Online POMDP Planning with Anytime Deterministic Guarantees - Supplementary

Anonymous Author(s)
Affiliation
Address
email

This document provides supplementary material to Online POMDP Planning with Anytime Deterministic Guarantees [1] and should not be considered a self-contained document. Throughout this report, all notations and definitions are in compliance with the ones presented in the main body of the paper.

Contents

1 Mathematical Analysis

We start by restating the definition of the simplified value function,

$$\bar{V}^{\pi}(\bar{b}_t) \triangleq r(\bar{b}_t, \pi_t) + \bar{\mathbb{E}}\left[\bar{V}(b_t)\right] \tag{1}$$

$$= \sum_{x_t} \bar{b}(x_t) r(x_t, \pi_t) + \sum_{z_t} \bar{\mathbb{P}}(z_{t+1} \mid H^-_{t+1}) \bar{V}(\bar{b}(z_{t+1})), \tag{2}$$

1.1 Theorem 1

Theorem 1 *Let b_t belief state at time t, and T be the last time step of the POMDP. Let $V^{\pi}(b_t)$ be the theoretical value function by following a policy π, and let $\bar{V}^{\pi}(b_t)$ be the simplified value function, as defined in* (1)*, by following the same policy. Then, for any policy π, the difference between the* *theoretical and simplified value functions is bounded as follows,*

$$\left|V^{\pi}(b_t)-\bar{V}^{\pi}(b_t)\right| \leq \mathcal{R}_{\max}\sum_{\tau=t+1}^{T}\left[1-\sum_{z_{t+1:\tau}}\sum_{x_{t:\tau}} b(x_t)\prod_{k=t+1}^{\tau}\overline{\mathbb{P}}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1},\pi_{k-1})\right] \triangleq \epsilon_z^{\pi}(b_t). \tag{3}$$

Proof 1 *For notational convenience, we derive the bounds for the value function by denoting the prior belief as* b_0*,*

$$V_0^{\pi}(b_0) = \mathbb{E}_{z_{1:T}}\left[\sum_{t=0}^{T} r(b_t, a_t)\right] \tag{4}$$

applying the belief update equation,

$$\begin{aligned} V_0^{\pi}(b_0) &= \sum_{z_{1:T}}\prod_{\tau=1}^{T}\mathbb{P}\left(z_\tau \mid H_\tau^-\right)\sum_{t=0}^{T}\left[\sum_{x_t}\frac{\mathbb{P}(z_t \mid x_t)\sum_{x_{t-1}}\mathbb{P}(x_t \mid x_{t-1},\pi_{t-1})b_{t-1}}{\mathbb{P}\left(z_t \mid H_t^-\right)}r(x_t,a_t)\right] && (5)\\ &= \sum_{z_{1:T}}\prod_{\tau=1}^{T}\mathbb{P}\left(z_\tau \mid H_\tau^-\right)\sum_{t=0}^{T}\left[\sum_{x_{0:t}}\frac{\prod_{k=1}^{t}\mathbb{P}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1},\pi_{k-1})b(x_0)}{\prod_{\tau=1}^{t}\mathbb{P}\left(z_\tau \mid H_\tau^-\right)}r(x_t,a_t)\right] && (6)\\ &= \sum_{t=0}^{T}\sum_{z_{1:T}}\sum_{x_{0:T}}\prod_{k=1}^{t}\mathbb{P}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1},\pi_{k-1})b(x_0)r(x_t,a_t) && (7)\end{aligned}$$

which applies similarly to the simplified value function,

$$\bar{V}_0^{\pi}(b_0) = \sum_{t=0}^{T}\sum_{z_{1:T}}\sum_{x_{0:T}}\prod_{k=1}^{t}\bar{\mathbb{P}}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1},\pi_{k-1})b(x_0)r(x_t,a_t). \tag{8}$$

We begin the derivation by focusing on a single time step, t*, and later generalize to the complete value function.*

$$\begin{aligned} &|\mathbb{E}_{z_{1:t}}[r(b_t)] - \overline{\mathbb{E}}_{z_{1:t}}[r(\bar{b}_t)]| && (9)\\ =&|\sum_{z_{1:t}}\sum_{x_{0:t}}[\prod_{k=1}^{t}\mathbb{P}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1},\pi_{k-1})b(x_0)r(x_t) - \prod_{k'=1}^{t}\overline{\mathbb{P}}(z_{k'} \mid x_{k'})\mathbb{P}(x_{k'} \mid x_{k'-1},\pi_{k'-1})b(x_0)r(x_t)]| && (10)\\ \leq&\sum_{z_{1:t}}\sum_{x_{0:t}}\left|r(x_t)\left[\prod_{k=1}^{t}\mathbb{P}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1},\pi_{k-1})b(x_0) - \prod_{k'=1}^{t}b(x_0)\,\overline{\mathbb{P}}(z_{k'} \mid x_{k'})\mathbb{P}(x_{k'} \mid x_{k'-1},\pi_{k'-1})\right]\right| && (11)\\ =&\sum_{z_{1:t}}\sum_{x_{0:t}}|r(x_t)|\left[\prod_{k=1}^{t}\mathbb{P}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1},\pi_{k-1})b(x_0) - \prod_{k'=1}^{t}b(x_0)\,\overline{\mathbb{P}}(z_{k'} \mid x_{k'})\,\mathbb{P}(x_{k'} \mid x_{k'-1},\pi_{k'-1})\right] && (12)\end{aligned}$$

where the second transition is due to triangle inequality, the third transition is equality by the construction, i.e. using the simplified observation models imply that the difference is nonnegative. We *add and subtract, followed by rearranging terms,*

$$
\begin{aligned}
&= \sum_{z_{1:t}} \sum_{x_{0:t}} |r(x_t)| \\
&\quad [\prod_{k=1}^{t} \mathbb{P}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) b(x_0) - \prod_{k=1}^{t-1} b(x_0) \overline{\mathbb{P}}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) \mathbb{P}(z_t, x_t \mid x_{t-1}, \pi_{t-1}) \\
&\quad + \prod_{k=1}^{t-1} \bar{b}(x_0) \overline{\mathbb{P}}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) \mathbb{P}(z_t, x_t \mid x_{t-1}, \pi_{t-1}) - \prod_{k'=1}^{t} b(x_0) \overline{\mathbb{P}}(z_{k'}, x_{k'} \mid x_{k'-1}, \pi_{k'-1})]
\end{aligned} \tag{13}
$$

$$
\begin{aligned}
&= \sum_{z_{1:t}} \sum_{x_{0:t}} |r(x_t)| \Big\{ \\
&\quad \mathbb{P}(z_t, x_t \mid x_{t-1}, \pi_{t-1}) \left[\prod_{k=1}^{t-1} \mathbb{P}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) b(x_0) - \prod_{k=1}^{t-1} b(x_0) \overline{\mathbb{P}}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) \right] \\
&\quad + \prod_{k=1}^{t-1} \bar{b}(x_0) \overline{\mathbb{P}}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) [\mathbb{P}(z_t, x_t \mid x_{t-1}, \pi_{t-1}) - \overline{\mathbb{P}}(z_t, x_t \mid x_{t-1}, \pi_{t-1})] \Big\}
\end{aligned} \tag{14}
$$

applying Holder's inequality,

$$
\begin{aligned}
&\leq \mathcal{R}_{\max} \sum_{z_{1:t}} \sum_{x_{0:t}} \mathbb{P}(z_t, x_t \mid x_{t-1}, \pi_{t-1}) \left[b(x_0) \prod_{k=1}^{t-1} \mathbb{P}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) - b(x_0) \prod_{k=1}^{t-1} \overline{\mathbb{P}}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) \right] \\
&\quad + \mathcal{R}_{\max} \sum_{z_{1:t}} \sum_{x_{0:t}} \prod_{k=1}^{t-1} \overline{\mathbb{P}}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) b(x_0) [\mathbb{P}(z_t, x_t \mid x_{t-1}, \pi_{t-1}) - \overline{\mathbb{P}}(z_t, x_t \mid x_{t-1}, \pi_{t-1})]
\end{aligned} \tag{15}
$$

$$
\begin{aligned}
&= \mathcal{R}_{\max} \sum_{z_{1:t}} \sum_{x_{0:t}} \mathbb{P}(z_t, x_t \mid x_{t-1}, \pi_{t-1}) \cdot \\
&\quad \left[b(x_0) \prod_{k=1}^{t-1} \mathbb{P}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) - b(x_0) \prod_{k=1}^{t-1} \overline{\mathbb{P}}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) \right] + \mathcal{R}_{\max} \delta_t
\end{aligned} \tag{16}
$$

$$
\begin{aligned}
&= \mathcal{R}_{\max} \sum_{z_{1:t-1}} \sum_{x_{0:t-1}} \left[b(x_0) \prod_{k=1}^{t-1} \mathbb{P}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) - b(x_0) \prod_{k=1}^{t-1} \overline{\mathbb{P}}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) \right] \\
&\quad + \mathcal{R}_{\max} \delta_t,
\end{aligned} \tag{17}
$$

following similar steps recursively,

$$
= \ldots = \mathcal{R}_{\max} \sum_{\tau=1}^{t} \delta_\tau. \tag{18}
$$

Finally, applying similar steps for every time step $t \in [1, T]$ results in,

$$
\left| V^\pi(b_t) - \bar{V}^\pi(b_t) \right| \leq \mathcal{R}_{\max} \sum_{t=1}^{T} \sum_{\tau=1}^{t} \delta_\tau \tag{19}
$$

where,

$$
\begin{aligned}
\delta_\tau &= \sum_{z_{1:\tau}} \sum_{x_{0:\tau}} \prod_{k=1}^{\tau-1} \overline{\mathbb{P}}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) b(x_0) [\mathbb{P}(z_\tau, x_\tau \mid x_{\tau-1}, \pi_{\tau-1}) - \overline{\mathbb{P}}(z_\tau, x_\tau \mid x_{\tau-1}, \pi_{\tau-1})] \\
&= \sum_{z_{1:\tau-1}} \sum_{x_{0:\tau-1}} \prod_{k=1}^{\tau-1} \overline{\mathbb{P}}(z_k, x_k \mid x_{k-1}, \pi_{k-1}) b(x_0) [1 - \sum_{z_\tau} \sum_{x_\tau} \overline{\mathbb{P}}(z_\tau, x_\tau \mid x_{\tau-1}, \pi_{\tau-1})]
\end{aligned} \tag{20}
$$

plugging the term in (20) *to* (19) *and expanding the terms results in the desired bound,*

$$
\left| V^\pi(b_t) - \bar{V}^\pi(b_t) \right| \leq \mathcal{R}_{\max} \sum_{\tau=t+1}^{T} \left[1 - \sum_{z_{t+1:\tau}} \sum_{x_{t:\tau}} b(x_t) \prod_{k=t+1}^{\tau} \overline{\mathbb{P}}(z_k \mid x_k) \mathbb{P}(x_k \mid x_{k-1}, \pi_{k-1}) \right] \tag{21}
$$

which concludes our derivation.

1.2 Lemma 1

Lemma 1 *The optimal value function can be bounded as*

$$V^{\pi*}(b_t) \leq \mathrm{UDB}^{\pi}(b_t), \tag{22}$$

where the policy π is determined according to Bellman optimality over the UDB, i.e.

$$\mathrm{UDB}^{\pi}(b_t) \triangleq \max_{a_t \in \mathcal{A}} [\bar{Q}^{\pi}(b_t, a_t) + \epsilon_z^{\pi}(b_t, a_t)] \tag{23}$$

$$= \max_{a_t \in \mathcal{A}} [r(b_t, a_t) + \bar{\mathbb{E}}_{z_{t+1}|b_t,a_t}[\bar{V}^{\pi}(b_{t+1})] + \epsilon_z^{\pi}(b_t, a_t)]. \tag{24}$$

Proof 2 *In the following, we prove by induction that applying the Bellman optimality operator on upper bounds to the value function in finite-horizon POMDPs will result in an upper bound on the optimal value function. The notations are the same as the ones presented in the main body of the paper. We restate some of the definitions from the paper for convenience.*

The policy $\pi_t(b_t)$ determined by applying Bellman optimality at belief b_t, i.e.,

$$\pi_t(b_t) = \arg\max_{a_t \in \mathcal{A}} [\bar{Q}^{\pi}(b_t, a_t) + \epsilon_z^{\pi}(b_t, a_t)]. \tag{25}$$

As it will be needed in the following proof, we also define the value of a belief which includes in its history at least one observation out of the simplified set, e.g. $H_t = \{a_0, z_1, \dots, z_k \notin \overline{\mathcal{Z}}, \dots, z_t\}$ as being equal to zero. Explicitly,

$$\overline{V}_t^{\pi}(\mathbb{P}(x_t \mid a_0, z_1, \dots, z_k \notin \overline{\mathcal{Z}}, \dots, z_t)) \equiv 0 \;\; \forall k \in [1, t]. \tag{26}$$

We also use the following simple bound,

$$V_{t,\max} \triangleq \mathcal{R}_{\max} \cdot (T - t - 1) \tag{27}$$

Base case $(t = T)$ *- At the final time step T, for each belief we set the value function to be equal to the reward value at that belief state, b_T and taking the action that maximizes the immediate reward,*

$$\mathrm{UDB}^{\pi}(b_T) = \max_{a_T} \{r(b_T, a_T) + \epsilon_z(b_T, a_T)\} = \arg\max_{a_T} \{r(b_T, a_T)\} \tag{28}$$

which provides an upper bound for the optimal value function for the final time step, $V_T^{\star}(b_T) \leq \mathrm{UDB}^{\pi}(b_T)$.

Induction hypothesis *- Assume that for a given time step, t, for all belief states the following holds,*

$$V_t^{\star}(b_t) \leq \mathrm{UDB}^{\pi}(b_t). \tag{29}$$

Induction step *- We will show that the hypothesis holds for time step $t-1$. By the induction hypothesis,*

$$V_t^{\star}(b_t) \leq \mathrm{UDB}^{\pi}(b_t) \;\; \forall b_t, \tag{30}$$

thus,

$$Q^{\star}(b_{t-1}, a_{t-1}) = r(b_{t-1}, a_{t-1}) + \sum_{z_t \in \mathcal{Z}} \mathbb{P}\left(z_t \mid H_t^{-}\right) V_t^{\star}(b(z_t)) \tag{31}$$

$$\leq r(b_{t-1}, a_{t-1}) + \sum_{z_t \in \mathcal{Z}} \mathbb{P}\left(z_t \mid H_t^{-}\right) \mathrm{UDB}^{\pi}(b(z_t)) \tag{32}$$

$$= r(b_{t-1}, a_{t-1}) + \sum_{z_t \in \mathcal{Z}} \mathbb{P}\left(z_t \mid H_t^{-}\right) \left[\overline{V}_t^{\pi}(b_t) + \epsilon_z^{\pi}(b_t)\right]. \tag{33}$$

For the following transition, we make use of lemma 2,

$$= r(b_{t-1}, a_{t-1}) + \overline{\mathbb{E}}_{z_t|b_{t-1},a_{t-1}} \left[\overline{V}_t^{\pi}(b_t)\right] + \epsilon_z^{\pi}(b_{t-1}, a_{t-1}) \tag{34}$$

$$\equiv \mathrm{UDB}^{\pi}(b_{t-1}, a_{t-1}). \tag{35}$$

Therefore, under the induction hypothesis, $Q^{\star}_{t-1}(b_{t-1}, a_{t-1}) \leq \mathrm{UDB}^{\pi}(b_{t-1}, a_{t-1})$. Taking the maximum over all actions a_t,

$$\mathrm{UDB}^{\pi}(b_{t-1}) = \max_{a_{t-1}\in\mathcal{A}} \{\mathrm{UDB}^{\pi}(b_{t-1}, a_{t-1})\} \\ \geq \max_{a_{t-1}\in\mathcal{A}} \left\{Q^{\star}_{t-1}(b_{t-1}, a_{t-1})\right\} = V^{\star}_{t-1}(b_{t-1}), \tag{36}$$

which completes the induction step and the required proof.

Lemma 2 *Let b_t denote a belief state and π_t a policy at time t. Let $\bar{\mathcal{P}}(z_t \mid x_t)$ be the simplified observation model which represents the likelihood of observing z_t given x_t. Then, the following terms are equivalent,*

$$\mathbb{E}_{z_t}\left[\overline{V}^{\pi}_t(b_t) + \epsilon^{\pi}_z(b_t)\right] = \overline{\mathbb{E}}_{z_t}\left[\overline{V}^{\pi}_t(b_t)\right] + \epsilon^{\pi}_z(b_{t-1}, a_{t-1}) \tag{37}$$

Proof 3

$$\mathbb{E}_{z_t}\left[\overline{V}^{\pi}_t(b_t) + \epsilon^{\pi}_z(b_t)\right] = \tag{38}$$

$$\mathbb{E}_{z_t}\left[\overline{V}^{\pi}_t(b_t)\right] + \mathbb{E}_{z_t}\left[\mathcal{R}_{\max}\sum_{\tau=t+1}^{T}\left[1 - \sum_{z_{t+1:\tau}}\sum_{x_{t:\tau}} b_t \prod_{k=t+1}^{\tau}\overline{\mathbb{P}}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1}, \pi_{k-1})\right]\right] \tag{39}$$

focusing on the second summand,

$$\sum_{z_t\in\mathcal{Z}} \mathbb{P}\left(z_t \mid H^-_t\right)\mathcal{R}_{\max}\sum_{\tau=t+1}^{T}\left[1 - \sum_{z_{t+1:\tau}}\sum_{x_{t:\tau}} b_t \prod_{k=t+1}^{\tau}\overline{\mathbb{P}}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1}, \pi_{k-1})\right] \tag{40}$$

$$= \mathcal{R}_{\max}\sum_{\tau=t+1}^{T}\left[1 - \sum_{z_t}\mathbb{P}\left(z_t \mid H^-_t\right)\sum_{z_{t+1:\tau}}\sum_{x_{t:\tau}} b(x_t)\prod_{k=t+1}^{\tau}\overline{\mathbb{P}}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1}, \pi_{k-1})\right] \tag{41}$$

by marginalizing over x_{t-1},

$$= \mathcal{R}_{\max}\sum_{\tau=t+1}^{T}[1 - \sum_{z_t}\mathbb{P}\left(z_t \mid H^-_t\right)\sum_{z_{t+1:\tau}}\sum_{x_{t-1:\tau}}\frac{\overline{\mathbb{P}}(z_t \mid x_t)\mathbb{P}(x_t \mid x_{t-1}, \pi_{t-1})b(x_{t-1})}{\mathbb{P}\left(z_t \mid H^-_t\right)}\cdot \\ \prod_{k=t+1}^{\tau}\overline{\mathbb{P}}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1}, \pi_{k-1})] \tag{42}$$

canceling out the denominator,

$$= \mathcal{R}_{\max}\sum_{\tau=t+1}^{T}[1 - \sum_{z_{t:\tau}}\sum_{x_{t-1:\tau}}\overline{\mathbb{P}}(z_t \mid x_t)\mathbb{P}(x_t \mid x_{t-1}, a_{t-1})b(x_{t-1})\cdot \\ \prod_{k=t+1}^{\tau}\overline{\mathbb{P}}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1}, \pi_{k-1})] \equiv \epsilon^{\pi}_z(b_{t-1}, a_{t-1}) \tag{43}$$

it is left to show that $\mathbb{E}_{z_t|b_{t-1},a_{t-1}}\left[\overline{V}^{\pi}_t(b_t)\right] = \overline{\mathbb{E}}_{z_t|b_{t-1},a_{t-1}}\left[\overline{V}^{\pi}_t(b_t)\right]$. By the definition of a value function of a belief not included in the simplified set, we have that,

$$
\begin{aligned}
\mathbb{E}_{z_t|b_{t-1},a_{t-1}}\left[\overline{V}^{\pi}_{t}(b_t)\right] &= \sum_{z_t\in\mathcal{Z}} \mathbb{P}\left(z_t \mid H_t^{-}\right) \overline{V}^{\pi}_{t}(b_t) && (44)\\
&= \sum_{z_t\in\overline{\mathcal{Z}}} \mathbb{P}\left(z_t \mid H_t^{-}\right) \overline{V}^{\pi}_{t}(b_t) + \sum_{z_t\in\mathcal{Z}\setminus\overline{\mathcal{Z}}} \mathbb{P}\left(z_t \mid H_t^{-}\right) \overline{V}^{\pi}_{t}(b_t) && (45)\\
&= \sum_{z_t\in\overline{\mathcal{Z}}} \overline{\mathbb{P}}\left(z_t \mid H_t^{-}\right) \cdot \overline{V}^{\pi}_{t}(b_t) + \sum_{z_t\in\mathcal{Z}\setminus\overline{\mathcal{Z}}} \mathbb{P}\left(z_t \mid H_t^{-}\right) \cdot 0 && (46)\\
&= \overline{\mathbb{E}}_{z_t|b_{t-1},a_{t-1}}\left[\overline{V}^{\pi}_{t}(b_t)\right], && (47)
\end{aligned}
$$

which concludes the derivation.

1.3 Corollary 1.1

We restate the definition of UDB exploration criteria,

$$a_t = \arg\max_{a_t\in\mathcal{A}}[\mathrm{UDB}^{\pi}(b_t, a_t)] = \arg\max_{a_t\in\mathcal{A}}[\bar{Q}^{\pi}(b_t, a_t) + \epsilon^{\pi}_{z}(b_t, a_t)]. \tag{48}$$

Corollary 1.1 *Using Lemma 1 and the exploration criteria defined in* (48) *guarantees convergence to the optimal value function.*

Proof 4 *Let us define a sequence of bounds,* $\mathrm{UDB}^{\pi}_{n}(b_t)$ *and a corresponding difference value between* UDB_n *and the simplified value function,*

$$\mathrm{UDB}^{\pi}_{n}(b_t) - \bar{V}^{\pi}_{n}(b_t) = \epsilon^{\pi}_{n,z}(b_t), \tag{49}$$

where $n \in [0, |\mathcal{Z}|]$ *corresponds to the number of unique observation instances within the simplified observation set,* $\overline{\mathcal{Z}}_n$*, and* $|\mathcal{Z}|$ *denotes the cardinality of the complete observation space. Additionally, for the clarity of the proof and notations, assume that by construction the simplified set is chosen such that* $\overline{\mathcal{Z}}_n(H_t) \equiv \overline{\mathcal{Z}}_n$ *remains identical for all time steps* t *and history sequences,* H_t *given* n*. By the definition of* $\epsilon^{\pi}_{n,z}(b_t)$*,*

$$\epsilon^{\pi}_{n,z}(b_t) = \mathcal{R}_{\max}\sum_{\tau=t+1}^{T}\left[1 - \sum_{z_{t+1:\tau}\in\overline{\mathcal{Z}}_n}\sum_{x_{t:\tau}} b(x_t)\prod_{k=t+1}^{\tau}\overline{\mathbb{P}}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1}, \pi_{k-1})\right], \tag{50}$$

we have that $\epsilon^{\pi}_{n,z}(b_t) \to 0$ *as* $n \to |\mathcal{Z}|$*, since*

$$\sum_{z_{t+1:\tau}\in\overline{\mathcal{Z}}_n}\sum_{x_{t:\tau}} b(x_t)\prod_{k=t+1}^{\tau}\overline{\mathbb{P}}(z_k \mid x_k)\mathbb{P}(x_k \mid x_{k-1}, \pi_{k-1}) \to 1 \tag{51}$$

as more unique observation elements are added to the simplified observation space, $\overline{\mathcal{Z}}_n$*, eventually recovering the entire support of the discrete observation distribution.*

From lemma 1 we have that, for all $n \in [0, |\mathcal{Z}|]$ *the following holds,*

$$V^{\pi*}(b_t) \leq \mathrm{UDB}^{\pi}_{n}(b_t) = \bar{V}^{\pi}_{n}(b_t) + \epsilon^{\pi}_{n,z}(b_t). \tag{52}$$

Additionally, from theorem 1 we have that,

$$\left|V^{\pi}(b_t) - \bar{V}^{\pi}_{n}(b_t)\right| \leq \epsilon^{\pi}_{n,z}(b_t), \tag{53}$$

for any policy π *and subset* $\overline{\mathcal{Z}}_n \subseteq \mathcal{Z}$*, thus,*

$$\bar{V}^{\pi}_{n}(b_t) - \epsilon^{\pi}_{n,z}(b_t) \leq V^{\pi}(b_t) \leq V^{\pi*}(b_t) \leq \bar{V}^{\pi}_{n}(b_t) + \epsilon^{\pi}_{n,z}(b_t). \tag{54}$$

Since $\epsilon^{\pi}_{n,z}(b_t) \to 0$ *as* $n \to |\mathcal{Z}|$*, and* $|\mathcal{Z}|$ *is finite, it is guaranteed that* $\mathrm{UDB}^{\pi}_{n}(b_t) \xrightarrow{n\to|\mathcal{Z}|} V^{\pi*}(b_t)$ *which completes our proof.*

Moreover, depending on the algorithm implementation, the number of iterations can be finite (e.g. by directly choosing actions and observations to minimize the bound). A stopping criteria can also be verified by calculating the difference between the upper and lower bounds. The optimal solution is obtained once the upper bound equals the lower bound.

2 Experiments

2.1 POMDP scenarios

We begin with a brief description of the Partially Observable Markov Decision Process (POMDP) scenarios implemented for the experiments. each scenario was bounded by a finite number of time steps used for every episode, where each action taken by the agent led to a decrement in the number of time steps left. After the allowable time steps ended, the simulation was reset to its initial state.

2.1.1 Tiger POMDP

The Tiger is a classic POMDP problem [2], involves an agent making decisions between two doors, one concealing a tiger and the other a reward. The agent needs to choose among three actions, either open each one of the doors or listen to receive an observation about the tiger position. In our experiments, the POMDP was limited horizon of 5 steps. The problem consists of 3 actions, 2 observations and 2 states.

2.1.2 Discrete Light Dark

Is an adaptation from [4]. In this setting the agent needs to travel on a 1D grid to reach a target location. The grid is divided into a dark region, which offers noisy observations, and a light region, which offers accurate localization observations. The agent receives a penalty for every step and a reward for reaching the target location. The key challenge is to balance between information gathering by traveling towards the light area, and moving towards the goal region.

2.1.3 Laser Tag POMDP

In the Laser Tag problem, [3], an agent has to navigate through a grid world, shoot and tag opponents by using a laser gun. The main goal is to tag as many opponents as possible within a given time frame. The grid is segmented into various sections that have varying visibility, characterized by obstacles that block the line of sight, and open areas. There are five possible actions, moving in four cardinal directions (North, South, East, West) and shooting the laser. The observation space cardinality is $|\mathcal{Z}| \approx 1.5 \times 10^6$, which is described as a discretized normal distribution and reflect the distance measured by the laser. The states reflect the agent's current position and the opponents' positions. The agent receives a reward for tagging an opponent and a penalty for every movement, encouraging the agent to make strategic moves and shots.

2.1.4 Baby POMDP

The Baby POMDP is a classic problem that represents the scenario of a baby and a caregiver. The agent, playing the role of the caregiver, needs to infer the baby's needs based on its state, which can be either crying or quiet. The states in this problem represent the baby's needs, which could be hunger, discomfort or no need. The agent has three actions to choose from: feeding, changing the diaper, or doing nothing. The observations are binary, either the baby is crying or not. The crying observation does not uniquely identify the baby's state, as the baby may cry due to hunger or discomfort, which makes this a partially observable problem. The agent receives a reward when it correctly addresses the baby's needs and a penalty when the wrong action is taken.

2.2 Hyperparameters

The hyperparameters for both DB-DESPOT and AR-DESPOT algorithms were selected through a grid search. We explored an array of parameters for AR-DESPOT, choosing the highest-performing configuration. Specifically, the hyperparameter K was varied across $\{10, 50, 500, 5000\}$, while λ was evaluated at $\{0, 0.01, 0.1\}$.

For upper and lower bounds used both by DB-DESPOT (which results in deterministic bounds) and AR-DESPOT (which result in probabilistic bounds); we used the maximal reward, multiplied by the remaining time steps of the episode, $\mathcal{R}_{\max} \cdot (\mathcal{T} - t - 1)$.

Finally, we provide our algorithm implementation in *[will be provided upon official publication of the paper]*.

References

[1] Anonymous Author(s). Online pomdp planning with anytime deterministic guarantees. Technical report.

[2] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.

[3] Adhiraj Somani, Nan Ye, David Hsu, and Wee Sun Lee. Despot: Online pomdp planning with regularization. In *NIPS*, volume 13, pages 1772–1780, 2013.

[4] Zachary Sunberg and Mykel Kochenderfer. Online algorithms for pomdps with continuous state, action, and observation spaces. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 28, 2018.