A Note on Belief Space Binning for POMDPs: Guarantees under Smoothness Condition

Youheng Zhu
CS Department
University of Illinois at Urbana Champaign
email: youheng@illinois.edu

Abstract

1 Introduction

2 Preliminaries

In this section, we introduce the model setup and algorithms for off-line POMDPs.

2.1 Infinite-horizon Discounted POMDP

An infinite-horizon discounted POMDP is a 7-tuple: $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, r, \gamma, \mathbb{O}, \mathbb{T} \rangle$ where $\gamma \in [0, 1)$ is the discount factor, $\mathcal{S}$ is the latent state space, $\mathcal{A}$ is the action space, $\mathcal{O}$ is the observation space, $r : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is the bounded reward function, $\mathbb{O} : \mathcal{S} \to \Delta(\mathcal{O})$ is the emission dynamic, and $\mathbb{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition dynamic. We use $\Delta(\cdot)$ to represent a probability distribution on a specific space, and $\mathbb{O}$ and $\mathbb{T}$ respectively indicates the probability of the observation given the current state, and that of the next-state given current state-action pair. We use $|\cdot|$ to denote the cardinality of a space. For simplicity yet without loosing the essence of the idea, we consider the spaces $\mathcal{S}, \mathcal{A}, \mathcal{O}$ to be discrete and finite.

The system dynamic can be uniquely demonstrated by the following procedure:

It is important to note that in a general POMDP, the latent state space $\mathcal{S}$ is unknown to the learner, and learners only get access to the trajectories sampled by a behavior policy, which is invariant under the off-line setting.

2.2 Belief State Space and Smoothness Condition

Since one cannot observe the latent state directly, the overall best prediction of the current state is by using the information from the entire history of observations and actions. We denote the history at time step h to be

$$\tau_h = (o_1, a_1, o_2, a_2, \cdots, o_{h-1}, a_{h-1}) \in \mathcal{H}, \tag{1}$$

and consequently one can predict the current state given the history data. The belief state $\mathbf{b}(\tau_h) = \Pr(s_h|\tau_h)$ is an element of $\Delta(\mathcal{S}) \subset \mathbb{R}^{|\mathcal{S}|}$ when $|\mathcal{S}| < \infty$. We use $\mathcal{B}$ to denote belief state space such that

$$\mathcal{B} = \{b : \exists h \in \mathbb{N}\ \exists \tau_h, \mathbf{b}(\tau_h) = b\} \tag{2}$$

Assumption 1. $\mathbf{b} : \mathcal{H} \to \mathcal{B}$ *is an injection (and thus a bijection).*

The assumption is especially natural when considering very large latent state space and therefore very high-dimensional belief state space. Consequently, $|\mathcal{B}| = \infty$ unless we truncate an infinitely long tail from the history that we consider and bares a reasonable truncation error. In a following example and section 7, we will discuss the case of $|\mathcal{B}| < \infty$, which gives us some worst-case properties.

With the assumption, we denote the policy of interest $\tilde{\pi}(\tau_h) = \pi(\mathbf{b}(\tau_h)) : \mathcal{H} \to \Delta(\mathcal{A})$, which is used to sample an action when given a history.

It is easy to see that a good belief state policy should treat two similar belief state similarly, and thus should have some smoothness condition with regard to the topology of the belief state space. We typically denote the type of policy we are interested in using the following assumption.

Assumption 2. ***(Lipchitz of Policy)*** $\exists L_\pi, \|\pi(b_1) - \pi(b_2)\|_1 \leq L_\pi \|b_1 - b_2\|_1$.

2.3 Off-line Data

A set of off-line data $\mathcal{D}$ is sampled using a behavior policy $\tilde{\pi}_b$ in the following manner: Independently sample n trajectories $(o_1, a_1 \cdots)$ from the POMDP, then select randomly a tuple $(\tau_{h_i}^{[i]}, o_{h_i}^{[i]}, a_{h_i}^{[i]})_i$ from each trajectory respectively. Eventually,

$$\mathcal{D} = \{(\tau_{h_i}^{[i]}, o_{h_i}^{[i]}, a_{h_i}^{[i]})_i\}_{i=1}^n. \tag{3}$$

3 Overall Analysis In a Nutshell

[**Youheng:** to be continue...]

[**Youheng:** I'll draw a graph to indicate the three steps I follow to derive a guarantee for the algorithm on the original POMDP.]

4 Off-Policy Evaluation under Smooth Conditions

4.1 Abstraction under Covering

Definition 1. *A ε-cover $\mathcal{C}_\varepsilon$ is a subspace of the belief state space which satisfies:*

$$\bigcup_{c \in \mathcal{C}_\varepsilon} \mathbf{B}(c, \varepsilon) \supset \mathcal{B} \tag{4}$$

where $\mathbf{B}(c, \varepsilon)$ stands for an open ball centered at c with radius ε. The cardinality of $\mathcal{C}_\varepsilon$ is called ε-covering number. For every ε-cover $\mathcal{C}_\varepsilon$, there exist a partition of the belief state space, where each $c \in \mathcal{C}_\varepsilon$ acts as the representation element of the bin.

Lemma 1. *For two belief states b_1 and b_2, $\forall a \in \mathcal{A}$, we have:*

$$|r(b_1, a) - r(b_2, a)| \leq R_{\max} \|b_1 - b_2\|_1. \tag{5}$$

Proof. This is easily obtained from:

$$\begin{aligned} |r(b_1, a) - r(b_2, a)| &= |\mathbb{E}_{s \sim b_1}[r(s, a)] - \mathbb{E}_{s \sim b_2}[r(s, a)]| \\ &= |\langle r(\cdot, a), b_1 - b_2 \rangle| \\ &\leq R_{\max} \|b_1 - b_2\|_1. \end{aligned}$$

And it shows that when treating POMDPs as belief space MDPs, there's intrinsic smoothness within the dynamic. □

For simplicity, we first adopt the standard bisimulation setting.

Assumption 3. ***(Bisimulation)*** *For two arbitrary belief points b_1 and b_2 in the bin represented by an element c in the ε-cover $\mathcal{C}_\varepsilon$, we have:*

$$\|\Phi P(b_1, a) - \Phi P(b_2, a)\|_1 \leq L_b \varepsilon.$$

This would put condition on the ε-cover $\mathcal{C}_\varepsilon$ and the partition it induces.

Then, we have the result from standard abstraction literature:

Lemma 2. *Under Assumption 3, for the abstract policy π_ϕ, which indicates the true policy π descending to the binned belief space, we have*

$$\|[V_{\text{bin}}^{\pi_\phi}]_{\text{true}} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq \frac{R_{\max}\varepsilon}{1-\gamma} + \frac{\gamma L_b R_{\max}\varepsilon}{2(1-\gamma)^2}. \tag{6}$$

However, Lemma 2 only controls the difference between the value function of the true system and that of the binned system for the abstracted policy π_ϕ, which is different from the true policy π. Before we fill this gap, one may notice that the assumption for bisimulation can be too strong and hard to understand in this context, and we'd like a weaker and more comprehensible result. To address this, we first put forward the following commonly-adopted assumption:

Assumption 4. ***(Belief Space Contraction)*** *Let b_1, b_2 be two belief points in the belief state space $\mathcal{B}$, we use the following notation to represent the next-state belief:*

$$b^{o,a} = \mathbf{b}(\mathbf{b}^{-1}(b) + o + a)$$

where $+$ represents concatenation. We also use b^{+1} instead of $b^{o,a}$ when we do not emphasise a specific (o,a) pair. The same notation works for $b^{+2}, b^{+3}, \cdots$.

The contraction assumption states that $\forall o, a$,

$$\|b_1^{o,a} - b_2^{o,a}\|_1 \leq \eta \|b_1 - b_2\|_1$$

for some uniform $\eta \in (0,1]$.

A weaker assumption for this is

Assumption 5. ***(Expected Belief Space Contraction)*** *Replace the contraction assumption in Assumption 4 by:*

$$\mathbb{E}_{\pi(b_1)}\left[\frac{\|b_1^{+k} - b_2^{+k}\|_1}{\|b_1 - b_2\|_1}\right] \leq \eta^k$$

which is a necessary condition for Assumption 4.

Lemma 3. *(Lemma 2 in [5]) For any two belief points b_1, b_2 satisfying $\|b_1 - b_2\|_1 \leq \varepsilon$, $|P(o|b_1,a) - P(o|b_2,a)| \leq \|b_1 - b_2\|_1 \leq \varepsilon$.*

Consequently, we put forward the following proposition.

Proposition 1. *For policy π satisfying Assumption 2, we have for $\forall o, a$*

$$|P(b_1^{o,a}|b_1) - P(b_2^{o,a}|b_2)| \leq (1 + L_\pi)\|b_1 - b_2\|_1. \tag{7}$$

Proof. We decompose our target function as

$$\begin{aligned}
&|P(b_1^{o,a}|b_1) - P(b_2^{o,a}|b_2)| \\
&= |P(o|b_1,a)\pi(a|b_1) - P(o|b_2,a)\pi(a|b_2)| \\
&= |P(o|b_1,a)\pi(a|b_1) - P(o|b_1,a)\pi(a|b_2) + P(o|b_1,a)\pi(a|b_2) - P(o|b_2,a)\pi(a|b_2)| \\
&\leq |P(o|b_1,a)(\pi(a|b_1) - \pi(a|b_2))| + |(P(o|b_1,a) - P(o|b_2,a))\pi(a|b_2)| \\
&\leq |P(o|b_1,a)| \cdot |\pi(a|b_1) - \pi(a|b_2)| + |P(o|b_1,a) - P(o|b_2,a)| \cdot |\pi(a|b_2)| \\
&\leq (1 + L_\pi)\|b_1 - b_2\|_1
\end{aligned} \tag{8}$$

where we used Lemma 3 for the last inequality. □

[**Youheng:** Here the bound is somehow loose (by a $|\mathcal{O}||\mathcal{A}|$ factor) in an expected manner.] A tighter result for Proposition 1 which will be useful later is

Proposition 2. *For any $o \in \mathcal{O}$,*

$$\left|\sum_{a\in\mathcal{A}}(P(b_1^{o,a}|b_1) - P(b_2^{o,a}|b_2))\right| \leq (1 + L_\pi)\|b_1 - b_2\|_1. \tag{9}$$

Proof. We follow the same idea in the proof of Proposition 1 and decompose the LHS as

$$
\begin{aligned}
&\left|\sum_{a\in\mathcal{A}}(P(b_1^{o,a}|b_1)-P(b_2^{o,a}|b_2))\right| \\
\leq &\left|\sum_{a\in\mathcal{A}}P(o|b_1,a)(\pi(a|b_1)-\pi(a|b_2))\right|+\left|\sum_{a\in\mathcal{A}}(P(o|b_1,a)-P(o|b_2,a))\pi(a|b_2)\right| \\
\leq &\max_{a\in\mathcal{A}}\left|P(o|b_1,a)\right|\cdot\left|\sum_{a\in\mathcal{A}}(\pi(a|b_1)-\pi(a|b_2))\right|+ \\
&\qquad\max_{a\in\mathcal{A}}\left|P(o|b_1,a)-P(o|b_2,a)\right|\cdot\left|\sum_{a\in\mathcal{A}}\pi(a|b_2)\right| \\
\leq &(1+L_\pi)\|b_1-b_2\|_1, \qquad (10)
\end{aligned}
$$

which proves the result. □

The one-step error is easy to control, however, without bisimulation, it is extremely difficult to control the accumulative error induced by infinite amount of steps. We illustrate first by the following proposition which shows that without the condition that b_1 and b_2 are close to each other, the approximation of the difference of expected reward at the k-th step.

Proposition 3.

$$
|\mathbb{E}_{a,o,\cdots\sim b_1}\left[r(b_1^{+k},a_k)\right]-\mathbb{E}_{a,o,\cdots\sim b_2}\left[r(b_1^{+k},a_k)\right]|\leq 4R_{\max}\eta^k \qquad (11)
$$

Proof.

$$
\begin{aligned}
&|\mathbb{E}_{a,o,\cdots\sim b_1}\left[r(b_1^{+k},a_k)\right]-\mathbb{E}_{a,o,\cdots\sim b_2}\left[r(b_1^{+k},a_k)\right]| \\
=&\,|\langle P_{[o,a]^k|b}(\cdot|b_1),r(b_1^{[\cdot]},a)\rangle-\langle P_{[o,a]^k|b}(\cdot|b_2),r(b_1^{[\cdot]},a)\rangle| \qquad (12)\\
=&\,|\langle P_{[o,a]^k|b}(\cdot|b_1),r(b_1^{[\cdot]},a)\rangle-\langle P_{[o,a]^k|b}(\cdot|b_1),\overline{R}\rangle \\
&+\langle P_{[o,a]^k|b}(\cdot|b_2),\overline{R}\rangle-\langle P_{[o,a]^k|b}(\cdot|b_2),r(b_1^{[\cdot]},a)\rangle| \qquad (13)\\
\leq&\,2\|\overline{R}-r(b_1^{+k},a)\|_\infty \qquad (14)\\
\leq&\,4R_{\max}\eta^k \qquad (15)
\end{aligned}
$$

□

The last inequality was the result of belief space contraction. In comparison, one can try to obtain an upper bound for k-step transition error:

Proposition 4. $|P(b_1^{+k}|b_1)-P(b_2^{+k}|b_2)|\leq(1+L_\pi)(1+\eta)^k\|b_1-b_2\|_1$

Proof. Using Proposition 1, we get

$$
|P(b_1^{+1}|b_1)-P(b_2^{+1}|b_2)|\leq(1+L_\pi)\|b_1-b_2\|_1. \qquad (16)
$$

Replacing b_1,b_2 with b_1^{+1},b_2^{+2} and using Assumption 4, we have

$$
|P(b_1^{+2}|b_1^{+1})-P(b_2^{+2}|b_2^{+1})|\leq(1+L_\pi)\eta\|b_1-b_2\|_1. \qquad (17)
$$

Therefore

$$
\begin{aligned}
&|P(b_1^{+2}|b_1) - P(b_2^{+2}|b_2)| \\
&= \left| P(b_1^{+2}|b_1) - \sum_{o,a} P(b_2^{+2}|b_2^{o,a}) P(b_1^{o,a}|b_1) + \sum_{o,a} P(b_2^{+2}|b_2^{o,a}) P(b_1^{o,a}|b_1) - P(b_2^{+2}|b_2) \right| && (18) \\
&\leq \left| P(b_1^{+2}|b_1) - \sum_{o,a} P(b_2^{+2}|b_2^{o,a}) P(b_1^{o,a}|b_1) \right| + \left| \sum_{o,a} P(b_2^{+2}|b_2^{o,a}) P(b_1^{o,a}|b_1) - P(b_2^{+2}|b_2) \right| && (19) \\
&= \left| \sum_{o,a} \left(P(b_1^{+2}|b_1^{o,a}) - P(b_2^{+2}|b_2^{o,a}) \right) P(b_1^{o,a}|b_1) \right| + \left| \sum_{o,a} P(b_2^{+2}|b_2^{o,a}) \left(P(b_1^{o,a}|b_1) - P(b_2^{o,a}|b_2) \right) \right| \\
&\leq \|P(b_1^{+2}|b_1^{[\cdot]}) - P(b_2^{+2}|b_2^{[\cdot]})\|_\infty \|P(b_1^{[\cdot]}|b_1)\|_1 + \|P(b_2^{+2}|b_2^{[\cdot]})\|_1 \|P(b_1^{[\cdot]}|b_1) - P(b_2^{[\cdot]}|b_2)\|_\infty && (20) \\
&\leq (1 + L_\pi)\eta \|b_1 - b_2\|_1 \cdot 1 + 1 \cdot (1 + L_\pi) \|b_1 - b_2\|_1 && (21) \\
&\leq (1 + L_\pi)(1 + \eta) \|b_1 - b_2\|_1 && (22)
\end{aligned}
$$

where in (21) we used the fact that $\|P(b_2^{+2}|b_2^{[\cdot]})\|_1 = P(b_2^{o_1,a_1,o_2,a_2}|b_2^{o_1,a_1}) \leq 1$. This is the consequence of Assumption 1 that every belief state has a unique history. **[Youheng:** The same as Proposition 1, every step there can be a $|\mathcal{O}||\mathcal{A}|$ factor loose.]

After that, we recursively repeat the procedure above, and using mathematical induction, we get the result. □

With Proposition 4, we can try to control the error propagated from the initial belief space difference:

$$
\begin{aligned}
&|\mathbb{E}_{a,o,\cdots \sim b_1}\left[r(b_1^{+k}, a_k)\right] - \mathbb{E}_{a,o,\cdots \sim b_2}\left[r(b_1^{+k}, a_k)\right]| \\
&= |\langle (P(b_1^{[\cdot]}|b_1) - P(b_2^{[\cdot]}|b_2)), r(b_1^{[\cdot]}, a) \rangle| && (23) \\
&\leq \|P(b_1^{+k}|b_1) - P(b_2^{+k}|b_2)\|_1 \cdot R_{\max}. && (24)
\end{aligned}
$$

Since we only get an L_∞ norm in Proposition 4, extending it to L_1 would need an extra $(|\mathcal{O}||\mathcal{A}|)^k$ expenses **[Youheng:** which is potentially tightable, but would need extra analysis. Even if this value is reasonable, we would still need the discounted factor γ to be small enough as stated later.], making the error propagation explode drastically, not to mention that our horizon would go to infinity. Unless $\gamma < 1/(1+\eta)|\mathcal{O}||\mathcal{A}|$, the error would be impossible to control in this analysis. This also tells us that we do need an assumption such as bisimulation to prevent the error from exploding. Meanwhile, bisimulation does more than that.

To conclude, bisimulation is the guarantee that the error throughout the horizon can be controlled, but it not only guarantees that. With this understanding, we would directly make it an assumption as a weaker alternative to bisimulation for further analysis.

Assumption 6. ***(Error Propagation Control)***

$$
\exists L_H,\ \forall k,\ |\mathbb{E}_{a,o,\cdots \sim b_1}\left[r(b_1^{+k}, a_k)\right] - \mathbb{E}_{a,o,\cdots \sim b_2}\left[r(b_1^{+k}, a_k)\right]| \leq L_H R_{\max} \|b_1 - b_2\|_1.
$$

*[**Youheng:** or replace b_2 by $\phi(b_1)$]*

The assumption above is almost identical to the following assumption about value functions. But there are differences as shown in the proof of the next theorem.

Assumption 7. ***(Lipchitz of Value Function)***

$$
\begin{aligned}
\exists L_V, |V^\pi(b_1) - V^\pi(b_2)| &\leq L_V \|b_1 - b_2\|_1 \\
|V^{\pi_\phi}(b_1) - V^{\pi_\phi}(b_2)| &\leq L_V \|b_1 - b_2\|_1
\end{aligned}
$$

*[**Youheng:** There are close relationships between this and the former assumption, which is also covered in the proof of the next theorem. Can elaborate it separately later.]*

With enough preparation, we now try to control the difference of value function for any abstract policy π_ϕ by proposing the following theorem.

Theorem 1. *Under Assumption 6 and Assumption 4, we have*

$$\|[\tilde{V}_{\text{bin}}^{\pi_\phi}]_{\text{true}} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq \frac{L_H R_{\max}\varepsilon}{1-\gamma} + \frac{R_{\max}\varepsilon}{1-\gamma\eta} \tag{25}$$

Replacing Assumption 4 by Assumption 5, we have

$$\|[\tilde{V}_{\text{bin}}^{\pi_\phi}]_{\text{true}} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq \frac{L_H R_{\max}\varepsilon}{1-\gamma} + R_{\max}\sqrt{\varepsilon}\cdot\left(\frac{1}{1-\gamma} + \frac{1}{1-\gamma\eta}\right) \tag{26}$$

Proof. We first control $\mathbb{E}_{a,o,\cdots\sim b_1}\big[\sum_{k=0}^\infty \gamma^k r(b_1^{+k}, a_k)\big] - \mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^\infty \gamma^k r(b_2^{+k}, a_k)\big]$. After we do this, we can reduce the problem to the one we need using

$$\begin{aligned}
&\left|\sum_{b_1\in\text{bin}(\phi(b))}\left[p_{\phi(b)}(b_1)\mathbb{E}_{a,o,\cdots\sim b_1}\big[\sum_{k=0}^\infty \gamma^k r(b_1^{+k}, a_k)\big]\right] - \mathbb{E}_{a,o,\cdots\sim b'}\big[\sum_{k=0}^\infty \gamma^k r(b'^{+k}, a_k)\big]\right| \\
\leq &\left|\mathbb{E}_{a,o,\cdots\sim b}\big[\sum_{k=0}^\infty \gamma^k r(b^{+k}, a_k)\big] - \mathbb{E}_{a,o,\cdots\sim b'}\big[\sum_{k=0}^\infty \gamma^k r(b'^{+k}, a_k)\big]\right|
\end{aligned} \tag{27}$$

which is already controlled.

To do this, we split the formula into two parts:

$$\begin{aligned}
&\left|\mathbb{E}_{a,o,\cdots\sim b_1}\big[\sum_{k=0}^\infty \gamma^k r(b_1^{+k}, a_k)\big] - \mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^\infty \gamma^k r(b_2^{+k}, a_k)\big]\right| \\
\leq &\left|\mathbb{E}_{a,o,\cdots\sim b_1}\big[\sum_{k=0}^\infty \gamma^k r(b_1^{+k}, a_k)\big] - \mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^\infty \gamma^k r(b_1^{+k}, a_k)\big]\right| + \\
&\left|\mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^\infty \gamma^k r(b_1^{+k}, a_k)\big] - \mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^\infty \gamma^k r(b_2^{+k}, a_k)\big]\right|.
\end{aligned} \tag{28}$$

We first look at the first term.

$$\begin{aligned}
&\left|\mathbb{E}_{a,o,\cdots\sim b_1}\big[\sum_{k=0}^\infty \gamma^k r(b_1^{+k}, a_k)\big] - \mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^\infty \gamma^k r(b_1^{+k}, a_k)\big]\right| \\
= &\left|\sum_{k=0}^\infty\left(\mathbb{E}_{a,o,\cdots\sim b_1}\big[\gamma^k r(b_1^{+k}, a_k)\big] - \mathbb{E}_{a,o,\cdots\sim b_2}\big[\gamma^k r(b_1^{+k}, a_k)\big]\right)\right|
\end{aligned} \tag{29}$$

which corresponds to the propagated error within each layer and summing them up. As discussed above, with Assumption 6, this term is dominated by $L_H R_{\max}\|b_1 - b_2\|_1/(1-\gamma)$.

Next, we look at the second term which is not covered by Assumption 6.

$$\begin{aligned}
&\left|\mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^\infty \gamma^k r(b_1^{+k}, a_k)\big] - \mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^\infty \gamma^k r(b_2^{+k}, a_k)\big]\right| \\
= &\left|\sum_{k=0}^\infty\left(\mathbb{E}_{a,o,\cdots\sim b_2}\big[\gamma^k r(b_1^{+k}, a_k) - \gamma^k r(b_2^{+k}, a_k)\big]\right)\right| \\
\leq &\frac{R_{\max}\|b_1 - b_2\|_1}{1-\gamma\eta}
\end{aligned} \tag{30}$$

where we used Lemma 1 and Assumption 4. If we apply a weaker assumption, Assumption 5, we have using Markov's inequality,

$$P\left(\frac{\|b_1^{+k} - b_2^{+k}\|_1}{\|b_1 - b_2\|_1} \geq \frac{1}{\sqrt{\varepsilon}}\right) \leq \sqrt{\varepsilon}\cdot\eta^k,\ \forall\varepsilon > 0. \tag{31}$$

Consequently,

$$\begin{aligned}
&\left|\mathbb{E}_{a,o,\cdots\sim b_2}\left[\gamma^k r(b_1^{+k},a_k)-\gamma^k r(b_2^{+k},a_k)\right]\right| \\
\le\ &\gamma^k\left|\mathbb{E}_{a,o,\cdots\sim b_2}\left[r(b_1^{+k},a_k)-r(b_2^{+k},a_k)\right]\mathbb{I}\left(\frac{\|b_1^{+k}-b_2^{+k}\|_1}{\|b_1-b_2\|_1}\ge\frac{1}{\sqrt{\varepsilon}}\right)\right|+ \\
&\gamma^k\left|\mathbb{E}_{a,o,\cdots\sim b_2}\left[r(b_1^{+k},a_k)-r(b_2^{+k},a_k)\right]\mathbb{I}\left(\frac{\|b_1^{+k}-b_2^{+k}\|_1}{\|b_1-b_2\|_1}<\frac{1}{\sqrt{\varepsilon}}\right)\right| \\
\le\ &\gamma^k R_{\max}\cdot\sqrt{\varepsilon}\cdot\eta^k+\gamma^k R_{\max}\cdot\frac{\|b_1-b_2\|_1}{\sqrt{\varepsilon}}.
\end{aligned}\tag{32}$$

Setting $\varepsilon=\|b_1-b_2\|_1$, we get

$$\left|\mathbb{E}_{a,o,\cdots\sim b_2}\left[\gamma^k r(b_1^{+k},a_k)-\gamma^k r(b_2^{+k},a_k)\right]\right|\le\gamma^k R_{\max}(1+\eta^k)\sqrt{\|b_1-b_2\|_1}.\tag{33}$$

Finally, summing up all the layers of horizon, we get the second term controlled as

$$\begin{aligned}
&\left|\mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^{\infty}\gamma^k r(b_1^{+k},a_k)\big]-\mathbb{E}_{a,o,\cdots\sim b_2}\big[\sum_{k=0}^{\infty}\gamma^k r(b_2^{+k},a_k)\big]\right| \\
\le\ &R_{\max}\sqrt{\|b_1-b_2\|_1}\cdot\left(\frac{1}{1-\gamma}+\frac{1}{1-\gamma\eta}\right).
\end{aligned}\tag{34}$$

which is also controllable, yet by a worse rate (square root). Combining the two terms, we prove the result using the fact that $\|b_1-b_2\|_1\le\varepsilon$ inside each bin. □

Next we control the gap between $\tilde{V}_{\text{bin}}^{\pi_\phi}$ and$V_{\text{bin}}^{\pi_\phi}$.

Theorem 2.

$$\|\tilde{V}_{\text{bin}}^{\pi_\phi}-V_{\text{bin}}^{\pi_\phi}\|_\infty\le\frac{R_{\max}\varepsilon}{1-\gamma}+\frac{R_{\max}L_V\varepsilon}{(1-\gamma)^2}\tag{35}$$

Proof. Our idea is to use chaining. Notice that

$$\tilde{V}_{\text{bin}}^{\pi_\phi}=\mathbb{E}_{b_1\sim\text{bin}(\phi(b))}[V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_1)]\tag{36}$$

and

$$V_{\text{bin}}^{\pi_\phi}=\mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots}}\left[r_\phi(\phi(b_1),a_1)+\gamma r_\phi(\phi(b_2),a_2)+\gamma^2 r_\phi(\phi(b_3),a_3)+\cdots\right]\tag{37}$$

Consider $V^{[k]}$ as

$$V^{[k]}=\mathbb{E}_{\substack{b_1\sim\text{bin}(\phi(b))\\ b_2\sim\text{bin}(\phi(b_1))\\ \cdots\\ b_k\sim\text{bin}(\phi(b_{k-1}))\\ b_{k+1}\sim b_k\\ \cdots}}\left[r_\phi(\phi(b_1),a_1)+\gamma r_\phi(\phi(b_2),a_2)+\gamma^2 r_\phi(\phi(b_3),a_3)+\cdots\right]\tag{38}$$

Then for $\forall b$,

$$|V^{[k+1]}(b) - V^{[k]}(b)| \tag{39}$$

$$= \Bigg| \mathbb{E}_{\substack{b_1\sim \mathrm{bin}(\phi(b)) \\ b_2\sim \mathrm{bin}(\phi(b_1)) \\ \cdots \\ b_k\sim \mathrm{bin}(\phi(b_{k-1})) \\ b_{k+1}\sim b_k \\ b_{k+2}\sim b_{k+1}\cdots}} [\gamma^k V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_{k+1})] -$$

$$\mathbb{E}_{\substack{b_1\sim \mathrm{bin}(\phi(b)) \\ b_2\sim \mathrm{bin}(\phi(b_1)) \\ \cdots \\ b_k\sim \mathrm{bin}(\phi(b_{k-1})) \\ b_{k+1}\sim \mathrm{bin}(\phi(b_k)) \\ b_{k+2}\sim b_{k+1}\cdots}} [\gamma^k r_\phi(\phi(b_{k+1}), a) + \gamma^{k+1} V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_{k+2})] \Bigg| \tag{40}$$

$$= \Bigg| \mathbb{E}_{\substack{b_1\sim \mathrm{bin}(\phi(b)) \\ b_2\sim \mathrm{bin}(\phi(b_1)) \\ \cdots \\ b_k\sim \mathrm{bin}(\phi(b_{k-1})) \\ b_{k+1}\sim b_k \\ b_{k+2}\sim b_{k+1}\cdots}} [\gamma^k V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_{k+1})] -$$

$$\mathbb{E}_{\substack{b_1\sim \mathrm{bin}(\phi(b)) \\ b_2\sim \mathrm{bin}(\phi(b_1)) \\ \cdots \\ b_k\sim \mathrm{bin}(\phi(b_{k-1})) \\ b_{k+1}\sim \mathrm{bin}(\phi(b_k)) \\ b_{k+2}\sim b_{k+1}\cdots}} [\gamma^k r_\phi(\phi(b_{k+1}), a) - \gamma^k r(b_{k+1}, a) + \gamma^k V_{\text{true}}^{[\pi_\phi]_{\text{true}}}(b_{k+1})] \Bigg| \tag{41}$$

$$\leq \gamma^k R_{\max}\varepsilon + \frac{\gamma^k}{1-\gamma} R_{\max} L_V \varepsilon \tag{42}$$

where the last inequality used the Lipchitz of value function since the next belief is sampled from the same bin and thus close enough.

Finally, we do the chaining, and sums up all the $V^{[k+1]} - V^{[k]}$ to get for $\forall \phi(b)$,

$$|\tilde{V}_{\text{bin}}^{\pi_\phi}(\phi(b)) - V_{\text{bin}}^{\pi_\phi}(\phi(b))| \tag{43}$$

$$= \left| \sum_{k=1}^{\infty} \left(V^{[k+1]}(b) - V^{[k]}(b) \right) \right| \tag{44}$$

$$\leq \sum_{k=1}^{\infty} \left| \gamma^k R_{\max}\varepsilon + \frac{\gamma^k}{1-\gamma} R_{\max} L_V \varepsilon \right| \tag{45}$$

$$\leq \frac{R_{\max}\varepsilon}{1-\gamma} + \frac{R_{\max} L_V \varepsilon}{(1-\gamma)^2} \tag{46}$$

□

Before ending this part, we'll need to fill the gap between the target policy and the abstracted policy to which the target policy descended. This is handled by the following theorem.

Theorem 3. *The following two inequalities hold simultaneously.*

$$\|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq \frac{R_{\max} L_\pi \varepsilon}{1-\gamma} + 2|\mathcal{O}||\mathcal{A}| \frac{\gamma R_{\max}}{(1-\gamma)^2}(1 + L_\pi)\varepsilon, \tag{47}$$

$$\|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq \frac{R_{\max} L_\pi \varepsilon}{1-\gamma} + |\mathcal{O}| \frac{\gamma R_{\max}}{(1-\gamma)^2}(1 + L_\pi)\varepsilon + |\mathcal{O}||\mathcal{A}| \frac{\gamma(1 + L_\pi) L_V \varepsilon^2}{1-\gamma}. \tag{48}$$

Proof. Using the fact that $V_{\text{true}}^{[\pi_\phi]_{\text{true}}} = \mathcal{T}^{[\pi_\phi]_{\text{true}}} V_{\text{true}}^{[\pi_\phi]_{\text{true}}}$,

$$\begin{aligned} \|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty &= \|V_{\text{true}}^{\pi} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V_{\text{true}}^{\pi} + \mathcal{T}^{[\pi_\phi]_{\text{true}}} V_{\text{true}}^{\pi} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \\ &\leq \|V_{\text{true}}^{\pi} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V_{\text{true}}^{\pi}\|_\infty + \gamma \|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty. \end{aligned} \tag{49}$$

Consequently,

$$\|V_{\text{true}}^{\pi} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty \leq \frac{1}{1-\gamma} \|V_{\text{true}}^{\pi} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V_{\text{true}}^{\pi}\|_\infty. \tag{50}$$

For any b, we have

$$\begin{aligned}
&|(V^\pi_{\text{true}} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V^\pi_{\text{true}})(b)| \\
&= |(\mathcal{T}^\pi V^\pi_{\text{true}} - \mathcal{T}^{[\pi_\phi]_{\text{true}}} V^\pi_{\text{true}})(b)| \\
&= \left| \mathbb{E}_{\substack{a\sim\pi(b) \\ b^{+1}\sim P(\cdot|b)}} \left[r + \gamma V^\pi_{\text{true}}(b^{+1}) \right] - \mathbb{E}_{\substack{a\sim\pi_\phi(\phi(b)) \\ b^{+1}\sim P(\cdot|b)}} \left[r + \gamma V^\pi_{\text{true}}(b^{+1}) \right] \right|
\end{aligned} \tag{51}$$

We first look at r,

$$\begin{aligned}
|\mathbb{E}_{a\sim\pi(b)}[r] - \mathbb{E}_{a\sim\pi_\phi(\phi(b))}[r]| &= |\mathbb{E}_{a\sim\pi(b)}[r] - \mathbb{E}_{a\sim\pi(\phi(b))}[r]| \\
&\le R_{\max} L_\pi \varepsilon
\end{aligned} \tag{52}$$

Then we look at V^π_{true},

$$\begin{aligned}
&\left| \mathbb{E}_{\substack{a\sim\pi(b) \\ b^{+1}\sim P(\cdot|b)}} \left[\gamma V^\pi_{\text{true}}(b^{+1}) \right] - \mathbb{E}_{\substack{a\sim\pi_\phi(\phi(b)) \\ b^{+1}\sim P(\cdot|b)}} \left[\gamma V^\pi_{\text{true}}(b^{+1}) \right] \right| \\
&= \gamma \left| \sum_{o\in\mathcal{O}} \sum_{a\in\mathcal{A}} \left[\left(P(b^{o,a}|b) - P(\phi(b)^{o,a}|\phi(b)) \right) \cdot V^\pi(b^{o,a}) \right] \right| \\
&\le \gamma \sum_{o\in\mathcal{O}} \left| \sum_{a\in\mathcal{A}} \left[\left(P(b^{o,a}|b) - P(\phi(b)^{o,a}|\phi(b)) \right) \cdot V^\pi(b^{o,a}) \right] \right| \\
&\le 2|\mathcal{O}||\mathcal{A}| \frac{\gamma R_{\max}}{1-\gamma}(1+L_\pi)\varepsilon \wedge \left(|\mathcal{O}| \frac{\gamma R_{\max}}{1-\gamma}(1+L_\pi)\varepsilon + \gamma |\mathcal{O}||\mathcal{A}|(1+L_\pi) L_V \varepsilon^2 \right)
\end{aligned} \tag{53}$$

where we used Proposition 1 or Proposition 2 combined with Assumption 7. [**Youheng:** We can't directly apply Assumption 6 here.] □

4.2 Algorithm on Belief Space MDP

Consider a Bellman error minimization algorithm using double sampling, whose optimization target can be written as

$$\hat{Q}^\pi = \arg\min_{f\in\mathcal{F}} \mathcal{E}(f, \pi) \tag{54}$$

where

$$\mathcal{E}(f,\pi) = \mathbb{E}_{\mathcal{D}}[(f(b,a) - (r + \gamma f(b'_A, \pi)))(f(b,a) - (r + \gamma f(b'_B, \pi)))]. \tag{55}$$

On the abstracted space, using the same data, the algorithm becomes

$$\hat{Q}^\pi_\phi = \arg\min_{f\in\mathcal{F}} \mathcal{E}_\phi(f, \pi) \tag{56}$$

where

$$\mathcal{E}_\phi(f,\pi) = \mathbb{E}_{\mathcal{D}}[(f(\phi(b),a) - (r_\phi + \gamma f(\phi(b'_A), \pi_\phi)))(f(\phi(b),a) - (r_\phi + \gamma f(\phi(b'_B), \pi_\phi)))]. \tag{57}$$

Despite the algorithm was computed in the true system, we consider a virtually executed algorithm, and adopts the following standard covering assumption.

Assumption 8. ***(Binned Policy Coverage)*** $\|d^{\pi_\phi}/d^{\mathcal{D}}\|_\infty \le C_\pi(\phi) < \infty$

Note that the coverage value is dependent of the abstraction mapping ϕ. The the best behaving data collection distribution $d^{\mathcal{D}}$ has a worst case coverage would scale as $|\mathcal{C}_\varepsilon|$, which indicates the covering number for ε.

Lemma 4. *In the binned system, we have the following telescoping error*

$$|J_{\hat{Q}^\pi_\phi}(\pi_\phi) - J(\pi_\phi)| \le \frac{\sqrt{C_\pi(\phi)}}{1-\gamma} \cdot \sqrt{\mathbb{E}_{d^{\mathcal{D}}}[(\hat{Q}^\pi_\phi - \mathcal{T}^{\pi_\phi}\hat{Q}^\pi_\phi)^2]} \tag{58}$$

[**Youheng:** The proof is standard textbook so I omitted it here.] Using Hoeffding's inequality, we get

Lemma 5. *With probability at least* $1-\delta$*, for* $\forall f \in \mathcal{F}$*,*

$$|\mathcal{E}_\phi(f,\pi) - \mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(f,\pi)]| \leq \sqrt{\frac{R_{\max}^2}{2n(1-\gamma)^2} \cdot \log \frac{2|\mathcal{F}|}{\delta}} \tag{59}$$

And we have the standard Bellman completeness assumption

Assumption 9. ***(Bellman Completeness)*** $\forall f \in \mathcal{F}, \mathcal{T}^\pi f \in \mathcal{F}$.

Consequently, we have

$$|\mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(\hat{Q}_\phi^\pi,\pi)]| \leq \sqrt{\frac{2R_{\max}^2}{n(1-\gamma)^2} \cdot \log \frac{2|\mathcal{F}|}{\delta}} \tag{60}$$

Theorem 4. *Under Assumption 9,*

$$|J_{\hat{Q}_\phi^\pi}(\pi_\phi) - J(\pi_\phi)| \leq \frac{\sqrt{C_\pi(\phi)}}{1-\gamma} \cdot \left(\frac{2R_{\max}^2}{n(1-\gamma)^2} \cdot \log \frac{2|\mathcal{F}|}{\delta}\right)^{\frac{1}{4}} \tag{61}$$

[**Youheng:** The proof is standard textbook so I omitted it here. Note that there times an extra factor 2. This may not be the best rate, but I'll leave it here for simplicity.]

[**Youheng:** Potentially, the analysis can be extended to other algorithms such as double-robust or MIS, and the analysis will also be quite standard, so I'll keep it this way.]

4.3 Gap Between True Algorithm and Virtual Algorithm

Noticed that we previously assumed the Lipchitz continuity of value function, whose equivalence to the Lipchitz continuity of Q-function at action a can be easily proven. We now assume the function class $\mathcal{F}$ we use to approximate Q-function is also Lipchitz with regard to belief state.

Assumption 10. ***(Lipchitz of Function Class)*** $\exists L_Q$, $\forall f \in \mathcal{F}$, $\forall a \in \mathcal{A}$, $|f(b_1,a) - f(b_2,a)| \leq L_Q \|b_1 - b_2\|_1$.

With the assumption on the function class, we can therefore control the differences between $\mathcal{E}_\phi(f,\pi)$ and $\mathcal{E}(f,\pi)$ for the very same fixed $f \in \mathcal{F}$, which is stated in the lemma. [**Youheng:** Before that, there's some problem with common abstraction literature that I'd like to point out. There's a little difference in setting between my Theorem 1 and standard abstraction. I'll draw a graph to indicate the subtle relation and difference.]

Lemma 6.

$$|\mathcal{E}(f,\pi) - \mathcal{E}_\phi(f,\pi)| \leq \frac{4R_{\max}}{1-\gamma} \cdot \left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon \tag{62}$$

Proof.

$$\begin{aligned}
&|\mathcal{E}(f,\pi) - \mathcal{E}_\phi(f,\pi)| \\
=\ &|\mathbb{E}_\mathcal{D}[(f(b,a) - (r+\gamma f(b'_A,\pi)))(f(b,a) - (r+\gamma f(b'_B,\pi)))] - \\
&\quad \mathbb{E}_\mathcal{D}[(f(\phi(b),a) - (r_\phi+\gamma f(\phi(b'_A),\pi_\phi)))(f(\phi(b),a) - (r_\phi+\gamma f(\phi(b'_B),\pi_\phi)))]| \\
\leq\ &|\mathbb{E}_\mathcal{D}[\{(f(b,a) - f(\phi(b),a)) - (r(b,a) - r_\phi(\phi(b),a)) - \gamma(f(b'_A,\pi) - f(\phi(b'_A),\pi_\phi)\} \\
&\qquad \cdot (f(b,a) - (r+\gamma f(b'_B,\pi)))]| + \\
&|\mathbb{E}_\mathcal{D}[\{(f(b,a) - f(\phi(b),a)) - (r(b,a) - r_\phi(\phi(b),a)) - \gamma(f(b'_B,\pi) - f(\phi(b'_B),\pi_\phi)\} \\
&\qquad \cdot (f(b,a) - (r+\gamma f(b'_A,\pi)))]|.
\end{aligned} \tag{63}$$

Using the fact that

$$\begin{aligned}
&|f(b,\pi) - f(\phi(b),\pi_\phi)| \\
=\ &|\mathbb{E}_{\pi(a|b)}[f(b,a)] - \mathbb{E}_{\pi(a|\phi(b))}[f(\phi(b),a)]| \\
\leq\ &|\mathbb{E}_{\pi(a|b)}[f(b,a)] - \mathbb{E}_{\pi(a|\phi(b))}[f(b,a)]| + |\mathbb{E}_{\pi(a|\phi(b))}[f(b,a)] - \mathbb{E}_{\pi(a|\phi(b))}[f(\phi(b),a)]| \\
\leq\ &\frac{R_{\max}}{1-\gamma}\varepsilon + L_Q\varepsilon
\end{aligned} \tag{64}$$

we have

$$\begin{aligned}&|\mathcal{E}(f,\pi)-\mathcal{E}_\phi(f,\pi)|\\ &\leq 2\cdot\frac{2R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q+\frac{R_{\max}}{1-\gamma}\right)\varepsilon\end{aligned}\tag{65}$$

□

Then, we look at how $\hat{Q}^\pi$ and $\hat{Q}^\pi_\phi$ differs on the very same empirical bellman error $\mathcal{E}_\phi(\cdot,\pi)$.

Theorem 5.

$$|\mathcal{E}_\phi(\hat{Q}^\pi,\pi)-\mathcal{E}_\phi(\hat{Q}^\pi_\phi,\pi)|\leq\frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q+\frac{R_{\max}}{1-\gamma}\right)\varepsilon\tag{66}$$

Proof.

$$\begin{aligned}&\mathcal{E}_\phi(\hat{Q}^\pi,\pi)-\mathcal{E}_\phi(\hat{Q}^\pi_\phi,\pi)+\mathcal{E}(\hat{Q}^\pi_\phi,\pi)-\mathcal{E}(\hat{Q}^\pi,\pi)\\ &=\mathcal{E}_\phi(\hat{Q}^\pi,\pi)-\mathcal{E}(\hat{Q}^\pi,\pi)+\mathcal{E}(\hat{Q}^\pi_\phi,\pi)-\mathcal{E}_\phi(\hat{Q}^\pi_\phi,\pi)\\ &\leq|\mathcal{E}_\phi(\hat{Q}^\pi,\pi)-\mathcal{E}(\hat{Q}^\pi,\pi)|+|\mathcal{E}(\hat{Q}^\pi_\phi,\pi)-\mathcal{E}_\phi(\hat{Q}^\pi_\phi,\pi)|\\ &\leq\frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q+\frac{R_{\max}}{1-\gamma}\right)\varepsilon\end{aligned}\tag{67}$$

where we employ Lemma 6 for the last inequality.

Using the fact that

$$\mathcal{E}_\phi(\hat{Q}^\pi,\pi)-\mathcal{E}_\phi(\hat{Q}^\pi_\phi,\pi)\geq 0\tag{68}$$

$$\mathcal{E}(\hat{Q}^\pi_\phi,\pi)-\mathcal{E}(\hat{Q}^\pi,\pi)\geq 0,\tag{69}$$

we have

$$|\mathcal{E}_\phi(\hat{Q}^\pi,\pi)-\mathcal{E}_\phi(\hat{Q}^\pi_\phi,\pi)|\leq\frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q+\frac{R_{\max}}{1-\gamma}\right)\varepsilon\tag{70}$$

□

To put things together, we have

Theorem 6.

$$|\mathbb{E}_{d^\mathcal{D}}[\mathcal{E}_\phi(\hat{Q}^\pi,\pi)]|\leq\sqrt{\frac{2R^2_{\max}}{n(1-\gamma)^2}\cdot\log\frac{2|\mathcal{F}|}{\delta}}+\frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q+\frac{R_{\max}}{1-\gamma}\right)\varepsilon\tag{71}$$

And consequently,

Theorem 7.

$$|J_{\hat{Q}^\pi}(\pi_\phi)-J(\pi_\phi)|\leq\frac{\sqrt{C_\pi(\phi)}}{1-\gamma}\cdot\sqrt{\sqrt{\frac{2R^2_{\max}}{n(1-\gamma)^2}\cdot\log\frac{2|\mathcal{F}|}{\delta}}+\frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q+\frac{R_{\max}}{1-\gamma}\right)\varepsilon}\tag{72}$$

Theorem 8.

$$|J(\pi_\phi)-J(\pi)|\leq\frac{(L_H+L_\pi+1)R_{\max}\varepsilon}{1-\gamma}+\frac{R_{\max}\varepsilon}{1-\gamma\eta}+\frac{R_{\max}L_V\varepsilon}{(1-\gamma)^2}+2|\mathcal{O}||\mathcal{A}|\frac{\gamma R_{\max}}{(1-\gamma)^2}(1+L_\pi)\varepsilon\tag{73}$$

Proof.

$$
\begin{aligned}
&|J(\pi_\phi) - J(\pi)| && (74)\\
&= \mathbb{E}_{b\sim d_0}[V_{\text{bin}}^{\pi_\phi}(\phi(b)) - V_{\text{true}}^{\pi}(b)]\\
&\leq \|[V_{\text{bin}}^{\pi_\phi}]_{\text{true}} - V_{\text{true}}^{\pi}\|_\infty\\
&\leq \|[V_{\text{bin}}^{\pi_\phi}]_{\text{true}} - [\tilde{V}_{\text{bin}}^{\pi_\phi}]_{\text{true}}\|_\infty + \|[\tilde{V}_{\text{bin}}^{\pi_\phi}]_{\text{true}} - V_{\text{true}}^{[\pi_\phi]_{\text{true}}}\|_\infty + \|V_{\text{true}}^{[\pi_\phi]_{\text{true}}} - V_{\text{true}}^{\pi}\|_\infty\\
&\leq \frac{(L_H + L_\pi + 1)R_{\max}\varepsilon}{1-\gamma} + \frac{R_{\max}\varepsilon}{1-\gamma\eta} + \frac{R_{\max}L_V\varepsilon}{(1-\gamma)^2} + 2|\mathcal{O}||\mathcal{A}|\frac{\gamma R_{\max}}{(1-\gamma)^2}(1+L_\pi)\varepsilon && (75)
\end{aligned}
$$

□

Theorem 9.

$$
|J_{\hat{Q}^\pi}(\pi) - J_{\hat{Q}^\pi}(\pi_\phi)| \leq \frac{R_{\max}}{1-\gamma}\varepsilon + L_Q\varepsilon \tag{76}
$$

Proof.

$$
\begin{aligned}
&|J_{\hat{Q}^\pi}(\pi) - J_{\hat{Q}^\pi}(\pi_\phi)|\\
&= |\mathbb{E}_{b\sim d_0}[\hat{Q}^\pi(b,\pi)] - \mathbb{E}_{b\sim d_0}[\hat{Q}^\pi(\phi(b),\pi_\phi)]|\\
&= |\mathbb{E}_{b\sim d_0}[\hat{Q}^\pi(b,\pi) - \hat{Q}^\pi(\phi(b),\pi_\phi)]|\\
&\leq \frac{R_{\max}}{1-\gamma}\varepsilon + L_Q\varepsilon && (77)
\end{aligned}
$$

□

Eventually

Theorem 10.

$$
\begin{aligned}
|J_{\hat{Q}^\pi}(\pi) - J(\pi)| \leq \frac{\sqrt{C_\pi(\varepsilon)}}{1-\gamma} \cdot \sqrt{\sqrt{\frac{2R_{\max}^2}{n(1-\gamma)^2}\cdot\log\frac{2|\mathcal{F}|}{\delta}} + \frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon}\\
+ \frac{(L_H + L_\pi + 2)R_{\max}\varepsilon}{1-\gamma} + \frac{R_{\max}\varepsilon}{1-\gamma\eta} + \frac{R_{\max}L_V\varepsilon}{(1-\gamma)^2}\\
+ 2|\mathcal{O}||\mathcal{A}|\frac{\gamma R_{\max}}{(1-\gamma)^2}(1+L_\pi)\varepsilon + L_Q\varepsilon \qquad (78)
\end{aligned}
$$

And we finish the total analysis. Notice that we will need to assume $\sqrt{C_\pi(\varepsilon)\cdot\varepsilon}$ would tend to zero when $\varepsilon \to 0$. In finite horizon setting (Section 7), this is automatically satisfied, but in infinite setting, we will need to assume that covering number has a increasing rate slower than $1/\varepsilon$ when $\varepsilon \ll 1$. Generally, covering number of belief space characterizes the hardness of OPE, and in infinite horizon cases, the analysis would depend on the rate of the covering number and even may not be able to control. But of course, this is assuming the most exploring data sampling distribution , and could be different when the exploring policy is good enough (i.e. can fit the target occupancy better.) [**Youheng:** Can change Lipchitz assumption of policy and value function...but the affect would be limited.]

For a sample complexity guarantee, one can first decide with probability $1-\delta$, the entire error should be below ϵ, then one can set an appropriate ε so that the error ϵ can be balanced onto the terms with ε, and the terms with $1/n$. Take a simple example, suppose

$$
\sqrt{\frac{2R_{\max}^2}{n(1-\gamma)^2}\cdot\log\frac{2|\mathcal{F}|}{\delta}} = 3\cdot\frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon. \tag{79}
$$

Then one may solve the equation

$$
\begin{aligned}
\epsilon = \frac{2\sqrt{C_\pi(\varepsilon)}}{1-\gamma}\cdot\sqrt{\frac{8R_{\max}}{1-\gamma}\cdot\left((1+\gamma)L_Q + \frac{R_{\max}}{1-\gamma}\right)\varepsilon} + \frac{(L_H + L_\pi + 2)R_{\max}\varepsilon}{1-\gamma}\\
+ \frac{R_{\max}\varepsilon}{1-\gamma\eta} + \frac{R_{\max}L_V\varepsilon}{(1-\gamma)^2} + 2|\mathcal{O}||\mathcal{A}|\frac{\gamma R_{\max}}{(1-\gamma)^2}(1+L_\pi)\varepsilon + L_Q\varepsilon \qquad (80)
\end{aligned}
$$

for the optimal $\varepsilon(\epsilon)$. After that, putting $\varepsilon(\epsilon)$ into (79) and one can solve the sample complexity n.

4.4 On the Rate of Covering Number

Consider the true belief state only spreading on a low dimensional manifold, we assume the covering number of the belief state space scales as

$$\mathcal{C}_\varepsilon = C \cdot \frac{1}{a\varepsilon^{\frac{1}{r}} + \varepsilon^d}. \tag{81}$$

Then the best exploring policy has a worst case scaling as

$$C_\pi(\varepsilon) = C \cdot \frac{1}{a\varepsilon^{\frac{1}{r}} + \varepsilon^d}. \tag{82}$$

Then according to (80), we get **[Youheng:** to be continue...]

5 Quick-Forgetting Function Class

In this section, consider a modified version of Assumption 4, which indicate that there exists an η', such that

$$\eta' \|b_1 - b_2\|_1 \leq \|b_1^{o,a} - b_2^{o,a}\|_1 \leq \eta \|b_1 - b_2\|_1$$

A possible example of a possible Lipchitz function class is the Quick-Forgetting function class $\mathcal{F}_q^{[m]}$, so that for $\forall f \in \mathcal{F}_q^{[m]}$,

$$f : \mathcal{H} \to [0, \frac{R_{\max}}{1-\gamma}] \tag{83}$$

$$f : \tau_h \mapsto V(\tau_{h-m:h}) \tag{84}$$

and

$$\forall \tau_h^{[1]}, \tau_h^{[2]},\ |f(\tau_h^{[1]}) - f(\tau_h^{[2]})| \leq L_F \cdot \eta_e^h \ \ s.t.\ \tau_{h-m:h}^{[1]} = \tau_{h-m:h}^{[2]} \tag{85}$$

for some $\eta_e \leq \eta'$.

In the sense that two distinct belief states will be close to each other after the same amount of history $\tau_{h-m:m}$ length m, by a factor of η^m, it is natural that the corresponding value function will be close enough under our assumption. Thus, by mapping two histories with the same m-step tail to the same value will be a good approximation.

With that said, we would also like to check out the Lipchitz guarantee that $\mathcal{F}_q^{[m]}$ provides. We first have the following lemma, which indicates the smoothness of $\mathcal{F}_q^{[m]}$ on the true belief states.

Lemma 7. *$\forall f \in \mathcal{F}_q^{[m]}$, we have*

$$\|f\|_{\text{lip}} = \max_{\substack{h_1, h_2 \in \mathcal{H} \\ h_1 \neq h_2}} \frac{|f(h_1) - f(h_2)|}{\|\mathbf{b}(h_1) - \mathbf{b}(h_2)\|_1} \leq F_m < \infty \tag{86}$$

for some uniform F_m.

Proof. **[Youheng:** Omitted here.] □

We also have the following extension lemma

Lemma 8. *Let $T \subset X$ be two metric spaces with $2 \leq |T| = k \leq \infty$. Let Y be a Banach space, $f : T \to Y$ be a function. Then there exists a function $g : X \to Y$ such that $g|_T = f$ and*

$$\|g\|_{\text{lip}} \leq K \cdot (\log k) \cdot \|f\|_{\text{lip}} \tag{87}$$

where K is an absolute constant.

According the the lemmas above, if $f(\tau_h) = f_{b\to R}(\mathbf{b}(\pi))$, $f_{b\to R}$ can be extended to the entire $\mathbb{R}^{|\mathcal{S}|}$ while being $F_m \cdot K \cdot m \cdot \log(|\mathcal{O}||\mathcal{A}|)$-Lipchitz.

6 Future Dependent Value Function and Belief Space POMDP

[**Youheng:** There's something wrong here, the behaviour policy in the belief space is not memoryless. And we don't need Belief space POMDP actually, directly changing s_h to $\phi(\mathbf{b}(\tau_h))$ will be fine, that way we won't need memoryless condition.]

7 Finite-Horizon Guarantees

7.1 Lipchitz Function Class Guarantee

For the finite horizon POMDP, any value function class is guaranteed to be Lipchitz under Assumption 1 with regard to some worst Lipchitz value.

This is because

$$\|f\|_{\text{lip}} = \max_{\substack{h_1,h_2\in\mathcal{H}\\ h_1\neq h_2}} \frac{|f(h_1)-f(h_2)|}{\|\mathbf{b}(h_1)-\mathbf{b}(h_2)\|_1} \tag{88}$$

is a finite number given that $\mathcal{B}$ is a finite state. Then one can follow similar steps from Section 5 and use Lemma 8 to get an upper bound for the Lipchitz parameter.

7.2 Covering Number Guarantees

For finite $\mathcal{B}$, the covering number is upper bounded by $|\mathcal{B}|$. However this could be exponential.

8 Belief Space Binning for FDVF

[**Youheng:** to be continue...]

References

[1] Kavosh Asadi, Dipendra Misra, and Michael Littman. Lipschitz continuity in model-based reinforcement learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 264–273. PMLR, 10–15 Jul 2018.

[2] Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011.

[3] Omer Gottesman, Kavosh Asadi, Cameron S. Allen, Samuel Lobel, George Konidaris, and Michael Littman. Coarse-grained smoothness for reinforcement learning in metric spaces. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 1390–1410. PMLR, 25–27 Apr 2023.

[4] Qianli Shen, Yan Li, Haoming Jiang, Zhaoran Wang, and Tuo Zhao. Deep reinforcement learning with robust and smooth policy, 2020.

[5] Zongzhang Zhang, Michael Littman, and Xiaoping Chen. Covering number as a complexity measure for pomdp planning and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1853–1859, 2012.