Simplified POMDP Planning with an Alternative Observation Space and Formal Performance Guarantees Supplementary Material

Da Kong[1] and Vadim Indelman[2]

[1] Technion Autonomous Systems Program
[2] Department of Aerospace Engineering
Technion - Israel Institute of Technology, Haifa 32000, Israel
da-kong@campus.technion.ac.il, vadim.indelman@technion.ac.il

This document provides supplementary material to [3]. Therefore, it should not be considered a self-contained document, but instead regarded as an appendix of [3]. Throughout this report, all notations and definitions are with compliance to the ones presented in [3].

1 Proof of Lemma 1

1.1 Preliminary proof

For a given transition model $\mathbb{P}(x_{i+1}|x_i, a_i)$ and a given observation model $\mathbb{P}(z_i|x_i)$, if we assume the history h_i^- and $h_i^{\tau-}$ are known, we can bound the expected state-dependent reward below between the original policy $\pi_i^{\tau Z}$ and a known policy π_i^τ for topology τ as:

$$\Big| \mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}}_{\bar{z}_i|x_i,h_i^-}{}^{\tau Z} \mathop{\mathbb{E}}_{x_{i+1}|x_i,\pi_i^{\tau Z}} r(x_{i+1}) - \mathop{\mathbb{E}}_{x_i|h_i^{\tau-}} \mathop{\mathbb{E}}_{\bar{z}_i|x_i,h_i^{\tau-}}{}^{\tau} \mathop{\mathbb{E}}_{x_{i+1}|x_i,\pi_i^{\tau}} r(x_{i+1}) \Big| \tag{1}$$

$$\leq \max_{\bar{\pi}_i} \Big| \mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}}_{\bar{z}_i|x_i,h_i^-}{}^{\tau} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \bar{\pi}_i(h_i^-, \bar{z}_i^\tau)) r(x_{i+1}) \tag{2}$$

$$- \mathop{\mathbb{E}}_{x_i|h_i^{\tau-}} \mathop{\mathbb{E}}_{\bar{z}_i|x_i,h_i^{\tau-}}{}^{\tau} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \pi_i^\tau) r(x_{i+1}) \Big| \tag{3}$$

Here, the policy $\bar{\pi}_i$ is a mapping : $\mathcal{H}_i^- \times \bar{\mathcal{Z}}_i(\mathcal{H}_i^-, \tau) \mapsto \mathcal{A}$.

Proof. **First, we consider the situation that the belief node uses the alternative observation space and model: $\beta^\tau(h_i^-) = 0$.**

For each $x_i \in \mathcal{X}$, we can find an action a_i that maximize and minimize the expected reward:

$$\int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \pi_i^{\tau Z}) r(x_{i+1}) dx_{i+1} \leq \max_{a_i(x_i) \in \mathcal{A}} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, a_i) r(x_{i+1}) dx_{i+1} \tag{4}$$

$$\int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \pi_i^{\tau_Z}) r(x_{i+1}) dx_{i+1} \geq \min_{a_i(x_i)\in\mathcal{A}} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, a_i) r(x_{i+1}) dx_{i+1} \tag{5}$$

From (4), we have:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}^{\tau_Z}}_{\bar{z}_i|x_i,h_i^-} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \pi_i^{\tau_Z}) r(x_{i+1}) dx_{i+1} \\
&\leq \mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}^{\tau_Z}}_{\bar{z}_i|x_i,h_i^-} \max_{a_i(x_i)\in\mathcal{A}} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, a_i) r(x_{i+1}) dx_{i+1} \qquad (6) \\
&= \mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}^{\tau}}_{\bar{z}_i|x_i,h_i^-} \max_{a_i(x_i)\in\mathcal{A}} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, a_i) r(x_{i+1}) dx_{i+1} \qquad (7)
\end{aligned}$$

Similarly, from (5), we will get:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}^{\tau_Z}}_{\bar{z}_i|x_i,h_i^-} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \pi_i^{\tau_Z}) r(x_{i+1}) dx_{i+1} \\
&\geq \mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}^{\tau_Z}}_{\bar{z}_i|x_i,h_i^-} \min_{a_i(x_i)\in\mathcal{A}} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, a_i) r(x_{i+1}) dx_{i+1} \qquad (8) \\
&= \mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}^{\tau}}_{\bar{z}_i|x_i,h_i^-} \min_{a_i(x_i)\in\mathcal{A}} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, a_i) r(x_{i+1}) dx_{i+1} \qquad (9)
\end{aligned}$$

Since we consider $\bar{z}_i^\tau = o_i$ at the belief node for the topology and the alternative observation model is fully observable, the actions in Equation (7) and (9) can be viewed as a policy: $a_i(x_i) = \pi_i, \pi_i \in \{\pi_i : \mathcal{H}_i^- \times \bar{\mathcal{Z}}_i^\tau \mapsto \mathcal{A}\} := \bar{\Pi}_i(\mathcal{H}_i^-, \bar{\mathcal{Z}}_i^\tau)$.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}^{\tau_Z}}_{\bar{z}_i|x_i,h_i^-} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \pi_i^{\tau_Z}) r(x_{i+1}) dx_{i+1} \\
&\leq \max_{\bar{\pi}_i\in\bar{\Pi}_i(\mathcal{H}_i^-,\bar{\mathcal{Z}}_i^\tau)} \mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}^{\tau}}_{\bar{z}_i|x_i,h_i^-} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \bar{\pi}_i(h_i^-, \bar{z}_i^\tau)) r(x_{i+1}) dx_{i+1} \qquad (10)
\end{aligned}$$

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}^{\tau_Z}}_{\bar{z}_i|x_i,h_i^-} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \pi_i^{\tau_Z}) r(x_{i+1}) dx_{i+1} \\
&\geq \min_{\bar{\pi}_i\in\bar{\Pi}_i(\mathcal{H}_i^-,\bar{\mathcal{Z}}_i^\tau)} \mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}^{\tau}}_{\bar{z}_i|x_i,h_i^-} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \bar{\pi}_i(h_i^-, \bar{z}_i^\tau)) r(x_{i+1}) dx_{i+1} \qquad (11)
\end{aligned}$$

Since different policy $\bar{\pi}_i$ upper and lower bound the expected closed-loop reward, we will also bound the distance between the expected closed-loop reward and a known value by exploring the policy space:

$$\begin{aligned}
&\Big| \mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}^{\tau_Z}}_{\bar{z}_i|x_i,h_i^-} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \pi_i^{\tau_Z}) r(x_{i+1}) dx_{i+1} - \mathop{\mathbb{E}}_{x_i|h_i^{\tau-}} \mathop{\mathbb{E}^{\tau}}_{\bar{z}_i|x_i,h_i^{\tau-}} \mathop{\mathbb{E}}_{x_{i+1}|x_i,\pi_i^\tau} r(x_{i+1}) \Big| \\
&\leq \max_{\bar{\pi}_i\in\bar{\Pi}_i(\mathcal{H}_i^-,\bar{\mathcal{Z}}_i^\tau)} \Big| \mathop{\mathbb{E}}_{x_i|h_i^-} \mathop{\mathbb{E}^{\tau}}_{\bar{z}_i|x_i,h_i^-} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \bar{\pi}_i(h_i^-, \bar{z}_i^\tau)) r(x_{i+1}) dx_{i+1} \\
&\qquad - \mathop{\mathbb{E}}_{x_i|h_i^{\tau-}} \mathop{\mathbb{E}^{\tau}}_{\bar{z}_i|x_i,h_i^{\tau-}} \mathop{\mathbb{E}}_{x_{i+1}|x_i,\pi_i^\tau} r(x_{i+1}) \Big| \qquad (12)
\end{aligned}$$

Then, we consider the belief node uses the original observation space and model: $\beta(h_i^-) = 1$.

We have:

$$\mathop{\mathbb{E}}_{x_i|h_i^- \bar{z}_i|x_i,h_i^-} \mathbb{E}^{\tau_Z} \mathop{\mathbb{E}}_{x_{i+1}|x_i,\pi_i^{\tau_Z}} r(x_{i+1}) = \mathop{\mathbb{E}}_{x_i|h_i^- \bar{z}_i|x_i,h_i^-} \mathbb{E}^{\tau} \mathop{\mathbb{E}}_{x_{i+1}|x_i,\pi_i^{\tau}} r(x_{i+1}) \tag{13}$$

Here, $\pi_i^{\tau Z}$ and π_i^{τ} share the same mapping, $\mathcal{H}_i^- \times \mathcal{Z}_i \mapsto \mathcal{A}$. We can say that the optimal and worst π_i^{τ} can upper and lower bound any $\pi_i^{\tau Z}$. Thus, there exists a policy $\bar{\pi}_i^{\tau}$ that can bound the distance in Equation (12). □

If we take a further step from the above claim and assume the propagated history h_i^- has the original POMDP topology τ_Z, we will have:

$$\begin{aligned}
&\max_{\bar{\pi}_i \in \bar{\Pi}_i(\mathcal{H}_i^-, \bar{\mathcal{Z}}_i^{\tau})} \Big| \mathop{\mathbb{E}}_{x_i|h_i^- \bar{z}_i|x_i,h_i^-} \mathbb{E}^{\tau} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \bar{\pi}_i^{\tau}) r(x_{i+1}) \\
&\qquad - \mathop{\mathbb{E}}_{x_i|h_i^{\tau-} \bar{z}_i|x_i,h_i^{\tau-}} \mathbb{E}^{\tau} \mathop{\mathbb{E}}_{x_{i+1}|x_i,\pi_i^{\tau}} r(x_{i+1}) \Big| \qquad (14) \\
&\triangleq \Big| \mathop{\mathbb{E}}_{x_i|h_i^- \bar{z}_i|x_i,h_i^-} \mathbb{E}^{\tau} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \bar{\pi}_i^{+}(h_i^-, \bar{z}_i^{\tau})) r(x_i) \\
&\qquad - \mathop{\mathbb{E}}_{x_i|h_i^{\tau-} \bar{z}_i|x_i,h_i^{\tau-}} \mathbb{E}^{\tau} \mathop{\mathbb{E}}_{x_{i+1}|x_i,\pi_i^{\tau}} r(x_{i+1}) \Big| \qquad (15) \\
&= \Big| \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^- \bar{z}_{i-1}|x_{i-1},h_{i-1}^-} \mathbb{E}^{\tau_Z} \mathop{\mathbb{E}}_{x_i|x_{i-1},\pi_{i-1}^{\tau_C} \bar{z}_i|x_i,h_i^-} \mathbb{E}^{\tau} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \bar{\pi}_i^{+}(h_i^-, \bar{z}_i^{\tau})) r(x_i) \\
&\qquad - \mathop{\mathbb{E}}_{x_i|h_i^{\tau-} \bar{z}_i|x_i,h_i^{\tau-}} \mathbb{E}^{\tau} \mathop{\mathbb{E}}_{x_{i+1}|x_i,\pi_i^{\tau}} r(x_i) \Big| \qquad (16)
\end{aligned}$$

i) If both $\beta(h_{i-1}^-) = 0$ **and** $\beta(h_i^-) = 0$, for every possible $x_i \in \mathcal{X}$ and $x_{i-1} \in \mathcal{X}$, we can find an action $a_i^+(x_i)$ and $a_{i-1}^+(x_{i-1})$ such that:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_i|x_{i-1},\pi_{i-1}^{\tau_C}} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \bar{\pi}_i^{+}) r(x_{i+1}) \\
&\leq \max_{a_{i-1}(x_{i-1})} \mathop{\mathbb{E}}_{x_i|x_{i-1},a_{i-1}(x_{i-1})} \max_{a_i(x_i)} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, a_i(x_i)) r(x_{i+1}) \qquad (17)
\end{aligned}$$

Then we have:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^- \bar{z}_{i-1}|x_{i-1},h_{i-1}^-} \mathbb{E}^{\tau_Z} \mathop{\mathbb{E}}_{x_i|x_{i-1},\pi_{i-1}^{\tau_C} \bar{z}_i|x_i,h_i^-} \mathbb{E}^{\tau} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, \bar{\pi}_i^{+}(h_i^-, \bar{z}_i^{\tau})) r(x_{i+1}) \\
&\leq \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^- \bar{z}_{i-1}|x_{i-1},h_{i-1}^-} \max_{a_{i-1}(x_{i-1})} \mathop{\mathbb{E}}_{x_i|x_{i-1},a_{i-1}(x_{i-1}) \bar{z}_i|x_i,\{h_{i-1},a_{i-1}\}} \mathbb{E}^{\tau} \\
&\qquad \max_{a_i(x_i)} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, a_i(x_i)) r(x_{i+1}) \qquad (18) \\
&= \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^- \bar{z}_{i-1}|x_{i-1},h_{i-1}^-} \mathbb{E}^{\tau} \max_{a_{i-1}(x_{i-1})} \mathop{\mathbb{E}}_{x_i|x_{i-1},a_{i-1}(x_{i-1}) \bar{z}_i|x_i,\{h_{i-1}^-, z_{i-1}^{\tau}, a_{i-1}\}} \mathbb{E}^{\tau} \\
&\qquad \max_{a_i(x_i)} \int_{x_{i+1}} \mathbb{P}(x_{i+1}|x_i, a_i(x_i)) r(x_{i+1}) \qquad (19)
\end{aligned}$$

Similarly, we can find the lower bound:

$$\mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^-\bar{z}_{i-1}|x_{i-1},h_{i-1}^-x_i|x_{i-1},\pi_{i-1}^{\tau_C}\bar{z}_i|x_i,h_i^-}\mathbb{E}^{\tau_Z}\mathop{\mathbb{E}}\mathbb{E}^{\tau}\int_{x_{i+1}}\mathbb{P}(x_{i+1}|x_i,\bar{\pi}_i^+(h_i^-,\bar{z}_i^\tau))r(x_{i+1}) \tag{20}$$

$$\geq \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^-\bar{z}_{i-1}|x_{i-1},h_{i-1}^-}\mathbb{E}^{\tau_Z}\min_{a_{i-1}(x_{i-1})}\mathop{\mathbb{E}}_{x_i|x_{i-1},a_{i-1}(x_{i-1})\bar{z}_i|x_i,\{h_i,a_{i-1}\}}\mathbb{E}^{\tau}$$

$$\min_{a_i(x_i)}\int_{x_{i+1}}\mathbb{P}(x_{i+1}|x_i,a_i(x_i))r(x_{i+1}) \tag{21}$$

$$= \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^-\bar{z}_{i-1}|x_{i-1},h_{i-1}^-}\mathbb{E}^{\tau}\min_{a_{i-1}(x_{i-1})}\mathop{\mathbb{E}}_{x_i|x_{i-1},a_{i-1}(x_{i-1})\bar{z}_i|x_i,\{h_i^-,z_{i-1}^\tau,a_{i-1}\}}\mathbb{E}^{\tau}$$

$$\min_{a_i(x_i)}\int_{x_{i+1}}\mathbb{P}(x_{i+1}|x_i,a_i(x_i))r(x_{i+1}) \tag{22}$$

The $a_{i-1}(x_{i-1})$ and $a_i(x_i)$ in the above equations can be viewed as policy respectively, $\pi_{i-1} \in \mathit{\Pi}_{i-1}(\mathcal{H}_{i-1}^-, \bar{\mathcal{Z}}_{i-1}^\tau)$ and $\pi_i \in \mathit{\Pi}_i(\mathcal{H}_{i-1}^-, \bar{\mathcal{Z}}_{i-1}^\tau, \pi_{i-1}, \bar{\mathcal{Z}}_i^\tau)$. Similar to the previous claim, we can say there exists a policy to bound the below distance:

$$\Big|\mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^-\bar{z}_{i-1}|x_{i-1},h_{i-1}^-x_i|x_{i-1},\pi_{i-1}^{\tau_C}\bar{z}_i|x_i,h_i^-}\mathbb{E}^{\tau_Z}\mathop{\mathbb{E}}\mathbb{E}^{\tau}\int_{x_{i+1}}\mathbb{P}(x_{i+1}|x_i,\bar{\pi}_i^+(h_i^-,\bar{z}_i^\tau))r(x_{i+1}) \tag{23}$$

$$-\mathop{\mathbb{E}}_{x_i|h_i^{\tau-}\bar{z}_i|x_i,h_i^{\tau-}x_{i+1}|x_i,\pi_i^\tau}\mathbb{E}^{\tau}\mathop{\mathbb{E}}r(x_{i+1})\Big|$$

$$\leq \max_{\bar{\pi}_{i-1}\in\bar{\mathit{\Pi}}_{i-1}(\mathcal{H}_{i-1}^-,\bar{\mathcal{Z}}_{i-1}^\tau),\bar{\pi}_i(h_{i-1}^-,z_{i-1}^\tau,\bar{\pi}_{i-1}(h_{i-1}^-,z_{i-1}^\tau),z_i^\tau)}$$

$$\Big|\mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^-\bar{z}_{i-1}|x_{i-1},h_{i-1}^-x_i|x_{i-1},\bar{\pi}_{i-1}\bar{z}_i|x_i,h_i^{p-}}\mathbb{E}^{\tau}\mathop{\mathbb{E}}\mathbb{E}^{\tau}\int_{x_{i+1}}\mathbb{P}(x_{i+1}|x_i,\bar{\pi}_i)r(x_{i+1}) \tag{24}$$

$$-\mathop{\mathbb{E}}_{x_i|h_i^{\tau-}\bar{z}_i|x_i,h_i^{\tau-}x_{i+1}|x_i,\pi_i^\tau}\mathbb{E}^{\tau}\mathop{\mathbb{E}}r(x_{i+1})\Big| \tag{25}$$

ii) If $\beta(h_{i-1}^-) = 0$ and $\beta(h_i^-) = 1$, for each possible $x_{i-1} \in \mathcal{X}$ and each $z_i \in \mathcal{Z}_i$:

$$\mathop{\mathbb{E}}_{x_i|x_{i-1},\pi_{i-1}^{\tau_C}\bar{z}_i|x_i}\int_{x_{i+1}}\mathbb{P}(x_{i+1}|x_i,\bar{\pi}_i^+(h_{i-1}^-,z_{i-1}^{\tau_Z},\pi_{i-1}^{\tau_C},z_i^\tau))r(x_{i+1})$$

$$\leq \max_{a_{i-1}(x_{i-1})}\mathop{\mathbb{E}}_{x_i|x_{i-1},a_{i-1}(x_{i-1})\bar{z}_i|x_i}\mathbb{E}\max_{a_i(z_i)}\int_{x_{i+1}}\mathbb{P}(x_{i+1}|x_i,a_i(z_i))r(x_{i+1}) \tag{26}$$

$$\mathop{\mathbb{E}}_{x_i|x_{i-1},\pi_{i-1}^{\tau_C}\bar{z}_i|x_i}\int_{x_{i+1}}\mathbb{P}(x_{i+1}|x_i,\bar{\pi}_i^+(h_{i-1}^-,z_{i-1}^{\tau_Z},\pi_{i-1}^{\tau_C},z_i^\tau))r(x_{i+1})$$

$$\geq \min_{a_{i-1}(x_{i-1})}\mathop{\mathbb{E}}_{x_i|x_{i-1},a_{i-1}(x_{i-1})\bar{z}_i|x_i}\mathbb{E}\min_{a_i(z_i)}\int_{x_{i+1}}\mathbb{P}(x_{i+1}|x_i,a_i(z_i))r(x_{i+1}) \tag{27}$$

The $a_{i-1}(x_{i-1})$ and $a_i(z_i)$ in the above equations can be viewed as policy respectively, $\pi_{i-1}(h_{i-1}^-, z_{i-1}^\tau)$ and $\pi_i(h_{i-1}^-, z_{i-1}^\tau, \pi_{i-1}(h_{i-1}^-, z_{i-1}^\tau), z_i^\tau)$. Then we can get the same result as Equation (25).

iii) If both $\beta(h_{i-1}^-) = 1$ and $\beta(h_i^-) = 0$, the result is similar.

iv) If $\beta(h_{i-1}^-) = 1$ and $\beta(h_i^-) = 1$, the policy $\bar{\pi}_{i-1,i}$ in Equation (25) shares the same mapping as the closed-loop policy $\pi_{i-1,i}^{\tau_C}$. So there exists the optimal and worst policy that upper and lower bound the expected reward, so we can also get the same result as Equation (25).

1.2 Main Proof of Lemma 1

$$\left| \mathop{\mathbb{E}}_{\bar{z}_{1:i}|b_0,\pi^\tau}^{\tau}(r(b_i^\tau)) - \mathop{\mathbb{E}}_{\bar{z}_{1:i}|b_0,\pi^{\tau_Z}}^{\tau_Z}(r(b_i^{\tau_Z})) \right| \tag{28}$$

Assume state-dependent reward

$$= \left| \mathop{\mathbb{E}}_{\bar{z}_{1:i}|b_0,\pi^\tau}^{\tau} \mathop{\mathbb{E}}_{x_i|b_i^\tau} r(x_i) - \mathop{\mathbb{E}}_{\bar{z}_{1:i}|b_0,\pi^{\tau_Z}}^{\tau_Z} \mathop{\mathbb{E}}_{x_i|b_i^{\tau_Z}} (r(x_i)) \right| \tag{29}$$

$$= \left| \mathop{\mathbb{E}}_{\bar{z}_{1:i-1}|b_0,\pi^\tau}^{\tau} \mathop{\mathbb{E}}_{x_i|h_i^{\tau-}} \mathop{\mathbb{E}}_{\bar{z}_i|x_i,h_i^{\tau-}}^{\tau} r(x_i) - \mathop{\mathbb{E}}_{\bar{z}_{1:i-1}|b_0,\pi^{\tau_Z}}^{\tau_Z} \mathop{\mathbb{E}}_{x_i|h_i^{\tau_Z-}} \mathop{\mathbb{E}}_{\bar{z}_i|x_i}^{\tau_Z} (r(x_i)) \right| \tag{30}$$

(Chain rule and cancel z_i)

$$= \left| \mathop{\mathbb{E}}_{\bar{z}_{1:i-1}|b_0,\pi^\tau}^{\tau} \mathop{\mathbb{E}}_{x_i|h_i^{\tau-}} r(x_i) - \mathop{\mathbb{E}}_{\bar{z}_{1:i-1}|b_0,\pi^{\tau_Z}}^{\tau_Z} \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^{\tau_Z} x_i|x_{i-1},\pi_{i-1}^{\tau_C}} (r(x_i)) \right| \tag{31}$$

(Bayes Rule)

$$= \left| \mathop{\mathbb{E}}_{\bar{z}_{1:i-1}|b_0,\pi^\tau}^{\tau} \mathop{\mathbb{E}}_{x_i|h_i^{\tau-}} r(x_i) - \mathop{\mathbb{E}}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau_Z}}^{\tau_Z} \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^{\tau_Z-}} \mathop{\mathbb{E}}_{\bar{z}_{i-1}|x_{i-1},h_{i-1}^{\tau_Z-}}^{\tau_Z} \mathop{\mathbb{E}}_{x_i|x_{i-1},\pi_{i-1}^{\tau_Z}} r(x_i) \right| \tag{32}$$

$$\begin{aligned} = \Bigg| & \mathop{\mathbb{E}}_{\bar{z}_{1:i-1}|b_0,\pi^\tau}^{\tau} \mathop{\mathbb{E}}_{x_i|h_i^{\tau-}} r(x_i) \\ & - \mathop{\mathbb{E}}_{x_0|b_0} \mathop{\mathbb{E}}_{x_1|x_0,\pi_0^{\tau_C}} \mathop{\mathbb{E}}_{\bar{z}_1|x_1,h_1^{p-}}^{\tau_Z} \ldots \mathop{\mathbb{E}}_{x_{i-1}|x_{i-2},\pi_{i-2}^{\tau_C} \bar{z}_{i-1}|x_{i-1},h_{i-1}^{p-} x_i|x_{i-1},\pi_{i-1}^{\tau_Z}} \mathop{\mathbb{E}}^{\tau_Z} \mathop{\mathbb{E}}_{} r(x_i) \Bigg| \end{aligned} \tag{33}$$

Here, $\bar{h}^{\tau-}$ is propagated history with topology τ. It can be easily got by just repeating the process in Section 1.1 from time $t = i$ to the initial time $t = 0$.

$$\begin{aligned} \leq \max_{\bar{\pi}^\tau} \Bigg| & \mathop{\mathbb{E}}_{\bar{z}_{1:i-1}|b_0,\pi^\tau}^{\tau} \mathop{\mathbb{E}}_{x_i|h_i^{\tau-}} r(x_i) \\ & - \mathop{\mathbb{E}}_{x_0|b_0} \mathop{\mathbb{E}}_{x_1|x_0,\bar{\pi}_0^\tau} \mathop{\mathbb{E}}_{\bar{z}_1|x_1,\bar{h}_1^{\tau-}}^{\tau} \ldots \mathop{\mathbb{E}}_{x_{i-1}|x_{i-2},\bar{\pi}_{i-2}^\tau \bar{z}_{i-1}|x_{i-1},\bar{h}_i^{\tau-} x_i|x_{i-1},\bar{\pi}_{i-1}^\tau} \mathop{\mathbb{E}}^{\tau} \mathop{\mathbb{E}}_{} r(x_i) \Bigg| \end{aligned} \tag{34}$$

2 Proof of Lemma 2

Proof. We can look into the difference in the Q function between two different topologies, which is defined as below:

$$\Delta Q(b_0, a_0, \pi^{\tau_Z}, \pi^{\tau}, \tau_Z, \tau) \triangleq |Q_{\tau_Z}^{\pi^{\tau_Z}}(b_0, a_0) - Q_{\tau}^{\pi^{\tau}}(b_0, a_0)| \tag{35}$$

$$= \Big| \sum_{i=1}^{L} \mathop{\mathbb{E}}_{\bar{z}_{1:i}|b_0,\pi^{\tau_Z}}^{\tau_Z}(r(b_i)) - \sum_{i=1}^{L} \mathop{\mathbb{E}}_{\bar{z}_{1:i}|b_0,\pi^{\tau}}^{\tau}(r(b_i)) \Big| \tag{36}$$

(Assume state-dependent reward and chain rule, same as Equation (32))

$$\begin{aligned} = \Big| \sum_{i=1}^{L} \Big[& \mathop{\mathbb{E}}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau_Z}}^{\tau_Z} \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^{\tau_Z -}} \mathop{\mathbb{E}}_{\bar{z}_{i-1}|x_{i-1},h_{i-1}^{\tau_Z -}}^{\tau_Z} \mathop{\mathbb{E}}_{x_i|x_{i-1},\pi_{i-1}^{\tau_Z}} r(x_i) \\ & - \mathop{\mathbb{E}}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau}}^{\tau} \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^{\tau -}} \mathop{\mathbb{E}}_{\bar{z}_{i-1}|x_{i-1},h_{i-1}^{\tau -}}^{\tau} \mathop{\mathbb{E}}_{x_i|x_{i-1},\pi_{i-1}^{\tau}} r(x_i) \Big] \Big| \end{aligned} \tag{37}$$

(Take out the expected reward $r(x_L)$ from the summation)

$$\begin{aligned} = \Big| \sum_{i=1}^{L-1} \Big[& \mathop{\mathbb{E}}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau_Z}}^{\tau_Z} \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^{\tau_Z -}} \mathop{\mathbb{E}}_{\bar{z}_{i-1}|x_{i-1},h_{i-1}^{\tau_Z -}}^{\tau_Z} \mathop{\mathbb{E}}_{x_i|x_{i-1},\pi_{i-1}^{\tau_Z}} r(x_i) \\ & - \mathop{\mathbb{E}}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau}}^{\tau} \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^{\tau -}} \mathop{\mathbb{E}}_{\bar{z}_{i-1}|x_{i-1},h_{i-1}^{\tau -}}^{\tau} \mathop{\mathbb{E}}_{x_i|x_{i-1},\pi_{i-1}^{\tau}} r(x_i) \Big] \\ & + \mathop{\mathbb{E}}_{\bar{z}_{1:L-2}|b_0,\pi^{\tau_Z}}^{\tau_Z} \mathop{\mathbb{E}}_{x_{L-1}|h_{L-1}^{\tau_Z -}} \mathop{\mathbb{E}}_{\bar{z}_{L-1}|x_{L-1},h_{L-1}^{p -}}^{\tau_Z} \mathop{\mathbb{E}}_{x_L|x_{L-1},\pi_{L-1}^{\tau_Z}} r(x_L) \\ & - \mathop{\mathbb{E}}_{\bar{z}_{1:L-2}|b_0,\pi^{\tau}}^{\tau} \mathop{\mathbb{E}}_{x_{L-1}|h_{L-1}^{\tau -}} \mathop{\mathbb{E}}_{\bar{z}_{L-1}|x_{L-1},h_{L-1}^{\tau -}}^{\tau} \mathop{\mathbb{E}}_{x_L|x_{L-1},\pi_{L-1}^{\tau}} r(x_L) \Big| \end{aligned} \tag{38}$$

(Bound the expected reward $r(x_L)$ using the same method in Section 1.1)

$$\begin{aligned} \leq \max_{\bar{\pi}_{L-1}^{\tau} \in \bar{\Pi}_{L-1}(\mathcal{H}_{L-1}^{\tau_Z -}, \bar{\mathcal{Z}}_{L-1}^{\tau})} \Big| \sum_{i=1}^{L-1} \Big[& \mathop{\mathbb{E}}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau_Z}}^{\tau_Z} \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^{\tau_Z -}} \mathop{\mathbb{E}}_{\bar{z}_{i-1}|x_{i-1},h_{i-1}^{\tau_Z -}}^{\tau_Z} \mathop{\mathbb{E}}_{x_i|x_{i-1},\pi_{i-1}^{\tau_Z}} r(x_i) \\ & - \mathop{\mathbb{E}}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau}}^{\tau} \mathop{\mathbb{E}}_{x_{i-1}|h_{i-1}^{\tau -}} \mathop{\mathbb{E}}_{\bar{z}_{i-1}|x_{i-1},h_{i-1}^{\tau -}}^{\tau} \mathop{\mathbb{E}}_{x_i|x_{i-1},\pi_{i-1}^{\tau}} r(x_i) \Big] \\ & + \mathop{\mathbb{E}}_{\bar{z}_{1:L-2}|b_0,\pi^{\tau_Z}}^{\tau_Z} \mathop{\mathbb{E}}_{x_{L-1}|h_{L-1}^{\tau_Z -}} \mathop{\mathbb{E}}_{\bar{z}_{L-1}|x_{L-1},h_{L-1}^{\tau_Z -}}^{\tau} \mathop{\mathbb{E}}_{x_L|x_{L-1},\bar{\pi}_{L-1}^{\tau}} r(x_L) \\ & - \mathop{\mathbb{E}}_{\bar{z}_{1:L-2}|b_0,\pi^{\tau}}^{\tau} \mathop{\mathbb{E}}_{x_{L-1}|h_{L-1}^{\tau -}} \mathop{\mathbb{E}}_{\bar{z}_{L-1}|x_{L-1},h_{L-1}^{\tau -}}^{\tau} \mathop{\mathbb{E}}_{x_L|x_{L-1},\pi_{L-1}^{\tau}} r(x_L) \Big| \end{aligned} \tag{39}$$

(Take out the expected reward $r(x_{L-1})$ from the summation, and denote the policy maximizing the distance as $\bar{\pi}^{+}_{L-1}$)

$$
\begin{aligned}
=\Big| \sum_{i=1}^{L-2} \Big[& \mathop{\mathbb{E}^{\tau_Z}}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau_Z}} \mathop{\mathbb{E}}_{x_{i-1}|h^{\tau_Z^-}_{i-1}\bar{z}_{i-1}|x_{i-1},h^{\tau_Z^-}_{i-1}x_i|x_{i-1},\pi^{\tau_Z}_{i-1}} \mathop{\mathbb{E}^{\tau_Z}}_{} \mathop{\mathbb{E}}_{} r(x_i) \\
& - \mathop{\mathbb{E}^{\tau}}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau}} \mathop{\mathbb{E}}_{x_{i-1}|h^{\tau^-}_{i-1}\bar{z}_{i-1}|x_{i-1},h^{\tau^-}_{i-1}x_i|x_{i-1},\pi^{\tau}_{i-1}} \mathop{\mathbb{E}^{\tau}}_{} \mathop{\mathbb{E}}_{} r(x_i) \ \Big] && (40) \\
& + \mathop{\mathbb{E}^{\tau_Z}}_{\bar{z}_{1:L-3}|b_0,\pi^{\tau_Z}} \mathop{\mathbb{E}}_{x_{L-2}|h^{\tau_Z^-}_{L-2}\bar{z}_{L-2}|x_{L-2},h^{\tau_Z^-}_{L-2}x_{L-1}|x_{L-2},\pi^{\tau_Z}_{L-2}} \mathop{\mathbb{E}}_{} r(x_{L-1}) \\
& - \mathop{\mathbb{E}^{\tau}}_{\bar{z}_{1:L-3}|b_0,\pi^{\tau}} \mathop{\mathbb{E}}_{x_{L-2}|h^{\tau^-}_{L-2}\bar{z}_{L-2}|h^{\tau^-}_{L-2},\tau x_{L-1}|x_{L-2},\pi^{\tau}_{L-2}} \mathop{\mathbb{E}}_{} r(x_{L-1}) && (41) \\
& + \mathop{\mathbb{E}^{\tau_Z}}_{\bar{z}_{1:L-2}|b_0,\pi^{\tau_Z}} \mathop{\mathbb{E}}_{x_{L-1}|h^{\tau_Z^-}_{L-1}\bar{z}_{L-1}|x_{L-1},h^{\tau_Z^-}_{L-1}x_L|x_{L-1},\bar{\pi}^{+}_{L-1}} \mathop{\mathbb{E}}_{} r(x_L) \\
& - \mathop{\mathbb{E}^{\tau}}_{\bar{z}_{1:L-2}|b_0,\pi^{\tau}} \mathop{\mathbb{E}}_{x_{L-1}|h^{\tau^-}_{L-1}\bar{z}_{L-1}|x_{L-1},h^{\tau^-}_{L-1}x_L|x_{L-1},\pi^{\tau}_{L-1}} \mathop{\mathbb{E}}_{} r(x_L) \ \Big| && (42) \\
=\Big| \sum_{i=1}^{L-2} \Big[& \mathop{\mathbb{E}^{\tau_Z}}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau_Z}} \mathop{\mathbb{E}}_{x_{i-1}|h^{\tau_Z^-}_{i-1}\bar{z}_{i-1}|x_{i-1},h^{\tau_Z^-}_{i-1}x_i|x_{i-1},\pi^{\tau_Z}_{i-1}} \mathop{\mathbb{E}^{\tau_Z}}_{} \mathop{\mathbb{E}}_{} r(x_i) \\
& - \mathop{\mathbb{E}^{\tau}}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau}} \mathop{\mathbb{E}}_{x_{i-1}|h^{\tau^-}_{i-1}\bar{z}_{i-1}|x_{i-1},h^{\tau^-}_{i-1}x_i|x_{i-1},\pi^{\tau}_{i-1}} \mathop{\mathbb{E}^{\tau}}_{} \mathop{\mathbb{E}}_{} r(x_i) \Big] && (43) \\
& + \mathop{\mathbb{E}^{\tau_Z}}_{\bar{z}_{1:L-3}|b_0,\pi^{\tau_Z}} \mathop{\mathbb{E}}_{x_{L-2}|h^{\tau_Z^-}_{L-2}\bar{z}_{L-2}|x_{L-2},h^{\tau_Z^-}_{L-2}x_{L-1}|x_{L-2},\pi^{\tau_Z}_{L-2}} \Big(r(x_{L-1}) \\
& \qquad + \mathop{\mathbb{E}^{\tau}}_{\bar{z}_{L-1}|x_{L-1},h^{\tau_Z^-}_{L-1}x_L|x_{L-1},\bar{\pi}^{+}_{L-1}(h^{\tau_Z^-}_{L-1},\bar{z}^{\tau}_{L-1})} \mathop{\mathbb{E}}_{} r(x_L) \Big) && (44) \\
& - \mathop{\mathbb{E}^{\tau}}_{\bar{z}_{1:L-3}|b_0,\pi^{\tau}} \mathop{\mathbb{E}}_{x_{L-2}|h^{\tau^-}_{L-2}\bar{z}_{L-2}|x_{L-2},h^{\tau^-}_{L-2}x_{L-1}|x_{L-2},\pi^{\tau}_{L-2}} \mathop{\mathbb{E}}_{} r(x_{L-1}) \\
& - \mathop{\mathbb{E}^{\tau}}_{\bar{z}_{1:L-2}|b_0,\pi^{\tau}} \mathop{\mathbb{E}}_{x_{L-1}|h^{\tau^-}_{L-1}\bar{z}_{L-1}|x_{L-1},h^{\tau^-}_{L-1}x_L|x_{L-1},\pi^{\tau}_{L-1}} \mathop{\mathbb{E}}_{} r(x_L) \ \Big| && (45)
\end{aligned}
$$

(Bound the term $\mathop{\mathbb{E}}_{x_{L-1}|x_{L-2},\pi^{\tau_Z}_{L-2}}(r(x_{L-1})+\mathop{\mathbb{E}}_{\bar{z}_{L-1}|x_{L-1},h^{\tau_Z-}_{L-1}x_L|x_{L-1},\bar{\pi}^{+}_{L-1}}r(x_L))$ using a similar method as Section 1.1)

$$\leq \max_{\bar{\pi}_{L-2,L-1}}\Big|\sum_{i=1}^{L-2}\Big[\mathop{\mathbb{E}}^{\tau_Z}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau_Z}}\mathop{\mathbb{E}}_{x_{i-1}|h^{\tau_Z-}_{i-1}\bar{z}_{i-1}|x_{i-1},h^{\tau_Z-}_{i-1}x_i|x_{i-1},\pi^{\tau_Z}_{i-1}}\mathop{\mathbb{E}}^{\tau_Z}r(x_i)$$

$$-\mathop{\mathbb{E}}^{\tau}_{\bar{z}_{1:i-2}|b_0,\pi^{\tau}}\mathop{\mathbb{E}}_{x_{i-1}|h^{\tau-}_{i-1}\bar{z}_{i-1}|x_{i-1},h^{\tau-}_{i-1}x_i|x_{i-1},\pi^{\tau}_{i-1}}\mathop{\mathbb{E}}^{\tau}r(x_i)\Big] \tag{46}$$

$$+\mathop{\mathbb{E}}_{\bar{z}_{1:L-3}|,b_0,\pi^{\tau_Z},\tau_Z x_{L-2}|h^{\tau_Z-}_{L-2}\bar{z}_{L-2}|x_{L-2},h^{\tau_Z-}_{L-2}x_{L-1}|x_{L-2},\bar{\pi}_{L-2}}\Big(r(x_{L-1})$$

$$+\mathop{\mathbb{E}}^{\tau}_{\bar{z}_{L-1}|x_{L-1},h^{p-}_{L-1}x_L|x_{L-1},\bar{\pi}_{L-1}}r(x_L)\Big) \tag{47}$$

$$-\mathop{\mathbb{E}}^{\tau}_{\bar{z}_{1:L-3}|b_0,\pi^{\tau}x_{L-2}|h^{\tau-}_{L-2}\bar{z}_{L-2}|x_{L-2},h^{\tau-}_{L-2}x_{L-1}|x_{L-2},\pi^{\tau}_{L-2}}\mathop{\mathbb{E}}^{\tau}r(x_{L-1})$$

$$-\mathop{\mathbb{E}}^{\tau}_{\bar{z}_{1:L-2}|b_0,\pi^{\tau}x_{L-1}|h^{\tau-}_{L-1}\bar{z}_{L-1}|x_{L-1},h^{\tau-}_{L-1}x_L|x_{L-1},\pi^{\tau}_{L-1}}\mathop{\mathbb{E}}^{\tau}r(x_L)\Big| \tag{48}$$

(We can do it in a recursive way)

$$\leq \max_{\bar{\pi}^{\tau}\in\Pi^{\tau}}|Q^{\bar{\pi}^{\tau}}_{\tau_Z}(b_0,a_0)-Q^{\pi^{\tau}}_{\tau}(b_0,a_0)| \tag{49}$$

Which completes the proof of Lemma 2.

3 Proof of Theorem 1

The proof can be simple.

Proof:

$$Q^{\pi^{\tau_Z}}_{\tau'}(b_k,a_k)\leq \min_{\pi^{\tau}\in\Pi^{\tau}}[Q^{\pi^{\tau}}_{\tau}(b_k,a_k)+\delta_Q(b_k,a_k,\pi^{\tau},\tau)] \tag{50}$$

$$=\min_{\pi^{\tau}\in\Pi^{\tau}}[Q^{\pi^{\tau}}_{\tau}(b_k,a_k)+\max_{\bar{\pi}^{\tau}\in\Pi^{\tau}}|Q^{\bar{\pi}^{\tau}}_{\tau}(b_k,a_k)-Q^{\pi^{\tau}}_{\tau}(b_k,a_k)|] \tag{51}$$

$$=\max_{\pi^{\tau}\in\Pi^{\tau}}[Q^{\pi^{\tau}}_{\tau}(b_k,a_k)] \tag{52}$$

The second direction in (21) can be proved similarly. □

4 Proof of Theorem 2

Proof. Let's consider the end of the planning horizon L.

If $\beta^{\tau}(h(b^{-}_{L-1}))=1$, we have:

$$\int_{z_{L-1}}p(z_{L-1}|x_{L-1})\max_{\pi^{\tau}_{L-1}\in\Pi^{\tau}}\int_{x_L}p(x_L|x_{L-1},\pi^{\tau}_{L-1})r(x_L) \tag{53}$$

$$=\int_{z_{L-1}}p(z_{L-1}|x_{L-1})\int_{x_L}p(x_L|x_{L-1},\pi^{\tau_Z*}_{L-1})r(x_L) \tag{54}$$

If $\beta^\tau(h(b^-_{L-1})) = 0$, we have:

$$\min_{\pi^\tau_{L-1}\in\Pi^\tau} \int_{x_L} p(x_L|x_{L-1},\pi^\tau_{L-1})r(x_L) \leq \int_{z_{L-1}} p(z_{L-1}|x_{L-1}) \int_{x_L} p(x_L|x_{L-1},\pi^{\tau_Z *}_{L-1})r(x_L) \tag{55}$$

We denote the left-hand side in Equation (54) and (55) as $LHS1(x_{L-1},\tau)$ and the right-hand side as $RHS1(x_{L-1},\tau_Z)$. In general, we can say $LHS1(x_{L-1},\tau) \leq RHS1(x_{L-1},\tau_Z)$.

Then, we take one more step earlier. Also if $\beta^\tau(h(b^-_{L-2})) = 1$, we have:

$$\begin{aligned}&\int_{z_{L-1}} p(z_{L-1}|x_{L-1}) \max_{\pi^\tau_{L-2}\in\Pi^\tau} \int_{x_{L-1}} p(x_{L-1}|x_{L-2},\pi^\tau_{L-2})[r(x_{L-1}) + LHS1(x_{L-1},\tau)] \\ &\leq \int_{z_{L-2}} p(z_{L-2}|h_{L-2}) \int_{x_{L-2}} p(x_{L-2}|x_{L-2},\pi^{\tau_Z *}_{L-2})[r(x_{L-1}) + RHS1(x_{L-1},\tau)]\end{aligned} \tag{56}$$

And if $\beta^\tau(h(b^-_{L-2})) = 0$, we have:

$$\begin{aligned}&\min_{\pi^\tau_{L-2}\in\Pi^\tau} \int_{x_{L-1}} p(x_{L-1}|x_{L-2},\pi^\tau_{L-2})[r(x_{L-1}) + LHS1(x_{L-1},\tau)] \\ &\leq \int_{z_{L-2}} p(z_{L-2}|h_{L-2}) \int_{x_{L-2}} p(x_{L-2}|x_{L-2},\pi^{\tau_Z *}_{L-2})[r(x_{L-1}) + RHS1(x_{L-1},\tau)]\end{aligned} \tag{57}$$

We also denote the left-hand side of the inequality in Equation (56) and (57) as $LHS2(x_{L-2},\tau)$ and the right-hand side as $RHS2(x_{L-2},\tau_Z)$. In general, we can say $LHS2(x_{L-2},\tau) \leq RHS2(x_{L-2},\tau_Z)$.

The left-hand side is the iterative way to calculate $lb(b_t,\pi_t,\tau)$. If we keep iterate the above process from the end to the beginning of the planning horizon, we will get the inequality at the top of the belief tree: $lb(b_0,a_0,\tau) \leq Q^{\pi^{\tau_Z *}}_{\tau_Z}(b_0,a_0)$.

5 Sparse Sampling

For a random variable $X \sim \mathcal{P}$, we use another sampling-based random variable, $Y \sim \mathcal{Q}$ to approximate the expectation of the original theoretic value. If we obtain some samples $\{y_i\}^N_{i=1}$ from a known distribution $\mathcal{Q}$, the calculation will be:

$$\mathop{\mathbb{E}}_{X\sim\mathcal{P}}[f(X)] = \int f(x)\mathcal{P}(x)dx = \int f(x)\frac{\mathcal{P}(x)}{\mathcal{Q}(x)}\mathcal{Q}(x)dx \simeq \frac{1}{N}\sum_{i=1}^{N}\frac{\mathcal{P}(y_i)}{\mathcal{Q}(y_i)}f(y_i) \tag{58}$$

Usually, the proposed distribution $\mathcal{Q}$ is the known motion model. The weighted sampling can be more efficient, where resampling will create more particles from the samples with higher weights. The belief density is approximated by $b(x) \simeq \frac{\sum_{i=1}^N w^i\delta(x-x^i)}{\sum_{i=1}^N w^i}$.

5.1 Proof of Theorem 3

Theorem 1 (Theorem 3). ***Bounded Estimation Error.*** *For all the depth $d = 0, ..., L-1$ and a, the following concentration bound holds with probability at least $1 - 2|A|(|A|C)^L \exp(\frac{-C\lambda^2}{2V_{\max}^2})$:*

$$\Delta\hat{ub}(b_0, a_0, \tau) \leq \frac{L(L-1)}{2}\lambda, \qquad \Delta\hat{lb}(b_0, a_0, \tau) \leq \frac{L(L-1)}{2}\lambda. \tag{59}$$

For the upper bound:

Proof. We will follow a similar way to prove it as [2, 5].

Firstly, we consider a general situation at a depth of d and try to bound the error of estimating the optimal Q function given a belief b_d and an action a_d:

$$\delta^{\tau}_{\hat{ub}}(d) \triangleq |Q^{\pi^{\tau *}}_{\tau}(b_d, a_d) - \hat{Q}^{\pi^{\tau *}}_{\tau}(b_d, a_d)| \tag{60}$$

$$= \Big| r(b_d, a_d) + \mathop{\mathbb{E}}_{b_{d+1}|b_d,a_d}[V^{\tau *}(b_{d+1})] - r(b_d, a_d) - \frac{1}{C}\sum_{i=1}^{C} \hat{V}^{\tau *}(b'^{[I_i]}_{d+1}) \Big| \tag{61}$$

$$= \Big| \mathop{\mathbb{E}}_{z_{d+1}|b_d,a_d,\tau}[V^{\tau *}(b_{d+1})] - \frac{1}{C}\sum_{i=1}^{C} \hat{V}^{\tau *}(\bar{b}^{\tau,[I_i]}_{d+1}) \Big| \tag{62}$$

$$\leq \Big| \mathop{\mathbb{E}}_{z_{d+1}|b_d,a_d,\tau}[V^{\tau *}(b_{d+1})] - \frac{1}{C}\sum_{i=1}^{C} V^{\tau *}(\bar{b}^{\tau,[I_i]}_{d+1}) \Big| + \Big| \frac{1}{C}\sum_{i=1}^{C} V^{\tau *}(\bar{b}^{\tau,[I_i]}_{d+1}) - \frac{1}{C}\sum_{i=1}^{C} \hat{V}^{\tau *}(\bar{b}^{\tau,[I_i]}_{d+1}) \Big| \tag{63}$$

Here, the propagated next step belief state samples for a given belief tree topology τ is denoted as $\bar{b}^{\tau,[I_i]}_{d+1}$. It is updated by the motion and observation models for the given topology τ:

$$\bar{b}^{\tau,[I_i]}_{d+1} \sim \psi^{\tau}(\bar{b}^{\tau}_{d+1}|b_d, a_d) = \mathbb{1}_{\beta(h^{-}_{d+1})=1}\alpha \cdot \mathbb{P}(z_{d+1}|x_{d+1})\mathbb{P}(x_{d+1}|b_d, a_d) + \mathbb{1}_{\beta(h^{-}_{d+1})=0}\mathbb{P}(x_{d+1}|b_d, a_d) \tag{64}$$

Here, α is a Bayes normalization factor. For the first term in Equation (63), we can directly use the Hoeffding's inequality:

$$\mathbb{P}\Big(\Big| \mathop{\mathbb{E}}_{z_{d+1}|b_d,a_d,\tau}[V^{\tau *}(b_{d+1})] - \frac{1}{C}\sum_{i=1}^{C} V^{\tau *}(b'^{[I_i]}_{d+1}) \Big| \leq \lambda \Big) \geq 1 - 2\exp(\frac{-C\lambda^2}{2V_{\max}^2}) \tag{65}$$

For the second term in Equation (63), we analyze it in an iterative way:

$$\Big| \frac{1}{C}\sum_{i=1}^{C} V^{\tau *}(\bar{b}^{\tau,[I_i]}_{d+1}) - \frac{1}{C}\sum_{i=1}^{C} \hat{V}^{\tau *}(\bar{b}^{\tau,[I_i]}_{d+1}) \Big| = \frac{1}{C}\Big| \sum_{i=1}^{C} Q^{\pi^{\tau *}}_{\tau}(\bar{b}^{\tau,[I_i]}_{d+1}, a^*_{d+1}) - \sum_{i=1}^{C} \hat{Q}^{\pi^{\tau *}}_{\tau}(\bar{b}^{\tau,[I_i]}_{d+1}, a^*_{d+1}) \Big| \tag{66}$$

$$\leq \delta^{\tau}_{\hat{ub}}(d+1) \tag{67}$$

So, the below iterative bound is satisfied with a probability of at least $1 - 2|A|(|A|C)^{L-d}\exp(\frac{-C\lambda^2}{2V_{\max}^2})$:

$$\delta^{\tau}_{\hat{ub}}(d) \leq \lambda + \delta^{\tau}_{\hat{ub}}(d+1) \tag{68}$$

Here, the probability is based on the worst case for iteration, where it requires all the child belief nodes generated are well estimated with the number of $|A|C$ child nodes for each time step.

The estimation error at the end of the planning horizon $L-1$ is:

$$\delta^{\tau}_{\hat{ub}}(L-1) = \lambda \tag{69}$$

Then, we can find out the estimation bound at the top of the given belief tree by calculating it from the bottom to the top by Equation (68).

For the estimation of lower bound:

Proof. Firstly, we consider a general situation at a depth of d and try to bound the error of estimating the lower bound function given a belief b_d and an action a_d:

$$\delta^{\tau}_{lb}(t) \triangleq |lb(b_t, a_t, \tau) - \hat{lb}(b_t, a_t, \tau)| \tag{70}$$

$$\begin{aligned} &= \Big| \mathbb{1}_{\beta(h(b^-_{t+1}))=1} \Big[\mathop{\mathbb{E}}_{z_{t+1}|h(b_t),a_t,\tau} \max_{\pi_{t+1}} lb(b_{t+1}, \pi_{t+1}, \tau) - \frac{1}{C}\sum_{i=1}^{C} \max_{\pi_{t+1}} \hat{lb}(\bar{b}^{I_i}_{t+1}, \pi_{t+1}, \tau) \Big] \\ &\quad + \mathbb{1}_{\beta(h(b^-_{t+1}))=0} \Big[\mathop{\mathbb{E}}_{z_{t+1}|h(b_t),a_t,\tau} \min_{\pi_{t+1}} lb(b_{t+1}, \pi_{t+1}, \tau) - \frac{1}{C}\sum_{i=1}^{C} \min_{\pi_{t+1}} \hat{lb}(\bar{b}^{I_i}_{t+1}, \pi_{t+1}, \tau) \Big] \Big| \end{aligned} \tag{71}$$

$$\begin{aligned} &\leq \Big| \mathbb{1}_{\beta(h(b^-_{t+1}))=1} \Big\{ \Big| \mathop{\mathbb{E}}_{z_{t+1}|h(b_t),a_t,\tau} \max_{\pi_{t+1}} lb(b_{t+1}, \pi_{t+1}, \tau) - \frac{1}{C}\sum_{i=1}^{C} \max_{\pi_{t+1}} lb(\bar{b}^{I_i}_{t+1}, \pi_{t+1}, \tau) \Big| \\ &\quad + \Big| \frac{1}{C}\sum_{i=1}^{C} \max_{\pi_{t+1}} lb(\bar{b}^{I_i}_{t+1}, \pi_{t+1}, \tau) - \frac{1}{C}\sum_{i=1}^{C} \max_{\pi_{t+1}} \hat{lb}(\bar{b}^{I_i}_{t+1}, \pi_{t+1}, \tau) \Big| \Big\} \\ &\quad + \mathbb{1}_{\beta(h(b^-_{t+1}))=0} \Big\{ \Big| \mathop{\mathbb{E}}_{z_{t+1}|h(b_t),a_t,\tau} \min_{\pi_{t+1}} lb(b_{t+1}, \pi_{t+1}, \tau) - \frac{1}{C}\sum_{i=1}^{C} \min_{\pi_{t+1}} lb(\bar{b}^{I_i}_{t+1}, \pi_{t+1}, \tau) \Big| \\ &\quad + \Big| \frac{1}{C}\sum_{i=1}^{C} \min_{\pi_{t+1}} lb(\bar{b}^{I_i}_{t+1}, \pi_{t+1}, \tau) - \frac{1}{C}\sum_{i=1}^{C} \min_{\pi_{t+1}} \hat{lb}(\bar{b}^{I_i}_{t+1}, \pi_{t+1}, \tau) \Big| \Big\} \Big| \end{aligned} \tag{72}$$

Here, we can use a technique similar to the previous proof. We can directly use the Hoeffding inequality to bound the first term in each indicator as:

$$\mathbb{P}\Big(\Big| \mathop{\mathbb{E}}_{z_{t+1}|h(b_t),a_t,\tau} \max_{\pi_{t+1}} lb(b_{t+1}, \pi_{t+1}, \tau) - \frac{1}{C}\sum_{i=1}^{C} \max_{\pi_{t+1}} lb(\bar{b}^{I_i}_{t+1}, \pi_{t+1}, \tau) \Big| \leq \lambda \Big) \geq 1 - 2\exp(\frac{-C\lambda^2}{2V_{\max}^2}) \tag{73}$$

$$\mathbb{P}\Big(\Big|\mathop{\mathbb{E}}_{z_{t+1}|h(b_t),a_t,\tau}\min_{\pi_{t+1}} lb(b_{t+1},\pi_{t+1},\tau)-\frac{1}{C}\sum_{i=1}^{C}\min_{\pi_{t+1}} lb(\bar{b}^{I_i}_{t+1},\pi_{t+1},\tau)\Big|\le\lambda\Big)\ge 1-2\exp\Big(\frac{-C\lambda^2}{2V^2_{\max}}\Big) \tag{74}$$

The second term is also iteratively bounded by $\delta^{\tau}_{lb}(t+1)$. So, combining the two probabilistic bounds together, we can get the iterative bound for the estimator satisfied with a probability of at least $1-2|A|(|A|C)^{L-ts}\exp(\frac{-C\lambda^2}{2V^2_{\max}})$:

$$\delta^{\tau}_{lb}(t)\le(72)\le \mathbb{1}_{\beta(h(b^{-}_{t+1}))=1}\big[\lambda+\delta^{\tau}_{lb}(t+1)\big]+\mathbb{1}_{\beta(h(b^{-}_{t+1}))=0}\big[\lambda+\delta^{\tau}_{lb}(t+1)\big] \tag{75}$$

$$=\lambda+\delta^{\tau}_{lb}(t+1) \tag{76}$$

The estimate error at the end of the planning horizon is:

$$\delta^{\tau}_{lb}(L-1)=\lambda \tag{77}$$

Now, we can find out the estimation bound at the top of the given belief tree by calculating it from the bottom to the top, from $L-1$ to 0.

6 Experiment Settings

6.1 Experiment 1

In our current implementation, the belief tree $\mathbb{T}^{\tau}$ that corresponds to an initial simplified topology τ is constructed by expanding the original observation space only at randomly chosen 15% of the propagated belief nodes and switching the rest to an alternative observation space $\mathcal{O}=\mathcal{X}$ (see (16)), i.e. considering the state space $\mathcal{X}$ instead of the observation space $\mathcal{Z}$. Then, at each iteration, if the condition (12) is not satisfied, we switch to a different (less simplified) topology τ' by turning 5 randomly-chosen propagated belief nodes in τ, that had an alternative observation space, back to the original observation space. Once (12) is satisfied, we are guaranteed to find the optimal action $a^{\star}_{k}$, and we terminate the process.

6.2 Experiment 2

The sparse sampling planner in a particle-belief POMDP setting has three parameters: planning horizon d, number of observation samples K, and number of the weighted state samples N that represent the particle belief. In our experiments, we evaluate our approach and the original sparse sampling solver using two sets of these parameters: $d=3, K=50, N=50$ and $d=3, K=90, N=90$.

6.3 Experiment 3

The beacon navigation problem serves as a common benchmark for evaluating POMDP solvers [1, 4].

The planning environment comprises a robot located at the initial belief, a beacon for generating observations, three obstacles, and a goal to reach. The locations of the robot, beacon, obstacles, and goal are defined in a 2D space as $\mathbf{x} \in \mathbb{R}^2$, $\mathbf{x}_b \in \mathbb{R}^2$, $\mathbf{x}_o \in \mathbb{R}^2$, and $\mathbf{x}_g \in \mathbb{R}^2$, respectively. The observation $\mathbf{z} \in \mathbb{R}^2$ is defined as the relative position with respect to the beacon, employing a Gaussian observation model defined as:

$$P(\mathbf{z}|\mathbf{x}) = \begin{cases} \mathcal{N}(\mathbf{x} - \mathbf{x}_b, I_{2\times 2}/(100||s - x_b||)), & \text{if } ||\mathbf{x} - \mathbf{x}_b|| > 1, \\ \mathcal{N}(\mathbf{x} - \mathbf{x}_b, I_{2\times 2}/100), & \text{otherwise.} \end{cases} \tag{78}$$

The reward is assumed to be state-dependent: $r(b, a) = \mathop{\mathbb{E}}_{x|b}[r(x, a)]$. The state-dependent reward function contains two parts, reward from reaching the goal r_g and penalty from the obstacles r_o, as $r(\mathbf{x}, \mathbf{a}) = r_g(\mathbf{x}, \mathbf{a}) + r_o(\mathbf{x}, \mathbf{a})$. The action $\mathbf{a}$ belongs to the action space $|\mathcal{A}| = \{[1.0, 0.0], [0.0, 1.0], [-1.0, 0.0], [0.0, -1.0]\}$ representing the movement of right, up, left, and down. The reward of reaching the goal is defined as:

$$r_g(\mathbf{x}, \mathbf{a}) = \frac{50}{||\mathbf{x} - \mathbf{x}_g|| + 0.001}. \tag{79}$$

The penalty from entering the nearby area of the three obstacles o_1, o_2, o_3 is defined as:

$$r_o(\mathbf{x}, \mathbf{a}) = -50, \text{ if } ||\mathbf{x} - \mathbf{x}_{o_i}|| \leq 1, \forall i = 1, 2, 3. \tag{80}$$

The motion model follows a Gaussian distribution:

$$P(\mathbf{x}'|\mathbf{x}, \mathbf{a}) = \mathcal{N}(\mathbf{x} + \mathbf{a}, \Sigma_T), \tag{81}$$

with the covariance $\Sigma_T = I_{2\times 2}/100$.

The trajectory of the goal-reaching tasks simulation is presented in Figure 1 below. The belief particles at each step are depicted by small dots of different colors. True positions of the robot are represented by blue dots and lines, while the gray circles indicate the obstacles with a penalty. The star symbolizes the goal to reach. This visualization serves to demonstrate the successful process by which our method reaches the goal. Our goal chooses the same optimal policy as the sparse sampling in the original full POMDP. Table 1 below demonstrates the total planning time for 10 steps until reach the goal using our proposed method and using sparse sampling in the original full problem. Own method achieves a significant speedup during the whole planning process.

Method	Total Planning Time for 10 Steps (s)
Proposed	**7.731**
Full Problem	17.720

Table 1: Comparison of methods for an exact calculation of the Q function.

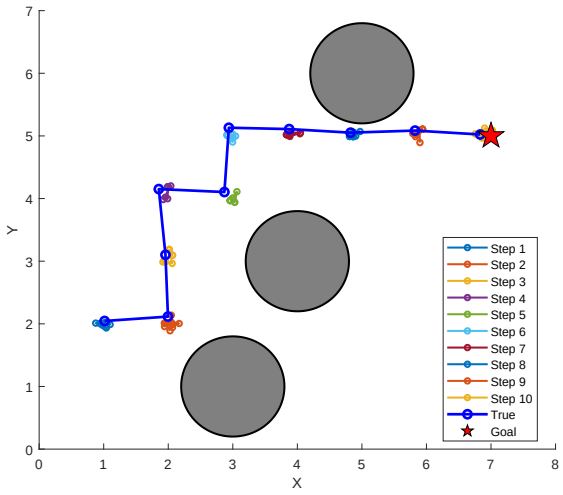

Fig. 1: Simulation Trajectory of the Goal-Reaching Task

6.4 Limitation

Our approach demonstrates effectiveness in scenarios characterized by a large observation space, which is consistent with the theoretical analysis of simplifying policy space. Despite requring multiple iterations, our simplification strategy is still able to reduce computation time.

During the experiment, we also encountered failure cases in which nearly all nodes need to be un-simplified and use the original observation space so that the optimal action at the root could be distinguished. This implies that our method will eventually converge to a topology resembling the original one after numerous iterations, resulting in increased time costs. We address this dilemma as the trade-off between simplification and direct calculation, which all the simplification problems will face.

In addressing this dilemma, it is crucial to have a method that can intelligently and adaptively determine whether a situation is amenable to simplification. By evaluating the potential level of simplification before incurring high costs of explicit calculating, we can prune inadequate simplification candidates and further simplify the iteration process. We will focus on this problem in future work.

References

1. M. Barenboim and V. Indelman. Online pomdp planning with anytime deterministic guarantees. In *Advances in Neural Information Processing Systems (NIPS)*, December 2023.
2. Michael Kearns, Yishay Mansour, and Andrew Y Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. volume 49, pages 193–208. Springer, 2002.

3. D. Kong and V. Indelman. Simplified pomdp with an alternative observation space and formal performance guarantees. In *Proc. of the Intl. Symp. of Robotics Research (ISRR)*, 2024. Submitted.
4. I. Lev-Yehudi, M. Barenboim, and V. Indelman. Simplifying complex observation models in continuous pomdp planning with probabilistic guarantees and practice. In *AAAI Conf. on Artificial Intelligence*, February 2024.
5. Michael H Lim, Tyler J Becker, Mykel J Kochenderfer, Claire J Tomlin, and Zachary N Sunberg. Optimality guarantees for particle belief approximation of pomdps. *Journal of Artificial Intelligence Research*, 77:1591–1636, 2023.