boldsymbol{x}_\ell) - \hat{\vartheta}^f_{h,\ell}[\phi(\boldsymbol{x}_h,\boldsymbol{a}_h),\cdot,\cdot]\right],$$

and note that $M \in \mathbb{R}^{d\times d}$ satisfies

$$\mathrm{Tr}(M) = \Delta_{h,\ell}(\pi_{h,\ell}). \tag{70}$$

Furthermore, the variables $(z_\ell, \tilde{z}_\ell, w_\ell)$ in Algorithm 5 satisfy

$$z_\ell \in \arg\max_{z\in\mathbb{B}(1)} |z^\top M z|, \quad \tilde{z}_\ell = \mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(z_\ell), \quad \text{and} \quad w_\ell = W_\ell^{-1}\tilde{z}_\ell.$$

Now, by definition of the trace, there exist unit vectors (eigenvectors in this case) $c_1, \dots, c_d$ such that $\mathrm{Tr}(M) = \sum_{i\in[d]} c_i^\top M c_i$. Combining this with (70) and (69), we get that

$$d \cdot |z_\ell^\top M z_\ell| \geq \left| \sum_{i\in[d]} c_i^\top M c_i \right| \geq \frac{8c\sqrt{\lambda}d\tilde{d}H}{\sqrt{\zeta}\mu} + \frac{8cd\nu L}{\mu\lambda}. \tag{71}$$

Now, let $\overline{M} \coloneqq \mathsf{sgn}(z_\ell^\top M z_\ell) \cdot M$ and note that $z_\ell \in \arg\max_{z\in\mathbb{B}(1)} z_\ell^\top \overline{M} z_\ell$. By (71), we have

$$z_\ell^\top \overline{M} z_\ell > \frac{8c\sqrt{\lambda}\tilde{d}H}{\sqrt{\zeta}\mu} + \frac{8c\nu L}{\mu\lambda}, \tag{72}$$

and so by Lemma J.7 and the definition of M, we have

$$\tilde{z}_\ell^\top \overline{M} \tilde{z}_\ell > \frac{8c\sqrt{\lambda}\tilde{d}H}{\sqrt{\zeta}\mu} + \frac{8(c-1)\nu L}{\mu\lambda}. \tag{73}$$

We use this to show the claims of the lemma.

Proving Item 1. We start by showing $\|w_\ell\| \leq \nu^{-1}$; that is, Item 1 of the lemma. Let $\lambda_1, \dots, \lambda_d$ be the eigenvalues of W_ℓ and let $w_1, \dots, w_d \in \mathbb{R}^d$ be an orthonormal basis such that $W_\ell w_i = \lambda_i w_i$, for all $i \in [d]$ (such a basis exists because $W_\ell \in \mathbb{S}_{++}^{d\times d}$). With this, we can write

$$W_\ell = \sum_{i=1}^d \lambda_i w_i w_i^\top.$$

Let P_ℓ be the matrix whose columns are $w_1, \dots, w_d$, and note that $P_\ell^\top P_\ell = I$ since $w_1, \dots, w_d$ are orthonormal. With this, we have

$$\begin{aligned}
W_\ell^{-1}\mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(z_\ell) &= W_\ell^{-1} \sum_{i\in[d]:\lambda_i\geq\nu} w_i w_i^\top z_\ell, \\
&= \left(\sum_{i=1}^d \lambda_i w_i w_i^\top\right)^{-1} \sum_{i\in[d]:\lambda_i\geq\nu} w_i w_i^\top z_\ell, \\
&= \left(\sum_{i=1}^d \lambda_i P_\ell e_i e_i^\top P_\ell^\top\right)^{-1} \sum_{i\in[d]:\lambda_i\geq\nu} P_\ell e_i e_i^\top P_\ell^\top z_\ell, \\
&= P_\ell \left(\sum_{i=1}^d \lambda_i e_i e_i^\top\right)^{-1} P_\ell^\top \sum_{i\in[d]:\lambda_i\geq\nu} P_\ell e_i e_i^\top P_\ell^\top z_\ell, \\
&= P_\ell \left(\sum_{i=1}^d \lambda_i e_i e_i^\top\right)^{-1} \sum_{i\in[d]:\lambda_i\geq\nu} e_i e_i^\top P_\ell^\top z_\ell, \\
&= P_\ell \mathrm{diag}(\lambda_1^{-1}, \dots, \lambda_d^{-1}) \sum_{i\in[d]:\lambda_i\geq\nu} e_i e_i^\top P_\ell^\top z_\ell, \\
&= P_\ell \sum_{i\in[d]:\lambda_i\geq\nu} \lambda_i^{-1} e_i e_i^\top P_\ell^\top z_\ell,
\end{aligned}$$

Thus, taking the norm and using that $\|P_\ell\|_{\mathrm{op}} = 1$, we get

$$\|w_\ell\| = \|W_\ell^{-1}\tilde{z}_\ell\| = \|W_\ell^{-1}\mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(z_\ell)\| \le \|P_\ell\|_{\mathrm{op}}\sqrt{\sum_{i\in[d]:\lambda_i\ge\nu}\lambda_i^{-2}(e_i^\top P_\ell z_\ell)^2},$$
$$\le \nu^{-1}\|P_\ell z_\ell\| \le \nu^{-1}\|P_\ell\|_{\mathrm{op}}\|z_\ell\| \le \nu^{-1}.$$

This shows Item 1.

Proving Item 2. To prove Item 2, we need to show that $\|W_\ell w_\ell\| = \|\tilde{z}_\ell\| \ge \frac{1}{2}$. Using that $z_\ell \in \arg\max_{z\in\mathbb{B}(1)} z^\top \overline{M} z$ and Lemma J.7,

$$\|\tilde{z}_\ell\|^{-2}\cdot \tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell \le z_\ell^\top \overline{M} z_\ell \le \tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell + \frac{8\nu L}{\mu\lambda},$$
$$\le \tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell + \frac{1}{c} z_\ell^\top \overline{M} z_\ell, \tag{74}$$

where the last inequality follows by (72). Rearranging (74) and using that $c \ge 3$, implies that $\|\tilde{z}_\ell\|^2 \ge \frac{2}{3}$. Therefore, get that

$$\|W_\ell w_\ell\|^2 = \|\tilde{z}_\ell\|^2 \ge \frac{2}{3},$$

satisfying the inequality in Item 2.

Proving Item 3. It remains to prove Item 3. We will start by showing that $\|(\tilde{z}_\ell)_\parallel\|^2 \le \frac{1}{c}$ and use this to prove Item 3. Since $\overline{M}$ is symmetric, we can decompose $\tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell$ as

$$\tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell = (\tilde{z}_\ell)_\parallel^\top \overline{M}\tilde{z}_\ell + (\tilde{z}_\ell)_\parallel^\top \overline{M}(\tilde{z}_\ell)_\perp + (\tilde{z}_\ell)_\perp^\top \overline{M}(\tilde{z}_\ell)_\perp. \tag{75}$$

Now, by Lemma J.6, we have that $(\tilde{z}_\ell)_\parallel^\top F(\cdot)z$ is α-admissible with $\alpha \coloneqq \frac{\sqrt{\zeta}\mu}{L}$, for all $z \in \mathbb{B}(1)$; by Lemma 2.2, this implies that for all $z \in \mathbb{B}(1)$, there exists $\theta_h^{f,z} \in \mathbb{B}(4\tilde{d}H/\alpha)$ such that

$$\mathbb{E}^{\widehat{\pi}_{h+1:H}}[(\tilde{z}_\ell)_\parallel^\top F(\boldsymbol{x}_\ell)z \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] = \phi(x,a)^\top \theta_h^{f,z}, \quad \forall (x,a) \in \mathcal{X}_h \times \mathcal{A}.$$

Thus, by the definitions of $\overline{M}$ and $\hat{\vartheta}_{h,\ell}^f$ in Algorithm 7, and Theorem O.1, we have that for all $z \in \mathbb{B}(1)$:

$$(\tilde{z}_\ell)_\parallel^\top \overline{M} z = \frac{4\sqrt{\lambda}\tilde{d}H}{\sqrt{\zeta}\mu},$$

where $\tilde{d} \coloneqq 5d\log(1 + 16H^4\nu^{-4})$. Instantiating this with $z \in \{\tilde{z}_\ell, (z_\ell)_\perp\}$ and using (75), we have

$$\tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell \le \frac{8\sqrt{\lambda}\tilde{d}H}{\sqrt{\zeta}\mu} + (\tilde{z}_\ell)_\perp^\top \overline{M}(\tilde{z}_\ell)_\perp. \tag{76}$$

Combining this with (73), we get that $\|(\tilde{z}_\ell)_\perp\| \ne 0$. Let $\bar{z}_\ell = (\tilde{z}_\ell)_\perp/\|(\tilde{z}_\ell)_\perp\|$. Since $z_\ell \in \arg\max_{z\in\mathbb{B}(1)} z^\top \overline{M} z$, we have

$$\|(\tilde{z}_\ell)_\perp\|^{-2}\cdot(\tilde{z}_\ell)_\perp^\top \overline{M}(\tilde{z}_\ell)_\perp = (\bar{z}_\ell)^\top \overline{M}\bar{z}_\ell \le z_\ell^\top \overline{M} z_\ell. \tag{77}$$

On the other hand, using Lemma J.7 and (76), we have

$$\begin{aligned} z_\ell^\top \overline{M} z_\ell &\leq \tilde{z}_\ell^\top \overline{M} \tilde{z}_\ell + \frac{8\nu L}{\mu\lambda}, \\ &\leq \frac{8\sqrt{\lambda}\tilde{d}H}{\sqrt{\zeta}\mu} + \frac{8\nu L}{\mu\lambda} + (\tilde{z}_\ell)_\perp^\top \overline{M} (\tilde{z}_\ell)_\perp, \\ &\leq \frac{1}{c} z_\ell^\top \overline{M} z_\ell + (\tilde{z}_\ell)_\perp^\top \overline{M} (\tilde{z}_\ell)_\perp, \end{aligned} \tag{78}$$

where the last inequality follows by (72). Rearranging (78) gives

$$z_\ell^\top \overline{M} z_\ell \leq \frac{c}{c-1} (\tilde{z}_\ell)_\perp^\top \overline{M} (\tilde{z}_\ell)_\perp.$$

Combining this with (77), dividing by $(\tilde{z}_\ell)_\perp^\top \overline{M} (\tilde{z}_\ell)_\perp$, and rearranging, we get that $\|(\tilde{z}_\ell)_\perp\|^2 \geq \frac{c-1}{c}$, and so

$$\|(\tilde{z}_\ell)_\parallel\|^2 \leq \frac{1}{c}, \tag{79}$$

since $1 = \|z_\ell\|^2 \geq \|\tilde{z}_\ell\|^2 = \|(\tilde{z}_\ell)_\parallel\|^2 + \|(\tilde{z}_\ell)_\perp\|^2$. We use (79) to prove Item 3. Note that since

$$G_\ell = W_\ell^{-1} \left(\sum_{\pi \in \operatorname{supp} \rho_\ell} \rho_\ell(\pi) \theta_\ell^\pi (\theta_\ell^\pi)^\top \right) W_\ell^{-1},$$

where ρ_ℓ is the approximate design for Θ_ℓ in Section 2.2, we have

$$\forall \theta \in W_\ell^{-1} \Theta_\ell, \quad \|\theta\|_{G_\ell^\dagger} \leq \sqrt{2d}. \tag{80}$$

With this, we have

$$\begin{aligned} \sup_{\theta \in \Theta_\ell} |\langle \theta, w_\ell \rangle| &= \sup_{\theta \in \Theta_\ell} |\langle \theta, W_\ell^{-1} \tilde{z}_\ell \rangle|, \\ &= \sup_{\theta \in W_\ell^{-1}\Theta_\ell} |\langle \theta, \tilde{z}_\ell \rangle|, \\ &\leq \sup_{\theta \in W_\ell^{-1}\Theta_\ell} \|\theta\| \|(\tilde{z}_\ell)_\parallel\| + \sup_{\theta \in W_\ell^{-1}\Theta_\ell} |\langle \theta, (\tilde{z}_\ell)_\perp \rangle|, \end{aligned}$$

and so by Lemma 2.4 and $d_\nu \coloneqq 5d \log(1 + 16H^4\nu^{-4})$,

$$\leq \sqrt{\frac{d_\nu}{c}} + \sup_{\theta \in W_\ell^{-1}\Theta_\ell} \|\theta\|_{G_\ell^\dagger} \cdot \|(\tilde{z}_\ell)_\perp\|_{G_\ell},$$

and so by (80), we have

$$\begin{aligned}
&\le \sqrt{\frac{d_\nu}{c}} + \sqrt{2d}\cdot\|(\tilde{z}_\ell)_\perp\|_{G_\ell}, \\
&\le \sqrt{\frac{d_\nu}{c}} + \sqrt{2d\cdot(\tilde{z}_\ell)_\perp^\top(\zeta I)(\tilde{z}_\ell)_\perp}, \quad \text{(see below)} \qquad (81)\\
&\le \sqrt{\frac{d_\nu}{c}} + \sqrt{2d\cdot\zeta}, \\
&= \sqrt{\frac{d_\nu}{c}} + \frac{1}{2}, \quad \text{(since } \zeta = (8d)^{-1}\text{)} \\
&\le 1,
\end{aligned}$$

where in (81) follows from the fact that $(\cdot)_\perp = \mathrm{Proj}_{\mathcal{S}(G_\ell,\zeta)^\perp}(\cdot)$; and the last inequality uses that $c = 20d\log(1+16\nu^{-4}H^4) = 4d_\nu$.

Case where $w_\ell = 0$. Note that $w_\ell = 0$ only if $\max_{\pi\in\Psi}|\Delta_{h,\ell}(\pi)| \le \frac{8c\sqrt{\lambda}d\tilde{d}H}{\sqrt{\zeta}\mu} + \frac{8cd\nu L}{\mu\lambda}$, which implies (68) and completes the proof. □

Appendix I. Guarantees of `UncertainPolicy`$_h$

In this section, we present the guarantees for the `UncertainPolicy` subroutine of `Optimistic-PSDP` (see Appendix C.1 for an intuitive explanation of the results). The following subsection outlines these guarantees, with the proofs deferred to Appendix I.2.

I.1. Statement of Guarantees

Lemma I.1 (Concentration result for `UncertainPolicy`$_h$)**.** *Let $h\in[H]$, $\Psi_{0:h-1}$, $\widehat{\pi}_{1:h}$, U_h, β, n be given and consider a call to* `UncertainPolicy`$_h(\Psi_{0:h-1},\widehat{\pi}_{1:h},U_h,\beta,n)$*. Then, for any $\delta\in(0,1)$, with probability at least $1-\delta$, the variables in Algorithm 6 satisfy for all $h\in[H]$, $i\in[d]$, $\ell\in[0\,..\,h-1]$, $(\pi,v)\in\Psi_\ell^{(t)}$, and $s\in\{-,+\}$:*

$$\begin{aligned}
&\left\|\mathbb{E}^{\pi\circ_{\ell+1}\widehat{\pi}_{\ell+1:h}}\left[\phi(\boldsymbol{x}_h,\pi_{h,i,s}(\boldsymbol{x}_h))\cdot\mathbb{I}\{s\cdot\phi(\boldsymbol{x}_h,\pi_{h,i,s}(\boldsymbol{x}_h))^\top v_{h,i}\ge 0\}\right] - u_{h,i,\ell,\pi,s}\right\| \\
&\le 2\sqrt{\frac{2\max_{j\in[H]}\log(4H^2d|\Psi_j|/\delta)}{n}}. \qquad (82)
\end{aligned}$$

Lemma I.2. *Consider a call to* `UncertainPolicy`$_h(\Psi_{0:h-1},\widehat{\pi}_{1:h},U_h,\beta,n)$ *(Algorithm 6) for some given $h\in[H]$, $\Psi_{0:h-1}$, $\widehat{\pi}_{1:h}$, U_h, β, and n. Then, for any $\delta'\in(0,1)$, with probability at least $1-\delta'$, the output $(u_h,\tilde{\pi}_{1:h},v_h)$ of* `UncertainPolicy`$_h$ *satisfies:*

$$\left\|\mathbb{E}^{\tilde{\pi}_{1:h}}\left[\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)\cdot\mathbb{I}\{\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top v_h\ge 0\}\right] - u_h\right\| \le 2\sqrt{\frac{2\max_{j\in[H]}\log(4H^2d|\Psi_j|/\delta')}{n}}, \qquad (83)$$

and for all $\ell\in[0\,..\,h-1]$ and $(\pi,v)\in\Psi_\ell$:

$$\mathbb{E}^{\pi\circ_{\ell+1}\widehat{\pi}_{\ell+1:H}}\left[\|\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)\|_{(\beta I+U_h)^{-1}}\right] \le 2\sqrt{d}\|u_h\|_{(\beta I+U_h)^{-1}} + \frac{4d}{\beta}\sqrt{\frac{2\max_{j\in[H]}\log(4H^2d|\Psi_j|/\delta')}{n}}. \qquad (84)$$

Lemma I.3 (Guarantee of `UncertainPolicy` for `Optimistic-PSDP`). *Let $\varepsilon, \delta \in (0,1)$ be given and consider a call to* $\texttt{Optimistic-PSDP}(\Pi_{\mathsf{Bench}}, \varepsilon, \delta)$ *(Algorithm 1). Then, for any $\delta \in (0,1)$, there is an event $\mathcal{E}^{\mathrm{hoff}}$ of probability at least $1-\delta/2$, under which for all $t \in [T]$ and $h \in [H]$, the output $(u_h^{(t)}, \tilde{\pi}_{1:h}^{(t)}, v_h^{(t)})$ of* $\texttt{UncertainPolicy}_h$ *in Line 9 satisfies:*

$$\left\| \mathbb{E}^{\tilde{\pi}_{1:h}^{(t)}} \left[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h) \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v_h^{(t)} \geq 0\} \right] - u_h^{(t)} \right\| \leq \varepsilon_{\mathrm{hoff}} \coloneqq 2\sqrt{\frac{2\log(4H^3dT/\delta)}{n_{\mathtt{traj}}}},$$

and for all $\ell \in [0\,..\,h-1]$ and $(\pi, v) \in \Psi_\ell^{(t)}$:

$$\mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}^{(t)}} \left[\max_{a \in \mathcal{A}} \|\phi(\boldsymbol{x}_h, a)\|_{(\beta I + U_h^{(t)})^{-1}} \right] \leq 2\sqrt{d} \|u_h^{(t)}\|_{(\beta I + U_h^{(t)})^{-1}} + 2d\beta^{-1}\varepsilon_{\mathrm{hoff}}.$$

Furthermore, $\Psi_h^{(t)}$ satisfies $\Psi_h^{(t)} = \{(\tilde{\pi}_{1:h}^{(\tau)}, v_h^{(\tau)}) \mid \tau \in [h-1]\}$.

Proof. The result follows from Lemma I.2 with $\delta = \delta'/(HT)$, Lemma O.4 (essentially the union bound over $t \in [T]$ and $h \in [H]$), and the fact that $|\Psi_h^{(t)}| \leq T$, for all $t \in [T]$ and $h \in [H]$. □

I.2. Proofs

Proof of Lemma I.1. For each $h \in [H]$, $\ell \in [0\,..\,h-1]$, and $(\pi, v) \in \Psi_\ell$, the dataset $\mathcal{D}_{\ell,\pi}$ in Algorithm 6 consists of i.i.d. trajectories sampled from $\mathbb{P}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:h}}$. What is more, by Line 14 and Line 15 in Algorithm 6, $u_{h,i,\ell,\pi,s}$ satisfies:

$$u_{h,i,\ell,\pi,s} = \frac{1}{n_{\mathtt{traj}}} \sum_{(x_{1:H}, a_{1:H}) \in \mathcal{D}_{\ell,\pi}} \phi(x_h, \pi_{h,i,s}(x_h)) \cdot \mathbb{I}\{s \cdot \phi(x_h, \pi_{h,i,s}(x_h))^\top v_{h,i} \geq 0\}.$$

Thus, by Hoeffding inequality, the fact that $\phi(\cdot,\cdot) \in \mathbb{B}(1)$, and the union bound over $h \in [H]$, $i \in [d]$, $\ell \in [0\,..\,h-1]$, $(\pi, v) \in \Psi_\ell$, and $s \in \{-,+\}$, there is an event of probability at least $1-\delta$ such that the desired inequality in (82) holds for all $h \in [H]$, $i \in [d]$, $\ell \in [0\,..\,h-1]$, $(\pi, v) \in \Psi_\ell$, and $s \in \{-,+\}$. □

Proof of Lemma I.2. Fix $\delta' \in (0,1)$. In this proof, we condition on the event of Lemma I.1 for $\delta = \delta'$; in particular, we assume that (82) holds for all $h \in [H]$, $i \in [d]$, $\ell \in [0\,..\,h-1]$, $(\pi, v) \in \Psi_\ell$, $s \in \{-,+\}$, and $\delta = \delta'$.

We note that (83) follows immediately from Lemma I.1 and the definitions of u_h, $\tilde{\pi}_{1:h}$, and v_h in Algorithm 6.

We now show (84). Fix $h \in [H]$, $\ell \in [0\,..\,h-1]$, and $(\pi, v) \in \Psi_\ell$. We have

$$
\begin{aligned}
&\mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}} \left[\max_{a \in \mathcal{A}} \|\phi(\boldsymbol{x}_h, a)\|_{(\beta I + U_h)^{-1}} \right] \\
&= \mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}} \left[\max_{a \in \mathcal{A}} \|(\beta I + U_h)^{-1/2} \phi(\boldsymbol{x}_h, a)\| \right] \\
&\le \sum_{i \in [d]} \mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}} \left[\max_{a \in \mathcal{A}} \left| e_i^\top (\beta I + U_h)^{-1/2} \phi(\boldsymbol{x}_h, a) \right| \right], \\
&= \sum_{i \in [d]} \mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}} \left[\max_{a \in \mathcal{A}} \left| v_{h,i}^\top \phi(\boldsymbol{x}_h, a) \right| \right], \quad \text{(by definition of } v_{h,i} \text{ in Algorithm 6)} \\
&= \sum_{i \in [d]} \mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}} \left[\max_{a \in \mathcal{A}} v_{h,i}^\top \phi(\boldsymbol{x}_h, a) \cdot \mathbb{I}\{ v_{h,i}^\top \phi(\boldsymbol{x}_h, a) \ge 0 \} \right] \\
&\quad + \sum_{i \in [d]} \mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}} \left[\max_{a \in \mathcal{A}} -v_{h,i}^\top \phi(\boldsymbol{x}_h, a) \cdot \mathbb{I}\{ v_{h,i}^\top \phi(\boldsymbol{x}_h, a) \le 0 \} \right], \\
&= \sum_{i \in [d]} \mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}} \left[v_{h,i}^\top \phi(\boldsymbol{x}_h, \pi_{h,i,+}(\boldsymbol{x}_h)) \cdot \mathbb{I}\{ v_{h,i}^\top \phi(\boldsymbol{x}_h, \pi_{h,i,+}(\boldsymbol{x}_h)) \ge 0 \} \right] \\
&\quad - \sum_{i \in [d]} \mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}} \left[v_{h,i}^\top \phi(\boldsymbol{x}_h, \pi_{h,i,-}(\boldsymbol{x}_h)) \cdot \mathbb{I}\{ v_{h,i}^\top \phi(\boldsymbol{x}_h, \pi_{h,i,-}(\boldsymbol{x}_h)) \le 0 \} \right], \quad \text{(by def. of } \pi_{h,i,\pm} \text{ in Alg. 6)}
\end{aligned}
$$

by Lemma I.1 and $\|v_{h,i}\| \le \beta^{-1}$,

$$
\begin{aligned}
&= \sum_{i \in [d]} \left(\langle v_{h,i}, u_{h,i,i,\ell,\pi,+} \rangle - \langle v_{h,i}, u_{h,i,i,\ell,\pi,-} \rangle \right) + \frac{4d}{\beta} \sqrt{\frac{2 \max_{j \in [H]} \log(4H^2 d |\Psi_j| / \delta')}{n_{\mathtt{traj}}}}, \\
&\le 2 \sum_{i \in [d]} |v_{h,i}^\top u_h| + \frac{4d}{\beta} \sqrt{\frac{2 \max_{j \in [H]} \log(4H^2 d |\Psi_j| / \delta')}{n_{\mathtt{traj}}}}, \quad \text{(by definition of } u_h \text{ in Algorithm 6)} \\
&\le 2\sqrt{d} \|u_h\|_{(\beta I + U_h)^{-1}} + \frac{4d}{\beta} \sqrt{\frac{2 \max_{j \in [H]} \log(4H^2 d |\Psi_j| / \delta')}{n_{\mathtt{traj}}}},
\end{aligned}
$$

where the last inequality follows by Jensen's inequality; in particular,

$$
\sum_{i \in [d]} |v_{h,i}^\top u_h| = \sum_{i \in [d]} |e_i^\top (\beta I + U_h)^{-1/2} u_h| = \|(\beta I + U_h)^{-1/2} u_h\|_1 \le \sqrt{d} \|(\beta I + U_h)^{-1/2} u_h\|_2 = \|u_h\|_{(\beta I + U_h)^{-1}}.
$$

This completes the proof. □

Appendix J. Guarantees of $\mathtt{FitValue}_h$

In this section, we present the guarantees for the `FitValue` subroutine of `Optimistic-PSDP` (see Appendix C.1 for an intuitive explanation of the results). The following subsection outlines these guarantees, with the proofs deferred to Appendix J. In Appendix J.3, we present additional structural results that are used in proving the guarantees.

J.1. Statement of Guarantees

Lemma J.1. *Let $h \in [H]$, $\Psi_{h-1}, \widehat{\pi}_{h+1:H}, U_{h+1:H}, W_{h+1:H}, \Pi', \mu, \nu, \lambda, \beta, \varepsilon, \delta, n$ be given and consider a call to $\texttt{FitValue}_h(\Psi_{h-1}, \widehat{\pi}_{h+1:H}, U_{h+1:H}, W_{h+1:H}; \Pi', \mu, \nu, \lambda, \beta, \varepsilon, \delta, n)$. Then, for any $\delta' \in (0,1)$, with probability at least $1-\delta'$, the variables in Algorithm 5 satisfy for all $\ell \in [h+1\,..\,H]$, $(\pi, v) \in \Psi_{h-1}$, and $\tilde{\pi} \in \Pi'$:*

$$\Big|\Delta_{h,\ell}(\pi, v, \tilde{\pi}) - \mathbb{E}^{\pi \circ_h \tilde{\pi} \circ_{h+1} \widehat{\pi}_{h+1:H}}\Big[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\} \cdot (b_\ell(\boldsymbol{x}_\ell, \boldsymbol{a}_\ell) - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}^{\mathsf{b}}_{h,\ell})\Big]\Big|$$
$$\leq \frac{4AH}{\lambda}\sqrt{\frac{\log(2H^2|\Psi_{h-1}|\mathcal{G}(\Pi', n)/\delta')}{n}},$$

where $\mathcal{G}$ is the growth function defined in Section 2.3.

Lemma J.2. *Let $h \in [H]$, $\Psi_{h-1}, \widehat{\pi}_{h+1:H}, U_{h+1:H}, W_{h+1:H}, \Pi', \mu, \nu, \lambda, \beta, \varepsilon, \delta, n$ be given and consider a call to $\texttt{FitValue}_h(\Psi_{h-1}, \widehat{\pi}_{h+1:H}, U_{h+1:H}, W_{h+1:H}; \Pi', \mu, \nu, \lambda, \beta, \varepsilon, \delta, n)$. Then, for any $\delta' \in (0,1)$, with probability at least $1-\delta'$, the variables in Algorithm 5 satisfy for all $(\pi, v) \in \Psi_{h-1}$ and $\tilde{\pi} \in \Pi'$:*

$$\Big|\mathbb{E}^{\pi \circ_h \tilde{\pi}}\Big[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\} \cdot \big(\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}^{\mathsf{r}}_h - Q^{\widehat{\pi}}_h(\boldsymbol{x}_h, \boldsymbol{a}_h)\big)\Big]\Big| \leq \varepsilon^{\mathsf{r}}_{\mathsf{reg}}(\delta', \Pi', |\Psi_{h-1}|, n),$$

where

$$\varepsilon^{\mathsf{r}}_{\mathsf{reg}}(\delta', \Pi', |\Psi|, n) := AH\sqrt{\lambda|\Psi|} + 2A\sqrt{\frac{d\log(\frac{2}{\delta'\lambda})}{n}} + \frac{4AH}{\lambda}\sqrt{\frac{\log(4|\Psi|\mathcal{G}(\Pi', n)/\delta')}{n}}, \tag{85}$$

where $\mathcal{G}$ is the growth function defined in Section 2.3.

Lemma J.3. *Let $h \in [H]$, $\Psi_{h-1}, \widehat{\pi}_{h+1:H}, U_{h+1:H}, W_{h+1:H}, \Pi', \mu, \nu, \lambda, \beta, \varepsilon, \delta, n$ be given and consider a call to $\texttt{FitValue}_h(\Psi_{h-1}, \widehat{\pi}_{h+1:H}, U_{h+1:H}, W_{h+1:H}; \Pi', \mu, \nu, \lambda, \beta, \varepsilon, \delta, n)$. Further, for any $\ell \in [h+1\,..\,H]$, let ρ_ℓ be the approximate design for Θ_ℓ in Section 2.2 and define*

$$G_\ell := W_\ell^{-1}\left(\sum_{\pi \in \operatorname{supp} \rho_\ell} \rho_\ell(\pi)\theta^\pi_\ell(\theta^\pi_\ell)^\top\right)W_\ell^{-1} \in \mathbb{R}^{d\times d}, \quad \text{and} \quad z_\| := \mathrm{Proj}_{\mathcal{S}(G_\ell,\zeta)}(z), \quad \forall z \in \mathbb{R}^d, \tag{86}$$

where $\zeta := (8d)^{-1}$. Then, for any $\delta' \in (0,1)$, with probability at least $1-\delta'$, the variables in Algorithm 5 satisfy for all $\ell \in [h+1\,..\,H]$, $(\pi, v) \in \Psi_{h-1}$, $\tilde{\pi} \in \Pi'$, and $z, y \in \mathbb{B}(1)$:

$$\left|\sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi}} \frac{A \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \geq 0, \tilde{\pi}(x_h) = a_h\}}{n} \cdot \Big(z_\|^\top B_\ell(x_\ell, a_\ell)y - \hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(x_h, a_h), z_\|, y]\Big)\right|$$
$$\leq \varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\delta', \Pi', |\Psi_{h-1}|, n),$$

where for $\tilde{d} := 4d\log\log d + 16$, $d_\nu := 5d\log(1+16H^4\nu^{-4})$, and $\mathcal{G}$ as in Section 2.3 (growth function):

$$\varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\delta', \Pi', |\Psi|, n) := \frac{\varepsilon}{256|\Psi|d_\nu d^3H^5} + \frac{8A\tilde{d}H^2}{\mu}\sqrt{2d\lambda|\Psi|} + 2A\sqrt{\frac{d\log(\frac{2}{\delta'\lambda})}{n}}$$
$$+\frac{16H^2A\tilde{d}d}{\mu}\sqrt{\frac{2\log(2^{12}d^3d_\nu H^6|\Psi|^2\mathcal{G}(\Pi', n)\varepsilon^{-1}/\delta')}{n}}. \tag{87}$$

Lemma J.4. *Let $h \in [H]$, $\Psi_{h-1}, \widehat{\pi}_{h+1:H}, U_{h+1:H}, W_{h+1:H}, \Pi', \mu, \nu, \lambda, \beta, \varepsilon, \delta, n$ such that $W_{h+1}, \dots, W_H \in \mathbb{S}^{d\times d}_{++}$ be given and consider a call to $\texttt{FitValue}_h(\Psi_{h-1}, \widehat{\pi}_{h+1:H}, U_{h+1:H}, W_{h+1:H}; \Pi', \mu, \nu, \lambda, \beta, \varepsilon, \delta, n)$. Then, with probability at least $1 - \frac{\delta}{2TH}$, the variables in Algorithm 5 are such that: if $\mathcal{H} \neq \varnothing$, then for all $\ell \in \mathcal{H}$:*

1. $\|w_\ell\| \le \nu^{-1}$;
2. $\|W_\ell w_\ell\| \ge 1/2$; *and*
3. $\sup_{\theta\in\Theta_\ell} |\langle \theta, w_\ell\rangle| \le 1$.

On the other hand, if $\mathcal{H} = \varnothing$, then for all $(\pi, v) \in \Psi_{h-1}$ and $\tilde{\pi} \in \Pi'$,

$$\begin{aligned}
&\left|\mathbb{E}^{\pi\circ_h\tilde{\pi}}\left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \ge 0\} \cdot (\widetilde{Q}_h(\boldsymbol{x}_h, \boldsymbol{a}_h) - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}_h)\right]\right| \\
&\le \frac{4AH^2}{\lambda}\sqrt{\frac{\log(4H^3T|\Psi_{h-1}|\mathcal{G}(\Pi', n)/\delta)}{n}} + \frac{32 d_\nu d \nu H^2 A}{\mu\lambda} \\
&\quad + 8 d_\nu d H \varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n) + \varepsilon^{\mathsf{r}}_{\mathsf{reg}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n),
\end{aligned} \tag{88}$$

where $d_\nu \coloneqq 5d\log(1 + 16H^4\nu^{-4})$, $\widetilde{Q}_h(x,a) = Q^{\widehat{\pi}}_h(x,a) + \mathbb{E}^{\widehat{\pi}}[\sum_{\ell=h+1}^H b_\ell(\boldsymbol{x}_h, \boldsymbol{a}_h) \mid \boldsymbol{x}_h = a, \boldsymbol{a}_h = a]$, for $(x,a) \in \mathcal{X}_h \times \mathcal{A}$, and $\varepsilon^{\mathsf{b}}_{\mathsf{reg}}$ [resp. $\varepsilon^{\mathsf{r}}_{\mathsf{reg}}$] is as in Lemma J.3 [resp. Lemma J.2].

Lemma J.5 (Guarantee of $\texttt{FitValue}_h$ for $\texttt{Optimistic-PSDP}$)**.** *Let $\varepsilon, \delta \in (0,1)$ be given and consider a call $\texttt{Optimistic-PSDP}(\Pi_{\mathsf{Bench}}, \varepsilon, \delta)$ (Algorithm 1) with Π_{Bench} as in Section 2.3. Further, let $\varepsilon^{\mathsf{b}}_{\mathsf{reg}}$ [resp. $\varepsilon^{\mathsf{r}}_{\mathsf{reg}}$] be as in Lemma J.3 [resp. Lemma J.2]. Then, there is an event $\mathcal{E}^{\mathsf{reg}}$ of probability at least $1 - \delta/2$, under which for all $t \in [T]$ and $h \in [H]$, the variable in Algorithm 1 satisfy*

- *$W_h^{(t)}$ is a valid ν-preconditioning for layer h, where ν is as in Algorithm 1; and*
- *If $W_{1:H}^{(t+1)} = W_{1:H}^{(t)}$ at the end of iteration t (i.e., the preconditioning is not updated during the calls to $\texttt{FitValue}_h$ for $h \in [H]$ at iteration t), then for all $(\pi, v) \in \Psi_{h-1}^{(t)}$ and $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$:*

$$\begin{aligned}
&\left|\mathbb{E}^{\pi\circ_h\tilde{\pi}}\left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \ge 0\} \cdot (Q_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h) - \widehat{Q}_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h))\right]\right| \\
&\le \varepsilon_{\mathsf{reg}} \coloneqq \frac{4AH^2}{\lambda}\sqrt{\frac{\log(4H^3T^2\mathcal{G}(\Pi_{\mathsf{Bench}}, n_{\mathsf{traj}})/\delta)}{n_{\mathsf{traj}}}} + \frac{160 d^2 \nu H^2 A \log(1 + 16H^4\nu^{-4})}{\mu\lambda} \\
&\quad + 50 d^2 H \varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\tfrac{\delta}{6HT}, \Pi_{\mathsf{Bench}}, T, n_{\mathsf{traj}}) \cdot \log(1 + 16H^4\nu^{-4}) + \varepsilon^{\mathsf{r}}_{\mathsf{reg}}(\tfrac{\delta}{6HT}, \Pi_{\mathsf{Bench}}, T, n_{\mathsf{traj}}),
\end{aligned}$$

where for all $(x,a) \in \mathcal{X}_h \times \mathcal{A}$, $Q_h^{(t)}(x,a) = Q^{\widehat{\pi}^{(t)}}_h(x,a) + \mathbb{E}^{\widehat{\pi}^{(t)}}[\sum_{\ell=h}^H b_\ell^{(t)}(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a]$, $\widehat{Q}_h^{(t)}(x,a) = \phi(x,a)^\top \hat{\theta}_h^{(t)} + b_h^{(t)}(x)$, and $b_h^{(t)}(x) \coloneqq \min\left(H, \frac{\varepsilon}{4H} \cdot \max_{a'\in\mathcal{A}} \|\phi(x,a')\|_{(\beta I + U_h^{(t)})^{-1}}\right) \cdot \mathbb{I}\{\|\varphi(x; W_h^{(t)})\| \ge \mu\}$.

J.2. Proofs

Proof of Lemma J.5. Fix $\delta \in (0,1)$. By the update rule in Line 5, we have that for all $t \in [T-1]$ and $h \in [H]$:

$$W_h^{(t+1)} = \left((W_h^{(t)})^{-2} + \sum_{\ell\in[h-1]} w_h^{(t,\ell)} (w_h^{(t,\ell)})^\top\right)^{-1/2}. \tag{89}$$

Now, for a given iteration $t \in [T-1]$ and $h \in [H]$, under the success event of Lemma J.4, the output $w_{h+1:H}^{(t,h)}$ of the call to $\texttt{FitValue}_h$ in Line 4 is such that

- For all $\ell \in [h+1\,..\,H]$ either $w_\ell^{(t,h)} = 0$; or

$$\|w_\ell^{(t,h)}\| \le \nu^{-1}, \qquad \|W_\ell^{(t)} w_\ell\| \ge 1/2, \qquad \text{and} \qquad \sup_{\theta \in \Theta_\ell} |\langle \theta, w_\ell^{(t,h)} \rangle| \le 1.$$

- When $w_\ell^{(t,h)} = 0$ for all $\ell \in [h+1\,..\,H]$ (corresponding to the case where $\mathcal{H}$ within the call to $\texttt{FitValue}_h$ is empty), we have that for all $(\pi, v) \in \Psi_{h-1}^{(t)}$ and $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$:

$$\left|\mathbb{E}^{\pi \circ_h \tilde{\pi}} \left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \ge 0\} \cdot (\widetilde{Q}_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h) - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}_h^{(t)})\right]\right| \le \varepsilon_{\mathsf{reg}},$$

  where $\varepsilon_{\mathsf{reg}}$ is as in the lemma statement; to get this bound we use (88) and the fact that $|\Psi_{h-1}^{(t)}| \le T$, for all $t \in [T]$.

Therefore, by Lemma J.4 and Lemma O.4 (essentially the union bound over $t \in [T-1]$ and $h \in [H]$), we get that there is event $\mathcal{E}^{\mathsf{reg}}$ of probability at least $1 - \delta/2$, under which for all $t \in [T-1]$ and $h \in [H]$:

1. For all $\ell \in [h+1\,..\,H]$ either $w_\ell^{(t,h)} = 0$; or

$$\|w_\ell^{(t,h)}\| \le \nu^{-1}, \quad \|W_\ell^{(t)} w_\ell\| \ge 1/2, \quad \text{and} \quad \sup_{\theta \in \Theta_\ell} |\langle \theta, w_\ell^{(t,h)} \rangle| \le 1;$$

2. When $w_\ell^{(t,h)} = 0$ for all $\ell \in [h+1\,..\,H]$ (corresponding to the case where $\mathcal{H}$ within the call to $\texttt{FitValue}_h$ is empty), we have that for all $(\pi, v) \in \Psi_{h-1}^{(t)}$ and $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$:

$$\left|\mathbb{E}^{\pi \circ_h \tilde{\pi}} \left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \ge 0\} \cdot (\widetilde{Q}_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h) - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}_h^{(t)})\right]\right| \le \varepsilon_{\mathsf{reg}}.$$

Item 1 together with (89) implies that for all $h \in [H]$ and $t \in [T-1]$, $W_h^{(t)}$ is a valid ν-preconditioning for layer h (Definition 2.5). Further, by Item 2, if $W_h^{(t+1)} = W_h^{(t)}$ at the end of iteration t, or equivalently if $w_\ell^{(t,h)} = 0$ for all $h \in [H]$ and $\ell \in [h+1\,..\,H]$, then for all $(\pi, v) \in \Psi_{h-1}^{(t)}$ and $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$:

$$\left|\mathbb{E}^{\pi \circ_h \tilde{\pi}} \left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \ge 0\} \cdot (\widetilde{Q}_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h) - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}_h^{(t)})\right]\right| \le \varepsilon_{\mathsf{reg}}.$$

Using the fact that $Q_h^{(t)}(x,a) = \widetilde{Q}_h^{(t)}(x,a) + b_h^{(t)}(x,a)$ for all $(x,a) \in \mathcal{X}_h \times \mathcal{A}$ and $h \in [H]$ completes the proof. □

Proof of Lemma J.1. By definition of $\Delta_{h,\ell}$ in Algorithm 5, we have for $(\pi, v) \in \Psi_{h-1}$ and $\tilde{\pi} \in \Pi'$:

$$\Delta_{h,\ell}(\pi, v, \tilde{\pi}) = \frac{A}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h) = a_h\} \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \ge 0\} \cdot \left(b_\ell(x_\ell) - \phi(x_h, a_h)^\top \hat{\theta}_{h,\ell}^{\mathsf{b}}\right),$$

where the trajectories in $\widehat{\mathcal{D}}_{h,\pi}$ are sampled i.i.d. according to $\mathbb{P}^{\pi \circ_h \pi_{\mathsf{unif}} \circ_{h+1} \widehat{\pi}_{h+1:H}}$. Note that we have the following:

- Thanks to the importance weight A and the indicator $\mathbb{I}\{\tilde{\pi}(\boldsymbol{x}_h) = \boldsymbol{a}_h\}$, we have
$$\mathbb{E}^{\pi\circ_h\pi_{\mathsf{unif}}\circ_{h+1}\widehat{\pi}_{h+1:H}}\left[\mathbb{I}\{\tilde{\pi}(\boldsymbol{x}_h) = \boldsymbol{a}_h\}\cdot\mathbb{I}\{\phi(\boldsymbol{x}_{h-1},\boldsymbol{a}_{h-1})^\top v \ge 0\}\cdot\left(b_\ell(\boldsymbol{x}_\ell) - \phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top\hat{\theta}^{\mathsf{b}}_{h,\ell}\right)\right]$$
$$= \mathbb{E}^{\pi\circ_h\tilde{\pi}\circ_{h+1}\widehat{\pi}_{h+1:H}}\left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1},\boldsymbol{a}_{h-1})^\top v \ge 0\}\cdot(b_\ell(\boldsymbol{x}_\ell,\boldsymbol{a}_\ell) - \phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top\hat{\theta}^{\mathsf{b}}_{h,\ell})\right];$$
- For all $x' \in \mathcal{X}_\ell$, we have $|b_\ell(x')| \le H$;
- We have $\|\hat{\theta}^{\mathsf{b}}_{h,\ell}\| \le \frac{H}{\lambda}$, (since $\hat{\theta}^{\mathsf{b}}_{h,\ell} = \Sigma_h^{-1}\sum_{(\pi,v)\in\Psi_{h-1}}\sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}}\phi(x_h,a_h)b_\ell(x_\ell)$ and $\sigma_{\min}(\Sigma_h) \ge \lambda n|\Psi_{h-1}|$) and so by Cauchy Schwarz inequality, we have for all $(x,a)\in\mathcal{X}_h\times\mathcal{A}$, $|\phi(x,a)^\top\hat{\theta}^{\mathsf{b}}_{h,\ell}| \le \frac{H}{\lambda}$.

Thus, by Lemma O.2 (generalization bound) and the union bound over $h\in[H]$, $\ell\in[h+1\,..\,H]$ and $(\pi,v)\in\Psi_{h-1}$, we get the desired result. □

Proof of Lemma J.2. All the variables in the proof are as in Algorithm 5. We start by showing that for all $(\pi,v)\in\Psi_{h-1}$, there is an event $\mathcal{E}(\pi,v)$ of probability at least $1-\delta'/|\Psi_{h-1}|$ under which (85) holds. Then, we apply the union bound to obtain the desired result.

Fix $(\pi,v)\in\Psi_{h-1}$. By the linear-Q^π assumption (Assumption 2.1), there is a $\theta^{\widehat{\pi}}_h\in\mathbb{B}(H)$ such that for all $(x,a)\in\mathcal{X}_h\times\mathcal{A}$:
$$Q^{\widehat{\pi}}_h(x,a) = \phi(x,a)^\top\theta^{\widehat{\pi}}_h. \tag{90}$$
Using this and Theorem O.1, we have that there is an event $\mathcal{E}$ of probability at least $1-\delta'/2$ such that
$$\|\hat{\theta}^{\mathsf{r}}_h - \theta^{\widehat{\pi}}_h\|_{\Sigma_h} \le \sqrt{\lambda n|\Psi_{h-1}|}\cdot\|\theta^{\widehat{\pi}}_h\| + \sqrt{2d\log(2/\delta') + \log\det(\Sigma_h\lambda^{-1}n^{-1}|\Psi_{h-1}|^{-1})}, \tag{91}$$
where
$$\Sigma_h \coloneqq \lambda n|\Psi_{h-1}|I + \sum_{(\pi',v')\in\Psi_{h-1}}\sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi'}}\phi(x_h,a_h)\phi(x_h,a_h)^\top,$$
and $\hat{\theta}^{\mathsf{r}}_h$ is as in Algorithm 6.

For the rest of the proof, we condition on $\mathcal{E}$. By (90) and Jensen's inequality, we have
$$\left|\frac{A}{n}\sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}}\mathbb{I}\{\tilde{\pi}(x_h) = a_h\}\cdot\mathbb{I}\{\phi(x_{h-1},a_{h-1})^\top v\ge 0\}\cdot\left(\phi(x_h,a_h)^\top\hat{\theta}^{\mathsf{r}}_h - Q^{\widehat{\pi}}_h(x_h,a_h)\right)\right|$$
$$\le A\sqrt{\frac{1}{n}\sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}}\mathbb{I}\{\tilde{\pi}(x_h) = a_h, \phi(x_{h-1},a_{h-1})^\top v\ge 0\}^2\cdot\langle\phi(x_h,a_h),\hat{\theta}^{\mathsf{r}}_h - \theta^{\widehat{\pi}}_h\rangle^2},$$
$$\le A\sqrt{\frac{1}{n}\sum_{(\pi',v')\in\Psi_{h-1}}\sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi'}}\langle\phi(x_h,a_h),\hat{\theta}^{\mathsf{r}}_h - \theta^{\widehat{\pi}}_h\rangle^2},$$
$$\le \frac{A}{\sqrt{n}}\|\hat{\theta}^{\mathsf{r}}_h - \theta^{\widehat{\pi}}_h\|_{\Sigma_h},$$
$$\le A\sqrt{\lambda|\Psi_{h-1}|}\|\theta^{\widehat{\pi}}_h\| + A\sqrt{\frac{2d\log(2/\delta')}{n} + \frac{\log\det(\Sigma_h\lambda^{-1}n^{-1}|\Psi_{h-1}|^{-1})}{n}}, \quad \text{(by (91))}$$
$$\le AH\sqrt{\lambda|\Psi_{h-1}|} + A\sqrt{\frac{2d\log(2/\delta')}{n} + \frac{\log\det(\Sigma_h\lambda^{-1}n^{-1}|\Psi_{h-1}|^{-1})}{n}}, \tag{92}$$

where the last inequality follows by the fact that $\theta_h^{\widehat{\pi}} \in \mathbb{B}(H)$. Now, by Jensen's inequality and the fact that $\|\Sigma_h\|_{\mathrm{op}} \leq (1+\lambda) \cdot |\Psi_{h-1}| \cdot n$, we have

$$\log\det(\Sigma_h \lambda^{-1} n^{-1} |\Psi_{h-1}|^{-1}) \leq d \log \mathrm{Tr}(\Sigma_h d^{-1} \lambda^{-1} n^{-1} |\Psi_{h-1}|^{-1}) \leq d \log(1+\lambda^{-1}).$$

Plugging this into (92), we get that

$$\begin{aligned}
&\left| \frac{A}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h) = a_h\} \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \geq 0\} \cdot \left(\phi(x_h, a_h)^\top \hat{\theta}_h^{\mathsf{r}} - Q_h^{\widehat{\pi}}(x_h, a_h)\right) \right| \\
&\leq AH\sqrt{\lambda |\Psi_{h-1}|} + 2A\sqrt{n^{-1} d \log(2\lambda^{-1}/\delta')},
\end{aligned} \tag{93}$$

Now, by Lemma O.2 (generalization bound) and the fact that $\hat{\theta}_h^{\mathsf{r}} \in \mathbb{B}(H/\lambda)$; since

$$\hat{\theta}_h^{\mathsf{r}} = \Sigma_h^{-1} \sum_{(\pi', v') \in \Psi_{h-1}} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi'}} \phi(x_h, a_h) \cdot \sum_{\ell=h}^{H} r_\ell$$

and $r_\ell \in [0,1]$ for all $\ell \in [H]$, there is an event $\mathcal{E}'(\pi, v)$ of probability at least $1 - \delta'/(2|\Psi_{h-1}|)$ under which for all $\tilde{\pi} \in \Pi'$:

$$\begin{aligned}
&\left| \mathbb{E}^{\pi \circ_h \tilde{\pi}} \left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\} \cdot \left(\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}_h^{\mathsf{r}} - Q_h^{\widehat{\pi}}(\boldsymbol{x}_h, \boldsymbol{a}_h)\right) \right] \right| \\
&\leq \left| \frac{A}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h) = a_h\} \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \geq 0\} \cdot \left(\phi(x_h, a_h)^\top \hat{\theta}_h^{\mathsf{r}} - Q_h^{\widehat{\pi}}(x_h, a_h)\right) \right| \\
&\quad + \frac{4AH}{\lambda} \sqrt{\frac{\log(4|\Psi_{h-1}| \mathcal{G}_h(\Pi', n)/\delta')}{n}}.
\end{aligned} \tag{94}$$

Thus, by combining (93) and (94), we get that under $\mathcal{E} \cap \mathcal{E}'(\pi, v)$:

$$\begin{aligned}
&\left| \mathbb{E}^{\pi \circ_h \tilde{\pi}} \left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\} \cdot \left(\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}_h^{\mathsf{r}} - Q_h^{\widehat{\pi}}(\boldsymbol{x}_h, \boldsymbol{a}_h)\right) \right] \right| \\
&\leq AH\sqrt{\lambda |\Psi_{h-1}|} + 2A\sqrt{\frac{d \log(\frac{2}{\delta' \lambda})}{n}} + \frac{4AH}{\lambda} \sqrt{\frac{\log(4|\Psi_{h-1}| \mathcal{G}_h(\Pi', n)/\delta')}{n}},
\end{aligned}$$

for all $\tilde{\pi} \in \Pi'$. Now, the desired result follows by the union bound over $(\pi, v) \in \Psi_{h-1}$. □

Proof of Lemma J.3. Let $\tilde{d}$, d_ν, and ζ be as in the lemma statement and define

$$\varepsilon_{\mathtt{disc}} \coloneqq \frac{\varepsilon \zeta \mu}{1024 |\Psi_{h-1}| d_\nu d^3 H^6 A}.$$

All other variables in this proof are as in Algorithm 5.

We start by showing that for all $\ell \in [h+1\,..\,H]$, $(\pi, v) \in \Psi_{h-1}$, and $z, y \in \mathbb{B}(1)$, there is an event $\mathcal{E}(\ell, \pi, v, z, y)$ of probability at least $1 - \delta'(\varepsilon_{\mathtt{disc}}/3)^{2d}/(2|\Psi_{h-1}|H)$ under which the inequality in the lemma's statement holds. Then, we apply the union bound to obtain the desired result.

Fix $\ell \in [h+1\,..\,H]$, $(\pi, v) \in \Psi_{h-1}$, and $z, y \in \mathbb{B}(1)$. Note that B_ℓ satisfies

$$\forall x \in \mathcal{X}, \quad B_\ell(x) = b_\ell(x) \cdot \frac{\varphi(x; W_\ell) \varphi(x; W_\ell)^\top}{\|\varphi(x; W_\ell)\|^2},$$

where b_ℓ is as in Line 7 of Algorithm 5. Combining this with the fact that $\|b_\ell\|_\infty \le H$ and Lemma J.6 implies that $F : x \mapsto z_\|^\top B_\ell(x) y$ is α-admissible (Definition 2.4) with $\alpha = \sqrt{\zeta}\mu/H$, and so by Lemma 2.2 there exists $\theta_{h,\ell} \in \mathbb{B}(4\tilde{d}H/\alpha)$ such that

$$\forall (x,a) \in \mathcal{X}_h \times \mathcal{A}, \quad \mathbb{E}^{\widehat{\pi}_{h+1:H}}[z_\|^\top B_\ell(\boldsymbol{x}_\ell) y \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] = \phi(x,a)^\top \theta_{h,\ell}. \tag{95}$$

Thus, by the law of total expectation and Lemma O.2 (generalization bound) with the facts that $\|\theta_{h,\ell}\| \le 4\tilde{d}H/\alpha$ and $\|B_\ell(x)\|_{\mathrm{op}} \le H$, for all $x \in \mathcal{X}_\ell$, there is an event $\mathcal{E}(\ell,\pi,v,z,y)$ of probability at least $1 - \delta' \cdot (\varepsilon_{\mathtt{disc}}/3)^{2d}/(2H|\Psi_{h-1}|)$ such that for all $\tilde{\pi} \in \Pi'$:

$$\left| \frac{A}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h) = a_h\} \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \ge 0\} \cdot \left(z_\|^\top B_\ell(x_\ell) y - \phi(x_h, a_h)^\top \theta_{h,\ell} \right) \right|$$
$$\le \frac{16HA\tilde{d}}{\alpha} \sqrt{\frac{2d \log(12H|\Psi_{h-1}|\mathcal{G}_h(\Pi', n)\varepsilon_{\mathtt{disc}}^{-1}/\delta')}{n}}, \tag{96}$$

On the other hand, by (95) and Theorem O.1, we have that there is an event $\mathcal{E}'$ of probability at least $1 - \delta'/2$ such that

$$\|\hat{\theta}_{h,\ell} - \theta_{h,\ell}\|_{\Sigma_h} \le \sqrt{\lambda n |\Psi_{h-1}|} \cdot \|\theta_{h,\ell}\| + \sqrt{2d \log(2/\delta') + \log\det(\Sigma_h \lambda^{-1} n^{-1} |\Psi_{h-1}|^{-1})}, \tag{97}$$

where

$$\Sigma_h \coloneqq \lambda n |\Psi_{h-1}| I + \sum_{(\pi',v') \in \Psi_{h-1}} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi'}} \phi(x_h, a_h) \phi(x_h, a_h),$$
$$\text{and} \quad \hat{\theta}_{h,\ell} \coloneqq \Sigma_h^{-1} \sum_{(\pi',v') \in \Psi_{h-1}} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi'}} \phi(x_h, a_h) \cdot z_\|^\top B_\ell(x_\ell) y.$$

We now condition of $\mathcal{E}(\ell,\pi,v,z,y) \cap \mathcal{E}'$. Using that $\hat{\theta}_{h,\ell} = \hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\cdot, z_\|, y]$, we get for all $\tilde{\pi} \in \Pi'$:

$$\left| \frac{A}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h) = a_h, \phi(x_{h-1}, a_{h-1})^\top v \ge 0\} \cdot \left(z_\|^\top B_\ell(x_\ell) y - \hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(x_h, a_h), z_\|, y] \right) \right|$$
$$= \left| \frac{A}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h) = a_h\} \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \ge 0\} \cdot \left(z_\|^\top B_\ell(x_\ell) y - \phi(x_h, a_h)^\top \hat{\theta}_{h,\ell} \right) \right|,$$

and so by the triangle inequality, we get

$$\le \left| \frac{A}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h) = a_h\} \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \ge 0\} \cdot \left(\phi(x_h, a_h)^\top \hat{\theta}_{h,\ell} - \phi(x_h, a_h)^\top \theta_{h,\ell} \right) \right|$$
$$+ \left| \frac{A}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h) = a_h\} \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \ge 0\} \cdot \left(z_\|^\top B_\ell(x_\ell) y - \phi(x_h, a_h)^\top \theta_{h,\ell} \right) \right|,$$
$$\le \left| \frac{A}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h) = a_h\} \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \ge 0\} \cdot \left(\phi(x_h, a_h)^\top \hat{\theta}_{h,\ell} - \phi(x_h, a_h)^\top \theta_{h,\ell} \right) \right|$$
$$+ \frac{16HA\tilde{d}}{\alpha} \sqrt{\frac{2d \log(12H|\Psi_{h-1}|\mathcal{G}_h(\Pi', n)\varepsilon_{\mathtt{disc}}^{-1}/\delta')}{n}}, \tag{98}$$

where the last inequality follows by (96). We now focus on bounding the first term on the right-hand side of (98) using (97). By Jensen's inequality, we have for all $\tilde{\pi} \in \Pi'$:

$$\begin{aligned}
&\left|\frac{A}{n}\sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h)=a_h\}\cdot\mathbb{I}\{\phi(x_{h-1},a_{h-1})^\top v\geq 0\}\cdot\left(\phi(x_h,a_h)^\top\hat{\theta}_{h,\ell}-\phi(x_h,a_h)^\top\theta_{h,\ell}\right)\right| \\
&\leq A\sqrt{\frac{1}{n}\sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h)=a_h\}\cdot\mathbb{I}\{\phi(x_{h-1},a_{h-1})^\top v\geq 0\}^2\cdot\langle\phi(x_h,a_h),\hat{\theta}_{h,\ell}-\theta_{h,\ell}\rangle^2}, \\
&\leq A\sqrt{\frac{1}{n}\sum_{(\pi',v')\in\Psi_{h-1}}\sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi'}}\langle\phi(x_h,a_h),\hat{\theta}_{h,\ell}-\theta_{h,\ell}\rangle^2}, \\
&= \frac{A}{\sqrt{n}}\|\hat{\theta}_{h,\ell}-\theta_{h,\ell}\|_{\Sigma_h}, \\
&\leq A\sqrt{\lambda|\Psi_{h-1}|}\|\theta_{h,\ell}\|+A\sqrt{\frac{2d\log(2/\delta')}{n}+\frac{\log\det(\Sigma_h\lambda^{-1}n^{-1}|\Psi_{h-1}|^{-1})}{n}}, \\
&\leq \frac{4A\tilde{d}H}{\alpha}\sqrt{\lambda|\Psi_{h-1}|}+A\sqrt{\frac{4d\log(2/\delta')}{n}+\frac{\log\det(\Sigma_h\lambda^{-1}n^{-1}|\Psi_{h-1}|^{-1})}{n}}, \qquad (99)
\end{aligned}$$

where the last inequality follows by the fact that $\theta_{h,\ell}\in\mathbb{B}(4H\tilde{d}/\alpha)$ (where we recall that $\alpha=\sqrt{\zeta}\mu/H$). Now, by Jensen's inequality and the fact that $\|\Sigma_h\|_{\mathrm{op}}\leq(1+\lambda)\cdot|\Psi_{h-1}|\cdot n$, we have

$$\log\det(\Sigma_h\lambda^{-1}n^{-1}|\Psi_{h-1}|^{-1})\leq d\log\mathrm{Tr}(\Sigma_h d^{-1}\lambda^{-1}n^{-1}|\Psi_{h-1}|^{-1})\leq d\log(1+\lambda^{-1}).$$

Plugging this into (99), we get that for all $\tilde{\pi}\in\Pi'$:

$$\begin{aligned}
&\left|\frac{A}{n}\sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h)=a_h\}\cdot\mathbb{I}\{\phi(x_{h-1},a_{h-1})^\top v\geq 0\}\cdot\left(\phi(x_h,a_h)^\top\hat{\theta}_{h,\ell}-\phi(x_h,a_h)^\top\theta_{h,\ell}\right)\right| \\
&\leq \frac{4A\tilde{d}H}{\alpha}\sqrt{\lambda|\Psi_{h-1}|}+2A\sqrt{\frac{d\log(\frac{2}{\lambda\delta'})}{n}},
\end{aligned}$$

Combining this with (98), we get that

$$\begin{aligned}
&\left|\frac{A}{n}\sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h)=a_h\}\cdot\mathbb{I}\{\phi(x_{h-1},a_{h-1})^\top v\geq 0\}\cdot\left(z_\parallel^\top B_\ell(x_\ell)y-\hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(x_h,a_h),z_\parallel,y]\right)\right| \\
&\leq \frac{4A\tilde{d}H}{\alpha}\sqrt{\lambda|\Psi_{h-1}|}+2A\sqrt{\frac{d\log(\frac{2}{\lambda\delta'})}{n}}+\frac{16HA\tilde{d}}{\alpha}\sqrt{\frac{2d\log(12H|\Psi_{h-1}|\mathcal{G}_h(\Pi',n)\varepsilon_{\mathtt{disc}}^{-1}/\delta')}{n}},
\end{aligned}$$

for all $\tilde{\pi}\in\Pi'$. Now, we apply Lemma O.5 instantiated with

- $\mathcal{K}\coloneqq\left\{\mathrm{Proj}_{\mathcal{S}(G_\ell,\zeta)}(z):z\in\mathbb{B}(1)\right\}$;
- $\varepsilon'=\varepsilon_{\mathtt{disc}}=\frac{\varepsilon\zeta\mu}{1024|\Psi_{h-1}|d_\nu d^3H^6A}$;

- $\mathcal{Z} = \Pi'$;
- $L = 2AH/\lambda$;
- and

$$\boldsymbol{M}^{\tilde{\pi}} = \frac{A}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h) = a_h\} \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \geq 0\} \cdot \left(B_\ell(x_\ell) - \hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(x_h, a_h), \cdot, \cdot]\right),$$

to get that there exists an event $\mathcal{E}(\ell, \pi, v)$ of probability at least $1 - \delta'/(2H|\Psi_{h-1}|)$ under which we have for all $z, y \in \mathbb{B}(1)$ and $\tilde{\pi} \in \Pi'$:

$$\left| \frac{A}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi}} \mathbb{I}\{\tilde{\pi}(x_h) = a_h\} \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v \geq 0\} \cdot \left(z_\parallel^\top B_\ell(x_\ell) y - \hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(x_h, a_h), z_\parallel, y]\right) \right|$$

$$\leq \frac{4AH\varepsilon_{\mathsf{disc}}}{\lambda} + \frac{4A\tilde{d}H}{\alpha}\sqrt{\lambda|\Psi_{h-1}|} + 2A\sqrt{\frac{d\log(\frac{2}{\delta'\lambda})}{n}} + \frac{16HA\tilde{d}}{\alpha}\sqrt{\frac{2d\log(12H|\Psi_{h-1}|\mathcal{G}_h(\Pi', n)\varepsilon^{-1}_{\mathsf{disc}}/\delta')}{n}}.$$

Note that the condition of Lemma O.5 that requires $\|\boldsymbol{M}^{\tilde{\pi}}\|_{\mathrm{op}} \leq L = 2AH/\lambda$, for all $\tilde{\pi} \in \Pi'$, is indeed satisfied by definition of $\hat{\vartheta}^{\mathsf{b}}_{h,\ell}$ in Algorithm 5 and the fact that $\sigma_{\min}(\Sigma_h) \geq \lambda|\Psi_{h-1}|n$. Now, the desired result follows by the union bound over $\ell \in [h+1\,..\,H]$ and $(\pi, v) \in \Psi_{h-1}$ and the facts that $\varepsilon_{\mathsf{disc}} = \frac{\varepsilon\zeta\mu}{1024|\Psi_{h-1}|d_\nu d^3 H^6 A}$ and $\zeta = \frac{1}{8d}$. □

Proof of Lemma J.4. In this proof, we condition on the intersection of the events of Lemma J.1, Lemma J.2, and Lemma J.3 for $\delta' = \delta/(6HT)$; by the union bound, the probability of this event intersection is at least $1 - \frac{\delta}{2HT}$. To simplify notation in this proof, we let $c \coloneqq 20d\log(1 + 16H^4\nu^{-4})$ and for any $z \in \mathbb{R}^d$, we write $z_\parallel \coloneqq \mathrm{Proj}_{\mathcal{S}(G_\ell,\zeta)}(z)$ and $z_\perp \coloneqq \mathrm{Proj}_{\mathcal{S}(G_\ell,\zeta)^\perp}(z)$, where G_ℓ is as in (86) and $\zeta \coloneqq \frac{1}{8d}$.

Case where $\mathcal{H} \neq \varnothing$. First, assume that $\mathcal{H} \neq \varnothing$. Fix $\ell \in \mathcal{H}$ and let

$$((\pi_{h,\ell}, v_{h,\ell}), \tilde{\pi}_{h,\ell}) \in \operatorname*{arg\,max}_{((\pi,v),\tilde{\pi}) \in \Psi_{h-1} \times \Pi'} \Delta_{h,\ell}(\pi, v, \tilde{\pi}),$$

where $\Delta_{h,\ell}$ is as in Algorithm 5. By the definition of $\mathcal{H}$, we have

$$|\Delta_{h,\ell}(\pi_{h,\ell}, v_{h,\ell}, \tilde{\pi}_{h,\ell})| \geq 2dc\varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n) + \frac{8cd\nu HA}{\mu\lambda}, \tag{100}$$

where $\varepsilon^{\mathsf{b}}_{\mathsf{reg}}$ is as in Lemma J.3. Moving forward, we let

$$M \coloneqq \frac{A}{n} \sum_{(x_{1:H}, a_{1:H}, r_{1:H}) \in \widehat{\mathcal{D}}_{h,\pi_{h,\ell}}} \mathbb{I}\{\tilde{\pi}_{h,\ell}(x_h) = a_h\} \cdot \mathbb{I}\{\phi(x_{h-1}, a_{h-1})^\top v_{h,\ell} \geq 0\} \cdot \left(B_\ell(x_\ell) - \hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(x_h, a_h), \cdot, \cdot]\right),$$

and note that $M \in \mathbb{R}^{d\times d}$ satisfies

$$\mathrm{Tr}(M) = \Delta_{h,\ell}(\pi_{h,\ell}, v_{h,\ell}, \tilde{\pi}_{h,\ell}). \tag{101}$$

Furthermore, the variables $(z_\ell, \tilde{z}_\ell, w_\ell)$ in Algorithm 5 satisfy

$$z_\ell \in \arg\max_{z\in\mathbb{B}(1)} |z^\top M z|, \quad \tilde{z}_\ell = \mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(z_\ell), \quad \text{and} \quad w_\ell = W_\ell^{-1}\tilde{z}_\ell.$$

Now, by definition of the trace, there exist unit vectors (eigenvectors in this case) $c_1, \dots, c_d$ such that $\mathrm{Tr}(M) = \sum_{i\in[d]} c_i^\top M c_i$. Combining this with (101) and (100), we get that

$$d\cdot|z_\ell^\top M z_\ell| \geq \left|\sum_{i\in[d]} c_i^\top M c_i\right| \geq 2dc\varepsilon_{\mathsf{reg}}^{\mathsf{b}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n) + \frac{8cd\nu HA}{\mu\lambda}. \tag{102}$$

Now, let $\overline{M} \coloneqq \mathsf{sgn}(z_\ell^\top M z_\ell)\cdot M$ and note that $z_\ell \in \arg\max_{z\in\mathbb{B}(1)} z_\ell^\top \overline{M} z_\ell$. By (102), we have

$$z_\ell^\top \overline{M} z_\ell > 2c\varepsilon_{\mathsf{reg}}^{\mathsf{b}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n) + \frac{8c\nu HA}{\mu\lambda}, \tag{103}$$

and so by Lemma J.8 and the definition of M, we have

$$\tilde{z}_\ell^\top \overline{M} \tilde{z}_\ell > 2c\varepsilon_{h}^{\mathsf{b}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n) + \frac{8(c-1)\nu HA}{\mu\lambda}. \tag{104}$$

We use this to show the claims of the lemma.

Proving Item 1. We start by showing $\|w_\ell\| \leq \nu^{-1}$; that is, Item 1 of the lemma. Let $\lambda_1, \dots, \lambda_d$ be the eigenvalues of W_ℓ and let $w_1, \dots, w_d \in \mathbb{R}^d$ be an orthonormal basis such that $W_\ell w_i = \lambda_i w_i$, for all $i \in [d]$ (such a basis exists because $W_\ell \in \mathbb{S}_{++}^{d\times d}$). With this, we can write

$$W_\ell = \sum_{i=1}^d \lambda_i w_i w_i^\top.$$

Let P_ℓ be the matrix whose columns are $w_1, \dots, w_d$, and note that $P_\ell^\top P_\ell = I$ since $w_1, \dots, w_d$ are orthonormal. With this, we have

$$\begin{aligned}
W_\ell^{-1}\mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(z_\ell) &= W_\ell^{-1} \sum_{i\in[d]:\lambda_i\geq\nu} w_i w_i^\top z_\ell, \\
&= \left(\sum_{i=1}^d \lambda_i w_i w_i^\top\right)^{-1} \sum_{i\in[d]:\lambda_i\geq\nu} w_i w_i^\top z_\ell, \\
&= \left(\sum_{i=1}^d \lambda_i P_\ell e_i e_i^\top P_\ell^\top\right)^{-1} \sum_{i\in[d]:\lambda_i\geq\nu} P_\ell e_i e_i^\top P_\ell^\top z_\ell, \\
&= P_\ell \left(\sum_{i=1}^d \lambda_i e_i e_i^\top\right)^{-1} P_\ell^\top \sum_{i\in[d]:\lambda_i\geq\nu} P_\ell e_i e_i^\top P_\ell^\top z_\ell, \\
&= P_\ell \left(\sum_{i=1}^d \lambda_i e_i e_i^\top\right)^{-1} \sum_{i\in[d]:\lambda_i\geq\nu} e_i e_i^\top P_\ell^\top z_\ell, \\
&= P_\ell \mathrm{diag}(\lambda_1^{-1}, \dots, \lambda_d^{-1}) \sum_{i\in[d]:\lambda_i\geq\nu} e_i e_i^\top P_\ell^\top z_\ell, \\
&= P_\ell \sum_{i\in[d]:\lambda_i\geq\nu} \lambda_i^{-1} e_i e_i^\top P_\ell^\top z_\ell,
\end{aligned}$$

Thus, taking the norm and using that $\|P_\ell\|_{\mathrm{op}} = 1$, we get

$$\|w_\ell\| = \|W_\ell^{-1}\tilde{z}_\ell\| = \|W_\ell^{-1}\mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(z_\ell)\| \le \|P_\ell\|_{\mathrm{op}}\sqrt{\sum_{i\in[d]:\lambda_i\ge\nu}\lambda_i^{-2}(e_i^\top P_\ell z_\ell)^2},$$
$$\le \nu^{-1}\|P_\ell z_\ell\| \le \nu^{-1}\|P_\ell\|_{\mathrm{op}}\|z_\ell\| \le \nu^{-1}.$$

This shows Item 1.

Proving Item 2. To prove Item 2, we need to show that $\|W_\ell w_\ell\| = \|\tilde{z}_\ell\| \ge \frac{1}{2}$. Using that $z_\ell \in \arg\max_{z\in\mathbb{B}(1)} z^\top \overline{M} z$ and Lemma J.8,

$$\|\tilde{z}_\ell\|^{-2}\cdot \tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell \le z_\ell^\top \overline{M} z_\ell \le \tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell + \frac{8\nu HA}{\mu\lambda},$$
$$\le \tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell + \frac{1}{c} z_\ell^\top \overline{M} z_\ell, \tag{105}$$

where the last inequality follows by (103). Rearranging (105) and using that $c \ge 3$, implies that $\|\tilde{z}_\ell\|^2 \ge \frac{2}{3}$. Therefore, we get that

$$\|W_\ell w_\ell\|^2 = \|\tilde{z}_\ell\|^2 \ge \frac{2}{3},$$

satisfying the inequality in Item 2.

Proving Item 3. It remains to prove Item 3. We will start by showing that $\|(\tilde{z}_\ell)_\parallel\|^2 \le \frac{1}{c}$ and use this to prove Item 3. Since $\overline{M}$ is symmetric, we can decompose $\tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell$ as

$$\tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell = (\tilde{z}_\ell)_\parallel^\top \overline{M}\tilde{z}_\ell + (\tilde{z}_\ell)_\parallel^\top \overline{M}(\tilde{z}_\ell)_\perp + (\tilde{z}_\ell)_\perp^\top \overline{M}(\tilde{z}_\ell)_\perp .$$

By the conditioning on the event of Lemma J.3 with $\delta' = \delta/(6HT)$, we have $(\tilde{z}_\ell)_\parallel^\top \overline{M}\tilde{z}_\ell \le \varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\frac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n)$ and $(\tilde{z}_\ell)_\parallel^\top \overline{M}(\tilde{z}_\ell)_\perp \le \varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\frac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n)$. Therefore, we have

$$\tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell \le 2\varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n) + (\tilde{z}_\ell)_\perp^\top \overline{M}(\tilde{z}_\ell)_\perp. \tag{106}$$

Combining this with (104), we get that $\|(\tilde{z}_\ell)_\perp\| \ne 0$. Let $\bar{z}_\ell = (\tilde{z}_\ell)_\perp/\|(\tilde{z}_\ell)_\perp\|$. Since $z_\ell \in \arg\max_{z\in\mathbb{B}(1)} z^\top \overline{M} z$, we have

$$\|(\tilde{z}_\ell)_\perp\|^{-2}\cdot(\tilde{z}_\ell)_\perp^\top \overline{M}(\tilde{z}_\ell)_\perp = (\bar{z}_\ell)^\top \overline{M}\bar{z}_\ell \le z_\ell^\top \overline{M} z_\ell. \tag{107}$$

On the other hand, using Lemma J.8 and (106), we have

$$z_\ell^\top \overline{M} z_\ell \le \tilde{z}_\ell^\top \overline{M}\tilde{z}_\ell + \frac{8\nu HA}{\mu\lambda},$$
$$\le 2\varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n) + \frac{8\nu HA}{\mu\lambda} + (\tilde{z}_\ell)_\perp^\top \overline{M}(\tilde{z}_\ell)_\perp,$$
$$\le \frac{1}{c} z_\ell^\top \overline{M} z_\ell + (\tilde{z}_\ell)_\perp^\top \overline{M}(\tilde{z}_\ell)_\perp, \tag{108}$$

where the last inequality follows by (103). Rearranging (108) gives

$$z_\ell^\top \overline{M} z_\ell \le \frac{c}{c-1}(\tilde{z}_\ell)_\perp^\top \overline{M}(\tilde{z}_\ell)_\perp.$$

Combining this with (107), dividing by $(\tilde{z}_\ell)_\perp^\top \overline{M}(\tilde{z}_\ell)_\perp$, and rearranging, we get that $\|(\tilde{z}_\ell)_\perp\|^2 \geq \frac{c-1}{c}$, and so

$$\|(\tilde{z}_\ell)_\parallel\|^2 \leq \frac{1}{c}, \tag{109}$$

since $1 = \|z_\ell\|^2 \geq \|\tilde{z}_\ell\|^2 = \|(\tilde{z}_\ell)_\parallel\|^2 + \|(\tilde{z}_\ell)_\perp\|^2$. We use (109) to prove Item 3. Note that since

$$G_\ell = W_\ell^{-1}\left(\sum_{\pi \in \operatorname{supp} \rho_\ell} \rho_\ell(\pi)\theta_\ell^\pi(\theta_\ell^\pi)^\top\right)W_\ell^{-1},$$

where ρ_ℓ is the approximate design for Θ_ℓ in Section 2.2, we have

$$\forall \theta \in W_\ell^{-1}\Theta_\ell, \quad \|\theta\|_{G_\ell^\dagger} \leq \sqrt{2d}. \tag{110}$$

With this, we have

$$\begin{aligned}
\sup_{\theta \in \Theta_\ell} |\langle \theta, w_\ell\rangle| &= \sup_{\theta \in \Theta_\ell} |\langle \theta, W_\ell^{-1}\tilde{z}_\ell\rangle|, \\
&= \sup_{\theta \in W_\ell^{-1}\Theta_\ell} |\langle \theta, \tilde{z}_\ell\rangle|, \\
&\leq \sup_{\theta \in W_\ell^{-1}\Theta_\ell} \|\theta\|\|(\tilde{z}_\ell)_\parallel\| + \sup_{\theta \in W_\ell^{-1}\Theta_\ell} |\langle \theta, (\tilde{z}_\ell)_\perp\rangle|,
\end{aligned}$$

and so by Lemma 2.4 and $d_\nu \coloneqq 5d\log(1 + 16H^4\nu^{-4})$,

$$\leq \sqrt{\frac{d_\nu}{c}} + \sup_{\theta \in W_\ell^{-1}\Theta_\ell} \|\theta\|_{G_\ell^\dagger} \cdot \|(\tilde{z}_\ell)_\perp\|_{G_\ell},$$

and so by (110), we have

$$\begin{aligned}
&\leq \sqrt{\frac{d_\nu}{c}} + \sqrt{2d} \cdot \|(\tilde{z}_\ell)_\perp\|_{G_\ell}, \\
&\leq \sqrt{\frac{d_\nu}{c}} + \sqrt{2d \cdot (\tilde{z}_\ell)_\perp^\top(\zeta I)(\tilde{z}_\ell)_\perp}, \quad \text{(see below)} \\
&\leq \sqrt{\frac{d_\nu}{c}} + \sqrt{2d \cdot \zeta}, \\
&= \sqrt{\frac{d_\nu}{c}} + \frac{1}{2}, \quad (\text{since } \zeta = (8d)^{-1}) \\
&\leq 1,
\end{aligned} \tag{111}$$

where in (111) follows from the fact that $(\cdot)_\perp = \mathrm{Proj}_{\mathcal{S}(G_\ell,\zeta)^\perp}(\cdot)$; and the last inequality uses that $c = 20d\log(1 + 16\nu^{-4}H^4) = 4d_\nu$.

Case where $\mathcal{H} = \varnothing$. We now consider the case where $\mathcal{H} = \varnothing$. By definition of $\mathcal{H}$ in Algorithm 5, we have that for all $\ell \in [h+1\,..\,H]$, $(\pi, v) \in \Psi_{h-1}$, and $\tilde{\pi} \in \Pi'$:

$$|\Delta_{h,\ell}(\pi, v, \tilde{\pi})| \le 2cd\varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n) + \frac{8cd\nu HA}{\mu\lambda}. \tag{112}$$

On the other hand, by the conditioning on the event of Lemma J.1 with $\delta' = \delta/(6HT)$, we have for all $\ell \in [h+1\,..\,H]$, $(\pi, v) \in \Psi_{h-1}$, and $\tilde{\pi} \in \Pi'$:

$$\begin{aligned}&\left|\Delta_{h,\ell}(\pi, v, \tilde{\pi}) - \mathbb{E}^{\pi\circ_h\tilde{\pi}\circ_{h+1}\widehat{\pi}_{h+1:H}}\left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \ge 0\} \cdot (b_\ell(\boldsymbol{x}_\ell) - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}^{\mathsf{b}}_{h,\ell})\right]\right| \\ &\le \frac{4AH}{\lambda}\sqrt{\frac{\log(2H^2|\Psi_{h-1}|\mathcal{G}_h(\Pi', n)/\delta)}{n}}.\end{aligned}$$

Thus, by the triangle inequality and (112), we have that for all $\ell \in [h+1\,..\,H]$, $(\pi, v) \in \Psi_{h-1}$, and $\tilde{\pi} \in \Pi'$:

$$\begin{aligned}&\left|\mathbb{E}^{\pi\circ_h\tilde{\pi}\circ_{h+1}\widehat{\pi}_{h+1:H}}\left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \ge 0\} \cdot (b_\ell(\boldsymbol{x}_\ell) - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}^{\mathsf{b}}_{h,\ell})\right]\right| \\ &\le \frac{4AH}{\lambda}\sqrt{\frac{\log(2H^2|\Psi_{h-1}|\mathcal{G}_h(\Pi', n)/\delta)}{n}} + 2cd\varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n) + \frac{8cd\nu HA}{\mu\lambda}.\end{aligned}$$

Thus, summing over $\ell \in [h+1\,..\,H]$ and using the triangle inequality, we get that for all $(\pi, v) \in \Psi_{h-1}$ and $\tilde{\pi} \in \Pi'$:

$$\begin{aligned}&\left|\mathbb{E}^{\pi\circ_h\tilde{\pi}\circ_{h+1}\widehat{\pi}_{h+1:H}}\left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \ge 0\} \cdot \left(\sum_{\ell=h+1}^{H} b_\ell(\boldsymbol{x}_\ell) - \sum_{\ell=h+1}^{H} \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \hat{\theta}^{\mathsf{b}}_{h,\ell}\right)\right]\right| \\ &\le \frac{4AH^2}{\lambda}\sqrt{\frac{\log(2H^2|\Psi_{h-1}|\mathcal{G}_h(\Pi', n)/\delta)}{n}} + 2cdH\varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n) + \frac{8cd\nu H^2A}{\mu\lambda}.\end{aligned}$$

Combining this with the conditioning on the event of Lemma J.2 with $\delta' = \delta/(6HT)$ and the triangle inequality, we get that for all $(\pi, v) \in \Psi_{h-1}$ and $\tilde{\pi} \in \Pi'$:

$$\begin{aligned}&\left|\mathbb{E}^{\pi\circ_h\tilde{\pi}\circ_{h+1}\widehat{\pi}_{h+1:H}}\left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \ge 0\} \cdot \left(\sum_{\ell=h}^{H} \boldsymbol{r}_h + \sum_{\ell=h+1}^{H} b_\ell(\boldsymbol{x}_\ell) - \phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \left(\hat{\theta}^{\mathsf{r}}_h + \sum_{\ell=h+1}^{H} \hat{\theta}^{\mathsf{b}}_{h,\ell}\right)\right)\right]\right| \\ &\le \frac{4AH^2}{\lambda}\sqrt{\frac{\log(2H^2|\Psi_{h-1}|\mathcal{G}_h(\Pi', n)/\delta)}{n}} + 2cdH\varepsilon^{\mathsf{b}}_{\mathsf{reg}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n) + \frac{8cd\nu H^2A}{\mu\lambda} \\ &\quad + \varepsilon^{\mathsf{r}}_{\mathsf{reg}}(\tfrac{\delta}{6HT}, \Pi', |\Psi_{h-1}|, n),\end{aligned}$$

where $\varepsilon^{\mathsf{r}}_{\mathsf{reg}}$ is as in Lemma J.2. Now, (88) follows from the definition of $\widetilde{Q}_h$ in the lemma's statement and the fact that $\hat{\theta}_h = \hat{\theta}^{\mathsf{r}}_h + \sum_{\ell=h+1}^{H} \hat{\theta}^{\mathsf{b}}_{h,\ell}$ and $c = 4d_\nu$. □

J.3. Additional Structural Results for the Proofs of `FitValue`

In this subsection, we present additional structural results we require for proving the guarantees of `FitValue`. These results are closely related to the structural results in (Weisz et al., 2024).

Lemma J.6. *Let $\ell \in [H]$, $L > 0$, $\mu, \zeta > 0$, $W_\ell \in \mathbb{S}_{++}^{d\times d}$, and $f : \mathcal{X}_\ell \to [-L, L]$ be given, and let $F : \mathcal{X}_\ell \to [-L, L]^{d\times d}$ be the function defined as*

$$F(x) = \mathbb{I}\{\|\varphi(x; W_\ell)\| \geq \mu\} \cdot \frac{\varphi(x; W_\ell)\varphi(x; W_\ell)^\top}{\|\varphi(x; W_\ell)\|^2} \cdot f(x),$$

where

$$\varphi(x; W_\ell) = W_\ell\phi(x, a_x) - W_\ell\phi(x, a'_x), \quad \text{and} \quad (a_x, a'_x) \in \operatorname*{arg\,max}_{(a,a')\in\mathcal{A}^2} \|W_\ell\phi(x, a) - W_\ell\phi(x, a')\|. \tag{113}$$

Further, let ρ_ℓ denote the approximate optimal design for Θ_ℓ in Section 2.3, and define

$$G_\ell \coloneqq W_\ell^{-1}\left(\sum_{\pi\in\operatorname{supp}\rho_\ell} \rho_\ell(\pi)\theta_\ell^\pi(\theta_\ell^\pi)^\top\right)W_\ell^{-1} \in \mathbb{R}^{d\times d}, \quad \text{and} \quad z_\| \coloneqq \operatorname{Proj}_{\mathcal{S}(G_\ell,\zeta)}(z), \quad \forall z \in \mathbb{R}^d.$$

Then, for all $z, y \in \mathbb{B}(1)$, the function $z_\|^\top F(\cdot)y$ is α-admissible with $\alpha \coloneqq \frac{\sqrt{\zeta}\mu}{L}$; that is, $z_\|^\top F(x)y \leq \mathrm{Rg}^{\mathsf{D}}(x)/\alpha$ for all $x \in \mathcal{X}_\ell$.

Proof of Lemma J.6. Aspects of this proof are inspired by the proof of (Weisz et al., 2024, Lemma 4.9).

Fix $x \in \mathcal{X}_\ell$ and let $\bar{\varphi}(x; W_\ell) \coloneqq \varphi(x; W_\ell)/\|\varphi(x; W_\ell)\|$. Using that $|f(x)| \leq L$, $z, y \in \mathbb{B}(1)$, and $\|\bar{\varphi}(x; W_\ell)\| \leq 1$, we obtain

$$\begin{aligned}
|z_\|^\top F(x)y| &\leq |\bar{\varphi}(x; W_\ell)^\top z_\| \cdot \bar{\varphi}(x; W_\ell)^\top y| \cdot \mathbb{I}\{\|\varphi(x; W_\ell)\| \geq \mu\} \cdot L, \\
&\leq |\varphi(x; W_\ell)^\top z_\|| \cdot \mu^{-1} \cdot L, \\
&\leq \|\varphi(x; W_\ell)_\|\| \cdot \mu^{-1} \cdot L.
\end{aligned} \tag{114}$$

On the other hand, we can write

$$\begin{aligned}
\mathrm{Rg}^{\mathsf{D}}(x)^2 &= \max_{\pi\in\operatorname{supp}(\rho_\ell)} \sup_{a,a'\in\mathcal{A}} \langle\phi(x, a) - \phi(x, a'), \theta_\ell^\pi\rangle^2, \\
&= \max_{\pi\in\operatorname{supp}(\rho_\ell)} \sup_{a,a'\in\mathcal{A}} \langle W_\ell\phi(x, a) - W_\ell\phi(x, a'), W_\ell^{-1}\theta_\ell^\pi\rangle^2, \\
&= \max_{\pi\in\operatorname{supp}(\rho_\ell)} \langle\varphi(x; W_\ell), W_\ell^{-1}\theta_\ell^\pi\rangle^2, \\
&= \max_{\pi\in\operatorname{supp}(\rho_\ell)} \varphi(x; W_\ell)^\top W_\ell^{-1}\theta_\ell^\pi(\theta_\ell^\pi)^\top W_\ell^{-1}\varphi(x; W_\ell), \\
&\geq \varphi(x; W_\ell)^\top G_\ell\varphi(x; W_\ell), \\
&\geq \varphi(x; W_\ell)_\|^\top G_\ell\varphi(x; W_\ell)_\|, \\
&\geq \|\varphi(x; W_\ell)_\|\|^2\zeta,
\end{aligned} \tag{115}$$

where the last inequality follows from the fact that the eigenvalues of G_ℓ corresponding to the subspace in which $\varphi(x; W_\ell)_\|$ lies are by definition at least ζ. Combining (115) with (114), we get that

$$\mathrm{Rg}^{\mathsf{D}}(x) \geq \sqrt{\zeta} \cdot \|\varphi(x; W_\ell)_\|\| \geq \frac{\sqrt{\zeta}\mu}{L} \cdot |z_\|^\top F(x)y|,$$

finishing the proof. □

Lemma J.7. *Let $\ell \in [H]$, $\mu, \nu > 0$, $W_\ell \in \mathbb{S}^{d\times d}_{++}$ be given. Further, let $\varphi(\cdot; W_\ell)$ and $\bar{\varphi}(\cdot; W_\ell)$ be as in Lemma J.6 and define*

$$M_\ell(x) \coloneqq \mathbb{I}\{\varphi(x; W_\ell) \geq \mu\} \cdot \bar{\varphi}(x; W_\ell)\bar{\varphi}(x; W_\ell)^\top, \quad \textit{where} \quad \bar{\varphi}(x; W_\ell) \coloneqq \frac{\varphi(x; W_\ell)}{\|\varphi(x; W_\ell)\|}. \tag{116}$$

Then, for any $z \in \mathbb{B}(1)$, $\tilde{z} \coloneqq \mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(z)$, and $x \in \mathcal{X}_\ell$:

$$|\tilde{z}^\top M_\ell(x)\tilde{z} - z^\top M_\ell(x) z| \leq \frac{4\nu}{\mu}.$$

Proof. Aspects of this proof are inspired by the proof of (Weisz et al., 2024, Lemma E.1).

Fix $z \in \mathbb{B}(1)$ and let $y = \mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)^\perp}(z)$ and note that $z = \tilde{z} + y$. By symmetry of $M_\ell(x)$, we have

$$\begin{aligned}
z^\top M_\ell(x) z &= (\tilde{z} + y)^\top M_\ell(x)(\tilde{z} + y), \\
&= \tilde{z}^\top M_\ell(x)(\tilde{z} + y) + y^\top M_\ell(x)(\tilde{z} + y), \\
&= \tilde{z}^\top M_\ell(x)\tilde{z} + y^\top M_\ell(x)(2\tilde{z} + y), \\
&= \tilde{z}^\top M_\ell(x)\tilde{z} + y^\top M_\ell(x)(\tilde{z} + z),
\end{aligned}$$

where the last step follows by the fact that $z = \tilde{z} + y$. Thus, it suffices to bound $y^\top M_\ell(x)(\tilde{z} + z)$. First, note that $\|z\| \leq 1$ and so $\|\tilde{z} + z\| \leq 2$. Therefore, we have

$$\begin{aligned}
|y^\top M_\ell(x)(\tilde{z} + z)| &= \mathbb{I}\{\varphi(x; W_\ell) \geq \mu\} \cdot |y^\top \bar{\varphi}(x; W_\ell)| \cdot |(\tilde{z} + z)^\top \bar{\varphi}(x; W_\ell)|, \\
&\leq 2\mathbb{I}\{\varphi(x; W_\ell) \geq \mu\} \cdot |y^\top \bar{\varphi}(x; W_\ell)|, \\
&= 2\mathbb{I}\{\varphi(x; W_\ell) \geq \mu\} \cdot \left| y^\top \frac{\varphi(x; W_\ell)}{\|\varphi(x; W_\ell)\|} \right|, \\
&\leq 2\mu^{-1} \cdot |y^\top \varphi(x; W_\ell)|, \\
&\leq 2\mu^{-1} \cdot \|\mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)^\perp}(\varphi(x; W_\ell))\|.
\end{aligned} \tag{117}$$

Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of W_ℓ and and let $w_1, \ldots, w_d \in \mathbb{R}^d$ be an orthonormal basis such that $W_\ell w_i = \lambda_i w_i$, for all $i \in [d]$ (such a basis exists because $W_\ell \in \mathbb{S}^{d\times d}_{++}$). With this, we can write

$$W_\ell = \sum_{i=1}^d \lambda_i w_i w_i^\top.$$

Therefore, since $\varphi(\tilde{x}; W_\ell) = W_\ell \varphi(x', a_{\tilde{x}}, a'_{\tilde{x}})$ with $(a_{\tilde{x}}, a'_{\tilde{x}})$ as in (113). we have

$$\mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)^\perp}(\varphi(x; W_\ell)) = \sum_{i=1}^d \lambda_i \cdot w_i^\top \varphi(x, a_x, a'_x) \cdot \mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)^\perp}(w_i),$$

and by definition of $\mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)^\perp}$:

$$\begin{aligned}
&= \sum_{i\in[d]:\lambda_i\leq\nu} \lambda_i \cdot w_i^\top \varphi(x, a_x, a'_x), \\
&\leq \nu \|\varphi(x, a_x, a_{x'})\|, \\
&\leq 2\nu.
\end{aligned}$$

Plugging this bound into (117) completes the proof. □

Lemma J.8. *For $h \in [H]$ and $\ell \in [h+1\,..\,H]$, let B_ℓ and $\hat{\vartheta}^{\mathsf{b}}_{h,\ell}$ be as in Algorithm 6. Then, for any $z \in \mathbb{B}(1)$ and $\tilde{z} \coloneqq \mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(z)$, we have:*

$$\forall x \in \mathcal{X}_\ell, \quad |\tilde{z}^\top B_\ell(x)\tilde{z} - z^\top B_\ell(x) z| \le \frac{4\nu H}{\mu}, \tag{118}$$

$$\forall (x,a) \in \mathcal{X}_h \times \mathcal{A}, \quad |\hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(x,a),\tilde{z},\tilde{z}] - \hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(x,a),z,z]| \le \frac{4\nu H}{\mu\lambda}. \tag{119}$$

Proof. Fix $x' \in \mathcal{X}_\ell$ and $z \in \mathbb{B}(1)$. Note that B_ℓ is of the form

$$B_\ell(x') \coloneqq \bar{b}_\ell(x') \cdot \mathbb{I}\{\|\varphi(x';W_\ell)\| \ge \mu\} \cdot \frac{\varphi(x';W_\ell)\varphi(x';W_\ell)^\top}{\|\varphi(x';W_\ell)\|^2}.$$

where $\bar{b}_\ell(x') \coloneqq \min\big(H, \frac{\varepsilon}{4H} \max_{a' \in \mathcal{A}} \|\phi(x',a')\|_{(\lambda I + U_\ell)^{-1}}\big)$. Using the notation of Lemma J.7; specifically using the matrix $M_\ell(x')$ in (116), we can write B_ℓ as

$$B_\ell(x') = \bar{b}_\ell(x') \cdot M_\ell(x').$$

Thus, applying Lemma J.7 with the fact that $\bar{b}_\ell(x') \le H$ implies that

$$|\tilde{z}^\top B_\ell(x')\tilde{z} - z^\top B_\ell(x') z| \le \frac{4\nu H}{\mu}, \tag{120}$$

where $\tilde{z} \coloneqq \mathrm{Proj}_{\mathcal{S}(W_\ell,\nu)}(z)$. This shows (118). We now show (119). Fix $(x,a) \in \mathcal{X}_h \times \mathcal{A}$. Using the definition of $\hat{\vartheta}^{\mathsf{b}}_{h,\ell}$ in Line 16 of Algorithm 5, we can write

$$\hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(x,a),\cdot,\cdot] = \phi(x,a)^\top \Sigma_h^{-1} \sum_{(\pi,v)\in\Psi_{h-1}} \sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \phi(x_h,a_h) \cdot B_\ell(x_\ell) \in \mathbb{R}^{d\times d},$$

where $\lambda n_{\mathtt{traj}}|\Psi_{h-1}|I + \sum_{(\pi,v)\in\Psi_{h-1}} \sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \phi(x_h,a_h)\phi(x_h,a_h)^\top$. Thus, using that

$$\left\| \Sigma_h^{-1} \sum_{(\pi,v)\in\Psi_{h-1}} \sum_{(x_{1:H},a_{1:H},r_{1:H})\in\widehat{\mathcal{D}}_{h,\pi}} \phi(x_h,a_h) \right\| \le \frac{1}{\lambda},$$

together with (120) and the fact that $\|\phi(x,a)\| \le 1$, we get that

$$|\hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(x,a),\tilde{z},\tilde{z}] - \hat{\vartheta}^{\mathsf{b}}_{h,\ell}[\phi(x,a),z,z]| \le \frac{4\nu H}{\mu\lambda}.$$

This completes the proof. □

Appendix K. Guarantee of `Evaluate`

Lemma K.1 (Guarantee of `Evaluate`). *Let $\widehat{\pi}_{h+1:H}$ and n be given and consider a call to $\mathtt{Evaluate}_h(\widehat{\pi}_{1:H}, n)$ (Algorithm 3). Then, for any $\delta' \in (0,1)$, with probability at least $1-\delta'$, the output J of Algorithm 3 is such that*

$$\left| J - \mathbb{E}^{\widehat{\pi}}\left[\sum_{h\in[H]} \boldsymbol{r}_h\right] \right| \le H\sqrt{\frac{2\log(1/\delta')}{n}}.$$

Proof. The result follows from Hoeffding's inequality and the fact that the rewards take values in $[0,1]$. □

Appendix L. Analysis: Proof of Theorem 3.1

Central to the proof of Theorem 3.1 is a change of measure argument (as described in Appendix C.1), which we now state.

Lemma L.1 (Change of measure). *Let $\varepsilon, \delta \in (0,1)$ be given and consider a call to* `Optimistic-PSDP`$(\Pi_{\mathsf{Bench}}, \varepsilon, \delta)$ *with Π_{Bench} as in Section 2.3. Further, let $(\varepsilon_{\mathrm{hoff}}, \mathcal{E}^{\mathrm{hoff}})$ be as in Lemma I.3. Then, under the event $\mathcal{E}^{\mathrm{hoff}}$, we have for all $\theta \in \mathbb{R}^d$, $t \in [T]$, $h \in [H]$ and $(x,a) \in \mathcal{X}_h \times \mathcal{A}$:*

$$|\theta^\top \phi(x,a)|$$
$$\leq \sqrt{d}\|\phi(x_h,a)\|_{(\beta I + U_h^{(t)})^{-1}} \cdot \left(\sqrt{\frac{\beta}{d}}\|\theta\| + t\varepsilon_{\mathrm{hoff}} + \sum_{(\pi,v)\in\Psi_h^{(t)}} \left| \mathbb{E}^\pi[\theta^\top \phi(\boldsymbol{x}_h, \boldsymbol{a}_h) \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top v \geq 0\}] \right| \right).$$

Proof. In this proof, we condition on the event $\mathcal{E}^{\mathrm{hoff}}$. First, we note that $U_h^{(t)} = \sum_{k=1}^{t-1} u_h^{(k)} (u_h^{(k)})^\top$, where $(u_h^{(k)})$ are as in Algorithm 1. Thus, by letting $A_h = \beta I + U_h^{(t)}$, we have that for all $\theta \in \mathbb{R}^d$, $t \in [T]$, $h \in [H]$, and $(x,a) \in \mathcal{X}_h \times \mathcal{A}$:

$$\begin{aligned}
|\theta^\top \phi(x,a)| &= |\theta^\top A_h A_h^{-1} \phi(x,a)|, \\
&\leq \left| \left(\beta\theta^\top + \sum_{k=1}^{t-1} \theta^\top u_h^{(k)} (u_h^{(k)})^\top \right) A_h^{-1/2} A_h^{-1/2} \phi(x,a) \right|, \\
&\leq \beta \|\theta\| \cdot \|A_h^{-1/2}\|_{\mathrm{op}} \cdot \|\phi(x,a)\|_{A_h^{-1}} + \sum_{k=1}^{t-1} |\theta^\top u_h^{(k)}| \cdot \|u_h^{(k)}\|_{A_h^{-1}} \cdot \|\phi(x,a)\|_{A_h^{-1}}, \\
&\leq \sqrt{\beta} \cdot \|\theta\| \cdot \|\phi(x,a)\|_{A_h^{-1}} + \sqrt{d} \sum_{k=1}^{t-1} |\theta^\top u_h^{(k)}| \cdot \|\phi(x,a)\|_{A_h^{-1}}, \qquad (121)
\end{aligned}$$

where the last inequality follows by the facts that $\|A_h^{-1/2}\|_{\mathrm{op}} \leq \beta^{-1/2}$ and $\|u_h^{(k)}\|_{A_h^{-1}}^2 \leq d$. Now, by Lemma I.3, the conditioning on $\mathcal{E}^{\mathrm{hoff}}$, and the triangle inequality, we have that for all $k \in [t-1]$:

$$|\theta^\top u_h^{(k)}| \leq \left| \mathbb{E}^{\tilde{\pi}_{1:h}^{(k)}} [\theta^\top \phi(\boldsymbol{x}_h, \boldsymbol{a}_h) \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top v_h^{(k)} \geq 0\}] \right| + \varepsilon_{\mathrm{hoff}}.$$

Plugging this into (121) and using that $\Psi_h^{(t)} = \bigcup_{k=1}^{t-1} \{(\tilde{\pi}_{1:h}^{(k)}, v_h^{(k)})\}$ (see the update rule in Line 10 of Algorithm 1), we get the desired result. □

The following key lemma consolidates the guarantees of `UncertainPolicy` and `FitValue`, presented in Appendix I and Appendix J, respectively, into a single statement. It also employs an elliptical potential argument to bound the expected bonuses under rollouts (see Appendix C.1 for a high-level explanation of this argument)

Lemma L.2. *Let $\varepsilon, \delta \in (0,1)$ be given and consider a call to* `Optimistic-PSDP`$(\Pi_{\mathsf{Bench}}, \varepsilon, \delta)$ *(Algorithm 1) with Π_{Bench} as in Section 2.3. Let $(\varepsilon_{\mathrm{hoff}}, \mathcal{E}^{\mathrm{hoff}})$ and $(\varepsilon_{\mathtt{reg}}, \mathcal{E}^{\mathtt{reg}})$ be as in Lemma I.3 and Lemma J.5, respectively. Then, under the event $\mathcal{E}^{\mathrm{hoff}} \cap \mathcal{E}^{\mathtt{reg}}$, there exists $\tau \in [T]$ such that for all $h \in [H]$, $\ell \in [0\,..\,h-1]$, and $(\pi, v) \in \Psi_\ell^{(\tau)}$:*

$$\mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}_{\ell+1:H}^{(\tau)}} \left[\max_{a \in \mathcal{A}} \|\phi(\boldsymbol{x}_h, a)\|_{(\beta I + U_h^{(\tau)})^{-1}} \right] \leq 1/2 + 2d\beta^{-1}\varepsilon_{\mathrm{hoff}},$$

and for all $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$ *and* $(\pi, v) \in \Psi_{h-1}^{(\tau)}$*:*

$$\mathbb{E}^{\pi \circ_h \tilde{\pi} \circ_{h+1} \widehat{\pi}_{h+1:H}^{(\tau)}} \left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\} \cdot (Q_h^{(\tau)}(\boldsymbol{x}_h, \boldsymbol{a}_h) - \widehat{Q}_h^{(\tau)}(\boldsymbol{x}_h, \boldsymbol{a}_h)) \right] \leq \varepsilon_{\mathsf{reg}},$$

where for all $(x, a) \in \mathcal{X}_h \times \mathcal{A}$, $Q_h^{(\tau)}(x, a) = Q_h^{\widehat{\pi}^{(\tau)}}(x, a) + \mathbb{E}^{\widehat{\pi}^{(\tau)}}[\sum_{\ell=h}^H b_\ell^{(\tau)}(\boldsymbol{x}_h) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a]$, $\widehat{Q}_h^{(\tau)}(x, a) \coloneqq \phi(x, a)^\top \hat{\theta}_h^{(\tau)} + b_h^{(\tau)}(x)$*, and* $b_h^{(\tau)}(x) \coloneqq \min\left(H, \frac{\varepsilon}{4H} \cdot \max_{a' \in \mathcal{A}} \|\phi(x, a')\|_{(\beta I + U_h^{(\tau)})^{-1}}\right) \cdot \mathbb{I}\{\|\varphi(x; W_h^{(\tau)})\| \geq \mu\}$.

Proof. Throughout the proof, we condition on the event $\mathcal{E}^{\mathrm{hoff}} \cap \mathcal{E}^{\mathsf{reg}}$. For $h \in [H]$, define

$$\mathcal{T}_h \coloneqq \{(t, \ell) \in [T] \times [h-1] \mid w_h^{(t,\ell)} \neq 0\},$$

where $(w_h^{(t,\ell)})$ are as in Algorithm 1. By Line 5 of Algorithm 1, we have that

$$W_h^{(T+1)} = \left(H^{-2} I + \sum_{(t,\ell) \in \mathcal{T}_h} w_h^{(t,\ell)} (w_h^{(t,\ell)})^\top\right)^{-1/2}.$$

By Lemma J.5 (and the conditioning on $\mathcal{E}^{\mathsf{reg}}$), $W_h^{(T+1)}$ is a valid ν-preconditioning, and so by Lemma 2.6, we must have that

$$|\mathcal{T}_h| \leq 4d \log(1 + 16\nu^{-4} H^4). \tag{122}$$

Now, let

$$\mathcal{T} \coloneqq \{t \in [T] \mid (t, \ell) \notin \mathcal{T}_h, \forall h \in [H], \forall \ell \in [h-1]\}.$$

By (122), we have that

$$\begin{aligned} |\mathcal{T}| &\geq T - 4Hd \log(1 + 16\nu^{-4} H^4), \\ &\geq T/2, \end{aligned} \tag{123}$$

where the last inequality follows by the choice of T in Algorithm 1. On the other hand, we have that

$$\sum_{t \in \mathcal{T}} \sum_{h \in [H]} \|u_h^{(t)}\|_{(\beta I + U_h^{(t)})^{-1}} \leq \sum_{h \in [H]} \sum_{t \in [T]} \|u_h^{(t)}\|_{(\beta I + U_h^{(t)})^{-1}} \leq H\sqrt{Td \log(1 + T/\beta)},$$

where the last inequality follows by Lemma O.6. Therefore, there is a $\tau \in \mathcal{T}$ such that for all $h \in [H]$:

$$\begin{aligned} \|u_h^{(\tau)}\|_{(\beta I + U_h^{(\tau)})^{-1}} \leq \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \sum_{h' \in [H]} \|u_{h'}^{(t)}\|_{(\beta I + U_{h'}^{(t)})^{-1}} &\leq \frac{H\sqrt{Td \log(1 + T/\beta)}}{|\mathcal{T}|}, \\ &\leq 2H\sqrt{\frac{d \log(1 + T/\beta)}{T}}, \quad \text{(by (123))} \\ &\leq 1/2, \end{aligned} \tag{124}$$

where the last inequality follows from the fact that $T \geq 16dH^2\log(1+T/\beta)$. Combining (124) with Lemma I.3 (and the conditioning on $\mathcal{E}^{\mathrm{hoff}}$), we have that for all $h \in [H]$, $\ell \in [0\,..\,h-1]$, and $(\pi, v) \in \Psi_\ell^{(\tau)}$:

$$\mathbb{E}^{\pi \circ_{\ell+1} \widehat{\pi}^{(\tau)}_{\ell+1:H}}\left[\|\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)\|_{(\beta I + U_h^{(\tau)})^{-1}}\right] \leq 2\sqrt{d}\|u_h^{(\tau)}\|_{(\beta I + U_h^{(\tau)})^{-1}} + 2d\beta^{-1}\varepsilon_{\mathrm{hoff}} \leq 1/2 + 2d\beta^{-1}\varepsilon_{\mathrm{hoff}}.$$

Now, by definition of $\mathcal{T}$, we have that $w_h^{(t,\ell)} = 0$ for all $h \in [H]$, and $\ell \in [h-1]$ at the end of each iteration $t \in \mathcal{T}$. This implies that $W_h^{(t+1)} = W_h^{(t)}$ at the end of any iteration $t \in \mathcal{T}$, and so by Lemma J.5 (and the conditioning on $\mathcal{E}^{\mathsf{reg}}$), we have that for all $\tilde{\pi} \in \Pi_{\mathsf{Bench}}$ and $(\pi, v) \in \Psi_{h-1}^{(\tau)}$:

$$\mathbb{E}^{\pi \circ_h \tilde{\pi} \circ_{h+1} \widehat{\pi}^{(t)}_{h+1:H}}\left[\mathbb{I}\{\phi(\boldsymbol{x}_{h-1}, \boldsymbol{a}_{h-1})^\top v \geq 0\} \cdot (Q_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h) - \widehat{Q}_h^{(t)}(\boldsymbol{x}_h, \boldsymbol{a}_h))\right] \leq \varepsilon_{\mathsf{reg}}.$$

Instantiating this with $t = \tau \in \mathcal{T}$ completes the proof. □

We still need a guarantee for the `Evaluate` subroutine within `Optimistic-PSDP`.

Lemma L.3 (Guarantee of `Evaluate` for `Optimistic-PSDP`)**.** *Let $\varepsilon, \delta \in (0,1)$ be given and consider a call* `Optimistic-PSDP`$(\Pi_{\mathsf{Bench}}, \varepsilon, \delta)$ *(Algorithm 1) with Π_{Bench} as in Section 2.3. Then, for any $\delta \in (0,1)$, there is an event $\mathcal{E}^{\mathrm{eval}}$ of probability at least $1-\delta$, under which for all $t \in [T]$, the variable τ' in Algorithm 1 satisfies:*

$$\max_{t \in [T]} J(\widehat{\pi}^{(t)}_{1:H}) - J(\widehat{\pi}^{(\tau')}_{1:H}) \leq \varepsilon_{\mathrm{eval}} \coloneqq H\sqrt{\frac{2\log(T/\delta)}{n_{\mathsf{traj}}}}.$$

Proof of Lemma L.3. The result follows from Lemma K.1 with $\delta' = \delta/T$ and Lemma O.4 (essentially the union bound over $t \in [T]$). □

We now have all the ingredients to prove Theorem 3.1.

Proof of Theorem 3.1. Let $(\varepsilon_{\mathrm{hoff}}, \mathcal{E}^{\mathrm{hoff}})$, $(\varepsilon_{\mathsf{reg}}, \mathcal{E}^{\mathsf{reg}})$, and $(\varepsilon_{\mathrm{eval}}, \mathcal{E}^{\mathrm{eval}})$ be as in Lemma I.3, Lemma J.5, and Lemma L.3, respectively. By the union bound, we have that $\mathbb{P}[\mathcal{E}^{\mathrm{hoff}} \cap \mathcal{E}^{\mathsf{reg}} \cap \mathcal{E}^{\mathrm{eval}}] \geq 1 - 2\delta$. Throughout this proof, we condition on $\mathcal{E}^{\mathrm{hoff}} \cap \mathcal{E}^{\mathsf{reg}} \cap \mathcal{E}^{\mathrm{eval}}$.

For μ, ν as in Algorithm 1, let $\gamma \coloneqq \mu/\sqrt{d_\nu}$, where $d_\nu \coloneqq 5d\log(1 + 16H^4\nu^{-4})$. Further, let $\tau \in [T]$ be as in Lemma L.2 and define

$$\widehat{\pi}^\star_h(\cdot) \coloneqq \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\cdot) < \gamma\} \cdot \widehat{\pi}^{(\tau)}_h(\cdot) + \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\cdot) \geq \gamma\} \cdot \pi^\star(\cdot),$$

for $h \in [H]$ and Rg^{D} as in Definition 2.3; note that by (8) and the fact that $\widehat{\pi}^{(\tau)}_h, \pi^\star \in \Pi_{\mathsf{Base}}$, where $\Pi_{\mathsf{Base}} = \{x \mapsto \arg\max_{a \in \mathcal{A}} \theta^\top \phi(x,a) \mid \theta \in \mathbb{B}(H)\}$, we have

$$\widehat{\pi}^\star_h \in \Pi_{\mathsf{Bench}}. \tag{125}$$

At a high-level, our strategy will be to show the sequence of inequalities:

$$\begin{aligned} J(\pi^\star) &\leq J(\widehat{\pi}^\star_{1:H}) + O(\varepsilon), \\ J(\widehat{\pi}^\star_{1:H}) &\leq J(\widehat{\pi}^{(\tau)}_{1:H}) + O(\varepsilon), \\ J(\widehat{\pi}^{(\tau)}_{1:H}) &\leq J(\widehat{\pi}_{1:H}) + O(\varepsilon). \end{aligned}$$

Summing up these inequalities and telescoping would imply the desired result.

Suboptimality of $\widehat{\pi}^\star_{1:H}$. First, by the performance difference lemma (see Lemma O.7), we have

$$\begin{aligned}
&J(\pi^\star) - J(\widehat{\pi}^\star_{1:H}) \\
&= \sum_{h=1}^H \mathbb{E}^{\widehat{\pi}^\star}\Big[Q_h^{\pi^\star}(\boldsymbol{x}_h, \pi^\star(\boldsymbol{x}_h)) - Q_h^{\pi^\star}(\boldsymbol{x}_h, \widehat{\pi}^\star_h(\boldsymbol{x}_h))\Big], \\
&= \sum_{h=1}^H \mathbb{E}^{\widehat{\pi}^\star}\Big[\langle \phi(\boldsymbol{x}_h, \pi^\star(\boldsymbol{x}_h)) - \phi(\boldsymbol{x}_h, \widehat{\pi}^\star_h(\boldsymbol{x}_h)), \theta_h^{\pi^\star}\rangle\Big], \\
&= \sum_{h=1}^H \mathbb{E}^{\widehat{\pi}^\star}\Big[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_h) < \gamma\} \cdot \langle \phi(\boldsymbol{x}_h, \pi^\star(\boldsymbol{x}_h)) - \phi(\boldsymbol{x}_h, \widehat{\pi}^{(\tau)}_h(\boldsymbol{x}_h)), \theta_h^{\pi^\star}\rangle\Big], \quad \text{(by definition of } \widehat{\pi}^\star) \\
&\le \sum_{h=1}^H \mathbb{E}^{\widehat{\pi}^\star}\Big[\mathbb{I}\{\mathrm{Rg}(\boldsymbol{x}_h) < \sqrt{2d}\gamma\} \cdot \langle \phi(\boldsymbol{x}_h, \pi^\star(\boldsymbol{x}_h)) - \phi(\boldsymbol{x}_h, \widehat{\pi}^{(\tau)}_h(\boldsymbol{x}_h)), \theta_h^{\pi^\star}\rangle\Big], \quad \text{(by Lemma 2.1)} \\
&\le \sum_{h=1}^H \mathbb{E}^{\widehat{\pi}^\star}\Big[\mathbb{I}\{\mathrm{Rg}(\boldsymbol{x}_h) < \sqrt{2d}\gamma\} \cdot \mathrm{Rg}(\boldsymbol{x}_h)\Big], \\
&\le H\sqrt{2d}\gamma = H\sqrt{2d/d_\nu}\mu. \qquad (126)
\end{aligned}$$

We now bound the suboptimality of $\widehat{\pi}^{(\tau)}_{1:H}$ relative to $\widehat{\pi}^\star_{1:H}$, where we recall that $\tau \in [T]$ is as in Lemma L.2.

Suboptimality of $\widehat{\pi}^{(\tau)}_{1:H}$. For this part of the proof, we need some additional notation for the bonuses and optimistic value functions; for $\ell \in [H]$, $(x,a) \in \mathcal{X}_\ell \times \mathcal{A}$, and $t \in [T]$, we define

$$\bar{b}_\ell^{(t)}(x) \coloneqq \min\left(H, \frac{\varepsilon}{4H} \cdot \max_{a' \in \mathcal{A}} \|\phi(x,a')\|_{(\beta I + U_\ell^{(t)})^{-1}}\right), \qquad b_\ell^{(t)}(x) \coloneqq \bar{b}_\ell^{(t)}(x) \cdot \mathbb{I}\{\|\varphi(x; W_\ell^{(t)})\| \le 1/2\}, \qquad (127)$$

$$Q_\ell^{(t)}(x,a) = Q_\ell^{\widehat{\pi}^{(t)}}(x,a) + \mathbb{E}^{\widehat{\pi}^{(t)}}\left[\sum_{k=\ell}^H b_k^{(t)}(\boldsymbol{x}_k) \mid \boldsymbol{x}_\ell = x, \boldsymbol{a}_\ell = a\right], \quad \text{and} \quad V_\ell^{(t)}(x) \coloneqq Q_\ell^{(t)}(x, \widehat{\pi}_\ell^{(t)}(x)). \qquad (128)$$

With this, we proceed via backward induction to show that for all $\ell = H+1, \dots, 1$ and $(x,a) \in \mathcal{X}_\ell \times \mathcal{A}$:

$$Q_\ell^{\widehat{\pi}^\star}(x,a) \le Q_\ell^{(\tau)}(x,a) + \bar{b}_\ell^{(\tau)}(x) \cdot \mathbb{I}\{\|\varphi(x; W_\ell^{(\tau)})\| < \mu\}, \qquad (129)$$

$$V_\ell^{\widehat{\pi}^\star}(x) \le V_\ell^{(\tau)}(x) + \xi_\ell^{(\tau)}(x, \widehat{\pi}^\star_\ell(x)) - \xi_\ell^{(\tau)}(x, \widehat{\pi}^{(\tau)}_\ell(x)) + \bar{b}_\ell^{(\tau)}(x) \cdot \mathbb{I}\{\|\varphi(x; W_\ell^{(\tau)})\| < \mu\}, \qquad (130)$$

where for $(\tilde{x}, \tilde{a}) \in \mathcal{X}_\ell \times \mathcal{A}$,

$$\xi_\ell^{(\tau)}(\tilde{x}, \tilde{a}) \coloneqq Q_\ell^{(\tau)}(\tilde{x}, \tilde{a}) - \widehat{Q}_\ell^{(\tau)}(\tilde{x}, \tilde{a}) \quad \text{and} \quad \widehat{Q}_\ell^{(t)}(\tilde{x}, \tilde{a}) = \phi(\tilde{x}, \tilde{a})^\top \hat{\theta}_\ell^{(t)} + b_\ell^{(t)}(\tilde{x}), \qquad (131)$$

with the convention that $Q^{\widehat{\pi}^\star}_{H+1} \equiv Q^{(\tau)}_{H+1} \equiv \widehat{Q}^{(\tau)}_{H+1} \equiv 0$.

The base case $\ell = H+1$ follows trivially by the convention that $Q^{\widehat{\pi}^\star}_{H+1} \equiv Q^{(\tau)}_{H+1} \equiv \widehat{Q}^{(\tau)}_{H+1} \equiv 0$. Now, let $h \in [H]$ and suppose that the induction hypothesis holds for all $\ell = [h+1 \,..\, H+1]$. We show that it holds for $\ell = h$.

We show (129) for $\ell = h$. Fix $(x,a) \in \mathcal{X}_h \times \mathcal{A}$. By Lemma O.8 instantiated with

- $(\pi'_k, \widehat{\pi}_k, \widehat{\pi}'_k) = (\pi^\star, \widehat{\pi}^{(\tau)}_k, \widehat{\pi}^\star_k)$;
- $\mathcal{K}_k = \{\tilde{x} \in \mathcal{X}_k \mid \mathrm{Rg}^{\mathsf{D}}(\tilde{x}) < \gamma\}$;

- $\tilde{r}_k \equiv R$ (R is the reward function); and
- $V_k = V_k^{\widehat{\pi}^\star}$,

for all $k \in [h+1 \,..\, H]$, we get that

$$\begin{aligned}
&\mathbb{E}[V_{h+1}^{\widehat{\pi}^\star}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] \\
&= \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \cdot \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot V_\ell^{\widehat{\pi}^\star}(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] \\
&\quad + \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\prod_{k=h+1}^{\ell} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot r_\ell(\boldsymbol{x}_\ell, \boldsymbol{a}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right].
\end{aligned} \tag{132}$$

We instantiate Lemma O.8 again, this time with

- $(\pi'_k, \widehat{\pi}_k, \widehat{\pi}'_k) = (\widehat{\pi}_k^{(\tau)}, \widehat{\pi}_k^{(\tau)}, \widehat{\pi}_k^{(\tau)})$;
- $\mathcal{K}_k = \{\tilde{x} \in \mathcal{X}_k \mid \mathrm{Rg}^{\mathsf{D}}(\tilde{x}) < \gamma\}$;
- $\tilde{r}_k(\tilde{x}, \tilde{a}) \coloneqq R(\tilde{x}, \tilde{a}) + b_k^{(\tau)}(\tilde{x})$ (with $b_k^{(\tau)}$ as in (127)); and
- $V_k = V_k^{(\tau)}$,

for all $k \in [h+1 \,..\, H]$, to get that

$$\begin{aligned}
&\mathbb{E}[V_{h+1}^{(\tau)}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] \\
&= \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \cdot \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot V_\ell^{(\tau)}(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] \\
&\quad + \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\prod_{k=h+1}^{\ell} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot \tilde{r}_\ell(\boldsymbol{x}_\ell, \boldsymbol{a}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right].
\end{aligned} \tag{133}$$

Thus, combining (132) and (133) and using that $\tilde{r}_\ell(\cdot) \geq r_\ell(\cdot)$, we get

$$\begin{aligned}
&\mathbb{E}[V_{h+1}^{\widehat{\pi}^\star}(\boldsymbol{x}_{h+1}) - V_{h+1}^{(\tau)}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] \\
&\leq \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \cdot \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot \left(V_\ell^{\widehat{\pi}^\star}(\boldsymbol{x}_\ell) - V_\ell^{(\tau)}(\boldsymbol{x}_\ell)\right) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right],
\end{aligned}$$

and so by the induction hypothesis; in particular (130) for $\ell \in [h+1 \,..\, H+1]$, we have

$$\begin{aligned}
&\mathbb{E}[V_{h+1}^{\widehat{\pi}^\star}(\boldsymbol{x}_{h+1}) - V_{h+1}^{(\tau)}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] \\
&\leq \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \left(\xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^\star(\boldsymbol{x}_\ell)) - \xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^{(\tau)}(\boldsymbol{x}_\ell))\right) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] \\
&\quad + \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot \mathbb{I}\{\|\varphi(\boldsymbol{x}_\ell; W_\ell^{(\tau)})\| < \mu\} \cdot \bar{b}_\ell^{(\tau)}(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right].
\end{aligned} \tag{134}$$

Now, by Lemma 2.5, we have that $\mathrm{Rg}^{\mathsf{D}}(\tilde{x}) \leq \sqrt{d_\nu} \cdot \|\varphi(\tilde{x}; W_\ell^{(\tau)})\|$ for all $\tilde{x} \in \mathcal{X}_\ell$. Thus, since $\mu = \sqrt{d_\nu} \cdot \gamma$, we have that for all $\tilde{x} \in \mathcal{X}_\ell$, $\mathbb{I}\{\|\varphi(\tilde{x}; W_\ell^{(\tau)})\| < \mu\} = 1$ only if $\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\tilde{x}) < \gamma\} = 1$. This implies that the second sum in (134) is zero, and so

$$\mathbb{E}[V_{h+1}^{\widehat{\pi}^\star}(\boldsymbol{x}_{h+1}) - V_{h+1}^{(\tau)}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] \leq \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[g_\ell^{(\tau)}(\boldsymbol{x}_{h+1:\ell}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right], \tag{135}$$

where

$$g_\ell^{(\tau)}(\boldsymbol{x}_{h+1:\ell}) \coloneqq \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\} \cdot \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \cdot \left(\xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^\star(\boldsymbol{x}_\ell)) - \xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^{(\tau)}(\boldsymbol{x}_\ell))\right). \tag{136}$$

Now, by definition of the bonuses and $Q_\ell^{(\tau)}$ in (127) and (128), we have $Q_\ell^{(\tau)} \in [0, H + H^2]$. On the other hand, since $\hat{\theta}_\ell^{(\tau)}$ in the definition of $\widehat{Q}_\ell^{(\tau)}$ in (131) satisfies $\hat{\theta}_\ell^{(\tau)} \in \mathbb{B}(H + H^2/\lambda)$; see Line 8-Line 11 of Algorithm 5, we have $\widehat{Q}_\ell^{(\tau)}(\tilde{x}, \tilde{a}) \in [-2H^2/\lambda, 2H^2/\lambda]$, for all $(\tilde{x}, \tilde{a}) \in \mathcal{X}_\ell \times \mathcal{A}$. Therefore,

$$\xi_\ell^{(\tau)}(\tilde{x}, \tilde{a}) = Q_\ell^{(\tau)}(\tilde{x}, \tilde{a}) - \widehat{Q}_\ell^{(\tau)}(\tilde{x}, \tilde{a}) \in [-3H^2/\lambda, 3H^2/\lambda], \tag{137}$$

for all $(\tilde{x}, \tilde{a}) \in \mathcal{X}_\ell \times \mathcal{A}$ and so by Lemma G.1 instantiated with $\pi = \widehat{\pi}^{(\tau)}$, $f(\cdot) = \xi_\ell^{(\tau)}(\cdot, \widehat{\pi}_\ell^\star(\cdot)) - \xi_\ell^{(\tau)}(\cdot, \widehat{\pi}_\ell^{(\tau)}(\cdot))$, and $L = 6H^2/\lambda$, we get that there exists

$$\theta_{h,\ell}^{(\tau)} \in \mathbb{B}(24\tilde{d}H^4/(\gamma\lambda)), \tag{138}$$

where $\tilde{d} = 4d \log\log d + 16$, such that

$$\mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[g_\ell^{(\tau)}(\boldsymbol{x}_{h+1:\ell}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a\right] = \phi(x, a)^\top \theta_{h,\ell}^{(\tau)}. \tag{139}$$

Thus, by Lemma L.1 and $\varepsilon_{\mathrm{hoff}}$ as in Lemma I.3, we have that

$$\begin{aligned}
&\left|\mathbb{E}^{\widehat{\pi}^{(\tau)}}\left[g_\ell^{(\tau)}(\boldsymbol{x}_{h+1:\ell}) \mid \boldsymbol{x}_h = x, a_h = a\right]\right| \\
&\leq \sqrt{d}\|\phi(x, a)\|_{(\beta I + U_h^{(\tau)})^{-1}} \\
&\quad \times \left(\sqrt{\frac{\beta}{d}}\|\theta_{h,\ell}^{(\tau)}\| + T\varepsilon_{\mathrm{hoff}} + \sum_{(\pi, v) \in \Psi_h^{(\tau)}} \left|\mathbb{E}^{\pi}[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \theta_{h,\ell}^{(\tau)} \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \geq 0\}]\right|\right).
\end{aligned} \tag{140}$$

Now, using (139) again and the law of total expectation, we get that for any $(\pi, v) \in \Psi_h^{(\tau)}$:

$$\left|\mathbb{E}^{\pi}[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \theta_{h,\ell}^{(\tau)} \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \geq 0\}]\right| = \left|\mathbb{E}^{\pi \circ_{h+1} \widehat{\pi}^{(\tau)}}[g_\ell^{(\tau)}(\boldsymbol{x}_{h+1:\ell}) \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \geq 0\}]\right|,$$

and so by letting $\mathbf{I}_{h,\ell,v} \coloneqq \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \geq 0\} \cdot \prod_{k=h+1}^{\ell-1} \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k) < \gamma\}$ and using the definition of $g_\ell^{(\tau)}$ in (136):

$$\begin{aligned}
&\left|\mathbb{E}^{\pi}[\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top \theta_{h,\ell}^{(\tau)} \cdot \mathbb{I}\{\phi(\boldsymbol{x}_h, \boldsymbol{a}_h)^\top v \geq 0\}]\right| \\
&= \left|\mathbb{E}^{\pi \circ_{h+1} \widehat{\pi}^{(\tau)}}\left[\mathbf{I}_{h,\ell,v} \cdot \mathbb{E}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \geq \gamma\} \cdot \left(\xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^\star(\boldsymbol{x}_\ell)) - \xi_\ell^{(\tau)}(\boldsymbol{x}_\ell, \widehat{\pi}_\ell^{(\tau)}(\boldsymbol{x}_\ell))\right) \mid \boldsymbol{x}_{\ell-1}, \boldsymbol{a}_{\ell-1}\right]\right]\right|.
\end{aligned} \tag{141}$$

Now, by (137), Lemma 2.3, and Lemma 2.2, there exists

$$\tilde{\theta}_{\ell-1}^{(\tau)} \in \mathbb{B}(24\tilde{d}H^3/(\gamma\lambda)) \tag{142}$$

such that for all $(\tilde{x},\tilde{a}) \in \mathcal{X}_{\ell-1}\times\mathcal{A}$:

$$\begin{aligned}
&\phi(\tilde{x},\tilde{a})^\top \tilde{\theta}_{\ell-1}^{(\tau)} \\
&= \mathbb{E}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell) \ge \gamma\}\cdot\left(\xi_\ell^{(\tau)}(\boldsymbol{x}_\ell,\widehat{\pi}_\ell^\star(\boldsymbol{x}_\ell)) - \xi_\ell^{(\tau)}(\boldsymbol{x}_\ell,\widehat{\pi}_\ell^{(\tau)}(\boldsymbol{x}_\ell))\right) \mid \boldsymbol{x}_{\ell-1}=\tilde{x},\boldsymbol{a}_{\ell-1}=\tilde{a}\right], \quad (143)\\
&= \mathbb{E}\left[\xi_\ell^{(\tau)}(\boldsymbol{x}_\ell,\widehat{\pi}_\ell^\star(\boldsymbol{x}_\ell)) - \xi_\ell^{(\tau)}(\boldsymbol{x}_\ell,\widehat{\pi}_\ell^{(\tau)}(\boldsymbol{x}_\ell)) \mid \boldsymbol{x}_{\ell-1}=\tilde{x},\boldsymbol{a}_{\ell-1}=\tilde{a}\right], \quad (144)
\end{aligned}$$

where the last equality follows by the fact that $\widehat{\pi}_\ell^\star(\cdot) = \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\cdot)<\gamma\}\cdot\widehat{\pi}_\ell^{(\tau)}(\cdot) + \mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\cdot)\ge\gamma\}\cdot\pi^\star(\cdot)$, by definition. Using (143) and Lemma L.1, we get that for all $(\tilde{x},\tilde{a}) \in \mathcal{X}_{\ell-1}\times\mathcal{A}$:

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_\ell)\ge\gamma\}\cdot\left(\xi_\ell^{(\tau)}(\boldsymbol{x}_\ell,\widehat{\pi}_\ell^\star(\boldsymbol{x}_\ell)) - \xi_\ell^{(\tau)}(\boldsymbol{x}_\ell,\widehat{\pi}_\ell^{(\tau)}(\boldsymbol{x}_\ell))\right) \mid \boldsymbol{x}_{\ell-1}=\tilde{x},\boldsymbol{a}_{\ell-1}=\tilde{a}\right]\right| \\
&\le \sqrt{d}\|\phi(\tilde{x},\tilde{a})\|_{(\beta I+U_{\ell-1}^{(\tau)})^{-1}}\cdot\left(\sqrt{\beta/d}\cdot\|\tilde{\theta}_{\ell-1}^{(\tau)}\| + T\varepsilon_{\mathrm{hoff}}\right) \\
&\quad + \sqrt{d}\|\phi(\tilde{x},\tilde{a})\|_{(\beta I+U_{\ell-1}^{(\tau)})^{-1}} \sum_{(\pi',v')\in\Psi_{\ell-1}^{(\tau)}} \left|\mathbb{E}^{\pi'}[\phi(\boldsymbol{x}_{\ell-1},\boldsymbol{a}_{\ell-1})^\top\tilde{\theta}_{\ell-1}^{(\tau)}\cdot\mathbb{I}\{\phi(\boldsymbol{x}_{\ell-1},\boldsymbol{a}_{\ell-1})^\top v'\ge 0\}]\right|,
\end{aligned}$$

and so using (144), the law of total expectation, and the triangle inequality, we get

$$\begin{aligned}
&\le \sqrt{d}\|\phi(\tilde{x},\tilde{a})\|_{(\beta I+U_{\ell-1}^{(\tau)})^{-1}}\cdot\left(\sqrt{\beta/d}\cdot\|\tilde{\theta}_{\ell-1}^{(\tau)}\| + T\varepsilon_{\mathrm{hoff}}\right) \\
&\quad + \sqrt{d}\|\phi(\tilde{x},\tilde{a})\|_{(\beta I+U_{\ell-1}^{(\tau)})^{-1}} \sum_{(\pi',v')\in\Psi_{\ell-1}^{(\tau)}} \left|\mathbb{E}^{\pi'}[\xi_\ell^{(\tau)}(\boldsymbol{x}_\ell,\widehat{\pi}_\ell^{(\tau)}(\boldsymbol{x}_\ell))\cdot\mathbb{I}\{\phi(\boldsymbol{x}_{\ell-1},\boldsymbol{a}_{\ell-1})^\top v'\ge 0\}]\right| \\
&\quad + \sqrt{d}\|\phi(\tilde{x},\tilde{a})\|_{(\beta I+U_{\ell-1}^{(\tau)})^{-1}} \sum_{(\pi',v')\in\Psi_{\ell-1}^{(\tau)}} \left|\mathbb{E}^{\pi'}[\xi_\ell^{(\tau)}(\boldsymbol{x}_\ell,\widehat{\pi}_\ell^\star(\boldsymbol{x}_\ell))\cdot\mathbb{I}\{\phi(\boldsymbol{x}_{\ell-1},\boldsymbol{a}_{\ell-1})^\top v'\ge 0\}]\right|, \\
&\le \sqrt{d}\|\phi(\tilde{x},\tilde{a})\|_{(\beta I+U_{\ell-1}^{(\tau)})^{-1}}\cdot\left(\sqrt{\beta/d}\cdot\|\tilde{\theta}_{\ell-1}^{(\tau)}\| + T\varepsilon_{\mathrm{hoff}} + 2T\varepsilon_{\mathsf{reg}}\right),
\end{aligned}$$

where the last inequality follows by Lemma L.2, the conditioning on $\mathcal{E}^{\mathsf{reg}}\cap\mathcal{E}^{\mathrm{hoff}}$, and the fact that $\widehat{\pi}^\star \in \Pi_{\mathsf{Bench}}$ (see (125)). Plugging this into (141), we get that for all $(\pi,v)\in\Psi_h^{(\tau)}$:

$$\begin{aligned}
&\left|\mathbb{E}^{\pi}[\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top\theta_{h,\ell}^{(\tau)}\cdot\mathbb{I}\{\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top v\ge 0\}]\right| \\
&\le \sqrt{d}\mathbb{E}^{\pi\circ_{h+1}\widehat{\pi}^{(\tau)}}\left[\mathbf{I}_{h,\ell,v}\cdot\|\phi(\boldsymbol{x}_{\ell-1},\boldsymbol{a}_{\ell-1})\|_{(\beta I+U_{\ell-1}^{(\tau)})^{-1}}\right]\cdot\left(\sqrt{\beta/d}\cdot\|\tilde{\theta}_{\ell-1}^{(\tau)}\| + T\varepsilon_{\mathrm{hoff}} + 2T\varepsilon_{\mathsf{reg}}\right),
\end{aligned}$$

where we recall that $\mathbf{I}_{h,\ell,v} = \mathbb{I}\{\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top v\ge 0\}\cdot\prod_{k=h+1}^{\ell-1}\mathbb{I}\{\mathrm{Rg}^{\mathsf{D}}(\boldsymbol{x}_k)<\gamma\}\le 1$. Thus, we have for all $(\pi,v)\in\Psi_h^{(\tau)}$:

$$\begin{aligned}
&\left|\mathbb{E}^{\pi}[\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top\theta_{h,\ell}^{(\tau)}\cdot\mathbb{I}\{\phi(\boldsymbol{x}_h,\boldsymbol{a}_h)^\top v\ge 0\}]\right| \\
&\le \sqrt{d}\mathbb{E}^{\pi\circ_{h+1}\widehat{\pi}^{(\tau)}}\left[\|\phi(\boldsymbol{x}_{\ell-1},\boldsymbol{a}_{\ell-1})\|_{(\beta I+U_{\ell-1}^{(\tau)})^{-1}}\right]\cdot\left(\sqrt{\beta/d}\cdot\|\tilde{\theta}_{\ell-1}^{(\tau)}\| + T\varepsilon_{\mathrm{hoff}} + 2T\varepsilon_{\mathsf{reg}}\right), \\
&\le \left(\sqrt{d}/2 + 2d^{3/2}\beta^{-1}\varepsilon_{\mathrm{hoff}}\right)\cdot\left(\sqrt{\beta/d}\cdot\|\tilde{\theta}_{\ell-1}^{(\tau)}\| + T\varepsilon_{\mathrm{hoff}} + 2T\varepsilon_{\mathsf{reg}}\right),
\end{aligned}$$

where the last inequality follows by Lemma L.2 and the conditioning on $\mathcal{E}^{\mathsf{reg}} \cap \mathcal{E}^{\mathrm{hoff}}$. Combining this with (140) and (135), we get that

$$
\begin{aligned}
&Q_h^{\widehat{\pi}^\star}(x,a) - \big(R(x,a) + \mathbb{E}[V_{h+1}^{(\tau)}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a]\big) \\
&= \mathbb{E}[V_{h+1}^{\widehat{\pi}^\star}(\boldsymbol{x}_{h+1}) - V_{h+1}^{(\tau)}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a], \\
&\leq \sqrt{d} \cdot \|\phi(x,a)\|_{(\beta I + U_h^{(\tau)})^{-1}} \cdot \left(\sqrt{\frac{\beta}{d}} \cdot \sum_{\ell=h+1}^{H} \|\theta_{h,\ell}^{(\tau)}\| + T(H-h)\varepsilon_{\mathrm{hoff}} \right) \\
&\quad + T \cdot \|\phi(x,a)\|_{(\beta I + U_h^{(\tau)})^{-1}} \cdot \left(\left(\frac{d}{2} + \frac{2d^2 \varepsilon_{\mathrm{hoff}}}{\beta} \right) \cdot \left(\sqrt{\frac{\beta}{d}} \cdot \sum_{\ell=h+1}^{H} \|\tilde{\theta}_{\ell-1}^{(\tau)}\| + T(H-h)\varepsilon_{\mathrm{hoff}} + 2T(H-h)\varepsilon_{\mathsf{reg}} \right) \right), \\
&\leq \|\phi(x,a)\|_{(\beta I + U_h^{(\tau)})^{-1}} \cdot \left(\left(\frac{24\sqrt{\beta}\tilde{d}H^4}{\gamma\lambda\sqrt{d}} + T\varepsilon_{\mathrm{hoff}} + T\varepsilon_{\mathsf{reg}} \right) \cdot \left(HTd + \frac{4d^2 HT\varepsilon_{\mathrm{hoff}}}{\beta} \right) \right), \qquad (145) \\
&\leq \|\phi(x,a)\|_{(\beta I + U_h^{(\tau)})^{-1}} \cdot \frac{\varepsilon}{4H}, \qquad (146)
\end{aligned}
$$

where (145) follows by the bounds on $\|\theta_{\ell-1}^{(\tau)}\|$ and $\|\tilde{\theta}_{\ell-1}^{(\tau)}\|$ from (138) and (142), respectively; and (146) follows from the choices of parameters $n_{\mathsf{traj}}, \mu, \nu, \lambda$, and β in Algorithm 6. On the other hand, we have that

$$
Q_h^{\widehat{\pi}^\star}(x,a) \leq H \quad \text{and} \quad R(x,a) + \mathbb{E}[V_{h+1}^{(\tau)}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] \geq 0.
$$

Combining this with (146) and the fact that $\min(H, c+b) \leq c + \min(H, b)$ for $c, b \geq 0$, we get

$$
\begin{aligned}
&Q_h^{\widehat{\pi}^\star}(x,a) \\
&\leq \min\left(H, R(x,a) + \mathbb{E}[V_{h+1}^{(\tau)}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] + \|\phi(x,a)\|_{(\beta I + U_h^{(\tau)})^{-1}} \cdot \frac{\varepsilon}{4H} \right), \\
&\leq R(x,a) + \mathbb{E}[V_{h+1}^{(\tau)}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] + \min\left(H, \|\phi(x,a)\|_{(\beta I + U_h^{(\tau)})^{-1}} \cdot \frac{\varepsilon}{4H} \right), \\
&\leq R(x,a) + \mathbb{E}[V_{h+1}^{(\tau)}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] + \min\left(H, \max_{\tilde{a} \in \mathcal{A}} \|\phi(x,\tilde{a})\|_{(\beta I + U_h^{(\tau)})^{-1}} \cdot \frac{\varepsilon}{4H} \right), \\
&= R(x,a) + \mathbb{E}[V_{h+1}^{(\tau)}(\boldsymbol{x}_{h+1}) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = a] + b_h^{(\tau)}(x) + \bar{b}_h^{(\tau)}(x) \cdot \mathbb{I}\{\|\varphi(x; W_h^{(\tau)})\| < \mu\}, \qquad (147) \\
&= Q_h^{(\tau)}(x,a) + \bar{b}_h^{(\tau)}(x) \cdot \mathbb{I}\{\|\varphi(x; W_h^{(\tau)})\| < \mu\},
\end{aligned}
$$

where (147) follows by the definitions of $b_h^{(\tau)}$ and $\bar{b}_h^{(\tau)}$ in (127), and the last inequality follows by the definition of $Q_h^{(\tau)}$ in (128). This shows (129) for $\ell = h$.
We show (130) for $\ell = h$. We have

$$
\begin{aligned}
&V_h^{\widehat{\pi}^\star}(x) - V_h^{(\tau)}(x) \\
&= Q_h^{\widehat{\pi}^\star}(x, \widehat{\pi}_h^\star(x)) - Q_h^{(\tau)}(x, \widehat{\pi}_h^{(\tau)}(x)), \\
&\leq Q_h^{\widehat{\pi}^\star}(x, \widehat{\pi}_h^\star(x)) - Q_h^{(\tau)}(x, \widehat{\pi}_h^{(\tau)}(x)) + \widehat{Q}_h^{(\tau)}(x, \widehat{\pi}_h^{(\tau)}(x)) - \widehat{Q}_h^{(\tau)}(x, \widehat{\pi}_h^\star(x)), \quad \text{(by definition of } \widehat{\pi}_h^{(\tau)}) \\
&= Q_h^{\widehat{\pi}^\star}(x, \widehat{\pi}_h^\star(x)) - Q_h^{(\tau)}(x, \widehat{\pi}_h^\star(x)) + Q_h^{(\tau)}(x, \widehat{\pi}_h^\star(x)) - Q_h^{(\tau)}(x, \widehat{\pi}_h^{(\tau)}(x)) \\
&\quad + \widehat{Q}_h^{(\tau)}(x, \widehat{\pi}_h^{(\tau)}(x)) - \widehat{Q}_h^{(\tau)}(x, \widehat{\pi}_h^\star(x)), \\
&\leq \xi_h^{(\tau)}(x, \widehat{\pi}_h^\star(x)) - \xi_h^{(\tau)}(x, \widehat{\pi}_h^{(\tau)}(x)) + \bar{b}_h^{(\tau)}(x) \cdot \mathbb{I}\{\|\varphi(x; W_h^{(\tau)})\| < \mu\},
\end{aligned}
$$

where the last inequality follows by (129) with $\ell = h$. This shows (130) for $\ell = h$ and completes the induction. Instantiation (130) with $\ell = 1$ and using the definition of $V_1^{(\tau)}$ in (128), we get that

$$
\begin{aligned}
&J(\widehat{\pi}^\star_{1:H}) - J(\widehat{\pi}^{(\tau)}_{1:H}) \\
&= \mathbb{E}\left[V_1^{\widehat{\pi}^\star}(\boldsymbol{x}_1)\right] - \mathbb{E}\left[V_1^{\widehat{\pi}^{(\tau)}}(\boldsymbol{x}_1)\right], \\
&= \mathbb{E}\left[V_1^{\widehat{\pi}^\star}(\boldsymbol{x}_1)\right] - \mathbb{E}\left[V_1^{(\tau)}(\boldsymbol{x}_1)\right] + \mathbb{E}\left[\sum_{h=1}^H b_h^{(\tau)}(\boldsymbol{x}_h)\right], \\
&\le \mathbb{E}\left[\xi_1^{(\tau)}(\boldsymbol{x}_1, \widehat{\pi}_1^\star(\boldsymbol{x}_1)) - \xi_1^{(\tau)}(\boldsymbol{x}_1, \widehat{\pi}_1^{(\tau)}(\boldsymbol{x}_1)) + \bar{b}_1^{(\tau)}(\boldsymbol{x}_1) \cdot \mathbb{I}\{\|\varphi(\boldsymbol{x}_1; W_1^{(\tau)})\| < \mu\}\right] + \mathbb{E}\left[\sum_{h=1}^H b_h^{(\tau)}(\boldsymbol{x}_h)\right], \\
&\le 2\varepsilon_{\mathtt{reg}} + \mathbb{E}\left[\bar{b}_1^{(\tau)}(\boldsymbol{x}_1) \cdot \mathbb{I}\{\|\varphi(\boldsymbol{x}_1; W_1^{(\tau)})\| < \mu\}\right] + \mathbb{E}\left[\sum_{h=1}^H b_h^{(\tau)}(\boldsymbol{x}_h)\right], \quad \text{(see below)} \qquad (148) \\
&\le 2\varepsilon_{\mathtt{reg}} + 2\mathbb{E}\left[\sum_{h=1}^H b_h^{(\tau)}(\boldsymbol{x}_h)\right],
\end{aligned}
$$

where (148) follows by Lemma L.2, the conditioning on $\mathcal{E}^{\mathtt{reg}} \cap \mathcal{E}^{\mathrm{hoff}}$, and the fact that $\widehat{\pi}^\star \in \Pi_{\mathtt{Bench}}$ (see (125)). Now, by the definition of $(b_h^{(\tau)})$ in (127), Lemma L.2, and the conditioning on $\mathcal{E}^{\mathrm{hoff}} \cap \mathcal{E}^{\mathtt{reg}}$, we get that

$$
J(\widehat{\pi}^\star_{1:H}) - J(\widehat{\pi}^{(\tau)}_{1:H}) \le 2\varepsilon_{\mathtt{reg}} + \frac{\varepsilon}{4} \cdot \left(1 + 4d\beta^{-1}\varepsilon_{\mathrm{hoff}}\right). \qquad (149)
$$

Suboptimality of $\widehat{\pi}_{1:H}$. Let τ' be as in Algorithm 1 just before the algorithm returns, and note that the policy $\widehat{\pi}_{1:H}$ satisfies $\widehat{\pi}_{1:H} = \widehat{\pi}^{(\tau')}_{1:H}$. Now, by Lemma L.3 and the conditioning on $\mathcal{E}^{\mathrm{eval}}$, we have that

$$
J(\widehat{\pi}^{(\tau)}_{1:H}) \le \max_{t \in [T]} J(\widehat{\pi}^{(t)}_{1:H}) \le J(\widehat{\pi}^{(\tau')}_{1:H}) + H\sqrt{\frac{2\log(T/\delta)}{n_{\mathtt{traj}}}}. \qquad (150)
$$

Combining (126), (149), and (150), we get

$$
J(\pi^\star) - J(\widehat{\pi}_{1:H}) \le H\sqrt{\frac{2\log(T/\delta)}{n_{\mathtt{traj}}}} + 2\varepsilon_{\mathtt{reg}} + \varepsilon \cdot \left(H + 2dH\beta^{-1}\varepsilon_{\mathrm{hoff}}\right) + H\sqrt{2d/d_\nu} \cdot \mu. \qquad (151)
$$

Using the expressions of $\varepsilon_{\mathrm{hoff}}$ and $\varepsilon_{\mathtt{reg}}$ in Lemma I.3 and Lemma J.5, respectively, and the choices of β, μ, ν, T, and $n_{\mathtt{traj}}$ in Algorithm 1, we get that the right-hand of side of (151) is at most ε, which completes the proof.

Number of oracle calls. A single call to the subroutine `FitOptVal` at iteration t makes Ht calls to the policy optimization oracle; this is because the problem in Line 13 of Algorithm 5 (i.e. $\min_{\tilde{\pi} \in \Pi_{\mathtt{Bench}}} \Delta_{h,\ell}(\pi, v, \tilde{\pi})$) has to be solved for all $\ell \in [h+1\,..\,H]$ and $(\pi, v) \in \Psi_{h-1}^{(t)}$, and we know that $|\Psi_{h-1}^{(t)}| \le t$. Thus, since Algorithm 1 calls `FitOptVal` TH times, the total number of calls to the policy optimization oracle is at most H^2T^2. Now the desired number of oracle calls follows by the choice of T in Algorithm 1. □

Appendix M. Implementation of the CSC oracle over Π_{Bench}

Lemma M.1. *Let $n \in \mathbb{N}$, $h \in [H]$, and $(c^{(1)}, x^{(1)}, a^{(1)}), \dots, (c^{(n)}, x^{(n)}, a^{(n)}) \in \mathbb{R} \times \mathcal{X}_h \times \mathcal{A}$ be given. Then, for the benchmark policy class Π_{Bench} in Section 2.3, it is possible to find*

$$\pi' \in \operatorname*{arg\,min}_{\pi \in \Pi_{\mathsf{Bench}}} \sum_{i=1}^{n} c^{(i)} \cdot \mathbb{I}\{\pi(x^{(i)}) = a^{(i)}\}, \tag{152}$$

in $O(\mathrm{poly}(n, d, A) \cdot (9n^2A^2/d)^{(d+1)^2})$ time.

Proof. Note that to solve (152), it suffices to enumerate points in the set

$$\mathrm{Val}_{\mathsf{Bench}} := \{(\pi(x^{(1)}), \dots, \pi(x^{(n)})) \mid \pi \in \Pi_{\mathsf{Bench}}\} \subseteq \mathcal{A}^n, \tag{153}$$

and take π' to be the policy corresponding to the point in $\mathrm{Val}_{\mathsf{Bench}}$ that minimizes the objective in (152). We know by Lemma 2.7 that the cardinality of $\mathrm{Val}_{\mathsf{Bench}}$ is at most $(9^2nA^2/d)^{(d+1)^2}$, but we still need a way of enumerating this set efficiently (in the dimension is constant). To do this, we will use the characterization of Π_{Bench} in the proof of Lemma 2.7 to reduce our problem to enumerating regions of a Euclidean space where certain polynomials change sign configurations.

As argued in the proof of Lemma 2.7, we have that for any $\pi \in \Pi_{\mathsf{Bench}}$, there exist $\theta_1, \dots, \theta_d$, $\tilde{\theta}_1, \tilde{\theta}_2 \in \mathbb{B}(H)$, and $\gamma > 0$ such that:

$$\pi(\cdot) = \bar{\pi}(\cdot; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma),$$

where

$$\begin{aligned}\bar{\pi}(\cdot; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma) := &\left(1 - \prod_{a,a' \in \mathcal{A}, i \in [d]} \mathbb{I}\left\{\varphi(\cdot, a, a')^\top \theta_i \geq \gamma\right\}\right) \cdot \tilde{\pi}(\cdot; \tilde{\theta}_1) \\ &+ \prod_{a,a' \in \mathcal{A}, i \in [d]} \mathbb{I}\left\{\varphi(\cdot, a, a')^\top \theta_i < \gamma\right\} \cdot \tilde{\pi}(\cdot; \tilde{\theta}_2),\end{aligned}$$

and $\tilde{\pi}(\cdot; \theta) = \arg\max_{a \in \mathcal{A}} \phi(\cdot, a)^\top \theta$. Furthermore, if we let $y_a^{(j)} := \phi(x^{(j)}, a)$ for $j \in [n]$ and $a \in \mathcal{A}$, then for any $\theta_{1:d}, \tilde{\theta}_{1:2} \in \mathbb{B}(H)$ and $\gamma \in \mathbb{R}$, the value of the tuple

$$(\bar{\pi}(x_1^{(1)}; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma), \dots, \bar{\pi}(x^{(n)}; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma)) \in \mathcal{A}^n$$

is completely determined by the signs of the linear functions (see proof of Lemma 2.7):

$$\begin{aligned}P_{i,j,a,a'}(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma) &:= \theta_i^\top y_a^{(j)} - \theta_i^\top y_{a'}^{(j)} - \gamma, \\ \widetilde{P}_{1,j,a,a'}(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma) &:= \tilde{\theta}_1^\top y_a^{(j)} - \tilde{\theta}_1^\top y_{a'}^{(j)}, \\ \widetilde{P}_{2,j,a,a'}(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma) &:= \tilde{\theta}_2^\top y_a^{(j)} - \tilde{\theta}_2^\top y_{a'}^{(j)}.\end{aligned}$$

In other words, if we let $(S_{i,j,a,a'}, \widetilde{S}_{1,j,a,a'}, \widetilde{S}_{2,j,a,a'}) := (\mathsf{sgn}(P_{i,j,a,a'}), \mathsf{sgn}(\widetilde{P}_{1,j,a,a'}), \mathsf{sgn}(\widetilde{P}_{2,j,a,a'}))$, then there is a surjective mapping from

$$\begin{aligned}&\mathrm{Sign}_{\mathsf{Bench}} \\ &:= \{(S_{i,j,a,a'}(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma), \widetilde{S}_{1,j,a,a'}(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma), \widetilde{S}_{2,j,a,a'}(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma))_{i \in [d], j \in [n], a,a' \in \mathcal{A}} \mid \theta_{1:d}, \tilde{\theta}_{1:2} \in \mathbb{B}(H), \gamma \in \mathbb{R}\}\end{aligned}$$

to the set $\text{Val}_{\text{Bench}}$ in (153). Note that the elements of the set $\text{Sign}_{\text{Bench}}$ take values in the finite set

$$(\{-1,1\}\times\{-1,1\}\times\{-1,1\})^{ndA^2},$$

and thus $\text{Sign}_{\text{Bench}}$ induces a partition over $\mathbb{B}(H)^{d+2}\times\mathbb{R}$, where each region of the partition corresponds to values of $(\theta_{1:d},\tilde{\theta}_{1:2},\gamma)$ mapping to a fixed sign tuple:

$$(S_{i,j,a,a'}(\theta_{1:d},\tilde{\theta}_{1:2},\gamma),\widetilde{S}_{1,j,a,a'}(\theta_{1:d},\tilde{\theta}_{1:2},\gamma),\widetilde{S}_{2,j,a,a'}(\theta_{1:d},\tilde{\theta}_{1:2},\gamma))_{i\in[d],j\in[n],a,a'\in\mathcal{A}}.$$

Thus, enumerating the elements of $\text{Val}_{\text{Bench}}$ reduces to enumerating the regions of this partition. Since $P_{i,j,a,a'},\widetilde{P}_{1,j,a,a'},\widetilde{P}_{2,j,a,a'}$ are linear functions, Lemma M.2 implies that enumerating the elements of the partition induced by $\text{Sign}_{\text{Bench}}$ can be done using $O(3ndA^2\cdot(9n^2A^2/d)^{(d+1)^2})$ calls to an oracle for checking the feasibility of a linear program over $\mathbb{R}^{(d+1)^2}$ with at most $3dnA^2$ constraints; this oracle can be implemented in $\text{poly}(d,A,n)$ time. We now explain what exactly it means to enumerate elements of the partition induced by $\text{Sign}_{\text{Bench}}$.

Enumerating partition elements. By calling Algorithm 8 with the input vectors corresponding to the linear functions $(P_{i,j,a,a'},\widetilde{P}_{1,j,a,a'},\widetilde{P}_{2,j,a,a'})_{i\in[d],j\in[n],a,a'\in\mathcal{A}}$, we get a set $\mathcal{G}$ of size at most $(9n^2A^2/d)^{(d+1)^2}$ such that for any tuple of signs $\text{tup}\in\text{Sign}_{\text{Bench}}$, there is a set $\mathcal{V}=\{v_1,\dots,v_N\}\in\mathcal{G}$ of vectors in $\mathbb{R}^K$, where $N\coloneqq 3dnA^2$ and $K\coloneqq(d+1)^2$ such that $\bigcap_{i\in[N]}\{\theta\mid\theta^\top v_i\ge 0\}\neq\varnothing$ and

$$\text{tup}=(\mathsf{sgn}(\theta^\top v_1),\dots,\mathsf{sgn}(\theta^\top v_N)),\quad \forall\theta\in\bigcap_{i\in[N]}\{\tilde{\theta}\mid\tilde{\theta}^\top v_i\ge 0\}.$$

This means that we can enumerate and evaluate the elements of $\text{Sign}_{\text{Bench}}$ (and thus also $\text{Val}_{\text{Bench}}$) by enumerating the elements of $\mathcal{G}$. □

Algorithm 8 Breath-First-Search for enumerating the state space regions corresponding to different combination of signs of a set of linear functions.

input: Vectors $v_1,\dots,v_N\in\mathbb{R}^K$. // These correspond to linear functions $\theta\mapsto v_i^\top\theta$.

Set $\mathcal{G}_1\leftarrow\varnothing$.
if $\{\theta\in\mathbb{R}^K\mid\theta^\top v_1\ge 0\}\neq\varnothing$ **then**
    $\mathcal{G}_1\leftarrow\mathcal{G}_1\cup\{\{v_1\}\}$.
if $\{\theta\in\mathbb{R}^N\mid\theta^\top v_1<0\}\neq\varnothing$ **then**
    $\mathcal{G}_1\leftarrow\mathcal{G}_1\cup\{\{-v_1\}\}$.
for $i=2,\dots,N$ **do**
    Set $\mathcal{G}_i\leftarrow\varnothing$.
    for $\mathcal{V}\in\mathcal{G}_{i-1}$ **do**
        if $\bigcap_{v\in\mathcal{V}\cup\{v_i\}}\{\theta\in\mathbb{R}^N\mid\theta^\top v\ge 0\}\neq\varnothing$ **then**
            $\mathcal{G}_i\leftarrow\mathcal{G}_i\cup\{\mathcal{V}\cup\{v_i\}\}$.
        if $\bigcap_{v\in\mathcal{V}\cup\{-v_i\}}\{\theta\in\mathbb{R}^N\mid\theta^\top v\ge 0\}\neq\varnothing$ **then**
            $\mathcal{G}_i\leftarrow\mathcal{G}_i\cup\{\mathcal{V}\cup\{-v_i\}\}$.
return $\mathcal{G}_N$.

Lemma M.2. *Let $N, K \in \mathbb{N}$ be given. Then, for any input vectors $v_1, \dots, v_N \in \mathbb{R}^K$, Algorithm 8 returns a set $\mathcal{G}_N$ such that for any tuple of signs*

$$tup \in \{(\mathsf{sgn}(v_1^\top \theta), \dots, \mathsf{sgn}(v_N^\top \theta)) \mid \theta \in \mathbb{R}^K\},$$

there exists an element $\mathcal{V} \in \mathcal{G}_N$ such that $\bigcap_{v \in \mathcal{V}} \{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\} \neq \varnothing$ and

$$tup = (\mathsf{sgn}(v_1^\top \theta), \dots, \mathsf{sgn}(v_N^\top \theta)), \quad \forall \theta \in \bigcap_{v \in \mathcal{V}} \{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\}.$$

Furthermore, the algorithm makes at most $N \cdot (8eN/K)^K$ calls to an oracle for checking the feasibility of a linear program over $\mathbb{R}^K$ with at most N constraints.

Lemma M.2 implies that the elements of set $\mathcal{G}_N$ returned by Algorithm 8 characterizes the partition over $\mathbb{R}^K$ that the set

$$\{(\mathsf{sgn}(v_1^\top \theta), \dots, \mathsf{sgn}(v_N^\top \theta)) \mid \theta \in \mathbb{R}^K\}$$

induces.

Proof of Lemma M.2. We will show by induction over $i = 1, \dots, N$ that the set $\mathcal{G}_i$ at the end of the i-th iteration of the outer loop of Algorithm 8 satisfies the property that for any tuple of signs

$$\text{tup} \in \{(\mathsf{sgn}(v_1^\top \theta), \dots, \mathsf{sgn}(v_i^\top \theta)) \mid \theta \in \mathbb{R}^K\},$$

there exists an element $\mathcal{V} \in \mathcal{G}_i$ such that $\bigcap_{v \in \mathcal{V}} \{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\} \neq \varnothing$ and

$$\text{tup} = (\mathsf{sgn}(v_1^\top \theta), \dots, \mathsf{sgn}(v_i^\top \theta)), \quad \forall \theta \in \bigcap_{v \in \mathcal{V}} \{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\}.$$

Instantiating this with $i = n$ implies the first claim of the lemma.

Base case $i = 1$. From Line 1-Line 5, the set $\mathcal{G}_1$ consists of:

- The elements $\{-v_i\}$ and $\{v_i\}$ if $\{\theta \in \mathbb{R}^K \mid \theta^\top v_i \geq 0\} \neq \varnothing$ and $\{\theta \in \mathbb{R}^K \mid \theta^\top v_i < 0\} \neq \varnothing$;
- The element $\{-v_i\}$ if $\{\theta \in \mathbb{R}^K \mid \theta^\top v_i \geq 0\} = \varnothing$ and $\{\theta \in \mathbb{R}^K \mid \theta^\top v_i < 0\} \neq \varnothing$;
- The element $\{v_i\}$ if $\{\theta \in \mathbb{R}^K \mid \theta^\top v_i \geq 0\} \neq \varnothing$ and $\{\theta \in \mathbb{R}^K \mid \theta^\top v_i < 0\} = \varnothing$.

In all cases, we have for any

$$\text{tup} \in \{\mathsf{sgn}(v_1^\top \theta) \mid \theta \in \mathbb{R}^K\},$$

there exists an element $\mathcal{V} \in \mathcal{G}_i$ such that $\bigcap_{v \in \mathcal{V}} \left\{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\right\} \neq \varnothing$ and

$$\text{tup} = \mathsf{sgn}(v_1^\top \theta), \quad \forall \theta \in \bigcap_{v \in \mathcal{V}} \left\{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\right\}.$$

General case. Now, let $i \in [N-1]$ and suppose that we have the property that for any

$$\text{tup} \in \{(\mathsf{sgn}(v_1^\top \theta), \dots, \mathsf{sgn}(v_i^\top \theta)) \mid \theta \in \mathbb{R}^K\},$$

there exists an element $\mathcal{V} \in \mathcal{G}_i$ such that $\bigcap_{v\in\mathcal{V}} \left\{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\right\} \neq \varnothing$ and

$$\text{tup} = (\text{sgn}(v_1^\top \theta), \ldots, \text{sgn}(v_i^\top \theta)), \quad \forall \theta \in \bigcap_{v\in\mathcal{V}} \left\{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\right\}.$$

We show that this property holds for i replaced by $i+1$. Let

$$\text{tup}' \in \{(\text{sgn}(v_1^\top \theta), \ldots, \text{sgn}(v_{i+1}^\top \theta)) \mid \theta \in \mathbb{R}^K\}.$$

Further, let $\theta' \in \mathbb{R}^K$ be such that $\text{tup}' \in (\text{sgn}(v_1^\top \theta'), \ldots, \text{sgn}(v_{i+1}^\top \theta'))$ and define

$$\overline{\text{tup}}' \coloneqq (\text{sgn}(v_1^\top \theta'), \ldots, \text{sgn}(v_i^\top \theta'));$$

Note that

$$\text{tup}' = (\overline{\text{tup}}'_1, \ldots, \overline{\text{tup}}'_i, \text{sgn}(v_{i+1}^\top \theta')). \tag{154}$$

By the induction hypothesis, there exists an element $\mathcal{V}' \in \mathcal{G}_i$ such that

$$\overline{\text{tup}}' = (\text{sgn}(v_1^\top \theta), \ldots, \text{sgn}(v_i^\top \theta)), \quad \forall \theta \in \bigcap_{v\in\mathcal{V}'} \left\{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\right\} \neq \varnothing. \tag{155}$$

On the other hand, by Line 7-Line 12 of Algorithm 8, the set $\mathcal{G}_{i+1}$ at the end of iteration $i+1$ contains:

- The elements $\mathcal{V}' \cup \{-v_{i+1}\}$ and $\mathcal{V}' \cup \{v_{i+1}\}$ if

  $$\bigcap_{v\in\mathcal{V}'\cup\{v_{i+1}\}} \{\theta \in \mathbb{R}^K \mid \theta^\top v \geq 0\} \neq \varnothing, \quad \text{and} \quad \bigcap_{v\in\mathcal{V}'\cup\{-v_{i+1}\}} \{\theta \in \mathbb{R}^K \mid \theta^\top v \geq 0\} \neq \varnothing; \tag{156}$$

  In this case, by (154) and the property of $\mathcal{V}'$ in (155), there exists $s \in \{-,+\}$ such that

  $$\text{tup}' = (\text{sgn}(v_1^\top \theta), \ldots, \text{sgn}(v_{i+1}^\top \theta)), \quad \forall \theta \in \bigcap_{v\in\mathcal{V}'\cup\{sv_{i+1}\}} \left\{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\right\} \overset{(156)}{\neq} \varnothing.$$

- The element $\mathcal{V}' \cup \{-v_{i+1}\}$ if

  $$\bigcap_{v\in\mathcal{V}'\cup\{v_{i+1}\}} \{\theta \in \mathbb{R}^K \mid \theta^\top v \geq 0\} = \varnothing, \quad \text{and} \quad \bigcap_{v\in\mathcal{V}'\cup\{-v_{i+1}\}} \{\theta \in \mathbb{R}^K \mid \theta^\top v \geq 0\} \neq \varnothing; \tag{157}$$

  In this case, by (154) and the property of $\mathcal{V}'$ in (155), we have $\text{sgn}(v_{i+1}^\top \theta') = -1$ and

  $$\text{tup}' = (\text{sgn}(v_1^\top \theta), \ldots, \text{sgn}(v_{i+1}^\top \theta)), \quad \forall \theta \in \bigcap_{v\in\mathcal{V}'\cup\{-v_{i+1}\}} \left\{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\right\} \overset{(157)}{\neq} \varnothing.$$

- The element $\mathcal{V}' \cup \{v_{i+1}\}$ if

  $$\bigcap_{v\in\mathcal{V}'\cup\{v_{i+1}\}} \{\theta \in \mathbb{R}^K \mid \theta^\top v \geq 0\} \neq \varnothing, \quad \text{and} \quad \bigcap_{v\in\mathcal{V}'\cup\{-v_{i+1}\}} \{\theta \in \mathbb{R}^K \mid \theta^\top v \geq 0\} = \varnothing. \tag{158}$$

  In this case, by (154) and the property of $\mathcal{V}'$ in (155), we have $\text{sgn}(v_{i+1}^\top \theta') = +1$ and

  $$\text{tup}' = (\text{sgn}(v_1^\top \theta), \ldots, \text{sgn}(v_{i+1}^\top \theta)), \quad \forall \theta \in \bigcap_{v\in\mathcal{V}'\cup\{v_{i+1}\}} \left\{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\right\} \overset{(158)}{\neq} \varnothing.$$

In all cases, the set $\mathcal{G}_{i+1}$ at the end of iteration $i+1$ will contain an element $\mathcal{V}$ such that $\bigcap_{v\in\mathcal{V}} \left\{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\right\} \neq \varnothing$ and

$$\text{tup}' = (\text{sgn}(v_1^\top \theta), \ldots, \text{sgn}(v_{i+1}^\top \theta)), \quad \forall \theta \in \bigcap_{v\in\mathcal{V}} \left\{\tilde{\theta} \in \mathbb{R}^K \mid v^\top \tilde{\theta} \geq 0\right\}.$$

This completes the induction step. We now bound the computational cost of running the algorithm.

Computational cost. The computational cost of Algorithm 8 is bounded by $N \cdot |\mathcal{G}_N|$ (due to the two for loops in the algorithm) times the cost of checking the feasibility of a linear program. Thus, it suffices to bound the number of elements in $\mathcal{G}_N$. The number of elements in $\mathcal{G}_N$ corresponds to the number of configurations of signs of the linear functions (polynomials of degree 1) $(\theta \mapsto v_i^\top \theta)_{i \in [N]}$. And so, by Lemma O.1, $|\mathcal{G}_N|$ is at most $(8eN/K)^K$. This completes the proof. □

Appendix N. Upper Bound on the Growth Function of Π_{Bench} (Proof of Lemma 2.7)

Proof of Lemma 2.7. Some aspects of this proof are inspired by proofs in (Jin, 2023).

Fix $n \in \mathbb{N}$ and $h \in [H]$. Note that by definition of Π_{Bench}, we have that for any $\pi \in \Pi_{\mathsf{Bench}}$, there exist $\theta_1, \dots, \theta_d, \tilde{\theta}_1, \tilde{\theta}_2 \in \mathbb{B}(H)$, and $\gamma > 0$ such that:

$$\pi(\cdot) = \bar{\pi}(\cdot; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma),$$

where

$$\begin{aligned}\bar{\pi}(\cdot; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma) \coloneqq &\left(1 - \prod_{a,a' \in \mathcal{A}, i \in [d]} \mathbb{I}\left\{\varphi(\cdot, a, a')^\top \theta_i \geq \gamma\right\}\right) \cdot \tilde{\pi}(\cdot; \tilde{\theta}_1) \\ &+ \prod_{a,a' \in \mathcal{A}, i \in [d]} \mathbb{I}\left\{\varphi(\cdot, a, a')^\top \theta_i < \gamma\right\} \cdot \tilde{\pi}(\cdot; \tilde{\theta}_2),\end{aligned}$$

and $\tilde{\pi}(\cdot; \theta) = \arg\max_{a \in \mathcal{A}} \phi(\cdot, a)^\top \theta$; note that the expression of $\bar{\pi}(\cdot; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma)$ is equivalent to the right-hand side of (7) with $\pi'(\cdot) \equiv \tilde{\pi}(\cdot; \tilde{\theta}_1)$ and $\pi''(\cdot) \equiv \tilde{\pi}(\cdot; \tilde{\theta}_2)$. Thus, we may write the growth function $\mathcal{G}_h(\Pi_{\mathsf{Bench}}, n)$ as

$$\mathcal{G}_h(\Pi_{\mathsf{Bench}}, n) = \max_{(x_1, \dots, x_n) \in \mathcal{X}_h^n} \left|\left\{(\bar{\pi}(x_1; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma), \dots, \bar{\pi}(x_d; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma)) \mid \theta_{1:d}, \tilde{\theta}_{1:2} \in \mathbb{B}(H), \gamma \in \mathbb{R}\right\}\right|.$$

Moving forward, we let

$$(x_1', \dots, x_n') \in \operatorname*{arg\,max}_{(x_1, \dots, x_n) \in \mathcal{X}_h^n} \left|\left\{(\bar{\pi}(x_1; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma), \dots, \bar{\pi}(x_d; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma)) \mid \theta_{1:d}, \tilde{\theta}_{1:2} \in \mathbb{B}(H), \gamma \in \mathbb{R}\right\}\right|,$$

and note that by definition of $\mathcal{G}_h(\Pi_{\mathsf{Bench}}, n)$, we have

$$\mathcal{G}_h(\Pi_{\mathsf{Bench}}, n) = \left|\left\{(\bar{\pi}(x_1'; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma), \dots, \bar{\pi}(x_d'; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma)) \mid \theta_{1:d}, \tilde{\theta}_{1:2} \in \mathbb{B}(H)\right\}\right|.$$

Finally, define the vectors $y_{j,a} \coloneqq \phi(x_j', a) \in \mathbb{R}^d$ for $j \in [n]$ and $a \in \mathcal{A}$. With this, observe that for any $\theta_{1:d}, \tilde{\theta}_{1:2} \in \mathbb{B}(H)$ and $\gamma \in \mathbb{R}$, the value of $(\bar{\pi}(x_1'; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma), \dots, \bar{\pi}(x_d'; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma))$ is fully determined by the signs of the following functions:

$$\begin{aligned}P_{i,j,a,a'}(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma) &\coloneqq \theta_i^\top y_{j,a} - \theta_i^\top y_{j,a'} - \gamma, \\ \widetilde{P}_{1,j,a,a'}(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma) &\coloneqq \tilde{\theta}_1^\top y_{j,a} - \tilde{\theta}_1^\top y_{j,a'}, \\ \widetilde{P}_{2,j,a,a'}(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma) &\coloneqq \tilde{\theta}_2^\top y_{j,a} - \tilde{\theta}_2^\top y_{j,a'}.\end{aligned}$$

In fact, we have that

- The signs of $P_{i,j,a,a'}(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma)$ for $i \in [d]$, $j \in [n]$, and $a, a' \in \mathcal{A}$ determine the value of

$$\prod_{a,a' \in \mathcal{A}, i \in [d]} \mathbb{I}\left\{\varphi(x'_j, a, a')^\top \theta_i < \gamma\right\}$$

  in the definition of $\bar{\pi}(x'_j; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma)$;

- The signs of $\widetilde{P}_{1,j,a,a'}(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma)$ for $j \in [n]$ and $a, a' \in \mathcal{A}$ determine the value of $\tilde{\pi}(\cdot; \tilde{\theta}_1)$ in the definition of $\bar{\pi}(\cdot; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma)$; and
- The signs of $\widetilde{P}_{2,j,a,a'}(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma)$ for $j \in [n]$, and $a, a' \in \mathcal{A}$ determine the value of $\tilde{\pi}(\cdot; \tilde{\theta}_2)$ in the definition of $\bar{\pi}(\cdot; \theta_{1:d}, \tilde{\theta}_{1:2}, \gamma)$.

Therefore, $\mathcal{G}_h(\Pi_{\mathsf{Bench}}, n)$ is bounded by the number of configuration of signs of the tuples

$$(P_{i,j,a,a'}, \widetilde{P}_{1,j,a,a'}, \widetilde{P}_{2,j,a,a'})_{i \in [d], j \in [n], a, a' \in \mathcal{A}}. \tag{159}$$

Since for each $i, j \in [d]$ and $a, a' \in \mathcal{A}$, $P_{i,j,a,a'}$, $\widetilde{P}_{1,j,a,a'}$, and $\widetilde{P}_{2,j,a,a'}$ are polynomials in $(\theta_{1:d}, \tilde{\theta}_{1:2}, \gamma)$ of degree 1, Lemma O.1 instantiated with $(p, N, K) = (1, 3dnA^2, d^2+2d+1)$ implies that the number of configuration of signs of the functions in (159) is at most $(24enA^2/d)^{d^2+2d+1}$, and so we have

$$\mathcal{G}_h(\Pi_{\mathsf{Bench}}, n) \le (24enA^2/d)^{d^2+2d} \le (9^2 nA^2/d)^{(d+1)^2}.$$

Combining this with the fact that $\mathcal{G}(\Pi_{\mathsf{Bench}}, n) = \max_{h \in [H]} \mathcal{G}_h(\Pi_{\mathsf{Bench}}, n)$ completes the proof. □

Appendix O. Helper Lemmas

Lemma O.1 (Goldberg and Jerrum (1993)). *Let $\{P_1, \dots, P_N\}$ be N polynomials of degree at most p in K reals variables with $N \geq K$, then the number of different configurations of signs of $\{P_1, \dots, P_N\}$ is at most $(8epN/K)^K$.*

Lemma O.2. *Let $\Pi' \subseteq \Pi$, $\pi \in \Pi$, $B > 0$, and $n \in \mathbb{N}$ be given. Further, let $(\boldsymbol{x}^{(i)}_{1:H}, \boldsymbol{a}^{(i)}_{1:H})_{i\in[n]}$ be n i.i.d. trajectories generated according to $\mathbb{P}^\pi$. Then, for any function $f : (\bigtimes_{h=1}^H \mathcal{X}_h) \times \mathcal{A}^H \to [-B, B]$, $h \in [H]$, and $\delta \in (0,1)$, there is an event of probability at least $1-\delta$ such that for all $\pi' \in \Pi'$:*

$$\left|\mathbb{E}^\pi[\mathbb{I}\{\pi'(\boldsymbol{x}_h) = \boldsymbol{a}_h\} \cdot f(\boldsymbol{x}_{1:H}, \boldsymbol{a}_{1:H})] - \frac{1}{n}\sum_{i\in[n]} \mathbb{I}\{\pi'(\boldsymbol{x}^{(i)}_h) = \boldsymbol{a}^{(i)}_h\} \cdot f(\boldsymbol{x}^{(i)}_{1:H}, \boldsymbol{a}^{(i)}_{1:H})\right|$$
$$\leq 4B\sqrt{\frac{\log(2\mathcal{G}_h(\Pi', n)/\delta)}{n}}.$$

Proof. Fix $f : (\bigtimes_{h=1}^H \mathcal{X}_h) \times \mathcal{A}^H \to [-B, B]$, $h \in [H]$, and $\delta \in (0,1)$. By a standard Rademacher complexity argument (see e.g. (Mohri et al., 2012, Section 11)), there is an event of probability at least $1-\delta$ such that for all $\pi' \in \Pi'$:

$$\left|\mathbb{E}^\pi[\mathbb{I}\{\pi'(\boldsymbol{x}_h) = \boldsymbol{a}_h\} \cdot f(\boldsymbol{x}_{1:H}, \boldsymbol{a}_{1:H})] - \frac{1}{n}\sum_{i\in[n]} \mathbb{I}\{\pi'(\boldsymbol{x}^{(i)}_h) = \boldsymbol{a}^{(i)}_h\} \cdot f(\boldsymbol{x}^{(i)}_{1:H}, \boldsymbol{a}^{(i)}_{1:H})\right|$$
$$\leq 2\mathfrak{R}_n + B\sqrt{\frac{\log(2/\delta)}{2n}}, \tag{160}$$

where

$$\mathfrak{R}_n \coloneqq \mathbb{E}_{(\boldsymbol{x}^{(i)}_{1:H}, \boldsymbol{a}^{(i)}_{1:H})_{i\in[d]}} \mathbb{E}_{\boldsymbol{\sigma}_{1:n}}\left[\sup_{\pi'\in\Pi'} \frac{1}{n}\sum_{i\in[n]} \boldsymbol{\sigma}_i \cdot \mathbb{I}\{\pi'(\boldsymbol{x}^{(i)}_h) = \boldsymbol{a}^{(i)}_h\} \cdot f(\boldsymbol{x}^{(i)}_{1:H}, \boldsymbol{a}^{(i)}_{1:H})\right], \tag{161}$$

and $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_n$ are i.i.d. Rademacher random variables. Now, by the weighted version of Massart's lemma (Lemma O.3), we have for any $(x^{(i)}_{1:H}, a^{(i)}_{1:H})_{i\in[n]}$:

$$\mathbb{E}_{\boldsymbol{\sigma}_{1:n}}\left[\sup_{\pi'\in\Pi'} \frac{1}{n}\sum_{i\in[n]} \boldsymbol{\sigma}_i \cdot \mathbb{I}\{\pi'(x^{(i)}_h) = a^{(i)}_h\} \cdot f(x^{(i)}_{1:H}, a^{(i)}_{1:H})\right] \leq B\sqrt{2n\log|\mathcal{J}|}, \tag{162}$$

where

$$\mathcal{J} \coloneqq \left\{\left(\mathbb{I}\{\pi'(x^{(1)}_h) = a^{(1)}_h\}, \dots, \mathbb{I}\{\pi'(x^{(n)}_h) = a^{(n)}_h\}\right) \mid (x^{(i)}_{1:H}, a^{(i)}_{1:H}) \in \left(\bigtimes_{h=1}^H \mathcal{X}_h\right) \times \mathcal{A}\right\}.$$

Note that $|\mathcal{J}| \leq \mathcal{G}_h(\Pi', n)$. Plugging this in (162) and using (160) and (161), we get the desired result. □

We now state a weighted version of Massart's lemma, which we use in the proof of Lemma O.2.

Lemma O.3. *Let $n \in \mathbb{N}$ and $B > 0$ be given and let $\mathcal{J} \subseteq \mathbb{R}^n$ be a finite set of points with $r = \max_{x_{1:n} \in \mathcal{J}} \|x\|$. Further, let $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_n$ be i.i.d. Rademacher random variables and let $\boldsymbol{w}_1, \dots, \boldsymbol{w}_n$ be i.i.d. random variables in $[-B, B]$. Then, we have*

$$\mathbb{E}_{\boldsymbol{w}_{1:n}} \mathbb{E}_{\boldsymbol{\sigma}_{1:n}} \left[\max_{x_{1:n} \in \mathcal{J}} \sum_{i=1}^{n} \boldsymbol{\sigma}_i \cdot x_i \cdot \boldsymbol{w}_i \right] \le rB\sqrt{2 \log |\mathcal{J}|}.$$

Proof. Fix $t > 0$. By Jensen's inequality, we have

$$\begin{aligned}
&\exp\left(t \mathbb{E}_{\boldsymbol{w}_{1:n}} \mathbb{E}_{\boldsymbol{\sigma}_{1:n}} \left[\max_{x_{1:n} \in \mathcal{J}} \sum_{i \in [n]} \boldsymbol{\sigma}_i x_i \boldsymbol{w}_i \right] \right) \\
&\le \mathbb{E}_{\boldsymbol{w}_{1:n}} \mathbb{E}_{\boldsymbol{\sigma}_{1:n}} \left[\exp\left(t \max_{x_{1:n} \in \mathcal{J}} \sum_{i \in [n]} \boldsymbol{\sigma}_i x_i \boldsymbol{w}_i \right) \right], \\
&\le \mathbb{E}_{\boldsymbol{w}_{1:n}} \mathbb{E}_{\boldsymbol{\sigma}_{1:n}} \left[\sum_{x_{1:n} \in \mathcal{J}} \exp\left(t \sum_{i \in [n]} \boldsymbol{\sigma}_i x_i \boldsymbol{w}_i \right) \right], \\
&= \sum_{x_{1:n} \in \mathcal{J}} \mathbb{E}_{\boldsymbol{w}_{1:n}} \mathbb{E}_{\boldsymbol{\sigma}_{1:n}} \left[\exp\left(t \sum_{i \in [n]} \boldsymbol{\sigma}_i x_i \boldsymbol{w}_i \right) \right], \\
&= \sum_{x_{1:n} \in \mathcal{J}} \mathbb{E}_{\boldsymbol{w}_{1:n}} \mathbb{E}_{\boldsymbol{\sigma}_{1:n}} \left[\prod_{i \in [n]} \exp\left(t \boldsymbol{\sigma}_i x_i \boldsymbol{w}_i \right) \right], \\
&= \sum_{x_{1:n} \in \mathcal{J}} \prod_{i \in [n]} \mathbb{E}_{\boldsymbol{w}_1} \mathbb{E}_{\boldsymbol{\sigma}_1} \left[\exp\left(t \boldsymbol{\sigma}_1 x_i \boldsymbol{w}_1 \right) \right], \quad \text{(by the i.i.d. assumption)} \\
&\le \sum_{x_{1:n} \in \mathcal{J}} \prod_{i \in [n]} \exp\left(\frac{(2tBx_i)^2}{8} \right), \quad \text{(Hoeffding's lemma)} \\
&= \sum_{x_{1:n} \in \mathcal{J}} \exp\left(\frac{t^2 B^2}{2} \sum_{i=1}^{n} x_i^2 \right), \\
&= |\mathcal{J}| \exp\left(t^2 B^2 r^2 / 2 \right), \quad \left(\text{since } r = \max_{x_{1:n} \in \mathcal{J}} \|x\| \right).
\end{aligned}$$

Taking the logarithm of both sides and dividing by t, we get

$$\mathbb{E}_{\boldsymbol{w}_{1:n}} \mathbb{E}_{\boldsymbol{\sigma}_{1:n}} \left[\max_{x_{1:n} \in \mathcal{J}} \sum_{i \in [n]} \boldsymbol{\sigma}_i x_i \boldsymbol{w}_i \right] \le \frac{\log |\mathcal{J}|}{t} + \frac{tB^2 r^2}{2}.$$

Taking $t = \frac{\sqrt{2 \log |\mathcal{J}|}}{Br}$ implies the desired result. □

Lemma O.4 (Union bound)**.** *Let $T, H \in \mathbb{N}$ and $\delta \in (0, 1)$ be given. Further, let $\mathcal{B}_1$ be an algorithm that runs in $T \in \mathbb{N}$ iterations. At each iteration, $\mathcal{B}_1$ makes a sequence of H calls to a subroutine $\mathcal{B}_2$. Let $\mathfrak{S}$ denote the state space of algorithm $\mathcal{B}_1$; the space capturing the values of all the internal variables of $\mathcal{B}_1$. Let $\mathbf{S}_{h,-}^{(t)} \in \mathfrak{S}$ denote the random state of $\mathcal{B}_1$ immediately before the hth call to $\mathcal{B}_2$ during the tth iteration; further, let $\mathbf{S}_{h,+}^{(t)} \in \mathfrak{S}$ denote the random state of $\mathcal{B}_1$ immediately*

after this call to $\mathcal{B}_2$. *Suppose that for any* $S^{(t)}_{h,-} \in \mathfrak{S}$, *there is an event* $\mathcal{E}^{(t)}_h(S^{(t)}_{h,-}) \subset \mathfrak{S}$ *such that* $\mathbb{P}[\mathbf{S}^{(t)}_{h,+} \in \mathcal{E}^{(t)}_h(S^{(t)}_{h,-})] \geq 1-\delta$. *Then, with probability at least* $1-\delta HT$, *for all* $t \in [T]$ *and* $h \in [H]$, *we have* $\mathbf{S}^{(t)}_{h,+} \in \mathcal{E}^{(t)}_h(\mathbf{S}^{(t)}_{h,-})$.

Proof. Let $\mathcal{E}$ be the event defined by

$$\mathcal{E} \coloneqq \left\{ \prod_{t=1}^{T} \prod_{h=1}^{H} \mathbb{I}\{\mathbf{S}^{(t)}_{h,+} \in \mathcal{E}^{(t)}_h(\mathbf{S}^{(t)}_{h,-})\} = 1 \right\}.$$

We need to show that $\mathbb{P}[\mathcal{E}] \geq 1-\delta HT$. To this end, we note that by the chain rule of probability, we have

$$\begin{aligned} \mathbb{P}[\mathcal{E}] &= \prod_{t=1}^{T} \prod_{h=1}^{H} \mathbb{E}\left[\mathbb{P}[\mathbf{S}^{(t)}_{h,+} \in \mathcal{E}^{(t)}_h(\mathbf{S}^{(t)}_{h,-}) \mid \mathbf{S}^{(t)}_{h,-}] \right], \\ &\geq \prod_{t=1}^{T} \prod_{h=1}^{H} (1-\delta), \qquad (163) \\ &\geq 1 - TH\delta, \end{aligned}$$

where (163) follows by the fact that $\mathbb{P}[\mathbf{S}^{(t)}_{h,+} \in \mathcal{E}^{(t)}_h(S^{(t)}_{h,-})] \geq 1-\delta$ for all $S^{(t)}_{h,-} \in \mathfrak{S}$, and the last inequality follows by the fact that for any sequence $x_1, \ldots, x_T \in (0,1)$, $\prod_{i \in [T]}(1-x_i) \geq 1 - \sum_{i \in [T]} x_i$. □

Lemma O.5. *Let* $L > 0$ *be given and consider a collection of random matrices* $\{\boldsymbol{M}^z\}_{z \in \mathcal{Z}}$, *where* $\mathcal{Z}$ *is some abstract set, such that* $\|\boldsymbol{M}^z\|_{\mathrm{op}} \leq L$, *for all* $z \in \mathcal{Z}$. *Suppose that for some* $\varepsilon', \delta \in (0,1)$ *and* $\mathcal{K} \subseteq \mathbb{B}(1)$, *we have that for all* $u \in \mathcal{K}$ *and* $v \in \mathbb{B}(1)$, *with probability at least* $1-\delta$:

$$\forall z \in \mathcal{Z}, \quad u^\top \boldsymbol{M}^z v \leq \varepsilon'. \qquad (164)$$

Then for any $\varepsilon'' \in (0,1)$, *with probability at least* $1 - \delta \cdot (3/\varepsilon'')^{2d}$, *we have for all* $u \in \mathcal{K}$, $v \in \mathbb{B}(1)$, *and* $z \in \mathcal{Z}$:

$$u^\top \boldsymbol{M}^z v \leq \varepsilon' + 2L\varepsilon''.$$

Proof. Let $\mathcal{N}(\mathbb{B}(1), \|\cdot\|, \varepsilon'')$ be an ε''-epsilon net of $\mathcal{B}(1)$ in $\|\cdot\|$. Further, let $\mathcal{K}' \coloneqq \mathcal{K} \cap \mathcal{N}(\mathbb{B}(1), \|\cdot\|, \varepsilon'')$. We note that

$$|\mathcal{K}'| \leq |\mathcal{N}(\mathbb{B}(1), \|\cdot\|, \varepsilon'')| \leq (3/\varepsilon'')^d.$$

By (164) and the union bound over elements in $\mathcal{K}' \times \mathcal{N}(\mathbb{B}(1), \|\cdot\|, \varepsilon'')$, there is an event $\mathcal{E}$ of probability at least $1 - \delta \cdot (3/\varepsilon'')^{2d}$ under which for all $u' \in \mathcal{K}'$, $v' \in \mathcal{N}(\mathbb{B}(1), \|\cdot\|, \varepsilon'')$, and $z \in \mathcal{Z}$:

$$(u')^\top \boldsymbol{M}^z v' \leq \varepsilon'. \qquad (165)$$

For the rest of the proof, we condition on $\mathcal{E}$. Since $\mathcal{K}'$ is an ε''-net of $\mathcal{K}$, we have that for any $u \in \mathcal{K}$ and $v \in \mathbb{B}(1)$:

$$\inf_{u' \in \mathcal{K}', v' \in \mathcal{N}(\mathbb{B}(1), \|\cdot\|, \varepsilon'')} \|u - u'\| \vee \|v - v'\| \leq \varepsilon''. \qquad (166)$$

Therefore, we have that for any $u \in \mathcal{K}$, $v \in \mathbb{B}(1)$, and $z \in \mathcal{Z}$:

$$\begin{aligned}
u^\top \boldsymbol{M}^z v &= \inf_{u' \in \mathcal{K}', v' \in \mathcal{N}(\mathbb{B}(1), \|\cdot\|, \varepsilon'')} \left\{ (u')^\top \boldsymbol{M}^z v' + (u')^\top \boldsymbol{M}^z (v - v') + (u - u')^\top \boldsymbol{M}^z v \right\}, \\
&\leq \sup_{u' \in \mathcal{K}', v' \in \mathcal{N}(\mathbb{B}(1), \|\cdot\|, \varepsilon'')} (u')^\top \boldsymbol{M}^z v' + \inf_{u' \in \mathcal{K}', v' \in \mathcal{N}(\mathbb{B}(1), \|\cdot\|, \varepsilon'')} \left\{ (u')^\top \boldsymbol{M}^z (v - v') + (u - u')^\top \boldsymbol{M}^z v \right\}, \\
&\leq \varepsilon' + \inf_{u' \in \mathcal{K}', v' \in \mathcal{N}(\mathbb{B}(1), \|\cdot\|, \varepsilon'')} \left\{ L\|v - v'\| + L\|u - u'\| \right\}, \quad \text{(by (165) and Cauchy Schwarz)} \\
&\leq \varepsilon' + 2L\varepsilon'',
\end{aligned}$$

where the last inequality follows by (166). This completes the proof. □

We take Assumption O.1 and Theorem O.1 from Lattimore and Szepesvári (2020).

Assumption O.1 (Prerequisites for Theorem O.1)**.** *Let $\lambda > 0$. For $k \in \mathbb{N}$, let Y_k be random variables taking values in $\mathbb{R}^d$. For some $\theta_\star \in \mathbb{R}^d$, let $Y_k = \langle X_k, \theta_\star \rangle + \eta_k$ for all $k \in \mathbb{N}$. Here, η_k is a conditionally 1-subgaussian random variable; that is, it satisfies:*

$$\text{for all } \alpha \in \mathbb{R} \text{ and } t \geq 1, \qquad \mathbb{E}[\exp(\alpha \eta_k) \,|\, \mathcal{F}_{k-1}] \leq \exp\left(\frac{\alpha^2}{2}\right) \quad a.s.,$$

where $\mathcal{F}_{k-1}$ is such that $X_1, Y_1, \ldots, X_{k-1}, Y_{k-1}, X_k$ are $\mathcal{F}_{k-1}$-measurable.

Theorem O.1 (Lattimore and Szepesvári (2020), Theorem 20.5)**.** *Let $\zeta \in (0, 1)$. Under Assumption O.1, with probability at least $1 - \zeta$, it holds that for all $k \in \mathbb{N}$,*

$$\|\hat{\theta}_k - \theta_\star\|_{\Sigma_k(\lambda)} < \sqrt{\lambda}\|\theta_\star\|_2 + \sqrt{2\log(1/\zeta) + \log\left(\frac{\det \Sigma_k(\lambda)}{\lambda^d}\right)},$$

where for $k \in \mathbb{N}$,

$$\Sigma_k(\lambda) = \lambda I + \sum_{s=1}^{k} X_s X_s^\top \quad \text{and} \quad \hat{\theta}_k = \Sigma_k(\lambda)^{-1} \sum_{s=1}^{k} X_s Y_s.$$

Lemma O.6. *Let $\beta > 0$ be given. Then, for any sequence of vectors $u^{(1)}, u^{(2)}, \ldots$ in $\mathbb{B}(1)$, we have for all $T \in \mathbb{N}$:*

$$\sum_{t \in [T]} \|u^{(t)}\|_{(\beta I + U^{(t)})^{-1}} \leq \sqrt{Td\log(1 + T/\beta)},$$

where $U^{(t)} \coloneqq \sum_{\tau \in [t-1]} u^{(\tau)} (u^{(\tau)})^\top$.

Proof. By (Hazan et al., 2007, Lemma 11), we have

$$\sum_{t \in [T]} \|u^{(t)}\|^2_{(\beta I + U^{(t)})^{-1}} = \sum_{t \in [T]} (u^{(t)})^\top (\beta I + U^{(t)})^{-1} u^{(t)} \leq d\log(1 + T/\beta). \tag{167}$$

Now, by Jensen's inequality, we have

$$\begin{aligned}\sum_{t\in[T]} \|u^{(t)}\|_{(\beta I+U^{(t)})^{-1}} &\leq T\cdot\frac{1}{T}\sum_{t\in[T]}\sqrt{\|u^{(t)}\|^2_{(\beta I+U^{(t)})^{-1}}},\\ &\leq T\sqrt{\frac{1}{T}\sum_{t\in[T]}\|u^{(t)}\|^2_{(\beta I+U^{(t)})^{-1}}},\\ &\leq \sqrt{Td\log(1+T/\beta)},\end{aligned}$$

where the last inequality follows by (167). This completes the proof. □

We require the classical performance difference lemma from Kakade (2003).

Lemma O.7 (Performance Difference Lemma)**.** *Let $\pi^\star, \pi \in \Pi$ be arbitrary, and Q_t^π be as defined in (2). Then, for any $h \geq 1$,*

$$\mathbb{E}^{\pi^\star}\left[\sum_{t=1}^h R(\boldsymbol{x}_t,\boldsymbol{a}_t)\right] - \mathbb{E}^{\pi}\left[\sum_{t=1}^h R(\boldsymbol{x}_t,\boldsymbol{a}_t)\right] = \sum_{t=1}^h \mathbb{E}^{\pi^\star}\left[Q_t^\pi(\boldsymbol{x}_t,\pi^\star(\boldsymbol{x}_t)) - Q_t^\pi(\boldsymbol{x}_t,\pi(\boldsymbol{x}_t))\right].$$

Lemma O.8 (Skip-step value decomposition)**.** *Let $\mathcal{K}_1 \subseteq \mathcal{X}_1, \ldots, \mathcal{K}_H \subseteq \mathcal{X}_H$ be arbitrary subsets and let $\pi'_{1:H}, \widehat{\pi}_{1:H}, \widehat{\pi}'_{1:H} \in \Pi$ be such that for all $h \in [H]$ and $x \in \mathcal{X}_h$, $\widehat{\pi}'_h(x) = \mathbb{I}\{x \in \mathcal{K}_h\}\cdot\widehat{\pi}_h(x) + \mathbb{I}\{x \notin \mathcal{K}_h\}\cdot\pi'_h(x)$. Further, let $(\tilde{r}_h : \mathcal{X}_h \times \mathcal{A} \to \mathbb{R})_{h\in[H]}$ be arbitrary functions and define*

$$V_h(x) \coloneqq \mathbb{E}^{\widehat{\pi}'}\left[\sum_{\ell=h}^H \tilde{r}_\ell(\boldsymbol{x}_\ell,\boldsymbol{a}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = \widehat{\pi}'_h(x)\right], \tag{168}$$

for $h \in [H]$ and $x \in \mathcal{X}_h$. Then, for all $h \in [H]$ and $x \in \mathcal{X}_h$, we have

$$\begin{aligned}V_h(x) = &\sum_{\ell=h}^H \mathbb{E}^{\widehat{\pi}}\left[\mathbb{I}\{\boldsymbol{x}_\ell \notin \mathcal{K}_\ell\}\cdot\prod_{k=h}^{\ell-1}\mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\}\cdot V_\ell(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = \widehat{\pi}'_h(x)\right]\\ &+ \sum_{\ell=h}^H \mathbb{E}^{\widehat{\pi}}\left[\prod_{k=h}^{\ell}\mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\}\cdot\tilde{r}_\ell(\boldsymbol{x}_\ell,\boldsymbol{a}_\ell) \mid \boldsymbol{x}_h = x, \boldsymbol{a}_h = \widehat{\pi}'_h(x)\right],\end{aligned} \tag{169}$$

with the convention that $\prod_{k=h}^{h-1}\mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} = 1$.

Proof. We prove the result via backward induction over $h = H, \ldots, 1$. Suppose that (169) holds for $h \in [H]$. We show that it holds for $h-1$. Fix $x \in \mathcal{X}_{h-1}$. If $x \notin \mathcal{K}_{h-1}$, we trivially have that

$$\sum_{\ell=h-1}^H \mathbb{E}^{\widehat{\pi}}\left[\mathbb{I}\{\boldsymbol{x}_\ell \notin \mathcal{K}_\ell\}\cdot\prod_{k=h-1}^{\ell-1}\mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\}\cdot V_\ell(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}'_{h-1}(x)\right] = V_{h-1}(x)$$

and

$$\sum_{\ell=h-1}^H \mathbb{E}^{\widehat{\pi}}\left[\prod_{k=h-1}^{\ell}\mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\}\cdot\tilde{r}_\ell(\boldsymbol{x}_\ell,\boldsymbol{a}_\ell) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}'_{h-1}(x)\right] = 0.$$

Therefore,

$$V_{h-1}(x) = \sum_{\ell=h-1}^{H} \mathbb{E}^{\widehat{\pi}}\left[\mathbb{I}\{\boldsymbol{x}_\ell \notin \mathcal{K}_\ell\} \cdot \prod_{k=h-1}^{\ell-1} \mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} \cdot V_\ell(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}'_{h-1}(x)\right] + \sum_{\ell=h-1}^{H} \mathbb{E}^{\widehat{\pi}}\left[\prod_{k=h-1}^{\ell} \mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} \cdot \tilde{r}_\ell(\boldsymbol{x}_\ell, \boldsymbol{a}_\ell) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}'_{h-1}(x)\right].$$

Now, suppose that $x \in \mathcal{K}_{h-1}$. Then, from (168) and the fact that $\widehat{\pi}'_h(x) = \mathbb{I}\{x \in \mathcal{K}_{h-1}\} \cdot \widehat{\pi}_{h-1}(x) + \mathbb{I}\{x \notin \mathcal{K}_{h-1}\} \cdot \pi'_{h-1}(x)$, we have

$$\begin{aligned}
&V_{h-1}(x) \\
&= \mathbb{I}\{x \in \mathcal{K}_{h-1}\} \cdot \tilde{r}_{h-1}(x, \widehat{\pi}_{h-1}(x)) \\
&\quad + \mathbb{E}^{\widehat{\pi}'}\left[\mathbb{I}\{\boldsymbol{x}_{h-1} \in \mathcal{K}_{h-1}\} \cdot \sum_{\ell=h}^{H} \tilde{r}_\ell(\boldsymbol{x}_\ell, \boldsymbol{a}_\ell) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right], \\
&= \mathbb{I}\{x \in \mathcal{K}_{h-1}\} \cdot \tilde{r}_{h-1}(x, \widehat{\pi}_{h-1}(x)) + \mathbb{E}\left[\mathbb{I}\{\boldsymbol{x}_{h-1} \in \mathcal{K}_{h-1}\} \cdot V_h(\boldsymbol{x}_h) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right], \\
&= \mathbb{I}\{x \in \mathcal{K}_{h-1}\} \cdot \tilde{r}_{h-1}(x, \widehat{\pi}_{h-1}(x)) \\
&\quad + \mathbb{E}\left[\mathbb{I}\{\boldsymbol{x}_{h-1} \in \mathcal{K}_{h-1}\} \cdot \mathbb{I}\{\boldsymbol{x}_h \notin \mathcal{K}_h\} \cdot V_h(\boldsymbol{x}_h) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right] \\
&\quad + \mathbb{E}\left[\mathbb{I}\{\boldsymbol{x}_{h-1} \in \mathcal{K}_{h-1}\} \cdot \mathbb{I}\{\boldsymbol{x}_h \in \mathcal{K}_h\} \cdot V_h(\boldsymbol{x}_h) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right]. \qquad (170)
\end{aligned}$$

Now, by the induction hypothesis, i.e. (169), we have that for all $x' \in \mathcal{K}_h$:

$$V_h(x') = \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}}\left[\mathbb{I}\{\boldsymbol{x}_\ell \notin \mathcal{K}_\ell\} \cdot \prod_{k=h}^{\ell-1} \mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} \cdot V_\ell(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_h = x', \boldsymbol{a}_h = \widehat{\pi}_h(x')\right] + \sum_{\ell=h}^{H} \mathbb{E}^{\widehat{\pi}}\left[\prod_{k=h}^{\ell} \mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} \cdot \tilde{r}_\ell(\boldsymbol{x}_\ell, \boldsymbol{a}_\ell) \mid \boldsymbol{x}_h = x', \boldsymbol{a}_h = \widehat{\pi}_h(x')\right].$$

Plugging this into (170) and using the law of total expectation, we get that

$$
\begin{aligned}
&V_{h-1}(x)\\
&= \mathbb{I}\{x \in \mathcal{K}_{h-1}\} \cdot \tilde{r}_{h-1}(x, \widehat{\pi}_{h-1}(x))\\
&\quad + \mathbb{E}^{\widehat{\pi}}\left[\mathbb{I}\{\boldsymbol{x}_{h-1} \in \mathcal{K}_{h-1}\} \cdot \mathbb{I}\{\boldsymbol{x}_h \notin \mathcal{K}_h\} \cdot V_h(\boldsymbol{x}_h) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right]\\
&\quad + \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}}\left[\mathbb{I}\{\boldsymbol{x}_{h-1} \in \mathcal{K}_{h-1}\} \cdot \mathbb{I}\{\boldsymbol{x}_h \in \mathcal{K}_h\} \cdot \mathbb{I}\{\boldsymbol{x}_\ell \notin \mathcal{K}_\ell\} \cdot \prod_{k=h}^{\ell-1} \mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} \cdot V_\ell(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right]\\
&\quad + \sum_{\ell=h}^{H} \mathbb{E}^{\widehat{\pi}}\left[\mathbb{I}\{\boldsymbol{x}_{h-1} \in \mathcal{K}_{h-1}\} \cdot \mathbb{I}\{\boldsymbol{x}_h \in \mathcal{K}_h\} \cdot \prod_{k=h}^{\ell} \mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} \cdot \tilde{r}_\ell(\boldsymbol{x}_\ell, \boldsymbol{a}_\ell) \mid x_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right],\\
&= \mathbb{I}\{x \in \mathcal{K}_{h-1}\} \cdot \tilde{r}_{h-1}(x, \widehat{\pi}_{h-1}(x))\\
&\quad + \mathbb{E}^{\widehat{\pi}}\left[\mathbb{I}\{\boldsymbol{x}_{h-1} \in \mathcal{K}_{h-1}\} \cdot \mathbb{I}\{\boldsymbol{x}_h \notin \mathcal{K}_h\} \cdot V_h(\boldsymbol{x}_h) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right]\\
&\quad + \sum_{\ell=h+1}^{H} \mathbb{E}^{\widehat{\pi}}\left[\mathbb{I}\{\boldsymbol{x}_\ell \notin \mathcal{K}_\ell\} \cdot \prod_{k=h-1}^{\ell-1} \mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} \cdot V_\ell(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right]\\
&\quad + \sum_{\ell=h}^{H} \mathbb{E}^{\widehat{\pi}}\left[\prod_{k=h-1}^{\ell} \mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} \cdot \tilde{r}_\ell(\boldsymbol{x}_\ell, \boldsymbol{a}_\ell) \mid x_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right],\\
&= \sum_{\ell=h}^{H} \mathbb{E}^{\widehat{\pi}}\left[\mathbb{I}\{\boldsymbol{x}_\ell \notin \mathcal{K}_\ell\} \cdot \prod_{k=h-1}^{\ell-1} \mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} \cdot V_\ell(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right]\\
&\quad + \sum_{\ell=h-1}^{H} \mathbb{E}^{\widehat{\pi}}\left[\prod_{k=h-1}^{\ell} \mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} \cdot \tilde{r}_\ell(\boldsymbol{x}_\ell, \boldsymbol{a}_\ell) \mid x_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right],\\
&= \sum_{\ell=h-1}^{H} \mathbb{E}^{\widehat{\pi}}\left[\mathbb{I}\{\boldsymbol{x}_\ell \notin \mathcal{K}_\ell\} \cdot \prod_{k=h-1}^{\ell-1} \mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} \cdot V_\ell(\boldsymbol{x}_\ell) \mid \boldsymbol{x}_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right]\\
&\quad + \sum_{\ell=h-1}^{H} \mathbb{E}^{\widehat{\pi}}\left[\prod_{k=h-1}^{\ell} \mathbb{I}\{\boldsymbol{x}_k \in \mathcal{K}_k\} \cdot r_\ell(\boldsymbol{x}_\ell, \boldsymbol{a}_\ell) \mid x_{h-1} = x, \boldsymbol{a}_{h-1} = \widehat{\pi}_{h-1}(x)\right],
\end{aligned}
$$

where the last inequality follows by the fact that $x \in \mathcal{K}_{h-1}$. This shows (169) with h replaced by $h-1$ and completes the induction. □