CONTENTS

A Table of Notations

Table 2: Table of Notations

Notation	Meaning
$J(\pi)$	Policy value $\mathbb{E}_\pi[\sum_{h=1}^H \gamma^{h-1} R_h]$
b_h^π, q_h^π	value bridge function, weight bridge function of π at step h
$\mathbf{b}^\pi$, $\mathbf{q}^\pi$	value bridge function vector, weight bridge function vector of π
$\mathrm{CR}^\pi(\xi)$	confidence region of $\mathbf{b}^\pi$, according to (3.12)
$\mathbf{b}$	an element in the confidence region $\mathrm{CR}^\pi(\xi)$
$F(\mathbf{b})$, $\widehat{F}(\mathbf{b})$	a mapping for identification with $J(\pi) = F(\mathbf{b}^\pi)$, according to (3.4)
ℓ_h^π	"Bellman residual" for bridge functions, according to (3.5)
$\mathcal{L}_h^\pi$	residual mean square loss for ℓ_h^π, according to (3.6)
$\Phi_{\pi,h}^\lambda$, $\widehat{\Phi}_{\pi,h}^\lambda$	a mapping for minimax estimation, according to (3.9)
$\widehat{b}_h(b_{h+1})$	minimax estimator of b_h^π given b_{h+1}, according to (3.11)
$\widehat{J}_{\text{pess}}(\pi)$	pessimistic estimator of $J(\pi)$, according to (3.13)
$\widehat{\pi}$	policy returned by `P3O` algorithm, according to (3.14)

In this section, we provide a comprehensive clarification on the use of notation in this paper.

We use lower case letters (i.e., s, a, o, and τ) to represent dummy variables and upper case letters (i.e., S, A, O, and Γ) to represent random variables. We use the variables in the calligraphic font (i.e., $\mathcal{S}$, $\mathcal{A}$, $\mathcal{O}$, and $\mathcal{H}$) to represent the spaces of variables, and the blackboard bold font (i.e., $\mathbb{P}$ and $\mathbb{O}$) to represent probability kernels.

We use $\mathcal{H} = \{\mathcal{H}_h\}_{h=0}^{H-1}$ to denote the space of observable history, where each element $\tau_h \in \mathcal{H}_h$ is a (partial) trajectory such that $\tau_h \subseteq \{(o_1, a_1), \cdots, (o_h, a_h)\}$. We use $\pi^b = \{\pi_h^b\}_{h=1}^H$ to denote the behavior policy, where $\pi_h^b : \mathcal{S} \mapsto \Delta(\mathcal{A})$. We use $\pi = \{\pi_h\}_{h=1}^H \in \Pi(\mathcal{H})$ to denote a history-dependent policy with $\pi_h : \mathcal{O} \times \mathcal{H}_{h-1} \mapsto \Delta(\mathcal{A})$. Also, we use $\pi^\star = \{\pi_h^\star\}_{h=1}^H$ to denote the optimal history-dependent policy. Offline data $\mathbb{D}$ is collected by π^b, as described in Section 2.2.

We use $\mathcal{P}^b = \{\mathcal{P}_h^b\}_{h=1}^H$ and $\mathcal{P}^\pi = \{\mathcal{P}_h^\pi\}_{h=1}^H$ to denote the distribution of trajectories under the policy π^b and π, respectively, where $\mathcal{P}_h^b$ and $\mathcal{P}_h^\pi$ denote the density of corresponding variables at step h. Also, we use $\mathbb{E}_{\pi^b}$ and $\mathbb{E}_\pi$ to denote the expectation w.r.t. the distribution $\mathcal{P}^b$ and $\mathcal{P}^\pi$. We use $\widehat{\mathbb{E}}_{\pi^b}$ to denote the empirical version of $\mathbb{E}_{\pi^b}$, which is calculated on data $\mathbb{D}$.

Through out the paper, we use $\mathcal{O}(\cdot)$ to hide problem-independent constants and use $\widetilde{\mathcal{O}}(\cdot)$ to hide problem-independent constants plus logarithm factors. The following table summaries the notations we used in our proposed algorithm design and theory.

B Further Discussion

B.1 Further Discussion on Related Work

Reinforcement learning in POMDPs. Our work is related to the recent line of research on developing provably efficient online RL methods for POMDPs (Guo et al., 2016; Krishnamurthy et al., 2016; Jin et al., 2020; Xiong et al., 2021; Jafarnia-Jahromi et al., 2021; Efroni et al., 2022; Liu et al., 2022). In the online setting, the actions are specified by history-dependent policies and thus the latent state does not directly affect the actions. Thus, the actions and observations in the online setting are not confounded by latent states. Consequently, although these work also conduct uncertainty quantification to encourage exploration, the confidence regions are not based on confounded data and are thus constructed differently.

Offline reinforcement learning and pessimism. Our work is also related to the literature on offline RL and particularly related to the works based on the pessimism principle (Antos et al., 2007; Munos and Szepesvári, 2008; Chen and Jiang, 2019; Buckman et al., 2020; Liu et al., 2020; Min et al., 2021; Jin et al., 2021; Zanette, 2021; Jin et al., 2021; Xie et al., 2021; Uehara and Sun, 2021; Yin and Wang, 2021; Rashidinejad et al., 2021; Zhan et al., 2022; Yin et al., 2022; Yan et al., 2022). Offline RL faces the challenge of the distributional shift between the behavior policy and the family of target policies. Without any coverage assumption on the offline data, the number of data needed to find a near-optimal policy can be exponentially large (Buckman et al., 2020; Zanette, 2021). To circumvent this problem, a few existing works study offline RL under a uniform coverage assumption, which requires the concentrability coefficients between the behavior and target policies are uniformly bounded. See, e.g., Antos et al. (2007); Munos and Szepesvári (2008); Chen and Jiang (2019) and the references therein. Furthermore, a more recent line of work aims to weaken the uniform coverage assumption by adopting the pessimism principle in algorithm design (Liu et al., 2020; Jin et al., 2021; Rashidinejad et al., 2021; Uehara and Sun, 2021; Xie et al., 2021; Yin and Wang, 2021; Zanette et al., 2021; Yin et al., 2022; Yan et al., 2022). In particular, these works proves theoretically that pessimism is effective in tackling the distributional shift of the offline dataset. In particular, by constructing pessimistic value function estimates, these works establish upper bounds on the suboptimality of the proposed methods based on significantly weaker partial coverage assumption. That is, these methods can find a near-optimal policy as long as the dataset covers the optimal policy. The efficacy of pessimism has also been validated empirically in Kumar et al. (2020); Kidambi et al. (2020); Yu et al. (2021); Janner et al. (2021). Compared with these works on pessimism, we focus on the more challenging setting of POMDP with a confounded dataset. To perform pessimism in the face of confounders, we conduct uncertainty quantification for the minimax estimation regarding the confounding bridge functions. Our work complements this line of research by successfully applying pessimism to confounded data.

OPE via causal inference. Our work is closely related to the line of research that employing tools from causal inference (Pearl, 2009) for studying OPE with unobserved confounders (Oberst and Sontag, 2019; Kallus and Zhou, 2020; Bennett et al., 2021; Kallus and Zhou, 2021; Mastouri et al., 2021; Shi et al., 2021; Bennett and Kallus, 2021; Shi et al., 2022). Among them, Bennett and Kallus (2021); Shi et al. (2021) are most relevant to our work. In particular, these works also leverage proximal causal inference (Lipsitch et al., 2010; Miao et al., 2018a;b; Cui et al., 2020; Tchetgen et al., 2020; Singh, 2020) to identify the value of the target policy in POMDPs. See Tchetgen et al. (2020) for a detailed survey of proximal causal inference. In comparison, this line of research only focuses on evaluating a single policy, whereas we focus on learning the optimal policy within a class of target policies. As a result, we need to handle a more challenging distributional shift problem between the behavior policy and an entire class of target policies, as opposed to a single target policy in OPE. However, thanks to the pessimism, we establish theory based on a partial coverage assumption that is similar to that in the OPE literature. To achieve such a goal, we conduct uncertainty quantification for the bridge function estimators, which is absent in the the works on OPE. As a result, our analysis is different from that in Bennett and Kallus (2021); Shi et al. (2021).

Relations between minimax-typed loss and least-square-typed loss (Xie et al., 2021). During the preparation of this paper, we find that *in the MDP setting*, the least-square-typed loss considered by (Xie et al., 2021) can be reformulated to the minimax-typed loss that we consider in this paper with a different dual function class. To see this, consider the MDP setting with a single transition tuple (S_h, A_h, S_{h+1}). The goal is to estimate the Bellman target $(\mathcal{B}V_{h+1})\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$, where $\mathcal{B}$ is the Bellman operator and $V_{h+1}\colon \mathcal{S}\to\mathbb{R}$ is a fixed state-value function. For each $(s,a)\in\mathcal{S}\times\mathcal{A}$, $(\mathcal{B}^\pi f_{h+1})(s,a)$ is given by

$$(\mathcal{B}f_{h+1})(s,a) = R_h(s,a) + \int_{\mathcal{S}} P_h(\mathrm{d}s'|s,a)V_{h+1}(s').$$

Here R_h is the reward function and we can assume it is known for now, and $P_h : \mathcal{S}\times\mathcal{A}\mapsto\Delta(\mathcal{S})$ is the unknown transition kernel. We use function class $\mathcal{F}$ to approximate the bellman target. Then based on the offline transition data $\mathbb{D} = \{(s_h^\tau, a_h^\tau, s_{h+1}^\tau)\}_{\tau=1}^N$, the least-square-typed loss function given in Equation (3.1) of (Xie et al., 2021) becomes

$$\begin{aligned}\widehat{\mathcal{L}}_h^{\mathrm{ls}}(f_h) = \widehat{\mathbb{E}}_{\mathbb{D}}\Big[\big(f_h(S_h,A_h) - R_h - V_{h+1}(S_{h+1})\big)^2\Big] \\ - \min_{f'_h\in\mathcal{F}} \widehat{\mathbb{E}}_{\mathbb{D}}\Big[\big(f'_h(S_h,A_h) - R_h - V_{h+1}(S_{h+1})\big)^2\Big],\end{aligned} \tag{B.1}$$

where R_h is an abbreviation for $R_h(S_h,A_h)$. Using the equality $x^2 - y^2 = (x+y)(x-y)$, we can rewrite the least-square-typed loss (B.1) as

$$\widehat{\mathcal{L}}_h^{\mathrm{ls}}(f_h) = \sup_{f'_h\in\mathcal{F}} \widehat{\mathbb{E}}_{\mathbb{D}}\Big[\Big((f_h + f'_h)(S_h,A_h) - 2R_h - 2V_{h+1}(S_{h+1})\Big)\Big((f_h - f'_h)(S_h,A_h)\Big)\Big].$$

For derivation, we further rewrite first term as

$$\begin{aligned}&(f_h + f'_h)(S_h,A_h) - 2R_h - 2V_{h+1}(S_{h+1}) \\ &\quad = \Big(2f_h(S_h,A_h) - 2R_h - 2V_{h+1}(S_{h+1})\Big) - \Big((f_h - f'_h)(S_h,A_h)\Big).\end{aligned}$$

With this, we can then rewrite the least-square-typed loss (B.1) as

$$\begin{aligned}\widehat{\mathcal{L}}_h^{\mathrm{ls}}(f_h) = \sup_{f'_h\in\mathcal{F}} \widehat{\mathbb{E}}_{\mathbb{D}}\Big[&\Big(2f_h(S_h,A_h) - 2R_h - 2V_{h+1}(S_{h+1})\Big)\Big((f_h - f'_h)(S_h,A_h)\Big) \\ &- \Big((f_h - f'_h)(S_h,A_h)\Big)^2\Big].\end{aligned}$$

Now by defining a new function class $\mathcal{G}_f$ *depending on* f as $\mathcal{G}_f = \{f - f' : f'\in\mathcal{F}\}$, we arrive that

$$\frac{1}{2}\widehat{\mathcal{L}}_h^{\mathrm{ls}}(f_h) = \sup_{g_h\in\mathcal{G}_{f_h}} \widehat{\mathbb{E}}_{\mathbb{D}}\Big[\Big(f_h(S_h,A_h) - R_h - V_{h+1}(S_{h+1})\Big)g_h(S_h,A_h) - \frac{1}{2}g_h(S_h,A_h)^2\Big]. \tag{B.2}$$

This shares the same form as the minimax-typed loss $\sup_{g_h\in\mathbb{G}} \widehat{\Phi}_{\pi,h}^{1/2}(b_h, b_{h+1}; g_h)$ we consider in our work, see (3.10) in the main text. But still there are differences. In (B.2), the dual function g_h lies in a dual function class $\mathcal{G}_{f_h}$ which depends on the primal function f_h. While in our minimax-typed loss, the dual function class does not depends on the primal function.

Finally, we need to point out that even the two losses share the same form, the form of the confidence region considered by our work is different from that considered by Xie et al. (2021). To see this, still using the previous notations, the confidence region in Xie et al. (2021) (Equation (3.2)) becomes

$$\mathrm{CR}_h(\xi) = \Big\{f_h\in\mathcal{F} : \widehat{\mathcal{L}}_h^{\mathrm{ls}}(f_h)\le\xi\Big\}.$$

Meanwhile, if we reduce our confidence region to the above MDP setting, our confidence region should be in the form of

$$\mathrm{CR}_h(\xi) = \Big\{f_h\in\mathcal{F} : \widehat{\mathcal{L}}_h^{\mathrm{mm}}(f_h) - \min_{f_h\in\mathcal{F}}\widehat{\mathcal{L}}_h^{\mathrm{mm}}(f_h)\le\xi\Big\},$$

where $\mathcal{L}_h^{\mathrm{mm}}(f_h)$ denotes the minimax-typed-loss. Our algorithm and theoretical analysis are based on the second form of confidence region, which is key to the derivation of fast statistical rates for elements in the confidence region based on minimax estimation.

B.2 DISCUSSION ABOUT THE PARTIAL COVERAGE

More about the partial coverage (Assumption 4.1). Our work assumes the partial coverage of $\mathbb{D}$ according to Assumption 4.1, where we implicitly requires that $\mathcal{P}_h^{\pi}\left(S_h, \Gamma_{h-1}\right) / \mathcal{P}_h^{b}\left(S_h, \Gamma_{h-1}\right) < +\infty$ for all $\pi \in \Pi(\mathcal{H})$ (we call it the finite-ratio condition from here). We note that this finite-ratio condition can NOT be regarded as the full coverage assumption. Instead, this is a regularity condition that arises from causal inference.

First of all, the finite-ratio condition is different from the full coverage assumption in standard MDPs. The Full coverage assumption in standard MDPs usually takes the form that

$$\max_{\pi \in \Pi} \frac{\mathcal{P}_h^{\pi}(s,a)}{\mathcal{P}_h^{b}(s,a)} < C,$$

for some fixed $C > 0$. This condition means the density ratio of the marginal distributions of (s, a) between any target policy π and the behavior policy π^b is uniformly bounded by a constant. This condition (or some similar form) is a common and widely accepted form of full coverage in the MDP literature, e.g. (Chen and Jiang, 2019; Xie and Jiang, 2020). Note that this constant C is a uniform upper bound over the candidate policy class. Very importantly, this constant u' appears in the final error bound. The partial coverage assmuption in MDP, on the other hand, is commonly formulated as

$$\frac{\mathcal{P}_h^{\pi^\star}(s,a)}{\mathcal{P}_h^{b}(s,a)} < C,$$

This condition means the density ratio of the marginal distributions of (s, a) between only the optimal policy $\pi^\star$ and the behavior policy π^b, is bounded by a constant. The form of this assumption is very close to Assumption 4.1 (Partial coverage) in our paper. In other words, our Assumption 4.1 is a version of the partial coverage assumption that is tailored to the POMDP case. Notably, this constant C in the partial coverage assumption also appears in the final error bound.

As a sharp comparison to both the full coverage and partial coverage assumptions, the finite-ratio condition that the quantity $\mathcal{P}_h^{\pi}\left(S_h, \Gamma_{h-1}\right) / \mathcal{P}_h^{b}\left(S_h, \Gamma_{h-1}\right) < +\infty$ for all $\pi \in \Pi(\mathcal{H})$ does not result in any constant factor that appears in the final error bound. In the case of infinite policy class $\Pi(\mathcal{H})$, we can allow the ratio to be arbitrarily large and that won't hurt our final error bound. Therefore, this is not a coverage assumption. Our finite-ratio condition is a regularity condition that arises from causal inference. This condition is needed to deal with the extra challenge of the confounding issue in our POMP setting. In related works studying OPE under confounded POMDP (Shi et al., 2021), this finite-ratio condition is also needed. Overall, our paper is indeed under partial coverage and the finite ration condition is not a kind of coverage assumption.

B.3 POTENTIAL APPLICATION: REAL-WORLD EXAMPLE OF PROXIMAL CAUSAL INFERENCE IN RL.

Let us consider the real-world example of applying the POMDP model to sepsis treatment studied by Tsoukalas et al. (2015). In such an example, the state, action, observation, and reward of the POMDP are given by the following:

- State variable S_h refers to the clinical state of the patient, e.g., sepsis, SIRS, Bacteremia, etc.
- Observable variable O_h refers to all the information one can read from a medical device, such as the heart rate, the respiratory rate, blood pressure, blood test result of infection, etc.
- Action A_h refers to certain treatment given to the patient. For example, each antibiotic combination can be considered as an action. As mentioned in Tsoukalas et al. (2015), a total of 48 antibiotics have been included in the patient's remedy.
- Reward/cost values need to be provided empirically by physicians, based on the severity of each state. In the example of Tsoukalas et al. (2015), the states and their corresponding rewards/costs include: Healthy (100,000), No SIRS (50,000), Probable Sepsis (PS, 5000), SIRS (-50), Bacteremia (-10,000), etc.
- Finally, a history trajectory is the record of antibiotic treatment received by the patient. The behavior policy is some treatment plans that have been applied to some patients to generate the dataset.

When using reactive policies (Example 2.1), the negative control action variable (Z_h) is just the observation variable O_{h-1} which reflects the patient's clinical state at the last treatment time step, and the negative control outcome variable (W_h) is just the observation variable O_h at the current time step. Furthermore, when the observation O contains enough information to reflect the underlying state S, which basically implies a certain full rank assumption, we can then use Example C.1 to guarantee the existence of the bridge functions (See Appendix C).

C Proximal Causal Inference

In this Section, we complement the discussion of proximal causal inference in Section 3.1.

C.1 Illustration of Examples

In this subsection, we give detailed discussions for the three examples of history-dependent policies mentioned in Section 2.3. In particular, we give causal graphs of the POMDP when adopting these policies. Also, we explain the choice of negative control variables for these policies in Section 3.1.

C.1.1 Reactive Policy (Example 2.1 revisited)

When the target policy is a reactive policy, it only depends on the current observation O_h. That is, $\mathcal{H}_{h-1} = \{\varnothing\}$ and $\Gamma_{h-1} = \varnothing$ for each $h \in [H]$. The causal graph for such a target policy is shown in Figure 2. In this case, we choose the negative control action as $Z_h = O_{h-1}$ (node in green) and the negative control outcome as $W_h = O_h$ (node in yellow). By this choice, we can check the independence condition in Assumption 3.1 via Figure 2, i.e., under $\mathcal{P}^b$,

$$O_{h-1} \perp O_h, R_h, O_{h+1} \,|\, S_h, A_h \quad O_h \perp A_h, S_{h-1} \,|\, S_h.$$

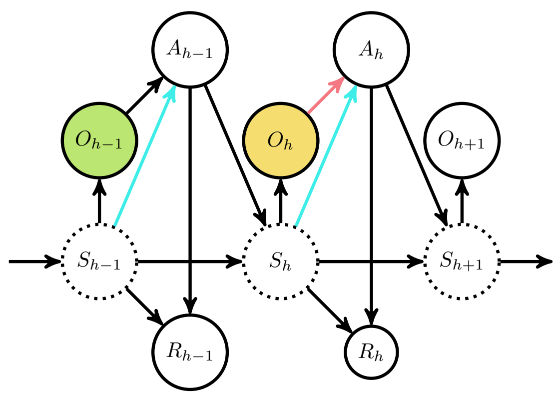

Figure 2: Causal graph for reactive policy. The dotted nodes indicate that the variables are not stored in the offline dataset. Solid arrows indicate the dependency among the variables. Specifically, The **red** arrows depict the dependence of the target policy on the observable variables. The **blue** arrows depict the dependence of the behavior policy on the latent state. The negative control action and outcome variables at the h-th step are filled in **green** and **yellow**, respectively.

C.1.2 Finite-history Policy (Example 2.2 revisited)

When the target policy a is finite-length history policy, it depends on the current observation and history of length at most k. That is, $\mathcal{H}_{h-1} = (\mathcal{O} \times \mathcal{A})^{\otimes \min\{k,h-1\}}$ for some $k \in \mathbb{N}$, $\Gamma_{h-1} = ((O_l, A_l), \cdots, (O_{h-1}, A_{h-1}))$ where the index $l = \max\{1, h-k\}$. The causal graph for such a target policy is shown in Figure 3. In this case, we choose the negative control action as $Z_h = O_{l-1}$ (node in green) and the negative control outcome as $W_h = O_h$ (node in yellow). By this choice, we can check the independence condition in Assumption 3.1 via Figure 3, i.e., under $\mathcal{P}^b$,

$$O_{l-1} \perp O_h, R_h, O_{h+1} \,|\, S_h, A_h, O_{h-1}, A_{h-1}, \cdots, O_l, A_l,$$
$$O_h \perp A_h, S_{h-1}, O_{h-1}, A_{h-1}, \cdots, O_l, A_l \,|\, S_h.$$

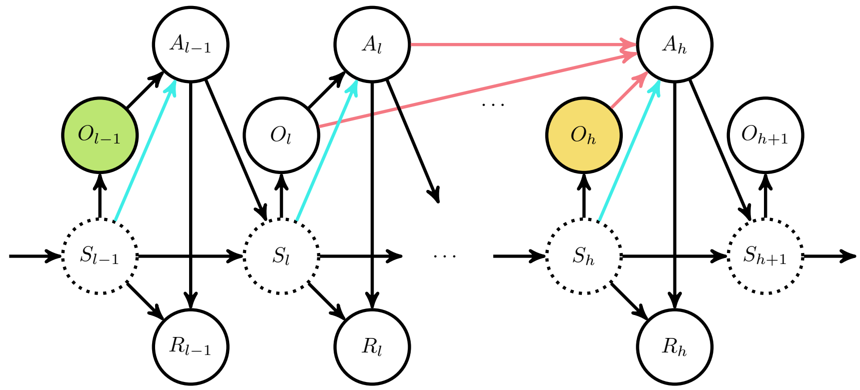

Figure 3: Causal graph for finite-length history policy. Index $l = \max\{1, h-k\}$. The dotted nodes indicate that the variables are not stored in the offline dataset. Solid arrows indicate the dependency among the variables. Specifically, The **red** arrows depict the dependence of the target policy on the observable variables. The **blue** arrows depict the dependence of the behavior policy on the latent state. The negative control action and outcome variables at step h are filled in **green** and **yellow**.

C.1.3 FULL-HISTORY POLICY (EXAMPLE 2.3 REVISITED)

When the target policy is a full-history policy, it depends on the current observation and the full history. That is, $\mathcal{H}_{h-1} = (\mathcal{O} \times \mathcal{A})^{\otimes(h-1)}$ and $\Gamma_{h-1} = ((O_1, A_1), \cdots, (O_{h-1}, A_{h-1}))$. The causal graph for such a target policy is shown in Figure 4. In this case, we choose the negative control action as $Z_h = O_0$ (node in green) and the negative control outcome as $W_h = O_h$ (node in yellow). By this choice, we can check the independence condition in Assumption 3.1 via Figure 4, i.e., under $\mathcal{P}^b$,

$$
\begin{aligned}
&O_0 \perp O_h, R_h, O_{h+1} \mid S_h, A_h, O_{h-1}, A_{h-1}, \cdots, O_1, A_1, \\
&O_h \perp A_h, S_{h-1}, O_{h-1}, A_{h-1}, \cdots, O_1, A_1 \mid S_h.
\end{aligned}
$$

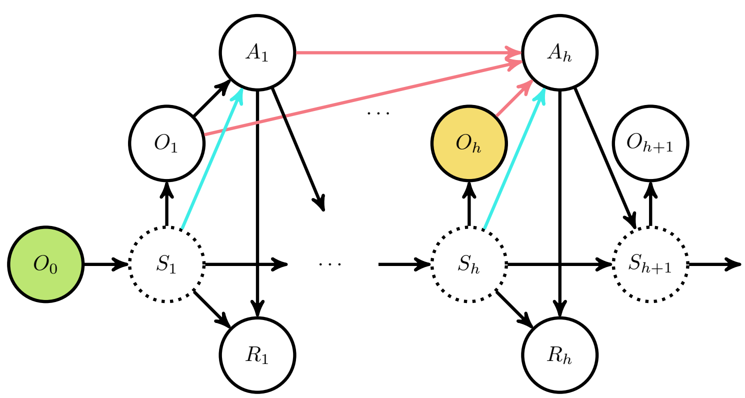

Figure 4: Causal graph for full-length history policy. The dotted nodes indicate that the variables are not stored in the offline dataset. Solid arrows indicate the dependency among the variables. Specifically, The **red** arrows depict the dependence of the target policy on the observable variables. The **blue** arrows depict the dependence of the behavior policy on the latent state. The negative control action and outcome variables at step h are filled in **green** and **yellow**, respectively.

C.2 EXAMINATION OF ASSUMPTION 3.2

In this subsection, we give concrete examples when the Assumption 3.2 holds, i.e., the confounding bridge functions exist.

Example C.1 (Example 2.1 revisited)**.** *For the tabular setting (i.e., $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{O}$ are finite spaces) and reactive policies (i.e., $\pi_h : \mathcal{O} \mapsto \Delta(\mathcal{A})$), the sufficient condition under which Assumption 3.2 holds is that*

$$\mathbf{rank}(\mathcal{P}_h^b(O_h|S_h)) = |\mathcal{S}|, \quad \mathbf{rank}(\mathcal{P}_h^b(O_{h-1}|S_h)) = |\mathcal{S}|, \tag{C.1}$$

where $\mathcal{P}_h(O_h|S_h)$ denote an $|\mathcal{S}| \times |\mathcal{O}|$ matrix whose (s, o)-th element is $\mathcal{P}_h^b(O_h = o|S_h = s)$, and $\mathcal{P}_h^b(O_{h-1}|S_h)$ is defined similarly.

Proof of Example C.1. Recall that for reactive policies, the history information $\Gamma_{h-1} = \varnothing$. We first show that under condition (C.1), there exist functions $\{b_h^\pi\}_{h=1}^H$ and $\{q_h^\pi\}_{h=1}^H$ which solve the following equations

$$\begin{aligned}
&\mathbb{E}_{\pi^b}\left[b_h^\pi(A_h, O_h)|A_h, S_h\right] \\
&\qquad = \mathbb{E}_{\pi^b}\Big[R_h\pi_h(A_h|O_h) + \gamma\sum_{a'} b_{h+1}^\pi(a', O_{h+1})\pi_h(A_h|O_h)\Big|A_h, S_h\Big],
\end{aligned} \tag{C.2}$$

$$\mathbb{E}_{\pi^b}\left[q_h^\pi(A_h, O_{h-1})|A_h, S_h\right] = \frac{\mu_h(S_h)}{\pi_h^b(A_h|S_h)}, \tag{C.3}$$

Then we show that the solutions to (C.2) and (C.3) also solve (3.2) and (3.3). The difference between (C.2) and (3.2) is that in (C.2) we condition on the latent state S_h rather than the observable negative control variable Z_h. In related literature (Bennett and Kallus, 2021; Shi et al., 2021), the solutions to (C.2) and (C.3) are referred to as *unlearnable bridge functions*.

We first show the existence of $\{b_h^\pi\}_{h=1}^H$ in a backward manner. Denote by b_{H+1}^π a zero function. Suppose that b_{h+1}^π exists, we show that b_h^π also exists. Since now spaces $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{O}$ are discrete, we adopt the notation of matrix. In particular, we denote by

$$\begin{aligned}
\mathbf{B} &\in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{O}|}, \qquad \mathbf{B}(a, o) = b_h(a, o), \\
\mathbf{O} &\in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{O}|}, \qquad \mathbf{O}(s, o) = \mathcal{P}_h^b(O_h = o|S_h = s),
\end{aligned}$$

$$\mathbf{R} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{S}|},\ \mathbf{R}(s, a) = \mathbb{E}_{\pi^b}\Big[R_h\pi_h(A_h|O_h) + \gamma\sum_{a'} b_{h+1}^\pi(a', O_{h+1})\pi_h(A_h|O_h)\Big|A_h = a, S_h = s\Big].$$

The existence of b_h^π satisfying (C.2) is equivalent to the existence of $\mathbf{B}$ solving the matrix equation

$$\mathbf{B}\,\mathbf{O}^\top = \mathbf{R}. \tag{C.4}$$

By condition (C.1), we known that the matrix $\mathbf{O}^\top$ is of full column rank, which indicates that (C.4) admits a solution $\mathbf{B}$. This proves the existence of b_h^π. For $\{q_h^\pi\}_{h=1}^H$, we use a similar method by considering

$$\begin{aligned}
\mathbf{Q} &\in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{O}|}, \qquad \mathbf{Q}(a, o) = q_h(a, o), \\
\mathbf{O}_- &\in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{O}|}, \qquad \mathbf{O}_-(s, o) = \mathcal{P}_h^b(O_{h-1} = o|S_h = s), \\
\mathbf{C} &\in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{S}|}, \qquad \mathbf{C}(s, a) = \frac{\mu_h(S_h = s)}{\pi_h^b(A_h = a|S_h = s)}.
\end{aligned}$$

The existence of q_h^π satisfying (C.3) is equivalent to the existence of $\mathbf{Q}$ solving the matrix equation

$$\mathbf{Q}\,\mathbf{O}_-^\top = \mathbf{C} \tag{C.5}$$

By condition (C.1), we known that the matrix $\mathbf{O}_-^\top$ is of full column rank, which indicates that (C.5) admits a solution $\mathbf{Q}$. This proves the existence of q_h^π. Thus we have shown that there exists $\{b_h^\pi\}_{h=1}^H$ and $\{q_h^\pi\}_{h=1}^H$ which solve equation (C.2) and (C.3). Finally, it holds that any solution to (C.2) and (C.3) also forms a solution to (3.2) and (3.3), which has been shown in Theorem 11 in Shi et al. (2021). This finishes the proof of Example C.1. ☐

Example C.2 (Example 2.2 revisited)**.** *For the tabular setting and finite length policies (i.e., $\pi_h : \mathcal{O} \times (\mathcal{O} \times \mathcal{A})^{\min\{k,h-1\}} \mapsto \Delta(\mathcal{A})$), the sufficient condition under which Assumption 3.2 holds is that, for any action $a \in \mathcal{A}$,*

$$\mathbf{rank}(\mathcal{P}_h^b(O_h|A_h = a, O_{h-k-1})) = |\mathcal{O}|, \quad \mathbf{rank}(\mathcal{P}_h^b(O_{h-k-1}|A_h = a, S_h, \Gamma_{h-1})) = |\mathcal{O}|, \tag{C.6}$$

where $\mathcal{P}_h^b(O_h|A_h = a, O_{h-k-1})$ is a $|\mathcal{O}| \times |\mathcal{O}|$ matrix with (o, o')-th element is $\mathcal{P}_h^b(O_h = o|A_h = a, O_{h-k-1} = o')$ and $\mathcal{P}_h^b(O_{h-k-1}|A_h = a, S_h, \Gamma_{h-1})$ is a $|\mathcal{S}||\mathcal{H}_{h-1}| \times |\mathcal{O}|$ matrix defined similarly.

Proof of Example C.2. To see this, we first prove the existence of $\{b_n^\pi\}$. For simplicity, we denote by

$$\mathbf{P}_a = \left(\mathcal{P}_h^b\left(O_h \mid A_h = a, O_{h-k-1}\right)\right) \in \mathbb{R}^{|\mathcal{O}|\times|\mathcal{O}|}$$

for each $a \in \mathcal{A}$. Also, we denote that

$$\mathbf{B}_a = \left(b_h\left(a, O_h\right)\right) \in \mathbb{R}^{|\mathcal{O}|\times 1},$$

$$\mathbf{R}_a = \left(\mathbb{E}_{\pi^b}\left[R_h \pi_h\left(A_h \mid O_h\right) + \gamma \sum_{a'} b_{h+1}^\pi\left(a', O_{h+1}\right)\pi_h\left(A_h \mid O_h\right) \mid A_h = a, O_{h-k-1}\right]\right) \in \mathbb{R}^{|\mathcal{O}|\times 1}.$$

Then for each $a \in \mathcal{A}$, the existence of $b_n^\pi(a, \cdot)$ is equivalent to the existence of the solution to

$$\mathbf{P}^a \mathbf{B}^a = \mathbf{R}^a.$$

Such a linear equation admits a solution due to our assumption on the matrix $\mathbf{P}_a$. This shows the existence of $\{b_h^\pi\}$. For $\{q_h^\pi\}$, the deduction is similar by considering for each $a \in \mathcal{A}$,

$$\mathbf{T}_a = \left(\mathcal{P}_h^b\left(O_{h-k-1} \mid A_h = a, S_h, \Gamma_{h-1}\right)\right) \in \mathbb{R}^{|\mathcal{S}||\mathcal{H}_{h-1}|\times|\mathcal{O}|},$$

$$\mathbf{Q}_a = \left(q_h\left(a, O_{h-k-1}\right)\right) \in \mathbb{R}^{|\mathcal{O}|\times 1},$$

$$\mathbf{C}_a = \left(\frac{\mu_h\left(S_h, \Gamma_{h-1}\right)}{\pi^b\left(a \mid S_h\right)}\right) \in \mathbb{R}^{|\mathcal{S}||\mathcal{H}_{h-1}|\times 1}.$$

By considering the equation that

$$\mathbf{T}_a \mathbf{Q}_a = \mathbf{C}_a$$

and using the full rank assumption on matrix $\mathbf{T}_a$, we can obtain the existence of $\{q_h^\pi\}$. This finishes the proof of Example C.2. □

D Proof Sketches of Main Theoretical Result

In this section, we sketch the proof of the main theoretical result Theorem 4.4, and we refer to Appendix G for a detailed proof. For simplicity, we denote that for any $\pi \in \Pi(\mathcal{H})$ and $\mathbf{b} \in \mathbb{B}^{\otimes H}$,

$$F(\mathbf{b}) := \mathbb{E}_{\pi^b}\left[\sum_{a\in\mathcal{A}} b_1(a, W_1)\right], \quad \widehat{F}(\mathbf{b}) := \widehat{\mathbb{E}}_{\pi^b}\left[\sum_{a\in\mathcal{A}} b_1(a, W_1)\right]. \tag{D.1}$$

By the definition (D.1) and Theorem 3.3, for any policy $\pi \in \Pi(\mathcal{H})$, it holds that $J(\pi) = F(\mathbf{b}^\pi)$, where we have denoted by $\mathbf{b}^\pi = (b_1^\pi, \cdots, b_H^\pi)$ the vector of true value bridge functions of π which are given in (3.2).

Our proof to Theorem 4.4 relies on the following three key lemmas. The first lemma relates the different values of mapping $F(\cdot)$ induced by a true value bridge function $\mathbf{b}^\pi$ and any other functions $\mathbf{b} \in \mathbb{B}^{\otimes H}$ to the RMSE loss which we aim to minimize by algorithm design. This indeed decomposes the suboptimality (2.2).

Lemma D.1 (Suboptimality decomposition)**.** *Under Assumption 3.1, 3.2, for any policy $\pi \in \Pi(\mathcal{H})$ and $\mathbf{b} \in \mathbb{B}^{\otimes H}$, it holds that*

$$F(\mathbf{b}^\pi) - F(\mathbf{b}) \leq \sum_{h=1}^{H} \gamma^{h-1}\sqrt{C^\pi} \cdot \sqrt{\mathcal{L}_h^\pi(b_h, b_{h+1})},$$

where the concentrability coefficient C^π is defined as $C^\pi := \sup_{h\in[H]} \mathbb{E}_{\pi^b}\left[(q_h^\pi(A_h, Z_h))^2\right]$.

Proof of Lemma D.1. See Appendix F.1 for a detailed proof. □

The following two lemmas characterize the theoretical properties of the confidence region $\mathrm{CR}^\pi(\xi)$. Specifically, Lemma D.2 shows that with high probability the confidence region of π contains the true value bridge function $\mathbf{b}^\pi$. Besides, Lemma D.3 shows that each bridge function vector $\mathbf{b} \in \mathrm{CR}^\pi(\xi)$ enjoys a fast statistical rate (Uehara et al., 2021) for its RMSE loss $\mathcal{L}_h^\pi$ defined in (3.6). To obtain such a fast rate, we develop novel proof techniques in Appendix F.3.

Lemma D.2 (Validity of confidence regions). *Under Assumption 3.2 and 4.2, for any $0 < \delta < 1$, by setting*

$$\xi = C_1(\lambda + 1/\lambda) \cdot M_{\mathbb{B}}^2 \cdot M_{\mathbb{G}}^2 \cdot \log(|\mathbb{B}||\Pi(\mathcal{H})|H/\zeta)/n,$$

for some problem-independent universal constant $C_1 > 0$ and $\zeta = \min\{\delta, 4c_1 \exp(-c_2 n \alpha_{\mathbb{G},n}^2)\}$, it holds with probability at least $1-\delta$ that $\mathbf{b}^\pi \in \mathrm{CR}^\pi(\xi)$ for any policy $\pi \in \Pi(\mathcal{H})$.

Proof of Lemma D.2. See Appendix F.2 for a detailed proof. □

Lemma D.3 (Accuracy of confidence regions). *Under Assumption 3.2, 4.2, and 4.3, by setting the same ξ as in Lemma D.2, with probability at least $1-\delta/2$, for any policy $\pi \in \Pi(\mathcal{H})$, $\mathbf{b} \in \mathrm{CR}^\pi(\xi)$, and step h,*

$$\sqrt{\mathcal{L}_h^\pi(b_h, b_{h+1})} \le \widetilde{C}_1 M_{\mathbb{B}} M_{\mathbb{G}} \sqrt{(\lambda + 1/\lambda) \cdot \log(|\mathbb{B}||\Pi(\mathcal{H})|H/\zeta)/n} + \widetilde{C}_1 \epsilon_{\mathbb{B}}^{1/4} M_{\mathbb{G}}^{1/2},$$

for some problem-independent universal constant $\widetilde{C}_1 > 0$, and $\zeta = \min\{\delta, 4c_1 \exp(-c_2 n \alpha_{\mathbb{G},n}^2)\}$.

Proof of Lemma D.3. See Appendix F.3 for a detailed proof. □

When $\alpha_{\mathbb{G},n} \in \mathcal{O}(n^{-1/2})$ and $\epsilon_{\mathbb{B}} = 0$, Lemma D.3 implies that $\mathcal{L}_h^\pi(b_h, b_{h+1}) \le \widetilde{\mathcal{O}}(n^{-1})$. Now with Lemma D.1, Lemma D.2, and Lemma D.3, by the choice of $\widehat{\pi}$ in `P3O`, we can show that

$$\begin{aligned}
J(\pi^\star) - J(\widehat{\pi}) &\le \widetilde{\mathcal{O}}(n^{-1/2}) + \max_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} F(\mathbf{b}) - \min_{\mathbf{b} \in \mathrm{CR}^{\widehat{\pi}}(\xi)} F(\mathbf{b}) \\
&\le \widetilde{\mathcal{O}}(n^{-1/2}) + \max_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} F(\mathbf{b}) - \min_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} F(\mathbf{b}) \\
&\le \widetilde{\mathcal{O}}(n^{-1/2}) + 2 \max_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} \left| F(\mathbf{b}) - F(\mathbf{b}^{\pi^\star}) \right| \\
&\le \widetilde{\mathcal{O}}(n^{-1/2}) + 2 \max_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} \sum_{h=1}^{H} \gamma^{h-1} \sqrt{C^{\pi^\star}} \cdot \sqrt{\mathcal{L}_h^{\pi^\star}(b_h, b_{h+1})},
\end{aligned} \tag{D.2}$$

where the first inequality holds by Lemma D.2, the second inequality holds from the optimality of $\widehat{\pi}$ in Algorithm 1, the third inequality holds directly, and the last inequality holds by Lemma D.1. Finally, by applying Lemma D.3 to the right hand side of (D.2), we conclude the proof of Theorem 4.4.

E PROOF OF THEOREM 3.3

Proof of Theorem 3.3. For any step h, we denote $J_h(\pi) := \mathbb{E}_\pi[R_h(S_h, A_h)]$. We have that

$$\begin{aligned}
J_h(\pi) &= \mathbb{E}_\pi[R_h(S_h, A_h)] \\
&= \mathbb{E}_\pi\left[\mathbb{E}_\pi[R_h(S_h, A_h)|O_h, S_h, \Gamma_{h-1}]\right] \\
&= \mathbb{E}_\pi\left[\sum_{a \in \mathcal{A}} R_h(S_h, a)\pi_h(a|O_h, \Gamma_{h-1})\right] \\
&= \mathbb{E}_\pi\left[\mathbb{E}_\pi\left[\sum_{a \in \mathcal{A}} R_h(S_h, a)\pi_h(a|O_h, \Gamma_{h-1}) \middle| S_h, \Gamma_{h-1}\right]\right],
\end{aligned}$$

where the second and the last equality follows from the tower property of conditional expectation. Using the definition of density ratio $\mu_h(S_h, \Gamma_{h-1})$ in Assumption 3.2, we can change the outer expectation to $\mathbb{E}_{\pi^b}$ by

$$
\begin{aligned}
J_h(\pi) &= \mathbb{E}_{\pi^b}\left[\mu_h(S_h,\Gamma_{h-1})\cdot\mathbb{E}_{\pi}\left[\sum_{a\in\mathcal{A}}R_h(S_h,a)\pi_h(a|O_h,\Gamma_{h-1})\middle|S_h,\Gamma_{h-1}\right]\right],\\
&= \mathbb{E}_{\pi^b}\left[\sum_{a\in\mathcal{A}}R_h(S_h,a)\cdot\pi_h(a|O_h,\Gamma_{h-1})\cdot\mu_h(S_h,\Gamma_{h-1})\right]\\
&= \mathbb{E}_{\pi^b}\left[\sum_{a\in\mathcal{A}}\pi_h^b(a|S_h)\cdot R_h(S_h,a)\cdot\frac{\pi_h(a|O_h,\Gamma_{h-1})}{\pi_h^b(a|S_h)}\cdot\mu_h(S_h,\Gamma_{h-1})\right]\\
&\stackrel{\text{(a)}}{=} \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[R_h(S_h,A_h)\cdot\frac{\pi_h(A_h|O_h,\Gamma_{h-1})}{\pi_h^b(A_h|S_h)}\cdot\mu_h(S_h,\Gamma_{h-1})\middle|S_h,O_h,\Gamma_{h-1}\right]\right]\\
&= \mathbb{E}_{\pi^b}\left[R_h(S_h,A_h)\cdot\pi_h(A_h|O_h,\Gamma_{h-1})\cdot\frac{\mu_h(S_h,\Gamma_{h-1})}{\pi_h^b(A_h|S_h)}\right],
\end{aligned}
$$

where step (a) follows from the fact that $A_h\sim\pi_h^b(\cdot|S_h)$ and satisfies $A_h\perp O_h,\Gamma_{h-1}|S_h$ under π^b. Now using the definition (3.3) of weight bridge function q_h^π in Assumption 3.2, we have that

$$
\begin{aligned}
J_h(\pi) &= \mathbb{E}_{\pi^b}\left[R_h(S_h,A_h)\cdot\pi_h(A_h|O_h,\Gamma_{h-1})\cdot\mathbb{E}_{\pi^b}\left[q_h^\pi(A_h,Z_h)|S_h,A_h,\Gamma_{h-1}\right]\right]\\
&\stackrel{\text{(a)}}{=} \mathbb{E}_{\pi^b}\left[R_h(S_h,A_h)\cdot\pi_h(A_h|O_h,\Gamma_{h-1})\cdot q_h^\pi(A_h,Z_h)\right]\\
&= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[R_h(S_h,A_h)\cdot\pi_h(A_h|O_h,\Gamma_{h-1})\cdot q_h^\pi(A_h,Z_h)|A_h,Z_h\right]\right]\\
&= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[R_h(S_h,A_h)\cdot\pi_h(A_h|O_h,\Gamma_{h-1})\cdot|A_h,Z_h\right]q_h^\pi(A_h,Z_h)\right],
\end{aligned}
$$

where step (a) follows from the assumption that $Z_h\perp O_h,R_h|S_h,A_h,\Gamma_{h-1}$ by Assumption 3.1. Now using the definition (3.2) of value bridge function b_h^π in Assumption 3.2, we have that

$$
\begin{aligned}
J_h(\pi) &= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[b_h^\pi(A_h,W_h)-\gamma\sum_{a'\in\mathcal{A}}b_{h+1}^\pi(a',W_{h+1})\pi_h(A_h|O_h,\Gamma_{h-1})\middle|A_h,Z_h\right]q_h^\pi(A_h,Z_h)\right]\\
&= \mathbb{E}_{\pi^b}\left[f(S_h,A_h,O_h,W_h,W_{h+1},\Gamma_{h-1})\cdot q_h^\pi(A_h,Z_h)\right]\\
&= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[f(S_h,A_h,O_h,W_h,W_{h+1},\Gamma_{h-1})\cdot q_h^\pi(A_h,Z_h)|S_h,A_h,O_h,W_h,W_{h+1},\Gamma_{h-1}\right]\right],\\
&= \mathbb{E}_{\pi^b}\left[f(S_h,A_h,O_h,W_h,W_{h+1},\Gamma_{h-1})\cdot\mathbb{E}_{\pi^b}\left[q_h^\pi(A_h,Z_h)|S_h,A_h,O_h,W_h,W_{h+1},\Gamma_{h-1}\right]\right],\\
&\stackrel{\text{(a)}}{=} \mathbb{E}_{\pi^b}\left[f(S_h,A_h,O_h,W_h,W_{h+1},\Gamma_{h-1})\cdot\mathbb{E}_{\pi^b}\left[q_h^\pi(A_h,Z_h)|S_h,A_h,\Gamma_{h-1}\right]\right],
\end{aligned}
$$

where for simplicity we have denoted that

$$
f(S_h,A_h,O_h,W_h,W_{h+1},\Gamma_{h-1}) = b_h^\pi(A_h,W_h)-\gamma\sum_{a'\in\mathcal{A}}b_{h+1}^\pi(a',W_{h+1})\pi_h(A_h|O_h,\Gamma_{h-1}),
$$

and step (a) follows from the assumption that $Z_h\perp O_h,W_h,W_{h+1}|S_h,A_h,\Gamma_{h-1}$ by Assumption 3.1. By the definition (3.3) of weight bridge function q_h^π in Assumption 3.2 again, we have that

$$
\begin{aligned}
J_h(\pi) &= \mathbb{E}_{\pi^b}\left[f(S_h,A_h,O_h,W_h,W_{h+1},\Gamma_{h-1})\cdot\frac{\mu_h(S_h,\Gamma_{h-1})}{\pi_h^b(A_h|S_h)}\right]\\
&\stackrel{\text{(a)}}{=} \mathbb{E}_{\pi^b}\left[\left(b_h^\pi(A_h,W_h)-\gamma\sum_{a'\in\mathcal{A}}b_{h+1}^\pi(a',W_{h+1})\pi_h(A_h|O_h,\Gamma_{h-1})\right)\cdot\frac{\mu_h(S_h,\Gamma_{h-1})}{\pi_h^b(A_h|S_h)}\right],
\end{aligned}
$$

where step (a) just applies the definition of f. Now sum $J_h(\pi)$ over $h\in[H]$, we have that

$$
J(\pi) = \sum_{h=1}^{H}\gamma^{h-1}J_h(\pi) = \underbrace{\mathbb{E}_{\pi^b}\left[\frac{\mu_1(S_1,\Gamma_0)}{\pi_1^b(A_1|S_1)}b_1^\pi(A_1,W_1)\right]}_{\text{(A)}} + \underbrace{\sum_{h=2}^{H}\gamma^{h-1}\Delta_h}_{\text{(B)}}, \tag{E.1}
$$

where for simplicity we define Δ_h for $h=2,\cdots,H$ as

$$
\Delta_h = \mathbb{E}_{\pi^b}\left[\frac{\mu_h(S_h,\Gamma_{h-1})}{\pi_h^b(A_h|S_h)}b_h^\pi(A_h,W_h)-\frac{\mu_{h-1}(S_{h-1},\Gamma_{h-2})}{\pi_{h-1}^b(A_{h-1}|S_{h-1})}\cdot\sum_{a'\in\mathcal{A}}b_h^\pi(a',W_h)\pi_{h-1}(A_{h-1}|O_{h-1},\Gamma_{h-1})\right].
$$

In the sequel, we deal with term (A) and (B) respectively. On the one hand, we have that

$$
\begin{aligned}
\text{(A)} &\overset{\text{(a)}}{=} \mathbb{E}_{\pi^b}\left[\frac{\mathcal{P}_1^\pi(S_1,\Gamma_0)}{\mathcal{P}_1^b(S_1,\Gamma_0)\pi_1^b(A_1|S_1)} b_1^\pi(A_1,W_1)\right] \\
&\overset{\text{(b)}}{=} \mathbb{E}_{\pi^b}\left[\frac{1}{\pi_1^b(A_1|S_1)} b_1^\pi(A_1,W_1)\right] \\
&= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\frac{1}{\pi_1^b(A_1|S_1)} b_1^\pi(A_1,W_1)\middle| S_1, W_1\right]\right] \\
&\overset{\text{(c)}}{=} \mathbb{E}_{\pi^b}\left[\sum_{a\in\mathcal{A}} \frac{\pi_1^b(a|S_1)}{\pi_1^b(a|S_1)} b_1^\pi(a,W_1)\right] \\
&= \mathbb{E}_{\pi^b}\left[\sum_{a\in\mathcal{A}} b_1^\pi(a,W_1)\right],
\end{aligned}
$$

where step (a) follows from the definition of $\mu_1(S_1,\Gamma_0)$ in Assumption 3.2, step (b) follows from the fact that at $h=1$, $\mathcal{P}_1^b(S_1,\Gamma_0)=\mathcal{P}_1^\pi(S_1,\Gamma_0)$, and step (c) follows from the assumption that $A_1 \perp W_1|S_1$ by Assumption 3.1. On the other hand, term (b) in (E.1) is actually 0, which we show by proving that $\Delta_h = 0$ for any $h \geq 2$. We denote by $\Delta_h = \Delta_h^1 - \Delta_h^2$ and we consider Δ_h^1 and Δ_h^2 respectively, where

$$
\begin{aligned}
\Delta_h^1 &= \mathbb{E}_{\pi^b}\left[\frac{\mu_h(S_h,\Gamma_{h-1})}{\pi_h^b(A_h|S_h)} \cdot b_h^\pi(A_h,W_h)\right], \\
\Delta_h^2 &= \mathbb{E}_{\pi^b}\left[\frac{\mu_{h-1}(S_{h-1},\Gamma_{h-2})}{\pi_{h-1}^b(A_{h-1}|S_{h-1})} \cdot \sum_{a'\in\mathcal{A}} b_h^\pi(a',W_h)\pi_{h-1}(A_{h-1}|O_{h-1},\Gamma_{h-1})\right].
\end{aligned}
$$

In the sequel, we prove that $\Delta_h^1 = \Delta_h^2$ for the three cases of T_h in Example 2.1, 2.2, and 2.3, respectively.

Case 1: Reactive policy (Example 2.1). We first focus on the simple case when policy π is reactive. Since for reactive policies $T_h = \varnothing$, we can equivalently write $\mu_h(S_h,\Gamma_{h-1})$ as $\mu_h(S_h) = \mathcal{P}_h^\pi(S_h)/\mathcal{P}_h^b(S_h)$. Now for Δ_h^1, we can rewrite it as

$$
\begin{aligned}
\Delta_h^1 &= \mathbb{E}_{\pi^b}\left[\frac{\mathcal{P}_h^\pi(S_h)}{\mathcal{P}_h^b(S_h)\pi_h^b(A_h|S_h)} \cdot b_h^\pi(A_h,W_h)\right] \\
&\overset{\text{(a)}}{=} \int_{\mathcal{S}} \cancel{\mathcal{P}_h^b(s_h)}\mathrm{d}s_h \sum_{a_h\in\mathcal{A}} \cancel{\pi_h^b(a_h|s_h)} \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h,a_h)\mathrm{d}w_h \cdot \frac{\mathcal{P}_h^\pi(s_h)}{\cancel{\mathcal{P}_h^b(s_h)}\cancel{\pi_h^b(a_h|s_h)}} b_h^\pi(a_h,w_h) \\
&\overset{\text{(b)}}{=} \sum_{a_h\in\mathcal{A}} \int_{\mathcal{S}} \mathcal{P}_h^\pi(s_h)\mathrm{d}s_h \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h)\mathrm{d}w_h \cdot b_h^\pi(a_h,w_h).
\end{aligned}
$$

Here step (a) expands the expectation by using integral against corresponding density functions, and step (b) follows from cancelling the same terms and the fact that $W_h \perp A_h|S_h$ under Assumption 3.1. For Δ_h^2, we can also rewrite it as

$$
\begin{aligned}
\Delta_h^2 &= \mathbb{E}_{\pi^b}\left[\frac{\mathcal{P}_{h-1}^\pi(S_{h-1})\pi_{h-1}(A_{h-1}|O_{h-1})}{\mathcal{P}_{h-1}^b(S_{h-1})\pi_{h-1}^b(A_{h-1}|S_{h-1})} \cdot \sum_{a'\in\mathcal{A}} b_h^\pi(a',W_h)\right] \\
&\overset{\text{(a)}}{=} \int_{\mathcal{S}} \cancel{\mathcal{P}_{h-1}^b(s_{h-1})}\mathrm{d}s_{h-1} \int_{\mathcal{O}} \mathbb{O}_{h-1}(o_{h-1}|s_{h-1})\mathrm{d}o_{h-1} \sum_{a_{h-1}\in\mathcal{A}} \cancel{\pi_{h-1}^b(a_{h-1}|s_{h-1})} \int_{\mathcal{S}} \mathbb{P}_h(s_h|s_{h-1},a_{h-1})\mathrm{d}s_h \\
&\qquad \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h,s_{h-1},a_{h-1},o_{h-1}) \cdot \frac{\mathcal{P}_{h-1}^\pi(s_{h-1})\pi_{h-1}(a_{h-1}|o_{h-1})}{\cancel{\mathcal{P}_{h-1}^b(s_{h-1})}\cancel{\pi_{h-1}^b(a_{h-1}|s_{h-1})}} \sum_{a_h\in\mathcal{A}} b_h^\pi(a_h,w_h)\mathrm{d}w_h.
\end{aligned}
$$

Here step (a) follows from expanding the expectation. It follows that

$$\begin{aligned}\Delta_h^2 &\overset{\text{(b)}}{=} \sum_{a_h \in \mathcal{A}} \int_{\mathcal{S}} \mathcal{P}_{h-1}^{\pi}(s_{h-1}) \mathrm{d}s_{h-1} \int_{\mathcal{O}} \mathbb{O}_{h-1}(o_{h-1}|s_{h-1}) \mathrm{d}o_{h-1} \sum_{a \in \mathcal{A}} \pi_{h-1}(a_{h-1}|o_{h-1}) \\ &\qquad \int_{\mathcal{S}} \mathbb{P}_h(s_h|s_{h-1}, a_{h-1}) \mathrm{d}s' \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h) \cdot b_h^{\pi}(a_h, w_h) \\ &\overset{\text{(c)}}{=} \sum_{a_h \in \mathcal{A}} \int_{\mathcal{S}} \mathcal{P}_h^{\pi}(s_h) \mathrm{d}s_h \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h) \cdot b_h^{\pi}(a_h, w_h) \mathrm{d}w_h.\end{aligned}$$

Here step (b) follows from cancelling the same terms and using the fact that $W_h \perp S_{h-1}, A_{h-1}, O_{h-1}|S_h$ by Assumption 3.1, and step (d) follows by marginalizing over $S_{h-1}, A_{h-1}, O_{j-1}$. Thus we have proved that $\Delta_h^1 = \Delta_h^2$ for reactive policies and consequently $\Delta_h = \Delta_h^1 - \Delta_h^2 = 0$.

Case 2: Finite-history policy (Example 2.2). Now we have that $\Gamma_{h-1} \cup \{A_h, O_h\} = \{A_{l-1}, O_{l-1}\} \cup T_h$, where the index $l = \max\{0, h-k\}$. Similarly, we can first rewrite Δ_h^1 as

$$\begin{aligned}\Delta_h^1 &= \mathbb{E}_{\pi^b}\left[\frac{\mathcal{P}_h^{\pi}(S_h, \Gamma_{h-1})}{\mathcal{P}_h^b(S_h, \Gamma_{h-1}) \pi_h^b(A_h|S_h)} b_h^{\pi}(A_h, W_h)\right] \\ &\overset{\text{(a)}}{=} \int_{\mathcal{S} \times \mathcal{H}_{h-1}} \mathcal{P}_h^b(s_h, \tau_{h-1}) \mathrm{d}s_h \mathrm{d}\tau_{h-1} \sum_{a_h \in \mathcal{A}} \pi_h^b(a_h|s_h) \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h, a_h, \tau_{h-1}) \mathrm{d}w_h \\ &\qquad \cdot \frac{\mathcal{P}_h^{\pi}(s_h, \tau_{h-1})}{\mathcal{P}_h^b(s_h, \tau_{h-1}) \pi_h^b(a_h|s_h, \tau_{h-1})} b_h^{\pi}(a_h, w_h) \\ &\overset{\text{(b)}}{=} \sum_{a_h \in \mathcal{A}} \int_{\mathcal{S} \times \mathcal{H}_{h-1}} \mathcal{P}_h^{\pi}(s_h, \tau_{h-1}) \mathrm{d}s_h \mathrm{d}\tau_{h-1} \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h) \mathrm{d}w_h \cdot b_h^{\pi}(a_h, w_h).\end{aligned}$$

Here step (a) follows from expanding the expectation, and step (b) follows from cancelling the same terms and using the fact that $W_h \perp A_h, \Gamma_{h-1}|S_h$ under Assumption 3.1. For Δ_h^2, we can also rewrite it as

$$\begin{aligned}\Delta_h^2 &= \mathbb{E}_{\pi^b}\left[\frac{\mathcal{P}_{h-1}^{\pi}(S_{h-1}, \Gamma_{h-2}) \pi_{h-1}(A_{h-1}|O_{h-1})}{\mathcal{P}_{h-1}^b(S_{h-1}, \Gamma_{h-2}) \pi_{h-1}^b(A_{h-1}|S_{h-1}, \Gamma_{h-2})} \sum_{a' \in \mathcal{A}} b_h^{\pi}(a', W_h)\right] \\ &\overset{\text{(a)}}{=} \int_{\mathcal{S} \times \mathcal{H}_{h-2}} \cancel{\mathcal{P}_{h-1}^b(s_{h-1}, \tau_{h-2})} \mathrm{d}s_{h-1} \mathrm{d}\tau_{h-2} \int_{\mathcal{O}} \mathbb{O}_{h-1}(o_{h-1}|s_{h-1}) \mathrm{d}o_{h-1} \sum_{a_{h-1} \in \mathcal{A}} \cancel{\pi_{h-1}^b(a_{h-1}|s_{h-1})} \\ &\qquad \int_{\mathcal{S}} \mathbb{P}_h(s_h|s_{h-1}, a_{h-1}) \mathrm{d}s_h \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h, s_{h-1}, a_{h-1}, o_{h-1}, \tau_{h-2}) \\ &\qquad \cdot \frac{\mathcal{P}_{h-1}^{\pi}(s_{h-1}, \tau_{h-2}) \pi_{h-1}(a_{h-1}|o_{h-1}, \tau_{h-2})}{\cancel{\mathcal{P}_{h-1}^b(s_{h-1}, \tau_{h-2})} \cancel{\pi_{h-1}^b(a_{h-1}|s_{h-1})}} \sum_{a_h \in \mathcal{A}} b_h^{\pi}(a_h, w_h) \\ &\overset{\text{(b)}}{=} \sum_{a_h \in \mathcal{A}} \int_{\mathcal{S} \times \mathcal{H}_{h-2}} \mathcal{P}_{h-1}^{\pi}(s_{h-1}, \widetilde{\tau}_{h-2}, a_l, o_l) \mathrm{d}s_{h-1} \mathrm{d}\widetilde{\tau}_{h-2} \mathrm{d}a_l \mathrm{d}o_l \int_{\mathcal{O}} \mathbb{O}_{h-1}(o_{h-1}|s_{h-1}) \mathrm{d}o_{h-1} \\ &\qquad \sum_{a_{h-1} \in \mathcal{A}} \pi_{h-1}(a_{h-1}|o_{h-1}, \tau_{h-2}) \int_{\mathcal{S}} \mathbb{P}_h(s_h|s_{h-1}, a_{h-1}) \mathrm{d}s_h \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h) \cdot b_h^{\pi}(a_h, w_h) \end{aligned} \tag{E.2}$$

$$\overset{\text{(c)}}{=} \sum_{a_h \in \mathcal{A}} \int_{\mathcal{S} \times \mathcal{H}_{h-1}} \mathcal{P}_h^{\pi}(s_h, \tau_{h-1}) \mathrm{d}s_h \mathrm{d}\tau_{h-1} \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h) \cdot b_h^{\pi}(a_h, w_h),$$

where the index $l = \max\{1, h-1-k\}$. In step (b), we have denoted by $\widetilde{\tau}_{h-2} = \tau_{h-2} \setminus \{a_l, o_l\}$ and it holds that $\tau_{h-1} = \widetilde{\tau}_{h-2} \cup \{o_{h-1}, a_{h-1}\}$. Here step (a) follows from expanding the expectation, step (b) follows from cancelling the same terms and using the fact that $W_h \perp S_{h-1}, A_{h-1}, \Gamma_{h-1}|S_h$ under Assumption 3.1, and step (c) follows by marginalizing S_{h-1}, A_l, O_l. Thus we have proved that $\Delta_h^1 = \Delta_h^2$ for finite-length history policies and consequently $\Delta_h = \Delta_h^1 - \Delta_h^2 = 0$.

Case 3: Full-history policy (Example 2.3). For full history information T_h, we have that $\Gamma_{h-1} \cup \{A_h, O_h\} = T_h$. Following the same argument as in Case 2 (Example 2.2), we can first show that

$$\Delta_h^1 = \sum_{a_h \in \mathcal{A}} \int_{\mathcal{S}\times\mathcal{H}_{h-1}} \mathcal{P}_h^\pi(s_h, \tau_{h-1}) \mathrm{d}s_h \mathrm{d}\tau_{h-1} \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h) \mathrm{d}w_h \cdot b_h^\pi(a_h, w_h).$$

Besides, for Δ_2, by a similar argument as in Case 2 except that we don't need marginalize over A_l, O_l in (E.2), we can show that

$$\begin{aligned}
\Delta_h^2 = & \sum_{a_h \in \mathcal{A}} \int_{\mathcal{S}\times\mathcal{H}_{h-2}} \mathcal{P}_{h-1}^\pi(s_{h-1}, \tau_{h-2}) \mathrm{d}s_{h-1} \mathrm{d}\tau_{h-2} \int_{\mathcal{O}} \mathbb{O}_{h-1}(o_{h-1}|s_{h-1}) \mathrm{d}o_{h-1} \sum_{a_{h-1}\in\mathcal{A}} \pi_{h-1}(a_{h-1}|o_{h-1}, \tau_{h-2}) \\
& \int_{\mathcal{S}} \mathbb{P}_h(s_h|s_{h-1}, a_{h-1}) \mathrm{d}s_h \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h, s_{h-1}, a_{h-1}, o_{h-1}, \tau_{h-2}) \cdot b_h^\pi(a_h, w_h) \\
= & \sum_{a_h \in \mathcal{A}} \int_{\mathcal{S}\times\mathcal{H}_{h-1}} \mathcal{P}_h^\pi(s_h, \tau_{h-1}) \mathrm{d}s_h \mathrm{d}\tau_{h-1} \int_{\mathcal{W}} \mathcal{P}_h^b(w_h|s_h) \mathrm{d}w_h \cdot b_h^\pi(a_h, w_h).
\end{aligned}$$

Therefore, we show that $\Delta_h^1 = \Delta_h^2$ for full history policies and consequently $\Delta_h = \Delta_h^1 - \Delta_h^2 = 0$.

Now we have shown that term (B) in (E.1) is actually 0 for Example 2.1, Example 2.2, and Example 2.3, respectively, which allows us to conclude that

$$J(\pi) = \text{(A)} = \mathbb{E}_{\pi^b}\left[\sum_{a\in\mathcal{A}} b_1^\pi(a, W_1)\right].$$

This finishes the proof of Theorem 3.3. $\square$

F Proof of Lemmas in Section D

We first review and define several notations and quantities that are useful in the proof of the lemmas in Section D. Firstly, we define mapping $\ell_h^\pi : \mathbb{B}\times\mathbb{B} \mapsto \{\mathcal{A}\times\mathcal{Z} \mapsto \mathbb{R}\}$ as

$$\begin{aligned}
\ell_h^\pi(b_h, b_{h+1})(A_h, Z_h) := \mathbb{E}_{\pi^b}\Big[& b_h(A_h, W_h) - R_h \pi_h(A_h|O_h, \Gamma_{h-1}) \\
& - \gamma \sum_{a'\in\mathcal{A}} b_{h+1}(a', W_{h+1}) \pi_h(A_h|O_h, \Gamma_{h-1}) \Big| A_h, Z_h \Big].
\end{aligned} \tag{F.1}$$

Furthermore, for each step $h \in [H]$, we define a joint space $\mathcal{I}_h = \mathcal{A}\times\mathcal{W}\times\mathcal{O}\times\mathcal{H}_{h-1}\times\mathcal{W}$ and define mapping $\varsigma_h^\pi : \mathbb{B}\times\mathbb{B} \mapsto \{\mathcal{I}_h \mapsto \mathbb{R}\}$ as

$$\begin{aligned}
\varsigma_h^\pi(b_h, b_{h+1})(A_h, W_h, O_h, \Gamma_{h-1}, W_{h+1}) := & b_h(A_h, W_h) - R_h \pi_h(A_h|O_h, \Gamma_{h-1}) \\
& - \gamma \sum_{a'\in\mathcal{A}} b_{h+1}(a', W_{h+1}) \pi_h(A_h|O_h, \Gamma_{h-1}).
\end{aligned} \tag{F.2}$$

When appropriate, we abbreviate $I_h = (A_h, W_h, O_h, \Gamma_{h-1}, W_{h+1}) \in \mathcal{I}_h$ in the sequel. Using definition (F.1) and (F.2), we further introduce two mappings $\Phi_{\pi,h}^\lambda, \Phi_{\pi,h} : \mathbb{B}\times\mathbb{B}\times\mathbb{G} \mapsto \mathbb{R}$ as defined by (3.9),

$$\Phi_{\pi,h}^\lambda(b_h, b_{h+1}; g) := \mathbb{E}_{\pi^b}\big[\ell_h^\pi(b_h, b_{h+1})(A_h, Z_h)\cdot g(A_h, Z_h) - \lambda g(A_h, Z_h)^2\big],$$

$$\Phi_{\pi,h}(b_h, b_{h+1}; g) := \Phi_{\pi,h}^0(b_h, b_{h+1}; g) = \mathbb{E}_{\pi^b}\big[\ell_h^\pi(b_h, b_{h+1})(A_h, Z_h)\cdot g(A_h, Z_h)\big],$$

where we define that $\Phi_{\pi,h} = \Phi_{\pi,h}^0$. Also, recall from (3.10) that the empirical version of $\Phi_{\pi,h}^\lambda, \Phi_{\pi,h}$ are defined by $\widehat{\Phi}_{\pi,h}^\lambda, \widehat{\Phi}_{\pi,h}$ as

$$\widehat{\Phi}_{\pi,h}^\lambda(b_h, b_{h+1}; g) := \widehat{\mathbb{E}}_{\pi^b}\big[\varsigma_h^\pi(b_h, b_{h+1})(I_h)\cdot g(A_h, Z_h) - \lambda g(A_h, Z_h)^2\big],$$

$$\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) := \widehat{\Phi}_{\pi,h}^0(b_h, b_{h+1}; g) = \widehat{\mathbb{E}}_{\pi^b}\big[\varsigma_h^\pi(b_h, b_{h+1})(I_h)\cdot g(A_h, Z_h)\big].$$

Recall from (3.11) that given function $b_{h+1} \in \mathbb{B}$, the minimax estimator $\widehat{b}_h(b_{h+1})$ is defined as

$$\widehat{b}_h(b_{h+1}) := \arg\min_{b \in \mathbb{B}} \max_{g \in \mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b, b_{h+1}; g).$$

Meanwhile, we define the following quantity for ease of theoretical analysis as

$$b^{\star}_h(b_{h+1}) := \arg\min_{b \in \mathbb{B}} \max_{g \in \mathbb{G}} \Phi^{\lambda}_{\pi,h}(b, b_{h+1}; g). \tag{F.3}$$

By the boundedness assumption on $\mathbb{B}$ in Assumption 4.2, we have that $|\ell^{\pi}_h|, |\varsigma^{\pi}_h| \leq 2M_{\mathbb{B}}$. By the completeness assumption on $\mathbb{G}$ in Assumption 4.3, we also know that $\ell^{\pi}_h(b_h, b_{h+1})/2\lambda \in \mathbb{G}$ for any $b_h, b_{h+1} \in \mathbb{B}$. Finally, for notational simplicity, we define for each $g \in \mathbb{G}$ that,

$$\|g\|_2^2 := \mathbb{E}_{\pi^b}[g(A_h, Z_h)^2],$$

and we denote by $\|g\|_{2,n}^2$ its empirical version, i.e.,

$$\|g\|_{2,n}^2 := \widehat{\mathbb{E}}_{\pi^b}[g(A_h, Z_h)^2].$$

We remark that we have dropped the dependence of $\|g\|_2^2$ on step h since it is clear from the context when used in the proofs and does not make any confusion.

F.1 Proof of Lemma D.1

Proof of Lemma D.1. By definition (D.1) of $F(\mathbf{b})$, for any policy $\pi \in \Pi(\mathcal{H})$ and vector of functions $\mathbf{b} \in \mathbb{B}^{\otimes H}$, it holds that

$$\begin{aligned}
F(\mathbf{b}^{\pi}) - F(\mathbf{b}) &\overset{\text{(a)}}{=} \mathbb{E}_{\pi^b}\left[\sum_{a \in \mathcal{A}} b^{\pi}_1(a, W_1) - b_1(a, W_1)\right] \\
&= \mathbb{E}_{\pi^b}\left[\sum_{a \in \mathcal{A}} \frac{\pi^b_1(a|S_1)}{\pi^b_1(a|S_1)}\left(b^{\pi}_1(a, W_1) - b_1(a, W_1)\right)\right] \\
&\overset{\text{(b)}}{=} \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\frac{1}{\pi^b_1(A_1|S_1)}\left(b^{\pi}_1(a, W_1) - b_1(a, W_1)\right)\middle| S_1, W_1\right]\right] \\
&= \mathbb{E}_{\pi^b}\left[\frac{1}{\pi^b_1(A_1|S_1)}\left(b^{\pi}_1(A_1, W_1) - b_1(A_1, W_1)\right)\right]
\end{aligned}$$

where step (a) follows from Theorem 3.3 and (D.1), and step (b) holds since $A_1 \perp W_1 \mid S_1$ by Assumption 3.1. Notice that by definition (3.3), at step $h = 1$, the weight bridge function q^{π}_h satisfies equation

$$\mathbb{E}_{\pi^b}[q^{\pi}_1(A_1, Z_1)|A_1, S_1, \Gamma_0] = \frac{\mathcal{P}^{\pi}_h(S_1, \Gamma_0)}{\mathcal{P}^{\pi}_h(S_1, \Gamma_0)\pi^b_1(A_1|S_1)} = \frac{1}{\pi^b_1(A_1|S_1)},$$

which further gives that

$$\begin{aligned}
F(\mathbf{b}^{\pi}) - F(\mathbf{b}) &= \mathbb{E}_{\pi^b}\big[\mathbb{E}_{\pi^b}\left[q^{\pi}_1(A_1, Z_1)|A_1, S_1, \Gamma_0\right]\left(b^{\pi}_1(A_1, W_1) - b_1(A_1, W_1)\right)\big] \\
&\overset{\text{(a)}}{=} \mathbb{E}_{\pi^b}\big[\mathbb{E}_{\pi^b}\left[q^{\pi}_1(A_1, Z_1)|A_1, S_1, W_1, \Gamma_0\right] \cdot \big(b^{\pi}_1(A_1, W_1) - b_1(A_1, W_1)\big)\big] \\
&= \mathbb{E}_{\pi^b}\big[q^{\pi}_1(A_1, Z_1)\big(b^{\pi}_1(A_1, W_1) - b_1(A_1, W_1)\big)\big],
\end{aligned}$$

where step (a)holds since $Z_1 \perp W_1 \mid A_1, S_1, \mathcal{H}_0$ by Assumption 3.1. Now we can further obtain that,

$$\begin{aligned}
F(\mathbf{b}^{\pi}) - F(\mathbf{b}) &= \mathbb{E}_{\pi^b}\big[q^{\pi}_1(A_1, Z_1)\mathbb{E}_{\pi^b}\left[b^{\pi}_1(A_1, W_1) - b_1(A_1, W_1)|A_1, Z_1\right]\big] \\
&\overset{\text{(a)}}{=} \mathbb{E}_{\pi^b}\left[q^{\pi}_1(A_1, Z_1)\left\{\mathbb{E}_{\pi^b}\left[R_1\pi_1(A_1|O_1, \Gamma_0) + \gamma\sum_{a'} b^{\pi}_2(a', W_2)\pi_1(A_1|O_1, \Gamma_0)\middle| A_1, Z_1\right]\right.\right. \\
&\qquad \left.\left. - \mathbb{E}_{\pi^b}\left[b_1(A_1, W_1)|A_1, Z_1\right]\right\}\right],
\end{aligned}$$

where step (a) follows from the definition in (3.2) of value bridge function b_1^π in Assumption 3.2. Now to relate the difference between $F(\mathbf{b}^\pi)$ and $F(\mathbf{b})$ with the RMSE loss $\mathcal{L}_1^\pi$ defined in (3.6), we rewrite the above equation as the following,

$$
\begin{aligned}
&F(\mathbf{b}^\pi) - F(\mathbf{b}) \\
&\quad = \mathbb{E}_{\pi^b}\Bigg[q_1^\pi(A_1, Z_1)\Bigg\{\mathbb{E}_{\pi^b}\Bigg[R_1\pi_h(A_1|O_1,\Gamma_0) + \gamma\sum_{a'} b_2^\pi(a', W_2)\pi_1(A_1|O_1,\Gamma_0)\Bigg|A_1, Z_1\Bigg] \\
&\qquad\quad - \mathbb{E}_{\pi^b}\Bigg[R_h\pi_1(A_1|O_1,\Gamma_0) + \gamma\sum_{a'} b_2(a', W_{h+1})\pi_1(A_1|O_1,\Gamma_0)\Bigg|A_1, Z_1\Bigg] \\
&\qquad\quad + \mathbb{E}_{\pi^b}\Bigg[R_1\pi_1(A_1|O_1,\Gamma_0) + \gamma\sum_{a'} b_2(a', W_2)\pi_1(A_1|O_1,\Gamma_0)\Bigg|A_1, Z_1\Bigg] \\
&\qquad\quad - \mathbb{E}_{\pi^b}\Bigg[b_1(A_1, W_1)\Bigg|A_1, Z_1\Bigg]\Bigg\}\Bigg] \\
&\quad = \mathbb{E}_{\pi^b}\Bigg[q_1^\pi(A_1, Z_1)\Bigg\{\gamma\mathbb{E}_{\pi^b}\Bigg[\sum_{a'}\Big(b_2^\pi(a', W_2) - b_2(a, W_2)\Big)\pi_1(A_1|O_1,\Gamma_0)\Bigg|A_1, Z_1\Bigg] \\
&\qquad\quad + \mathbb{E}_{\pi^b}\Bigg[R_1\pi_h(A_1|O_1,\Gamma_0) + \gamma\sum_{a'} b_2(a', W_2)\pi_h(A_1|O_1,\Gamma_0) - b_1(A_1, W_1)\Bigg|A_1, Z_1\Bigg]\Bigg\}\Bigg].
\end{aligned}
\tag{F.4}
$$

We deal with the two terms in the right-hand side of (F.4) respectively. On the one hand, the first term equals to

$$
\begin{aligned}
&\gamma\mathbb{E}_{\pi^b}\Bigg[q_1^\pi(A_1, Z_1)\mathbb{E}_{\pi^b}\Bigg[\sum_{a'}\Big(b_2^\pi(a', W_2) - b_2(a, W_2)\Big)\pi_1(A_1|O_1,\Gamma_0)\Bigg|A_1, Z_1\Bigg]\Bigg] \\
&\quad = \gamma\mathbb{E}_{\pi^b}\Bigg[q_1^\pi(A_1, Z_1)\sum_{a'}\Big(b_2^\pi(a', W_2) - b_2(a, W_2)\Big)\pi_1(A_1|O_1,\Gamma_0)\Bigg] \\
&\quad = \gamma\mathbb{E}_{\pi^b}\Bigg[\mathbb{E}_{\pi^b}\Bigg[q_1^\pi(A_1, Z_1)\Bigg|S_1, A_1, \Gamma_0, O_1, W_2\Bigg]\sum_{a'}\Big(b_2^\pi(a', W_2) - b_2(a, W_2)\Big)\pi_1(A_1|O_1,\Gamma_0)\Bigg] \\
&\quad \overset{\text{(a)}}{=} \gamma\mathbb{E}_{\pi^b}\Bigg[\mathbb{E}_{\pi^b}\Bigg[q_1^\pi(A_1, Z_1)\Bigg|S_1, A_1, \Gamma_0\Bigg]\sum_{a'}\Big(b_2^\pi(a', W_2) - b_2(a, W_2)\Big)\pi_1(A_1|O_1,\Gamma_0)\Bigg] \\
&\quad \overset{\text{(b)}}{=} \gamma\mathbb{E}_{\pi^b}\Bigg[\frac{\mu_1(S_1,\Gamma_0)}{\pi_1^b(A_1|S_1)}\sum_{a'}\Big(b_2^\pi(a', W_2) - b_2(a, W_2)\Big)\pi_1(A_1|O_1,\Gamma_0)\Bigg],
\end{aligned}
$$

where step (a) follows from the fact that $Z_1 \perp O_1, W_2 | S_1, A_1, \Gamma_0$ according to Assumption 3.1, and step (b) follows from the definition (3.3) of weight bridge function q_1^π in Assumption 3.2. Now following the same argument as in showing $\Delta_h = 0$ in the proof of Theorem 3.3, we can show that

$$
\begin{aligned}
&\mathbb{E}_{\pi^b}\Bigg[\frac{\mu_1(S_1,\Gamma_0)}{\pi_1^b(A_1|S_1)}\sum_{a'}\Big(b_2^\pi(a', W_2) - b_2(a, W_2)\Big)\pi_1(A_1|O_1,\Gamma_0)\Bigg] \\
&\quad = \mathbb{E}_{\pi^b}\Bigg[q_2^\pi(A_2, Z_2)\Big(b_2^\pi(A_2, W_2) - b_2(A_2, W_2)\Big)\Bigg].
\end{aligned}
\tag{F.5}
$$

On the other hand, the second term in the R.H.S. of (F.4) can be rewritten and bounded by

$$
\begin{aligned}
&\mathbb{E}_{\pi^b}\Bigg[q_1^\pi(A_1, Z_1)\mathbb{E}_{\pi^b}\Bigg[R_1\pi_1(A_1|O_1,\Gamma_0) + \gamma\sum_{a'} b_2(a', W_2)\pi_1(A_1|O_1,\Gamma_0) - b_1(A_1, W_1)\Bigg|A_1, Z_1\Bigg]\Bigg] \\
&\quad \le \sqrt{C^\pi}\mathbb{E}_{\pi^b}\Bigg[\Bigg\{\mathbb{E}_{\pi^b}\Bigg[R_1\pi_1(A_1|O_1,\Gamma_0) + \gamma\sum_{a'} b_2(a', W_2)\pi_1(A_1|O_1,\Gamma_0) - b_1(A_1, W_1)\Bigg|A_1, Z_1\Bigg]\Bigg\}^{1/2}\Bigg] \\
&\quad = \sqrt{C^\pi}\cdot\sqrt{\mathcal{L}_1^\pi(b_1, b_2)},
\end{aligned}
\tag{F.6}
$$

where C^π is defined as $C^\pi := \sup_{h\in[H]} \mathbb{E}_{\pi^b}\left[(q_h^\pi(A_h, Z_h))^2\right]$, the inequality follows from Cauchy-Schwarz inequality, and the equality follows from the definition of $\mathcal{L}_1^\pi$ in (3.6). Combining (F.4), (F.5) with (F.6), we can obtain that

$$
\begin{aligned}
&F(\mathbf{b}^\pi) - F(\mathbf{b}) \\
&\quad \le \sqrt{C^\pi} \cdot \sqrt{\mathcal{L}_1^\pi(b_1, b_2)} + \gamma \mathbb{E}_{\pi^b}\left[q_2^\pi(A_2, Z_2)\Big(b_2^\pi(A_2, W_2) - b_2(A_2, W_2)\Big)\right]. \qquad \text{(F.7)}
\end{aligned}
$$

Now applying the above argument on the second term in the R.H.S. of (F.7) recursively, we can obtain that

$$
F(\mathbf{b}^\pi) - F(\mathbf{b}) \le \sum_{h=1}^{H} \gamma^{h-1}\sqrt{C^\pi} \cdot \sqrt{\mathcal{L}_h^\pi(b_h, b_{h+1})}.
$$

This finishes the proof of Lemma D.1. $\square$

F.2 Proof of Lemma D.2

Proof of Lemma D.2. By the definition of the confidence region $\mathrm{CR}^\pi(\alpha)$ in (3.12), we need to show for any policy $\pi \in \Pi(\mathcal{H})$ and step $h \in [H]$, it holds that,

$$
\max_{g\in\mathbb{G}} \widehat{\Phi}_{\pi,h}^\lambda(b_h^\pi, b_{h+1}^\pi; g) - \max_{g\in\mathbb{G}} \widehat{\Phi}_{\pi,h}^\lambda(\widehat{b}_h(b_{h+1}^\pi), b_{h+1}^\pi; g) \le \xi. \qquad \text{(F.8)}
$$

Notice that by Assumption 4.2, the function class $\mathbb{G}$ is symmetric and star-shaped, which indicates that

$$
\max_{g\in\mathbb{G}} \widehat{\Phi}_{\pi,h}^\lambda(\widehat{b}_h(b_{h+1}^\pi), b_{h+1}^\pi; g) \ge \widehat{\Phi}_{\pi,h}^\lambda(\widehat{b}_h(b_{h+1}^\pi), b_{h+1}^\pi; 0) = 0.
$$

Therefore, in order to prove (F.8), it suffices to show that

$$
\max_{g\in\mathbb{G}} \widehat{\Phi}_{\pi,h}^\lambda(b_h^\pi, b_{h+1}^\pi; g) \le \xi. \qquad \text{(F.9)}
$$

To relate the empirical expectation $\widehat{\Phi}_{\pi,h}^\lambda(b_h^\pi, b_{h+1}^\pi; g) = \widehat{\Phi}_{\pi,h}(b_h^\pi, b_{h+1}^\pi; g) - \lambda\|g\|_{2,n}^2$ to its population version, we need two localized uniform concentration inequalities. On the one hand, to relate $\|g\|_2^2$ and $\|g\|_{2,n}^2$, by Lemma I.1 (Theorem 14.1 of Wainwright (2019)), for some absolute constants $c_1, c_2 > 0$, it holds with probability at least $1 - \delta/2$ that,

$$
\big|\|g\|_{2,n}^2 - \|g\|_2^2\big| \le \frac{1}{2}\|g\|_2^2 + \frac{M_{\mathbb{G}}^2 \log(2c_1/\zeta)}{2c_2 n}, \quad \forall g \in \mathbb{G}, \qquad \text{(F.10)}
$$

where $\zeta = \min\{\delta, 2c_1\exp(-c_2 n\alpha_{\mathbb{G},n}^2/M_{\mathbb{G}}^2)\}$ and $\alpha_{\mathbb{G},n}$ is the critical radius of function class $\mathbb{G}$ defined in Assumption 4.2. On the other hand, to relate $\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g)$ and $\Phi_{\pi,h}(b_h, b_{h+1}; g)$ we invoke Lemma I.2 (Lemma 11 of (Foster and Syrgkanis, 2019)). Specifically, for any given $b_h, b_{h+1} \in \mathbb{B}$, $\pi \in \Pi(\mathcal{H})$, and $h \in [H]$, in Lemma I.2 we choose $\mathcal{F} = \mathbb{G}$, $\mathcal{X} = \mathcal{A}\times\mathcal{Z}$, $\mathcal{Y} = \mathcal{I}_h$, and loss function $\ell(g(A_h, Z_h), I_h) := \varsigma_h^\pi(b_h, b_{h+1})(I_h)\cdot g(A_h, Z_h)$ where ς_h^π is defined in (F.1), $I_h \in \mathcal{I}_h$ is defined in the beginning of Appendix F. It holds that ℓ is L-Lipschitz continuous in the first argument since for any $g, g' \in \mathbb{G}$, $(A_h, Z_h) \in \mathcal{A}\times\mathcal{Z}$, it holds that

$$
\begin{aligned}
\big|\ell(g(A_h, Z_h), I_h) - \ell(g'(A_h, Z_h), I_h)\big| &= |\varsigma_h^\pi(b_h, b_{h+1})(I_h)| \cdot |g(A_h, Z_h) - g'(A_h, Z_h)| \\
&\le 2M_{\mathbb{B}} \cdot |g(A_h, Z_h) - g'(A_h, Z_h)|,
\end{aligned}
$$

which indicates that $L = 2M_{\mathbb{B}}$. Now setting $f^\star = 0$ in Lemma I.2, we have that δ_n in Lemma I.2 coincides with $\alpha_{\mathbb{G},n}$ in Assumption 4.2. Then we can conclude that for some absolute constants $c_1, c_2 > 0$, it holds with probability at least $1 - \delta/(2|\mathbb{B}|^2|\Pi(\mathcal{H})|H)$ that

$$
\begin{aligned}
&\Big|\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \Phi_{\pi,h}(b_h, b_{h+1}; g)\Big| \\
&\quad = \Big|\widehat{\mathbb{E}}_{\pi^b}[\ell(g(A_h, Z_h), I_h)] - \mathbb{E}_{\pi^b}[\ell(g(A_h, Z_h), I_h)]\Big| \\
&\quad \le 18L\|g\|_2\sqrt{\frac{M_{\mathbb{G}}^2 \log\big(2c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta'\big)}{c_2 n}} + \frac{18LM_{\mathbb{G}}^2 \log\big(2c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta'\big)}{c_2 n}, \quad \forall g \in \mathbb{G},
\end{aligned}
\qquad \text{(F.11)}
$$

where $\zeta' = \min\{\delta, 2c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H\exp(-c_2 n\alpha_{\mathbb{G},n}^2/M_{\mathbb{G}}^2)\}$. Applying a union bound argument over $b_h, b_{h+1} \in \mathbb{B}$, $\pi \in \Pi(\mathcal{H})$, and $h \in [H]$, we then have that (F.11) holds for any $b_h, b_{h+1} \in \mathbb{B}$, $g \in \mathbb{G}$, $\pi \in \Pi(\mathcal{H})$, and $h \in [H]$ with probability at least $1 - \delta/2$. Now using these two concentration inequalities (F.10) and (F.11), we can further deduce that, for some absolute constants $c_1, c_2 > 0$, with probability at least $1 - \delta$,

$$
\begin{aligned}
&\max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b_h^\pi, b_{h+1}^\pi; g) \\
&\quad = \max_{g\in\mathbb{G}} \left\{ \widehat{\Phi}_{\pi,h}(b_h^\pi, b_{h+1}^\pi; g) - \lambda\|g\|_{2,n}^2 \right\} \\
&\quad \le \max_{g\in\mathbb{G}} \Bigg\{ \Phi_{\pi,h}(b_h^\pi, b_{h+1}^\pi; g) - \lambda\|g\|_2^2 + \frac{\lambda}{2}\|g\|_2^2 + \frac{\lambda M_{\mathbb{G}}^2 \log(2c_1/\zeta)}{2c_2 n}, \\
&\qquad\qquad + 18L\|g\|_2 \sqrt{\frac{M_{\mathbb{G}}^2 \log(2c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{c_2 n}} + \frac{18LM_{\mathbb{G}}^2 \log(2c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{c_2 n} \Bigg\},
\end{aligned}
$$

where ζ is given as $\zeta = \min\{\delta, 2c_1\exp(-c_2 n\alpha_{\mathbb{G},n}^2/M_{\mathbb{G}}^2)\}$ and ζ' is given as $\zeta' = \min\{\delta, 2c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H\exp(-c_2 n\alpha_{\mathbb{G},n}^2/M_{\mathbb{G}}^2)\}$ for any policy $\pi \in \Pi(\mathcal{H})$ and step h. Then we can further bound the right-hand side of the above inequality as

$$
\begin{aligned}
&\max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b_h^\pi, b_{h+1}^\pi; g) \\
&\quad \le \max_{g\in\mathbb{G}} \Phi_{\pi,h}(b_h^\pi, b_{h+1}^\pi; g) + \max_{g\in\mathbb{G}} \left\{ -\frac{\lambda}{2}\|g\|_2^2 + 18L\|g\|_2 \sqrt{\frac{M_{\mathbb{G}}^2 \cdot \log(2c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{c_2 n}} \right\} \\
&\qquad + \frac{\lambda M_{\mathbb{G}}^2 \cdot \log(2c_1/\zeta)}{2c_2 n} + \frac{18LM_{\mathbb{G}}^2 \cdot \log(2c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{c_2 n} \\
&\quad \le \frac{728L^2 \cdot M_{\mathbb{G}}^2 \cdot \log(2c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{\lambda n} + \frac{\lambda M_{\mathbb{G}}^2 \cdot \log(2c_1/\zeta)}{2c_2 n} \\
&\qquad + \frac{18LM_{\mathbb{G}}^2 \cdot \log(2c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{c_2 n}.
\end{aligned}
$$

Here the last inequality holds from the fact that $\Phi_{\pi,h}(b_h^\pi, b_{h+1}^\pi; g) = 0$ since b_h^π and b_{h+1}^π are true bridge functions, and the fact that $\sup_{\|g\|_2}\{a\|g\|_2 - b\|g\|_2^2\} \le a^2/4b$ for any $b > 0$. Now according to the choice of ξ in Lemma D.2, using the fact that $\zeta < \zeta'$ and $L = 2M_{\mathbb{B}}$, we can conclude that, with probability at least $1 - \delta$,

$$
\begin{aligned}
&\max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b_h^\pi, b_{h+1}^\pi; g) \\
&\quad \le \frac{728L^2 M_{\mathbb{G}}^2 \cdot \log(2c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{\lambda n} + \frac{\lambda M_{\mathbb{G}}^2 \cdot \log(2c_1/\zeta)}{2c_2 n} + \frac{18LM_{\mathbb{G}}^2 \cdot \log(2c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{c_2 n} \\
&\quad \lesssim \mathcal{O}\left(\frac{(\lambda + 1/\lambda) \cdot M_{\mathbb{B}}^2 M_{\mathbb{G}}^2 \cdot \log(|\mathbb{B}||\Pi(\mathcal{H})|H/\zeta)}{n} \right) \lesssim \xi.
\end{aligned}
$$

This proves (F.9), and thus further indicates (F.8). Therefore, we finish the proof of Lemma D.2. $\square$

F.3 Proof of Lemma D.3

We first give the high-level idea for proving Lemma D.3 as following. In order to achieve the fast rate for the whole confidence region, we took a series of novel proof steps.

We first introduce the following lemma, which claims that for any $b_{h+1} \in \mathbb{B}$, the $b^\star(b_{h+1})$ defined in (F.3) satisfies that $\max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b^\star(b_{h+1}), b_{h+1}; g)$ is well-bounded. The proof of lemma follows the same argument as in the proof of Lemma D.2, which we defer to Appendix F.4.

Then given any bridge function in the confidence region, we identify a key term (term $(\star)$ in (F.12)) which is related to the RMSE of this bridge function. By carefully upper & lower bound this term, where Lemma F.1 is applied, we eventually obtain a quadratic inequality that the RMSE of this bridge function satisfies. By solving this inequality, we can derive an upper bound on the RMSE loss which is *uniform* over the bridge functions in the confidence region, which is exactly the fast rate of the whole confidence region.

Lemma F.1. *For any function $b_{h+1} \in \mathbb{B}$, policy $\pi \in \Pi(\mathcal{H})$, and step $h \in [H]$, it holds with probability at least $1-\delta/2$ that*

$$\max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) \le \xi + \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}},$$

where $b^\star(b_{h+1})$ is defined in (F.3) and ξ is defined in Lemma D.3.

Proof of Lemma F.1. See Appendix F.4 for a detailed proof. $\square$

With Lemma F.1, we are now ready to give the proof of Lemma D.3.

Proof of Lemma D.3. Let's consider that for any $b_h, b_{h+1} \in \mathrm{CR}^\pi(\xi)$, we have that

$$\begin{aligned}\max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b_h, b_{h+1}; g) = \max_{g\in\mathbb{G}} \Big\{ &\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \widehat{\Phi}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) - 2\lambda\|g\|^2_{2,n} \\ &+ \widehat{\Phi}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) + \lambda\|g\|^2_{2,n}\Big\}.\end{aligned}$$

We further write the above as

$$\begin{aligned}\max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b_h, b_{h+1}; g) &\ge \max_{g\in\mathbb{G}} \Big\{ \widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \widehat{\Phi}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) - 2\lambda\|g\|^2_{2,n}\Big\} \\ &\quad + \min_{g\in\mathbb{G}} \Big\{\widehat{\Phi}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) + \lambda\|g\|^2_{2,n}\Big\} \\ &\overset{\text{(a)}}{=} \underbrace{\max_{g\in\mathbb{G}} \Big\{ \widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \widehat{\Phi}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) - 2\lambda\|g\|^2_{2,n}\Big\}}_{(\star)} \\ &\quad - \max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g). \end{aligned} \tag{F.12}$$

Here step (a) follows from that $\mathbb{G}$ is symmetric, $\widehat{\Phi}_{\pi,h}(b_h, h_{h+1}; -g) = -\widehat{\Phi}_{\pi,h}(b_h, h_{h+1}; g)$, and that

$$\begin{aligned}\min_{g\in\mathbb{G}} \Big\{\widehat{\Phi}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) + \lambda\|g\|^2_{2,n}\Big\} &= \min_{g\in\mathbb{G}} \Big\{ - \widehat{\Phi}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; -g) + \lambda\|g\|^2_{2,n}\Big\} \\ &= \min_{g\in\mathbb{G}} \Big\{ - \widehat{\Phi}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) + \lambda\|g\|^2_{2,n}\Big\} \\ &= -\max_{g\in\mathbb{G}} \Big\{\widehat{\Phi}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) - \lambda\|g\|^2_{2,n}\Big\} \\ &= -\max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g).\end{aligned}$$

In the sequel, we upper and lower bound term $(\star)$ respectively.

Upper bound of term $(\star)$. By inequality (F.12), after rearranging terms, we can arrive that

$$\begin{aligned}(\star) &\le \max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) + \max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b_h, b_{h+1}; g) \\ &\le \max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) \\ &\qquad + \max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b_h, b_{h+1}; g) - \max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(\widehat{b}_h(b_{h+1}), b_{h+1}; g) \\ &\qquad + \max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(\widehat{b}_h(b_{h+1}), b_{h+1}; g)\end{aligned}$$

On the one hand, by Lemma F.1, we have that with probability at least $1-\delta/2$,

$$\max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) \le \xi + \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}}, \tag{F.13}$$

and by the definition of $\widehat{b}_h(b_{h+1})$ in (3.11), it holds simultaneously that

$$\max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(\widehat{b}_h(b_{h+1}), b_{h+1}; g) \le \max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) \le \xi + \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}}. \tag{F.14}$$

On the other hand, by the choice of $\mathrm{CR}^\pi(\xi)$, it holds that

$$\max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b_h, b_{h+1}; g) - \max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(\widehat{b}_h(b_{h+1}), b_{h+1}; g) \le \xi. \tag{F.15}$$

Consequently, by combining (F.13), (F.14), and (F.15), we conclude that with probability at least $1-\delta/2$,

$$(\star) \le 3\xi + 2\epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}}. \tag{F.16}$$

Lower bound of term $(\star)$. For lower bound, we need two localized uniform concentration inequalities similar to (F.10) and (F.11) in the proof of Lemma D.2. On the one hand, by Lemma I.1, for some absolute constants $c_1, c_2 > 0$, it holds with probability at least $1-\delta/4$ that,

$$\big|\|g\|_{2,n}^2 - \|g\|_2^2\big| \le \frac{1}{2}\|g\|_2^2 + \frac{M_{\mathbb{G}}^2 \log(4c_1/\zeta)}{2c_2 n}, \quad \forall g \in \mathbb{G}, \tag{F.17}$$

where $\zeta = \min\{\delta, 4c_1\exp(-c_2 n\alpha_{\mathbb{G},n}^2/M_{\mathbb{G}}^2)\}$ and $\alpha_{\mathbb{G},n}$ is the critical radius of $\mathbb{G}$ defined in Assumption 4.2. On the other hand, following the same argument as in deriving (F.11), for any given $b_h, b_h', b_{h+1} \in \mathbb{B}$, $\pi \in \Pi(\mathcal{H})$, and $h \in [H]$, in Lemma I.2 we choose $\mathcal{F} = \mathbb{G}$, $\mathcal{X} = \mathcal{A}\times\mathcal{Z}$, $\mathcal{Y} = \mathcal{I}$, and loss function

$$\ell(g(A_h, Z_h), I_h) := \varsigma_h^\pi(b_h, b_{h+1})(I_h) g(A_h, Z_h) - \varsigma_h^\pi(b_h', b_{h+1})(I_h) g(A_h, Z_h),$$

where ς_h^π is defined in (F.1) and $I_h \in \mathcal{I}_h$ is defined in the beginning of Appendix F. It holds that ℓ is L-Lipschitz continuous in its first argument with $L = 2M_{\mathbb{B}}$. Now setting $f^\star = 0$ in Lemma I.2, we have that δ_n in Lemma I.2 coincides with $\alpha_{\mathbb{G},n}$ in Assumption 4.2. Then we have that for some absolute constants $c_1, c_2 > 0$, it holds with probability at least $1-\delta/(4|\mathbb{B}|^3|\Pi(\mathcal{H})|H)$ that

$$\begin{aligned}
&\Big|\Big(\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \widehat{\Phi}_{\pi,h}(b_h', b_{h+1}; g)\Big) - \Big(\Phi_{\pi,h}(b_h, b_{h+1}; g) - \Phi_{\pi,h}(b_h', b_{h+1}; g)\Big)\Big| \\
&\quad = \Big|\widehat{\mathbb{E}}_{\pi^b}[\ell(g(A_h, Z_h), I_h)] - \mathbb{E}_{\pi^b}[\ell(g(A_h, Z_h), I_h)]\Big| \\
&\quad \le 18L\|g\|_2\sqrt{\frac{M_{\mathbb{G}}^2\cdot\log(4c_1|\mathbb{B}|^3|\Pi(\mathcal{H})|H/\zeta')}{c_2 n}} + \frac{18L\cdot M_{\mathbb{G}}^2\cdot\log\big(4c_1|\mathbb{B}|^3|\Pi(\mathcal{H})|H/\zeta'\big)}{c_2 n}, \quad \forall g \in \mathbb{G},
\end{aligned} \tag{F.18}$$

where $\zeta' = \min\{\delta, 4c_1|\mathbb{B}|^3|\Pi(\mathcal{H})|H\exp(-c_2 n\alpha_{\mathbb{G},n}^2/M_{\mathbb{G}}^2)\}$. Applying a union bound argument over $b_h, b_h', b_{h+1} \in \mathbb{B}$, $\pi \in \Pi(\mathcal{H})$, and $h \in [H]$, we have that (F.18) holds for any $b_h, b_h', b_{h+1} \in \mathbb{B}$, $g \in \mathbb{G}$, $\pi \in \Pi(\mathcal{H})$, and $h \in [H]$ with probability at least $1-\delta/4$. Finally, for simplicity, we denote that

$$\iota_n := \sqrt{\frac{M_{\mathbb{G}}^2\cdot\log(4c_1|\mathbb{B}|^3|\Pi(\mathcal{H})|H/\zeta')}{c_2 n}}, \quad \iota_n' := \sqrt{\frac{M_{\mathbb{G}}^2\cdot\log(4c_1/\zeta)}{2c_2 n}} \tag{F.19}$$

Now we are ready to prove the lower bound on term $(\star)$. For simplicity, given fixed $b_h, b_{h+1} \in \mathbb{B}$, we denote

$$g_h^\pi := \frac{1}{2\lambda}\ell_h^\pi(b_h, b_{h+1}) \in \mathbb{G},$$

where ℓ_h^π is defined in (F.1) and $g_h^\pi \in \mathbb{G}$ due to Assumption 4.3. Now consider that

$$\begin{aligned}
(\star) &= \max_{g\in\mathbb{G}}\Big\{\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \widehat{\Phi}_{\pi,h}(b_h^\star(b_{h+1}), b_{h+1}; g) - 2\lambda\|g\|_{2,n}^2\Big\} \\
&\ge \widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g_h^\pi/2) - \widehat{\Phi}_{\pi,h}(b_h^\star(b_{h+1}), b_{h+1}; g_h^\pi/2) - \frac{\lambda}{2}\|g_h^\pi\|_{2,n}^2,
\end{aligned}$$

where the inequality follows from the fact that $\mathbb{G}$ is star-shaped and consequently $g_h^\pi/2 \in \mathbb{G}$. Then by applying concentration inequality (F.17) and (F.18), we have that

$$\begin{aligned}(\star) &\geq \Phi_{\pi,h}(b_h, b_{h+1}; g_h^\pi/2) - \Phi_{\pi,h}(b_h^\star(b_{h+1}), b_{h+1}; g_h^\pi/2) - 18L\iota_n\|g_h^\pi\|_2 - 18L\iota_n^2 \\ &\quad - \frac{\lambda}{2}\left(\frac{3}{2}\|g_h^\pi\|_2^2 + \iota_n'^2\right) \\ &\geq \lambda\|g_h^\pi\|_2^2 - 18L\iota_n\|g_h^\pi\|_2 - \epsilon_{\mathbb{B}}^{1/2}M_{\mathbb{G}} - 18L\iota_n^2 - \frac{\lambda}{2}\left(\frac{3}{2}\|g_h^\pi\|_2^2 + \iota_n'^2\right) \\ &= \frac{\lambda}{4}\|g_h^\pi\|_2^2 - 18L\iota_n\|g_h^\pi\|_2 - 18L\iota_n^2 - \frac{\lambda}{2}\iota_n'^2 - \epsilon_{\mathbb{B}}^{1/2}M_{\mathbb{G}}, \end{aligned} \tag{F.20}$$

where the second inequality follows from that $\Phi_{\pi,h}(b^\star(b_{h+1}), b_{h+1}; g_h^\pi/2) \leq \epsilon_{\mathbb{B}}^{1/2}M_{\mathbb{G}}$ (we prove this inequality by (F.25) in the proof of Lemma F.1) and the fact that

$$\Phi_{\pi,h}(b_h, b_{h+1}; g_h^\pi/2) = \frac{1}{4\lambda}\mathbb{E}_{\pi^b}[\ell_h^\pi(b_h, b_{h+1})(A_h, Z_h)^2] = \lambda\|g_h^\pi\|_2^2.$$

Combining upper bound and lower bound of term $(\star)$. Now we are ready to combine the upper bound and lower bound of $(\star)$ to derive the bound on $\mathcal{L}_h^\pi(b_h, b_{h+1})$. By combining upper bound (F.16) and lower bound (F.20), we have that with probability at least $1-\delta$, for any $b_h, b_{h+1} \in \mathbb{B}$, $\pi \in \Pi(\mathcal{H})$, and $h \in [H]$,

$$\frac{\lambda}{4}\|g_h^\pi\|_2^2 - 18L\iota_n\|g_h^\pi\|_2 - 18L\iota_n^2 - \frac{\lambda}{2}\iota_n'^2 - \epsilon_{\mathbb{B}}^{1/2}M_{\mathbb{G}} \leq 3\xi + 2\epsilon_{\mathbb{B}}^{1/2}M_{\mathbb{G}}, \tag{F.21}$$

This gives a quadratic inequality on $\|g_h^\pi\|_2$, i.e.,

$$\lambda\|g_h^\pi\|_2^2 - \underbrace{72L\iota_n}_{\text{(A)}}\|g_h^\pi\|_2 - \underbrace{4\left(18L\iota_n^2 + \frac{\lambda}{2}\iota_n'^2 + 3\xi + 3\epsilon_{\mathbb{B}}^{1/2}M_{\mathbb{G}}\right)}_{\text{(B)}} \leq 0.$$

By solving this quadratic equation, we have that

$$\|g_h^\pi\|_2 \leq \frac{1}{2\lambda}\mathrm{A} + \frac{1}{2\lambda}\sqrt{\mathrm{A}^2 + 4\mathrm{B}} \leq \frac{\mathrm{A}}{\lambda} + \frac{\sqrt{\mathrm{B}}}{\lambda}.$$

Applying the definition of A and B, we conclude that, with probability at least $1-\delta$,

$$\begin{aligned}\|g_h^\pi\|_2 &\leq \frac{72}{\lambda}L\iota_n + \frac{2}{\lambda}\left(18L\iota_n^2 + \frac{\lambda}{2}\iota_n'^2 + 3\xi + 3\epsilon_{\mathbb{B}}^{1/2}M_{\mathbb{G}}\right)^{1/2} \\ &\leq \frac{72}{\lambda}L\iota_n + \frac{6\sqrt{2}}{\lambda}L^{1/2}\iota_n + \frac{\sqrt{2}}{\sqrt{\lambda}}\iota_n' + \frac{2\sqrt{3}}{\lambda}\xi^{1/2} + \frac{2\sqrt{3}}{\lambda}\epsilon_{\mathbb{B}}^{1/4}M_{\mathbb{G}}^{1/2}\end{aligned}$$

Therefore, we can bound the RMSE loss $\mathcal{L}_h^\pi(b_h, b_{h+1})$ by

$$\sqrt{\mathcal{L}_h^\pi(b_h, b_{h+1})} = 2\lambda\|g_h^\pi\|_2 \leq (144L + 12\sqrt{2}L^{1/2})\iota_n + 2\sqrt{2\lambda}\iota_n' + 4\sqrt{3}\xi^{1/2} + 4\sqrt{3}\epsilon_{\mathbb{B}}^{1/4}M_{\mathbb{G}}^{1/2}. \tag{F.22}$$

Plugging in the definition of ι_n, ι_n' in (F.19), ξ in Lemma D.3, and that $L = 2M_{\mathbb{B}}$, we have that

$$\begin{aligned}&\sqrt{\mathcal{L}_h^\pi(b_h, b_{h+1})} \\ &\quad\leq (144L + 12\sqrt{2L})\cdot\sqrt{\frac{M_{\mathbb{G}}^2\cdot\log(4c_1|\mathbb{B}|^3|\Pi(\mathcal{H})|H/\zeta')}{c_2 n}} + 2\sqrt{2\lambda}\cdot\sqrt{\frac{M_{\mathbb{G}}^2\cdot\log(4c_1/\zeta)}{2c_2 n}} \\ &\qquad + 4\sqrt{3}\cdot\sqrt{\frac{C_1(\lambda+1/\lambda)\cdot M_{\mathbb{B}}^2M_{\mathbb{G}}^2\cdot\log(|\mathbb{B}||\Pi(\mathcal{H})|H/\zeta')}{n}} + 4\sqrt{3}\cdot\epsilon_{\mathbb{B}}^{1/4}M_{\mathbb{G}}^{1/2} \\ &\quad\leq \widetilde{C}_1 M_{\mathbb{B}}M_{\mathbb{G}}\sqrt{\frac{(\lambda+1/\lambda)\cdot\log(|\mathbb{B}||\Pi(\mathcal{H})|H/\zeta)}{n}} + \widetilde{C}_1\epsilon_{\mathbb{B}}^{1/4}M_{\mathbb{G}}^{1/2}.\end{aligned}$$

for some problem-independent constant $\widetilde{C}_1 > 0$ and $\zeta = \min\{\delta, 4c_1\exp(-c_2 n\alpha_{\mathbb{G},n}^2/M_{\mathbb{G}}^2)\}$. Here in the second inequality we have used the fact that $\zeta < \zeta'$. This finishes the proof of Lemma D.3. $\square$

F.4 Proof of Lemma F.1

Proof of Lemma F.1. Following the proof of Lemma D.2, we first relate $\widehat{\Phi}^{\lambda}_{\pi,h}(b_h, b_{h+1}; g) = \widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \lambda\|g\|^2_{2,n}$ and its population version $\Phi^{\lambda}_{\pi,h}(b_h, b_{h+1}; g)$ via two localized uniform concentration inequalities. On the one hand, to relate $\|g\|_2^2$ and $\|g\|^2_{2,n}$, by Lemma I.1 (Theorem 14.1 of Wainwright (2019)), for some absolute constants $c_1, c_2 > 0$, it holds with probability at least $1 - \delta/4$ that

$$\big|\|g\|^2_{2,n} - \|g\|_2^2\big| \le \frac{1}{2}\|g\|_2^2 + \frac{M_{\mathbb{G}}^2 \cdot \log(4c_1/\zeta)}{2c_2 n}, \quad \forall g \in \mathbb{G}, \tag{F.23}$$

where $\zeta = \min\{\delta, 4c_1\exp(-c_2 n\alpha^2_{\mathbb{G},n}/M^2_{\mathbb{G}})\}$ and $\alpha_{\mathbb{G},n}$ is the critical radius of function class $\mathbb{G}$ defined in Assumption 4.2. On the other hand, to relate $\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g)$ and $\Phi_{\pi,h}(b_h, b_{h+1}; g)$, we invoke Lemma I.2 (Lemma 11 of (Foster and Syrgkanis, 2019)). Specifically, for any given $b_h, b_{h+1} \in \mathbb{B}$, $\pi \in \Pi(\mathcal{H})$, and step h, in Lemma I.2 we choose $\mathcal{F} = \mathbb{G}$, $\mathcal{X} = \mathcal{A} \times \mathcal{Z}$, $\mathcal{Y} = \mathcal{I}_h$, and loss function $\ell(g(A_h, Z_h), I_h) := \varsigma_h^{\pi}(b_h, b_{h+1})(I_h) g(A_h, Z_h)$ where ℓ_h^{π} is defined in (F.1) and $I_h \in \mathcal{I}_h$ is defined in the beginning of Appendix F. We can see that ℓ is L-Lipschitz continuous in the first argument since for any $g, g' \in \mathbb{G}$, $(A_h, Z_h) \in \mathcal{A} \times \mathcal{Z}$, it holds that

$$\begin{aligned}\big|\ell(g(A_h, Z_h), I_h) - \ell(g'(A_h, Z_h), I_h)\big| &= |\varsigma_h^{\pi}(b_h, b_{h+1})(I_h)| \cdot |g(A_h, Z_h) - g'(A_h, Z_h)| \\ &\le 2M_{\mathbb{B}} \cdot |g(A_h, Z_h) - g'(A_h, Z_h)|,\end{aligned}$$

which indicates that $L = 2M_{\mathbb{B}}$. Now setting $f^\star = 0$ in Lemma I.2, we have that δ_n in Lemma I.2 coincides with $\alpha_{\mathbb{G},n}$ in Assumption 4.2. Then we can conclude that for some absolute constants $c_1, c_2 > 0$, it holds with probability at least $1 - \delta/(4|\mathbb{B}|^2|\Pi(\mathcal{H})|H)$ that, for all $g \in \mathbb{G}$,

$$\begin{aligned}&\Big|\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \Phi_{\pi,h}(b_h, b_{h+1}; g)\Big| \\ &\quad = \Big|\widehat{\mathbb{E}}_{\pi^b}[\ell(g(A_h, Z_h), A_h, Z_h)] - \mathbb{E}_{\pi^b}[\ell(g(A_h, Z_h), A_h, Z_h)]\Big| \\ &\quad \le 18L\|g\|_2\sqrt{\frac{M_{\mathbb{G}}^2 \cdot \log\big(4c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta\big)}{c_2 n}} + \frac{18L \cdot M_{\mathbb{G}}^2 \cdot \log\big(4c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta\big)}{c_2 n},\end{aligned} \tag{F.24}$$

where $\zeta = \min\{\delta, 4c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H\exp(-c_2 n\alpha^2_{\mathbb{G},n}/M^2_{\mathbb{G}})\}$. Applying a union bound argument over $b_h, b_{h+1} \in \mathbb{B}$, $\pi \in \Pi(\mathcal{H})$, and $h \in [H]$, we then have that (F.11) holds for any $b_h, b_{h+1} \in \mathbb{B}$, $g \in \mathbb{G}$, $\pi \in \Pi(\mathcal{H})$, and $h \in [H]$ with probability at least $1 - \delta/4$. Now using these two concentration inequalities (F.23) and (F.24), we can further deduce that, for some absolute constants $c_1, c_2 > 0$, with probability at least $1 - \delta/2$,

$$\begin{aligned}&\max_{g\in\mathbb{G}} \widehat{\Phi}^{\lambda}_{\pi,h}(b_h^\star(b_{h+1}), b^{\pi}_{h+1}; g) \\ &\quad = \max_{g\in\mathbb{G}}\Big\{\widehat{\Phi}_{\pi,h}(b_h^\star(b_{h+1}), b_{h+1}; g) - \lambda\|g\|^2_{2,n}\Big\} \\ &\quad \le \max_{g\in\mathbb{G}}\Bigg\{\Phi_{\pi,h}(b_h^\star(b_{h+1}), b_{h+1}; g) - \lambda\|g\|_2^2 + \frac{\lambda}{2}\|g\|_2^2 + \frac{\lambda M_{\mathbb{G}}^2\log(4c_1/\zeta)}{2c_2 n}, \\ &\qquad + 18L\|g\|_2\sqrt{\frac{M_{\mathbb{G}}^2 \cdot \log(4c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{c_2 n}} + \frac{18L \cdot M_{\mathbb{G}}^2 \cdot \log(4c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{c_2 n}\Bigg\} \\ &\quad \le \max_{g\in\mathbb{G}} \Phi_{\pi,h}(b_h^\star(b_{h+1}), b_{h+1}; g) + \max_{g\in\mathbb{G}}\Bigg\{-\frac{\lambda}{2}\|g\|_2^2 + 18L\|g\|_2\sqrt{\frac{M_{\mathbb{G}}^2 \cdot \log(4c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{c_2 n}}\Bigg\} \\ &\qquad + \frac{\lambda M_{\mathbb{G}}^2\log(4c_1/\zeta)}{2c_2 n} + \frac{18L \cdot M_{\mathbb{G}}^2 \cdot \log(4c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{c_2 n} \\ &\quad \le \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}} + \frac{728L^2 \cdot M_{\mathbb{G}}^2 \cdot \log(4c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{\lambda n} + \frac{\lambda M_{\mathbb{G}}^2 \cdot \log(4c_1/\zeta)}{2c_2 n} \\ &\qquad + \frac{18L \cdot sM_{\mathbb{G}}^2 \cdot \log(4c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{c_2 n},\end{aligned}$$

where ζ is given as $\zeta = \min\{\delta, 4c_1 \exp(-c_2 n\alpha_{\mathbb{G},n}^2/M_{\mathbb{G}}^2)\}$ and ζ' is given as $\zeta' = \min\{\delta, 4c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H \exp(-c_2 n\alpha_{\mathbb{G},n}^2/M_{\mathbb{G}}^2)\}$ for any policy $\pi \in \Pi(\mathcal{H})$ and step $h \in [H]$. Here the last inequality holds from the fact that

$$\max_{g\in\mathcal{G}} \Phi_{\pi,h}(b_h^\star(b_{h+1}), b_{h+1}; g) \le \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}}, \tag{F.25}$$

and that $\sup_{\|g\|_2}\{a\|g\|_2 - b\|g\|_2^2\} \le a^2/4b$. Note that inequality (F.25) holds according to Assumption 4.3 and 4.3. In fact, by Assumption 4.3, we can first obtain by quadratic optimization that for $\lambda > 0$,

$$\max_{g\in\mathbb{G}} \Phi_{\pi,h}^\lambda(b_h, b_{h+1}) = \frac{1}{4\lambda}\mathcal{L}_h^\pi(b_h, b_{h+1}),$$

for any functions $b_h, b_{h+1} \in \mathbb{B}$. Thus we can equivalently express $b_h^\star(b_{h+1})$ as

$$b_h^\star(b_{h+1}) = \arg\min_{b\in\mathbb{B}} \frac{1}{4\lambda}\mathcal{L}_h^\pi(b, b_{h+1}) = \arg\min_{b\in\mathbb{B}} \mathcal{L}_h^\pi(b, b_{h+1}).$$

This further indicates the following bound on $\max_{g\in\mathbb{G}} \Phi_{\pi,h}(b_h^\star(b_{h+1}), b_{h+1}; g)$ that

$$\max_{g\in\mathbb{G}} \Phi_{\pi,h}(b_h^\star(b_{h+1}), b_{h+1}; g) \le \max_{g\in\mathbb{G}} \sqrt{\mathcal{L}_h(b_h^\star(b_{h+1}), b_{h+1}) \cdot \mathbb{E}_{\pi^b}[g(A_h, Z_h)^2]} \le \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}},$$

by Cauchy-Schwarz inequality and Assumption 4.3. Now according to the choice of ξ in Lemma D.2, using the fact that $\zeta < \zeta'$ and $L = 2M_{\mathbb{B}}$, we can conclude that, with probability at least $1 - \delta/2$,

$$\begin{aligned}
&\max_{g\in\mathbb{G}} \widehat{\Phi}_{\pi,h}^\lambda(b_h^\star(b_{h+1}), b_{h+1}^\pi; g) \\
&\quad \le \frac{728L^2 \cdot M_{\mathbb{G}}^2 \cdot \log(4c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{\lambda n} + \frac{\lambda M_{\mathbb{G}}^2 \cdot \log(4c_1/\zeta)}{2c_2 n} \\
&\qquad + \frac{18L \cdot M_{\mathbb{G}}^2 \cdot \log(4c_1|\mathbb{B}|^2|\Pi(\mathcal{H})|H/\zeta')}{c_2 n} + \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}} \\
&\quad \lesssim \mathcal{O}\left(\frac{(\lambda + 1/\lambda) \cdot M_{\mathbb{B}}^2 M_{\mathbb{G}}^2 \cdot \log(|\mathbb{B}||\Pi(\mathcal{H})|H/\zeta)}{n}\right) + \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}} \lesssim \xi + \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}}.
\end{aligned}$$

Therefore, we conclude the proof of Lemma F.1. □

G Proof of Theorem 4.4

Proof of Theorem 4.4. By the definition of $F(\mathbf{b})$ and $\widehat{F}(\mathbf{b})$ in (D.1) and the fact that $J(\pi) = F(\mathbf{b}^\pi)$ according to Theorem 3.3, we first have that

$$\begin{aligned}
&J(\pi^\star) - J(\widehat{\pi}) \\
&\quad = F(\mathbf{b}^{\pi^\star}) - F(\mathbf{b}^{\widehat{\pi}}) \\
&\quad = \underbrace{\left(F(\mathbf{b}^{\pi^\star}) - \widehat{F}(\mathbf{b}^{\pi^\star})\right)}_{\text{(i)}} + \underbrace{\left(F(\mathbf{b}^{\pi^\star}) - \widehat{F}(\mathbf{b}^{\widehat{\pi}})\right)}_{\text{(ii)}} + \underbrace{\left(\widehat{F}(\mathbf{b}^{\widehat{\pi}}) - F(\mathbf{b}^{\widehat{\pi}})\right)}_{\text{(iii)}}.
\end{aligned}$$

We can bound term (i) and term (iii) via uniform concentration inequalities, which we present latter. For term (ii), via Lemma D.2, with probability at least $1 - \delta$, $\mathbf{b}^{\pi^\star} \in \mathrm{CR}^{\pi^\star}(\xi)$ and $\mathbf{b}^{\widehat{\pi}} \in \mathrm{CR}^{\widehat{\pi}}(\xi)$, which indicates that

$$\text{(ii)} = \widehat{F}(\mathbf{b}^{\pi^\star}) - \widehat{F}(\mathbf{b}^{\widehat{\pi}}) \le \max_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)} \widehat{F}(\mathbf{b}) - \min_{\mathbf{b}\in\mathrm{CR}^{\widehat{\pi}}(\xi)} \widehat{F}(\mathbf{b}). \tag{G.1}$$

From (G.1), we can further bound term (ii) as

$$\begin{aligned}
\text{(ii)} &\le \max_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)} \widehat{F}(\mathbf{b}) - \max_{\pi\in\Pi(\mathcal{H})} \min_{\mathbf{b}\in\mathrm{CR}^{\pi}(\xi)} \widehat{F}(\mathbf{b}) \\
&\le \max_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)} \widehat{F}(\mathbf{b}) - \min_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)} \widehat{F}(\mathbf{b}) \\
&= \max_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)} \widehat{F}(\mathbf{b}) - \widehat{F}(\mathbf{b}^{\pi^\star}) + \widehat{F}(\mathbf{b}^{\pi^\star}) - \min_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)} \widehat{F}(\mathbf{b}) \\
&\le 2 \max_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)} \left|\widehat{F}(\mathbf{b}) - \widehat{F}(\mathbf{b}^{\pi^\star})\right|.
\end{aligned} \tag{G.2}$$

Here the first inequality holds because $\max_{\pi\in\Pi(\mathcal{H})}\min_{\mathbf{b}\in\mathrm{CR}^{\pi}(\xi)}\widehat{F}(\mathbf{b})=\min_{\mathbf{b}\in\mathrm{CR}^{\widehat{\pi}}(\xi)}\widehat{F}(\mathbf{b})$ by the definition of $\widehat{\pi}$ from (3.14). The second inequality holds because by definition $\pi^\star$ is the optimal policy in $\Pi(\mathcal{H})$. The third inequality is trivial. Now to further bound (G.2) by the RMSE loss defined in (3.6), we consider

$$
\begin{aligned}
&2\max_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)}\left|\widehat{F}(\mathbf{b})-\widehat{F}(\mathbf{b}^{\pi^\star})\right| \\
&\quad\le \underbrace{2\max_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)}\left|\widehat{F}(\mathbf{b})-F(\mathbf{b})\right|}_{\text{(iv)}}+\underbrace{2\max_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)}\left|F(\mathbf{b})-F(\mathbf{b}^{\pi^\star})\right|}_{\text{(v)}}+\underbrace{2\left|F(\mathbf{b}^{\pi^\star})-\widehat{F}(\mathbf{b}^{\pi^\star})\right|}_{\text{(vi)}},
\end{aligned}
$$

where we can bound term (iv) and term (vi) via uniform concentration inequalities, which we present latter. For term (v), we invoke Lemma D.1 and obtain that

$$
\text{(v)}\le 2\max_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)}\sum_{h=1}^{H}\gamma^{h-1}\sqrt{C^{\pi^\star}}\cdot\sqrt{\mathcal{L}_h^{\pi^\star}(b_h,b_{h+1})}\le 2\sqrt{C^{\pi^\star}}\sum_{h=1}^{H}\gamma^{h-1}\max_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)}\sqrt{\mathcal{L}_h^{\pi^\star}(b_h,b_{h+1})}.
$$

Now invoking Lemma D.3, with probability at least $1-\delta$, $\max_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)}\sqrt{\mathcal{L}_h^{\pi^\star}(b_h,b_{h+1})}$ is bounded by

$$
\max_{\mathbf{b}\in\mathrm{CR}^{\pi^\star}(\xi)}\sqrt{\mathcal{L}_h^{\pi}(b_h,b_{h+1})}\le\widetilde{C}_1M_{\mathbb{B}}M_{\mathbb{G}}\sqrt{\frac{(\lambda+1/\lambda)\log(|\mathbb{B}||\Pi(\mathcal{H})|H/\zeta)}{n}}+\widetilde{C}_1\epsilon_{\mathbb{B}}^{1/4}M_{\mathbb{G}}^{1/2},\quad\text{(G.3)}
$$

for each step $h\in[H]$, where $\zeta=\min\{\delta,c_1\exp(-c_2n\alpha_{\mathbb{G},n}^2)\}$. In the sequel, we turn to deal with term (i), (iii), (iv), and (vi), respectively. To this end, it suffices to apply uniform concentration inequalities to bound $F(\mathbf{b})$ and $\widehat{F}(\mathbf{b})$ uniformly over $\mathbf{b}\in\mathbb{B}^{\otimes H}$. By Hoeffding inequality, we have that, with probability at least $1-\delta$,

$$
\left|J(\pi,\mathbf{b})-\widehat{J}(\pi,\mathbf{b})\right|\le\sqrt{\frac{2M_{\mathbb{B}}^2\log(|\mathbb{B}|/\delta)}{n}},\quad\forall\pi\in\Pi(\mathcal{H}),\quad\forall\mathbf{b}\in\mathbb{B}^{\otimes H}.\qquad\text{(G.4)}
$$

Consequently, all of (i), (iii), (iv), and (vi) are bounded by the right hand side of (G.4). Finally, by combining (G.3) and (G.4), with probability at least $1-3\delta$, it holds that

$$
\begin{aligned}
J(\pi^\star)-J(\widehat{\pi})&\le\text{(i)}+\text{(iii)}+\text{(iv)}+\text{(vi)}+\text{(v)}\\
&\le 2\sqrt{C^{\pi^\star}}\sum_{h=1}^{H}\gamma^{h-1}\left(\widetilde{C}_1M_{\mathbb{B}}M_{\mathbb{G}}\sqrt{\frac{(\lambda+1/\lambda)\log(|\mathbb{B}||\Pi(\mathcal{H})|H/\zeta)}{n}}+\widetilde{C}_1\epsilon_{\mathbb{B}}^{1/4}M_{\mathbb{G}}^{1/2}\right)\\
&\qquad+4\sqrt{\frac{2M_{\mathbb{B}}^2\log(|\mathbb{B}|/\delta)}{n}}\\
&\le C_1'\sqrt{C^{\pi^\star}}\,(\lambda+1/\lambda)^{1/2}\,HM_{\mathbb{B}}M_{\mathbb{G}}\sqrt{\frac{\log(|\mathbb{B}||\Pi(\mathcal{H})|H/\zeta)}{n}}+C_1'\sqrt{C^{\pi^\star}}H\epsilon_{\mathbb{B}}^{1/4}M_{\mathbb{G}}^{1/2},
\end{aligned}
$$

for some problem-independent constant $C_1'>0$. We finish the proof of Theorem 4.4 by taking $\lambda=1$. □

H Details for Linear Function Approximation

H.1 Main Result for Linear Function Approximation

In this subsection, we extend Theorem 4.4 to primal function class $\mathbb{B}$, dual function class $\mathbb{G}$, and policy class $\Pi(\mathcal{H})$ with linear structures. The linear structure assumption is commonly considered in the RL literature (Jin et al., 2021; Xie et al., 2021; Zanette et al., 2021; Duan et al., 2021; Min et al., 2022a;b; Fei and Xu, 2022; Huang et al., 2023), to mention a few. And it can be viewed as an extension of linear bandits (Auer, 2002; Dani et al., 2008; Li et al., 2010; Abbasi-Yadkori et al., 2011; He et al., 2022) to multiple-horizon setting. Note that the exact detail of the linear structure assumption might change across different works. In our case, we consider linear function classes $\mathbb{B}_{\mathrm{lin}}$, $\mathbb{G}_{\mathrm{lin}}$ and Π_{lin}, which is characterized by the following definition.

Definition H.1 (Linear function approximation)**.** *Let $\phi : \mathcal{A} \times \mathcal{W} \to \mathbb{R}^d$ be a feature mapping for some integer $d \in \mathbb{N}$. We let the primal function class be $\mathbb{B} = \mathbb{B}_{\text{lin}}$ where*

$$\mathbb{B}_{\text{lin}} := \left\{ b \,\middle|\, b(\cdot, \cdot) = \langle \phi(\cdot, \cdot), \theta \rangle, \theta \in \mathbb{R}^d, \|\theta\|_2 \le L_b, \sup_{w \in \mathcal{W}} |\sum_{a \in \mathcal{A}} b(a, w)| \le M_{\mathbb{B}} \right\}.$$

Let $\psi = \{\psi_h : \mathcal{A} \times \mathcal{O} \times \mathcal{H}_{h-1} \to \mathbb{R}^d\}_{h=1}^H$ be H feature mappings. We let the policy function class be $\Pi(\mathcal{H}) = \Pi_{\text{lin}}$ where $\Pi_{\text{lin}} = \{\Pi_{\text{lin},h}\}_{h=1}^H$ and each $\Pi_{\text{lin},h}$ is defined as

$$\Pi_{\text{lin},h} := \left\{ \pi_h \,\middle|\, \pi_h(a|o, \tau) = \frac{\exp\left(\langle \psi_h(a, o, \tau), \beta \rangle\right)}{\sum_{a' \in \mathcal{A}} \exp(\langle \psi_h(a', o, \tau), \beta \rangle)}, \ \beta \in \mathbb{R}^d, \ \|\beta\|_2 \le L_\pi \right\}.$$

Finally, let $\nu : \mathcal{A} \times \mathcal{Z} \to \mathbb{R}^d$ be another feature mapping. We let the dual function class be $\mathbb{G} = \mathbb{G}_{\text{lin}}$ where

$$\mathbb{G}_{\text{lin}} := \left\{ g \mid g(\cdot, \cdot) = \langle \nu(\cdot, \cdot), \omega \rangle, \omega \in \mathbb{R}^d, \|\omega\|_2 \le L_g \right\}.$$

Assume without loss of generality that these feature mappings are normalized, i.e., $\|\phi\|_2, \|\psi\|_2, \|\nu\|_2 \le 1$.

We note that Definition H.1 is consistent with Assumption 4.2. One can see that $\mathbb{B}_{\text{lin}}$ and $\mathbb{G}_{\text{lin}}$ is uniformly bounded, $\mathbb{G}_{\text{lin}}$ is symmetric and star-shaped. And for other more detailed theoretical properties of $\mathbb{B}_{\text{lin}}$, $\mathbb{G}_{\text{lin}}$, and Π_{lin}, we refer the readers to Appendix H.2 for corresponding results.

Under linear function approximation, we can extend Theorem 4.4 to the following corollary, which characterizes the suboptimality (2.2) of $\widehat{\pi}$ found by `P3O` when using $\mathbb{B}_{\text{lin}}$, $\mathbb{G}_{\text{lin}}$, and Π_{lin} as function classes.

Corollary H.2 (Suboptimality analysis: linear function approximation)**.** *With linear function approximation (Definition H.1), under Assumption 3.1, 3.2, 4.1, and 4.3, by setting the regularization parameter λ and the confidence parameter ξ as $\lambda = 1$ and*

$$\xi = C_2 M_{\mathbb{B}}^2 \cdot M_{\mathbb{G}}^2 \cdot dH \cdot \log\left(1 + L_b L_\pi H n/\delta\right)/n,$$

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

$$\text{SubOpt}(\widehat{\pi}) \le C_2' \sqrt{C^{\pi^\star}} H M_{\mathbb{B}} L_g \sqrt{dH \log\left(1 + L_b L_\pi H n/\delta\right)/n} + C_2' \sqrt{C^{\pi^\star} L_g} H \epsilon_{\mathbb{B}}^{1/4}.$$

Here C_2 and C_2' are problem-independent universal constants.

Proof of Corollary H.2. See Appendix H.3 for a detailed proof. □

The guarantee of Corollary H.2 is structurally similar to that of Theorem 4.4, except that we can explicitly compute the complexity of the linear function classes and policy class. When $\epsilon_{\mathbb{B}} = 0$, `P3O` algorithm enjoys a $\widetilde{\mathcal{O}}(\sqrt{C^{\pi^\star} H^3 d/n})$ suboptimality under the linear function approximation. Compared to Theorem 4.4, Corollary H.2 does not explicitly assume Assumption 4.2 since it is implicitly satisfied by Definition H.1.

H.2 Auxiliary Results for Linear Function Approximation

Here we present results that bound the complexity of certain functions classes in the case of linear function approximation (Definition H.1).

Recall the definition of the bridge function class $\mathbb{B}^{\otimes H}$ where $\mathbb{B} = \mathbb{B}_{\text{lin}}$ is defined as

$$\mathbb{B}_{\text{lin}} := \left\{ b \,\middle|\, b(\cdot, \cdot) = \langle \phi(\cdot, \cdot), \theta \rangle, \theta \in \mathbb{R}^d, \|\theta\|_2 \le L_b, \sup_{w \in \mathcal{W}} \Big|\sum_{a \in \mathcal{A}} b(a, w)\Big| \le M_{\mathbb{B}} \right\}.$$

Denote by $\mathcal{N}_\epsilon^\infty(\mathbb{B})$ the ϵ-covering number of $\mathbb{B}$ with respect to the ℓ_∞ norm. That is, there exists a collection of functions $\{b_i\}_{i=1}^N$ with $N \le \mathcal{N}_\epsilon^\infty(\mathbb{B})$ such that for any $b \in \mathbb{B}$, we can find some $b' \in \{b_i\}_{i=1}^N$ satisfying

$$\|b - b'\|_\infty := \sup_{a \in \mathcal{A}, w \in \mathcal{W}} |b(a, w) - b'(a, w)| \le \epsilon.$$

Recall the policy function class $\Pi(\mathcal{H}) = \Pi_{\text{lin}}^{\otimes H}$ where Π_{lin} is defined as

$$\Pi_{\text{lin}} := \left\{ \pi \;\middle|\; \pi(a|o,\tau) = \frac{e^{\langle \psi(a,o,\tau),\beta \rangle}}{\sum_{a' \in \mathcal{A}} e^{\langle \psi(a',o,\tau),\beta \rangle}},\ \beta \in \mathbb{R}^d,\ \|\beta\|_2 \le L_\pi \right\}.$$

Denote by $\mathcal{N}_\epsilon^{\infty,1}(\Pi_{\text{lin}})$ the ϵ-covering number of Π_{lin} with respect to the $\ell_{\infty,1}$ norm, i.e.,

$$\|\pi - \pi'\|_{\infty,1} := \sup_{o \in \mathcal{O}, \tau \in \mathcal{H}} \sum_{a \in \mathcal{A}} |\pi(a|o,\tau) - \pi'(a|o,\tau)|.$$

The upper bounds for these covering numbers are given by the following lemma.

Lemma H.3 (Lemma 6 in Zanette et al. 2021)**.** *For any $\epsilon \in (0,1)$, we have*

$$\log \mathcal{N}_\epsilon^{\infty}(\mathbb{B}) \le d \log\left(1 + \frac{2L_b}{\epsilon}\right),$$
$$\log \mathcal{N}_\epsilon^{\infty,1}(\Pi_{\text{lin}}) \le d \log\left(1 + \frac{16L_\pi}{\epsilon}\right).$$

The ϵ-nets for the product function classes In the rest of Appendix H, due to the proof, we need to consider ϵ-nets defined for the product function classes $\mathbb{B}^{\otimes H}$ and $\Pi(\mathcal{H}) = \Pi_{\text{lin}}^{\otimes H}$. Specifically, for $\mathbb{B}^{\otimes H}$, we consider an ϵ-net of $\mathbb{B}^{\otimes H}$ defined in the following way: for any $\mathbf{b} = \{b_h\}_{h=1}^H \in \mathbb{B}^{\otimes H}$, there exists an $\mathbf{b}' = \{b'_h\}_{h=1}^H$ in the ϵ-net, such that

$$\|b_h - b'_h\|_\infty \le \epsilon.$$

By Lemma H.3, the cardinality of this ϵ-net is upper bounded by

$$\log \mathcal{N}_\epsilon^{\infty}(\mathbb{B}^{\otimes H}) \le dH \log\left(1 + \frac{2L_b}{\epsilon}\right).$$

Similarly, we consider an ϵ-net defined for $\Pi(\mathcal{H})$ defined as the following: for any $\pi = \{\pi_h\}_{h=1}^H \in \Pi(\mathcal{H})$, there exists an $\pi' = \{\pi'_h\}_{h=1}^H$ in the ϵ-net such that

$$\|\pi_h - \pi'_h\|_{\infty,1} \le \epsilon.$$

Then by Lemma H.3, the cardinality of this ϵ-net is upper bounded by

$$\log \mathcal{N}_\epsilon^{\infty,1}(\Pi(\mathcal{H})) \le dH \log\left(1 + \frac{16L_\pi}{\epsilon}\right)$$

For the dual function class $\mathbb{G}_{\text{lin}}$, recall the definition of the critical radius $\alpha_{\mathbb{G},n}$ in Assumption 4.2. The next lemma bound the critical radius of the linear dual function class $\mathbb{G} = \mathbb{G}_{\text{lin}}$.

Lemma H.4 (Lemma D.3 in Duan et al. 2021)**.** *For the function class $\mathbb{G}_{\text{lin}}$ defined in Definition H.1, its critical radius $\alpha_{\mathbb{G},n}$ satisfies*

$$\alpha_{\mathbb{G},n} = M_{\mathbb{G}} \sqrt{\frac{2d}{n}},$$

where $M_{\mathbb{G}} := \sup_{g \in \mathbb{G}_{\text{lin}}} \|g\|_\infty$.

H.3 Proof of Corollary H.2

We first introduce some lemmas needed for proving Corollary H.2. Their proof is deferred to Appendix H.4.1 and H.4.2.

Lemma H.5 (Alternative of Lemma D.2 in the linear case)**.** *Let the function, policy and dual function class $\mathbb{B} = \mathbb{B}_{\text{lin}}$, $\Pi(\mathcal{H}) = \Pi_{\text{lin}}$ and $\mathbb{G} = \mathbb{G}_{\text{lin}}$ be defined as in Definition H.1. Then under Assumption 3.2, 4.2, and 4.3, by setting ξ such that*

$$\xi = C_2 \cdot \left(\lambda + \frac{1}{\lambda}\right) \cdot \frac{M_{\mathbb{B}}^2 M_{\mathbb{G}}^2 dH \log(1 + L_b L_\pi Hn/\delta)}{n},$$

for some problem-independent universal constant $C_2 > 0$, it holds with probability at least $1-\delta$ that $\mathbf{b}^\pi \in \mathrm{CR}^\pi(\xi)$ for any policy $\pi \in \Pi(\mathcal{H})$.

Lemma H.6 (Alternative of Lemma D.3 in the linear case)**.** *Under Assumption 3.2, 4.2, 4.3, and 4.3, by setting the same ξ as in Lemma H.5, with probability at least $1-\delta$, for any policy $\pi \in \Pi(\mathcal{H})$, $\mathbf{b} \in \mathrm{CR}^{\pi}(\xi)$, and step h,*

$$\sqrt{\mathcal{L}_h^{\pi}(b_h, b_{h+1})} \leq \widetilde{C}_2 \cdot (1+\lambda)\, M_{\mathbb{B}} M_{\mathbb{G}} \cdot \sqrt{dH \log\left(1 + L_b L_\pi H n/\delta\right)/n} + \widetilde{C}_2 \cdot M_{\mathbb{G}}^{1/2} \epsilon_{\mathbb{B}}^{1/4},$$

for some problem-independent universal constant $\widetilde{C}_2 > 0$.

We are now ready to prove Corollary H.2.

Proof of Corollary H.2. We follow the proof of Theorem 4.4 and write

$$\begin{aligned} &J(\pi^\star) - J(\widehat{\pi}) \\ &\quad = \underbrace{\left(J(\pi^\star, \mathbf{b}^{\pi^\star}) - \widehat{J}(\pi^\star, \mathbf{b}^{\pi^\star})\right)}_{\text{(i)}} + \underbrace{\left(\widehat{J}(\pi^\star, \mathbf{b}^{\pi^\star}) - \widehat{J}(\widehat{\pi}, \mathbf{b}^{\widehat{\pi}})\right)}_{\text{(ii)}} + \underbrace{\left(\widehat{J}(\widehat{\pi}, \mathbf{b}^{\widehat{\pi}}) - J(\widehat{\pi}, \mathbf{b}^{\widehat{\pi}})\right)}_{\text{(iii)}}. \end{aligned} \tag{H.1}$$

We deal with term (ii) first. By Lemma H.5, with probability at least $1-\delta/2$, $\mathbf{b}^{\pi^\star} \in \mathrm{CR}^{\pi^\star}(\xi)$ and $\mathbf{b}^{\widehat{\pi}} \in \mathrm{CR}^{\widehat{\pi}}(\xi)$, which indicates that

$$\text{(ii)} = \widehat{J}(\pi^\star, \mathbf{b}^{\pi^\star}) - \widehat{J}(\widehat{\pi}, \mathbf{b}^{\widehat{\pi}}) \leq \max_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} \widehat{J}(\pi^\star, \mathbf{b}) - \min_{\mathbf{b} \in \mathrm{CR}^{\widehat{\pi}}(\xi)} \widehat{J}(\widehat{\pi}, \mathbf{b}).$$

Then following (G.2), we can upper bound term (ii) by

$$\begin{aligned} \text{(ii)} &\leq 2 \max_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} \left| \widehat{J}(\pi^\star, \mathbf{b}) - \widehat{J}(\pi^\star, \mathbf{b}^{\pi^\star}) \right| \\ &\leq \underbrace{2 \max_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} \left| \widehat{J}(\pi^\star, \mathbf{b}) - J(\pi^\star, \mathbf{b}) \right|}_{\text{(iv)}} + \underbrace{2 \max_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} \left| J(\pi^\star, \mathbf{b}) - J(\pi^\star, \mathbf{b}^{\pi^\star}) \right|}_{\text{(v)}} \\ &\quad + \underbrace{2 \left| J(\pi^\star, \mathbf{b}^{\pi^\star}) - \widehat{J}(\pi^\star, \mathbf{b}^{\pi^\star}) \right|}_{\text{(vi)}}. \end{aligned} \tag{H.2}$$

To bound term (v), we invoke Lemma D.1 which holds regardless of the underlying function classes and obtain that

$$\text{(v)} = 2 \max_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} \sum_{h=1}^{H} \gamma^{h-1} \sqrt{C^{\pi}} \cdot \sqrt{\mathcal{L}_h^{\pi}(b_h, b_{h+1})} \leq 2\sqrt{C^{\pi^\star}} \sum_{h=1}^{H} \gamma^{h-1} \max_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} \sqrt{\mathcal{L}_h^{\pi}(b_h, b_{h+1})}.$$

Now by Lemma H.6, with probability at least $1-\delta$, $\max_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} \sqrt{\mathcal{L}_h^{\pi}(b_h, b_{h+1})}$ is bounded by

$$\begin{aligned} &\max_{\mathbf{b} \in \mathrm{CR}^{\pi^\star}(\xi)} \sqrt{\mathcal{L}_h^{\pi}(b_h, b_{h+1})} \\ &\quad \leq \widetilde{C}_2 \cdot (1+\lambda)\, M_{\mathbb{B}} M_{\mathbb{G}} \cdot \sqrt{\frac{dH \log\left(1 + L_b L_\pi H n/\delta\right)}{n}} + \widetilde{C}_2 \cdot M_{\mathbb{G}}^{1/2} \epsilon_{\mathbb{B}}^{1/4}, \quad \forall h \in [H]. \end{aligned} \tag{H.3}$$

Now we deal with the term (i), (iii), (iv), and (vi), respectively. To this end, we apply uniform concentration inequalities to bound $J(\pi, \mathbf{b})$ and $\widehat{J}(\pi, \mathbf{b})$ uniformly over the ϵ-net of π and $\mathbf{b}$ as described in the proof of Lemma H.5. By Hoeffding's inequality, we have that, with probability at least $1-\delta$, for all π and $\mathbf{b}$ in their ϵ-nets,

$$\left| J(\pi, \mathbf{b}) - \widehat{J}(\pi, \mathbf{b}) \right| \leq \sqrt{\frac{2M_{\mathbb{B}}^2 \log(\mathcal{N}_{\epsilon,\mathbf{b}} \mathcal{N}_{\epsilon,\pi}/\delta)}{n}},$$

where $\mathcal{N}_{\epsilon,\pi}$ and $\mathcal{N}_{\epsilon,\mathbf{b}}$ are the covering numbers defined in Appendix H.2. Here we use the regularity assumption that $|\sum_{a \in \mathcal{A}} b_1^{\pi}(a, w)| \leq M_{\mathbb{B}}$ for all $w \in \mathcal{W}$ and the definition of $J(\pi, \mathbf{b})$ from (D.1). Consequently, for all $\pi \in \Pi(\mathcal{H})$ and $\mathbf{b} \in \mathbb{B}^{\otimes H}$, we have

$$\left| J(\pi, \mathbf{b}) - \widehat{J}(\pi, \mathbf{b}) \right| \leq \sqrt{\frac{2M_{\mathbb{B}}^2 \log(\mathcal{N}_{\epsilon,\mathbf{b}} \mathcal{N}_{\epsilon,\pi}/\delta)}{n}} + 2M_{\mathbb{B}}\epsilon. \tag{H.4}$$

Next, all of (i), (iii), (iv), and (vi) are bounded by the R.H.S. of (H.4). Finally, by (H.1), (H.2), (H.3) and (H.4), we have that

$$
\begin{aligned}
&J(\pi^\star) - J(\widehat{\pi}) \\
&\quad \leq \text{(i)} + \text{(iii)} + \text{(iv)} + \text{(vi)} + \text{(v)} \\
&\quad \leq 2\sqrt{C^{\pi^\star}} \sum_{h=1}^{H} \gamma^{h-1} \left[\widetilde{C}_2 \cdot (1+\lambda)\, M_{\mathbb{B}} M_{\mathbb{G}} \cdot \sqrt{\frac{dH \log\left(1 + L_b L_\pi H n/\delta\right)}{n}} + \widetilde{C}_2 \cdot M_{\mathbb{G}}^{1/2} \epsilon_{\mathbb{B}}^{1/4} \right] \\
&\qquad + 4 \left[\sqrt{\frac{2 M_{\mathbb{B}}^2 \log(\mathcal{N}_{\epsilon,\mathbf{b}} \mathcal{N}_{\epsilon,\pi}/\delta)}{n}} + 2 M_{\mathbb{B}} \epsilon \right].
\end{aligned}
$$

Finally, by taking $\epsilon = 1/n^2$, and plugging in the values of $\mathcal{N}_{\epsilon,\mathbf{b}}$ and $\mathcal{N}_{\epsilon,\pi}$ from Lemma H.3, we get

$$
\begin{aligned}
&J(\pi^\star) - J(\widehat{\pi}) \\
&\quad \leq 2\sqrt{C^{\pi^\star}} \sum_{h=1}^{H} \gamma^{h-1} \left[\widetilde{C}_2 \cdot (1+\lambda)\, M_{\mathbb{B}} M_{\mathbb{G}} \cdot \sqrt{\frac{dH \log\left(1 + L_b L_\pi H n/\delta\right)}{n}} + \widetilde{C}_2 \cdot M_{\mathbb{G}}^{1/2} \epsilon_{\mathbb{B}}^{1/4} \right] \\
&\qquad + C_3 M_{\mathbb{B}} \sqrt{\frac{dH \log(1 + L_b L_\pi n/\delta)}{n}},
\end{aligned}
$$

where C_3 is some problem-independent universal constant. We then simplify the expression and use the fact that

$$
M_{\mathbb{G}} = \sup_{a,z} |g(a,z)| = \sup_{a,z} |\langle \nu(a,z), \omega \rangle| \leq \sup_{a,z} \|\nu(a,z)\|_2 \cdot \|\omega\|_2 \leq L_g.
$$

This gives the result of Corollary H.2. □

H.4 Proof of Lemmas in Appendix H

H.4.1 Proof of Lemma H.5

Proof of Lemma H.5. First, for any $\epsilon \in (0,1)$, consider arbitrary $\pi = \{\pi_h\}_{h=1}^H$ and $\pi' = \{\pi'_h\}_{h=1}^H$ in Π_{lin} such that $\|\pi_h - \pi'_h\|_{\infty,1} \leq \epsilon$ for all $h \in [H]$. And consider arbitrary $\mathbf{b} = \{b_h\}_{h=1}^H$ and $\mathbf{b}' = \{b'_h\}_{h=1}^H$ in $\mathbb{B}^{\otimes H}$ such that $\|b_h - b'_h\|_\infty \leq \epsilon$ for all $h \in [H]$. Then by definition of $\Phi^\lambda_{\pi,h}(b_h, b_{h+1}; g)$ in (3.9) and $\widehat{\Phi}^\lambda_{\pi,h}(b_h, b_{h+1}; g)$ in (3.10), and that $\Phi^\lambda_{\pi,h} = \Phi^0_{\pi,h}$ and $\widehat{\Phi}^\lambda_{\pi,h} = \widehat{\Phi}^0_{\pi,h}$, one can easily get that

$$
\begin{aligned}
\left| \Phi_{\pi,h}(b_h, b_{h+1}; g) - \Phi_{\pi',h}(b'_h, b'_{h+1}; g) \right| &\leq \left[2\epsilon + \gamma \cdot (\epsilon + \epsilon M_{\mathbb{B}}) \right] \cdot M_{\mathbb{G}} \leq 4 M_{\mathbb{B}} M_{\mathbb{G}} \epsilon, \\
\left| \widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \widehat{\Phi}_{\pi',h}(b'_h, b'_{h+1}; g) \right| &\leq \left[2\epsilon + \gamma \cdot (\epsilon + \epsilon M_{\mathbb{B}}) \right] \cdot M_{\mathbb{G}} \leq 4 M_{\mathbb{B}} M_{\mathbb{G}} \epsilon,
\end{aligned} \tag{H.5}
$$

for all $g \in \mathbb{G}$.

Now, same as in the proof of Lemma D.2, we want to show: for any $\pi \in \Pi(\mathcal{H})$,

$$
\max_{g \in \mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(b^\pi_h, b^\pi_{h+1}; g) \leq \xi.
$$

The rest of the proof would be very similar to that of Lemma D.2 with an additional covering argument. To begin with, we again write $\widehat{\Phi}^\lambda_{\pi,h}(b^\pi_h, b^\pi_{h+1}; g) = \widehat{\Phi}_{\pi,h}(b^\pi_h, b^\pi_{h+1}; g) - \lambda \|g\|^2_{2,n}$.

Same as (F.10), we have that with probability at least $1 - \delta/2$,

$$
\left| \|g\|^2_{2,n} - \|g\|^2_2 \right| \leq \frac{1}{2} \|g\|^2_2 + \frac{M_{\mathbb{G}}^2 \log(2c_1/\zeta)}{2c_2 n}, \quad \forall g \in \mathbb{G}, \tag{H.6}
$$

where $\zeta = \min\{\delta, 2c_1 \exp(-c_2 n \alpha^2_{\mathbb{G},n}/M^2_{\mathbb{G}})\}$ and c_1, c_2 are some universal constants.

Next, we upper bound $|\widehat{\Phi}_{\pi',h}(b_h, b_{h+1}; g) - \Phi_{\pi',h}(b_h, b_{h+1}; g)|$ for any $\pi \in \Pi(\mathcal{H})$, and $\mathbf{b} \in \mathbb{B}^{\otimes H}$. We first prove this for a fixed ϵ-net of $\Pi(\mathcal{H})$ and $\mathbb{B}^{\otimes H}$. Specifically, choose an ϵ-net of $\Pi(\mathcal{H})$ such that for any $\pi = \{\pi_h\}_{h=1}^H$ and $\pi' = \{\pi'_h\}_{h=1}^H$ in this ϵ-net, it holds that $\|\pi_h - \pi'_h\|_{\infty,1} \leq \epsilon$ for all

h. Also choose an ϵ-net of $\mathbb{B}^{\otimes H}$ such that for any $\mathbf{b} = \{b_h\}_{h=1}^H$ and $\mathbf{b}' = \{b'_h\}_{h=1}^H$ in the ϵ-net, it holds that $\|b_h - b'_h\|_\infty \le \epsilon$ for all h. Denote the cardinality of these two ϵ-net by $\mathcal{N}_{\epsilon,\pi}$ and $\mathcal{N}_{\epsilon,\mathbf{b}}$, respectively. Then by the same argument behind (F.11), we get that, with probability at least $1-\delta/2$, for any π and $\mathbf{b}$ in their ϵ-nets, and for any $g \in \mathbb{G}$,

$$\begin{aligned}&\left|\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \Phi_{\pi,h}(b_h, b_{h+1}; g)\right| \\ &\qquad \le 18L\|g\|_2\sqrt{\frac{M_{\mathbb{G}}^2 \log\left(2c_1\mathcal{N}_{\epsilon,\mathbf{b}}^2\mathcal{N}_{\epsilon,\pi}H/\zeta\right)}{c_2 n}} + \frac{18LM_{\mathbb{G}}^2 \log\left(2c_1\mathcal{N}_{\epsilon,\mathbf{b}}^2\mathcal{N}_{\epsilon,\pi}H/\zeta\right)}{c_2 n},\end{aligned} \tag{H.7}$$

where $\zeta' = \min\{\delta, 2c_1\mathcal{N}_{\epsilon,\mathbf{b}}^2\mathcal{N}_{\epsilon,\pi}H\exp(-c_2 n\alpha_{\mathbb{G},n}^2/M_{\mathbb{G}}^2)\}$.

Now for any $\pi \in \Pi(\mathcal{H})$ and $\mathbf{b} \in \mathbb{B}^{\otimes H}$, by our construction of the ϵ-nets, we can find a π' and $\mathbf{b}'$ in the ϵ-nets such that $\|\pi_h - \pi'_h\|_{\infty,1} \le \epsilon$ and $\|b_h - b'_h\|_\infty \le \epsilon$ for all h. Then we have that with probability at least $1-\delta/2$, for any $\pi \in \Pi(\mathcal{H})$ and $\mathbf{b} \in \mathbb{B}^{\otimes H}$, and for any $g \in \mathbb{G}$,

$$\begin{aligned}&\left|\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \Phi_{\pi,h}(b_h, b_{h+1}; g)\right| \\ &\quad \le |\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \widehat{\Phi}_{\pi',h}(b'_h, b'_{h+1}; g)| + |\widehat{\Phi}_{\pi',h}(b'_h, b'_{h+1}; g) - \Phi_{\pi',h}(b'_h, b'_{h+1}; g)| \\ &\qquad + |\Phi_{\pi',h}(b'_h, b'_{h+1}; g) - \Phi_{\pi,h}(b_h, b_{h+1}; g)| \\ &\quad \le 8M_{\mathbb{B}}M_{\mathbb{G}}\cdot\epsilon + 18L\|g\|_2\sqrt{\frac{M_{\mathbb{G}}^2 \log\left(2c_1\mathcal{N}_{\epsilon,\mathbf{b}}^2\mathcal{N}_{\epsilon,\pi}H/\zeta\right)}{c_2 n}} + \frac{18LM_{\mathbb{G}}^2 \log\left(2c_1\mathcal{N}_{\epsilon,\mathbf{b}}^2\mathcal{N}_{\epsilon,\pi}H/\zeta\right)}{c_2 n},\end{aligned} \tag{H.8}$$

where the first step is by the triangle inequality and the second steps is by (H.5) and (H.7).

Now combine (H.6) and (H.8) with a union bound, we conclude that, with probability at least $1-\delta$, for any $\pi \in \Pi(\mathcal{H})$,

$$\begin{aligned}&\max_{g\in\mathbb{G}} \widehat{\Phi}_{\pi,h}^\lambda(b_h^\pi, b_{h+1}^\pi; g) \\ &\quad = \max_{g\in\mathbb{G}}\left\{\widehat{\Phi}_{\pi,h}(b_h^\pi, b_{h+1}^\pi; g) - \lambda\|g\|_{2,n}^2\right\} \\ &\quad \le \max_{g\in\mathbb{G}}\Bigg\{\Phi_{\pi,h}(b_h^\pi, b_{h+1}^\pi; g) - \lambda\|g\|_2^2 + \frac{\lambda}{2}\|g\|_2^2 + \frac{\lambda M_{\mathbb{G}}^2\log(2c_1/\zeta)}{2c_2 n}, \\ &\qquad + 18L\|g\|_2\sqrt{\frac{M_{\mathbb{G}}^2 \log\left(2c_1\mathcal{N}_{\epsilon,\mathbf{b}}^2\mathcal{N}_{\epsilon,\pi}H/\zeta\right)}{c_2 n}} + \frac{18LM_{\mathbb{G}}^2 \log\left(2c_1\mathcal{N}_{\epsilon,\mathbf{b}}^2\mathcal{N}_{\epsilon,\pi}H/\zeta\right)}{c_2 n}\Bigg\} + 8M_{\mathbb{B}}M_{\mathbb{G}}\epsilon \\ &\quad \le \max_{g\in\mathbb{G}}\Phi_{\pi,h}(b_h^\pi, b_{h+1}^\pi; g) + \max_{g\in\mathbb{G}}\left\{-\frac{\lambda}{2}\|g\|_2^2 + 18L\|g\|_2\sqrt{\frac{M_{\mathbb{G}}^2 \log\left(2c_1\mathcal{N}_{\epsilon,\mathbf{b}}^2\mathcal{N}_{\epsilon,\pi}H/\zeta\right)}{c_2 n}}\right\} \\ &\qquad + \frac{\lambda M_{\mathbb{G}}^2\log(2c_1/\zeta)}{2c_2 n} + \frac{18LM_{\mathbb{G}}^2 \log\left(2c_1\mathcal{N}_{\epsilon,\mathbf{b}}^2\mathcal{N}_{\epsilon,\pi}H/\zeta\right)}{c_2 n} + 8M_{\mathbb{B}}M_{\mathbb{G}}\epsilon \\ &\quad \le \frac{728L^2M_{\mathbb{G}}^2\log(2c_1\mathcal{N}_{\epsilon,\mathbf{b}}^2\mathcal{N}_{\epsilon,\pi}H/\zeta')}{\lambda n} + \frac{\lambda M_{\mathbb{G}}^2\log(2c_1/\zeta)}{2c_2 n} \\ &\qquad + \frac{18LM_{\mathbb{G}}^2\log(2c_1\mathcal{N}_{\epsilon,\mathbf{b}}^2\mathcal{N}_{\epsilon,\pi}H/\zeta')}{c_2 n} + 8M_{\mathbb{B}}M_{\mathbb{G}}\epsilon,\end{aligned} \tag{H.9}$$

with $\zeta = \min\{\delta, 2c_1\exp(-c_2 n\alpha_{\mathbb{G},n}^2/M_{\mathbb{G}}^2)\}$ and $\zeta' = \min\{\delta, 2c_1\mathcal{N}_{\epsilon,\mathbf{b}}^2\mathcal{N}_{\epsilon,\pi}H\exp(-c_2 n\alpha_{\mathbb{G},n}^2/M_{\mathbb{G}}^2)\}$ for any policy $\pi \in \Pi(\mathcal{H})$ and step h. Here the first inequality is by (H.6) and (H.8), the second inequality is trivial, and the last inequality holds from the fact that $\Phi_{\pi,h}(b_h^\pi, b_{h+1}^\pi; g) = 0$ and the fact that $\sup_{\|g\|_2}\{a\|g\|_2 - b\|g\|_2^2\} \le a^2/4b$.

Now by Definition H.1, we apply Lemma H.3 with $\|\theta_h\|_2 \le L_b$ and $\|\beta_h\| \le L_\pi$ and get that

$$\log \mathcal{N}_{\epsilon,\pi} \le dH \log\left(1 + \frac{16 L_\pi}{\epsilon}\right),$$
$$\log \mathcal{N}_{\epsilon,b} \le dH \log\left(1 + \frac{2 L_b}{\epsilon}\right). \tag{H.10}$$

Now we pick $\epsilon = 1/n^2$, and together with (H.9) and (H.10), we get that

$$\max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(b^\pi_h, b^\pi_{h+1}; g) \le C \cdot \frac{(\lambda + 1/\lambda)\, M_{\mathbb{B}}^2 M_{\mathbb{G}}^2 \left[dH \log\left(1 + L_b L_\pi H n/\delta\right) + n\alpha^2_{\mathbb{G},n}/M_{\mathbb{G}}^2\right]}{n} + C \cdot \frac{M_{\mathbb{B}} M_{\mathbb{G}}}{n^2},$$

where C is some universal constant. Here we have plugged in the value of ζ, ζ' and $L = 2M_{\mathbb{B}}$. Finally, by plugging in the value of $\alpha_{\mathbb{G},n}$ from Lemma H.4, we conclude that

$$\max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(b^\pi_h, b^\pi_{h+1}; g) \le C_1 \cdot \left(\lambda + \frac{1}{\lambda}\right) \cdot \frac{M_{\mathbb{B}}^2 M_{\mathbb{G}}^2 dH \log\left(1 + L_b L_\pi H n/\delta\right)}{n} + C_1 \cdot \frac{M_{\mathbb{B}} M_{\mathbb{G}}}{n^2},$$

where C_1 is some problem-independent constant. Note that second term on the right hand side is smaller than the first term. Then the result follows from our choice of ξ in Lemma H.5.

□

H.4.2 Proof of Lemma H.6

Proof of Lemma H.6. Consider any $\pi \in \Pi(\mathcal{H})$ and $\mathbf{b} = \{b_h\}_{h=1}^H \in \mathrm{CR}^\pi(\xi)$. Same as (F.12), we have

$$\begin{aligned}
&\max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(b_h, b_{h+1}; g) \\
&\quad \ge \underbrace{\max_{g\in\mathbb{G}} \left\{\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \widehat{\Phi}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) - 2\lambda\|g\|^2_{2,n}\right\}}_{(\star)} \\
&\qquad - \max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g).
\end{aligned}$$

We again upper and lower bound term $(\star)$ respectively.

Upper bound of term $(\star)$. By the same argument as in the proof of Lemma F.1, we have that: for any $\mathbf{b} \in \mathbb{B}^{\otimes H}$, $\pi \in \Pi(\mathcal{H})$, and $h \in [H]$, it holds with probability at least $1 - \delta/2$ that

$$\max_{g\in\mathbb{G}} \widehat{\Phi}_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) \le \xi + \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}},$$

where $b^\star_h(b_{h+1})$ is defined in (F.3) and ξ is defined in Lemma H.5. We then get

$$\max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(\widehat{b}_h(b_{h+1}), b_{h+1}; g) \le \max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) \le \xi + \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}}, \tag{H.11}$$

where the first inequality follows from the definition of $\widehat{b}_h(b_{h+1})$ in (3.11). Also note that, by the construction of the confidence region $\mathrm{CR}^\pi(\xi)$, we have

$$\max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(b_h, b_{h+1}; g) - \max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(\widehat{b}_h(b_{h+1}), b_{h+1}; g) \le \xi. \tag{H.12}$$

Furthermore, we can write

$$\begin{aligned}
(\star) &\le \max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) + \max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(b_h, b_{h+1}; g) \\
&\le \max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(b^\star_h(b_{h+1}), b_{h+1}; g) \\
&\quad + \max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(b_h, b_{h+1}; g) - \max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(\widehat{b}_h(b_{h+1}), b_{h+1}; g) \\
&\quad + \max_{g\in\mathbb{G}} \widehat{\Phi}^\lambda_{\pi,h}(\widehat{b}_h(b_{h+1}), b_{h+1}; g).
\end{aligned}$$

Combining with (H.11) and (H.12), we get that, with probability at least $1-\delta/2$,

$$(\star) \leq 3\xi + 2\epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}}. \tag{H.13}$$

Lower bound of term $(\star)$. First of all, same as (F.17), it holds with probability at least $1-\delta/4$ that,

$$\big| \|g\|_{2,n}^2 - \|g\|_2^2 \big| \leq \frac{1}{2}\|g\|_2^2 + \frac{M_{\mathbb{G}}^2 \log(4c_1/\zeta)}{2c_2 n}, \quad \forall g \in \mathbb{G}, \tag{H.14}$$

where $\zeta = \min\{\delta, 4c_1 \exp(-c_2 n \alpha_{\mathbb{G},n}^2 / M_{\mathbb{G}}^2)\}$ for some absolute constants c_1 and c_2, and $\alpha_{\mathbb{G},n}$ is the critical radius of $\mathbb{G}$ defined in Assumption 4.2.

Second, we fix an ϵ-net of $\Pi(\mathcal{H})$ and an ϵ-net of $\mathbb{B}^{\otimes H}$, as described in Appendix H.2. Denote by $\mathcal{N}_{\epsilon,\pi}$ and $\mathcal{N}_{\epsilon,\mathbf{b}}$ their respective covering numbers. Then by the same argument behind (F.18) and a union bound, we get that, with probability at least $1-\delta/4$, for all $\pi = \{\pi_h\}_{h=1}^H$, $\mathbf{b} = \{b_h\}_{h=1}^H$ and $\mathbf{b}' = \{b_h'\}_{h=1}^H$ in their ϵ-nets, and for all $g \in \mathbb{G}$,

$$\begin{aligned} &\Big| \Big(\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \widehat{\Phi}_{\pi,h}(b_h', b_{h+1}; g) \Big) - \Big(\Phi_{\pi,h}(b_h, b_{h+1}; g) - \Phi_{\pi,h}(b_h', b_{h+1}; g) \Big) \Big| \\ &\quad \leq 18L\|g\|_2 \sqrt{\frac{M_{\mathbb{G}}^2 \log(4c_1 \mathcal{N}_{\epsilon,\mathbf{b}}^3 \mathcal{N}_{\epsilon,\pi} H/\zeta')}{c_2 n}} + \frac{18LM_{\mathbb{G}}^2 \log\big(4c_1 \mathcal{N}_{\epsilon,\mathbf{b}}^3 \mathcal{N}_{\epsilon,\pi} H/\zeta'\big)}{c_2 n}, \end{aligned} \tag{H.15}$$

where $\zeta' = \min\{\delta, 4c_1 \mathcal{N}_{\epsilon,\mathbf{b}}^3 \mathcal{N}_{\epsilon,\pi} H \exp(-c_2 n \alpha_{\mathbb{G},n}^2 / M_{\mathbb{G}}^2)\}$.

We then use (H.5), and conclude that, with probability at least $1-\delta/4$, for all $\pi \in \Pi(\mathcal{H})$, and $\mathbf{b}$, $\mathbf{b}' \in \mathbb{B}^{\otimes H}$, and $g \in \mathbb{G}$,

$$\begin{aligned} &\Big| \Big(\widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \widehat{\Phi}_{\pi,h}(b_h', b_{h+1}; g) \Big) - \Big(\Phi_{\pi,h}(b_h, b_{h+1}; g) - \Phi_{\pi,h}(b_h', b_{h+1}; g) \Big) \Big| \\ &\quad \leq 18L\|g\|_2 \sqrt{\frac{M_{\mathbb{G}}^2 \log(4c_1 \mathcal{N}_{\epsilon,\mathbf{b}}^3 \mathcal{N}_{\epsilon,\pi} H/\zeta')}{c_2 n}} + \frac{18LM_{\mathbb{G}}^2 \log\big(4c_1 \mathcal{N}_{\epsilon,\mathbf{b}}^3 \mathcal{N}_{\epsilon,\pi} H/\zeta'\big)}{c_2 n} + 8M_{\mathbb{B}} M_{\mathbb{G}} \epsilon. \end{aligned} \tag{H.16}$$

In the sequel, for simplicity, we denote that

$$\iota_n := \sqrt{\frac{M_{\mathbb{G}}^2 \log(4c_1 \mathcal{N}_{\epsilon,\mathbf{b}}^3 \mathcal{N}_{\epsilon,\pi} H/\zeta')}{c_2 n}}, \quad \iota_n' := \sqrt{\frac{M_{\mathbb{G}}^2 \log(4c_1/\zeta)}{2c_2 n}}, \tag{H.17}$$

where ζ and ζ' are same as in (H.14) and (H.15). Furthermore, given fixed $b_h, b_{h+1} \in \mathbb{B}$, we denote

$$g_h^\pi := \frac{1}{2\lambda} \ell_h^\pi(b_h, b_{h+1}) \in \mathbb{G}, \tag{H.18}$$

where ℓ_h^π is defined by (F.1) and $g_h^\pi \in \mathbb{G}$ follows from Assumption 4.3. We then have

$$\begin{aligned} (\star) &= \max_{g \in \mathbb{G}} \Big\{ \widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g) - \widehat{\Phi}_{\pi,h}(b_h^\star(b_{h+1}), b_{h+1}; g) - 2\lambda\|g\|_{2,n}^2 \Big\} \\ &\geq \widehat{\Phi}_{\pi,h}(b_h, b_{h+1}; g_h^\pi/2) - \widehat{\Phi}_{\pi,h}(b_h^\star(b_{h+1}), b_{h+1}; g_h^\pi/2) - \frac{\lambda}{2}\|g_h^\pi\|_{2,n}^2, \end{aligned}$$

where the inequality holds because $g_h^\pi/2 \in \mathbb{G}$.

Together with (H.14) and (H.16), we have

$$\begin{aligned} (\star) &\geq \Phi_{\pi,h}(b_h, b_{h+1}; g_h^\pi/2) - \Phi_{\pi,h}(b_h^\star(b_{h+1}), b_{h+1}; g_h^\pi/2) - 18L\iota_n\|g_h^\pi\|_2 - 18L\iota_n^2 \\ &\qquad - 8M_{\mathbb{B}} M_{\mathbb{G}} \epsilon - \frac{\lambda}{2}\left(\frac{3}{2}\|g_h^\pi\|_2^2 + \iota_n'^2 \right) \\ &\geq \lambda\|g_h^\pi\|_2^2 - 18L\iota_n\|g_h^\pi\|_2 - \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}} - 18L\iota_n^2 - 8M_{\mathbb{B}} M_{\mathbb{G}} \epsilon - \frac{\lambda}{2}\left(\frac{3}{2}\|g_h^\pi\|_2^2 + \iota_n'^2 \right) \\ &= \frac{\lambda}{4}\|g_h^\pi\|_2^2 - 18L\iota_n\|g_h^\pi\|_2 - \epsilon_{\mathbb{B}}^{1/2} M_{\mathbb{G}} - 18L\iota_n^2 - 8M_{\mathbb{B}} M_{\mathbb{G}} \epsilon - \frac{\lambda}{2}\iota_n'^2, \end{aligned} \tag{H.19}$$

where the second inequality follows from the same reason as in (F.20).

Finally, combine (H.19) and (H.13) and we get

$$\frac{\lambda}{4}\|g_h^\pi\|_2^2 - 18L\iota_n\|g_h^\pi\|_2 - \epsilon_{\mathbb{B}}^{1/2}M_{\mathbb{G}} - 18L\iota_n^2 - 8M_{\mathbb{B}}M_{\mathbb{G}}\epsilon - \frac{\lambda}{2}\iota_n'^2 \le 3\xi + 2\epsilon_{\mathbb{B}}^{1/2}M_{\mathbb{G}}.$$

This gives the following quadratic inequality w.r.t. $\|g_h^\pi\|_2$

$$\lambda\|g_h^\pi\|_2^2 - \underbrace{72L\iota_n}_{\mathrm{A}}\|g_h^\pi\|_2 - \underbrace{4\left[18L\iota_n^2 + \frac{\lambda}{2}{\iota_n'}^2 + 3\xi + 8M_{\mathbb{B}}M_{\mathbb{G}}\epsilon + 3\epsilon_{\mathbb{B}}^{1/2}M_{\mathbb{G}}\right]}_{\mathrm{B}} \le 0.$$

By the fact that $x^2 - \mathrm{A}x - \mathrm{B} \le 0$ implies $x \le (\mathrm{A} + \sqrt{\mathrm{A}^2 + 4\mathrm{B}})/2 \le \mathrm{A} + \sqrt{\mathrm{B}}$, we have

$$\|g_h^\pi\|_2 \le \frac{72L\iota_n}{\lambda} + \sqrt{\frac{4}{\lambda}\left[18L\iota_n^2 + \frac{\lambda}{2}{\iota_n'}^2 + 3\xi + 8M_{\mathbb{B}}M_{\mathbb{G}}\epsilon + 3\epsilon_{\mathbb{B}}^{1/2}M_{\mathbb{G}}\right]}.$$

We then plug in the values of ι_n and ι_n' from (H.17), ξ from Lemma H.5, ζ and ζ' from below (H.14) and (H.15), $\mathcal{N}_{\epsilon,\mathbf{b}}$ and $\mathcal{N}_{\epsilon,\pi}$ from Lemma H.3, $\alpha_{\mathbb{G},n}$ from Lemma H.4, and set $\epsilon = 1/n^2$. Simplify the expression and we get

$$\|g_h^\pi\|_2 \le C\cdot\left(1+\frac{1}{\lambda}\right)M_{\mathbb{B}}M_{\mathbb{G}}\cdot\sqrt{\frac{dH\log\left(1+L_bL_\pi Hn/\delta\right)}{n}} + C\cdot\frac{M_{\mathbb{G}}^{1/2}\epsilon_{\mathbb{B}}^{1/4}}{\lambda},$$

where C is some problem-independent universal constant. By (H.18) and (3.6), we have $\mathcal{L}_h^\pi(b_h, b_{h+1}) = \|2\lambda g_h^\pi\|_2^2$. It follows that

$$\sqrt{\mathcal{L}_h^\pi(b_h, b_{h+1})} = 2\lambda\|g_h^\pi\|_2 \le \widetilde{C}_2\cdot(1+\lambda)M_{\mathbb{B}}M_{\mathbb{G}}\cdot\sqrt{\frac{dH\log\left(1+L_bL_\pi Hn/\delta\right)}{n}} + \widetilde{C}_2\cdot M_{\mathbb{G}}^{1/2}\epsilon_{\mathbb{B}}^{1/4},$$

for some constant $\widetilde{C}_2$. This finishes the proof. □

I Auxiliary Lemmas

We introduce some useful lemmas for the uniform concentration over function classes. Before we present the lemmas, we first introduce several notations. For a function class $\mathcal{F}$ on a probability space $(\mathcal{X}, P)$, we denote by $\|f\|_2^2$ the expectation of $f(X)^2$, that is $\|f\|_2^2 = \mathbb{E}_{X\sim\mathcal{P}}[f(X)^2]$. Also, we denote by

$$\mathcal{R}_n(\mathcal{F},\delta) := \mathbb{E}\left[\sup_{f\in\mathcal{F}:\|f\|_2\le\delta}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] \tag{I.1}$$

the localized Rademacher complexity of $\mathcal{F}$ with scale $\delta > 0$ and size $n \in \mathbb{N}$. Here $\{\epsilon_i\}_{i=1}^b$ and $\{X_i\}_{i=1}^n$ are i.i.d. and independent. Each ϵ_i is uniformly distributed on $\{+1, -1\}$ and each X_i is distributed according to P. Finally, we denote by star($\mathcal{F}$) the star-shaped set induced by set $\mathcal{F}$ as

$$\mathrm{star}(\mathcal{F}) = \{\alpha f : \alpha\in[0,1],\ f\in\mathcal{F}\}. \tag{I.2}$$

Now we are ready to present the lemmas for uniform concentration inequalities.

Lemma I.1 (Localized Uniform Concentration 1 (Wainwright, 2019))**.** *Given a star-shaped and b-uniformly bounded function class $\mathcal{F}$, let δ_n be any positive solution of the inequality*

$$\mathcal{R}_n(\mathcal{F};\delta) \le \frac{\delta^2}{b}.$$

Then for any $t \ge \delta_n$, we have that

$$\left|\|f\|_n^2 - \|f\|_2^2\right| \le \frac{1}{2}\|f\|_2^2 + \frac{1}{2}t^2, \quad \forall f\in\mathcal{F}$$

with probability at least $1 - c_1\exp(-c_2nt^2/b^2)$. If in addition $n\delta_n^2 \ge 2\log\left(4\log\left(1/\delta_n\right)\right)/c_2$, then we have that

$$\left|\|f\|_n - \|f\|_2\right| \le c_0\delta_n, \quad \forall f\in\mathcal{F}$$

with probability at least $1 - c_1'\exp(-c_2'n\delta_n^2/b^2)$.

Proof of Lemma I.1. See Theorem 14.1 of Wainwright (2019) for a detailed proof. □

Lemma I.2 (Localized Uniform Concentration 2 (Foster and Syrgkanis, 2019))**.** *Consider a star-shaped function class $\mathcal{F} : \mathcal{X} \mapsto \mathbb{R}$ with $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq b$, and pick any $f^\star \in \mathcal{F}$. Also, consider a loss function $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}$ which is L-Lipschitz in its first argument with respect to the $\|\cdot\|_2$-norm. Now let $\delta_n^2 \geq 4 \log(41 \log(2c_2 n)) / (c_2 n)$ be any solution to the inequality:*

$$\mathcal{R}_n(\mathrm{star}(\mathcal{F} - f^\star); \delta) \leq \frac{\delta^2}{b}.$$

Then for any $t \geq \delta_n$ and some absolute constants $c_1, c_2 > 0$, with probability $1 - c_1 \exp(-c_2 n t^2 / b^2)$ it holds that

$$\begin{aligned} &\left| \left(\widehat{\mathbb{E}}_n[\ell(f(x), y)] - \widehat{\mathbb{E}}_n\left[\ell\left(f^\star(x), y\right)\right] \right) - \left(\mathbb{E}[\ell(f(x), y)] - \mathbb{E}\left[\ell\left(f^\star(x), y\right)\right] \right) \right| \\ &\qquad \leq 18 L t \left(\|f - f^\star\|_2 + t \right), \end{aligned} \tag{I.3}$$

for any $f \in \mathcal{F}$. If furthermore, the loss function ℓ is linear in f, i.e., $\ell((f + f')(x), y) = \ell(f(x), y) + \ell(f'(x), y)$ and $\ell(\alpha f(x), y) = \alpha \ell(f(x), z)$, then the lower bound on δ_n^2 is not required.

Proof of Lemma I.2. See Lemma 11 of Foster and Syrgkanis (2019) for a detailed proof. □

Remark I.3. *We remark that in the original Lemma 11 of Foster and Syrgkanis (2019), inequality (I.3) only holds for δ_n, and we extend it to any $t \geq \delta_n$ since according to Lemma 13.6 of Wainwright (2019) we know that $\mathcal{R}_n(\mathcal{F}; \delta)/\delta$ is a non-increasing function of δ on $(0, +\infty)$, which indicates that $t \geq \delta_n$ also solves the inequality.*