# Appendix

In the subsequent sections, we begin by highlighting related works, establishing the context and relevance of this paper within the broader academic discourse in Section A. The experimental intricacies pertaining to the motivating example introduced in Section 3 are detailed in Section B. The pseudo-code for our PDRVI-L algorithm is delineated in Section C. Within Section D, we elucidate the methodologies and findings from our experiments on the American option and CartPole. Section E presents the rigorous proof for Theorem 4. Meanwhile, Section F offers detailed proofs for Theorem 5.2, Theorem 5.3, and Theorem 5.1.

## A Related Work

**Offline RL:** Recent research interests arouse to design offline RL algorithms with fewer dataset requirements based on a shared intuition called pessimism, i.e., the agent can act conservatively in the face of state-action pairs that the dataset has not covered. Empirical evidence has emerged (Fujimoto et al., 2019; Wu et al., 2019; Kumar et al., 2019; Fujimoto & Gu, 2021; Kumar et al., 2020; Kostrikov et al., 2021; Wu et al., 2021; Wang et al., 2018; Chen et al., 2020; Yang et al., 2021b; Kostrikov et al., 2022). Jin et al. (2021) prove that a pessimistic variant of the value iteration algorithm can achieve sample-efficient suboptimality under a mild data coverage assumption. Xie et al. (2021) introduce the notion of Bellman consistent pessimism to design a general function approximation algorithm. Rashidinejad et al. (2021) design the lower confidence bound algorithm utilizing pessimism in the face of uncertainty and show it is almost adaptively optimal in MDPs.

**Linear Function Approximation:** Research interests on the provable efficient RL under the linear model representations have emerged in recent years. Yang & Wang (2019) propose a parametric $Q$-learning algorithm to find an approximate-optimal policy with access to a generative model. Zanette et al. (2021) considers the Linear Bellman Complete model and designs the efficient actor-critic algorithm that achieves improvement in dependence on $d$. Yin et al. (2022) designs the variance-aware pessimistic value iteration to improve the suboptimality bounds over the best-known existing results. On the other hand, Wang et al. (2020); Zanette (2021) prove the statistical hardness of offline RL with linear representations by proving that the sample sizes could be exponential in the problem horizon for the value estimation task of any policy.

**Robust MDP and RL:** The robust optimization approach has been used to address the parameters uncertainty in MDPs first by Satia & Lave Jr (1973) and later by Xu & Mannor (2010); Iyengar (2005); Nilim & El Ghaoui (2005); Wiesemann et al. (2013); Kaufman & Schaefer (2013); Ho et al. (2018; 2021); Wiesemann et al. (2013). Although flourishing in the supervised learning (Namkoong & Duchi, 2017; Bertsimas et al., 2018; Duchi & Namkoong, 2021; Duchi et al., 2021), few works consider computing the optimal robust policy for RL. For online RL, a line of work has considered learning the optimal MDP policy under worst-case perturbations of the observation or environmental dynamics (Rajeswaran et al., 2016; Pattanaik et al., 2017; Huang et al., 2017; Pinto et al., 2017; Zhang et al., 2020). For offline RL, Zhou et al. (2021b) studies the distributionally robust policy with the offline dataset, where they focus on the KL ambiguity set and $(s, a)$-rectangular assumption and develop a value iteration algorithm. Yang et al. (2021a) improve the results in Zhou et al. (2021b) and extend the algorithms to other uncertainty sets. However, current theoretical advances mainly focus on tabular settings.

Among the previous work, one of the closest works to ours is Tamar et al. (2014), which develops a robust ADP method based on a projected Bellman equation. Based on this, Badrinath & Kalathil (2021) address the model-free robust RL with large state spaces by the proposed robust least squares policy iteration algorithm. While both provide the convergence guarantee for their algorithm, as shown in Section 3, their robustify-then-approximate (RTA) design fail to exploit the latent structure of the problem and may lead to the conflict between robustness and approximator, which finally yields suboptimal decisions. Besides, Panaganti et al. (2022a) considers the robust RL problem with general function approximator. Their algorithmic design highly depend on the choice of the ambiguity set and have no theoretical guarantee under weaker data coverage condition. The other closed work to ours is Goyal & Grand-Clement (2022), which considers a more general assumption for the ambiguity set, called $d$-rectangular [1] for MDPs with low dimensional linear representation. They mainly focus

[1] Goyal & Grand-Clement (2022) call it $r$-rectangular.

on the optimal policy structure for robust MDPs and the computational cost given the ambiguity set. In contrast, we study the offline RL setting and focus on the linear function approximation with a provable finite-sample guarantee for the suboptimality.

# B Details of the motivating example

For convenience, given a r.v. $X$, we denote its distributional robust value (refer Lemma 4.1) as $g(X, \rho) = \sup_{\beta \geq 0}\{-\beta \log(\mathbb{E}_{X\sim P}[e^{X/\beta}]) - \beta \cdot \rho\}$. For any action $a$, the corresponding reward distribution is

$$r_a \sim \begin{cases} \mathcal{N}(1,1) & \text{w.p.} \quad 1-a \\ \mathcal{N}(0,0.5) & \text{w.p.} \quad a \end{cases} \tag{7}$$

Based on the definition of $(s,a)$-rectangular, the robust action-value function $Q_{\mathrm{sa}}(a) = g(r_a, 1)$. The projected value function is approximated with a linear function $Q_{\mathrm{proj}}(a) = [1-a; a]^\top w$, where $w = \arg\min_u \mathbb{E}_{a\sim U[0,1]}(V_{\mathrm{sa}}(a) - [1-a;a]^T u)^2$. The $d$-rectangular robust action-value function is $Q_{\mathrm{d}}(a) = (1-a)g(r_0, 1) + ag(r_1, 1)$.

# C Algorithm Design for the PDRVI-L

**Algorithm 2** PDRVI-L

```
**Input:** $\underline{\beta}, \mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{N,H}$.
**Init:** $\widehat{V}_H = 0$.
**for** step $h = H$ **to** $1$ **do**
    $\Lambda_h \leftarrow \sum_{\tau=1}^N \phi\left(s_h^\tau, a_h^\tau\right)\phi\left(s_h^\tau, a_h^\tau\right)^\top + \lambda I$
    $\widehat{\theta}_h \leftarrow \Lambda_h^{-1}\left[\sum_{\tau=1}^N \phi(s_h^\tau, a_h^\tau) r_h^\tau\right]$
    **if** $h = H$ **then**
        $\widehat{w}_H \leftarrow 0$
    **else**
        Update $\widehat{w}_{h,i}$ with Equation 6.
    **end if**
    $\widehat{\nu}_h = \min(\widehat{\theta}_h + \widehat{w}_h, H - h + 1)_+$
    $\widehat{Q}_h(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^\top \widehat{\nu}_h - \gamma_h \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}$
    $\widehat{\pi}_h(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h} \langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\rangle_{\mathcal{A}}$
    $\widehat{V}_h(\cdot) \leftarrow \langle \widehat{Q}_h(\cdot,\cdot), \widehat{\pi}_h(\cdot \mid \cdot)\rangle_{\mathcal{A}}$
**end for**
```

# D Experiment Setup

## D.1 American Option Pricing

We set the feature dimension $d = 31$ and collect a dataset with 1000 trajectories as the offline dataset. We assume that the price follows Bernoulli distribution Cox et al. (1979),

$$s_{h+1} = \begin{cases} c_u s_h, & \text{w.p. } p_0, \\ c_d s_h, & \text{w.p. } 1 - p_0, \end{cases} \tag{8}$$

where the $c_u$ and $c_d$ are the price up and down factors and $p_0$ is the probability that the price goes up. The initial price $s_0$ is uniformly sampled from $[\kappa - \epsilon, \kappa + \epsilon]$, where $\kappa = 100$ is the strike price and $\epsilon = 5$ in our simulation. The agent can take an action to exercise the option ($a_h = 1$) or not exercise ($a_h = 0$) at the time step $h$. If exercising the option, the agent receives a reward $\max(0, \kappa - s_h)$ and the state transits into an exit state. Otherwise, the price will fluctuate based on the above model and no reward will be assigned. In our experiments, we set $H = 20$, $c_u = 1.02$, $c_d = 0.98$. We limit the price in $[80, 140]$ and discretize with the precision of 1 decimal place. Thus the state space size $|\mathcal{S}| = 602$.

Since $Q(s_h, a = 1) = \max(0, \kappa - s_h)$ is known in advance, we do not need to do any approximation for $a_h = 1$, and only need to estimate $\widehat{Q}(s_h, a = 0) = \phi(s_h)^\top \widehat{w}_h$. The features are chosen as $\phi(s_h) = [\varphi(s_h, s_1), \ldots, \varphi(s_h, s_d)]^\top$, where $s_1, \ldots, s_d$ are selected anchor states and $\varphi(s_h, s_i), \forall i \in [d]$ is the pairwise similarity measure. In particular, we set $s_1 = s_{\min} = 80, s_d = s_{\max=}140$, and $\Delta = s_{i+1} - s_i = (s_{\max} - s_{\min})/(d-1), \forall i \in [d-1]$. The similarity measure $\varphi(s_h, s_i) = \max(0, 1 - |s_h - s_i|/\Delta), \forall i \in [d]$, which is the partition to the nearest anchor states. Before training the agent, we collect data with a fixed behavior policy for $N$ trajectories. Since taking $a_h = 1$ will terminate the episode, it is helpless for learning the transition model. Hence, we use a fixed policy to collect data, which always chooses $a_h = 0$. All the experiments are finished on a server with an AMD EPYC 7702 64-Core Processor CPU.

### D.2 CARTPOLE

For our PDRVI-L algorithm, we construct the feature map $\phi(s, a) = (\phi_1, \phi_2, \cdots, \phi_d) \in \mathbb{R}^d$ using the gaussian kernel where for action $a \in \{0, 1\}$, the feature map with bandwidth $\sigma$ is defined as

$$\phi_{a\cdot d/2+i}(s, a; \sigma) = \exp\left(-\frac{\|s - s_i\|^2}{2\sigma^2}\right), \forall i \in [d/2]. \tag{9}$$

We generate the anchor states $s_i$ by uniformly sampling from the state space $\mathcal{S} = \mathbb{R}^4$ with each component sample from a uniform distribution $U[-1, 1]$. The feature dimension is $d = 512$.We choose LCB coefficient $\gamma = 0.03$ and the $\rho = 0.1$. Note that Cartpole is a inifinite-horizon sequential decision process and we set the discount factoris as $0.95$. We iterate the algorithm until the infinity norm of the difference between two consecutive value functions is less than $10^{-4}$.

Now we discuss the offline dataset used in the training of PDRVI-L and RFQI. The setup is the same as that in Panaganti et al. (2022a) for fair comparison. To be specific, we train proximal policy optimization (PPO) Schulman et al. (2017) algorithm using RL baseline zoo Raffin (2020) with default parameters. Then we generate the Cartpole dataset with $10^5$ samples using an $\varepsilon$-greedy version of the PPO trained policy with $\varepsilon = 0.3$. The preparation of the dataset can be finished by simply running the Github Repo of Panaganti et al. (2022a).

Next we introduce the details of the baseline algorithms. We directly adopt the implementation of RFQI from Panaganti et al. (2022a) and use the default parameters. For the implementation of PDRVI, please refer to Jin et al. (2021).

### D.3 REPRODUCTION OF RAPI

In this part, we introduce the implementation of RAPI Tamar et al. (2014) in our setting. Since the original RAPI focuses on the online setting with general uncertainty set, we instantiate RAPI with the episode MDP and KL-divergence as the uncertainty measure. Similar to our method, we incorporate Lemma 4.1 to robust problem for each $(s, a)$:

$$\sigma_{sa}(V) = \sup_{\beta \in [0,\infty)} \left\{-\beta \log\left(\mathbb{E}_{P(\cdot|s,a)} e^{-V(\cdot)/\beta}\right) - \rho\beta\right\}. \tag{10}$$

In the offline setting, the main challenge is to estimate the $\mathbb{E}_{P(\cdot|s,a)} e^{-V(\cdot)/\beta}$ from data. Since the ordinary least squares (OLS) has the close form solution, we can estimate Equation 10 with

$$\hat{\sigma}_{sa}(V) = \sup_{\beta \in [0,\infty)} \left\{-\beta \log\left(\phi(s, a)^\top \Lambda_h^{-1} \left(\sum_{\tau=1}^{K} \phi\left(s_h^\tau, a_h^\tau\right) \cdot \left(e^{-V(s_{h+1}^\tau)/\beta}\right)\right)\right) - \rho\beta\right\}. \tag{11}$$

Plugging it into the template of RAPI, we have the algorithm in Algorithm 3.

## E PROOF OF SECTION 4

Before we prove the main theorem, we first introduce the following lemma.

**Algorithm 3** RPVI with KL-divergence

```
Input: $\underline{\beta}, \mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{N,H}$.
Init: $\widehat{V}_H = 0$.
for step $h = H$ to $1$ do
    $\Lambda_h \leftarrow \sum_{\tau=1}^{N} \phi(s_h^\tau, a_h^\tau)\, \phi(s_h^\tau, a_h^\tau)^\top + \lambda I$
    if $h = H$ then
        $\widehat{w}_H \leftarrow \Lambda_H^{-1}\left[\sum_{\tau=1}^{N} \phi(s_h^\tau, a_h^\tau) r_h^\tau\right]$
    else
        $\widehat{w}_h \leftarrow \Lambda_h^{-1}\left[\sum_{\tau=1}^{N} \phi(s_h^\tau, a_h^\tau)\left(r_h^\tau + \hat{\sigma}_{sa}(\hat{V}_{h+1})\right)\right]$
    end if
    $\widehat{Q}_h(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^\top \widehat{w}_h$
    $\widehat{\pi}_h(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h} \langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\rangle_{\mathcal{A}}$
    $\widehat{V}_h(\cdot) \leftarrow \langle \widehat{Q}_h(\cdot,\cdot), \widehat{\pi}_h(\cdot \mid \cdot)\rangle_{\mathcal{A}}$
end for
```

**Lemma E.1** (Robust Extended Value Difference)**.** *Let $\pi = \{\pi_h\}_{h=1}^H$ and $\pi' = \{\pi'_h\}_{h=1}^H$ as two different policies. Denote $\mathcal{P}_1$ and $\mathcal{P}_2$ as two different ambiguity sets. Finally we use $V_{\mathcal{P}}^\pi = \{V_{h,\mathcal{P}}^\pi\}_{h\in[H]}$ denotes the value function using policy $\pi$ and the ambiguity set $\mathcal{P}$. For all $s \in \mathcal{S}$, we have*

$$
\begin{aligned}
V_{1,\mathcal{P}_1}^{\pi}(s) - V_{1,\mathcal{P}_2}^{\pi'}(s) = \sum_{h=1}^{H} & \mathbb{E}_{P_1^\star,\pi}[\iota_h(s_h,a_h)|s_1 = s] \\
& + \sum_{h=1}^{H} \mathbb{E}_{P_1^\star,\pi}[\langle Q_{h,\mathcal{P}_2}^{\pi'}(s,\cdot), \pi_h(\cdot|s_h) - \pi'(\cdot|s_h)\rangle_{\mathcal{A}}|s_1 = s],
\end{aligned}
$$

*where*

$$
\iota_h(s,a) = \inf_{P_1 \in \mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_1}[V_{h+1,\mathcal{P}_1}^{\pi}(s')] - \inf_{P_2 \in \mathcal{P}_{2,h}(s,a)} \mathbb{E}_{P_2}[V_{h+1,\mathcal{P}_2}^{\pi}(s')],
$$

*and $P_1^\star = \{P_{1,h}^\star\}_{h=1}^H$ for some $P_{1,h}^\star \in \mathcal{P}_{1,h}$.*

*Proof.*

$$
\begin{aligned}
V_{h,\mathcal{P}_1}^{\pi}(s) - V_{h,\mathcal{P}_2}^{\pi'}(s) &= \langle Q_{h,\mathcal{P}_1}^{\pi}(s,\cdot), \pi_h(\cdot|s_h)\rangle_{\mathcal{A}} - \langle Q_{h,\mathcal{P}_2}^{\pi'}(s,\cdot), \pi'_h(\cdot|s_h)\rangle_{\mathcal{A}} \\
&= \langle Q_{h,\mathcal{P}_1}^{\pi}(s,\cdot) - Q_{h,\mathcal{P}_2}^{\pi'}(s,\cdot), \pi_h(\cdot|s_h)\rangle_{\mathcal{A}} + \langle Q_{h,\mathcal{P}_2}^{\pi'}(s,\cdot), \pi_h(\cdot|s_h) - \pi'(\cdot|s_h)\rangle_{\mathcal{A}}.
\end{aligned} \tag{12}
$$

Note that

$$
Q_{h,\mathcal{P}_1}^{\pi}(s,a) - Q_{h,\mathcal{P}_2}^{\pi'}(s,a) = \inf_{P_{1,h} \in \mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_{1,h}}[V_{h+1,\mathcal{P}_1}^{\pi}(s')] - \inf_{P_{2,h} \in \mathcal{P}_{2,h}(s,a)} \mathbb{E}_{P_{2,h}}[V_{h+1,\mathcal{P}_2}^{\pi'}(s')]. \tag{13}
$$

We set

$$
\begin{aligned}
P_{1,h}^\star &= \arg\min_{P_{1,h} \in \mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_1}[V_{h+1,\mathcal{P}_1}^{\pi}(s')], \\
P_{2,h}^\star &= \arg\min_{P_{2,h} \in \mathcal{P}_{2,h}(s,a)} \mathbb{E}_{P_2}[V_{h+1,\mathcal{P}_2}^{\pi}(s')].
\end{aligned}
$$

Then

$$
\begin{aligned}
\text{Equation 13} &= \inf_{P_{1,h} \in \mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_{1,h}}[V_{h+1,\mathcal{P}_1}^{\pi}(s')] - \inf_{P_{2,h} \in \mathcal{P}_{2,h}(s,a)} \mathbb{E}_{P_2}[V_{h+1,\mathcal{P}_2}^{\pi'}(s')] \\
&= \inf_{P_{1,h} \in \mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_{1,h}}[V_{h+1,\mathcal{P}_1}^{\pi}(s')] - \inf_{P_{2,h} \in \mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_2}[V_{h+1,\mathcal{P}_2}^{\pi'}(s')] \\
&+ \inf_{P_{2,h} \in \mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_2}[V_{h+1,\mathcal{P}_2}^{\pi'}(s')] - \inf_{P_{2,h} \in \mathcal{P}_{2,h}(s,a)} \mathbb{E}_{P_2}[V_{h+1,\mathcal{P}_2}^{\pi'}(s')] \qquad (14) \\
&= \mathbb{E}_{P_{1,h}^\star}[V_{h+1,\mathcal{P}_1}^{\pi}(s') - V_{h+1,\mathcal{P}_2}^{\pi'}(s')] \\
&+ \inf_{P_{2,h} \in \mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_2}[V_{h+1,\mathcal{P}_2}^{\pi'}(s')] - \inf_{P_{2,h} \in \mathcal{P}_{2,h}(s,a)} \mathbb{E}_{P_2}[V_{h+1,\mathcal{P}_2}^{\pi'}(s')].
\end{aligned}
$$

Now we prove the last equality, i.e., there exists $P^\star_{1,h} \in \mathcal{P}_{1,h}(s,a)$ such that

$$\inf_{P_{1,h}\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_{1,h}}[V^{\pi}_{h+1,\mathcal{P}_1}(s')] - \inf_{P_{2,h}\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_2}[V^{\pi'}_{h+1,\mathcal{P}_2}(s')] = \mathbb{E}_{P^\star_{1,h}}[V^{\pi}_{h+1,\mathcal{P}_1}(s') - V^{\pi'}_{h+1,\mathcal{P}_2}(s')].$$

Define $h(P) = \mathbb{E}_P[V^{\pi}_{h+1,\mathcal{P}_1}(s') - V^{\pi'}_{h+1,\mathcal{P}_2}(s')]$. First we have

$$\begin{aligned}\inf_{P\in\mathcal{P}_{1,h}(s,a)} h(P) &= \inf_{P\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_P[V^{\pi}_{h+1,\mathcal{P}_1}(s') - V^{\pi'}_{h+1,\mathcal{P}_2}(s')] \\ &\leq \inf_{P\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_P[V^{\pi}_{h+1,\mathcal{P}_1}(s')] - \inf_{P\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_P[V^{\pi'}_{h+1,\mathcal{P}_2}(s')].\end{aligned}$$

On the other hand,

$$\begin{aligned}&\inf_{P_{1,h}\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_{1,h}}[V^{\pi}_{h+1,\mathcal{P}_1}(s')] - \inf_{P_{2,h}\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_2}[V^{\pi'}_{h+1,\mathcal{P}_2}(s')] \\ =& \inf_{P_{1,h}\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_{1,h}}[V^{\pi}_{h+1,\mathcal{P}_1}(s')] - \mathbb{E}_{P^\star_{2,h}}[V^{\pi'}_{h+1,\mathcal{P}_2}(s')] \\ \leq& \mathbb{E}_{P^\star_{2,h}}[V^{\pi}_{h+1,\mathcal{P}_1}(s')] - \mathbb{E}_{P^\star_{2,h}}[V^{\pi'}_{h+1,\mathcal{P}_2}(s')] \\ \leq& \sup_{P\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_P[V^{\pi}_{h+1,\mathcal{P}_1}(s') - V^{\pi'}_{h+1,\mathcal{P}_2}(s')].\end{aligned}$$

In summary,

$$\inf_{P\in\mathcal{P}_{1,h}(s,a)} h(P) \leq \inf_{P_{1,h}\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_{1,h}}[V^{\pi}_{h+1,\mathcal{P}_1}(s')] - \inf_{P_{2,h}\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_2}[V^{\pi'}_{h+1,\mathcal{P}_2}(s')] \leq \sup_{P\in\mathcal{P}_{1,h}(s,a)} h(P).$$

We know the ambiguity set constructed by the Cressie-Read divergence is a convex set, thus it is connected. Moreover, $h$ function is continuous with respect to the finite-dimension probability $P$. By the general intermediate value theorem in topological spaces, we know there exists $P^\star_{1,h} \in \mathcal{P}_{1,h}(s,a)$ such that

$$\inf_{P_{1,h}\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_{1,h}}[V^{\pi}_{h+1,\mathcal{P}_1}(s')] - \inf_{P_{2,h}\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_2}[V^{\pi'}_{h+1,\mathcal{P}_2}(s')] = \mathbb{E}_{P^\star_{1,h}}[V^{\pi}_{h+1,\mathcal{P}_1}(s') - V^{\pi'}_{h+1,\mathcal{P}_2}(s')].$$

Iteratively preceeding with Equation 12 and 14, combining with the initization that $V^{\pi}_{H+1,\mathcal{P}_1} = V^{\pi'}_{H+1,\mathcal{P}_2} = \mathbf{0}$, we have

$$\begin{aligned}V^{\pi}_{1,\mathcal{P}_1}(s) - V^{\pi'}_{1,\mathcal{P}_2}(s) = &\sum_{h=1}^{H} \mathbb{E}_{P^\star_1,\pi}[\iota_h(s_h,a_h)|s_1 = s] \\ &+ \sum_{h=1}^{H} \mathbb{E}_{P^\star_1,\pi}[\langle Q^{\pi'}_{h,\mathcal{P}_2}(s,\cdot), \pi_h(\cdot|s_h) - \pi'(\cdot|s_h)\rangle_{\mathcal{A}}|s_1 = s],\end{aligned}$$

where

$$\iota_h(s,a) = \inf_{P_{1,h}\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_{1,h}}[V^{\pi'}_{h,\mathcal{P}_2}(s')] - \inf_{P_{2,h}\in\mathcal{P}_{1,h}(s,a)} \mathbb{E}_{P_{2,h}}[V^{\pi'}_{h,\mathcal{P}_2}(s')].$$

□

**Lemma E.2** (Decomposition of Suboptimality (DRO version))**.**

$$\begin{aligned}\mathrm{SubOpt}(\widehat\pi;\mathcal{P}) = &\sum_{h=1}^{H} \mathbb{E}_{P^*,\pi^*}[\langle Q^{\widehat\pi}_{h,\widehat{\mathcal{P}}}(s_h,\cdot), \pi^*_h(\cdot|s_h) - \widehat\pi_h(\cdot|s_h)\rangle|s_1 \sim \mu] \\ + &\sum_{h=1}^{H} \mathbb{E}_{P^*,\pi^*}[\iota_h(s_h,a_h)|s_1 \sim \mu] - \sum_{h=1}^{H} \mathbb{E}_{P^\star,\widehat\pi}[\iota_h(s_h,a_h)|s_1 \sim \mu],\end{aligned}$$

*where*

$$\iota_h(s,a) := \mathbb{E}_{P^\star_h}[V^{\widehat\pi}_{h+1,\widehat{\mathcal{P}}}(s')] - \mathbb{E}_{\widehat{P}^\star_h}[V^{\widehat\pi}_{h+1,\widehat{\mathcal{P}}}(s')],$$

$$P^\star_h = \arg\min_{P_h\in\mathcal{P}_h(s,a)} \mathbb{E}_{P_h}[V^{\pi}_{h+1,\mathcal{P}}(s')], \quad \widehat{P}^\star_h = \arg\min_{\widehat{P}_h\in\widehat{\mathcal{P}}_h(s,a)} \mathbb{E}_{\widehat{P}_h}[V^{\pi}_{h+1,\widehat{\mathcal{P}}}(s')].$$

*Proof.* By the definition above, the suboptimality of the policy $\hat{\pi}$ can be decomposed as

$$\mathrm{SubOpt}(\widehat{\pi};\mathcal{P}) = \underbrace{(\mathbb{E}_{s\sim\mu}[V_1^*(s)] - \mathbb{E}_{s\sim\mu}[\widehat{V}_1(s)])}_{\text{I}} + \underbrace{(\mathbb{E}_{s\sim\mu}[\widehat{V}_1(s)] - \mathbb{E}_{s\sim\mu}[V_1^{\widehat{\pi}}(s)])}_{\text{II}}, \tag{15}$$

where $\{\widehat{V}_h\}_{h=1}^H$ are the estimated value functions constructed by any algorithm.

To clarify the proof, we write explicitly the dependence of the value function with respect to the policy, the ambiguity set: $\widehat{V}_1(s) = V_{1,\widehat{\mathcal{P}}}^{\widehat{\pi}}(s)$ and $V_1^*(s) = V_{1,\mathcal{P}}^{\pi^\star}(s)$. Correspondingly, we have $\widehat{Q}_1(s,a) = Q_{1,\widehat{\mathcal{P}}}^{\widehat{\pi}}(s,a)$ and $Q_1^*(s,a) = Q_{1,\mathcal{P}}^{\pi^\star}(s,a)$. Apply Lemma E.1 to the I term in equation 15 with $\pi = \pi^*$, $\pi' = \widehat{\pi}$ and $\mathcal{P}_1 = \mathcal{P}$ as the ambiguity set centered around the training transition model while $\mathcal{P}_2 = \widehat{\mathcal{P}}$ as the ambiguity set centered around the empirical model. Thus we have

$$\begin{aligned} V_{1,\mathcal{P}}^{\pi^\star}(s) - V_{1,\widehat{\mathcal{P}}}^{\widehat{\pi}}(s) = &\sum_{h=1}^H \mathbb{E}_{P^*,\pi^*}[\langle Q_{h,\widehat{\mathcal{P}}}^{\widehat{\pi}}(s_h,\cdot), \pi_h^*(\cdot|s_h) - \widehat{\pi}_h(\cdot|s_h)\rangle | s_1 = s] \\ &+ \sum_{h=1}^H \mathbb{E}_{P^*,\pi^*}[\iota_h(s_h,a_h)|s_1 = s], \end{aligned} \tag{16}$$

where

$$\iota_h(s_h,a_h) = \mathbb{E}_{P_h^\star}[V_{h+1,\widehat{\mathcal{P}}}^{\widehat{\pi}}(s')] - \mathbb{E}_{\widehat{P}_h^\star}[V_{h+1,\widehat{\mathcal{P}}}^{\widehat{\pi}}(s')].$$

Similarily, apply Lemma E.1 to the II term in equation 15, we have

$$\begin{aligned} V_{1,\widehat{\mathcal{P}}}^{\widehat{\pi}}(s) - V_{1,\mathcal{P}}^{\widehat{\pi}}(s) &= -(V_{1,\mathcal{P}}^{\widehat{\pi}}(s) - V_{1,\widehat{\mathcal{P}}}^{\widehat{\pi}}(s)) \\ &= -\sum_{h=1}^H \mathbb{E}_{P^\star,\widehat{\pi}}[\iota_h(s_h,a_h)|s_1 = s]. \end{aligned} \tag{17}$$

Putting equation 16 and equation 17 into equation 15 we yield the desired conclusions. □

*Proof of Lemma 4.2 .* Recall the Bellman equation for the $d$-rectangular robust MDP (Equation 4):

$$\begin{aligned} (\mathbb{B}_h V)(s,a) &= r(s,a) + \inf_{P_h \in \mathcal{P}^{\mathrm{KL}}(\psi_h;\rho)} \mathbb{E}_{s'\sim P_h(\cdot|s,a)}[V(s')] \\ &= \sum_{i\in[d]} \phi_i(s,a)\theta_{h,i} + \sum_{i\in[d]} \phi_i(s,a) \min_{\psi'_{h,i}\in\mathcal{P}^{\mathrm{KL}}(\psi_{h,i};\rho)} \psi'^{\top}_{h,i} V \end{aligned} \tag{18}$$

Since $M \in \mathcal{M}^{\mathrm{rob}}$, from the proof of above that for any $f \in \mathcal{F}$, we have $\mathbb{B}_h f \in \mathcal{F}$ for any $h \in [H]$, which finish the second part of the proof of lemma 4.2. □

**Lemma E.3.** *For any fix $h \in [H]$ and $i \in [d]$, we denote*

$$\beta_{h,i}^* \in \operatorname*{arg\,max}_{\beta_{h,i}\geq 0} \{-\beta_{h,i} \cdot \mathbb{E}_{\psi_{h,i}}[e^{-V_{h+1}(s')/\beta_{h,i}}] - \beta_{h,i}\rho\}.$$

*Then $\beta_{h,i}^* \leq \overline{\beta} := \frac{H-h+1}{\rho}$.*

*Proof.* This proof is by invoking the part 2 in the Lemma 4 in Zhou et al. (2021b) with $M = H$. □

Thus in the following, we consider the variant of the dual form of the KL optimization,

$$\sup_{\beta_{h,i}\in[0,\overline{\beta}]} \{-\beta_{h,i} \cdot \mathbb{E}_{\psi_{h,i}}[e^{-V_{h+1}(s')/\beta_{h,i}}] - \beta_{h,i}\rho\}.$$

**Algorithm 4** PDRVI-L-META

```
Input: $\underline{\beta}, \mathcal{D} = \{(s_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{N,H}$.
Init: $\widehat{V}_H = 0$.
for step $h = H$ to $1$ do
    $\Lambda_h \leftarrow \sum_{\tau=1}^N \phi\left(s_h^\tau, a_h^\tau\right) \phi\left(s_h^\tau, a_h^\tau\right)^\top + \lambda I$
    $\widehat{\theta}_h \leftarrow \Lambda_h^{-1}\left[\sum_{\tau=1}^N \phi(s_h^\tau, a_h^\tau) r_h^\tau\right]$
    if $h = H$ then
        $\widehat{w}_H \leftarrow 0$
    else
        $\widehat{\psi}_h\left(s'\right) \leftarrow \Lambda_h^{-1}\left[\sum_{\tau=1}^N \phi\left(s_h^\tau, a_h^\tau\right) \mathbf{1}\left(s_{h+1}^\tau = s'\right)^\top\right]$, for $s' \in \mathcal{S}$
        $\bar{w}_{h,i} \leftarrow \widehat{\theta}_{h,i} + \sup_{\beta\in[0,\infty)}\left\{-\beta\log\left(\widehat{\psi}_{h,i}^\top\left(e^{-\widehat{V}_{h+1}/\beta} - \mathbf{1}\right) + 1\right) - \rho\beta\right\}$
    end if
    $\widehat{\nu}_h = \min(\widehat{\theta}_h + \widehat{w}_h, H - h + 1)_+$
    $\widehat{Q}_h(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^\top \widehat{\nu}_h - \gamma_h \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}$
    $\widehat{\pi}_h(\cdot \mid \cdot) \leftarrow \arg\max_{\pi_h} \langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot \mid \cdot)\rangle_{\mathcal{A}}$
    $\widehat{V}_h(\cdot) \leftarrow \langle \widehat{Q}_h(\cdot,\cdot), \widehat{\pi}_h(\cdot \mid \cdot)\rangle_{\mathcal{A}}$
end for
```

# F Proof of Theorem 5.2

In the following, we rewrite our algorithms into the form of PDRVI-L-META. The proposed PDRVI-L-META algorithm is to facilitate the presentation of our proof. In specific, we introduce $\widehat{\psi}_h$ as the estimator for $\psi_h$ in the update of $\widehat{w}_h$. We ignore the rewritten for Algorithm 1 as it is similar.

Before we start the proof, it is obvious to note that $\sum_{i=1}^d \phi_i(s,a) = 1, \forall (s,a) \in \mathcal{S} \times \mathcal{A}$ under the Assumption 4.1. In this section, we mainly prove the Theorem 5.2. By setting the model mis-specification $\xi = 0$, we can recover the results in Theorem 4.1.

**Proposition F.1.** *With probability at least* $1 - \delta$*, for all* $h \in [H]$*, we have*

$$\begin{aligned}\iota_h(s,a) \leq & \underline{\beta}(e^{H/\underline{\beta}} - 1)(2\xi\sqrt{d} + 10\sqrt{d\zeta_1}) \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \\ & + 2\sqrt{2}\sqrt{\underline{\beta}}(e^{H/\underline{\beta}} - 1)\sqrt{H\zeta_2} \sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} + (H-h)\xi,\end{aligned}$$

*for* $\zeta_1 = \log(2N + 16Nd^{3/2}H^2e^{H/\underline{\beta}})$ *and* $\zeta_2 = \log(\frac{2dNH^3}{\delta\rho})$.

*Proof.* From the DRO Bellman optimality equation and Lemma 4.1, we denote

$$\begin{aligned}(\mathbb{B}_h\widehat{V}_{h+1})(s,a) &= r_h(s,a) + \inf_{P_{h+1}\in\mathcal{P}_{h+1}} E_{P_{h+1}(\cdot|s,a)}[\widehat{V}_{h+1}(s')] \\ &= r_h(s,a) + \inf_{\tilde{P}_{h+1}\in\tilde{\mathcal{P}}_{h+1}} E_{\tilde{P}_{h+1}(\cdot|s,a)}[\widehat{V}_{h+1}(s')] \\ &\quad + (\inf_{P_{h+1}\in\mathcal{P}_{h+1}} E_{P_{h+1}(\cdot|s,a)}[\widehat{V}_{h+1}(s')] - \inf_{\tilde{P}_{h+1}\in\tilde{\mathcal{P}}_{h+1}} E_{\tilde{P}_{h+1}(\cdot|s,a)}[\widehat{V}_{h+1}(s')]) \\ &= \sum_{i=1}^d \phi_i(s,a)\theta_{h,i} + (H-h)\xi \\ &\quad + \sum_{i=1}^d \phi_i(s,a) \max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\{-\beta_{h,i}\cdot\log(\mathbb{E}_{\psi_{h,i}}[e^{-\widehat{V}_{h+1}(s')/\beta_{h,i}}]) - \beta_{h,i}\rho\}.\end{aligned}$$

Combined with the empirical Bellman operator in our algorithm 1,

$$\begin{aligned}\iota_h(s,a) &= (\mathbb{B}_h \widehat{V}_{h+1})(s,a) - (\widehat{\mathbb{B}}_h \widehat{V}_{h+1})(s,a) \\ &= \phi(s,a)^\top(\theta_h - \widehat{\theta}_h) + \phi(s,a)^\top(w_h - \widetilde{w}_h) + (H-h)\xi,\end{aligned} \tag{19}$$

where $w_{h,i} \coloneqq \max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\{-\beta_{h,i}\log(\mathbb{E}_{\psi_{h,i}}[e^{-\widehat{V}_{h+1}(s')/\beta_{h,i}}]) - \beta_{h,i}\rho\}$ and $\widetilde{w}_{h,i} \coloneqq \max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\{-\beta_{h,i}\log(\widehat{\psi}_{h,i}^\top[e^{-\widehat{V}_{h+1}(s')/\beta_{h,i}} - 1] + 1) - \beta_{h,i}\rho\}$.

**Step 1:** we analyze the error in the reward estimation, i.e., $\phi(s,a)^\top(\theta_h - \widehat{\theta}_h)$.

$$\begin{aligned}\phi(s,a)^\top(\theta_h - \widehat{\theta}_h) &= \phi(s,a)^\top\Lambda_h^{-1}\Lambda_h\theta_h - \phi(s,a)^\top\Lambda_h^{-1}[\sum_{\tau=1}^N \phi(s_h^\tau,a_h^\tau)r_h(s_h^\tau,a_h^\tau)] \\ &= \phi(s,a)^\top\Lambda_h^{-1}\Lambda_h\theta_h - \phi(s,a)^\top\Lambda_h^{-1}[\sum_{\tau=1}^N \phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top\theta_h] \\ &= \phi(s,a)^\top\Lambda_h^{-1}\Lambda_h\theta_h - \phi(s,a)^\top\Lambda_h^{-1}(\Lambda_h - \lambda I)\theta_h \\ &= \lambda\phi(s,a)^\top\Lambda_h^{-1}\theta_h \\ &\le \lambda\|\theta_h\|_{\Lambda_h^{-1}}\|\phi(s,a)\|_{\Lambda_h^{-1}} \\ &\le \sqrt{d\lambda}\|\phi(s,a)\|_{\Lambda_h^{-1}} \\ &\le \sqrt{d\lambda}\sum_{i=1}^d \|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}.\end{aligned} \tag{20}$$

Here the last inequality is from

$$\|\theta_h\|_{\Lambda_h^{-1}} = \sqrt{\theta_h^\top\Lambda_h^{-1}\theta_h} \le \|\Lambda_h^{-1}\|^{1/2}\|\theta_h\| \le \sqrt{d/\lambda},$$

by using the fact that $\|\Lambda_h^{-1}\| \le \lambda^{-1}$ and the Definition 2.1.

**Step 2:** we turn to the estimation error from the transition model, i.e., $\phi(s,a)^\top(w_h - \widehat{w}_h)$. We define two auxiliary functions:

$$\hat{g}_{h,i}(\beta) \coloneqq -\beta\cdot\log(\widehat{\psi}_{h,i}^\top[e^{-\widehat{V}_{h+1}(s')/\beta} - 1] + 1) - \beta\rho,$$

and

$$g_{h,i}(\beta) \coloneqq -\beta\cdot\log(\mathbb{E}_{\psi_{h,i}}[e^{-\widehat{V}_{h+1}(s')/\beta}]) - \beta\rho.$$

Then

$$\begin{aligned}&|\sum_{i\in[d]}\phi_i(s,a)(w_{h,i} - \widetilde{w}_{h,i})| \\ &\le \sum_{i\in[d]}|\phi_i(s,a)(w_{h,i} - \widetilde{w}_{h,i})| \\ &= \sum_{i\in[d]}\phi_i(s,a)|\max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]} g_{h,i}(\beta_{h,i}) - \max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\hat{g}_{h,i}(\beta_{h,i})| \\ &\le \sum_{i\in[d]}\phi_i(s,a)\max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}|g_{h,i}(\beta_{h,i}) - \hat{g}_{h,i}(\beta_{h,i})| \\ &= \sum_{i\in[d]}\phi_i(s,a)\max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\{|\beta_{h,i}\cdot(\log(\widehat{\psi}_{h,i}^\top[(e^{-\widehat{V}_{h+1}(s')/\beta_{h,i}} - 1)] + 1) - \log(\mathbb{E}_{\psi_{h,i}}[e^{-\widehat{V}_{h+1}(s')/\beta_{h,i}}]))|\} \\ &= \sum_{i\in[d]}\phi_i(s,a)\max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\{|\beta_{h,i}\cdot(\log(\widehat{\psi}_{h,i}^\top[e^{(H-\widehat{V}_{h+1}(s'))/\beta_{h,i}} - e^{H/\beta_{h,i}}] + e^{H/\beta_{h,i}}) - \log(\mathbb{E}_{\psi_{h,i}}[e^{(H-\widehat{V}_{h+1}(s'))/\beta_{h,i}}]))|\} \\ &\le \sum_{i\in[d]}\phi_i(s,a)\max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\{|\beta_{h,i}\cdot(\widehat{\psi}_{h,i}^\top[e^{(H-\widehat{V}_{h+1}(s'))/\beta_{h,i}} - e^{H/\beta_{h,i}}] + e^{H/\beta_{h,i}} - \mathbb{E}_{\psi_{h,i}}[e^{(H-\widehat{V}_{h+1}(s'))/\beta_{h,i}}])|\},\end{aligned}$$

$$\tag{22}$$

where the last inequality follows from the fact I.2 by setting $x = \beta_{h,i} \log(\widehat{\psi}_{h,i}^\top[e^{(H-\widehat{V}_{h+1}(s))/\beta_{h,i}} - e^{H/\beta_{h,i}}] + e^{H/\beta_{h,i}})$ and $y = \beta_{h,i} \log \mathbb{E}_{\psi_{h,i}}[e^{(H-\widehat{V}_{h+1}(s))/\beta_{h,i}}]$. To ease the presentation, we index the finite states from 1 to $S$ and introduce the vector $\widehat{V}_{h+1} \in \mathbb{R}^S$ where $[\widehat{V}_{h+1}]_j = \widehat{V}_{h,i}(s_j)$ and $[H - \widehat{V}_{h+1}]_j = H - \widehat{V}_{h+1}(s_j)$.

We denote $\mathbb{1}_j \in \mathbb{R}^{d\times 1}$ with the $j$-the component as 1 and the other components are 0 and $\mathbf{1} \in \mathbb{R}^S$ is a all-one vector. Notice that

$$\begin{aligned}
&\mathbb{E}_{\psi_{h,i}}[e^{(H-\widehat{V}_{h+1}(s'))/\beta_{h,i}}] \\
&= \mathbb{E}_{\psi_{h,i}}[e^{(H-\widehat{V}_{h+1}(s'))/\beta_{h,i}} - e^{H/\beta_{h,i}}] + e^{H/\beta_{h,i}} \\
&= \psi_{h,i}^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}} - e^{H/\beta_{h,i}}) + e^{H/\beta_{h,i}} \\
&= \mathbb{1}_i^\top \psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}} - e^{H/\beta_{h,i}}) + e^{H/\beta_{h,i}} \\
&= \mathbb{1}_i^\top \Lambda_h^{-1}\Lambda_h \psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}} - e^{H/\beta_{h,i}}) + e^{H/\beta_{h,i}} \\
&= \mathbb{1}_i^\top \Lambda_h^{-1}(\sum_{\tau=1}^N \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda I)\psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}} - e^{H/\beta_{h,i}}) + e^{H/\beta_{h,i}} \\
&= \mathbb{1}_i^\top \Lambda_h^{-1}\sum_{\tau=1}^N \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top \psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}} - e^{H/\beta_{h,i}}) \\
&\quad + \lambda \mathbb{1}_i^\top \Lambda_h^{-1}\psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}} - e^{H/\beta_{h,i}}) + e^{H/\beta_{h,i}} \\
&= \mathbb{1}_i^\top \Lambda_h^{-1}\sum_{\tau=1}^N \phi(s_h^\tau, a_h^\tau)\tilde{P}_{h+1}(\cdot|s_h^\tau, a_h^\tau)^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}} - e^{H/\beta_{h,i}}) \\
&\quad + \lambda \mathbb{1}_i^\top \Lambda_h^{-1}\psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}} - e^{H/\beta_{h,i}}) + e^{H/\beta_{h,i}},
\end{aligned}$$

and

$$\begin{aligned}
&\widehat{\psi}_{h,i}^\top[(e^{(H-\widehat{V}_{h+1}(s'))/\beta_{h,i}} - e^{H/\beta_{h,i}})] + e^{H/\beta_{h,i}} \\
&= \widehat{\psi}_{h,i}^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}} - e^{H/\beta_{h,i}}) + e^{H/\beta_{h,i}} \\
&= \mathbb{1}_i^\top \Lambda_h^{-1}\sum_{\tau=1}^N \phi(s_h^\tau, a_h^\tau)\mathbb{1}(s_{h+1}^\tau)^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}} - e^{H/\beta_{h,i}}) + e^{H/\beta_{h,i}},
\end{aligned}$$

where $\mathbb{1}(s_{h+1}^\tau) \in \mathbb{R}^{d\times 1}$ with the correponding component for the $s_{h+1}^\tau$ being 1 and the other being 0 and $\psi_h = [\psi_{h,1}, \psi_{h,2}, \cdots, \psi_{h,d}] \in \mathbb{R}^{d\times S}$. Now we are ready for controlling the error in Equation 21. In particular, we aim to control the error within the maximum of Equation 21 for any given $\beta_{h,i} \in [\underline{\beta}, \overline{\beta}]$, that is

$$\left|\beta_{h,i}\left(\widehat{\psi}_{h,i}^\top[(e^{(H-\widehat{V}_{h+1}(s'))/\beta_{h,i}} - e^{H/\beta_{h,i}})] + e^{H/\beta_{h,i}} - \mathbb{E}_{\psi_{h,i}}[e^{(H-\widehat{V}_{h+1}(s'))/\beta_{h,i}}]\right)\right|.$$

By using the above decomposition, we have

$$\begin{aligned}
&\left|\beta_{h,i}\left(\widehat{\psi}_{h,i}^\top[(e^{(H-\widehat{V}_{h+1}(s'))/\beta_{h,i}} - e^{H/\beta_{h,i}})] + e^{H/\beta_{h,i}} - \mathbb{E}_{\psi_{h,i}}[e^{(H-\widehat{V}_{h+1}(s'))/\beta_{h,i}}]\right)\right| \\
&= \left|\beta_{h,i}\mathbb{1}_i^\top \Lambda_h^{-1}\left(\sum_{\tau=1}^N \phi(s_h^\tau, a_h^\tau)\cdot(\mathbb{1}(s_{h+1}^\tau) - \tilde{P}_{h+1}(\cdot|s_h^\tau, a_h^\tau))^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}} - e^{H/\beta_{h,i}})\right)\right| \\
&\quad + \left|\lambda\beta_{h,i}\mathbb{1}_i^\top \Lambda_h^{-1}\psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}} - e^{H/\beta_{h,i}})\right|.
\end{aligned}$$

We further decompose the difference,

$$\begin{aligned}
&\left|\beta_{h,i}\mathbb{1}_i^\top\Lambda_h^{-1}\left(\sum_{\tau=1}^N\phi(s_h^\tau,a_h^\tau)\cdot(\mathbb{1}(s_{h+1}^\tau)-\tilde{P}_{h+1}(\cdot|s_h^\tau,a_h^\tau))^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}}-e^{H/\beta_{h,i}})\right)\right| \\
&\quad+\left|\lambda\beta_{h,i}\mathbb{1}_i^\top\Lambda_h^{-1}\psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}}-e^{H/\beta_{h,i}})\right| \\
&=\left|\beta_{h,i}\mathbb{1}_i^\top\Lambda_h^{-1}\left(\sum_{\tau=1}^N\phi(s_h^\tau,a_h^\tau)\cdot(P_{h+1}(\cdot|s_h^\tau,a_h^\tau)-\tilde{P}_{h+1}(\cdot|s_h^\tau,a_h^\tau))^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}}-e^{H/\beta_{h,i}})\right)\right| \\
&\quad+\left|\beta_{h,i}\mathbb{1}_i^\top\Lambda_h^{-1}\left(\sum_{\tau=1}^N\phi(s_h^\tau,a_h^\tau)\cdot(\mathbb{1}(s_{h+1}^\tau)-P_{h+1}(\cdot|s_h^\tau,a_h^\tau))^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}}-e^{H/\beta_{h,i}})\right)\right| \\
&\quad+\left|\lambda\beta_{h,i}\mathbb{1}_i^\top\Lambda_h^{-1}\psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}}-e^{H/\beta_{h,i}})\right| \\
&\le\left|\xi\beta_{h,i}(e^{H/\beta_{h,i}}-1)\cdot\mathbb{1}_i^\top\Lambda_h^{-1}\sum_{\tau=1}^N\phi(s_h^\tau,a_h^\tau)+\beta_{h,i}\mathbb{1}_i^\top\Lambda_h^{-1}\left(\sum_{\tau=1}^N\phi(s_h^\tau,a_h^\tau)\cdot\epsilon_h^\tau(\beta_{h,i},\widehat{V}_{h+1})\right)\right| \\
&\quad+\left|\lambda\beta_{h,i}\mathbb{1}_i^\top\Lambda_h^{-1}\psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}}-e^{H/\beta_{h,i}})\right|,
\end{aligned}$$

where $\epsilon_h^\tau(\beta,V):=(P_{h+1}(\cdot|s_h^\tau,a_h^\tau)-\mathbb{1}(s_{h+1}^\tau))^\top(e^{(H-V)/\beta_{h,i}}-e^{H/\beta_{h,i}})$ and $|e^{(H-V)/\beta_{h,i}}-e^{H/\beta_{h,i}}|\le(e^{H/\beta_{h,i}}-1)$.

Plug into Equation 22 we have

$$\begin{aligned}
&|\phi(s,a)^\top(w_h-\widetilde{w}_h)| \\
&=|\sum_{i=1}^d\phi_i(s,a)(w_{h,i}-\widetilde{w}_{h,i})| \\
&\le\underbrace{\xi\sum_{i=1}^d\max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\phi_i(s,a)|\beta_{h,i}(e^{H/\beta_{h,i}}-1)\cdot\mathbb{1}_i^\top\Lambda_h^{-1}\sum_{\tau=1}^N\phi(s_h^\tau,a_h^\tau)|}_{\mathrm{I}} \\
&\quad+\underbrace{\sum_{i=1}^d\phi_i(s,a)|\mathbb{1}_i^\top\Lambda_h^{-1}(\sum_{\tau=1}^N\phi(s_h^\tau,a_h^\tau)\cdot\max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\beta_{h,i}\epsilon_h^\tau(\beta_{h,i},\widehat{V}_{h+1}))|}_{\mathrm{II}} \\
&\quad+\underbrace{\sum_{i=1}^d\max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\phi_i(s,a)|\lambda\beta_{h,i}\mathbb{1}_i^\top\Lambda_h^{-1}\psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}}-e^{H/\beta_{h,i}}))|}_{\mathrm{III}}.
\end{aligned} \tag{23}$$

For I term in 23, for any $\beta_{h,i}\in[\underline{\beta},\overline{\beta}]$ we know,

$$\begin{aligned}
&\xi\sum_{i=1}^d|\beta_{h,i}(e^{H/\beta_{h,i}}-1)\cdot\phi_i(s,a)\mathbb{1}_i^\top\Lambda_h^{-1}\sum_{\tau=1}^N\phi(s_h^\tau,a_h^\tau)| \\
&\le\xi\sum_{i=1}^d\|\beta_{h,i}(e^{H/\beta_{h,i}}-1)\phi_i(s,a)\cdot\mathbb{1}_i\|_{\Lambda_h^{-1}}\|\sum_{\tau=1}^N\phi(s_h^\tau,a_h^\tau)\|_{\Lambda_h^{-1}}.
\end{aligned}$$

Note that

$$\sum_{i=1}^d\|\beta_{h,i}(e^{H/\beta_{h,i}}-1)\phi_i(s,a)\cdot\mathbb{1}_i\|_{\Lambda_h^{-1}}\le\underline{\beta}(e^{H/\underline{\beta}}-1)\sum_{i=1}^d\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}.$$

Then we turn to control $\|\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau)\|_{\Lambda_h^{-1}}$ as follow,

$$\begin{aligned}
\|\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau)\|_{\Lambda_h^{-1}} &= \sqrt{(\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau))^\top\Lambda_h^{-1}(\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau))} \\
&= \sqrt{\mathrm{Tr}(\Lambda_h^{-1}(\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau))(\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau))^\top)} \\
&= \sqrt{\mathrm{Tr}(\Lambda_h^{-1}(\Lambda_h-\lambda\cdot I))} \\
&\leq \sqrt{\mathrm{Tr}(\Lambda_h^{-1}\Lambda_h)} \\
&= \sqrt{d}.
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{I} =&\xi\sum_{i=1}^{d}\max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\phi_i(s,a)|\beta_{h,i}(e^{H/\beta_{h,i}}-1)\cdot\mathbb{1}_i^\top\Lambda_h^{-1}\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau)| \\
\leq&\xi\sqrt{d}\underline{\beta}(e^{H/\underline{\beta}}-1)\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}.
\end{aligned}$$

For term III and any $\beta_{h,i}\in[\underline{\beta},\overline{\beta}]$,

$$\begin{aligned}
&\sum_{i=1}^{d}|\lambda\beta_{h,i}\phi_i(s,a)\mathbb{1}_i^\top\Lambda_h^{-1}\psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}}-e^{H/\beta_{h,i}})| \\
&\leq\sum_{i=1}^{d}\lambda\|\phi_i(s,a)\mathbb{1}_i^\top\Lambda_h^{-1}\|_1\|\beta_{h,i}\psi_h^\top(e^{(H-\widehat{V}_{h+1})/\beta_{h,i}}-e^{H/\beta_{h,i}})\|_\infty \\
&\leq\lambda\underline{\beta}(e^{H/\underline{\beta}}-1)\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i^\top\Lambda_h^{-1}\|_1 \\
&\leq\sqrt{d}\lambda\underline{\beta}(e^{H/\underline{\beta}}-1)\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i^\top\Lambda_h^{-1}\|_2 \\
&\leq\sqrt{d}\lambda\underline{\beta}(e^{H/\underline{\beta}}-1)\|\Lambda_h^{-1/2}\|_2\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \\
&\leq\sqrt{d\lambda}\underline{\beta}(e^{H/\underline{\beta}}-1)\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}},
\end{aligned}$$

the second inequality is from $\|e^{(H-\widehat{V}_{h+1})/\beta_{h,i}}-e^{H/\beta_{h,i}}\|_\infty\leq(e^{H/\underline{\beta}}-1)$ and the three inequality is from $\|x\|_1\leq\sqrt{d}\|x\|_2$ for any $x\in\mathbb{R}^d$.

To control II term, we invoke Lemma G.1 with the choice of $\lambda = 1$ and then with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{i=1}^{d} |\phi_i(s,a)\mathbb{1}_i^\top \Lambda_h^{-1}(\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau)\cdot \max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\beta_{h,i}\epsilon_h^\tau(\beta_{h,i},\widehat{V}_{h+1}))| \\
&\leq \sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}\|\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau)\cdot \max_{\beta_{h,i}\in[\underline{\beta},\overline{\beta}]}\beta_{h,i}\epsilon_h^\tau(\beta_{h,i},\widehat{V}_{h+1})\|_{\Lambda_h^{-1}} \\
&\leq (4\sqrt{d}\underline{\beta}(e^{H/\underline{\beta}}-1)\sqrt{\log(2N+16Nd^{3/2}H^2e^{H/\underline{\beta}})} \\
&\quad + 2\sqrt{2}\underline{\beta}(e^{H/\underline{\beta}}-1)(\sqrt{\frac{H}{\underline{\beta}}\cdot\log(\frac{2dNH^3}{\delta\rho})}+\sqrt{2}))\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \\
&\leq (e^{H/\underline{\beta}}-1)(8\underline{\beta}\sqrt{d\zeta_1}+2\sqrt{2}\sqrt{\underline{\beta}H\zeta_2})\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}},
\end{aligned}$$

where the first inequality is from $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$ for $x,y\geq 0$. The second inequality is by $\sqrt{d\zeta_1}\geq 1$ where $\zeta_1=\log(2N+16Nd^{3/2}H^2e^{H/\underline{\beta}})$ and $\zeta_2=\log(\frac{2dNH^3}{\delta\rho})$. Thus we have

$$\begin{aligned}
&|\phi(s,a)^\top(w_h-\widetilde{w}_h)| \\
&\leq \underline{\beta}(e^{H/\underline{\beta}}-1)(\xi\sqrt{d}+8\cdot\sqrt{d\zeta_1}+\sqrt{d})\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}+2\sqrt{2}\sqrt{\underline{\beta}}(e^{H/\underline{\beta}}-1)\sqrt{H\zeta_2}\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \\
&\quad +\xi\sqrt{d}\underline{\beta}(e^{H/\underline{\beta}}-1)\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \\
&\leq \underline{\beta}(e^{H/\underline{\beta}}-1)(\xi\sqrt{d}+9\cdot\sqrt{d\zeta_1})\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}+2\sqrt{2}\sqrt{\underline{\beta}}(e^{H/\underline{\beta}}-1)\sqrt{H\zeta_2}\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \\
&\quad +\xi\sqrt{d}\underline{\beta}(e^{H/\underline{\beta}}-1)\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}},
\end{aligned} \tag{24}$$

where the second inequality is by noticing that $f(\beta)=\beta(e^{H/\beta}-1)$ and $g(\beta)=\sqrt{\beta}(e^{H/\beta}-1)$ are both monotonically decreasing with $\beta>0$ and the last inequality is from $\sqrt{d\zeta_1}\geq 1$.

Plug Equation 20 and 24 into Equation 19 to finally upper bound the Bellman error,

$$\begin{aligned}
&|\iota_h(s,a)| \\
&=|(\mathbb{B}_h\widehat{V}_{h+1})(s,a)-(\hat{\mathbb{B}}_h\widehat{V}_{h+1})(s,a)| \\
&\leq\sqrt{d}\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}+\underline{\beta}(e^{H/\underline{\beta}}-1)(\xi\sqrt{d}+9\cdot\sqrt{d\zeta_1})\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \\
&\quad +2\sqrt{2}\sqrt{\underline{\beta}}(e^{H/\underline{\beta}}-1)\sqrt{H\zeta_2}\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}+\xi\sqrt{d}\underline{\beta}(e^{H/\underline{\beta}}-1)\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}+(H-h)\xi \\
&\leq\underline{\beta}(e^{H/\underline{\beta}}-1)(2\xi\sqrt{d}+10\sqrt{d\zeta_1})\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}+2\sqrt{2}\sqrt{\underline{\beta}}(e^{H/\underline{\beta}}-1)\sqrt{H\zeta_2}\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \\
&\quad +(H-h)\xi,
\end{aligned}$$

where the last inequality is from the fact that $f(\beta)=\beta(e^{H/\beta}-1)\geq H\geq 1$ and and $\sqrt{d\zeta_1}>1$. □

*Proof of Theorem 5.2.* Our policy is the greedy policy w.r.t. to the estimated Q-function, thus we can reduce the suboptimality reduction in Lemma E.2 into

$$\operatorname{SubOpt}(\widehat{\pi};\mathcal{P}) \leq \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\iota_h\left(s_h, a_h\right) \mid s_1 \sim \mu\right] - \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\iota_h\left(s_h, a_h\right) \mid s_1 \sim \mu\right]. \tag{25}$$

Putting Proposition F.1 in the Equation 25 we know that with probability at least $1-\delta/2$,

$$\begin{aligned}
&\operatorname{SubOpt}(\widehat{\pi};\mathcal{P}) \\
&\leq (e^{H/\underline{\beta}}-1)(2\xi\sqrt{d}\underline{\beta} + 10\cdot\sqrt{d\zeta_1}\underline{\beta} + 2\sqrt{2}\sqrt{\underline{\beta}}\sqrt{H\zeta_2})\cdot\sum_{h=1}^{H}(\mathbb{E}_{\widehat{\pi}}[\|\Lambda_h^{-1}\|_{\operatorname{tr}(\phi(s,a))}] + \mathbb{E}_{\pi^*}[\|\Lambda_h^{-1}\|_{\operatorname{tr}(\phi(s,a))}]) + \cdots \\
&\quad + (H-1)H\xi/2.
\end{aligned}$$

Based on the Assumption 5.1, Definition 2.1 and using the similar steps in the proof of Corollar 4.6 in Jin et al. (2021), we can conclude that when $N$ is sufficiently large so that $N \geq 40/\underline{c}\cdot\log(4dH/\delta)$, for all $h\in[H]$, it holds that with probability at least $1-\delta/2$,

$$\begin{aligned}
\|\Lambda_h^{-1}\|_{\operatorname{tr}(\phi(s,a))} &= \sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}} \\
&\leq \sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|\|\Lambda_h^{-1}\|^{-1/2} \\
&\leq \sqrt{\frac{2}{\underline{c}N}} \coloneqq c/\sqrt{N}, \quad \forall (s,a)\in\mathcal{S}\times\mathcal{A}, \forall h\in[H],
\end{aligned}$$

where the second inequality is from Assumption 4.1. In conclusions, when $N \geq 40/\underline{c}\cdot\log(4dH/\delta)$, we have probability at least $1-\delta$,

$$\begin{aligned}
\operatorname{SubOpt}(\widehat{\pi};\mathcal{P}) \leq{}& c_1 H\underline{\beta}(e^{H/\underline{\beta}}-1)(\xi\sqrt{d}+\sqrt{d\zeta_1})/N^{1/2} + c_2\sqrt{\underline{\beta}}(e^{H/\underline{\beta}}-1)\sqrt{\zeta_2}H^{3/2}/N^{1/2} \\
&+ (H-1)H\xi,
\end{aligned}$$

for some absolute constants $c_1$ and $c_2$ that only depend on $\underline{c}$. □

## G AUXILIARY LEMMAS FOR THE PROOF FOR THEOREM 5.2

**Lemma G.1.** *For all $i\in[d]$, $\beta_{h,i}\in[\underline{\beta},\overline{\beta}]$, and the estimator $\{\widehat{V}_h\}_{h=1}^{H}$ constructed from Algorithm 1, we have the following holds with probability at least $1-\delta$,*

$$\|\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau)\cdot\epsilon_h^\tau(\beta_{h,i},\widehat{V}_{h+1})\|_{\Lambda_h^{-1}}^2 \leq 16d(e^{H/\beta_{h,i}}-1)^2\zeta_1 + 8(e^{H/\beta_{h,i}}-1)^2(\frac{H}{\beta_{h,i}}\zeta_2+2),$$

*for $\zeta_1=\log(2N+16Nd^{3/2}H^2e^{H/\underline{\beta}})$, $\zeta_2=\log(\frac{2dNH^3}{\delta\rho})$ and some absolute constant $c>1$.*

*Proof.* For the fixed $h\in[H]$ and fixed $\tau\in[N]$, we define the $\sigma$-algebra,

$$\mathcal{F}_h^\tau \coloneqq \sigma(\{s_h^{\tau'},a_h^{\tau'}\}_{\tau'=1}^{\tau}\cup\{r_h^{\tau'},s_{h+1}^{\tau'}\}_{\tau'=1}^{(\tau-1)\vee 0}),$$

i.e., $\mathcal{F}_h^\tau$ is the filtration generated by the samples $\{s_h^\tau,a_h^\tau\}_{\tau'=1}^{\tau}\cup\{r_h^\tau,s_{h+1}^\tau\}_{\tau'=1}^{(\tau-1)\vee 0}$. Notice that $\mathbb{E}[\epsilon_h^\tau(\beta_{h,i},\widehat{V}_{h+1})|\mathcal{F}_h^\tau]=0$. However, $\widehat{V}_{h+1}$ depends on $\{(s_h^\tau,a_h^\tau)\}_{\tau\in[N]}$ via $\{(s_{h'}^\tau,a_{h'}^\tau)\}_{h'>h,\tau\in[N]}$ and thus we cannot directly apply vanilla concentration bounds to control $\|\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau)\cdot\epsilon_h^\tau(\beta_{h,i},\widehat{V}_{h+1})\|_{\Lambda_h^{-1}}$.

To tackle this point, we consider the function family $V_h(R)$. In specific,

$$V_h(R)=\{V_h(x;\tilde{w}):\mathcal{S}\to[0,H-h+1]|\|\tilde{w}\|\leq R\},$$

and

$$V_h(x;\tilde{w}) = \max_{a\in\mathcal{A}}\{\phi(s,a)^\top \tilde{w}\}.$$

We let $\mathcal{N}_\epsilon(R)$ be the minimal $\epsilon$-cover of $\mathcal{V}_h(R)$ with respect to the supremum norm, i.e., for any function $V \in \mathcal{V}_h(R)$, there exists a function $V' \in \mathcal{N}_\epsilon(R)$ such that

$$\sup_{s\in\mathcal{S}}|V(s) - V'(s)| \le \epsilon.$$

Hence, for $\widehat{V}_{h+1}$, we have $V^\dagger_{h+1} \in \mathcal{N}_\epsilon(R)$ such that

$$\sup_{s\in\mathcal{S}}|\widehat{V}_{h+1}(s) - V^\dagger_{h+1}(s)| \le s.$$

For the ease of presentation, we ignore the subscript of $\beta_{h,i}$ and use $\beta$ in the following. Next we denote $\mathcal{N}_\epsilon(\underline{\beta},\overline{\beta})$ the minimal $\epsilon$-cover of the $[\underline{\beta},\overline{\beta}]$ with respect to the absolute value, i.e., for any $\beta \in [\underline{\beta},\overline{\beta}]$, there exists $\mathcal{N}_\epsilon(\beta) \in \mathcal{N}_\epsilon(\underline{\beta},\overline{\beta})$ such that

$$|\beta - \mathcal{N}_\epsilon(\beta)| \le \epsilon.$$

We proceed our analysis for all $i \in [d]$ and $\beta \in [\underline{\beta},\overline{\beta}]$

$$\begin{aligned}
&\|\sum_{\tau\in[N]}\phi(s_h^\tau,a_h^\tau)\epsilon_h^\tau(\beta,\widehat{V}_{h+1})\|^2_{\Lambda_h^{-1}}\\
&\le 2\|\sum_{\tau\in[N]}\phi(s_h^\tau,a_h^\tau)\epsilon_h^\tau(\beta,V^\dagger_{h+1})\|^2_{\Lambda_h^{-1}} + 2\|\sum_{\tau\in[N]}\phi(s_h^\tau,a_h^\tau)(\epsilon_h^\tau(\beta,\widehat{V}_{h+1}) - \epsilon_h^\tau(\beta,V^\dagger_{h+1}))\|^2_{\Lambda_h^{-1}}\\
&\le \sup_{V\in\mathcal{N}_\epsilon(R)} 2\|\sum_{\tau\in[N]}\phi(s_h^\tau,a_h^\tau)\epsilon_h^\tau(\beta,V)\|^2_{\Lambda_h^{-1}} + 2\|\sum_{\tau\in[N]}\phi(s_h^\tau,a_h^\tau)(\epsilon_h^\tau(\beta,\widehat{V}_{h+1}) - \epsilon_h^\tau(\beta,V^\dagger_{h+1}))\|^2_{\Lambda_h^{-1}}\\
&\le \underbrace{\sup_{V\in\mathcal{N}_\epsilon(R)} 4\|\sum_{\tau\in[N]}\phi(s_h^\tau,a_h^\tau)\epsilon_h^\tau(N_\epsilon(\beta),V)\|^2_{\Lambda_h^{-1}}}_{\text{I}} + \cdots\\
&\quad + \underbrace{\sup_{V\in\mathcal{N}_\epsilon(R)} 4\|\sum_{\tau\in[N]}(\phi(s_h^\tau,a_h^\tau)(\epsilon_h^\tau(\beta,V) - \epsilon_h^\tau(N_\epsilon(\beta),V)))\|^2_{\Lambda_h^{-1}}}_{\text{II}} + \cdots\\
&\quad + \underbrace{2\|\sum_{\tau\in[N]}\phi(s_h^\tau,a_h^\tau)(\epsilon_h^\tau(\beta,\widehat{V}_{h+1}) - \epsilon_h^\tau(\beta,V^\dagger_{h+1}))\|^2_{\Lambda_h^{-1}}}_{\text{III}},
\end{aligned} \tag{26a}$$

where the first and second inequality is from the fact that $\|a+b\|^2_\Lambda \le 2\|a\|^2_\Lambda + 2\|b\|^2_\Lambda$ for any vectors $a, b \in \mathbb{R}^d$ and any positive definite matrix $\Lambda \in \mathbb{R}^{d\times d}$. Here we decompose it into three parts: I term represents the error within a finite ball $\mathcal{N}_\epsilon(R)$ which can be controlled via classical finite-sample error. II term is the discretion error from the $\beta$, which can be controlled by choose proper $\epsilon$. III term is also the discretion error from the $V$ function space.

We invoke Lemma G.2, Lemma I.7 and Lemma I.8 for the term I, II and III respectively and have with probability at least $1-\delta$, for all $h \in [H]$,

$$\begin{aligned}
&\| \sum_{\tau \in [N]} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\beta, \widehat{V}_{h+1}) \|_{\Lambda_h^{-1}}^2 \\
&\leq 4d \left(e^{H/\beta} - 1\right)^2 \log\left(1 + \frac{N}{\lambda}\right) \\
&+ 8 \left(e^{H/\beta} - 1\right)^2 \cdot \log\left(\frac{H \left|\mathcal{N}_\epsilon(R, B, \lambda)\right| \left|\mathcal{N}_\epsilon(\underline{\beta}, \bar{\beta})\right|}{\delta}\right) \\
&\quad + \frac{16N^2(H-h)^2\epsilon^2}{\lambda\beta^4} e^{2H/\beta} + \frac{8N^2\epsilon^2}{\lambda\beta^2} e^{2H/\beta} \\
&\leq 4d(e^{H/\beta}-1)^2 \log(1+\frac{N}{\lambda}) + 8(e^{H/\beta}-1)^2 \cdot \log(\frac{H^2}{\epsilon\delta\rho}) + 8d(e^{H/\beta}-1)^2 \cdot \log(1+\frac{4R}{\epsilon}) \\
&\quad + \frac{16N^2(H-h)^2\epsilon^2}{\lambda\beta^4} e^{2H/\beta} + \frac{8N^2\epsilon^2}{\lambda\beta^2} e^{2H/\beta},
\end{aligned}$$

where the second inequality is from Lemma I.1 and the fact that $\mathcal{N}_\epsilon(\underline{\beta}, \overline{\beta}) \leq \frac{H}{\rho\epsilon}$.

By choosing $R = 2Hd^{3/2}$, $\lambda = 1$, $\epsilon = \frac{2}{4NHe^{H/\underline{\beta}}}$, then for all $\beta \in [\underline{\beta}, \overline{\beta}]$

$$\begin{aligned}
&\| \sum_{\tau \in [N]} \phi(s_h^\tau, a_h^\tau) \epsilon_h^\tau(\beta, \widehat{V}_{h+1}) \|_{\Lambda_h^{-1}}^2 \\
&\leq 4d(e^{H/\beta}-1)^2 \cdot \log(2N) + 8(e^{H/\beta}-1)^2 \cdot \frac{H}{\beta} \log(\frac{2dNH^3}{\delta\rho}) + \frac{1}{\beta^4} + \frac{1}{2H^2\beta^2} \\
&\quad + 8d(e^{H/\beta}-1)^2 \cdot \log(1 + 16Nd^{3/2}H^2 e^{\frac{H}{\underline{\beta}}}) \\
&\leq 16d(e^{H/\beta}-1)^2 \log(2N + 16Nd^{3/2}H^2e^{H/\underline{\beta}}) \\
&\quad + 8(e^{H/\beta}-1)^2 \cdot \frac{H}{\beta} \log(\frac{2dNH^3}{\delta\rho}) + \frac{1}{\beta^4} + \frac{1}{2H^2\beta^2} \\
&\leq 16d(e^{H/\beta}-1)^2 \log(2N + 16Nd^{3/2}H^2e^{H/\underline{\beta}}) + 8(e^{H/\beta}-1)^2(\frac{H}{\beta}\log(\frac{2dNH^3}{\delta\rho}) + 2),
\end{aligned}$$

where the second inequality is from $2N > 1$ and the third inequality is from $(e^{H/\beta}-1)^2 \geq (\frac{H}{\beta} + \frac{H^2}{\beta^2})^2 \geq \frac{H^2}{\beta^2} + \frac{H^4}{\beta^4} \geq \frac{1}{2H^2\beta^2} + \frac{1}{\beta^4}$. □

**Lemma G.2** (Concentration of Self-Normalized Processes)**.** *Let $V : \mathcal{S} \to [0, H]$ be any fixed function. For any fixed $h \in [H]$, any $0 < \delta < 1$, all $V \in \mathcal{N}_\epsilon(R)$ and all $\beta \in \mathcal{N}_\epsilon(\underline{\beta}, \overline{\beta})$, we have the following holds with probability at least $1-\delta$,*

$$\|\sum_{\tau=1}^N \phi(s_h^\tau, a_h^\tau)\epsilon_h^\tau(\beta, V)\|_{\Lambda_h^{-1}}^2 > d\left(e^{H/\beta}-1\right)^2 \log\left(1+\frac{N}{\lambda}\right) + 2\left(e^{H/\beta}-1\right)^2 \cdot \log\left(\frac{H\left|\mathcal{N}_\epsilon(R,B,\lambda)\right|\left|\mathcal{N}_\epsilon(\underline{\beta},\bar{\beta})\right|}{\delta}\right).$$

*Proof.* Recall the definition of the filtration $\mathcal{F}_h^\tau$ and note that $\phi(s_h^\tau, a_h^\tau) \in \mathcal{F}_h^\tau$. Moreover, for any fixed function $V : \mathcal{S} \to [0, H]$ and any fixed $\beta \in [\underline{\beta}, \overline{\beta}]$, we have $\epsilon_h^\tau(\beta, V) \in \mathcal{F}_h^\tau$ and $\mathbb{E}[\epsilon_h^\tau(\beta, V)|\mathcal{F}_h^\tau] = \mathbb{E}[(P(\cdot|s_h^\tau, a_h^\tau) - \mathbb{1}(s_{h+1}^\tau))^\top(e^{(H-V)/\beta} - e^{H/\beta})|\mathcal{F}_h^\tau] = 0$. Moreover, as we have $|\epsilon_h(V)| \in [0, e^{H/\beta}-1]$, for all fixed $h \in [H]$ and all $\tau \in [N]$, $\epsilon_h^\tau(\beta, V)$ is mean zero and $(e^{H/\beta}-1)$-sub-Gaussian conditional on $\mathcal{F}_h^\tau$.

We invoke Lemma I.3 with $V = \lambda \cdot I$, $X_t = \phi(s_h^\tau, a_h^\tau)$, $\eta_\tau = \epsilon_h^\tau(\beta, V)$ and $R = e^{H/\beta}-1$, we have

$$\mathbb{P}\left(\|\sum_{\tau=1}^N \phi(x_h^\tau, a_h^\tau) \cdot \epsilon_h(\beta, V)\|_{\Lambda_h^{-1}}^2 > 2(e^{H/\beta}-1)^2 \cdot \log(\frac{\det(\Lambda_h)^{1/2}}{\delta \cdot \det(\lambda \cdot I)^{1/2}})\right) \leq \delta,$$

for any $\delta > 0$. Moreover, $\det(\lambda \cdot I) = \lambda^d$, and

$$
\begin{aligned}
\|\Lambda_h\| &= \|\lambda \cdot I + \sum_{\tau=1}^{N} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\| && (27)\\
&\le \lambda + \sum_{\tau=1}^{N} \|\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\| && (28)\\
&\le \lambda + N, && (29)
\end{aligned}
$$

$\det(\Lambda_h) \le (\lambda + N)^d$ for $\Lambda_h$ is a positive-definite matrix. Hence we know

$$
\begin{aligned}
&2(e^{H/\beta} - 1)^2 \cdot \log\left(\frac{\det(\Lambda_h)^{1/2}}{\delta \cdot \det(\lambda \cdot I)^{1/2}}\right)\\
&= 2(e^{H/\beta} - 1)^2 \cdot \frac{d}{2} \log\left(1 + \frac{N}{\lambda}\right) + 2(e^{H/\beta} - 1)^2 \cdot \log(1/\delta)\\
&\le d(e^{H/\beta} - 1)^2 \log\left(1 + \frac{N}{\lambda}\right) + 2(e^{H/\beta} - 1)^2 \cdot \log(1/\delta),
\end{aligned}
$$

which implies

$$
\mathbb{P}\left(\|\sum_{\tau=1}^{N} \phi(s_h^\tau, a_h^\tau)\epsilon_h^\tau(\beta, V)\|^2_{\Lambda_h^{-1}} > d(e^{H/\beta} - 1)^2 \log(1 + \frac{N}{\lambda}) + 2(e^{H/\beta} - 1)^2 \cdot \log(1/\delta)\right) \le \delta.
$$

Finally we know by the union bound that for all $h \in [H]$, the following holds with probability at least $1 - \delta$, all $V \in \mathcal{N}_\epsilon(R)$ and all $\beta \in \mathcal{N}_\epsilon(\underline{\beta}, \overline{\beta})$,

$$
\begin{aligned}
\|\sum_{\tau=1}^{N} \phi(s_h^\tau, a_h^\tau)\epsilon_h^\tau(\beta, V)\|^2_{\Lambda_h^{-1}} >& d\left(e^{H/\beta} - 1\right)^2 \log\left(1 + \frac{N}{\lambda}\right)\\
&+ 2\left(e^{H/\beta} - 1\right)^2 \cdot \log\left(\frac{H\left|\mathcal{N}_\epsilon(R, B, \lambda)\right|\left|\mathcal{N}_\epsilon(\underline{\beta}, \bar{\beta})\right|}{\delta}\right).
\end{aligned}
$$

□

## H Proof of Theorem 5.3

In this section, we mainly prove the Theorem 5.3. By setting the model mis-specification $\xi = 0$, we can recover the results in Theorem 5.1.

*Proof of Theorem 5.1.* Following the same argument, $\widehat{V}_{h+1}$ depends on $\{(s_h^\tau, a_h^\tau)\}_{\tau \in [N]}$ via $\{(s_{h'}^\tau, a_{h'}^\tau)\}_{h' > h, \tau \in [N]}$ and thus we cannot directly apply vanilla concentration bounds to control $\|\sum_{\tau=1}^{N} \phi(s_h^\tau, a_h^\tau) \cdot \epsilon_h^\tau(\beta_{h,i}, \widehat{V}_{h+1})\|_{\Lambda_h^{-1}}$.

To tackle this point, we consider the function family $V_h(R)$. In specific,

$$
V_h(R) = \{V_h(x; w, \gamma, \Lambda) : \mathcal{S} \to [0, H - h + 1] | \|w\| \le R, \gamma \in [0, B], \Lambda \succeq \lambda \cdot I\},
$$

and

$$
V_h(x; w, \gamma, \Lambda) = \max_{a \in \mathcal{A}}\{\max\{\phi(s, a)^\top w - \gamma \cdot \sum_{i=1}^{d} \sqrt{(\phi_i(s, a)\mathbb{1}_i)^\top \Lambda^{-1}(\phi_i(s, a)\mathbb{1}_i)}, 0\}\}.
$$

We continue from the Equation 26a in Subsection G, i.e.,

$$
\begin{aligned}
\| \sum_{\tau\in[N]} \phi(s_h^\tau, a_h^\tau)\epsilon_h^\tau(\beta, \widehat{V}_{h+1})\|^2_{\Lambda_h^{-1}} \le & \underbrace{\sup_{V\in\mathcal{N}_\epsilon(R)} 4\| \sum_{\tau\in[N]} \phi(s_h^\tau, a_h^\tau)\epsilon_h^\tau(N_\epsilon(\beta), V)\|^2_{\Lambda_h^{-1}}}_{\text{I}} + \cdots \\
& + \underbrace{\sup_{V\in\mathcal{N}_\epsilon(R)} 4\| \sum_{\tau\in[N]} (\phi(s_h^\tau, a_h^\tau)(\epsilon_h^\tau(\beta, V) - \epsilon_h^\tau(N_\epsilon(\beta), V)))\|^2_{\Lambda_h^{-1}}}_{\text{II}} + \cdots \\
& + \underbrace{2\| \sum_{\tau\in[N]} \phi(s_h^\tau, a_h^\tau)(\epsilon_h^\tau(\beta, \widehat{V}_{h+1}) - \epsilon_h^\tau(\beta, V_{h+1}^\dagger))\|^2_{\Lambda_h^{-1}}}_{\text{III}},
\end{aligned}
$$

We invoke Lemma G.2, Lemma I.7 and Lemma I.8 for the term I, II and III respectively and for all $h \in [H]$ and any $\beta \in [\underline{\beta}, \overline{\beta}]$ we have the following holds with probability at least $1-\delta$, ,

$$
\begin{aligned}
& \| \sum_{\tau\in[N]} \phi(s_h^\tau, a_h^\tau)\epsilon_h^\tau(\beta, \widehat{V}_{h+1})\|^2_{\Lambda_h^{-1}} \\
& \le 4d\left(e^{H/\beta}-1\right)^2 \log\left(1+\frac{N}{\lambda}\right) \\
& + 8\left(e^{H/\beta}-1\right)^2 \cdot \log\left(\frac{H\left|\mathcal{N}_\epsilon(R,B,\lambda)\right| \left|\mathcal{N}_\epsilon(\underline{\beta},\bar{\beta})\right|}{\delta}\right) \\
& \quad + \frac{16N^2(H-h)^2\epsilon^2}{\lambda\beta^4}e^{2H/\beta} + \frac{8N^2\epsilon^2}{\lambda\beta^2}e^{2H/\beta} \\
& \le 4d(e^{H/\beta}-1)^2\log(1+\frac{N}{\lambda}) + 8(e^{H/\beta}-1)^2\cdot\log(\frac{H^2}{\epsilon\delta\rho}) + 8d(e^{H/\beta}-1)^2\cdot\log(1+\frac{4R}{\epsilon}) \\
& \quad + \frac{16N^2(H-h)^2\epsilon^2}{\lambda\beta^4}e^{2H/\beta} + \frac{8N^2\epsilon^2}{\lambda\beta^2}e^{2H/\beta} + 8(e^{H/\beta}-1)^2\cdot d^2\log(1+\frac{8\sqrt{d}B^2}{\lambda\epsilon^2}),
\end{aligned}
$$

where the second inequality is from Lemma I.2 and the fact that $|\mathcal{N}_\epsilon(\underline{\beta},\overline{\beta})| \le \frac{H}{\rho\epsilon}$.

By choosing $R = 2Hd^{3/2}$, $\zeta_2 = \log(\frac{2dNH^3}{\delta\rho})$, $\zeta_3 = \log(2N + 32N^2H^3d^{5/2}\zeta e^{2H/\beta})$, $\lambda = 1$, $\gamma = \underline{\beta}(e^{H/\underline{\beta}}-1)(\xi\sqrt{d} + c_1 d\sqrt{\zeta_3}) + c_2(e^{H/\underline{\beta}}-1)\sqrt{H\zeta_2}$, $B = 2\gamma$, $\epsilon = \frac{2}{4NHe^{H/\beta}} \le \frac{\beta}{2}$, then

$$
\begin{aligned}
& \| \sum_{\tau\in[N]} \phi(s_h^\tau, a_h^\tau)\epsilon_h^\tau(\beta, \widehat{V}_{h+1})\|^2_{\Lambda_h^{-1}} \\
& \le 4d(e^{H/\beta}-1)^2\cdot\log(2N) + 8(e^{H/\beta}-1)^2\cdot\frac{H}{\beta}\log(\frac{2dNH^3}{\delta\rho}) + \frac{1}{\beta^4} + \frac{1}{2H^2\beta^2} \\
& \quad + 8d(e^{H/\beta}-1)^2\cdot\log(1+8NH^2e^{\frac{H}{\underline{\beta}}}) + 8d^2(e^{H/\beta}-1)^2\cdot\log(1+32(c')^2N^2H^3d^{5/2}\gamma e^{2H/\underline{\beta}}) \\
& \le 16d(e^{H/\beta}-1)^2\log(2N+8NH^2e^{H/\underline{\beta}}) + \frac{1}{\beta^4} + \frac{1}{2H^2\beta^2} \\
& \quad + 8(e^{H/\beta}-1)^2\cdot\frac{H}{\beta}\log(\frac{2dNH^3}{\delta\rho}) + 8d^2(e^{H/\beta}-1)^2\cdot\log(1+32(c')^2N^2H^3d^{5/2}\gamma e^{2H/\underline{\beta}}) \\
& \le 16d(e^{H/\beta}-1)^2\log(2N+16Nd^{3/2}H^2e^{H/\underline{\beta}}) + 8(e^{H/\beta}-1)^2(\frac{H}{\beta}\log(\frac{2dNH^3}{\delta\rho})+2) \\
& \quad + 8d^2(e^{H/\beta}-1)^2\cdot\log(1+32(c')^2N^2H^3d^{5/2}\gamma e^{2H/\underline{\beta}}) \\
& \le 32d^2(e^{H/\beta}-1)^2\cdot\log(2N+32(c')^2N^2H^3d^{5/2}\gamma e^{2H/\underline{\beta}}) + 8(e^{H/\beta}-1)^2\cdot(\frac{H}{\beta}\log(\frac{2dNH^3}{\delta\rho})+2) \\
& \le 64d^2(e^{H/\beta}-1)^2\cdot\log(2N+32(c')^2N^2H^3d^{5/2}\gamma e^{2H/\underline{\beta}}) + 8(e^{H/\beta}-1)^2\cdot\frac{H}{\beta}\log(\frac{2dNH^3}{\delta\rho})
\end{aligned}
$$

Thus we have

$$\begin{aligned}
&\max_{i\in[d],\beta_{h,i}\in[\underline{\beta},\overline{\beta}]} \beta_{h,i}\Big\|\sum_{\tau=1}^{N}\phi(s_h^\tau,a_h^\tau)\cdot\epsilon_h^\tau(\beta_{h,i},\widehat{V}_{h+1})\Big\|_{\Lambda_h^{-1}}\\
&\le \max_{i\in[d],\beta_{h,i}\in[\underline{\beta},\overline{\beta}]} 8d\beta_{h,i}(e^{H/\beta_{h,i}}-1)\cdot\sqrt{\log(2N+32N^2H^3d^{5/2}\zeta e^{2H/\beta})}\\
&\quad+\max_{i\in[d],\beta_{h,i}\in[\underline{\beta},\overline{\beta}]} 2\sqrt{2}\beta_{h,i}(e^{H/\beta_{h,i}}-1)\cdot\sqrt{\frac{H}{\beta_{h,i}}\log(\frac{2dNH^3}{\delta\rho})}\\
&\le \max_{i\in[d],\beta_{h,i}\in[\underline{\beta},\overline{\beta}]} 8d\beta_{h,i}(e^{H/\beta_{h,i}}-1)\cdot\sqrt{\log(2N+32N^2H^3d^{5/2}\zeta e^{2H/\beta})}\\
&\quad+\max_{i\in[d],\beta_{h,i}\in[\underline{\beta},\overline{\beta}]} 2\sqrt{2\beta_{h,i}}(e^{H/\beta_{h,i}}-1)\cdot\sqrt{H\log(\frac{2dNH^3}{\delta\rho})}\\
&\le 8d\underline{\beta}(e^{H/\underline{\beta}}-1)\sqrt{\zeta_3}+2\sqrt{2}\sqrt{\underline{\beta}}(e^{H/\underline{\beta}}-1)\sqrt{H\zeta_2},
\end{aligned}$$

holds for probability at least $1-\delta$ for all $h\in[H]$ for some constant $c>1$ and $\zeta_2=\log(\frac{2dNH^3}{\delta\rho})$ and $\zeta_3=\log(2N+32N^2H^3d^{5/2}\zeta e^{2H/\beta})$. Thus we have

$$\begin{aligned}
&|\phi(s,a)^\top(w_h-\widehat{w}_h)|\\
&\le \max_{i\in[d],\beta_{h,i}\in[\underline{\beta},\overline{\beta}]} \beta_{h,i}(e^{H/\beta_{h,i}}-1)(\xi\sqrt{d}+8d\sqrt{\zeta_3}+\sqrt{d})\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}\\
&\quad+2\sqrt{2}\sqrt{\beta_{h,i}}(e^{H/\beta_{h,i}}-1)\sqrt{H\zeta_2}\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}\\
&\le \underline{\beta}(e^{H/\underline{\beta}}-1)(\xi\sqrt{d}+9d\sqrt{\zeta_3})\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}+2\sqrt{2}\sqrt{\underline{\beta}}(e^{H/\underline{\beta}}-1)\sqrt{H\zeta_2}\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}},
\end{aligned}\tag{30}$$

where the last inequality is by noticing that $f(\beta)=\beta(e^{H/\beta}-1)$ and $g(\beta)=\sqrt{\beta}(e^{H/\beta}-1)$ are both monotonically decreasing with $\beta>0$.

Plug Equation 30 and 20 into Equation 19 to finally upper bound the Bellman error with the choice $\lambda=1$,

$$\begin{aligned}
|\iota_h(s,a)|&\le\sqrt{d}\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}+\underline{\beta}(e^{H/\underline{\beta}}-1)(\xi\sqrt{d}+9d\sqrt{\zeta_3})\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}\\
&\quad+2\sqrt{2}\sqrt{\underline{\beta}}(e^{H/\underline{\beta}}-1)\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}+(H-h)\xi\\
&\le\underline{\beta}(e^{H/\underline{\beta}}-1)(\xi\sqrt{d}+10d\sqrt{\zeta_3})\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}+2\sqrt{2}\sqrt{\underline{\beta}}(e^{H/\underline{\beta}}-1)\sqrt{H\zeta_2}\sum_{i=1}^{d}\|\phi_i(s,a)\mathbb{1}_i\|_{\Lambda_h^{-1}}\\
&\quad+(H-h)\xi,
\end{aligned}$$

where the last inequality is from the fact that $f(\beta)=\beta(e^{H/\beta}-1)\ge H\ge1$ and and $\sqrt{d\zeta_3}>1$ as $N\ge e/2$.

Using the similar steps in the proof of Corollar 4.5 in Jin et al. (2021), we know that

$$\begin{aligned}
\mathrm{SubOpt}(\widehat{\pi};\mathcal{P}) &\le c' \sum_{h=1}^{H} \mathbb{E}[\sum_{i=1}^{d} \sqrt{(\phi_i(s,a)\mathbb{1}_i)^\top \Lambda_h^{-1}(\phi_i(s,a)\mathbb{1}_i)}] + H(H-1)\xi/2 \\
&\le c' \sum_{h=1}^{H} \mathbb{E}[\sum_{i=1}^{d} \sqrt{\mathrm{Tr}(\Lambda_h^{-1}(\phi_i(s,a)\mathbb{1}_i)(\phi_i(s,a)\mathbb{1}_i)^\top)}] + H(H-1)\xi/2 \\
&\le c' \sum_{h=1}^{H} \mathbb{E}[\sum_{i=1}^{d} \sqrt{\frac{\lambda_{h,i,j}}{1+c^\dagger \cdot N \cdot \lambda_{h,i,j}}}] + H(H-1)\xi/2 \\
&\le c' \sum_{h=1}^{H} \mathbb{E}[\sum_{i=1}^{d} \sqrt{\frac{1}{1+c^\dagger \cdot N}}] + H(H-1)\xi/2 \\
&\le c' \cdot d^{1/2} H/\sqrt{N} + H(H-1)\xi/2,
\end{aligned}$$

where $c' := c_1\underline{\beta}(e^{H/\underline{\beta}}-1)(\xi d + d^{3/2}\sqrt{\zeta_3})H/ + c_2\sqrt{\underline{\beta}}(e^{H/\underline{\beta}}-1)d^{1/2}H^{3/2}\sqrt{\zeta_2}$ and $c_1, c_2$ are some absolute constants only depend on $c^\dagger$.

In conclusions, we have probability at least $1-\delta$,

$$\begin{aligned}
\mathrm{SubOpt}(\widehat{\pi};\mathcal{P}) \le & c_1\underline{\beta}(e^{H/\underline{\beta}}-1)(\xi d + d^{3/2}\sqrt{\zeta_3})H/N^{1/2} + c_2\sqrt{\underline{\beta}}(e^{H/\underline{\beta}}-1)d^{1/2}H^{3/2}\sqrt{\zeta_2}/N^{1/2} \\
& + H(H-1)\xi/2.
\end{aligned}$$

□

## I Auxiliary Lemma

Before we proceed our analysis, we need the following lemmas.

**Fact I.1.** *For $x, y \ge 0$ and $b > 0$, we have $|e^{-by} - e^{-bx}| \le b|x-y|$.*

**Fact I.2.** *For $x, y \ge 0$ and $b > 0$, we have $b|x-y| \le |e^{bx} - e^{by}|$.*

**Lemma I.1** ($\epsilon$-Covering Number Jin et al. (2020))**.** *Let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with following parametric form*

$$V(s) = \max_a w^\top \boldsymbol{\phi}(s,a),$$

*where the parameters $w$ satisfy $\|w\| \le L, \beta \in [0,B]$ Assume $\|\phi(x,a)\| \le 1$ for all $(x,a)$ pairs, and let $\mathcal{N}_\varepsilon$ be the $\varepsilon$-covering number of $\mathcal{V}$ with respect to the distance* $\operatorname{dist}(V,V') = \sup_x |V(x) - V'(x)|$. *Then*

$$\log|\mathcal{N}_\epsilon(R,B,\lambda)| \le ds\log(1+4R/\epsilon).$$

**Lemma I.2.** *Let $\mathcal{V}$ denote a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with following parametric form*

$$V(s) = \max_a w^\top \boldsymbol{\phi}(s,a) + \beta \sum_{i=1}^{d} \sqrt{(\phi_i(s,a)\mathbb{1}_i)^\top \Lambda^{-1}(\phi_i(s,a)\mathbb{1}_i)},$$

*where the parameters $(w,\beta,\Lambda)$ satisfy $\|w\| \le L, \beta \in [0,B]$ and the minimum eigenvalue satisfies $\lambda_{\min}(\Lambda) \ge \lambda$. Assume $\|\phi(x,a)\| \le 1$ for all $(x,a)$ pairs, and let $\mathcal{N}_\varepsilon$ be the $\varepsilon$-covering number of $\mathcal{V}$ with respect to the distance* $\operatorname{dist}(V,V') = \sup_x |V(x) - V'(x)|$. *Then*

$$\log\mathcal{N}_\varepsilon \le d\log(1+4L/\varepsilon) + d^2\log\left[1 + 8d^{1/2}B^2/\left(\lambda\varepsilon^2\right)\right].$$

*Proof.* Equivalently, we can reparametrize the function class $\mathcal{V}$ by let $A = \beta^2\Lambda^{-1}$, so we have

$$V(s) = \max_a w^\top \phi(s,a) + \sum_{i=1}^{d} \sqrt{(\phi_i(s,a)\mathbb{1}_i)^\top A(\phi_i(s,a)\mathbb{1}_i)},$$

for $\|w\| \leq L$ and $\|A\| \leq B^2\lambda^{-1}$. For any two functions $V_1, V_2 \in \mathcal{V}$, let them take the form in Eq. (27) with parameters $(w_1, A_1)$ and $(w_2, A_2)$, respectively. Then, since both $\min\{\cdot, H\}$ and $\max_a$ are contraction maps, we have

$$
\begin{aligned}
&\operatorname{dist}(V_1, V_2) \\
&\leq \sup_{x,a} \Bigg| \left[ w_1^\top \phi(x,a) + \sum_{i=1}^{d} \sqrt{(\phi_i(s,a)\mathbb{1}_i)^\top A_2 (\phi_i(s,a)\mathbb{1}_i)} \right] - \\
&\qquad\qquad \left[ w_2^\top \phi(x,a) + \sum_{i=1}^{d} \sqrt{(\phi_i(s,a)\mathbb{1}_i)^\top A_1 (\phi_i(s,a)\mathbb{1}_i)} \right] \Bigg| \\
&\leq \sup_{\phi:\|\phi\|\leq 1} \left| \left[ w_1^\top \phi + \sum_{i=1}^{d} \sqrt{(\phi_i(s,a)\mathbb{1}_i)^\top A_1 (\phi_i(s,a)\mathbb{1}_i)} \right] - \left[ w_2^\top \phi + \sum_{i=1}^{d} \sqrt{(\phi_i(s,a)\mathbb{1}_i)^\top A_2 (\phi_i(s,a)\mathbb{1}_i)} \right] \right| \\
&\leq \sup_{\phi:\|\phi\|\leq 1} \left| (w_1 - w_2)^\top \phi \right| + \sup_{\phi:\|\phi\|\leq 1} \sum_{i=1}^{d} \sqrt{\left| (\phi_i(s,a)\mathbb{1}_i)^\top (A_1 - A_2) (\phi_i(s,a)\mathbb{1}_i) \right|} \\
&= \|w_1 - w_2\| + \sqrt{\|A_1 - A_2\|} \sup_{\phi:\|\phi\|\leq 1} \sum_{i=1}^{d} \|\phi_i(s,a)\mathbb{1}_i\| \\
&\leq \|w_1 - w_2\| + \sqrt{\|A_1 - A_2\|_F},
\end{aligned}
$$

where the second last inequality follows from the fact that $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x-y|}$ holds for any $x, y \geq 0$. For matrices, $\|\cdot\|$ and $\|\cdot\|_F$ denote the matrix operator norm and Frobenius norm respectively.

Let $\mathcal{C}_w$ be an $\varepsilon/2$-cover of $\left\{ w \in \mathbb{R}^d \mid \|w\| \leq L \right\}$ with respect to the 2-norm, and $\mathcal{C}_A$ be an $\varepsilon^2/4$-cover of $\left\{ A \in \mathbb{R}^{d\times d} \mid \|A\|_F \leq d^{1/2}B^2\lambda^{-1} \right\}$ with respect to the Frobenius norm. By Lemma D.5, we know:

$$
|\mathcal{C}_w| \leq (1 + 4L/\varepsilon)^d, \quad |\mathcal{C}_A| \leq \left[ 1 + 8d^{1/2}B^2 / \left(\lambda\varepsilon^2\right) \right]^{d^2}.
$$

By Eq. (28), for any $V_1 \in \mathcal{V}$, there exists $w_2 \in \mathcal{C}_w$ and $A_2 \in \mathcal{C}_A$ such that $V_2$ parametrized by $(w_2, A_2)$ satisfies $\operatorname{dist}(V_1, V_2) \leq \varepsilon$. Hence, it holds that $\mathcal{N}_\varepsilon \leq |\mathcal{C}_w| \cdot |\mathcal{C}_A|$, which gives:

$$
\log \mathcal{N}_\varepsilon \leq \log |\mathcal{C}_w| + \log |\mathcal{C}_A| \leq d \log(1 + 4L/\varepsilon) + d^2 \log \left[ 1 + 8d^{1/2}B^2 / \left(\lambda\varepsilon^2\right) \right]
$$

This concludes the proof. □

**Lemma I.3** (Self-Normalized Bound for Vector-Valued Martingales Abbasi-Yadkori et al. (2011))**.** *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that $\eta_t$ is $\mathcal{F}_t$-measurable and $\eta_t$ is conditionally $R$-sub-gaussian for some $R \geq 0$, i.e.,*

$$
\forall \lambda \in \mathbb{R}, \mathbb{E}[e^{\lambda\eta_t} | \mathcal{F}_{t-1}] \leq e^{\frac{\lambda^2 R^2}{2}}.
$$

*Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochstic process such that $X_t$ is $\mathcal{F}_{t-1}$-measurable. Assume that $V$ is a $d \times d$ positive definite matrix. For any $t \geq 0$, define*

$$
V_t = V + \sum_{s=1}^{\top} X_s X_s^\top, \quad S_t = \sum_{s=1}^{\top} \eta_s X_s.
$$

*Then for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$,*

$$
\|S_t\|_{V_t^{-1}}^2 \leq 2R^2 \log\left(\frac{\det(V_t)^{1/2}\det(V)^{-1/2}}{\delta}\right).
$$

**Lemma I.4.** *For any $0 \leq x \leq H$, $|\beta' - \beta| \leq \epsilon$ for some $\epsilon > 0$, we have*

$$
|e^{(H-x)/\beta'} - e^{(H-x)/\underline{\beta}}| \leq e^{2H/\underline{\beta}} \cdot \frac{4H}{\beta^2}\epsilon.
$$

*Proof.* We denote $f(z) = e^{(H-x)/z}$. Then

$$\begin{aligned}
|f'(z)| &= |e^{(H-x)/z} \cdot \frac{(H-x)}{z^2}| \\
&\leq e^{H/z} \cdot \frac{H}{z^2} \\
&\leq e^{2H/\beta} \cdot \frac{4H}{\beta^2},
\end{aligned}$$

where the last inequality is from the fact that $z \geq \beta - \epsilon \geq \frac{\beta}{2}$. Thus by the mean value theorem we know

$$|e^{(H-x)/\beta'} - e^{(H-x)/\beta}| \leq e^{2H/\underline{\beta}} \cdot \frac{4H}{\beta^2}\epsilon.$$

□

**Lemma I.5.** *For any $\beta \in [\underline{\beta}, \overline{\beta}]$, and any $\epsilon$-net over $[\underline{\beta}, \overline{\beta}]$, i.e., $\mathcal{N}_\epsilon(\beta)$ for some $\epsilon > 0$, we have*

$$|\epsilon(\beta, V) - \epsilon(\mathcal{N}_\epsilon(\beta), V)| \leq 2e^{2H/\underline{\beta}} \cdot \frac{4H}{\beta^2}\epsilon.$$

*Proof.*

$$\begin{aligned}
&|\epsilon(\beta, V) - \epsilon(\mathcal{N}_\epsilon(\beta), V)| \\
&= |\mathbb{E}_{s'}[e^{(H-V(s'))/\beta}] - e^{(H-V(s_{h+1}^\tau))/\beta} - (\mathbb{E}_{s'}[e^{(H-V(s'))/\mathcal{N}_\epsilon(\beta)}] - e^{(H-V(s_{h+1}^\tau))/\mathcal{N}_\epsilon(\beta)})| \\
&= |\mathbb{E}_{s'}[e^{(H-V(s'))/\beta} - e^{(H-V(s'))/\mathcal{N}_\epsilon(\beta)}]| + |(e^{(H-V(s_{h+1}^\tau))/\beta} - e^{(H-V(s_{h+1}^\tau))/\mathcal{N}_\epsilon(\beta)})| \\
&\leq 2e^{2H/\underline{\beta}} \cdot \frac{4H}{\beta^2}\epsilon,
\end{aligned}$$

where the last inequality is form Lemma I.4. □

**Lemma I.6.** *For any $V^\dagger$ and $V^\ddagger : \mathcal{S} \to [0, H]$, and $\|V^\dagger - V^\ddagger\|_\infty \leq \epsilon$, for some $\epsilon > 0$, we have*

$$|\epsilon(\beta, V^\dagger) - \epsilon(\beta, V^\ddagger)| \leq e^{H/\underline{\beta}}\frac{2\epsilon}{\beta}.$$

*Proof.*

$$\begin{aligned}
&|\epsilon(\beta, V^\dagger) - \epsilon(\beta, V^\ddagger)| \\
&= |\mathbb{E}_{s'}[e^{(H-V^\dagger(s'))/\beta}] - e^{(H-V^\dagger(s'))/\beta} - (\mathbb{E}_{s'}[e^{(H-V^\ddagger(s'))/\beta}] - e^{(H-V^\ddagger(s'))/\beta})| \\
&= |\mathbb{E}_{s'}[e^{(H-V^\dagger(s'))/\beta} - e^{(H-V^\ddagger(s'))/\beta}]| + |(e^{(H-V^\dagger(s'))/\beta} - e^{(H-V^\ddagger(s'))/\beta})| \\
&= e^H|\mathbb{E}_{s'}[e^{(-V^\dagger(s'))/\beta} - e^{(-V^\ddagger(s'))/\beta}| + |(e^{(-V^\dagger(s'))/\beta} - e^{(-V^\ddagger(s'))/\beta})| \\
&\leq e^{H/\underline{\beta}}\frac{2\epsilon}{\beta},
\end{aligned}$$

where the last inequality is form Fact I.1. □

**Lemma I.7.**

$$\|\sum_{\tau \in [N]} (\phi(s_h^\tau, a_h^\tau)\,(\epsilon_h^\tau(\beta, V) - \epsilon_h^\tau(N_\epsilon(\beta), V)))\|_{\Lambda_h^{-1}}^2 \leq \frac{4N^2(H-h)^2\epsilon^2}{\lambda\underline{\beta}^4}e^{2H/\underline{\beta}}.$$

*Proof.*

$$
\begin{aligned}
&\| \sum_{\tau \in [N]} \left(\phi(s_h^\tau, a_h^\tau) \left(\epsilon_h^\tau(\beta, V) - \epsilon_h^\tau(N_\epsilon(\beta), V)\right)\right)\|^2_{\Lambda_h^{-1}} \\
&\leq \sum_{\tau,\tau'}^{N} \phi(s_h^\tau, a_h^\tau)^\top \Lambda_h^{-1} \phi(s_h^\tau, a_h^\tau) \cdot (\epsilon_h^\tau(\beta, V) - \epsilon_h^\tau(N_\epsilon(\beta), V)) \cdot (\epsilon_h^{\tau'}(\beta, V) - \epsilon_h^{\tau'}(N_\epsilon(\beta), V) \\
&\leq \frac{4(H-h)^2\epsilon^2}{\underline{\beta}^4} e^{2H/\beta} \sum_{\tau,\tau'}^{N} |\phi(s_h^\tau, a_h^\tau)^\top \Lambda_h^{-1} \phi(s_h^\tau, a_h^\tau)| \\
&\leq \frac{4(H-h)^2\epsilon^2}{\underline{\beta}^4} e^{2H/\beta} \sum_{\tau,\tau'}^{N} \|\phi(s_h^\tau, a_h^\tau)\|^2 \|\Lambda_h^{-1}\| \\
&\leq \frac{4N^2(H-h)^2\epsilon^2}{\lambda\underline{\beta}^4} e^{2H/\underline{\beta}},
\end{aligned}
$$

where the first inequality is from Lemma I.5. □

**Lemma I.8.**

$$
\| \sum_{\tau \in [N]} \phi(s_h^\tau, a_h^\tau)(\epsilon_h^\tau(\beta, \widehat{V}_{h+1}) - \epsilon_h^\tau(\beta, V_{h+1}^\dagger))\|^2_{\Lambda_h^{-1}} \leq \frac{4N^2\epsilon^2}{\lambda\underline{\beta}^2} e^{2H/\underline{\beta}}.
$$

*Proof.*

$$
\begin{aligned}
&\| \sum_{\tau \in [N]} \phi(s_h^\tau, a_h^\tau)(\epsilon_h^\tau(\beta, \widehat{V}_{h+1}) - \epsilon_h^\tau(\beta, V_{h+1}^\dagger))\|^2_{\Lambda_h^{-1}} \\
&\leq \sum_{\tau,\tau'}^{N} \phi(s_h^\tau, a_h^\tau)^\top \Lambda_h^{-1} \phi(s_h^\tau, a_h^\tau) \cdot (\epsilon_h^\tau(\beta, \widehat{V}_{h+1}) - \epsilon_h^\tau(\beta, V_{h+1}^\dagger)) \cdot (\epsilon_h^{\tau'}(\beta, \widehat{V}_{h+1}) - \epsilon_h^{\tau'}(\beta, V_{h+1}^\dagger) \\
&\leq \frac{4\epsilon^2}{\underline{\beta}^2} e^{2H/\underline{\beta}} \sum_{\tau,\tau'}^{N} |\phi(s_h^\tau, a_h^\tau)^\top \Lambda_h^{-1} \phi(s_h^{\tau'}, a_h^{\tau'})| \\
&\leq \frac{4\epsilon^2}{\underline{\beta}^2} e^{2H/\underline{\beta}} \sum_{\tau,\tau'}^{N} \|\phi(s_h^\tau, a_h^\tau)\| \|\phi(s_h^{\tau'}, a_h^{\tau'})\| \|\Lambda_h^{-1}\| \\
&\leq \frac{4N^2\epsilon^2}{\lambda\underline{\beta}^2} e^{2H/\underline{\beta}},
\end{aligned}
$$

□

where the first inequality is from Lemma I.6 and the last inequality is from the fact that $\|x\|_2 \leq \|x\|_1$ for any $x \in \mathbb{R}^d$ and Assumption 4.2 that $\|\phi(s,a)\|_1 = 1$ and also the fact that $\|\Lambda_h^{-1}\| \leq \frac{1}{\lambda}$.