LIST OF NOTATION

In the sequel, we present a list of notations in the paper.

Notation	Explanation
$\mathcal{S}$, $\mathcal{A}$, $\mathcal{O}$	The state, action, and observation spaces, respectively.
A, H	The capacity of action space $\lvert\mathcal{A}\rvert$ and the length of an episode, respectively.
Φ	The embedding of trajectory defined in (3.1).
$\mathbb{P}_h(s_{h+1} \mid s_h, a_h)$	The transition probability from (s_h, a_h) to s_{h+1}.
$\mathbb{O}_h(o_h \mid s_h)$	The emission probability of observing o_h given s_h.
$[H]^+$	The set of steps $\{1-\ell, \ldots, H+k\}$ of the extended POMDP.
a_h^{h+k-1}, o_h^{h+k}	The sequences of actions and observations $\{a_h, \ldots, a_{h+k-1}\}$ and $\{o_h, \ldots, o_{h+k}\}$, respectively.
τ_h^{h+k}	The sequence of interactions $\{o_h, a_h, \ldots, o_{h+k-1}, a_{h+k-1}, o_{h+k}\}$ from the h-th step to the $(h+k)$-th step.
$\underline{\tau}_h^{h+k}$	The sequence of interactions $\{o_h, a_h, \ldots, o_{h+k-1}, a_{h+k-1}, o_{h+k}, a_{h+k}\}$ from the h-th step to the $(h+k)$-th step, including the $(h+k)$-th action.
$\mathbb{P}(\tau_h^{h+k})$, $\mathbb{P}(\tau_h^{h+k} \mid s_h)$	The conditional densities $\mathbb{P}(o_h^{h+k} \mid a_h^{h+k-1})$ and $\mathbb{P}(o_h^{h+k} \mid s_h, a_h^{h+k-1})$, respectively.
z_h, w_{h-1}	The shorthand for the sequences of interactions τ_h^{h+k} and $\underline{\tau}_{h-\ell}^{h-1}$, respectively, on page 7 of the paper.
ϕ^*, ψ^*	The unknown features of the low-rank POMDP in Assumption 3.1.
ϕ^θ, ψ^θ, $\mathbb{O}_h^\theta$	The parameterized features and emission kernel in Definition 3.2.
$\mathbb{P}^\theta$, $\mathbb{P}^{\theta,\pi}$	The probability densities corresponding to the transition dynamics defined by $\{\psi^\theta, \phi^\theta, \mathbb{O}_h^\theta\}$ and the policy π, repsectively.
$\mathbb{U}_h^\theta$, $\mathbb{U}_h^{\theta,\dagger}$	The forward emission operator and its pseudo-inverse defined in Definition 3.4 and Lemma 3.6, respectively.
M_h^θ, $M_h^{\theta,\dagger}$	The d-by-d matrix and its inverse defined in Assumption 3.5.
$\mathbb{B}_h^\theta$	The Bellman operator defined in Definition 3.7.

Notation	Explanation
$\mathfrak{E}(\cdot)$	The density estimation oracle defined in Assumption 4.1.
$w_{\mathfrak{E}}$, ν, γ	The parameters in Assumptions 4.1, 5.1, and 5.2, respectively.
$\mathbb{X}_h^{\theta,\pi}$, $\mathbb{Y}_h^{\theta,\pi}$	The density mappings defined in (3.6) and (3.7), respectively.
$b_1^\theta(\tau_1^H)$	The density of initial trajectory $\mathbb{P}^\theta(\tau_1^H)$.
$\mathcal{D}_h^t$, π^t, $\mathcal{C}^t$	The dataset, policy, and confidence set of parameters, respectively, in the t-th iteration of Algorithm 1.
$\widehat{\mathbb{X}}_h^t$, $\widehat{\mathbb{Y}}_h^t$, $\widehat{b}_1^t$	The estimated density mappings and initial trajectory density, respectively, in the t-th iteration of Algorithm 1.
L_h^t	The objective function defined in (4.1).

A Conclusion, Limitation, and Future Study

In this paper, we propose Represent to Control (RTC) as a unified framework for embedding and control in POMDPs. In particular, by exploiting the low-rank transition and the future sufficiency condition, we decompose the embedding learning into the learning of Bellman operators across multiple steps. By assembling the Bellman operators, we identify a sufficient embedding for the control in the POMDP. Moreover, we identify a confidence set of parameters fitting the Bellman operators, which further allows us to conduct exploration. Our analysis shows that RTC attains the $\mathcal{O}(1/\epsilon^2)$ sample complexity to attain an ϵ-suboptimal policy. To our best knowledge, we provide the first sample efficiency analysis for representation learning in POMDPs with infinite observation and state spaces.

A key to our analysis is the decomposition of the embedding across multiple steps, which hinges on the future sufficiency condition. It remains unclear if wearker conditions are possible for such a decomposition. In addition, our sample efficiency analysis hinges on the additional past sufficiency condition. It remain unclear whether such past sufficiency is necessary as our decomposition of embedding does not require such a condition. In our future study, we aim to tackle such challenges by recent advances in the tabular and low-rank POMDPs (Cai et al., 2022; Liu et al., 2022).

B Related Work on Latent State Space Models and MDPs

Our work is related to the previous study of latent state space models. Coates et al. (2008) recovers a class of latent state space models from observations by the expectation-maximization (EM) algorithm. In contrast, the spectral method (Rosencrantz et al., 2004; Hefny et al., 2015; Azizzadenesheli et al., 2016; Sun et al., 2016; Jin et al., 2020a) proposes to conduct filtering and prediction by solving a system of integral equations directly. In particular, previous works utilize predictive state representations (PSRs) (Hefny et al., 2015; Sun et al., 2016) as a sufficient representation of interaction history of fixed length and aim to conduct filtering on such predictive states. Our embedding strategy is inspired by the spectral algorithms with predictive states. In particular, our embedding of interaction history is also a predictive state. Nevertheless, unlike the previous analysis of predictive states (Hefny et al., 2015; Sun et al., 2016), we do not cast assumptions explicitly on the transition of predictive states (the filtration). Moreover, we focus on the sample efficiency of learning predictive states via iteratively exploring the environment, whereas previous works typically study PSRs with a fixed trajectory generator (Hefny et al., 2015; Sun et al., 2016).

To learn a sufficient embedding for control, we utilize the low-rank transition of POMDPs. Our idea is motivated by the previous analysis of low-rank MDPs (Cai et al., 2020; Jin et al., 2020b; Ayoub et al., 2020; Agarwal et al., 2020; Modi et al., 2021; Uehara et al., 2021). In particular, the state transition of a low-rank MDP aligns with that in our low-rank POMDP model. Nevertheless,

we remark that such states are observable in a low-rank MDP but are unobservable in POMDPs with the low-rank transition. Such unobservability makes solving a low-rank POMDP much more challenging than solving a low-rank MDP.

C Algorithm Description of RTC

In the sequel, we decribe the procedure of RTC. In summary, RTC iteratively (i) interacts with the environment to collect observations, (ii) fits the density mappings defined in (3.6) and (3.7), respectively, by observations, (iii) identifies a confidence set of parameters by fitting the Bellman equations according to (3.8), and (iv) conducts optimistic planning based on the fitted embeddings and the associated the confidence set.

To conduct RTC, we first initialize a sequence of datasets indexed by the step $h \in [H]$ and the action sequences $a_{h-\ell}^{h+k} \in \mathcal{A}^{k+\ell+1}$,

$$\mathcal{D}_h^0(a_{h-\ell}^{h+k}) = \varnothing.$$

Meanwhile, we initialize a policy $\pi^0 \in \Pi$, where Π is the class of all deterministic policies. In the sequel, we introduce the update procedure of RTC in the t-th iterate.

C.1 Data Collection

We first introduce the data collecting process of an agent with the policy π^{t-1} in the t-th iterate. For each of the step $h \in [H]$ and the action sequence $a_{h-\ell}^{h+k} \in \mathcal{A}^{k+\ell+1}$, the agent first execute the policy π^{t-1} till the $(h-\ell)$-th step, and collects a sequence of actions and observations as follows,

$${}^t a_{1-\ell}^{h-\ell-1} = \big\{{}^t a_{1-\ell}, \ldots, {}^t a_{h-\ell-1}\big\}, \quad {}^t o_{1-\ell}^{h-\ell} = \big\{{}^t o_{1-\ell}, \ldots, {}^t o_{h-\ell}\big\}.$$

Here we use the superscript t to denote the observations and actions acquired in the t-th iterate. Correspondingly, we denote by ${}^t\tau_{h-\ell}^{h-1} = \{{}^t a_{1-\ell}^{h-\ell-1}, {}^t o_{1-\ell}^{h-\ell}\}$ the interaction history from the $(h-\ell)$-th step to the $(h-1)$-th step. Then, the agent execute $a_{h-\ell}^{h+k}$ regardless of the observations and collect the following observation sequence,

$${}^t o_{h-\ell+1}^{h+k+1} = \big\{{}^t o_{h-\ell+1}, \ldots, {}^t o_{h+k+1}\big\}.$$

Finally, we store the observation sequence ${}^t o_{h-\ell}^{h+k+1}$ generated by fixing the action sequence $a_{h-\ell}^{h+k}$ into a dataset indexed by such action sequence, namely,

$$\mathcal{D}_h^t(a_{h-\ell}^{h+k}) \leftarrow \mathcal{D}_h^{t-1}(a_{h-\ell}^{h+k}) \cup \big\{{}^t o_{h-\ell}^{h+k+1}\big\}.$$

C.2 Density Estimation

Upon collecting the data, we follow the embedding learning procedure and fit the density mappings for the estimation of Bellman operator. In practice, various approaches are available in fitting the density by observations, including the maximum likelihood estimation (MLE), the generative adversial approaches, and the reproducing kernel Hilbert space (RKHS) density estimation. In what follows, we unify such density estimation approaches by a density estimation oracle.

Assumption C.1 (Density Estimation Oracle)**.** We assume that we have access to a density estimation oracle $\mathfrak{E}(\cdot)$. Moreover, for all $\delta > 0$ and dataset $\mathcal{D}$ drawn from the density p of size n following a martingale process, we assume that

$$\|\mathfrak{E}(\mathcal{D}) - p\|_1 \le C \cdot \sqrt{w_{\mathfrak{E}} \cdot \log(1/\delta)/n}$$

with probability at least $1-\delta$. Here $C > 0$ is an absolute constant and $w_{\mathfrak{E}}$ is a parameter that depends on the density estimation oracle $\mathfrak{E}(\cdot)$.

We highlight that such convergence property can be achieved by various density estimations. In particular, when the function approximation space $\mathcal{P}$ of $\mathfrak{E}(\cdot)$ is finite, Assumption 4.1 holds for the maximum likelihood estimation (MLE) and the generative adversial approach with $w_{\mathfrak{E}} = \log|\mathcal{P}|$ (Geer et al., 2000; Zhang, 2006; Agarwal et al., 2020). Meanwhile, $w_{\mathfrak{E}}$ scales with the entropy integral of $\mathcal{P}$ endowed with the Hellinger distance if $\mathcal{P}$ is infinite (Geer et al., 2000; Zhang, 2006). In addition, Assumption 4.1 holds for the RKHS density estimation (Gretton et al., 2005; Smola et al., 2007; Cai et al., 2022) with $w_{\mathfrak{E}} = \mathrm{poly}(d)$, where d is rank of the low-rank transition (Cai et al., 2022).

We now fit the density mappings based on the density estimation oracle. For each step $h \in [H]$ and action sequence $a_{h-\ell}^{h+k} \in \mathcal{A}^{k+\ell+1}$, we first fit the density of trajectory as follows,

$$\widehat{\mathbb{P}}_h^t(\cdot \,|\, a_{h-\ell}^{h+k}) = \mathfrak{E}\big(\mathcal{D}_h^t(a_{h-\ell}^{h+k})\big),$$

where the dataset $\mathcal{D}_h^t$ is updated based on the data collection procedure described in §C.1. Meanwhile, we define the following density mappings for the estimation of Bellman operators,

$$\big[\widehat{\mathbb{X}}_h^t(\underline{\tau}_{h-\ell}^{h-1})\big](\tau_h^{h+k}) = \widehat{\mathbb{P}}_h^t(\tau_{h-\ell}^{h+k}), \tag{C.1}$$

$$\big[\widehat{\mathbb{Y}}_h^t(\underline{\tau}_{h-\ell}^{h})\big](\tau_{h+1}^{h+k+1}) = \widehat{\mathbb{P}}_h^t(\tau_{h-\ell}^{h+k+1}). \tag{C.2}$$

Here recall that we define the trajectories $\underline{\tau}_{h-\ell}^h = \{a_{h-\ell}^h, o_{h-\ell}^h\}$ and $\tau_{h-\ell}^{h+k+1} = \{a_{h-\ell}^{h+k}, o_{h-\ell}^{h+k+1}\}$. Meanwhile, we write $\widehat{\mathbb{P}}_h^t(\tau_{h-\ell}^{h+k+1}) = \widehat{\mathbb{P}}^t(o_{h-\ell}^{h+k+1} \,|\, a_{h-\ell}^{h+k})$ for notational simplicity. We remark that the density mappings $\widehat{\mathbb{X}}_h^t$ and $\widehat{\mathbb{Y}}_h^t$ are estimations of the density mappings defined in (3.6) and (3.7), respectively, under the true parameter θ^*. We then estimate the Bellman operators by minimizing the following objective,

$$L_h^t(\theta) = \sup_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \int_{\mathcal{O}^{\ell+1}} \|\mathbb{B}_h^\theta(a_h, o_h)\widehat{\mathbb{X}}_h^t(\underline{\tau}_{h-\ell}^{h-1}) - \widehat{\mathbb{Y}}_h^t(\underline{\tau}_{h-\ell}^{h})\|_1 \mathrm{d}o_{h-\ell}^h. \tag{C.3}$$

We remark that the objective defined in (4.1) is motivated by the identity in (3.8). In what follows, we introduce an exploration procedure based on the objective defined in (4.1). In addition, we acquire the estimation of initial trajectory density $\widehat{b}_1^t(\tau_1^k) = \widehat{\mathbb{P}}_1^t(\tau_1^k)$ by marginalizing the dummy past trajectory $\tau_{1-\ell}^0$ of $\widehat{\mathbb{P}}_1^t$.

C.3 Optimistic Planning

We remark that the objective defined in (4.1) encapsulates the uncertainty in the estimation of the corresponding Bellman operator $\mathbb{B}_h^\theta(a_h, o_h)$. In particular, a smaller objective $L_h^t(\theta)$ yields a higher confidence that θ is close to the true parameter θ^*. Thus, we define the following confidence set of parameters,

$$\mathcal{C}^t = \Big\{\theta \in \Theta : \max\big\{\|b_1^\theta - \widehat{b}_1^t\|_1, L_h^t(\theta)\big\} \le \beta_t \cdot \sqrt{1/t}, \quad \forall h \in [H]\Big\}, \tag{C.4}$$

where β_t is the tuning parameter in the t-th iterate. Meanwhile, for each parameter $\theta \in \Theta$, we can estimate the embedding

$$\Phi^\theta(\tau_1^H) = \mathbb{P}^\theta(\tau_1^H)$$

based on the Bellman operators $\{\mathbb{B}_h^\theta\}_{h\in[H]}$ and Lemma 3.8. Such embedding further allows us to evaluate a policy as follows,

$$V^\pi(\theta) = \int_{\mathcal{O}^H} r(o_1^H) \cdot \mathbb{P}^\theta\big(o_1^H \,|\, (a^\pi)_1^H\big)\mathrm{d}o_1^H = \int_{\mathcal{O}^H} r(o_1^H) \cdot \Phi^\theta\big(o_1^H, (a^\pi)_1^H\big)\mathrm{d}o_1^H,$$

where we define $V^\pi(\theta)$ as the cumulative rewards of π in the POMDP induced by the parameter $\theta \in \Theta$. Meanwhile, we define $(a^\pi)_1^H = (a_1^\pi, \ldots, a_H^\pi)$, where the actions a_h^π are the action taken by the deterministic policy π in the h-th step given the observations.

To conduct optimistic planning, we seek for the policy that maximizes the return among all parameters $\theta \in \mathcal{C}^t$ and the corresponding features. The update of policy takes the following form,

$$\pi^t \leftarrow \operatorname*{argmax}_{\pi\in\Pi} \max_{\theta\in\mathcal{C}^t} V^\pi(\theta),$$

where we denote by Π the set of all deterministic policies. We summarize RTC in Algorithm 1.

D Proof of Preliminary Result

In the sequel, we present the proof of preliminary results in §3.

D.1 PROOF OF LEMMA 3.6

Proof. It holds for all time step $h \in [H]$, policy $\pi \in \Pi$, parameter $\theta \in \Theta$ that

$$\begin{aligned}\mathbb{P}_h^{\theta,\pi}(s_h) &= \int_{\mathcal{S}\times\mathcal{A}} \mathbb{P}_{h-1}^{\theta}(s_h \,|\, s_{h-1}, a_{h-1}) \cdot \mathbb{P}^{\theta,\pi}(s_{h-1}, a_{h-1}) \mathrm{d}s_{h-1}, a_{h-1} \\ &= \psi_{h-1}^{\theta}(s_h)^\top \int_{\mathcal{S}\times\mathcal{A}} \phi_{h-1}^{\theta}(s_{h-1}, a_{h-1}) \cdot \mathbb{P}^{\theta,\pi}(s_{h-1}, a_{h-1}) \mathrm{d}s_{h-1}, a_{h-1} \\ &= \psi_{h-1}^{\theta}(s_h)^\top W_{h-1}(\theta, \pi), \end{aligned} \tag{D.1}$$

where we define

$$W_{h-1}(\theta,\pi) = \int_{\mathcal{S}\times\mathcal{A}} \phi_{h-1}^{\theta}(s_{h-1}, a_{h-1}) \cdot \mathbb{P}^{\theta,\pi}(s_{h-1}, a_{h-1}) \mathrm{d}s_{h-1}, a_{h-1}.$$

Meanwhile, recall that we define the following linear operator in Lemma 3.6,

$$(\mathbb{U}_h^{\theta,\dagger} f)(s_h) = \int_{\mathcal{A}^k\times\mathcal{O}^{k+1}} \psi_{h-1}^{\theta}(s_h)^\top z_h^{\theta}(\tau_h^{h+k}) \cdot f(\tau_h^{h+k}) \mathrm{d}\tau_h^{h+k}, \quad \forall f \in L^1(\mathcal{A}^k \times \mathcal{O}^{k+1}),\ \forall s_h \in \mathcal{S},$$

where we define

$$z_h^{\theta}(\tau_h^{h+k}) = M_h^{\theta,\dagger}(\mathbb{U}_h^{\theta}\psi_{h-1}^{\theta})(\tau_h^{h+k}), \quad \forall \tau_h^{h+k} \in \mathcal{A}^k \times \mathcal{O}^{k+1},$$

It thus follows from (D.1) that

$$\begin{aligned}&\int_{\mathcal{A}^k\times\mathcal{O}^{k+1}} z_h^{\theta}(\tau_h^{h+k})(\mathbb{U}_h^{\theta}\mathbb{P}_h^{\theta,\pi})(\tau_h^{h+k}) \mathrm{d}\tau_h^{h+k} \\ &\quad = M_h^{\theta,\dagger} \int_{\mathcal{A}^k\times\mathcal{O}^{k+1}} (\mathbb{U}_h^{\theta}\psi_{h-1}^{\theta})(\tau_h^{h+k})(\mathbb{U}_h^{\theta}\psi_{h-1}^{\theta})(\tau_h^{h+k})^\top W_{h-1}(\theta,\pi) \mathrm{d}\tau_h^{h+k} \\ &\quad = M_h^{\theta,\dagger} M_h^{\theta} W_{h-1}(\theta,\pi) = W_{h-1}(\theta,\pi). \end{aligned} \tag{D.2}$$

Here recall that we define

$$M_h^{\theta} = \int_{\mathcal{A}^k\times\mathcal{O}^{k+1}} (\mathbb{U}_h^{\theta}\psi_{h-1}^{\theta})(\tau_h^{h+k})(\mathbb{U}_h^{\theta}\psi_{h-1}^{\theta})(\tau_h^{h+k})^\top \mathrm{d}\tau_h^{h+k} \in \mathbb{R}^{d\times d}$$

and $M_h^{\theta,\dagger}$ as the inverse of M_h^{θ} in Assumption 3.5. Thus, we have

$$\begin{aligned}\mathbb{U}_h^{\theta,\dagger}\mathbb{U}_h^{\theta}\big(\mathbb{P}_h^{\theta,\pi}(\cdot)\big) &= \psi_{h-1}^{\theta}(\cdot)^\top \int_{\mathcal{A}^k\times\mathcal{O}^{k+1}} z_h^{\theta}(\tau_h^{h+k})(\mathbb{U}_h^{\theta}\mathbb{P}_h^{\theta,\pi})(\tau_h^{h+k}) \mathrm{d}\tau_h^{h+k} \\ &= \psi_{h-1}^{\theta}(\cdot)^\top W_{h-1}(\theta,\pi) = \mathbb{P}_h^{\theta,\pi}(\cdot), \end{aligned} \tag{D.3}$$

which completes the proof of Lemma 3.6. □

D.2 PROOF OF EQUATION 3.8

Proof. By the definition of Bellman operators in Definition 3.7, we have

$$\big(\mathbb{B}_h^{\theta}(a_h, o_h)\mathbb{X}_h(\underline{\tau}_{h-\ell}^{h-1})\big)(\tau_{h+1}^{h+k+1}) = \int_{\mathcal{S}} \mathbb{P}^{\theta}(\tau_h^{h+k+1} \,|\, s_h) \cdot \big(\mathbb{U}_h^{\theta,\dagger}\mathbb{X}_h^{\theta}(\underline{\tau}_{h-\ell}^{h-1})\big)(s_h) \mathrm{d}s_h. \tag{D.4}$$

Meanwhile, by the definition of $\mathbb{X}_h^{\theta}$ and $\mathbb{U}_h^{\theta}$ in (3.6) and (3.2), respectively, we have

$$\begin{aligned}\big[\mathbb{X}_h^{\theta}(\underline{\tau}_{h-\ell}^{h-1})\big](\tau_h^{h+k}) = \mathbb{P}^{\theta}(\tau_{h-\ell}^{h+k}) &= \int_{\mathcal{S}} \mathbb{P}^{\theta}(o_{h-\ell}^{h-1}, s_h \,|\, a_{h-\ell}^{h-1}) \cdot \mathbb{P}^{\theta}(\tau_h^{h+k} \,|\, s_h) \mathrm{d}s_h \\ &= \big(\mathbb{U}_h^{\theta}\mathbb{P}^{\theta}(o_{h-\ell}^{h-1}, s_h = \cdot \,|\, a_{h-\ell}^{h-1})\big)(\tau_h^{h+k}). \end{aligned}$$

Thus, by Lemma 3.6, it holds that

$$\mathbb{U}_h^{\theta,\dagger}\mathbb{X}_h^{\theta}(\underline{\tau}_{h-\ell}^{h-1}) = \mathbb{U}_h^{\theta,\dagger}\mathbb{U}_h^{\theta}\mathbb{P}^{\theta}(o_{h-\ell}^{h-1}, s_h = \cdot \,|\, a_{h-\ell}^{h-1}) = \mathbb{P}^{\theta}(o_{h-\ell}^{h-1}, s_h = \cdot \,|\, a_{h-\ell}^{h-1}). \tag{D.5}$$

Plugging (D.5) into (D.4), we conclude that

$$\big(\mathbb{B}_h^{\theta}(a_h, o_h)\mathbb{X}_h(\underline{\tau}_{h-\ell}^{h-1})\big)(\tau_{h+1}^{h+k+1}) = \int_{\mathcal{S}} \mathbb{P}^{\theta}(\tau_h^{h+k+1} \,|\, s_h) \cdot \mathbb{P}^{\theta}(s_h, o_{h-\ell}^{h-1} \,|\, a_{h-\ell}^{h-1}) \mathrm{d}s_h = \mathbb{P}^{\theta}(\tau_{h-\ell}^{h+k+1}),$$

where the second equality follows from the fact that the past observations $o_{h-\ell}^{h-1}$ is independent of the forward observations o_{h+1}^{h+k+1} given the current state s_h. Thus, by the definition of $\mathbb{Y}_h^{\theta}$ in (3.6), we conclude the proof of equation 3.8. □

D.3 Proof of Lemma 3.8

Proof. We first define the following density function of initial trajectory,

$$b_1^\theta(\tau_1^{1+k}) = (\mathbb{U}_1^\theta \mu_1)(\tau_1^{1+k}) = \mathbb{P}^\theta(\tau_1^{1+k}) \in L^1(\mathcal{A}^k \times \mathcal{O}^{k+1}). \tag{D.6}$$

Thus, it holds from the definition of Bellman operators in Definition 3.7 that

$$\begin{aligned}\big[\mathbb{B}_1^\theta(a_1, o_1) b_1^\theta\big](\tau_2^{k+2}) &= \int_{\mathcal{S}} \mathbb{P}^\theta(\tau_1^{k+2} \,|\, s_1) \cdot (\mathbb{U}_1^{\theta,\dagger} b_1^\theta)(s_1) \mathrm{d}s_1 \\ &= \int_{\mathcal{S}} \mathbb{P}^\theta(\tau_1^{k+2} \,|\, s_1) \cdot \mu_1(s_1) \mathrm{d}s_1 = \mathbb{P}(\tau_1^{k+2}),\end{aligned} \tag{D.7}$$

where μ_1 is the initial state density of the POMDP. Here the second equality follows from the left invertibility of the forward emission operator $\mathbb{U}_1^\theta$ in Lemma 3.6 and the definition of b_1^θ in (D.6). Thus, by the recursive computation following (D.7), we obtain that

$$\big[\mathbb{B}_H^\theta(o_H, a_H) \ldots \mathbb{B}_1^\theta(o_1, a_1) b_1^\theta\big](\tau_{H+1}^{H+k+1}) = \mathbb{P}^\theta(\tau_1^{H+k+1}).$$

Finally, by marginalizing over the dummy future trajectories

$$\tau_{H+1}^{H+k+1} = \{a_{H+1}, o_{H+1}, \ldots, a_{H+k}, o_{H+k+1}\} \in \mathcal{A}^k \times \mathcal{O}^{k+1},$$

we conclude that

$$\begin{aligned}\mathbb{P}^\theta(\tau_1^H) &= \frac{1}{A^k} \cdot \int_{\mathcal{A}^k \times \mathcal{O}^{k+1}} \mathbb{P}^\theta(\tau_1^{H+k+1}) \mathrm{d}\tau_{H+1}^{H+K+1} \\ &= \frac{1}{A^k} \cdot \int_{\mathcal{A}^k \times \mathcal{O}^{k+1}} \big[\mathbb{B}_H^\theta(a_H, o_H) \ldots \mathbb{B}_1^\theta(a_1, o_1) b_1^\theta\big](\tau_H^{H+k}) \mathrm{d}o_h^{h+k}.\end{aligned}$$

Thus, we complete the proof of Lemma 3.8. □

E Proof of Main Result

In the sequel, we present the proof of the main result in §5.

E.1 Computing the Performance Difference

In the sequel, we present lemmas for the sample efficiency analysis of RTC. Our analysis is motivated by previous work (Jin et al., 2020a; Cai et al., 2022). We first define linear operators $\{\mathbb{T}_h^\theta, \widetilde{\mathbb{O}}_h^\theta\}_{h \in [H]}$ as follows,

$$\big(\mathbb{T}_h^\theta(a_h) f\big)(s_{h+1}) = \int_{\mathcal{S}} \mathbb{P}_h^\theta(s_{h+1} \,|\, s_h, a_h) \cdot f(s_h) \mathrm{d}s_h, \quad \forall f \in L^1(\mathcal{S}),\ a_h \in \mathcal{A}, \tag{E.1}$$

$$\big(\widetilde{\mathbb{O}}_h^\theta(o_h) f\big)(s_h) = \mathbb{O}_h^\theta(o_h \,|\, s_h) \cdot f(s_h), \quad \forall f \in L^1(\mathcal{S}),\ o_h \in \mathcal{O}. \tag{E.2}$$

It thus holds that

$$\mathbb{B}_h^\theta(a_h, o_h) = \mathbb{U}_{h+1}^\theta \mathbb{T}_h^\theta(a_h) \widetilde{\mathbb{O}}_h^\theta(o_h) \mathbb{U}_h^{\theta,\dagger}. \tag{E.3}$$

To see such a fact, note that we have for all $f \in L^1(\mathcal{A}^k \times \mathcal{O}^{k+1})$ that

$$\begin{aligned}&\big(\mathbb{U}_{h+1}^\theta \mathbb{T}_h^\theta(a_h) \widetilde{\mathbb{O}}_h(o_h) \mathbb{U}_h^{\theta,\dagger} f\big)(\tau_{h+1}^{h+k+1}) \\ &\quad= \bigg(\mathbb{U}_{h+1}^\theta \int_{\mathcal{S}} \mathbb{P}_h^\theta(\cdot \,|\, s_h, a_h) \cdot \mathbb{O}_h(o_h \,|\, s_h) (\mathbb{U}_h^{\theta,\dagger} f)(s_h) \mathrm{d}s_h\bigg)(\tau_{h+1}^{h+k+1}) \\ &\quad= \int_{\mathcal{S}^2} \mathbb{P}^\theta(\tau_{h+1}^{h+k+1} \,|\, s_{h+1}) \cdot \mathbb{P}_h^\theta(s_{h+1} \,|\, s_h, a_h) \cdot \mathbb{O}_h(o_h \,|\, s_h) (\mathbb{U}_h^{\theta,\dagger} f)(s_h) \mathrm{d}s_h \mathrm{d}s_{h+1} \\ &\quad= \int_{\mathcal{S}^2} \mathbb{P}^\theta(o_h^{h+k+1}, s_{h+1} \,|\, s_h, a_h^{h+k}) (\mathbb{U}_h^{\theta,\dagger} f)(s_h) \mathrm{d}s_h \mathrm{d}s_{h+1},\end{aligned} \tag{E.4}$$

where the first and second equalities follow from the definitions of $\mathbb{T}_h^\theta$, $\widetilde{\mathbb{O}}_h$, and $\mathbb{U}_{h+1}^\theta$ in (E.1), (E.2), and (3.2), respectively. Meanwhile, the third equality follows from the fact that the POMDP is Markov with respect to the state and action pairs (s_{h+1}, a_{h+1}). Marginalizing over the state s_{h+1}

on the right-hand side of (E.4), we obtain for all $(\tau_{h+1}^{h+k+1}) \in \mathcal{A}^k \times \mathcal{O}^{k+1}$ that

$$\begin{aligned}\big(\mathbb{U}_{h+1}^{\theta}\mathbb{T}_h^{\theta}(a_h)\widetilde{\mathbb{O}}_h^{\theta}(o_h)\mathbb{U}_h^{\theta,\dagger}f\big)(\tau_{h+1}^{h+k+1}) &= \int_{\mathcal{S}} \mathbb{P}^{\theta}(\tau_h^{h+k+1} \,|\, s_h)\cdot(\mathbb{U}_h^{\theta,\dagger}f)(s_h)\mathrm{d}s_h \\ &= \big(\mathbb{B}_h^{\theta}(a_h,o_h)f\big)(\tau_{h+1}^{h+k+1}),\end{aligned}$$

where the second equality follows from the definition of Bellman operator $\mathbb{B}_h^{\theta}$ in Definition 3.7. Thus, we complete the proof of (E.3).

Lemma E.1 (Performance Difference)**.** It holds for all policy $\pi \in \Pi$ and parameters $\theta, \theta' \in \Theta$ that

$$|V^{\pi}(\theta) - V^{\pi}(\theta')| \le H\cdot\nu\cdot\sum_{h=1}^{H-1}\sum_{a_{h-\ell}^h\in\mathcal{A}^{\ell+1}}\int_{\mathcal{O}}\sum_{q_{h-1}\in[d]}\|u_{h,q_{h-1}}\|_1\mathrm{d}o_h + H\cdot\nu\cdot\|b_1^{\theta}-b_1^{\theta'}\|_1,$$

where we define

$$u_{h,q_{h-1}} = \big(\mathbb{B}_h^{\theta}(a_h,o_h)-\mathbb{B}_h^{\theta'}(a_h,o_h)\big)\mathbb{U}_h^{\theta'}\mathbb{P}_h^{\theta'}(s_h=\cdot\,|\,q_{h-1})\cdot\mathbb{P}^{\pi}(q_{h-1}\,|\,a_{h-\ell}^{h-1}).$$

Proof. See §F.1 for a detailed proof. □

E.2 Confidence Set Analysis

We first present the following norm bound on Bellman operators.

Lemma E.2 (Norm Bound of Bellman Operator)**.** Under Assumptions 3.1, 3.5, and 5.1, it holds for all $h \in [H]$, $\theta \in \Theta$, and $(a_h, o_h) \in \mathcal{A}\times\mathcal{O}$ that $\|\mathbb{B}_h^{\theta}(a_h,o_h)\|_{1\mapsto 1} \le \nu\cdot A^k$.

Proof. It holds for all $f \in L^1(\mathcal{A}^k\times\mathcal{O}^{k+1})$ that

$$\begin{aligned}\|\mathbb{B}_h^{\theta}(a_h,o_h)f\|_1 &\le \int_{\mathcal{A}^k\times\mathcal{O}^{k+1}}\int_{\mathcal{S}}\mathbb{P}^{\theta}(\tau_h^{h+k+1}\,|\,s_h)\cdot|\mathbb{U}_h^{\theta,\dagger}f(s_h)|\mathrm{d}s_h\mathrm{d}\tau_{h+1}^{h+k+1} \\ &\le A^k\cdot\int_{\mathcal{S}}|\mathbb{U}_h^{\theta,\dagger}f(s_h)|\mathrm{d}s_h. \qquad \text{(E.5)}\end{aligned}$$

Meanwhile, by the definition of $\mathbb{U}_h^{\theta,\dagger}$ in (3.3) and Assumption 5.1, it holds that

$$\int_{\mathcal{S}}|\mathbb{U}_h^{\theta,\dagger}f(s_h)|\mathrm{d}s_h = \|\mathbb{U}_h^{\theta,\dagger}f(s_h)\|_1 \le \nu\cdot\|f\|_1. \qquad \text{(E.6)}$$

Combining (E.5) and (E.6), we conclude that

$$\|\mathbb{B}_h^{\theta}(a_h,o_h)f\|_1 \le \nu\cdot A^k\cdot\|f\|_1,$$

which completes the proof of Lemma E.2. □

In what follows, we recall the definition of the reverse emission operator.

Definition E.3 (Reverse Emission)**.** We define for all $h \in [H]$ the following linear operator $\mathbb{F}_h^{\theta,\pi}$: $\mathbb{R}^d \mapsto L^1(\mathcal{O}^{\ell}\times\mathcal{A}^{\ell})$ for all $h \in [H]$, $\pi \in \Pi$, and $\theta \in \Theta$,

$$(\mathbb{F}_h^{\theta,\pi}v)(\underline{\tau}_{h-\ell}^{h-1}) = \sum_{q_{h-1}\in[d]}[v]_{q_{h-1}}\cdot\mathbb{P}^{\theta,\pi}(o_{h-\ell}^{h-1}\,|\,q_{h-1},a_{h-\ell}^{h-1}),\quad \forall v\in\mathbb{R}^d,$$

where $(\underline{\tau}_{h-\ell}^{h-1}) \in \mathcal{A}^{\ell}\times\mathcal{O}^{\ell}$.

In addition, we define the following visitation measure of mix policy in the t-th iteration,

$$\mathbb{P}^t = \frac{1}{t}\cdot\sum_{\omega=0}^{t-1}\mathbb{P}^{\pi^{\omega}},$$

where $\{\pi^{\omega}\}_{\omega\in[t]}$ is the set of policy returned by Algorithm 1. We remark that the data collected by our data collection process in Algorithm 1 follow the trajectory density induced by $\mathbb{P}^t$ in the t-th iterate. Hence, the estimated density $\widehat{\mathbb{P}}^t$ returned by our density estimator $\mathfrak{E}(\mathcal{D}^t)$ in the t-th iterate aligns closely to $\mathbb{P}^t$. Meanwhile, recall that we define the following estimators in (C.1) and (C.2), respectively,

$$\big[\widehat{\mathbb{X}}_h^t(\underline{\tau}_{h-\ell}^{h-1})\big](\tau_h^{h+k}) = \widehat{\mathbb{P}}_h^t(\tau_{h-\ell}^{h+k}),\qquad \big[\widehat{\mathbb{Y}}_h^t(\underline{\tau}_{h-\ell}^{h})\big](\tau_{h+1}^{h+k+1}) = \widehat{\mathbb{P}}_h^t(\tau_{h-\ell}^{h+k+1}).$$

Recall that we define the confidence set as follows,

$$\mathcal{C}^t = \Big\{\theta \in \Theta : \int_{\mathcal{O}^{\ell+1}} \|\mathbb{B}_h^\theta(a_h, o_h)\widehat{\mathbb{X}}_h^t(a_{h-\ell}^{h-1}) - \widehat{\mathbb{Y}}_h^t(\underline{\tau}_{h-\ell}^h)\|_1 \mathrm{d}o_{h-\ell}^h \le \beta_t, \quad \forall a_{h-\ell}^h \in \mathcal{A}^{\ell+1}\Big\},$$

where we select

$$\beta_t = (\nu+1)\cdot A^{2k}\cdot\sqrt{w_{\mathfrak{E}}\cdot(k+\ell)\cdot\log(H\cdot A\cdot T)/t}.$$

In the sequel, we denote by θ^t the parameter selected in optimistic planning. The following lemma guarantees that the true parameter θ^* is included by our confidence set $\mathcal{C}^t$ with high probability. Moreover, we show that initial density and the Bellman operators $\{\mathbb{B}^{\theta^t}\}$ corresponding to the parameter θ^t aligns closely to that corresponding to the true parameter θ^*.

Lemma E.4 (Good Event Probability)**.** Under Assumptions 3.1, 3.5, and 5.2, it holds with probability at least $1-\delta$ that $\theta^* \in \mathcal{C}^t$. In addition, it holds for all $h \in [H]$ and $t \in [T]$ with probability at least $1-\delta$ that

$$\|b_1^{\theta^t} - b_1^{\theta^*}\|_1 = \mathcal{O}(\nu\cdot A^{2k}\cdot\sqrt{w_{\mathfrak{E}}\cdot(k+\ell)\cdot\log(H\cdot A\cdot T)/t}), \tag{E.7}$$

$$\sum_{a_{h-\ell}^h\in\mathcal{A}^{\ell+1}}\int_{\mathcal{O}}\sum_{q_{h-1}\in[d]}\big\|\big(\mathbb{B}_h^{\theta^t}(a_h,o_h)-\mathbb{B}_h^{\theta^*}(a_h,o_h)\big)\mathbb{U}_h^{\theta^*}\mathbb{P}_h^{\theta^*}(s_h=\cdot\,|\,q_{h-1})\big\|_1$$
$$\cdot\,\mathbb{P}^{\theta^*,t}(q_{h-1}\,|\,a_{h-\ell}^{h-1})\mathrm{d}o_h = \mathcal{O}\big(\gamma\cdot\nu\cdot A^{2k+\ell}\cdot\sqrt{w_{\mathfrak{E}}\cdot(k+\ell)\cdot\log(H\cdot A\cdot T)/t}\big), \tag{E.8}$$

where we define

$$\mathbb{P}^{\theta^*,t}(q_{h-1}\,|\,a_{h-\ell}^{h-1}) = \frac{1}{t}\cdot\sum_{\omega=0}^{t-1}\mathbb{P}^{\theta^*,\pi^\omega}(q_{h-1}\,|\,a_{h-\ell}^{h-1}).$$

Proof. See §F.2 for a detailed proof. □

E.3 Proof of Theorem 5.3

We are now ready to present the sample complexity analysis of Algorithm 1.

Proof. It holds that

$$V^*(\theta^*) - V^{\overline{\pi}^T}(\theta^*) = \frac{1}{T}\cdot\sum_{t=1}^T V^*(\theta^*) - V^{\pi^t}(\theta^*). \tag{E.9}$$

It suffices to upper bound the performance difference

$$V^*(\theta^*) - V^{\pi^t}(\theta^*)$$

for all $t \in [T]$. By Lemma E.4, it holds with probability at least $1-\delta$ that $\theta^* \in \mathcal{C}^t$ for all $t \in [T]$. Thus, by the update of π^t in Algorithm 1, it holds with probability at least $1-\delta$ that

$$V^*(\theta^*) - V^{\pi^t}(\theta^*) \le V^{\pi^t}(\theta^t) - V^{\pi^t}(\theta^*). \tag{E.10}$$

It now suffices to upper bound the performance difference on the right-hand side of (E.10). By Lemma E.1, it holds that

$$|V^{\pi^t}(\theta^t) - V^{\pi^t}(\theta^*)| \le \underbrace{H\cdot\nu\cdot\sum_{h=1}^{H-1}\sum_{a_{h-\ell}^h\in\mathcal{A}^{\ell+1}}\int_{\mathcal{O}}\sum_{q_{h-1}\in[d]}\|u_{h,q_{h-1}}^t\|_1\cdot\mathbb{P}^{\theta^*,\pi^t}(q_{h-1}\,|\,a_{h-\ell}^{h-1})\mathrm{d}o_h}_{\text{(i)}}$$
$$+\underbrace{H\cdot\nu\cdot\|b_1^{\theta^t}-b_1^{\theta^*}\|_1}_{\text{(ii)}}, \tag{E.11}$$

where we write

$$u_{h,q_{h-1}}^t = \big(\mathbb{B}_h^{\theta^t}(a_h,o_h)-\mathbb{B}_h^{\theta^*}(a_h,o_h)\big)\mathbb{U}_h^{\theta^*}\mathbb{P}_h^{\theta^*}(s_h=\cdot\,|\,q_{h-1}) \tag{E.12}$$

for notational simplicity. We remark that the summation in term (i) is different from that in (E.8) of Lemma E.4. In particular, by Lemma E.4, it holds for all $h \in [H]$ and $t \in [T]$ that

$$\sum_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \int_{\mathcal{O}} \sum_{q_{h-1} \in [d]} \|w_{h,q_{h-1}}^t\|_1 \cdot \mathbb{P}^{\theta^*,t}(q_{h-1} \,|\, a_{h-\ell}^{h-1}) \mathrm{d}o_h$$
$$= \mathcal{O}(\gamma \cdot \nu \cdot d \cdot A^{2k+\ell} \cdot \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t}) \tag{E.13}$$

with probability at least $1-\delta$, where we define

$$\mathbb{P}^{\theta^*,t}(q_{h-1} \,|\, a_{h-\ell}^{h-1}) = \frac{1}{t} \cdot \sum_{\omega=0}^{t-1} \mathbb{P}^{\theta^*,\pi^\omega}(q_{h-1} \,|\, a_{h-\ell}^{h-1}).$$

The only difference between the left-hand side of (E.13) and the term (i) in (E.11) is the conditional density of the bottleneck factor q_{h-1}, which follows the visitation of π^t, namely, $\mathbb{P}^{\theta^*,\pi^t}$, in (E.11) but the mixture of visitation $\mathbb{P}^{\theta^*,t}$ in (E.13). To upper bound (E.13) by (E.11), we utilize the calculation trick proposed by Jin et al. (2020a). In particular, we utilize the following lemma.

Lemma E.5 (Lemma 16 of Jin et al. (2020a))**.** Let $0 \le z_t \le C_z$ and $0 \le w_t \le C_w$ for all $t \in [T]$. We define $S_t = (1/t) \cdot \sum_{i=1}^t w_i$ and $S_0 = 0$. Given

$$z_t \cdot S_{t-1} \le C_z \cdot C_w \cdot C \cdot \sqrt{1/t}$$

for all $t \in [T]$, it holds that

$$\frac{1}{T} \cdot \sum_{t=1}^T z_t \cdot w_t \le 2C_z \cdot C_w \cdot (C+1) \cdot \sqrt{1/K} \cdot \log T.$$

Here $C > 0$ is an absolute constant.

Proof. See Jin et al. (2020a) for a detailed proof. □

It thus follows from (E.13) and Lemma E.5 that

$$\text{(i)} = \mathcal{O}\big(\gamma \cdot \nu^2 \cdot d \cdot H \cdot A^{2k+\ell} \cdot \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log T \cdot \log(H \cdot A \cdot T)/t}\big) \tag{E.14}$$

with probability at least $1-\delta$. Meanwhile, by (E.7) of Lemma E.4, it holds that

$$\text{(ii)} = H \cdot \nu \cdot \|b_1^{\theta^t} - b_1^{\theta^*}\|_1 = \mathcal{O}\big(H \cdot \nu^2 \cdot A^{2k} \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t}\big) \tag{E.15}$$

with probability at least $1-\delta$. Finally, by plugging (E.14) and (E.15) into (E.13), it holds for all $t \in [T]$ that

$$|V^{\pi^t}(\theta^t) - V^{\pi^t}(\theta^*)|$$
$$= \mathcal{O}\big(\gamma \cdot \nu^2 \cdot d \cdot H \cdot A^{2k+\ell} \cdot \log T \cdot \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t}\big) \tag{E.16}$$

with probability at least $1-\delta$. Combining (E.10) and (E.16), it holds with probability at least $1-\delta$ that

$$V^*(\theta^*) - V^{\overline{\pi}^T}(\theta^*)$$
$$= \mathcal{O}\big(\gamma \cdot \nu^2 \cdot d \cdot H \cdot A^{2k+\ell} \cdot \log T \cdot \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/T}\big).$$

Thus, by setting

$$T = \mathcal{O}\big(\gamma^2 \cdot \nu^4 \cdot d^2 \cdot H^2 \cdot A^{2(2k+\ell)} \cdot (k+\ell) \cdot \log(H \cdot A/\epsilon)/\epsilon^2\big),$$

it holds with probability at least $1-\delta$ that $V^*(\theta^*) - V^{\overline{\pi}^T}(\theta^*) \le \epsilon$. Thus, we complete the proof of Theorem 5.3.

□

F Proof of Auxiliary Result

In the sequel, we present the proof of the auxiliary results in §E.

F.1 PROOF OF LEMMA E.1

Proof. By Lemma G.7, it holds for all policy $\pi \in \Pi$ and parameter $\theta \in \Theta$ that

$$V^{\pi}(\theta) = \int_{\mathcal{O}^H} r(o_1^{H-1}) \cdot \mathbb{P}^{\pi,\theta}(o_1^{H-1}) \mathrm{d}o_1^{H-1}$$
$$= \int_{\mathcal{O}^H} r(o_1^{H-1}) \cdot \mathbb{P}^{\theta}\big(o_1^{H-1} \,\big|\, (a^{\pi})_1^H\big) \mathrm{d}o_1^{H-1},$$

where $(a^{\pi})_1^H = (a_1^{\pi}, \ldots, a_H^{\pi})$ and the actions $a_h^{\pi} = \pi(o_1^h, (a^{\pi})_1^{h-1})$ are taken by the policy π for all $h \in [H]$. Following from the fact that $0 \le r(o_1^{H-1}) \le H$ for all observation array $o_1^{H-1} \in \mathcal{O}^H$, we obtain for all policy $\pi \in \Pi$ and parameters $\theta, \theta' \in \Theta$ that

$$|V^{\pi}(\theta) - V^{\pi}(\theta')| \le H \cdot \int_{\mathcal{O}^H} \big|\mathbb{P}^{\theta}\big(o_1^{H-1} \,\big|\, (a^{\pi})_1^H\big) - \mathbb{P}^{\theta'}\big(o_1^{H-1} \,\big|\, (a^{\pi})_1^H\big)\big| \mathrm{d}o_1^{H-1}, \tag{F.1}$$

where the actions a_h^{π} are taken by the policy π for all $h \in [H]$. In the sequel, we utilize a slight modification of Lemma 3.8. In particular, following the same calculation as the proof of Lemma 3.8 in §D.3, we have

$$\mathbb{P}^{\theta}(\tau_1^{H+k}) = \big[\mathbb{B}_{H-1}^{\theta}(a_{H-1}, o_{H-1}) \ldots \mathbb{B}_1^{\theta}(a_1, o_1) b_1^{\theta}\big](\tau_H^{H+k}).$$

Thus, by marginalizing over the dummy future observations o_{H+1}^{H+k} and fixing the final observation o_H, we obtain for all dummy future actions a_H^{H+k-1} that

$$\mathbb{P}^{\theta}(\tau_1^H) = \int_{\mathcal{O}^k} \mathbb{1}_{o_H, a_H^{H+k-1}}(\tau_1^{H+k}) \cdot \mathbb{P}^{\theta}(\tau_1^{H+k}) \mathrm{d}o_{H+1}^{H+K}, \tag{F.2}$$

where we define $\mathbb{1}_{o_H, a_H^{H+k-1}}$ the indicator that takes value one at the final observation o_H and the fixed dummy future actions a_H^{H+k-1}. By plugging (F.2) into (F.1), we have

$$|V^{\pi}(\theta) - V^{\pi}(\theta')| \le H \cdot \int_{\mathcal{O}^{H+k}} |f^{\theta} - f^{\theta'}|\big(o_h^{h+k}, (a^{\pi})_H^{H+k-1}\big) \mathrm{d}o_1^{H+k}, \tag{F.3}$$

where $(a^{\pi})_H^{H+k-1} = (a_H^{\pi}, \ldots, a_{H+k-1}^{\pi})$ and the actions a_h^{π} are taken by the policy π. Here we define

$$f^{\theta} = \mathbb{B}_{H-1}^{\theta}(a_{H-1}^{\pi}, o_{H-1}) \ldots \mathbb{B}_1^{\theta}(a_1^{\pi}, o_1) b_1^{\theta},$$

where b_1^{θ} is the initial trajectory distribution for the first k steps defined in Lemma 3.8. Meanwhile, by the linearity of Bellman operators, we have

$$f^{\theta} - f^{\theta'} = \sum_{h=0}^{H-1} \mathbb{B}_{H-1}^{\theta}(a_{H-1}^{\pi}, o_{H-1}) \ldots \mathbb{B}_{h+1}^{\theta}(a_{h+1}^{\pi}, o_{h+1}) v_h, \tag{F.4}$$

where we define $v_0 = b_1^{\theta} - b_1^{\theta'}$ and

$$v_h = \big(\mathbb{B}_h^{\theta}(a_h^{\pi}, o_h) - \mathbb{B}_h^{\theta'}(a_h^{\pi}, o_h)\big) \mathbb{B}_{h-1}^{\theta'}(a_{h-1}^{\pi}, o_{h-1}) \ldots \mathbb{B}_1^{\theta'}(a_1^{\pi}, o_1) b_1^{\theta'}, \quad h \in [H-1]. \tag{F.5}$$

By combining (F.3) and (F.4), we have

$$|V^{\pi}(\theta) - V^{\pi}(\theta')| \le H \cdot \sum_{h=1}^{H-1} \int_{\mathcal{O}^{H+k}} |\mathbb{B}_{H-1}^{\theta}(a_{H-1}^{\pi}, o_{H-1}) \ldots \mathbb{B}_{h+1}^{\theta}(a_{h+1}^{\pi}, o_{h+1}) v_h| \mathrm{d}o_1^{H+k}$$
$$+ \int_{\mathcal{O}^{H+k}} |\mathbb{B}_{H-1}^{\theta}(a_{H-1}^{\pi}, o_{H-1}) \ldots \mathbb{B}_1^{\theta}(a_1^{\pi}, o_1)(b_1^{\theta} - b_1^{\theta'})| \mathrm{d}o_1^{H+k}. \tag{F.6}$$

The following lemma upper bounds the right-hand side of (F.6).

Lemma F.1. Under Assumption 3.5, it holds for all $h \in [H]$, $\pi \in \Pi$, and $v_h \in L^1(\mathcal{A}^k \times \mathcal{O}^{k+1})$ that

$$\int_{\mathcal{O}^{H+k-h}} |\mathbb{B}_{H-1}^{\theta}(o_{H-1}, a_{H-1}) \ldots \mathbb{B}_{h+1}^{\theta}(a_{h+1}, o_{h+1}) v_h| \mathrm{d}o_{h+1}^{H+k} \le \nu \cdot \|v_h\|_1.$$

Proof. See §F.3 for a detailed proof. □

By Lemma F.1, it follows from (F.6) that

$$|V^{\pi}(\theta) - V^{\pi}(\theta')| \le H \cdot \nu \cdot \sum_{h=0}^{H-1} \int_{\mathcal{O}^h} \|v_h\|_1 \mathrm{d}o_1^h + H \cdot \nu \cdot \|b_1^{\theta} - b_1^{\theta'}\|_1, \tag{F.7}$$

where we define $v_0 = b_1^{\theta} - b_1^{\theta'}$ and

$$v_h = \big(\mathbb{B}_h^{\theta}(a_h^{\pi}, o_h) - \mathbb{B}_h^{\theta'}(a_h^{\pi}, o_h)\big)\mathbb{B}_{h-1}^{\theta'}(a_{h-1}^{\pi}, o_{h-1}) \ldots \mathbb{B}_1^{\theta'}(a_1^{\pi}, o_1) b_1^{\theta'}, \quad h \in [H-1].$$

Meanwhile, the following lemma upper bound the L^1-norm of v_h for $h = 2, \ldots, H$.

Lemma F.2. It holds for all $\pi \in \Pi$ and $h \in [H-1]$ that

$$\int_{\mathcal{O}^h} \|v_h\|_1 \mathrm{d}o_1^h \le \sum_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \int_{\mathcal{O}} \sum_{q_{h-1} \in [d]} u_h \mathrm{d}o_h,$$

where we define

$$v_h = \big(\mathbb{B}_h^{\theta}(a_h^{\pi}, o_h) - \mathbb{B}_h^{\theta'}(a_h^{\pi}, o_h)\big)\mathbb{B}_{h-1}^{\theta'}(a_{h-1}^{\pi}, o_{h-1}) \ldots \mathbb{B}_1^{\theta'}(a_1^{\pi}, o_1) b_1^{\theta'},$$
$$u_h = \big\|\big(\mathbb{B}_h^{\theta}(a_h, o_h) - \mathbb{B}_h^{\theta'}(a_h, o_h)\big)\mathbb{U}_h^{\theta'}\mathbb{P}_h^{\theta'}(s_h = \cdot \,|\, q_{h-1}) \cdot \mathbb{P}^{\pi}(q_{h-1} \,|\, a_{h-\ell}^{h-1})\big\|_1,$$

for all $h \in [H-1]$.

Proof. See §F.4 for a detailed proof. □

Combining (F.7) and Lemma F.2, we conclude that

$$|V^{\pi}(\theta) - V^{\pi}(\theta')| \le H \cdot \nu \cdot \sum_{h=1}^{H-1} \sum_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \sum_{q_{h-1} \in [d]} \int_{\mathcal{O}} \|u_{h,q_{h-1}}\|_1 \mathrm{d}o_h + H \cdot \nu \cdot \|b_1^{\theta} - b_1^{\theta'}\|_1,$$

where we define

$$u_{h,q_{h-1}} = \big(\mathbb{B}_h^{\theta}(a_h, o_h) - \mathbb{B}_h^{\theta'}(a_h, o_h)\big)\mathbb{U}_h^{\theta'}\mathbb{P}_h^{\theta'}(s_h = \cdot \,|\, q_{h-1}) \cdot \mathbb{P}^{\theta',\pi}(q_{h-1} \,|\, a_{h-\ell}^{h-1}).$$

Thus, we complete the proof of Lemma E.1. □

F.2 Proof of Lemma E.4

Proof. We first show that $\theta^* \in \mathcal{C}^t$ with probability at least $1 - \delta$. By Assumption 4.1, it holds for all $t \in [T]$ that

$$\|\widehat{b}_1^t - b_1^{\theta^*}\|_1 \le \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t} \tag{F.8}$$

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

$$\begin{aligned}
&\int_{\mathcal{O}^{\ell+1}} \|\mathbb{B}_h^{\theta^*}(a_h, o_h)\widehat{\mathbb{X}}_h^t(\underline{\tau}_{h-\ell}^{h-1}) - \widehat{\mathbb{Y}}_h^t(\underline{\tau}_{h-\ell}^h)\|_1 \mathrm{d}o_{h-\ell}^h \\
&\quad \le \int_{\mathcal{O}^{\ell+1}} \|\mathbb{B}_h^{\theta^*}(a_h, o_h)\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}_{h-\ell}^{h-1}) - \mathbb{Y}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}_{h-\ell}^h)\|_1 \mathrm{d}o_{h-\ell}^h \\
&\qquad + \int_{\mathcal{O}} \big\|\mathbb{B}_h^{\theta^*}(a_h, o_h)\big(\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}_{h-\ell}^{h-1}) - \widehat{\mathbb{X}}_h^t(\underline{\tau}_{h-\ell}^{h-1})\big)\big\|_1 \mathrm{d}o_h \\
&\qquad + \int_{\mathcal{O}^{\ell+1}} \big\|\big(\mathbb{Y}_h^{\theta^*,\overline{\pi}^t} - \widehat{\mathbb{Y}}_h^t\big)(\underline{\tau}_{h-\ell}^h)\big\|_1 \mathrm{d}o_{h-\ell}^h.
\end{aligned} \tag{F.9}$$

We now upper bound the right-hand side of (F.9). According to the identity in (3.8), we have

$$\mathbb{B}_h^{\theta^*}(a_h, o_h)\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(a_{h-\ell}^{h-1}) - \mathbb{Y}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}_{h-\ell}^h) = 0 \tag{F.10}$$

for all $h \in [H]$ and $(a_{h-\ell}^h, o_{h-\ell}^h) \in \mathcal{A}^{\ell+1} \times \mathcal{O}^{\ell+1}$. Meanwhile, by Assumption 4.1 and the update of density estimators $\widehat{\mathbb{Y}}_h^t$ in (C.2), it holds for all $h \in [H]$, $t \in [T]$, and $\underline{\tau}_{h-\ell}^h \in \mathcal{A}^\ell \times \mathcal{O}^\ell$ that

$$\begin{aligned}\int_{\mathcal{O}^{\ell+1}} \big\|\big(\mathbb{Y}_h^{\theta^*,\overline{\pi}^t} - \widehat{\mathbb{Y}}_h^t\big)(\underline{\tau}_{h-\ell}^h)\big\|_1 \mathrm{d}o_{h-\ell}^h &= \sum_{a_{h+1}^{h+k} \in \mathcal{A}^k} \int_{\mathcal{O}^{k+\ell+1}} |(\widehat{\mathbb{P}}_h^t - \mathbb{P}_h^{\theta^*,\overline{\pi}^t})(\tau_{h-\ell}^{h+k+1})| \mathrm{d}o_{h-\ell}^{h+k+1}. \\ &= \sum_{a_{h+1}^{h+k} \in \mathcal{A}^k} \|(\widehat{\mathbb{P}}_h^t - \mathbb{P}_h^{\theta^*,\overline{\pi}^t})(\cdot \,|\, a_{h-\ell}^{h+k})\|_1 \\ &\le A^k \cdot \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t} \end{aligned} \tag{F.11}$$

with probability at least $1-\delta$. Similarly, by Assumption 4.1, Lemma E.2, and the update of density estimators $\widehat{\mathbb{X}}_h^t$ in (C.1), we further obtain for all $h \in [H]$ that

$$\begin{aligned}&\int_{\mathcal{O}} \big\|\mathbb{B}_h^{\theta^*}(a_h, o_h)\big(\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}_{h-\ell}^{h-1}) - \widehat{\mathbb{X}}_h^t(\underline{\tau}_{h-\ell}^{h-1})\big)\big\|_1 \mathrm{d}o_h \\ &\qquad \le \nu \cdot A^{2k} \cdot \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t} \end{aligned} \tag{F.12}$$

with probability at least $1-\delta$. Plugging (F.10), (F.11), and (F.12) into (F.9), we obtain for all $h \in [H]$ that

$$\int_{\mathcal{O}^{\ell+1}} \|\mathbb{B}_h^{\theta^*}(a_h, o_h)\widehat{\mathbb{X}}_h^t(\underline{\tau}_{h-\ell}^{h-1}) - \widehat{\mathbb{Y}}_h^t(\underline{\tau}_{h-\ell}^h)\|_1 \mathrm{d}o_{h-\ell}^h \le \beta_t \cdot \sqrt{1/t} \tag{F.13}$$

with probability at least $1-\delta$. Thus, combining (F.9) and (F.13), it holds that $\theta^* \in \mathcal{C}^t$ with probability at least $1-\delta$. In what follows, we prove (E.7) and (E.8), respectively.

Part I: Proof of Upper Bound in (E.7)**.** By the definition of confidence set $\mathcal{C}^t$, it holds for all $t \in [T]$, $(a_h^{h+k-1}, a_{h-\ell}^h) \in \mathcal{A}^{k+\ell}$, and $h \in [H]$ that

$$\|b_1^{\theta^t} - \widehat{b}_1^t\|_1 \le (1+\nu) \cdot A^{2k} \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t}$$

with probability at least $1-\delta$. Thus, by (F.8) and triangle inequality, it holds for all $t \in [T]$, $(a_h^{h+k-1}, a_{h-\ell}^h) \in \mathcal{A}^{k+\ell}$, and $h \in [H]$ that

$$\begin{aligned}\|b_1^{\theta^t} - b_1^{\theta^*}\|_1 &\le \|b_1^{\theta^t} - \widehat{b}_1^t\|_1 + \|\widehat{b}_1^t - b_1^{\theta^*}\|_1 \\ &= \mathcal{O}(\nu \cdot A^{2k} \cdot \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t}) \end{aligned}$$

with probability at least $1-\delta$. Thus, we complete the proof of the upper bound in (E.7).

Part II: Proof of Upper Bound in (E.8)**.** It suffices to upper bound the following term for all $h \in [H]$ and $t \in [T]$,

$$G_h^t = \sum_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \bigg(\sum_{q_{h-1} \in [d]} \int_{\mathcal{O}} \|u_{h,q_{h-1}}^t\|_1 \mathrm{d}o_h \bigg), \tag{F.14}$$

where we write

$$u_{h,q_{h-1}}^t(\cdot) = \big(\mathbb{B}_h^{\theta^t}(a_h, o_h) - \mathbb{B}_h^{\theta^*}(a_h, o_h)\big)\mathbb{U}_h^{\theta^*}\mathbb{P}_h^{\theta^*}(s_h = \cdot \,|\, q_{h-1}) \cdot \mathbb{P}^{\theta^*,\overline{\pi}^t}(q_{h-1} \,|\, a_{h-\ell}^{h-1}) \tag{F.15}$$

for notational simplicity. We remark that $u_{h,q_{h-1}}^t \in L^1(\mathcal{A}^k \times \mathcal{O}^{k+1})$ is a function in the space $L^1(\mathcal{A}^k \times \mathcal{O}^{k+1})$ by the definition of Bellman operators $\mathbb{B}_h^{\theta^t}$ and $\mathbb{B}_h^{\theta^*}$. In the sequel, we define the vector-valued function

$$u_h^t = [u_{h,1}^t, \ldots, u_{h,d}^t] \in \mathbb{R}^d.$$

It thus holds that

$$G_h^t = \sum_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \int_{\mathcal{O}^{k+2}} \|u_h^t(\tau_{h+1}^{h+k+1})\|_1 \mathrm{d}o_h^{h+k+1}. \tag{F.16}$$

Here the integration and summation are taken with respect to the domain $o_{h+1}^{h+k+1} \in \mathcal{O}^{k+1}$ of u_h^t, the action sequence $a_{h-\ell}^{h-1} \in \mathcal{A}^\ell$ in (F.15), and the action and observation pair $(a_h, o_h) \in \mathcal{A} \times \mathcal{O}$ in the Bellman operators that defines $u_{h,i}^t$ in (F.15). We remark that in (F.16), we abuse the notation slightly and write

$$\|u_h^t(\tau_{h+1}^{h+k+1})\|_1 = \sum_{q_{h-1} \in [d]} |u_{h,q_{h-1}}^t(\tau_{h+1}^{h+k+1})|,$$

where $u^t_{h,q_{h-1}}(\tau^{h+k+1}_{h+1})$ is defined in (F.15). In the sequel, we upper bound the right-hand side of (F.16). By Assumption 5.2, it holds that

$$\|u^t_h\|_1 = \|\mathbb{F}_h^{\theta^*,\overline{\pi}_t,\dagger}\mathbb{F}_h^{\theta^*,\overline{\pi}_t}u^t_h\|_1 \le \gamma\cdot\|\mathbb{F}^{\theta^*,\pi_t}u^t_h\|_1 \tag{F.17}$$

Meanwhile, by the definition of $\mathbb{F}_h^{\theta^*,\overline{\pi}^t}$, we have

$$\mathbb{F}_h^{\theta^*,\overline{\pi}^t}u^t_h = \sum_{q_{h-1}\in[d]} u^t_{h,q_{h-1}}\cdot\mathbb{P}^{\theta^*,\overline{\pi}^t}(o^{h-1}_{h-\ell}\,|\,q_{h-1},a^{h-1}_{h-\ell}). \tag{F.18}$$

By the definition of $u^t_{h,q_{h-1}}$ in (F.15), we further obtain that

$$\begin{aligned}
u^t_{h,q_{h-1}}\cdot\mathbb{P}^{\theta^*,\overline{\pi}^t}(o^{h-1}_{h-\ell}\,|\,q_{h-1},a^{h-1}_{h-\ell}) &= u^t_{h,q_{h-1}}\cdot\frac{\mathbb{P}^{\theta^*,\overline{\pi}^t}(o^{h-1}_{h-\ell},q_{h-1}\,|\,a^{h-1}_{h-\ell})}{\mathbb{P}^{\theta^*,\overline{\pi}^t}(q_{h-1}\,|\,a^{h-1}_{h-\ell})}\\
&= \big(\mathbb{B}_h^{\theta^t}(a_h,o_h)-\mathbb{B}_h^{\theta^*}(a_h,o_h)\big)\mathbb{U}_h^{\theta^*}\mathbb{P}_h^{\theta^*}(s_h=\cdot\,|\,q_{h-1})\cdot\mathbb{P}^{\theta^*,\overline{\pi}^t}(o^{h-1}_{h-\ell},q_{h-1}\,|\,a^{h-1}_{h-\ell})\\
&= \big(\mathbb{B}_h^{\theta^t}(a_h,o_h)-\mathbb{B}_h^{\theta^*}(a_h,o_h)\big)\mathbb{U}_h^{\theta^*}\mathbb{P}^{\theta^*,\overline{\pi}^t}(o^{h-1}_{h-\ell},q_{h-1},s_h=\cdot\,|\,a^{h-1}_{h-\ell}).
\end{aligned}$$

Thus, it follows from (F.18) and the linearity of Bellman operators $\mathbb{B}^\theta_h$ and $\mathbb{U}^\theta_h$ that

$$\begin{aligned}
(\mathbb{F}^{\theta^*,\pi_t}u^t_h)(\underline{\tau}^{h-1}_{h-\ell}) &= \sum_{q_{h-1}\in[d]}\big(\mathbb{B}_h^{\theta^t}(a_h,o_h)-\mathbb{B}_h^{\theta^*}(a_h,o_h)\big)\mathbb{U}_h^{\theta^*}\mathbb{P}^{\theta^*,\overline{\pi}^t}(o^{h-1}_{h-\ell},q_{h-1},s_h=\cdot\,|\,a^{h-1}_{h-\ell})\\
&= \big(\mathbb{B}_h^{\theta^t}(a_h,o_h)-\mathbb{B}_h^{\theta^*}(a_h,o_h)\big)\mathbb{U}_h^{\theta^*}\mathbb{P}^{\theta^*,\overline{\pi}^t}(o^{h-1}_{h-\ell},s_h=\cdot\,|\,a^{h-1}_{h-\ell})\\
&= \big(\mathbb{B}_h^{\theta^t}(a_h,o_h)-\mathbb{B}_h^{\theta^*}(a_h,o_h)\big)\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}^{h-1}_{h-\ell}),
\end{aligned}\tag{F.19}$$

where we marginalize the bottleneck factor q_{h-1} in the second equality. Here recall that $\mathbb{X}_h^{\theta^*,\overline{\pi}^t}$ is the density mapping defined in (3.6) and $\overline{\pi}^t$ is the mixed policy in the t-th iteration. Plugging (F.19) into (F.17), we obtain that

$$\|u^t_h\|_1 \le \gamma\cdot\sum_{a^{h-1}_{h-\ell}\in\mathcal{A}^\ell}\int_{\mathcal{O}^\ell}|\big(\mathbb{B}_h^{\theta^t}(a_h,o_h)-\mathbb{B}_h^{\theta^*}(a_h,o_h)\big)\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}^{h-1}_{h-\ell})|\mathrm{d}o^{h-1}_{h-\ell}. \tag{F.20}$$

Plugging (F.20) into (F.16), we obtain that

$$G^t_h \le \gamma\cdot\sum_{a^h_{h-\ell}\in\mathcal{A}^{\ell+1}}\int_{\mathcal{O}^{\ell+1}}\big\|\big(\mathbb{B}_h^{\theta^t}(a_h,o_h)-\mathbb{B}_h^{\theta^*}(a_h,o_h)\big)\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}^{h-1}_{h-\ell})\big\|_1\mathrm{d}o^h_{h-\ell}. \tag{F.21}$$

It remains to upper bound the right-hand side of (F.21). By triangle inequality, we have

$$\begin{aligned}
&\int_{\mathcal{O}^{\ell+1}}\big\|\big(\mathbb{B}_h^{\theta^t}(a_h,o_h)-\mathbb{B}_h^{\theta^*}(a_h,o_h)\big)\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}^{h-1}_{h-\ell})\big\|_1\mathrm{d}o^h_{h-\ell} \qquad\text{(F.22)}\\
&\quad\le\int_{\mathcal{O}^{\ell+1}}\|\mathbb{B}_h^{\theta^t}(a_h,o_h)\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}^{h-1}_{h-\ell})-\mathbb{Y}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}^{h}_{h-\ell})\|_1\mathrm{d}o^h_{h-\ell}\\
&\qquad+\int_{\mathcal{O}^{\ell+1}}\|\mathbb{B}_h^{\theta^*}(a_h,o_h)\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}^{h-1}_{h-\ell})-\mathbb{Y}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}^{h}_{h-\ell})\|_1\mathrm{d}o^h_{h-\ell}.
\end{aligned}$$

By the identity in (3.8), it holds that

$$\mathbb{B}_h^{\theta^*}(a_h,o_h)\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}^{h-1}_{h-\ell})-\mathbb{Y}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}^{h}_{h-\ell})=0 \tag{F.23}$$

In the sequel, we upper bound the term

$$\int_{\mathcal{O}^{\ell+1}}\|\mathbb{B}_h^{\theta^t}(a_h,o_h)\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}^{h-1}_{h-\ell})-\mathbb{Y}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}^{h}_{h-\ell})\|_1\mathrm{d}o^h_{h-\ell}$$

on the right-hand side of (F.22). The calculation is similar to that of the derivation of (F.13). It holds for all $h \in [H]$ and $t \in [T]$ that

$$\begin{aligned}
&\int_{\mathcal{O}^{\ell+1}} \|\mathbb{B}_h^{\theta^t}(a_h, o_h)\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}_{h-\ell}^{h-1}) - \mathbb{Y}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}_{h-\ell}^{h})\|_1 \mathrm{d}o_{h-\ell}^h \\
&\quad \le \int_{\mathcal{O}^{\ell+1}} \|\mathbb{B}_h^{\theta^t}(a_h, o_h)\widehat{\mathbb{X}}_h^t(\underline{\tau}_{h-\ell}^{h-1}) - \widehat{\mathbb{Y}}_h^t(\underline{\tau}_{h-\ell}^{h})\|_1 \mathrm{d}o_{h-\ell}^h \\
&\qquad + \int_{\mathcal{O}^{\ell+1}} \big\|\big(\mathbb{Y}_h^{\theta^*,\overline{\pi}^t} - \widehat{\mathbb{Y}}_h^t\big)(\underline{\tau}_{h-\ell}^{h})\big\|_1 \mathrm{d}o_{h-\ell}^h \\
&\qquad + \int_{\mathcal{O}^{\ell+1}} \big\|\mathbb{B}_h^{\theta^t}(a_h, o_h)\big(\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}_{h-\ell}^{h-1}) - \widehat{\mathbb{X}}_h^t(\underline{\tau}_{h-\ell}^{h-1})\big)\big\|_1 \mathrm{d}o_{h-\ell}^h. \qquad \text{(F.24)}
\end{aligned}$$

We now upper bound the right-hand side of (F.24). By the definition of confidence set $\mathcal{C}^t$, it holds for all $h \in [H]$ and $t \in [T]$ that

$$\int_{\mathcal{O}^{\ell+1}} \|\mathbb{B}_h^{\theta^t}(a_h, o_h)\widehat{\mathbb{X}}_h^t(\underline{\tau}_{h-\ell}^{h-1}) - \widehat{\mathbb{Y}}_h^t(\underline{\tau}_{h-\ell}^{h})\|_1 \mathrm{d}o_{h-\ell}^h \le \beta_t \cdot \sqrt{1/t}. \qquad \text{(F.25)}$$

Meanwhile, by Assumption 4.1 and the update of density estimators $\widehat{\mathbb{Y}}_h^t$ in (C.2), it holds for all $h \in [H]$ and $t \in [T]$ that

$$\begin{aligned}
\int_{\mathcal{O}^{\ell+1}} \big\|\big(\mathbb{Y}_h^{\theta^*,\overline{\pi}^t} - \widehat{\mathbb{Y}}_h^t\big)(\underline{\tau}_{h-\ell}^{h})\big\|_1 \mathrm{d}o_{h-\ell}^h &= \sum_{a_{h+1}^{h+k} \in \mathcal{A}^k} \int_{\mathcal{O}^{k+\ell+1}} |(\widehat{\mathbb{P}}_h^t - \mathbb{P}_h^{\theta^*,\overline{\pi}^t})(\tau_{h-\ell}^{h+k+1})| \mathrm{d}o_{h-\ell}^{h+k+1} \\
&= \sum_{a_{h+1}^{h+k} \in \mathcal{A}^k} \|(\widehat{\mathbb{P}}_h^t - \mathbb{P}_h^{\theta^*,\overline{\pi}^t})(\cdot \,|\, a_{h-\ell}^{h+k})\|_1 \\
&\le A^k \cdot \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t} \qquad \text{(F.26)}
\end{aligned}$$

with probability at least $1 - \delta$. Similarly, by Assumption 4.1, Lemma E.2, and the update of density estimators $\widehat{\mathbb{X}}_h^t$ in (C.1), we further obtain for all $h \in [H]$ and $t \in [T]$ that

$$\begin{aligned}
&\int_{\mathcal{O}^{\ell+1}} \big\|\mathbb{B}_h^{\theta^t}(a_h, o_h)\big(\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}_{h-\ell}^{h-1}) - \widehat{\mathbb{X}}_h^t(\underline{\tau}_{h-\ell}^{h-1})\big)\big\|_1 \mathrm{d}o_{h-\ell}^h \\
&\quad \le \nu \cdot A^{2k} \cdot \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t} \qquad \text{(F.27)}
\end{aligned}$$

with probability at least $1 - \delta$. Plugging (F.25)–(F.27) into (F.24), we obtain for all $h \in [H]$ and $t \in [T]$ that

$$\begin{aligned}
&\int_{\mathcal{O}^{\ell+1}} \|\mathbb{B}_h^{\theta^*}(a_h, o_h)\widehat{\mathbb{X}}_h^t(\underline{\tau}_{h-\ell}^{h-1}) - \widehat{\mathbb{Y}}_h^t(\underline{\tau}_{h-\ell}^{h})\|_1 \mathrm{d}o_{h-\ell}^h \\
&\quad = \mathcal{O}(\nu \cdot A^{2k} \cdot \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t}) \qquad \text{(F.28)}
\end{aligned}$$

with probability at least $1 - \delta$. By plugging (F.23) and (F.28) into (F.22), we obtain for all $h \in [H]$, $t \in [T]$, and $(a_h^{h+k-1}, a_{h-\ell}^h) \in \mathcal{A}^{k+\ell}$ that

$$\begin{aligned}
&\int_{\mathcal{O}^{\ell+1}} \big\|\big(\mathbb{B}_h^{\theta^t}(a_h, o_h) - \mathbb{B}_h^{\theta^*}(a_h, o_h)\big)\mathbb{X}_h^{\theta^*,\overline{\pi}^t}(\underline{\tau}_{h-\ell}^{h-1})\big\|_1 \mathrm{d}o_{h-\ell}^h \\
&\quad = \mathcal{O}(\nu \cdot A^{2k} \cdot \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t}) \qquad \text{(F.29)}
\end{aligned}$$

with probability at least $1 - \delta$. Plugging (F.29) into (F.21), we obtain for all $h \in [H]$ and $t \in [T]$ that

$$G_h^t = \mathcal{O}\big(\gamma \cdot \nu \cdot A^{2k+\ell} \cdot \sqrt{w_{\mathfrak{E}} \cdot (k+\ell) \cdot \log(H \cdot A \cdot T)/t}\big)$$

with probability at least $1 - \delta$. Here G_h^t is defined in (F.14). Thus, we complete the proof of the upper bound in (E.8). □

F.3 Proof of Lemma F.1

Proof. Recall that we define linear operators $\{\mathbb{T}_h^\theta, \widetilde{\mathbb{O}}_h^\theta\}_{h \in [H]}$ as follows,

$$\begin{aligned}
\big(\mathbb{T}_h^\theta(a_h)f\big)(s_{h+1}) &= \int_{\mathcal{S}} \mathbb{P}_h^\theta(s_{h+1} \,|\, s_h, a_h) \cdot f(s_h) \mathrm{d}s_h, \quad \forall f \in L^1(\mathcal{S}),\ a_h \in \mathcal{A}, \\
\big(\widetilde{\mathbb{O}}_h^\theta(o_h)f\big)(s_h) &= \mathbb{O}_h^\theta(o_h \,|\, s_h) \cdot f(s_h), \quad \forall f \in L^1(\mathcal{S}),\ o_h \in \mathcal{O}.
\end{aligned}$$

Recall that we have

$$\mathbb{B}_h^\theta(a_h, o_h) = \mathbb{U}_{h+1}^\theta \mathbb{T}_h^\theta(a_h) \widetilde{\mathbb{O}}_h^\theta(o_h) \mathbb{U}_h^{\theta,\dagger}.$$

Thus, following from Lemma 3.6 and the fact that $\mathbb{T}_h^\theta(a_h) f$ is linear in ψ_h^θ, it further holds for all $h \in [H]$ and $v_h \in L^1(\mathcal{A}^k \times \mathcal{O}^{k+1})$ that

$$\begin{aligned} &B_{H-1}^\theta(o_{H-1}, a_{H-1}) \ldots \mathbb{B}_{h+1}^\theta(a_{h+1}, o_{h+1}) v_h \\ &\quad = \mathbb{U}_h^\theta \mathbb{T}_{H-1}^\theta(a_{H-1}) \widetilde{\mathbb{O}}_{H-1}^\theta(o_{H-1}) \ldots \mathbb{T}_{h+1}^\theta(a_{h+1}) \widetilde{\mathbb{O}}_{h+1}^\theta(o_{h+1}) \mathbb{U}_{h+1}^{\theta,\dagger} v_h. \end{aligned} \tag{F.30}$$

We now prove Lemma F.1 in the sequel. To begin with, it holds for all $h \in [H]$, $a_{h+1} \in \mathcal{A}$, and $f \in L^1(\mathcal{S})$ that

$$\begin{aligned} &\int_{\mathcal{O}} \|\mathbb{T}_{h+1}^\theta(a_{h+1}) \widetilde{\mathbb{O}}_{h+1}^\theta(o_{h+1}) f\|_1 \mathrm{d}o_{h+1} \\ &\quad \le \int_{\mathcal{S}^2 \times \mathcal{O}} \mathbb{P}^\theta(s_{h+2} \,|\, s_{h+1}, a_{h+1}) \cdot \mathbb{O}_{h+1}^\theta(o_{h+1} \,|\, s_{h+1}) \cdot |f(s_{h+1})| \mathrm{d}o_{h+1} \mathrm{d}s_{h+1} \mathrm{d}s_{h+2} \\ &\quad = \int_{\mathcal{S}} |f(s_{h+1})| \mathrm{d}s_{h+1} = \|f\|_1. \end{aligned}$$

Inductively, it holds for all $h \in [H]$, $a_{h+1}^{H-1} \in \mathcal{A}^{H-h-1}$, and $f \in L^1(\mathcal{S})$ that

$$\int_{\mathcal{O}^{H-h-1}} \|\mathbb{T}_{H-1}^\theta(a_{H-1}) \widetilde{\mathbb{O}}_{H-1}^\theta(o_{H-1}) \ldots \mathbb{T}_{h+1}^\theta(a_{h+1}) \widetilde{\mathbb{O}}_{h+1}^\theta(o_{h+1}) f\|_1 \mathrm{d}o_{h+1}^{H-1} \le \|f\|_1. \tag{F.31}$$

Meanwhile, by the definition of $\mathbb{U}_H^\theta$ in (3.2) of Definition 3.4, it holds for all $f \in L^1(\mathcal{S})$ and $a_h^{h+k-1} \in \mathcal{A}^k$ that

$$\begin{aligned} \int_{\mathcal{O}^{k+1}} |(\mathbb{U}_H^\theta f)(o_H^{H+k}, a_H^{H+k-1})| \mathrm{d}o_H^{H+k} &\le \int_{\mathcal{S} \times \mathcal{O}^{k+1}} \mathbb{P}^\theta(o_H^{H+k} \,|\, s_H, a_H^{H+k-1}) \cdot |f(s_H)| \mathrm{d}o_H^{H+k} \mathrm{d}s_H \\ &= \int_{\mathcal{S}} |f(s_H)| \mathrm{d}s_H = \|f\|_1. \end{aligned} \tag{F.32}$$

Combining (F.30), (F.31), and (F.32) with $h = H$, we obtain that

$$\begin{aligned} &\int_{\mathcal{O}^{H+k-h}} |\mathbb{B}_{H-1}^\theta(o_{H-1}, a_{H-1}) \ldots \mathbb{B}_{h+1}^\theta(a_{h+1}, o_{h+1}) v_h| \mathrm{d}o_{h+1}^{H+k} \\ &\quad = \int_{\mathcal{O}^{H+k-h}} |\mathbb{U}_H^\theta \mathbb{T}_{H-1}^\theta(a_{H-1}) \widetilde{\mathbb{O}}_{H-1}^\theta(o_{H-1}) \ldots \mathbb{T}_{h+1}^\theta(a_{h+1}) \widetilde{\mathbb{O}}_{h+1}^\theta(o_{h+1}) \mathbb{U}_{h+1}^{\theta,\dagger} v_h| \mathrm{d}o_{h+1}^{H+k} \\ &\quad \le \int_{\mathcal{O}^{H-h-1}} \|\mathbb{T}_{H-1}^\theta(a_{H-1}) \widetilde{\mathbb{O}}_{H-1}^\theta(o_{H-1}) \ldots \mathbb{T}_{h+1}^\theta(a_{h+1}) \widetilde{\mathbb{O}}_{h+1}^\theta(o_{h+1}) \mathbb{U}_{h+1}^{\theta,\dagger} v_h\|_1 \mathrm{d}o_{h+1}^{H-1} \\ &\quad \le \|\mathbb{U}_{h+1}^{\theta,\dagger} v_h\|_1. \end{aligned} \tag{F.33}$$

Finally, by Assumption 3.5, it holds that

$$\|\mathbb{U}_{h+1}^{\theta,\dagger} v_h\|_1 \le \nu \cdot \|v_h\|_1.$$

Thus, we completes the proof of Lemma F.1. □

F.4 Proof of Lemma F.2

Proof. Recall that we define

$$v_h = \big(\mathbb{B}_h^\theta(a_h^\pi, o_h) - \mathbb{B}_h^{\theta'}(a_h^\pi, o_h)\big) \mathbb{B}_{h-1}^{\theta'}(a_{h-1}^\pi, o_{h-1}) \ldots \mathbb{B}_1^{\theta'}(a_1^\pi, o_1) b_1^{\theta'}, \quad \forall h \in [H],$$

where the actions a_h^π are taken by the policy π for all $h \in [H]$. To accomplish the proof, we first handle the dependency of the actions a_j^π on policy π for $h - \ell \le j \le h$. To this end, we utilize the following upper bound,

$$\begin{aligned} \int_{\mathcal{O}^h} \|v_h\|_1 \mathrm{d}o_1^h \le \sum_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \int_{\mathcal{O}^H} \Big\| &\big(\mathbb{B}_h^\theta(a_h^\pi, o_h) - \mathbb{B}_h^{\theta'}(a_h, o_h)\big) \mathbb{B}_{h-1}^{\theta'}(a_{h-1}, o_{h-1}) \ldots \\ &\ldots \mathbb{B}_{h-\ell}^{\theta'}(a_{h-\ell}, o_{h-\ell}) \mathbb{B}_{h-\ell-1}^{\theta'}(a_{h-\ell-1}^\pi, o_{h-\ell-1}) \ldots b_1^{\theta'} \Big\|_1 \mathrm{d}o_1^h. \end{aligned} \tag{F.34}$$

Here we abuse the notation of index slightly for simplicity. We remark that the sequence of product of Bellman operators $B_j^{\theta'}(a_j, o_j)$ ends at the index $j = 1$. Recall that we have

$$\mathbb{B}_h^\theta(a_h, o_h) = \mathbb{U}_{h+1}^\theta \mathbb{T}_h^\theta(a_h) \widetilde{\mathbb{O}}_h^\theta(o_h) \mathbb{U}_h^{\theta,\dagger},$$

where the linear operators $\{\mathbb{T}_h^\theta, \widetilde{\mathbb{O}}_h^\theta\}_{h\in[H]}$ are defined in (E.1) and (E.2), respectively. Thus, by Lemma 3.6, it holds for the right-hand side of (F.34) that

$$\begin{aligned}&\mathbb{B}_{h-1}^{\theta'}(a_{h-1}, o_{h-1}) \ldots \mathbb{B}_{h-\ell}^{\theta'}(a_{h-\ell}, o_{h-\ell}) \mathbb{B}_{h-\ell-1}^{\theta'}(a_{h-\ell-1}^\pi, o_{h-\ell-1}) \ldots b_1^{\theta'} \\ &\quad = \mathbb{U}_h^{\theta'} \mathbb{T}_{h-1}^{\theta'}(a_{h-1}) \widetilde{\mathbb{O}}_{h-1}^{\theta}(o_{h-1}) \ldots \widetilde{\mathbb{O}}_1^{\theta'}(o_1) \mu_1, \end{aligned} \tag{F.35}$$

where $\mu_1 \in L^1(\mathcal{S})$ is the initial state probability density function. By the definition of linear operators $\{\mathbb{T}_h^\theta, \widetilde{\mathbb{O}}_h^\theta\}_{h\in[H]}$ in (E.1) and (E.2), respectively, it further holds that

$$\mathbb{T}_{h-1}^{\theta'}(a_{h-1}) \widetilde{\mathbb{O}}_{h-1}^{\theta'}(o_{h-1}) \mathbb{T}_{h-2}^{\theta'}(a_{h-2}) \ldots \widetilde{\mathbb{O}}_1^{\theta'}(o_1) \mu_1 = \mathbb{P}^{\theta', \pi}(o_1^{h-1}, s_h = \cdot \,|\, a_{h-\ell}^{h-1}) \in L^1(\mathcal{S}). \tag{F.36}$$

By plugging (F.35) and (F.36) into (F.34), we obtain that

$$\int_{\mathcal{O}^h} \|v_h\|_1 \mathrm{d}o_1^h \le \sum_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \int_{\mathcal{O}^h} \big\| \big(\mathbb{B}_h^\theta(a_h, o_h) - \mathbb{B}_h^{\theta'}(a_h, o_h)\big) \mathbb{U}_h^{\theta'} \mathbb{P}^{\theta', \pi}(o_1^{h-1}, s_h = \cdot \,|\, a_{h-\ell}^{h-1}) \big\|_1 \mathrm{d}o_1^h. \tag{F.37}$$

Meanwhile, it holds for all $s_h \in \mathcal{S}$ that

$$\mathbb{P}^{\theta', \pi}(o_1^{h-1}, s_h \,|\, a_{h-\ell}^{h-1}) = \sum_{q_{h-1}\in[d]} \mathbb{P}_{h-1}^{\theta'}(s_h \,|\, q_{h-1}) \cdot \mathbb{P}_{h-1}^{\theta', \pi}(q_{h-1}, o_{h-\ell}^{h-1} \,|\, a_{h-\ell}^{h-1}).$$

Thus, it follows from Jensen's inequality that

$$\begin{aligned}&\int_{\mathcal{O}^H} \big\| \big(\mathbb{B}_h^\theta(a_h^\pi, o_h) - \mathbb{B}_h^{\theta'}(a_h, o_h)\big) \mathbb{U}_h^{\theta'} \mathbb{P}^{\theta', \pi}(o_1^{h-1}, s_h = \cdot \,|\, a_{h-\ell}^{h-1}) \big\|_1 \mathrm{d}o_1^{H-1} \\ &\quad \le \int_{\mathcal{O}^H} \sum_{q_{h-1}\in[d]} w_h \cdot \mathbb{P}^{\theta', \pi}(o_1^{h-1}, q_{h-1} \,|\, a_{h-\ell}^{h-1}) \mathrm{d}o_1^{H-1} \\ &\quad = \int_{\mathcal{O}} \sum_{q_{h-1}\in[d]} w_h \cdot \mathbb{P}^{\theta', \pi}(q_{h-1} \,|\, a_{h-\ell}^{h-1}) \mathrm{d}o_h, \end{aligned} \tag{F.38}$$

where we write

$$w_h = \big\| \big(\mathbb{B}_h^\theta(a_h^\pi, o_h) - \mathbb{B}_h^{\theta'}(a_h, o_h)\big) \mathbb{U}_h^{\theta'} \mathbb{P}_{h-1}^{\theta'}(s_h = \cdot \,|\, q_{h-1}) \big\|_1$$

for notational simplicity. By plugging (F.38) into (F.37), we complete the proof of Lemma F.2. □

G Analysis for the Tabular POMDPs

In the sequel, we present an analysis for the tabular POMDPs. We remark that our analysis extends the previous analysis of undercomplete POMDPs (Azizzadenesheli et al., 2016; Guo et al., 2016; Jin et al., 2020a), where the emission matrices are left invertible. In particular, our analysis handles the overcomplete POMDPs with $O < S$, where O and S are the size of observation and state spaces $\mathcal{O}$ amd $\mathcal{S}$, respectively.

G.1 Bellman Operator

We first introduce notations for matrices to simplify the discussions of POMDPs.

Notation. We denote by $M = [f(i,j)]_{i,j} \in \mathbb{R}^{n\times m}$ the n-by-m matrix, where $f(i,j) \in \mathbb{R}$ is the element in the i-th row and j-th column of M. In addition, for a matrix M, we denote by $M_{i,j}$ the (i,j)-th element of M.

In addition, recall that we denote by $\tau_h^{h+k} = \{o_h, a_h, \ldots, a_{h+k-1}, o_{h+k}\}$ the trajectory from the h-th observation o_h to the $(h+k)$-th observation o_{h+k}. Similarly, we denote by $\underline{\tau}_h^k = (o_h, a_h, \ldots, o_{h+k}, a_{h+k})$ the trajectory from the h-th observation o_h to the $(h+k)$-th action a_{h+k}. We denote by $a_h^{h+k-1} = (a_h, \ldots, a_{h+k-1})$ and $o_h^{h+k} = (o_h, \ldots, o_{h+k})$ the action and observation sequences, respectively. Meanwhile, recall that we write

$$\begin{aligned}\mathbb{P}^\pi(\tau_h^{h+k}) &= \mathbb{P}^\pi(o_h, \ldots, o_{h+k} \,|\, a_h, \ldots, a_{h+k-1}) = \mathbb{P}^\pi(o_h^{h+k} \,|\, a_h^{h+k-1}), \\ \mathbb{P}^\pi(\tau_h^{h+k} \,|\, s_h) &= \mathbb{P}^\pi(o_h, \ldots, o_{h+k} \,|\, s_h, a_h, \ldots, a_{h+k-1}) = \mathbb{P}^\pi(o_h^{h+k} \,|\, s_h, a_h^{h+k-1}).\end{aligned}$$

for notational simplicity.

Forward Emission Operator. In the sequel, we define several matrices that describes the transition and emission in POMDPs. We define

$$\widetilde{\mathbb{O}}_h(o_h) = \mathbb{D}\big(\mathbb{O}_h(o_h \,|\, \cdot)\big) = \mathbb{D}\Big(\big[\mathbb{O}(o_h \,|\, s_h)\big]_{s_h}\Big) \in \mathbb{R}^{S\times S},$$
$$\mathbb{T}_h(a_h) = \mathbb{P}_h(\cdot \,|\, \cdot, a_h) = \big[\mathbb{P}_h(s_{h+1} \,|\, s_h, a_h)\big]_{s_h, s_{h+1}} \in \mathbb{R}^{S\times S},$$
$$\mathbb{O}_h = \mathbb{O}_h(\cdot \,|\, \cdot) = \big[\mathbb{O}_h(o_h \,|\, s_h)\big]_{o_h, s_h} \in \mathbb{R}^{O\times S},$$

where we denote by $\mathbb{D}(v) \in \mathbb{R}^{S\times S}$ the diagonal matrix where the diagonal entries aligns with the vector $v \in \mathbb{R}^S$.

Definition G.1 (Forward Emission Operator)**.** For all $h \in [H]$ and $k > 0$, we define the following forward emission operator,

$$\begin{aligned}\mathbb{U}_h &= \mathbb{O}_{h+k}\mathbb{T}_{h+k-1}(\cdot)\widetilde{\mathbb{O}}_{h+k-1}(\cdot)\cdots\mathbb{T}_h(\cdot)\widetilde{\mathbb{O}}_h(\cdot)\\ &= \big[\mathbb{1}(o_{h+k})^\top\mathbb{O}_{h+k}\mathbb{T}_{h+k-1}(a_{h+k-1})\widetilde{\mathbb{O}}_{h+k-1}(o_{h+k-1})\\ &\qquad \cdots\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\,\mathbb{1}(s_h)\big]_{\tau_h^{h+k}, s_h} \in \mathbb{R}^{(O^{k+1}\cdot A^k)\times S},\end{aligned}$$

where we index the column of $\mathbb{U}_h$ by the state $s_h \in \mathcal{S}$, and the row of $\mathbb{U}_h$ by the observation and action arrays

$$a_h^{h+k-1} = (a_h, \ldots, a_{h+k-1}) \in \mathcal{A}^k, \quad o_h^{h+k} = (o_h, \ldots, o_{h+k}) \in \mathcal{O}^{k+1}.$$

Lemma G.2 (Forward Emission Operator)**.** It holds for all $h \in [H]$ that

$$\mathbb{U}_h = \big[\mathbb{P}(\tau_h^{h+k} \,|\, s_h)\big]_{\tau_h^{h+k}, s_h} \in \mathbb{R}^{(O^{k+1}\cdot A^k)\times S}.$$

Proof. See §H.1 for a detailed proof. □

Lemma G.2 characterizes the semantic meaning of the forward emission operator $\mathbb{U}_{h+1}$. More specifically, Lemma G.2 allows us to write $\mathbb{U}_{h+1}$ in the following operator form,

$$\mathbb{U}_h = \mathbb{P}(\underbrace{o_h^{h+k} = \cdot}_{\text{(a)}} \,|\, \underbrace{s_h = \cdot}_{\text{(b)}}, \underbrace{a_h^{h+k-1} = \cdot}_{\text{(c)}}) \in \mathbb{R}^{(O^{k+1}\cdot A^k)\times S},$$

where (a) and (c) correspond to the row indices, and (b) corresponds to the column indices of the forward emission matrix $\mathbb{U}_h$. For a state distribution vector $\mu_h \in \mathbb{R}^S$ of state s_h, it holds that

$$\mathbb{U}_h\mu_h = \mathbb{P}(o_h^{h+k} = \cdot \,|\, a_h^{h+k-1} = \cdot) \in \mathbb{R}^{O^{k+1}\cdot A^k},$$

which corresponds to the forward emission probability of $o_h^{h+k} = (o_h, \ldots, o_{h+k})$ given an action sequence $a_h^{h+k-1} = (a_h, \ldots, a_{h+k-1})$.

Assumption G.3 (Future Sufficiency)**.** We assume for some $k > 0$ that the forward emission operator $\mathbb{U}_h$ has full column rank for all $h \in [H]$. We denote by $\mathbb{U}_h^\dagger$ the left inverse of $\mathbb{U}_h$ for all $h \in [H]$. We assume further that $\|\mathbb{U}_h^\dagger\|_{1\mapsto 1} \le \nu$ for all $h \in [H]$.

Planning with Bellman Operator. In the sequel, we introduce the Bellman operators $\{\mathbb{B}_h\}_{h\in[H]}$, which plays a central role in solving the overcomplete POMDP. We define the Bellman operators as follows.

Definition G.4 (Bellman Operator)**.** For all $h \in [H]$, we define the Bellman operator $\mathbb{B}_h$ as follows,

$$\mathbb{B}_h(a_h, o_h) = \mathbb{U}_{h+1}\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\mathbb{U}_h^\dagger, \quad \forall (a_h, o_h) \in \mathcal{A}\times\mathcal{O}.$$

Given the Bellman operators $\{\mathbb{B}_h\}_{h\in[H]}$, we are able to estimate the probability of any given observation sequence $(o_1, a_1, \ldots, o_H)$. In particular, the following lemma holds.

Lemma G.5. It holds for the Bellman operator defined in Definition G.4 that

$$\mathbb{P}(\tau_1^{H-1}) = \mathbb{1}(o_H)^\top\mathbb{B}_{H-1}(a_{H-1}, o_{H-1})\ldots\mathbb{B}_1(a_1, o_1)b_1, \quad b_1 = \mathbb{U}_1\mu_1.$$

Here $\mathbb{1}(o_H)$ is an indicator vector that takes value one at the indices (o_H^{H+k}, a_H^{H+k-1}) for any dummy observation sequence o_{H+1}^{H+k} and a randomly fixed action sequence a_H^{H+k-1}. In addition, $\mu_1 \in \mathbb{R}^S$

is the probability array of initial state distribution, and $b_1 \in \mathbb{R}^{O^{k+1} \cdot A^k}$ is the probability distribution of the first k steps, namely,

$$b_1 = \mathbb{U}_1 \mu_1 = \big[\mathbb{P}(\tau_1^k)\big]_{\tau_h^k} \in \mathbb{R}^{O^{k+1} \cdot A^k}.$$

Proof. See §H.2 for a detailed proof. □

Lemma G.5 allows us to estimate the probability of any given trajectory. In addition, for a deterministic policy π, it further holds that

$$\mathbb{P}^\pi(o_1^H) = \mathbb{P}\big(o_1^H \,|\, (a^\pi)_1^{H-1}\big), \tag{G.1}$$

where $(a^\pi)_1^{H-1} = (a_1^\pi, \ldots, a_{H-1}^\pi)$ is the action sequence determined the observation sequence o_1^{H-1} and the deterministic policy π. Thus, for a given deterministic policy π, one can evaluate the policy π based on the Bellman operators as follows,

$$\begin{aligned} V^\pi &= \sum_{o_1^{H-1} \in \mathcal{O}^H} \mathbb{P}^\pi(o_1^H) \cdot \sum_{h=1}^H r(o_h) \\ &= \sum_{o_1^{H-1} \in \mathcal{O}^H} \sum_{h=1}^H r(o_h) \cdot \mathbb{1}(o_H)^\top \mathbb{B}_{H-1}(a_{H-1}, o_{H-1}) \ldots \mathbb{B}_1(a_1, o_1) b_1. \end{aligned}$$

Estimating the Bellman Operator. To estimate the Bellman operators based on interactions, we utilize the following identity of Bellman operators,

$$\mathbb{B}_h(a_h, o_h)\mathbb{X}_h(o_h^{h+k}) = \mathbb{Y}_h(a_{h-\ell}^{h+k}, o_h). \tag{G.2}$$

Here we define the probability tensors $\mathbb{X}_h$ and $\mathbb{Y}_h$ as follows,

$$\begin{aligned} \mathbb{X}_h(a_{h-\ell}^{h-1}) &= \mathbb{U}_h \mathbb{T}_{h-1}(a_{h-1}) \widetilde{\mathbb{O}}_{h-1}(\cdot) \ldots \mathbb{T}_{h-\ell}(a_{h-\ell}) \widetilde{\mathbb{O}}_{h-\ell}(\cdot) \mu_{h-\ell}, \\ \mathbb{Y}_h(a_{h-\ell}^{h}, o_h) &= \mathbb{U}_{h+1} \mathbb{T}_h(a_h) \widetilde{\mathbb{O}}_h(o_h) \mathbb{T}_{h-1}(a_{h-1}) \widetilde{\mathbb{O}}_{h-1}(\cdot) \ldots \mathbb{T}_{h-\ell}(a_{h-1}) \widetilde{\mathbb{O}}_{h-\ell}(\cdot) \mu_{h-\ell}, \end{aligned}$$

where $\mu_{h-\ell} \in \mathbb{R}^S$ is a probability density array for the state $s_{h-\ell}$. The following lemma characterizes the semantic meaning of the probability tensors $\mathbb{X}_h$ and $\mathbb{Y}_h$ for all $h \in [H]$.

Lemma G.6. Let $\mathbb{P}_{h-\ell}(s_{h-\ell} = \cdot) = \mu_{h-1} \in \mathbb{R}^S$ be a probability density array for the state $s_{h-\ell}$. It holds for all $h \in [H]$ that

$$\begin{aligned} \mathbb{X}_h(a_{h-\ell}^{h-1}) &= \big[\mathbb{P}(\tau_{h-\ell}^{h+k})\big]_{\tau_h^{h+k}, o_{h-\ell}^{h-1}} \in \mathbb{R}^{(A^{k+1} \cdot O^k) \times O^\ell}, \\ \mathbb{Y}_h(a_{h-\ell}^{h}, o_h) &= \big[\mathbb{P}(\tau_{h-\ell}^{h+k+1})\big]_{\tau_{h+1}^{h+k+1}, o_{h-\ell}^{h-1}} \in \mathbb{R}^{(A^{k+1} \cdot O^k) \times O^\ell}. \end{aligned}$$

Proof. See §H.3 for a detailed proof. □

G.2 Algorithm

We now introduce RTC under the tabular POMDPs. In particular, RTC iteratively (i) collects data and fit the density of visitation trajectory, (ii) fits the Bellman operators and construct confidence sets, and (iii) conducts optimistic planning. See Algorithm 2 for the summary.

We remark that the data collection process is identical to that for the low-rank POMDPs. Meanwhile, in the tabular POMDPs, we estimate the density of visitation trajectory by count-based estimators

as follows.

$$\widehat{b}_1^t = \frac{1}{t} \cdot \sum_{a_1^k \in \mathcal{A}^k} \Bigg(\sum_{\tau_1^{k+1} \in \mathcal{D}^t(a_1^k)} \mathbb{1}(\tau_1^{k+1}) \Bigg), \tag{G.3}$$

$$\widehat{\mathbb{X}}_h^t(a_{h-\ell}^{h-1}) = \frac{1}{t} \cdot \sum_{a_h^{h+k-1} \in \mathcal{A}^k} \Bigg(\sum_{\tau_{h-\ell}^{h+k} \in \mathcal{D}^t(a_{h-\ell}^{h+k-1})} \mathbb{1}(\tau_h^{h+k}) \, \mathbb{1}(o_{h-\ell}^{h-1})^\top \Bigg), \tag{G.4}$$

$$\widehat{\mathbb{Y}}_h^t(a_{h-\ell}^{h}, o_h) = \frac{1}{t} \cdot \sum_{a_{h+1}^{h+k} \in \mathcal{A}^k} \Bigg(\sum_{\tau_{h-\ell}^{h+k+1} \in \mathcal{D}^t(a_{h-\ell}^{h+k})} \mathbb{1}(\tau_{h+1}^{h+k+1}) \, \mathbb{1}(o_{h-\ell}^{h-1})^\top \Bigg), \tag{G.5}$$

In the sequel, we summarize the estimations of initial trajectory density and Bellman operators in the t-th iterate by the parameter θ^t. Accordingly, we estimate the Bellman operator in the t-th iterate by minimizing the following objective,

$$\widehat{L}_h^t = \sup_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \|\mathbb{B}_h^\theta(a_h, o_h)\widehat{\mathbb{X}}_h^t(a_{h-\ell}^{h-1}) - \mathbb{Y}_h^t(a_{h-\ell}^h, o_h)\|_1$$

We define the following confidence set of the parameter θ in the t-th iteration.

$$\mathcal{C}^t = \Big\{ \theta \in \Theta : \max_{h \in [H]} \big\{ \|b_1^\theta - \widehat{b}_1^t\|_1, \widehat{L}_h^t \big\} \le \beta_t \cdot \sqrt{1/t}, \forall h \in [H] \Big\}. \tag{G.6}$$

where we set

$$\beta_t = (1+\nu) \cdot (k+\ell) \cdot \sqrt{A^{5k+1} \cdot O^{k+\ell} \cdot \log(O \cdot A \cdot T \cdot H/\delta)/t}.$$

Note that the initial density b_1^θ and Bellman operators $\{\mathbb{B}_h^\theta\}_{h \in [H]}$ are sufficient for policy evaluation since they recovers the visitation density of an arbitrary deterministic policy (Lemma G.7). We conduct optimistic planning in the t-th iteration as follows,

$$\pi^t = \operatorname*{argmax}_{\pi \in \Pi, \theta^t \in \mathcal{C}^t} V^\pi(\theta^t),$$

where $V^\pi(\theta^t)$ is the policy evaluation of π with parameter θ^t and Π is the set of all deterministic policies.

Algorithm 2 Represent to Control for Tabular POMDP

Require: Number of iterates $T \in \mathbb{N}$. A set of tuning parameters $\{\beta_t\}_{t \in [T]}$.

1: **Initialization:** Set π_0 as a deterministic policy. Set the dataset $\mathcal{D}_h^0(a_{h-\ell}^{h+k})$ as an empty set for all $(h, a_{h-\ell}^{h+k}) \in [H] \times \mathcal{A}^{k+\ell+1}$.
2: **for** $t \in [T]$ **do**
3: **for** $(h, a_{h-\ell}^{h+k}) \in [H] \times \mathcal{A}^{k+\ell+1}$ **do**
4: Start a new episode from the $(1-\ell)$-th step.
5: Execute policy π^{t-1} till the $(h-\ell)$-th step and receive the observations ${}^t o_{1-\ell}^{h-\ell}$.
6: Execute the action sequence $a_{h-\ell}^{h+k}$ regardless of the observations and receive the observations ${}^t o_{h-\ell+1}^{h+k+1}$.
7: Update the dataset $\mathcal{D}_h^t(a_{h-\ell}^{h+k}) \leftarrow \mathcal{D}_h^{t-1}(a_{h-\ell}^{h+k}) \cup \big\{ {}^t o_{h-\ell+1}^{h+k+1} \big\}$.
8: **end for**
9: Update the density mappings $\widehat{\mathbb{X}}_h^t$ and $\widehat{\mathbb{Y}}_h^t$ by (G.4) and (G.5), respectively.
10: Update the initial density estimation $\widehat{b}_1^t(\tau_1^H) \leftarrow$ by (G.3).
11: Update the confidence set $\mathcal{C}^t$ by (G.6).
12: Update the policy $\pi^t \leftarrow \operatorname{argmax}_{\pi \in \Pi} \max_{\theta \in \mathcal{C}^t} V^\pi(\theta)$.
13: **end for**
14: **Output:** policy set $\{\pi^t\}_{t \in [T]}$.

G.3 THEORY

In the sequel, we present the sample efficiency analysis of RTC for the tabular POMDPs.

Calculating the Performance Difference. Similar to the analysis under the low-rank POMDPs, we first calculate the performance difference of a policy between two different POMDPs defined by the parameter θ and θ', respectively. The following lemma is adopted from Jin et al. (2020a).

Lemma G.7 (Trajectory Density (Jin et al., 2020a))**.** It holds that

$$\mathbb{P}^{\theta,\pi}(o_1^{H-1}) = \mathbb{P}^{\theta}\big(o_1^{H-1} \,|\, (a^{\pi})_1^{H}\big),$$

where $(a^{\pi})_1^H = (a_1^{\pi}, \ldots, a_H^{\pi})$ and $a_h^{\pi} = \pi(a_1^{h-1}, o_1^h)$ is the action taken by π in the h-th step for all $h \in [H]$.

Proof. See Jin et al. (2020b) for a detailed proof. □

We now calculate the performance difference in the following lemma.

Lemma G.8 (Performance Difference)**.** It holds for any policy π that

$$\begin{aligned}
&|V^{\pi}(\theta) - V^{\pi}(\theta')| \\
&\quad\le \nu\cdot\sqrt{S}\cdot H\cdot\sum_{h=2}^{H-1}\sum_{o_1^h\in\mathcal{O}^h}\big\|\big(\mathbb{B}_h^{\theta}(a_h^{\pi},o_h)-\mathbb{B}_h^{\theta'}(a_h^{\pi},o_h)\big)\mathbb{B}_{h-1}^{\theta'}(a_{h-1}^{\pi},o_{h-1})\ldots\mathbb{B}_1^{\theta'}(a_1^{\pi},o_1)b_1^{\theta'}\big\|_1 \\
&\qquad+\nu\cdot\sqrt{S}\cdot H\cdot\sum_{a_1\in\mathcal{A}}\sum_{o_1\in\mathcal{O}}\big\|\big(\mathbb{B}_1^{\theta}(a_1^{\pi},o_1)-\mathbb{B}_1^{\theta'}(a_1^{\pi},o_1)\big)b_1^{\theta'}\big\|_1+\nu\cdot\sqrt{S}\cdot H\cdot\|b_1^{\theta}-b_1^{\theta'}\|_1,
\end{aligned}$$

where $\mathbb{B}_h^{\theta}$ is the Bellman operator corresponding to the parameter θ for all $h \in [H]$, and $b_1^{\theta} = \mathbb{U}_{1,k}^{\theta}\mu_1$ is the initial trajectory distribution corresponding to the parameter θ. Here the action $a_h^{\pi} = \pi((a^{\pi})_1^{H-1}, o_1^{H-1})$ is the action taken by π in the h-th step for all $h \in [H]$.

Proof. See §H.4 for a detailed proof. □

We define the following state density array,

$$\begin{aligned}
\mu_{h-1}^{\theta}(a_{h-\ell}^{h-1}, o_1^{h-1}; \pi) &= \underbrace{\widetilde{\mathbb{O}}_{h-1}^{\theta}(o_{h-1})\mathbb{T}_{h-2}^{\theta}(a_{h-2})\cdots\mathbb{T}_{h-\ell}^{\theta}(a_{h-\ell})\widetilde{\mathbb{O}}_{h-\ell}^{\theta}(o_{h-\ell})}_{\text{(i)}} \\
&\quad\underbrace{\cdot\mathbb{T}_{h-\ell-1}^{\theta}(a_{h-\ell-1}^{\pi})\cdots\mathbb{T}_1^{\theta}(a_1^{\pi})\widetilde{\mathbb{O}}_1^{\theta}(o_1)\mu_1}_{\text{(ii)}} \\
&= \big[\mathbb{P}^{\theta,\pi}(s_h, o_1^{h-1} \,|\, a_{h-\ell}^{h-1})\big]_{s_h\in\mathcal{S}} \in \mathbb{R}^S. \qquad \text{(G.7)}
\end{aligned}$$

Here the actions $a_{h-\ell-1}^{\pi}, \ldots, a_1^{\pi}$ in (ii) of (G.7) is determined by the observations array $o_1^{h-\ell-2}$ and the policy π. Meanwhile, the action array $a_{h-\ell}^{h-1}$ is the fixed action array that defines the state density array $\mu_{h-1}^{\theta'}(\pi, a_{h-\ell}^{h-1}, o_1^{h-1})$. In addition, we denote by $\mathbb{P}^{\theta}$ the probability that corresponds to the transition dynamics defined by the operators $\{\widetilde{\mathbb{O}}_h^{\theta}, \mathbb{T}_h^{\theta}\}_{h\in[H]}$. Based on (G.7), we further define the following marginal state density array,

$$\begin{aligned}
\widetilde{\mu}_{h-1}^{\theta}(a_{h-\ell}^{h-1}; \pi) &= \sum_{o_1^{h-1}\in\mathcal{O}^{h-1}} \mu_{h-1}^{\theta}(\pi, a_{h-\ell}^{h-1}, o_1^{h-1}) \\
&= \big[\mathbb{P}^{\theta,\pi}(s_{h-1} \,|\, a_{h-\ell}^{h-1})\big]_{s_h\in\mathcal{S}} \in \mathbb{R}^S. \qquad \text{(G.8)}
\end{aligned}$$

The marginal state density array $\widetilde{\mu}_{h-1}^{\theta}(a_{h-\ell}^{h-1}; \pi)$ captures the state distribution of s_{h-1} given the following interaction protocol: (i) starting with the initial observation, interacting with the environment based on policy π till the $(h-\ell)$-th step and observing $o_{h-\ell}$, and (ii) interacting with the environment with a fixed action sequence $a_{h-\ell}^{h-1}$ regardless of the observations till the $(h-1)$-th step and observing o_{h-1}. We remark that such interaction protocol is identical to the sampling process in Line 4–6 of Algorithm 2. The following lemma upper bounds the performance difference calculated in Lemma G.8.

Lemma G.9 (Upper Bound of Performance Difference)**.** It holds for all π and $h > 1$ that

$$\sum_{o_1^h \in \mathcal{O}^h} \big\| \big(\mathbb{B}_h^{\theta}(a_h^{\pi}, o_h) - \mathbb{B}_h^{\theta'}(a_h^{\pi}, o_h)\big) \mathbb{B}_{h-1}^{\theta'}(a_{h-1}^{\pi}, o_{h-1}) \ldots \mathbb{B}_1^{\theta'}(a_1^{\pi}, o_1) b_1^{\theta'} \big\|_1$$
$$\leq \sum_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \sum_{o_h \in \mathcal{O}} \sum_{s_{h-1} \in \mathcal{S}} \big\| \big(\mathbb{B}_h^{\theta}(a_h, o_h) - \mathbb{B}_h^{\theta'}(a_h, o_h)\big) \mathbb{U}_{h,k}^{\theta'} \mathbb{T}_{h-1}^{\theta'}(s_{h-1}, a_{h-1}) \big\|_1 \cdot \mathbb{P}^{\pi}(s_{h-1} \,|\, a_{h-\ell}^{h-1}),$$

where the action $a_h^{\pi} = \pi((a^{\pi})_1^{h-1}, o_1^h)$ is the action taken by π in the h-th step for all $h \in [H]$.Here $\mathbb{T}_{h-1}^{\theta'}(s_{h-1}, a_{h-1}) \in \mathbb{R}^S$ is the state distribution array $[\mathbb{T}_{h-1}^{\theta'}(s_h \,|\, s_{h-1}, a_{h-1})]_{s_h} \in \mathbb{R}^S$ for all $h > 1$.

Proof. See §H.5 for a detailed proof. □

Confidence Set Analysis. We now analyze the confidence set utilized for optimistic planning. We define the following visitation measure of mix policy in the t-th iteration for all $t > 0$,

$$\mathbb{P}^t(s_h) = \frac{1}{t} \cdot \sum_{\omega=0}^{t-1} \mathbb{P}^{\pi^{\omega}}(s_h),$$

where $\{\pi^{\omega}\}_{\omega \in [t]}$ is the set of policy returned by Algorithm 2. Meanwhile, recall that we define the empirical density estimators,

$$\widehat{b}_1^t = \frac{1}{t} \cdot \sum_{a_1^k \in \mathcal{A}^k} \Bigg(\sum_{\tau_1^{k+1} \in \mathcal{D}^t(a_1^k)} \mathbb{1}(\tau_1^{k+1}) \Bigg),$$

$$\widehat{\mathbb{X}}_h^t(a_{h-\ell}^{h-1}) = \frac{1}{t} \cdot \sum_{a_h^{h+k-1} \in \mathcal{A}^k} \Bigg(\sum_{\tau_{h-\ell}^{h+k} \in \mathcal{D}^t(a_{h-\ell}^{h+k-1})} \mathbb{1}(\tau_h^{h+k}) \, \mathbb{1}(o_{h-\ell}^{h-1})^{\top} \Bigg),$$

$$\widehat{\mathbb{Y}}_h^t(a_{h-\ell}^h, o_h) = \frac{1}{t} \cdot \sum_{a_{h+1}^{h+k} \in \mathcal{A}^k} \Bigg(\sum_{\tau_{h-\ell}^{h+k+1} \in \mathcal{D}^t(a_{h-\ell}^{h+k})} \mathbb{1}(\tau_{h+1}^{h+k+1}) \, \mathbb{1}(o_{h-\ell}^{h-1})^{\top} \Bigg),$$

where we denote by $\mathbb{1}(x)$ the indicator vector that takes value one at the index x. Recall that we define the confidence set as follows,

$$\mathcal{C}^t = \Big\{ \theta \in \Theta : \max\big\{ \|b_1^{\theta} - \widehat{b}_1^t\|_1, \widehat{L}_h^t \big\} \leq \beta_t \cdot \sqrt{1/t}, \forall h \in [H] \Big\}.$$

where we set

$$\beta_t = A^k \cdot (k + \ell) \cdot \sqrt{\log(O \cdot A \cdot T \cdot H/\delta)}.$$

Recall that in the t-th iteration of Algorithm 2, we update the policy π^t as follows,

$$\pi^t = \operatorname*{argmax}_{\pi \in \Pi, \theta \in \mathcal{C}^t} V^{\pi}(\theta).$$

The following lemma shows that the empirical estimations aligns closely to the true density corresponding to the exploration.

Lemma G.10 (Concentration Bound of Density Estimation)**.** It holds for all $t \in [T]$ with probability at least $1 - \delta$ that

$$\max\big\{ \|\mathbb{X}_h^t(a_{h-\ell}^{h-1}) - \widehat{\mathbb{X}}_h^t(a_{h-\ell}^{h-1})\|_F, \|b_1 - \widehat{b}_1^t\|_2, \|\mathbb{Y}_h^t(a_{h-\ell}^h, o_h) - \widehat{\mathbb{Y}}_h^t(a_{h-\ell}^h, o_h)\|_F \big\}$$
$$= \mathcal{O}\big(A^k \cdot (k + \ell) \cdot \sqrt{\log(O \cdot A \cdot T \cdot H/\delta)/t} \big)$$

with probability at least $1 - \delta$.

Proof. See §H.6 for a detailed proof. □

In what follows, we define the reverse emission operators for the tabular POMDPs.

Definition G.11 (Revserse Emission)**.** For all $1 < h \le H$ and $a_{h-\ell}^{h-2} \in \mathcal{A}^{\ell-1}$, we define

$$\begin{aligned} A^{\pi}_{h-1,\ell}(a_{h-\ell}^{h-2}) &= \widetilde{\mathbb{O}}_{h-1}(o_{h-1} = \cdot)\mathbb{T}_{h-2}(a_{h-2}) \ldots \cdot \widetilde{\mathbb{O}}_{h-\ell}(o_{h-\ell} = \cdot)\mathbb{D}\big(\mathbb{P}^{\pi}(s_{h-\ell} = \cdot)\big) \\ &= \widetilde{\mathbb{O}}_{h-1}(o_{h-1} = \cdot)\big(\Pi_{i=h-2}^{h-\ell}\mathbb{T}_i(a_i)\widetilde{\mathbb{O}}_i(o_i = \cdot)\big)\widetilde{\mathbb{O}}_{h-\ell}(o_{h-\ell} = \cdot)\mathbb{D}\big(\mathbb{P}^{\pi}(s_{h-\ell} = \cdot)\big) \\ &\in \mathbb{R}^{S \times O^{\ell}}. \end{aligned}$$

By Definition G.11 and the identity in Lemma G.6, we have the following identity,

$$\mathbb{X}_h^t(a_{h-\ell}^{h-1}) = \mathbb{U}_h\mathbb{T}_{h-1}(a_{h-1})A^{\overline{\pi}^t}_{h-1,\ell}(a_{h-\ell}^{h-2}), \qquad \forall a_{h-\ell}^{h-1} \in \mathcal{A}^{\ell}. \tag{G.9}$$

Here we denote by $\overline{\pi}^t$ the mixed policy induced by the policies $\{\pi^{\omega}\}_{\omega \in [t]}$ obtained till the t-th iteration of Algorithm 2.

Assumption G.12 (Past Sufficiency)**.** We define the following matrix for all policy π, $a_{h-\ell}^{h-2} \in \mathcal{A}^{\ell-1}$, and $0 < h \le H$,

$$C^{\pi}_{h-1,\ell}(a_{h-\ell}^{h-1}) = \mathbb{D}(\mathbb{P}^{\pi}(s_{h-1} \,|\, a_{h-\ell}^{h-1}))^{-1}A^{\pi}_{h-1,\ell} \in \mathbb{R}^{S \times (A^{\ell-1} \cdot O^{\ell})}.$$

Here recall that $\mathbb{D}(v)$ is the diagonal matrix where the diagonal entries align with the vector v. We assume that $C^{\pi}_{h-1,\ell}$ has full row rank for all π, $a_{h-\ell}^{h-1} \in \mathcal{A}^{\ell-1}$, and $0 < h \le H$. We denote by $C^{\pi,\dagger}_{h-1,\ell}(a_{h-\ell}^{h-1})$ the right inverse of $C^{\pi}_{h-1,\ell}(a_{h-\ell}^{h-1})$. We assume further that

$$\|C^{\pi,\dagger}_{h-1,\ell}(a_{h-\ell}^{h-1})^{\top}\|_{1 \mapsto 1} \le \gamma$$

for an absolute constant $\gamma > 0$ for all π, $a_{h-\ell}^{h-1} \in \mathcal{A}^{\ell-1}$, and $0 < h \le H$.

Lemma G.13 (Good Event Probability)**.** Under Assumptions G.3 and G.12, it holds with probability at least $1 - \delta$ that $\theta^* \in \mathcal{C}_t$. Moreover, it holds for all $t \in [T]$ with probability at least $1 - \delta$ that

$$\|b_1 - \widehat{b}_1^{\theta^t}\|_1 = \mathcal{O}\Big(\nu \cdot (k+\ell) \cdot \sqrt{A^{5k+1} \cdot O^{k+\ell} \cdot \log(O \cdot A \cdot T \cdot H/\delta)/t}\Big), \tag{G.10}$$

$$\big\|\big(\mathbb{B}_1^{\theta^t}(a_1, o_1) - \mathbb{B}_1^{\theta^*}(a_1, o_1)\big)b_1\big\|_1 = \mathcal{O}\Big(\nu \cdot (k+\ell) \cdot \sqrt{A^{5k+1} \cdot O^{k+\ell} \cdot \log(O \cdot A \cdot T \cdot H/\delta)/t}\Big), \tag{G.11}$$

Meanwhile, it holds for all $1 < h \le H$ and $t \in [T]$ with probability at least $1 - \delta$ that

$$\begin{aligned} &\sum_{s_{h-1} \in \mathcal{S}} \big\|\big(\mathbb{B}_h^{\theta^t}(a_h, o_h) - \mathbb{B}_h^{\theta^*}(a_h, o_h)\big)\mathbb{U}_{h,k}^{\theta^*}\mathbb{T}_{h-1}^{\theta^*}(s_{h-1}, a_{h-1})\big\|_1 \cdot \mathbb{P}^t(s_{h-1} \,|\, a_{h-\ell}^{h-1}) \\ &\quad = \mathcal{O}\Big(\gamma \cdot \nu \cdot (k+\ell) \cdot \sqrt{A^{5k+\ell} \cdot O^{k+1} \cdot \log(O \cdot A \cdot T \cdot H/\delta)/t}\Big) \end{aligned} \tag{G.12}$$

for all $1 < h \le H$. Here $\{\mathbb{B}_h^{\theta^t}\}_{h \in [H]}$ and b_1^t are the updated Bellman operators and initial trajectory distribution, respectively, in the t-th iterate of Algorithm 2.

Proof. See §H.7 for a detailed proof. □

Sample Complexity Analysis. We are now ready to present the sample efficiency analysis of RTC under the tabular POMDPs.

Theorem G.14. Let

$$T = \mathcal{O}\Big(\mathrm{poly}(S, A, O, H) \cdot \gamma^2 \cdot \nu^4 \cdot (k+\ell)^2 \cdot \log(O \cdot A \cdot H/\delta)/\epsilon^2\Big).$$

Under Assumptions G.3 and G.12, it holds with probability at least $1 - \delta$ that $\overline{\pi}^T$ is ϵ-optimal. Here $\mathrm{poly}(S, A, O, H)$ is a polynomial that takes the following order,

$$\mathrm{poly}(S, A, O, H) = \mathcal{O}(S^2 \cdot A^{10k+2\ell} \cdot O^{2k+2\ell} \cdot H^2).$$

Proof. It holds that

$$V^*(\theta^*) - V^{\overline{\pi}^t}(\theta^*) = \frac{1}{T} \cdot \sum_{t=1}^{T} V^*(\theta^*) - V^{\pi^t}(\theta^*). \tag{G.13}$$

It suffices to upper bound the performance difference

$$V^*(\theta^*) - V^{\pi^t}(\theta^*)$$

for all $t \in [T]$. By Lemma G.13, it holds with probability at least $1-\delta$ that $\theta^* \in \mathcal{C}^t$ for all $t \in [T]$. Thus, by the update of π^t in Algorithm 2, it holds with probability at least $1-\delta$ that

$$V^*(\theta^*) - V^{\pi^t}(\theta^*) \le V^{\pi^t}(\theta^t) - V^{\pi^t}(\theta^*). \tag{G.14}$$

It now suffices to upper bound the performance difference on the right-hand side of (G.14). By Lemma G.8, we obtain

$$\begin{aligned}
&|V^{\pi^t}(\theta^t) - V^{\pi}(\theta^*)| \qquad\qquad (G.15)\\
&\le \nu \cdot \sqrt{S} \cdot H \cdot \underbrace{\sum_{h=2}^{H-1} \sum_{o_1^{H-1} \in \mathcal{O}^h} \big\| \big(\mathbb{B}_h^{\theta^t}(a_h, o_h) - \mathbb{B}_h^{\theta^*}(a_h, o_h)\big) \mathbb{B}_{h-1}^{\theta^*}(a_{h-1}, o_{h-1}) \ldots \mathbb{B}_1^{\theta^*}(a_1, o_1) b_1^{\theta^*} \big\|_1}_{\text{(i)}} \\
&\quad + \underbrace{\nu \cdot \sqrt{S} \cdot H \cdot \sum_{a_1 \in \mathcal{A}} \sum_{o_1 \in \mathcal{O}} \big\| \big(\mathbb{B}_1^{\theta^t}(a_1, o_1) - \mathbb{B}_1^{\theta^*}(a_1, o_1)\big) b_1^{\theta^*} \big\|_1}_{\text{(ii)}} + \underbrace{\nu \cdot \sqrt{S} \cdot H \cdot \|b_1^{\theta^t} - b_1^{\theta^*}\|_1}_{\text{(iii)}}.
\end{aligned}$$

In the sequel, we upper bound terms (i), (ii), and (iii) on the right-hand side of (G.15). By Lemma G.13, it holds for all $t \in [T]$ with probability at least $1-\delta$ that

$$\begin{aligned}
&\text{(ii)} = \big\| \big(\mathbb{B}_1^{\theta^t}(a_1, o_1) - \mathbb{B}_1^{\theta^*}(a_1, o_1)\big) b_1 \big\|_1 = \mathcal{O}\Big(\nu \cdot (k+\ell) \cdot \sqrt{A^{5k+1} \cdot O^{k+\ell} \cdot \log(O \cdot A \cdot T \cdot H/\delta)/t}\Big), \\
&\text{(iii)} = \|b_1 - \widehat{b}_1^{\theta^t}\|_1 = \mathcal{O}\Big(\nu \cdot (k+\ell) \cdot \sqrt{A^{5k+1} \cdot O^{k+\ell} \cdot \log(O \cdot A \cdot T \cdot H/\delta)/t}\Big). \qquad\qquad (G.16)
\end{aligned}$$

It remains to upper bound term (i) on the right-hand side of (G.15). By Lemma G.9, we obtain that

$$\begin{aligned}
\text{(i)} &= \sum_{o_1^{H-1} \in \mathcal{O}^h} \big\| \big(\mathbb{B}_h^{\theta^t}(a_h, o_h) - \mathbb{B}_h^{\theta^*}(a_h, o_h)\big) \mathbb{B}_{h-1}^{\theta^*}(a_{h-1}, o_{h-1}) \ldots \mathbb{B}_1^{\theta^*}(a_1, o_1) b_1^{\theta'} \big\|_1 \qquad\qquad (G.17)\\
&\le \sum_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \sum_{o_h \in \mathcal{O}} \sum_{s_{h-1} \in \mathcal{S}} \big\| \big(\mathbb{B}_h^{\theta^t}(a_h, o_h) - \mathbb{B}_h^{\theta^*}(a_h, o_h)\big) \mathbb{U}_{h,k}^{\theta^*} \mathbb{T}_{h-1}^{\theta^*}(s_{h-1}, a_{h-1}) \big\|_1 \cdot \mathbb{P}^{\pi^t}(s_{h-1} \,|\, a_{h-\ell}^{h-1})
\end{aligned}$$

Meanwhile, by Lemma G.13, it holds for all $h \in [T]$ and $t \in [T]$ with probability at least $1-\delta$ that

$$\begin{aligned}
&\sum_{s_{h-1} \in \mathcal{S}} \big\| \big(\mathbb{B}_h^{\theta^t}(a_h, o_h) - \mathbb{B}_h^{\theta^*}(a_h, o_h)\big) \mathbb{U}_{h,k}^{\theta^*} \mathbb{T}_{h-1}^{\theta^*}(s_{h-1}, a_{h-1}) \big\|_1 \cdot \mathbb{P}^t(s_{h-1} \,|\, a_{h-\ell}^{h-1}) \qquad\qquad (G.18)\\
&\quad = \mathcal{O}\Big(\gamma \cdot \nu \cdot (k+\ell) \cdot \sqrt{A^{5k+\ell} \cdot O^{k+1} \cdot \log(O \cdot A \cdot T \cdot H/\delta)/t}\Big),
\end{aligned}$$

where we define

$$\mathbb{P}^t = \frac{1}{t} \cdot \sum_{\omega=0}^{t-1} \mathbb{P}^{\pi^\omega} \tag{G.19}$$

for $t > 1$. We remark that the upper bound in (G.18) does not match the right-hand side of (G.15). The only difference is the probaility density of s_{h-1}, which is $\mathbb{P}^{\pi^t}$ on the right-hand side of (G.15) but $\mathbb{P}^t$ defined in (G.19) on the left-hand side of (G.18), respectively. To this end, we utilize the same calculation trick as §E.3 and adopt Lemma E.5. By the upper bound in (G.18) and Lemma E.5 with

$$\begin{aligned}
z_t &= \big\| \big(\mathbb{B}_h^{\theta^t}(a_h, o_h) - \mathbb{B}_h^{\theta^*}(a_h, o_h)\big) \mathbb{U}_{h,k}^{\theta^*} \mathbb{T}_{h-1}^{\theta^*}(s_{h-1}, a_{h-1}) \big\|_1, \\
w_t &= \mathbb{P}^{\pi^t}(s_{h-1} \,|\, a_{h-\ell}^{h-1}),
\end{aligned}$$

we obtain for all $s_{h-1} \in \mathcal{S}$ that

$$\begin{aligned}
&\frac{1}{T} \cdot \sum_{t=1}^{T} \big\| \big(\mathbb{B}_h^{\theta^t}(a_h, o_h) - \mathbb{B}_h^{\theta^*}(a_h, o_h)\big) \mathbb{U}_{h,k}^{\theta^*} \mathbb{T}_{h-1}^{\theta^*}(s_{h-1}, a_{h-1}) \big\|_1 \cdot \mathbb{P}^{\pi^t}(s_{h-1} \,|\, a_{h-\ell}^{h-1}) \\
&\quad = \mathcal{O}\Big(\gamma \cdot \nu^2 \cdot (k+\ell) \cdot \sqrt{A^{5k+\ell} \cdot O^{k+1} \cdot \log T \cdot \log(O \cdot A \cdot T \cdot H/\delta)/T}\Big). \qquad\qquad (G.20)
\end{aligned}$$

Thus,combining (G.14), (G.16), (G.17), and (G.20), we obtain

$$\frac{1}{T}\cdot\sum_{t=1}^{T}|V^{\pi^t}(\theta^t)-V^{\pi}(\theta^*)|$$
$$=\mathcal{O}\Big(\text{poly}(S,A,O,H)\cdot\gamma\cdot\nu^2\cdot(k+\ell)\cdot\sqrt{\log T\cdot\log(O\cdot A\cdot T\cdot H/\delta)/T}\Big),$$

with probability at least $1-\delta$, where we define

$$\text{poly}(S,A,O,H)=H\cdot\sqrt{S^3\cdot A^{5k+3\ell}\cdot O^{k+3}}.$$

By (G.13), it further holds with probability at least $1-\delta$ that

$$V^*(\theta^*)-V^{\overline{\pi}^T}(\theta^*)=\mathcal{O}\Big(\text{poly}(S,A,O,H)\cdot\gamma\cdot\nu^2\cdot(k+\ell)\cdot\sqrt{\log T\cdot\log(O\cdot A\cdot T\cdot H/\delta)/T}\Big).$$

Hence, by setting

$$T=\mathcal{O}\Big(\text{poly}(S,A,O,H)\cdot\gamma^2\cdot\nu^4\cdot(k+\ell)^2\cdot\log(O\cdot A\cdot H/\delta)/\epsilon^2\Big),$$

it holds with probability at least $1-\delta$ that $V^*(\theta^*)-V^{\overline{\pi}^T}(\theta^*)\le\epsilon$, which completes the proof of Theorem G.14. □

H Proof of Tabular POMDP

In this section, we present the proof of the auxiliary results in §G.

H.1 Proof of Lemma G.2

Proof. Recall that we define for all $h\in[H]$ the following operators,

$$\widetilde{\mathbb{O}}_h(o_h)=\mathbb{D}\big(\mathbb{O}_h(o_h\,|\,\cdot)\big)=\mathbb{D}\Big(\big[\mathbb{O}(o_h\,|\,s_h)\big]_{s_h}\Big)\in\mathbb{R}^{S\times S},$$
$$\mathbb{T}_h(a_h)=\mathbb{P}_h(\cdot\,|\,\cdot,a_h)=\big[\mathbb{P}_h(s_{h+1}\,|\,s_h,a_h)\big]_{s_h,s_{h+1}}\in\mathbb{R}^{S\times S},$$
$$\mathbb{O}_h=\mathbb{O}_h(\cdot\,|\,\cdot)=\big[\mathbb{O}_h(o_h\,|\,s_h)\big]_{o_h,s_h}\in\mathbb{R}^{O\times S},$$

where we denote by $\mathbb{D}(v)\in\mathbb{R}^{S\times S}$ the diagonal matrix where the diagonal entries aligns with the vector $v\in\mathbb{R}^S$. Thus, it holds that

$$\widetilde{\mathbb{O}}_h(o_{h+1})\,\mathbb{1}(s_h)=\mathbb{O}_h(o_h\,|\,s_h)\cdot\mathbb{1}(s_h)\in\mathbb{R}^S.$$

By further calculation, we have

$$\begin{aligned}\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\,\mathbb{1}(s_h)&=\big[\mathbb{P}_h(s_{h+1}\,|\,a_h,s_h)\cdot\mathbb{O}_h(o_h\,|\,s_h)\big]_{s_{h+1}}\\&=\big[\mathbb{P}(s_{h+1},o_h\,|\,a_h,s_h)\big]_{s_{h+1}}\in\mathbb{R}^S,\end{aligned}$$

where the second equality holds since we have $O_h\perp\!\!\!\perp S_{h+1}\,|\,s_h,a_h$. It then holds that

$$\begin{aligned}\widetilde{\mathbb{O}}_{h+1}(o_{h+1})\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\,\mathbb{1}(s_h)&=\big[\mathbb{O}_{h+1}(o_{h+1}\,|\,s_{h+1})\cdot\mathbb{P}(s_{h+1},o_h\,|\,a_h,s_h)\big]_{s_{h+1}}\\&=\big[\mathbb{P}(s_{h+1},o_{h+1},o_h\,|\,a_h,s_h)\big]_{s_{h+1}}\in\mathbb{R}^S,\end{aligned}\tag{H.1}$$

where the second equality holds since the observation O_{h+1} is independent of all the other random variables in (H.1) given the state $S_{h+1}=s_{h+1}$. By further multiplying the right-hand side of (H.1) by $\mathbb{T}_{h+1}(a_{h+1})$, we obtain that

$$\begin{aligned}&\mathbb{T}_{h+1}(a_{h+1})\widetilde{\mathbb{O}}_{h+1}(o_{h+1})\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\,\mathbb{1}(s_h)\\&=\bigg[\sum_{s_{h+1}\in\mathcal{S}}\mathbb{P}_{h+1}(s_{h+2}\,|\,s_{h+1},a_{h+1})\cdot\mathbb{P}(s_{h+1},o_{h+1},o_h\,|\,a_h,s_h)\bigg]_{s_{h+2}}\\&=\bigg[\sum_{s_{h+1}\in\mathcal{S}}\mathbb{P}(s_{h+2},s_{h+1},o_{h+1},o_h\,|\,a_{h+1},a_h,s_h)\bigg]_{s_{h+2}}\\&=\big[\mathbb{P}(s_{h+2},o_{h+1},o_h\,|\,a_{h+1},a_h,s_h)\big]_{s_{h+2}}\in\mathbb{R}^S,\end{aligned}\tag{H.2}$$

where the second equality holds since the state S_{h+2} is independent of all the other random variables in (H.2) given the previous state $S_{h+1}=s_{h+1}$ and action $A_{h+1}=a_{h+1}$. By an iterative calculation

similar to (H.1) and (H.2), we obtain that

$$\begin{aligned}&\mathbb{T}_{h+k-1}(a_{h+k-1})\widetilde{\mathbb{O}}_{h+k-1}(o_{h+k-1})\cdots\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\,\mathbb{1}(s_h)\\&\quad=\big[\mathbb{P}(s_{h+k},o_{h+k-1},\ldots,o_h\,|\,a_{h+k-1},\ldots,a_h,s_h)\big]_{s_{h+k}}\in\mathbb{R}^S.\end{aligned}$$

By further calculation, we obtain that

$$\begin{aligned}&\mathbb{O}_{h+k}\mathbb{T}_{h+k-1}(a_{h+k-1})\widetilde{\mathbb{O}}_{h+k-1}(o_{h+k-1})\cdots\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\,\mathbb{1}(s_h)\\&\quad=\big[\mathbb{P}(o_{h+k},o_{h+k-1},\ldots,o_h\,|\,a_{h+k-1},\ldots,a_h,s_h)\big]_{o_{h+k}}\in\mathbb{R}^O.\end{aligned}\tag{H.3}$$

Finally, by multiplying the right-hand side of (H.3) with the indicator vector $\mathbb{1}(o_{h+k})$, we conclude that

$$\begin{aligned}\mathbb{U}_h&=\big[\mathbb{1}(o_{h+k})^\top\mathbb{O}_{h+k}\mathbb{T}_{h+k-1}(a_{h+k-1})\widetilde{\mathbb{O}}_{h+k-1}(o_{h+k-1})\\&\qquad\cdots\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\,\mathbb{1}(s_h)\big]_{(o_h^{h+k},a_h^{h+k-1}),s_h}\\&=\big[\mathbb{P}(o_{h+k},o_{h+k-1},\ldots,o_h\,|\,a_{h+k-1},\ldots,a_h,s_h)\big]_{(o_h^{h+k},a_h^{h+k-1}),s_h}\\&=\big[\mathbb{P}(\tau_h^{h+k}\,|\,s_h)\big]_{\tau_h^{h+k},s_h}\in\mathbb{R}^{(A^k\cdot O^{k+1})\times S},\end{aligned}$$

which completes the proof of Lemma G.2. □

H.2 Proof of Lemma G.5

Proof. The proof is similar to that of Lemma G.2. By the definition of Bellman operators in Definition G.4, it holds that

$$\begin{aligned}\mathbb{1}(o_H)^\top\mathbb{B}_{H-1}(a_{H-1},o_{H-1})\ldots\mathbb{B}_1(a_1,o_1)\mathbb{U}_1&=\Big(\Pi_{h=1}^{H-1}\mathbb{U}_{h+1}\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\mathbb{U}_h^\dagger\Big)\mathbb{U}_1\\&=\mathbb{1}(o_H)^\top\mathbb{U}_{H-1}\Pi_{h=1}^{H-1}\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h),\end{aligned}\tag{H.4}$$

where μ_1 is the probability array of initial state distribution. Following the same computation as the proof of Lemma G.2 in §H.1, we obtain that

$$\mathbb{U}_{H-1}\Pi_{h=1}^{H-1}\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\mu_1=\big[\mathbb{P}(o_1^{H+k}\,|\,a_1^{H+k-1})\big]_{(o_1^{H+k},a_1^{H+k-1})}.\tag{H.5}$$

Thus, multiplying the right-hand side of (H.5) by the indicator $\mathbb{1}(o_H)$ that takes value 1 for all indices that contain o_H and a fixed action sequence a_h^{h+k-1}, we obtain that

$$\begin{aligned}\mathbb{1}(o_H)^\top\mathbb{U}_{H-1}\Pi_{h=1}^{H-1}\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\mu_1&=\mathbb{1}(o_H)^\top\big[\mathbb{P}(o_1^{H+k}\,|\,a_1^{H+k-1})\big]_{(o_1^{H+k},a_1^{H+k-1})}\\&=\mathbb{P}(o_1^H\,|\,a_1^{H-1}),\end{aligned}\tag{H.6}$$

which is as desired. Thus, combining (H.4) and (H.6), we complete the proof of Lemma G.5. □

H.3 Proof of Lemma G.6

Proof. The proof is similar to that of Lemma G.2. Recall that we define for all $h\in[H]$ the following operators,

$$\begin{aligned}\widetilde{\mathbb{O}}_h(o_h)&=\mathbb{D}\big(\mathbb{O}_h(o_h\,|\,\cdot)\big)=\mathbb{D}\Big(\big[\mathbb{O}(o_h\,|\,s_h)\big]_{s_h}\Big)\in\mathbb{R}^{S\times S},\\\mathbb{T}_h(a_h)&=\mathbb{P}_h(s_{h+1}=\cdot\,|\,\cdot,a_h)=\big[\mathbb{P}_h(s_{h+1}\,|\,s_h,a_h)\big]_{s_h,s_{h+1}}\in\mathbb{R}^{S\times S},\\\mathbb{O}_h&=\mathbb{O}_h(\cdot\,|\,\cdot)=\big[\mathbb{O}_h(o_h\,|\,s_h)\big]_{o_h,s_h}\in\mathbb{R}^{O\times S}.\end{aligned}$$

Recall that we define

$$\begin{aligned}\mathbb{X}_h(a_{h-\ell}^{h-1})&=\mathbb{U}_h\mathbb{T}_{h-1}(a_{h-1})\widetilde{\mathbb{O}}_{h-1}(\cdot)\ldots\mathbb{T}_{h-\ell}(a_{h-\ell})\widetilde{\mathbb{O}}_{h-\ell}(\cdot)\mu_{h-\ell},\\\mathbb{Y}_h(a_{h-\ell}^h,o_h)&=\mathbb{U}_{h+1}\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\mathbb{T}_{h-1}(a_{h-1})\widetilde{\mathbb{O}}_{h-1}(\cdot)\ldots\mathbb{T}_{h-\ell}(a_{h-1})\widetilde{\mathbb{O}}_{h-\ell}(\cdot)\mu_{h-\ell}.\end{aligned}$$

We first show that the following equation holds,

$$\begin{aligned}&\mathbb{T}_{h-1}(a_{h-1})\widetilde{\mathbb{O}}_{h-1}(\cdot)\ldots\mathbb{T}_{h-\ell}(a_{h-\ell})\widetilde{\mathbb{O}}_{h-\ell}(\cdot)\mu_{h-\ell}\\&\quad=\mathbb{P}(s_h=\cdot,o_{h-1}=\cdot,\ldots,o_{h-\ell}=\cdot\,|\,a_{h-1},\ldots,a_{h-\ell}).\end{aligned}\tag{H.7}$$

To see such a fact, note that for all $o_{h-\ell}\in\mathcal{O}$, we have

$$\widetilde{\mathbb{O}}_{h-\ell}(o_{h-\ell})\mu_{h-\ell}=\big[\mathbb{O}_{h-\ell}(o_{h-\ell}\,|\,s_{h-\ell})\cdot\mathbb{P}(s_{h-\ell})\big]_{s_{h-\ell}}\in\mathbb{R}^S.\tag{H.8}$$

It thus holds that

$$\mathbb{T}_{h-\ell}(a_{h-\ell})\widetilde{\mathbb{O}}_{h-\ell}(o_{h-\ell})\mu_{h-\ell} = \Big[\sum_{s_{h-\ell}\in\mathcal{S}} \mathbb{P}(s_{h-\ell+1}\,|\,s_{h-\ell},a_{h-\ell})\cdot\mathbb{O}_{h-\ell}(o_{h-\ell}\,|\,s_{h-\ell})\cdot\mathbb{P}(s_{h-\ell})\Big]_{s_{h-\ell+1}}$$
$$= \big[\mathbb{P}(s_{h-\ell+1},o_{h-\ell}\,|\,a_{h-\ell})\big]_{s_{h-\ell+1}}, \tag{H.9}$$

where the second equality follows from the fact that $O_{h-\ell} \perp\!\!\!\perp S_{h-\ell+1}\,|\,s_{h-\ell}$. Thus, by recursive computation similar to (H.8) and (H.9), we obtain (H.7), namely,

$$\begin{aligned}&\mathbb{T}_{h-1}(a_{h-1})\widetilde{\mathbb{O}}_{h-1}(\cdot)\ldots\mathbb{T}_{h-\ell}(a_{h-\ell})\widetilde{\mathbb{O}}_{h-\ell}(\cdot)\mu_{h-\ell}\\ &\quad= \mathbb{P}(s_h=\cdot,o_{h-1}=\cdot,\ldots,o_{h-\ell}=\cdot\,|\,a_{h-1},\ldots,a_{h-\ell})\\ &\quad= \big[\mathbb{P}(\tau_{h-\ell}^{h-1}\,|\,s_h)\big]_{s_h,o_{h-\ell}^{h-1}}\in\mathbb{R}^{S\times O^\ell}\end{aligned} \tag{H.10}$$

Meanwhile, by Lemma G.2, we have

$$\mathbb{U}_h = \big[\mathbb{P}(\tau_h^{h+k}\,|\,s_h)\big]_{\tau_h^{h+k},s_h}\in\mathbb{R}^{(O^{k+1}\cdot A^k)\times S}. \tag{H.11}$$

Thus, it holds for all $h\in[H]$ that

$$\begin{aligned}\mathbb{X}_h(a_{h-\ell}^{h-1}) &= \mathbb{U}_h\mathbb{T}_{h-1}(a_{h-1})\widetilde{\mathbb{O}}_{h-1}(\cdot)\ldots\mathbb{T}_{h-\ell}(a_{h-\ell})\widetilde{\mathbb{O}}_{h-\ell}(\cdot)\mu_{h-\ell}\\ &= \Big[\sum_{s_h\in\mathcal{S}}\mathbb{P}(\tau_h^{h+k}\,|\,s_h)\cdot\mathbb{P}(s_h,o_{h-\ell}^{h-1}\,|\,a_{h-\ell}^{h-1})\Big]_{(o_h^{h+k},a_h^{h+k-1}),o_{h-\ell}^{h-1}}\\ &= \big[\mathbb{P}(\tau_{h-\ell}^{h+k})\big]_{\tau_h^{h+k},o_{h-\ell}^{h-1}}\in\mathbb{R}^{(O^{k+1}\cdot A^k)\times O^\ell},\end{aligned}$$

where the second equality follows from the fact that $o_h^{h+k} \perp\!\!\!\perp o_{h-\ell}^{h-1}\,|\,s_h$. The computation of $\mathbb{X}_h(a_{h-\ell}^{h-1})$ is identical to that of $\mathbb{X}_h(a_{h-\ell}^{h-1})$. In conclusion, we have

$$\begin{aligned}\mathbb{Y}_h(a_{h-\ell}^h,o_h) &= \mathbb{U}_{h+1}\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)\mathbb{T}_{h-1}(a_{h-1})\widetilde{\mathbb{O}}_{h-1}(\cdot)\ldots\mathbb{T}_{h-\ell}(a_{h-1})\widetilde{\mathbb{O}}_{h-\ell}(\cdot)\mu_{h-\ell}.\\ &= \big[\mathbb{P}(\tau_h^{h+k+1})\big]_{\tau_{h+1}^{h+k+1},o_{h-\ell}^{h-1}}\in\mathbb{R}^{(O^{k+1}\cdot A^k)\times O^\ell},\end{aligned}$$

which completes the proof of Lemma G.6.

□

H.4 Proof of Lemma G.8

Proof. By Lemma G.5 and (G.1), we have

$$\begin{aligned}V^\pi(\theta) &= \sum_{o_1^{H-1}\in\mathcal{O}^H} r(o_1^{H-1})\cdot\mathbb{P}^\pi(o_1^{H-1})\\ &= \sum_{o_1^{H-1}\in\mathcal{O}^H} r(o_1^{H-1})\cdot\mathbb{1}(o_H)^\top\mathbb{B}_{H-1}^\theta(a_{H-1}^\pi,o_{H-1})\ldots\mathbb{B}_1^\theta(a_1^\pi,o_1)b_1^\theta.\end{aligned} \tag{H.12}$$

Here the actions a_h^π are taken based on the policy π and the past observations and actions taken $(o_1,a_1^\pi,\ldots,a_{h-1}^\pi,o_h)$. In addition, recall that we define $b_1^\theta=\mathbb{U}_1^\theta\mu_1$, where μ_1 is the probability array of initial state distribution. It thus follows from (H.12) that

$$\begin{aligned}V^\pi(\theta)-V^\pi(\theta') = \sum_{o_1^{H-1}\in\mathcal{O}^H}\sum_{h=1}^{H-1} &r(o_1^{H-1})\cdot\mathbb{1}(o_H)^\top\mathbb{B}_{H-1}^\theta(a_{H-1}^\pi,o_{H-1})\cdots\\ &\cdots\big(\mathbb{B}_h^\theta(a_h^\pi,o_h)-\mathbb{B}_h^{\theta'}(a_h^\pi,o_h)\big)\mathbb{B}_{h-1}^{\theta'}(a_{h-1}^\pi,o_{h-1})\cdots\mathbb{B}_1^{\theta'}(a_1^\pi,o_1)b_1^{\theta'}\\ +\sum_{o_1^{H-1}\in\mathcal{O}^H} &r(o_1^{H-1})\cdot\mathbb{1}(o_H)^\top\mathbb{B}_{H-1}^\theta(a_{H-1}^\pi,o_{H-1})\ldots\mathbb{B}_1^\theta(a_1^\pi,o_1)(b_1^\theta-b_1^{\theta'}).\end{aligned} \tag{H.13}$$

In the sequel, we upper bound the absolute value of the right-hand side of (H.13). We define the following vectors for all $h\in[H-1]$ for notational simplicity,

$$\begin{aligned}v_h &= \big(\mathbb{B}_h^\theta(a_h^\pi,o_h)-\mathbb{B}_h^{\theta'}(a_h^\pi,o_h)\big)\mathbb{B}_{h-1}^{\theta'}(a_{h-1}^\pi,o_{h-1})\ldots\mathbb{B}_1^{\theta'}(a_1^\pi,o_1)b_1^{\theta'},\\ v_0 &= b_1^\theta-b_1^{\theta'}.\end{aligned}$$

Since $0 \le r(o_1^{H-1}) \le H$ for all observation sequences $o_1^{H-1} \in \mathcal{O}^H$, we obtain that

$$\Big| \sum_{o_1^{H-1} \in \mathcal{O}^H} \sum_{h=1}^{H-1} r(o_1^{H-1}) \cdot \mathbb{1}(o_H)^\top \mathbb{B}_{H-1}^\theta(a_{H-1}^\pi, o_{H-1}) \dots \mathbb{B}_{h+1}^\theta(a_{h+1}^\pi, o_{h+1}) v_h \Big|$$

$$\le H \cdot \Big| \sum_{o_1^{H-1} \in \mathcal{O}^H} \sum_{h=1}^{H-1} \mathbb{1}(o_H)^\top \mathbb{B}_{H-1}^\theta(a_{H-1}^\pi, o_{H-1}) \dots \mathbb{B}_{h+1}^\theta(a_{h+1}^\pi, o_{h+1}) v_h \Big|.$$

Moreover, since the vector $\mathbb{1}(o_H)$ takes value in $\{0, 1\}$ for all the indices, we further obtain that

$$H \cdot \Big| \sum_{o_1^{H-1} \in \mathcal{O}^H} \sum_{h=1}^{H-1} \mathbb{1}(o_H)^\top \mathbb{B}_{H-1}^\theta(a_{H-1}, o_{H-1}) \dots \mathbb{B}_{h+1}^\theta(a_{h+1}^\pi, o_{h+1}) v_h \Big|$$

$$\le H \cdot \sum_{o_1^{H-1} \in \mathcal{O}^H} \sum_{h=1}^{H-1} \|\mathbb{B}_{H-1}^\theta(a_{H-1}^\pi, o_{H-1}) \dots \mathbb{B}_{h+1}^\theta(a_{h+1}^\pi, o_{h+1}) v_h\|_1 \tag{H.14}$$

It now suffices to upper bound the right-hand side of (H.14). By the definition of Bellman operators in Definition G.4, we obtain for all $h \in \{0, \dots, H-1\}$ that

$$\|\mathbb{B}_{H-1}^\theta(a_{H-1}^\pi, o_{H-1}) \dots \mathbb{B}_{h+1}^\theta(a_{h+1}^\pi, o_{h+1}) v_h\|_1$$

$$= \|\mathbb{U}_h \mathbb{T}_{H-1}(a_{H-1}^\pi) \widetilde{\mathbb{O}}_{H-1}(o_{H-1}) \dots \mathbb{T}_h(a_h^\pi) \widetilde{\mathbb{O}}_h(o_h) \mathbb{U}_h^\dagger v_h\|_1 \tag{H.15}$$

The following lemma upper bound the right-hand side of (H.15).

Lemma H.1. It holds for all $a_h^{H-1} \in \mathcal{A}^{H-h}$, $h \in [H]$, and $u \in \mathbb{R}^S$ that

$$\sum_{o_h^{H-1} \in \mathcal{O}^{H-h}} \|\mathbb{U}_h \mathbb{T}_{H-1}(a_{H-1}) \widetilde{\mathbb{O}}_{H-1}(o_{H-1}) \dots \mathbb{T}_h(a_h) \widetilde{\mathbb{O}}_h(o_h) u\|_1 \le \|u\|_1.$$

Proof. See §H.8 for a detailed proof. □

Meanwhile, by Assumption G.3 and the fact that $\|A\|_{1\mapsto 1} \le \sqrt{S}\|A\|_2$ for any matrix $A \in \mathbb{R}^{S \times N}$, we obtain that

$$\|\mathbb{U}_h^\dagger v_h\|_1 \le \|\mathbb{U}_h^\dagger\|_{1\mapsto 1} \cdot \|v_h\|_1 \le \sqrt{S} \cdot \|\mathbb{U}_h^\dagger\|_2 \cdot \|v_h\|_1 \le \nu \cdot \sqrt{S} \cdot \|v_h\|_1 \tag{H.16}$$

for all $h \in [H]$. Combining Lemma H.1, (H.14), (H.15), and (H.16), we obtain that

$$|V^\pi(\theta) - V^\pi(\theta')| \le \nu \cdot \sqrt{S} \cdot H \cdot \sum_{h=0}^{H-1} \sum_{o_1^h \in \mathcal{O}^h} \|v_h\|_1$$

$$= \nu \cdot \sqrt{S} \cdot H \cdot \sum_{h=1}^{H-1} \sum_{o_1^h \in \mathcal{O}^h} \big\| \big(\mathbb{B}_h^\theta(a_h, o_h) - \mathbb{B}_h^{\theta'}(a_h, o_h)\big) \mathbb{B}_{h-1}^{\theta'}(a_{h-1}, o_{h-1}) \dots \mathbb{B}_1^{\theta'}(a_1, o_1) b_1^{\theta'} \big\|_1$$

$$+ \nu \cdot \sqrt{S} \cdot H \cdot \big\| \big(\mathbb{B}_1^\theta(a_1, o_1) - \mathbb{B}_1^{\theta'}(a_1, o_1)\big) b_1^{\theta'} \big\|_1 + \nu \cdot \sqrt{S} \cdot H \cdot \|b_1^\theta - b_1^{\theta'}\|_1,$$

which completes the proof of Lemma G.8. □

H.5 Proof of Lemma G.9

Proof. It holds for all policy π and the corresponding action sequence $a_{1:h}^\pi$ generated by π that

$$\sum_{o_1^h \in \mathcal{O}^h} \big\| \big(\mathbb{B}_h^\theta(a_h^\pi, o_h) - \mathbb{B}_h^{\theta'}(a_h^\pi, o_h)\big) \mathbb{B}_{h-1}^{\theta'}(a_{h-1}^\pi, o_{h-1}) \dots \mathbb{B}_1^{\theta'}(a_1^\pi, o_1) b_1^{\theta'} \big\|_1 \tag{H.17}$$

$$\le \sum_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \sum_{o_1^h \in \mathcal{O}^h} \big\| \big(\mathbb{B}_h^\theta(a_h, o_h) - \mathbb{B}_h^{\theta'}(a_h, o_h)\big) \mathbb{B}_{h-1}^{\theta'}(a_{h-1}, o_{h-1}) \dots \mathbb{B}_1^{\theta'}(a_1, o_1) b_1^{\theta'} \big\|_1$$

$$= \sum_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \sum_{o_1^h \in \mathcal{O}^h} \big\| \big(\mathbb{B}_h^\theta(a_h, o_h) - \mathbb{B}_h^{\theta'}(a_h, o_h)\big) \mathbb{U}_{h,k}^{\theta'} \mathbb{T}_{h-1}^{\theta'}(a_{h-1}) \mu_{h-1}^{\theta'}(\pi, a_{h-\ell}^{h-1}, o_1^{h-1}) \big\|_1,$$

where $\mu_{h-1}^{\theta'}(\pi, a_{h-\ell}^{h-1}, o_1^{h-1})$ is the state distribution array defined in (G.7). Note that on the right-hand side of (H.17), the state distribution array $\mu_{h-1}^{\theta'}(\pi, a_{h-\ell}^{h-1}, o_1^{h-1})$ is the only term that is related to policy π. In what follows, we define

$$M_h(\theta, \theta', a_h, a_{h-1}, o_h) = \big(\mathbb{B}_h^{\theta}(a_h, o_h) - \mathbb{B}_h^{\theta'}(a_h, o_h)\big)\mathbb{U}_{h,k}^{\theta'}\mathbb{T}_{h-1}^{\theta'}(a_{h-1}) \in \mathbb{R}^{(A^{k+1}\cdot O^k)\times S} \tag{H.18}$$

for notational simplicity. It holds that

$$\begin{aligned}
&\sum_{o_1^h \in \mathcal{O}^h} \|M_h(\theta, \theta', a_h, a_{h-1}, o_h)\mu_{h-1}^{\theta'}(\pi, a_{h-\ell}^{h-1}, o_1^{h-1})\|_1 \\
&\quad \le \sum_{o_1^h \in \mathcal{O}^h} \sum_{s_{h-1} \in \mathcal{S}} \big\| \big[M_h(\theta, \theta', a_h, a_{h-1}, o_h)\big]_{s_{h-1}} \|_1 \cdot \big[\mu_{h-1}^{\theta'}(\pi, a_{h-\ell}^{h-1}, o_1^{h-1})\big]_{s_{h-1}},
\end{aligned} \tag{H.19}$$

where we denote by $[M_h(\theta, \theta', a_h, a_{h-1}, o_h)]_{s_{h-1}}$ and $[\mu_{h-1}^{\theta'}(\pi, a_{h-\ell}^{h-1}, o_1^{h-1})]_{s_{h-1}}$ the s_{h-1}-th column of $M_h(\theta, \theta', a_h, a_{h-1}, o_h)$ and the s_{h-1}-th entry of $\mu_{h-1}^{\theta'}(\pi, a_{h-\ell}^{h-1}, o_1^{h-1})$, respectively. Recall that we have

$$\big[\mu_{h-1}^{\theta'}(\pi, a_{h-\ell}^{h-1}, o_1^{h-1})\big]_{s_{h-1}} = \mathbb{P}^{\theta}(s_h, o_1^{h-1} \,|\, a_{h-\ell}^{h-1}, \pi).$$

Thus, by marginalizing over the observation sequence o_1^{h-1}, it holds that

$$\sum_{o_1^{h-1} \in \mathcal{O}^{h-1}} \big[\mu_{h-1}^{\theta'}(\pi, a_{h-\ell}^{h-1}, o_1^{h-1})\big]_{s_{h-1}} = \mathbb{P}^{\theta'}(s_h \,|\, a_{h-\ell}^{h-1}, \pi). \tag{H.20}$$

Plugging (H.20) into (H.19), we obtain that

$$\begin{aligned}
&\sum_{o_1^h \in \mathcal{O}^h} \|M_h(\theta, \theta', a_h, a_{h-1}, o_h)\mu_{h-1}^{\theta'}(\pi, a_{h-\ell}^{h-1}, o_1^{h-1})\|_1 \\
&\quad \le \sum_{o_h \in \mathcal{O}} \sum_{s_{h-1} \in \mathcal{S}} \big\| \big[M_h(\theta, \theta', a_h, a_{h-1}, o_h)\big]_{s_{h-1}} \|_1 \cdot \mathbb{P}^{\theta'}(s_h \,|\, a_{h-\ell}^{h-1}, \pi) \\
&\quad = \sum_{a_{h-\ell}^h \in \mathcal{A}^{\ell+1}} \sum_{o_h \in \mathcal{O}} \big\| \big(\mathbb{B}_h^{\theta}(a_h, o_h) - \mathbb{B}_h^{\theta'}(a_h, o_h)\big)\mathbb{U}_{h,k}^{\theta'}\mathbb{T}_{h-1}^{\theta'}(a_{h-1})\mathbb{D}\big(\widetilde{\mu}_{h-1}^{\theta'}(\pi, a_{h-\ell}^{h-1})\big)\big\|_1.
\end{aligned}$$

Thus, we complete the proof of Lemma G.9. □

H.6 Proof of Lemma G.10

Proof. Recall that we aim to recover the following density matrices in the estimation of Bellman operators in §G.1,

$$\begin{aligned}
b_1 &= \big[\mathbb{P}(\tau_1^k)\big]_{\tau_1^k} \in \mathbb{R}^{\mathcal{A}^k\cdot\mathcal{O}^{k+1}}, \\
\mathbb{X}_h^t(a_{h-\ell}^{h-1}) &= \big[\mathbb{P}^t(\tau_{h-\ell}^{h+k})\big]_{\tau_h^{h+k}, o_{h-\ell}^{h-1}} \in \mathbb{R}^{(\mathcal{A}^k\cdot\mathcal{O}^{k+1})\times\mathcal{O}^\ell}, &\text{(H.21)} \\
\mathbb{Y}_h^t(a_{h-\ell}^{h}, o_h) &= \big[\mathbb{P}^t(\tau_{h-\ell}^{h+k+1})\big]_{\tau_{h+1}^{h+k+1}, o_{h-\ell}^{h-1}} \in \mathbb{R}^{(\mathcal{A}^k\cdot\mathcal{O}^{k+1})\times\mathcal{O}^\ell}. &\text{(H.22)}
\end{aligned}$$

Here we denote by $\mathbb{P}^t$ the visitation measure of the mixed policy $\{\pi^\omega\}_{\omega\in[t]}$ generated by Algorithm 2. Alternatively, we can write the densities in (H.21) in the following vector product form,

$$\mathbb{X}_h^t(a_{h-\ell}^{h-1}) = \sum_{\tau_h^{h+k} \in \mathcal{A}^k\times\mathcal{O}^{k+1}} \bigg(\sum_{o_{h-\ell}^{h-1} \in \mathcal{O}^\ell} \mathbb{1}(\tau_h^{h+k})\, \mathbb{1}(o_{h-\ell}^{h-1})^\top \cdot \mathbb{P}^t(\tau_{h-\ell}^{h+k}) \bigg).$$

Recall that we adopt the following estimator of the probability density defined in (H.21),

$$\widehat{\mathbb{X}}_h^t(a_{h-\ell}^{h-1}) = \frac{1}{t} \cdot \sum_{a_h^{h+k-1} \in \mathcal{A}^k} \bigg(\sum_{\tau_{h-\ell}^{h+k} \in \mathcal{D}^t(a_{h-\ell}^{h+k-1})} \mathbb{1}(\tau_{h-\ell}^{h+k})\, \mathbb{1}(o_{h-\ell}^{h-1})^\top \bigg). \tag{H.23}$$

By the martingale concentration inequality (see e.g., Jin et al. (2019)) and the fact that

$$\| \mathbb{1}(\tau_{h-\ell}^{h+k})\, \mathbb{1}(o_{h-\ell}^{h-1})^\top \|_F \le 1$$

for all trajectory $\tau_{h-\ell}^{h+k} \in \mathcal{A}^{k+\ell} \times \mathcal{O}^{k+\ell+1}$ and observation sequence $o_{h-\ell}^{h-1} \in \mathcal{O}^{\ell}$, we obtain for all $h \in [H]$ and $a_{h-\ell}^{h+k-1} \in \mathcal{A}^{k+\ell}$ that

$$\Bigg\| \sum_{o_{h-\ell}^{h+k} \in \mathcal{O}^{k+\ell+1}} \mathbb{1}(\tau_h^{h+k}) \, \mathbb{1}(o_{h-\ell}^{h-1})^\top \cdot \mathbb{P}^t(\tau_{h-\ell}^{h+k}) - \frac{1}{t-1} \sum_{\tau_{h-\ell}^{h+k} \in \mathcal{D}^t(a_{h-\ell}^{h+k-1})} \mathbb{1}(\tau_{h-\ell}^{h+k}) \, \mathbb{1}(o_{h-\ell}^{h-1})^\top \Bigg\|_F \leq C \cdot (k+\ell) \cdot \sqrt{\log(O \cdot A \cdot T/\delta)/t}$$

with probability at least $1 - \delta/A^k$, where C is a positive absolute constant. It thus holds for all $h \in [H]$, $t \in [T]$, and $a_{h-\ell}^{h-1} \in \mathcal{A}^{\ell}$ that

$$\begin{aligned} \|\mathbb{X}_h^t(a_{h-\ell}^{h-1}) - \widehat{\mathbb{X}}_h^t(a_{h-\ell}^{h-1})\|_F \leq \sum_{a_h^{h+k-1} \in \mathcal{A}^k} \Bigg\| & \sum_{o_{h-\ell}^{h+k} \in \mathcal{O}^{k+\ell+1}} \mathbb{1}(\tau_h^{h+k}) \, \mathbb{1}(o_{h-\ell}^{h-1})^\top \cdot \mathbb{P}^t(\tau_{h-\ell}^{h+k}) \\ & - \frac{1}{t-1} \sum_{\tau_{h-\ell}^{h+k} \in \mathcal{D}^t(a_{h-\ell}^{h+k-1})} \mathbb{1}(\tau_{h-\ell}^{h+k}) \, \mathbb{1}(o_{h-\ell}^{h-1})^\top \Bigg\|_F \\ = \mathcal{O}\big(A^k \cdot (k+\ell) \cdot & \sqrt{\log(O \cdot A \cdot T \cdot H/\delta)/t}\big) \end{aligned}$$

with probability at least $1 - \delta/2$. Meanwhile, recall that we estimate $\mathbb{Y}_h^t$ in (H.22) by the following estimator,

$$\widehat{\mathbb{Y}}_h^t(a_{h-\ell}^h, o_h) = \frac{1}{t-1} \cdot \sum_{a_{h+1}^{h+k} \in \mathcal{A}^k} \sum_{\tau_{h-\ell}^{h+k+1} \in \mathcal{D}^t(a_{h-\ell}^{h+k})} \mathbb{1}(\tau_{h+1}^{h+k+1}) \, \mathbb{1}(o_{h-\ell}^{h-1})^\top, \tag{H.24}$$

Following a similar computation, it holds for all $h \in [H]$, $t \in [T]$, $a_{h-\ell}^{h-1} \in \mathcal{A}^{\ell}$, $a_h \in \mathcal{A}$, and $o_h \in \mathcal{O}$ that

$$\|\mathbb{Y}_h^t(a_{h-\ell}^h, o_h) - \widehat{\mathbb{Y}}_h^t(a_{h-\ell}^h, o_h)\|_F = \mathcal{O}\big(A^k \cdot (k+\ell) \cdot \sqrt{\log(O \cdot A \cdot T/\delta)/t}\big)$$

with probability at least $1 - \delta/2$. Similarly, it also holds for all $h \in [H]$ and $t \in [T]$ that

$$\|b_1 - \widehat{b}_1^t\|_2 = \mathcal{O}(A^k \cdot k \cdot \sqrt{\log(O \cdot A \cdot T/\delta)/t})$$

with probability at least $1 - \delta/2$. Thus, we complete the proof of Lemma G.10. □

H.7 Proof of Lemma G.13

Proof. In what follows, we prove (G.10)–(G.12) separately.

Part I: Proof of Upper Bound in (G.10)**.** By Lemma G.10, it holds with probability at least $1 - \delta$ that

$$\|b_1 - \widehat{b}_1^t\|_1 \leq \sqrt{A^k \cdot O^{k+1}} \cdot \|b_1 - \widehat{b}_1^t\|_2 = \mathcal{O}\Big((k+\ell) \cdot \sqrt{A^{3k} \cdot O^{k+1} \cdot \log(O \cdot A \cdot T \cdot H/\delta)/t}\Big). \tag{H.25}$$

Meanwhile, by the update of $b_1^{\theta^t}$ in Algorithm 2, it holds with probability at least $1 - \delta$ that $b_1^{\theta^t} \in \mathcal{C}^t$, namely,

$$\|b_1^{\theta^t} - \widehat{b}_1^t\|_2 \leq C \cdot \nu \cdot (k+\ell) \cdot \sqrt{A^{5k+1} \cdot O^{k+\ell} \cdot \log(O \cdot A \cdot T \cdot H/\delta)/t}. \tag{H.26}$$

Combining (H.25) and (H.26), it holds with probability at least $1 - \delta$ that

$$\begin{aligned} \|b_1 - b_1^{\theta^t}\|_1 \leq \|b_1 - \widehat{b}_1^t\|_1 + \|b_1^{\theta^t} - \widehat{b}_1^t\|_1 &\leq \sqrt{A^k \cdot O^{k+1}} \cdot \big(\|b_1 - \widehat{b}_1^t\|_2 + \|b_1^{\theta^t} - \widehat{b}_1^t\|_2\big) \\ &= \mathcal{O}\Big(\nu \cdot (k+\ell) \cdot \sqrt{A^{5k+1} \cdot O^{k+\ell} \cdot \log(O \cdot A \cdot T \cdot H/\delta)/t}\Big), \end{aligned}$$

which completes the proof of (G.10).

Part II: Proof of Upper Bound in (G.11)**.** Recall that we define

$$\begin{aligned} b_1 &= \mathbb{U}_1 \mu_1 = \big[\mathbb{P}(\tau_1^{k+1})\big]_{\tau_1^{k+1}} \in \mathbb{R}^{O^{k+1} \cdot A^k}, \\ \mathbb{X}_1^t(a_{1-\ell}^0) &= \big[\mathbb{P}^t(\tau_{1-\ell}^{k+1})\big]_{\tau_1^{k+1}, o_{1-\ell}^0} \in \mathbb{R}^{(O^{k+1} \cdot A^k) \times O^{\ell}}. \end{aligned}$$

Thus, it holds for all action array $a^0_{1-\ell}$ and $t \in [T]$ that

$$[b_1]_{\tau_1^{k+1}} = \mathbb{P}(\tau_1^{k+1}) = \sum_{o^0_{1-\ell} \in \mathcal{O}^\ell} \mathbb{P}^t(\underline{\tau}^0_{1-\ell}) \cdot \mathbb{P}^t(\tau^{k+1}_{1-\ell}). \tag{H.27}$$

It thus holds for all $a^0_{1-\ell} \in \mathcal{A}^\ell$ that

$$\begin{aligned}
&\big\|\big(\mathbb{B}_1^{\theta^t}(a_1, o_1) - \mathbb{B}_1^{\theta^*}(a_1, o_1)\big) b_1\|_1 \\
&\quad = \bigg\|\big(\mathbb{B}_1^{\theta^t}(a_1, o_1) - \mathbb{B}_1^{\theta^*}(a_1, o_1)\big) \sum_{o^0_{1-\ell} \in \mathcal{O}^\ell} \mathbb{P}^t(\underline{\tau}^0_{1-\ell}) \cdot \big[\mathbb{X}_1^t(a^0_{1-\ell})\big]_{o^0_{1-\ell}}\bigg\|_1 \\
&\quad \le \bigg\| \sum_{o^0_{1-\ell} \in \mathcal{O}^\ell} \big(\mathbb{B}_1^{\theta^t}(a_1, o_1) - \mathbb{B}_1^{\theta^*}(a_1, o_1)\big) \big[\mathbb{X}_1^t(a^0_{1-\ell})\big]_{o^0_{1-\ell}}\bigg\|_1 \\
&\quad = \big\|\big(\mathbb{B}_1^{\theta^t}(a_1, o_1) - \mathbb{B}_1^{\theta^*}(a_1, o_1)\big)\mathbb{X}_1^t(a^0_{1-\ell})\big\|_1,
\end{aligned} \tag{H.28}$$

Here with a slight abuse of notation, we denote by $\big[\mathbb{X}_1^t(a^0_{1-\ell})\big]_{o^0_{1-\ell}}$ the $o^0_{1-\ell}$-th column of the matrix $\mathbb{X}_1^t(a^0_{1-\ell}) \in \mathbb{R}^{(O^{k+1}\cdot A^k)\times O^\ell}$. Meanwhile, the inequality follows from the fact that $0 \le \mathbb{P}^t(\underline{\tau}^0_{1-\ell}) \le 1$ for all $t \in [T]$ and $\underline{\tau}^0_{1-\ell} \in \mathcal{A}^\ell \times \mathcal{O}^\ell$. It remains to establish high confidence bound for the right-hand side of (H.28). To this end, we first obtain by triangle inequality that

$$\big\|\big(\mathbb{B}_1^{\theta^t}(a_1, o_1) - \mathbb{B}_1^{\theta^*}(a_1, o_1)\big)\mathbb{X}_1^t(a^0_{1-\ell})\big\|_1 \le \text{(i)} + \text{(ii)} + \text{(iii)} + \text{(iv)}, \tag{H.29}$$

where we define

$$\begin{aligned}
\text{(i)} &= \big\|\mathbb{B}_1^{\theta^t}(a_1, o_1)\big(\widehat{\mathbb{X}}_1^t(a^0_{1-\ell}) - \mathbb{X}_1^t(a^0_{1-\ell})\big)\big\|_1, \\
\text{(ii)} &= \|\mathbb{B}_1^{\theta^t}(a_1, o_1)\widehat{\mathbb{X}}_1^t(a^0_{1-\ell}) - \widehat{\mathbb{Y}}_1^t(a^1_{1-\ell}, o_1)\|_1, \\
\text{(iii)} &= \|\widehat{\mathbb{Y}}_1^t(a^1_{1-\ell}, o_1) - \mathbb{Y}_1^t(a^1_{1-\ell}, o_1)\|_1, \\
\text{(iv)} &= \|\mathbb{B}_1^{\theta^*}(a_1, o_1)\mathbb{X}_1^t(a^0_{1-\ell}) - \mathbb{Y}_1^t(a^1_{1-\ell}, o_1)\|_1.
\end{aligned}$$

In what follows, we upper bound terms (i)–(iv) on the right-hand side of (H.29). By the definition of parameter space and the concentration inequality in Lemma G.10, we obtain that

$$\begin{aligned}
\text{(i)} &= \big\|\mathbb{B}_1^{\theta^t}(a_1, o_1)\big(\widehat{\mathbb{X}}_1^t(a^0_{1-\ell}) - \mathbb{X}_1^t(a^0_{1-\ell})\big)\big\|_1 \\
&\le \|\mathbb{B}_1^{\theta^t}(a_1, o_1)\|_{1\mapsto 1} \cdot \|\widehat{\mathbb{X}}_1^t(a^0_{1-\ell}) - \mathbb{X}_1^t(a^0_{1-\ell})\|_1 \\
&= \mathcal{O}\big(\nu \cdot A^{2k} \cdot \sqrt{A^{k+1}\cdot O^{k+\ell}} \cdot (k+\ell) \cdot \sqrt{\log(O\cdot A\cdot T\cdot H/\delta)/T}\big),
\end{aligned} \tag{H.30}$$

where the third equality follows from the fact that $\|\mathbb{B}_1^{\theta^t}(a_1, o_1)\|_{1\mapsto 1} \le \nu \cdot A^k$ and the fact that $\|x\|_1 \le \sqrt{mn}\cdot\|x\|_2$ for $x \in \mathbb{R}^{m\times n}$. Meanwhile, by the definition of confidence set $\mathcal{C}^t$ in (G.6) and the fact that $\theta^t \in \mathcal{C}^t$, we obtain that

$$\begin{aligned}
\text{(ii)} &= \|\mathbb{B}_1^{\theta^t}(a_1, o_1)\widehat{\mathbb{X}}_1^t(a^0_{1-\ell}) - \widehat{\mathbb{Y}}_h^t(a^h_{h-\ell}, o_h)\|_1 \\
&= \mathcal{O}\big(\nu \cdot A^{2k} \cdot \sqrt{A^{k+1}\cdot O^{k+\ell}} \cdot (k+\ell) \cdot \sqrt{\log(O\cdot A\cdot T\cdot H/\delta)/T}\big).
\end{aligned} \tag{H.31}$$

By the concentration inequality in Lemma G.10, we further obtain that

$$\begin{aligned}
\text{(iii)} &= \|\widehat{\mathbb{Y}}_h^t(a^h_{h-\ell}, o_h) - \mathbb{Y}_h^t(a^h_{h-\ell}, o_h)\|_1 \\
&= \mathcal{O}\big(\sqrt{A^{k+1}\cdot O^{k+\ell}} \cdot A^k \cdot (k+\ell) \cdot \sqrt{\log(O\cdot A\cdot T\cdot H/\delta)/T}\big).
\end{aligned} \tag{H.32}$$

Finally, by the identity of Bellman operators in (G.2), we have

$$\text{(iv)} = \|\mathbb{B}_1^{\theta^*}(a_1, o_1)\mathbb{X}_1^t(a^0_{1-\ell}) - \mathbb{Y}_h^t(a^h_{h-\ell}, o_h)\|_F = 0. \tag{H.33}$$

Plugging (H.30), (H.31), (H.32), and (H.33) into (H.29), it holds for all $a_1 \in \mathcal{A}$, $o_1 \in \mathcal{O}$, and $a^0_{1-\ell} \in \mathcal{A}^\ell$ with probability at least $1-\delta$ that

$$\begin{aligned}
&\big\|\big(\mathbb{B}_1^{\theta^t}(a_1, o_1) - \mathbb{B}_1^{\theta^*}(a_1, o_1)\big)\mathbb{X}_1^t(a^0_{1-\ell})\big\|_1 \\
&\qquad = \mathcal{O}\big(\nu \cdot (k+\ell) \cdot \sqrt{A^{5k+1}\cdot O^{k+\ell}\cdot \log(O\cdot A\cdot T\cdot H/\delta)/t}\big),
\end{aligned}$$

which completes the proof of (G.11).

Part III: Proof of Upper Bound in (G.12)**.** Under Assumption G.12, it holds for all $1 < h \le H$ that

$$\begin{aligned}
&\sum_{s_{h-1}\in\mathcal{S}} \big\|\big(\mathbb{B}_h^{\theta^t}(a_h,o_h)-\mathbb{B}_h^{\theta^*}(a_h,o_h)\big)\mathbb{U}_{h,k}^{\theta^*}\mathbb{T}_{h-1}^{\theta^*}(s_{h-1},a_{h-1})\big\|_1\cdot\mathbb{P}^t(s_{h-1}\,|\,a_{h-\ell}^{h-1})\\
&\quad=\big\|\big(\mathbb{B}_h^{\theta}(a_h,o_h)-\mathbb{B}_h^{\theta'}(a_h,o_h)\big)\mathbb{U}_{h,k}^{\theta'}\mathbb{T}_{h-1}^{\theta'}(a_{h-1})\mathbb{D}\big(\widetilde{\mu}_{h-1}^{\theta'}(\overline{\pi}^t,a_{h-\ell}^{h-1})\big)\big\|_1\\
&\quad=\big\|\big(\mathbb{B}_h^{\theta}(a_h,o_h)-\mathbb{B}_h^{\theta'}(a_h,o_h)\big)\mathbb{U}_{h,k}^{\theta'}\mathbb{T}_{h-1}^{\theta'}(a_{h-1})A_{h-1,\ell}^{\overline{\pi}^t}(a_{h-\ell}^{h-2})C_{h-1,\ell}^{\overline{\pi}^t,\dagger}(a_{h-\ell}^{h-1})\big\|_1, \qquad \text{(H.34)}
\end{aligned}$$

where $A_{h-1,\ell}^{\overline{\pi}^t}$ is the reverse emission operator defined in (G.11) and $C_{h-1,\ell}^{\overline{\pi}^t,\dagger}$ is the right inverse of $C_{h-1,\ell}^{\overline{\pi}^t}$ in Assumption G.12. Meanwhile, by Assumption G.12, it holds that

$$\begin{aligned}
&\big\|\big(\mathbb{B}_h^{\theta}(a_h,o_h)-\mathbb{B}_h^{\theta'}(a_h,o_h)\big)\mathbb{U}_{h,k}^{\theta'}\mathbb{T}_{h-1}^{\theta'}(a_{h-1})A_{h-1,\ell}^{\overline{\pi}^t}(a_{h-\ell}^{h-2})C_{h-1,\ell}^{\overline{\pi}^t,\dagger}(a_{h-\ell}^{h-1})\big\|_1\\
&\quad\le\gamma\cdot\big\|\big(\mathbb{B}_h^{\theta}(a_h,o_h)-\mathbb{B}_h^{\theta'}(a_h,o_h)\big)\mathbb{U}_{h,k}^{\theta'}\mathbb{T}_{h-1}^{\theta'}(a_{h-1})A_{h-1,\ell}^{\overline{\pi}^t}(a_{h-\ell}^{h-1})\big\|_1. \qquad \text{(H.35)}
\end{aligned}$$

By the identity of $\mathbb{X}_h^t(a_{h-\ell}^{h-1})$ in (G.9), we further obtain that

$$\begin{aligned}
&\big\|\big(\mathbb{B}_h^{\theta}(a_h,o_h)-\mathbb{B}_h^{\theta'}(a_h,o_h)\big)\mathbb{U}_{h,k}^{\theta'}\mathbb{T}_{h-1}^{\theta'}(a_{h-1})A_{h-1,\ell}^{\overline{\pi}^t}\big\|_1\\
&\quad=\big\|\big(\mathbb{B}_h^{\theta}(a_h,o_h)-\mathbb{B}_h^{\theta'}(a_h,o_h)\big)\mathbb{X}_h^t(a_{h-\ell}^{h-1})\big\|_1. \qquad \text{(H.36)}
\end{aligned}$$

We now upper bound the right-hand side of (H.36). The calculation is similar to that in Part II of the proof. By triangle inequality, we obtain

$$\big\|\big(\mathbb{B}_h^{\theta}(a_h,o_h)-\mathbb{B}_h^{\theta'}(a_h,o_h)\big)\mathbb{X}_h^t(a_{h-\ell}^{h-1})\big\|_1\le\text{(v)}+\text{(vi)}+\text{(vii)}+\text{(viii)}, \qquad \text{(H.37)}$$

where we define

$$\begin{aligned}
\text{(v)}&=\big\|\mathbb{B}_h^{\theta^t}(a_h,o_h)\big(\widehat{\mathbb{X}}_h^t(a_{h-\ell}^{h-1})-\mathbb{X}_h^t(a_{h-\ell}^{h-1})\big)\big\|_1,\\
\text{(vi)}&=\|\mathbb{B}_h^{\theta^t}(a_h,o_h)\widehat{\mathbb{X}}_h^t(a_{h-\ell}^{h-1})-\widehat{\mathbb{Y}}_h^t(a_{h-\ell}^{h},o_h)\|_1,\\
\text{(vii)}&=\|\widehat{\mathbb{Y}}_h^t(a_{h-\ell}^{h},o_h)-\mathbb{Y}_h^t(a_{h-\ell}^{h},o_h)\|_1,\\
\text{(viii)}&=\|\mathbb{B}_h^{\theta^*}(a_h,o_h)\mathbb{X}_h^t(a_{h-\ell}^{h-1})-\mathbb{Y}_h^t(a_{h-\ell}^{h},o_h)\|_1.
\end{aligned}$$

In what follows, we upper bound terms (v)–(viii) on the right-hand side of (H.37). By the definition of parameter space and the concentration inequality in Lemma G.10, we obtain that

$$\begin{aligned}
\text{(v)}&=\big\|\mathbb{B}_h^{\theta^t}(a_h,o_h)\big(\widehat{\mathbb{X}}_1^t(a_{h-\ell}^{h-1})-\mathbb{X}_h^t(a_{h-\ell}^{h-1})\big)\big\|_1\\
&\le\|\mathbb{B}_h^{\theta^t}(a_h,o_h)\|_{1\mapsto1}\cdot\|\widehat{\mathbb{X}}_h^t(a_{h-\ell}^{h-1})-\mathbb{X}_h^t(a_{h-\ell}^{h-1})\|_1\\
&=\mathcal{O}\big(\nu\cdot\sqrt{A^{5k+1}\cdot O^{k+\ell}}\cdot(k+\ell)\cdot\sqrt{\log(O\cdot A\cdot T\cdot H/\delta)/T}\big). \qquad \text{(H.38)}
\end{aligned}$$

Meanwhile, by the definition of confidence set $\mathcal{C}^t$ in (G.6) and the fact that $\theta^t\in\mathcal{C}^t$, we obtain that

$$\begin{aligned}
\text{(vi)}&=\|\mathbb{B}_h^{\theta^t}(a_h,o_h)\widehat{\mathbb{X}}_h^t(a_{h-\ell}^{h-1})-\widehat{\mathbb{Y}}_h^t(a_{h-\ell}^{h},o_h)\|_1\\
&=\mathcal{O}\big(\nu\cdot\sqrt{A^{5k+1}\cdot O^{k+\ell}}\cdot(k+\ell)\cdot\sqrt{\log(O\cdot A\cdot T\cdot H/\delta)/T}\big). \qquad \text{(H.39)}
\end{aligned}$$

By the concentration inequality in Lemma G.10, we further obtain that

$$\begin{aligned}
\text{(vii)}&=\|\widehat{\mathbb{Y}}_h^t(a_{h-\ell}^{h},o_h)-\mathbb{Y}_h^t(a_{h-\ell}^{h},o_h)\|_1\\
&=\mathcal{O}\big(\sqrt{A^{3k+1}\cdot O^{k+\ell}}\cdot(k+\ell)\cdot\sqrt{\log(O\cdot A\cdot T\cdot H/\delta)/T}\big). \qquad \text{(H.40)}
\end{aligned}$$

Finally, by the identity of Bellman operators in (G.2), we have

$$\text{(viii)}=\|\mathbb{B}_h^{\theta^*}(a_h,o_h)\mathbb{X}_h^t(a_{h-\ell}^{h-1})-\mathbb{Y}_h^t(a_{h-\ell}^{h},o_h)\|_1=0. \qquad \text{(H.41)}$$

Plugging (H.38), (H.39), (H.40), and (H.41) into (H.37), it holds for all $a_h\in\mathcal{A}$, $o_h\in\mathcal{O}$, and $a_{h-\ell}^{h-1}\in\mathcal{A}^\ell$ with probability at least $1-\delta$ that

$$\begin{aligned}
&\big\|\big(\mathbb{B}_h^{\theta}(a_h,o_h)-\mathbb{B}_h^{\theta'}(a_h,o_h)\big)\mathbb{X}_h^t(a_{h-\ell}^{h-1})\big\|_1\\
&\quad=\mathcal{O}\big(\nu\cdot\sqrt{A^{5k+1}\cdot O^{k+\ell}}\cdot(k+\ell)\cdot\sqrt{\log(O\cdot A\cdot T\cdot H/\delta)/T}\big). \qquad \text{(H.42)}
\end{aligned}$$

Combining (H.34), (H.35), (H.36), and (H.42), we conclude that

$$\begin{aligned}
&\sum_{s_{h-1}\in\mathcal{S}} \big\|\big(\mathbb{B}_h^{\theta^t}(a_h,o_h)-\mathbb{B}_h^{\theta^*}(a_h,o_h)\big)\mathbb{U}_{h,k}^{\theta^*}\mathbb{T}_{h-1}^{\theta^*}(s_{h-1},a_{h-1})\big\|_1\cdot\mathbb{P}^t(s_{h-1}\,|\,a_{h-\ell}^{h-1})\\
&\quad=\big\|\big(\mathbb{B}_h^{\theta}(a_h,o_h)-\mathbb{B}_h^{\theta'}(a_h,o_h)\big)\mathbb{U}_{h,k}^{\theta'}\mathbb{T}_{h-1}^{\theta'}(a_{h-1})A_{h-1,\ell}^{\pi}(a_{h-\ell}^{h-2})C_{h-1,\ell}^{\pi,\dagger}(a_{h-\ell}^{h-2})\big\|_1\\
&\quad\le\gamma\cdot\big\|\big(\mathbb{B}_h^{\theta}(a_h,o_h)-\mathbb{B}_h^{\theta'}(a_h,o_h)\big)\mathbb{X}_h^t(a_{h-\ell}^{h-1})\big\|_1\\
&\quad=\mathcal{O}\Big(\gamma\cdot\nu\cdot(k+\ell)\cdot\sqrt{A^{5k+\ell}\cdot O^{k+1}\cdot\log(O\cdot A\cdot T\cdot H/\delta)/t}\Big),
\end{aligned}$$

where the first inequality follows from (H.35) and (H.36). Thus, we complete the proof of (G.11).

□

H.8 Proof of Lemma H.1

Proof. Recall that we define

$$\widetilde{\mathbb{O}}_h(o_h)=\mathbb{D}\big(\mathbb{O}_h(o_h\,|\,\cdot)\big)=\mathbb{D}\Big(\big[\mathbb{O}(o_h\,|\,s_h)\big]_{s_h}\Big)\in\mathbb{R}^{S\times S},$$

where we denote by $\mathbb{D}(v)\in\mathbb{R}^{S\times S}$ the diagonal matrix where the diagonal entries aligns with the vector $v\in\mathbb{R}^S$. Thus, it holds for all $h\in[H]$ that

$$\begin{aligned}
\sum_{o_h\in\mathcal{O}}\|\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)u\|_1&\le\sum_{o_h\in\mathcal{O}}\|\widetilde{\mathbb{O}}_h(o_h)u\|_1=\sum_{o_h\in\mathcal{O}}\sum_{s_h\in\mathcal{S}}\mathbb{O}_h(o_h\,|\,s_h)\cdot|u(s_h)|\\
&=\sum_{s_h\in\mathcal{S}}|u(s_h)|=\|u\|_1,
\end{aligned}$$

where the first inequality follows from the fact that $\mathbb{T}_h(a_h)$ is a transition matrix. Here we denote by $u(s_h)$ the s_h-th entry of $u\in\mathbb{R}^S$. Inductively, we obtain that

$$\sum_{o_h^{H-1}\in\mathcal{O}^{H-h}}\|\mathbb{T}_{H-1}(a_{H-1})\widetilde{\mathbb{O}}_{H-1}(o_{H-1})\ldots\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)u\|_1\le\|u\|_1. \tag{H.43}$$

Meanwhile, note that $\mathbb{U}_h$ is a transition matrix. Thus, it holds for all $h\in[H]$ that

$$\begin{aligned}
&\|\mathbb{U}_h\mathbb{T}_{H-1}(a_{H-1})\widetilde{\mathbb{O}}_{H-1}(o_{H-1})\ldots\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)u\|_1\\
&\quad\le\|\mathbb{T}_{H-1}(a_{H-1})\widetilde{\mathbb{O}}_{H-1}(o_{H-1})\ldots\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)u\|_1
\end{aligned}\tag{H.44}$$

Combining (H.43) and (H.44), we conclude

$$\sum_{o_h^{H-1}\in\mathcal{O}^{H-h}}\|\mathbb{U}_h\mathbb{T}_{H-1}(a_{H-1})\widetilde{\mathbb{O}}_{H-1}(o_{H-1})\ldots\mathbb{T}_h(a_h)\widetilde{\mathbb{O}}_h(o_h)u\|_1\le\|u\|_1,$$

which completes the proof of Lemma H.1. □