Adaptive Information Belief Space Planning

Moran Barenboim[1] and **Vadim Indelman**[2]
[1]Technion Autonomous Systems Program
[2]Department of Aerospace Engineering
Technion - Israel Institute of Technology, Haifa 32000, Israel
moranbar@campus.technion.ac.il, vadim.indelman@technion.ac.il

A Proofs

A.1 Lemma 1

The proof is provided for continuous state space; The discrete case obtained similarly by changing integrals to summations.

Proof.

$$\sum_{n=1}^{N_o} \bar{\mathbb{P}}\left(o^n \mid H^-\right) \mathcal{H}(\bar{b}) =$$
$$-\sum_{n=1}^{N_o} \bar{\mathbb{P}}\left(o^n \mid H^-\right) \int_s \bar{\mathbb{P}}(s \mid H) \cdot log(\bar{\mathbb{P}}(s \mid H)) \quad (1)$$

applying Bayes' rule for $\bar{\mathbb{P}}(s \mid H)$,

$$-\sum_{n=1}^{N_o} \int_s \bar{Z}(o^n \mid s)\mathbb{P}\left(s \mid H^-\right) \quad (2)$$
$$\cdot log\left(\frac{\bar{Z}(o^n \mid s)\mathbb{P}\left(s \mid H^-\right)}{\int_{s'} \bar{Z}(o^n \mid s')\mathbb{P}\left(s' \mid H^-\right)}\right)$$

Splitting summation to follow the partitioning of the abstract observation model,

$$-\sum_{c=1}^{C} \sum_{k=K(c-1)+1}^{Kc} \int_s \bar{Z}(o^k \mid s)\mathbb{P}\left(s \mid H^-\right) \quad (3)$$
$$\cdot log\left(\frac{\bar{Z}(o^k \mid s)\mathbb{P}\left(s \mid H^-\right)}{\int_{s'} \bar{Z}(o^k \mid s')\mathbb{P}\left(s' \mid H^-\right)}\right)$$

By construction, $\bar{Z}(o \mid s)$ has uniform distribution for o^k, where $k \in [K(c-1)+1, Kc]$. Thus,

$$-\sum_{c=1}^{C} \left[\sum_{k=K(c-1)+1}^{Kc} 1\right] \int_s \bar{Z}(o^{Kc} \mid s)\mathbb{P}\left(s \mid H^-\right) \quad (4)$$
$$\cdot log\left(\frac{\bar{Z}(o^{Kc} \mid s)\mathbb{P}\left(s \mid H^-\right)}{\int_{s'} \bar{Z}(o^{Kc} \mid s')\mathbb{P}\left(s' \mid H^-\right)}\right) =$$
$$-\sum_{c=1}^{C} K \cdot \int_s \bar{Z}(o^{Kc} \mid s)\mathbb{P}\left(s \mid H^-\right) \quad (5)$$
$$\cdot log\left(\frac{\bar{Z}(o^{Kc} \mid s)\mathbb{P}\left(s \mid H^-\right)}{\int_{s'} \bar{Z}(o^{Kc} \mid s')\mathbb{P}\left(s' \mid H^-\right)}\right) =$$
$$K \cdot \sum_{c=1}^{C} \bar{\mathbb{P}}\left(o^{Kc} \mid H^-\right) \mathcal{H}(\bar{b})$$

which concludes the proof. □

A.2 Lemma 2

Proof. We begin with,

$$\bar{\mathbb{E}}_o \left[\mathbb{E}_{s \sim \bar{b}} \left[r\left(s,a\right) \right] \right] \tag{6}$$

by definition of $\bar{\mathbb{E}}_o \left[\cdot\right]$, $\bar{b}(s)$,

$$\sum_{n=1}^{N_o} \bar{\mathbb{P}} \left(o^n \mid H^- \right) \left[\sum_{s \in S} \bar{\mathbb{P}}(s \mid o^n, H^-) r\left(s,a\right) \right] \tag{7}$$

applying chain rule,

$$\sum_{n=1}^{N_o} \sum_{s \in S} \bar{\mathbb{P}}(s, o^n \mid H^-) r\left(s,a\right) = \tag{8}$$

$$\sum_{n=1}^{N_o} \sum_{s \in S} \bar{Z}(o^n \mid s) b^-(s) r\left(s,a\right)$$

we split the sum over the observations to comply with the abstraction partitioning and use the the abstract observation model definition, (4),

$$\sum_{s \in S} \sum_{c=1}^{C} \sum_{k=K(c-1)+1}^{Kc} \frac{\sum_{m=K(c-1)+1}^{Kc} Z(o^m \mid s)}{K} b^-(s) r\left(s,a\right) \tag{9}$$

we then arrive at the desired result,

$$\sum_{s \in S} \sum_{n=1}^{N_o} Z(o^n \mid s) b^-(s) r\left(s,a\right) = \tag{10}$$

$$\mathbb{E}_o \left[\mathbb{E}_{s \sim b} \left[r\left(s,a\right) \right] \right] \tag{11}$$

□

A.3 Theorem 1

Proof. For clarity, we omit the time index in the derivation, the result holds for any time step. We use H^- to denote past history while excluding last observation. We also use b and $\mathbb{P}(s \mid o, H^-)$ interchangeably. Rearranging the abstraction from (1),

$$\sum_{k=1}^{K} \bar{Z}\left(o^k \mid s\right) = K \cdot \bar{Z}(o^b \mid s) \doteq \sum_{k=1}^{K} Z\left(o^k \mid s\right) \quad \forall b \in [1, K]$$

Plugging it to the expected entropy term, (11),

$$\bar{\mathbb{E}}_o \left[\mathcal{H}\left(\bar{b}\right) \right] - \mathbb{E}_o \left[\mathcal{H}\left(b\right) \right] = \tag{12}$$

$$\sum_{i=1}^{N_o} \bar{\mathbb{P}}\left(o_i \mid H^-\right) \mathcal{H}\left(\bar{b}\right) - \sum_{i=1}^{N_o} \mathbb{P}\left(o_i \mid H^-\right) \mathcal{H}\left(b\right) \tag{13}$$

expanding the entropy term,

$$-\sum_{i=1}^{N_o} \bar{\mathbb{P}}\left(o_i \mid H^-\right) \int_s \bar{\mathbb{P}}\left(s \mid o_i, H^-\right) log\left(\bar{b}\right) \tag{14}$$

$$+\sum_{i=1}^{N_o} \mathbb{P}\left(o_i \mid H^-\right) \int_s \mathbb{P}\left(s \mid o_i, H^-\right) log\left(b\right)$$

by Bayes' rule,

$$-\sum_{i=1}^{N_o} \int_s \bar{Z}\left(o_i \mid s\right) \mathbb{P}\left(s \mid H^-\right) log\left(\bar{b}\right) \tag{15}$$

$$+\sum_{i=1}^{N_o} \int_s Z\left(o_i \mid s\right) \mathbb{P}\left(s \mid H^-\right) log\left(b\right)$$

a change in the order of summation and integral and a split of $N_o = C \cdot K$ result in,

$$-\int_s \sum_{c=1}^{C} \sum_{k=K(c-1)+1}^{Kc} \bar{Z}\left(o_k \mid s\right) \mathbb{P}\left(s \mid H^-\right) log\left(\bar{b}\right) \tag{16}$$

$$+\int_s \sum_{c=1}^{C} \sum_{k=K(c-1)+1}^{Kc} Z\left(o_k \mid s\right) \mathbb{P}\left(s \mid H^-\right) log\left(b\right)$$

By plugging-in the definition of the abstract model,

$$-\int_s \sum_{c=1}^{C}\left[\sum_{k=K(c-1)+1}^{Kc}\frac{\sum_{\bar{k}=K(c-1)+1}^{Kc} Z\left(o_{\bar{k}} \mid s\right)}{K}\right] \tag{17}$$
$$\mathbb{P}\left(s \mid H^{-}\right) log\left(\bar{b}\right)$$
$$+\int_s \sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{Kc} Z\left(o_k \mid s\right)\mathbb{P}\left(s \mid H^{-}\right) log\left(b\right) =$$
$$-\int_s \sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{Kc} Z\left(o_k \mid s\right)\mathbb{P}\left(s \mid H^{-}\right) log\left(\bar{b}\right) \tag{18}$$
$$+\int_s \sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{Kc} Z\left(o_k \mid s\right)\mathbb{P}\left(s \mid H^{-}\right) log\left(b\right) =$$
$$\sum_{i=1}^{N_o}\mathbb{P}\left(o_i \mid H^{-}\right)\int_s b\cdot log\left(\frac{b}{\bar{b}}\right) = \tag{19}$$
$$E_o\left[\mathcal{D}_{KL}\left(b||\bar{b}\right)\right] \geq 0$$

(19) obtained by applying similar steps in reverse order. The last equality holds since KL-divergence is non-negative and so is its expectation. It is left to prove the upper bound; Applying Bayes rule to the nominator and denominator of (19),

$$\sum_{i=1}^{N_o}\mathbb{P}\left(o_i \mid H^{-}\right)\int_s b log\left(\frac{Z\left(o_i \mid s\right)}{\bar{Z}\left(o_i \mid s\right)}\right) \tag{20}$$
$$+\sum_{i=1}^{N_o}\mathbb{P}\left(o_i \mid H^{-}\right) log\left(\frac{\bar{\mathbb{P}}\left(o_i \mid H^{-}\right)}{\mathbb{P}\left(o_i \mid H^{-}\right)}\right)\int_s bds$$

By construction of the abstract observation model,

$$\sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{Kc}\mathbb{P}\left(o_k \mid H^{-}\right)\int_s b\cdot log\left(\frac{Z\left(o_k \mid s\right)\cdot K}{\sum_{k=K(c-1)+1}^{Kc} Z\left(o_{\bar{k}} \mid s\right)}\right)ds$$
$$+\sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{Kc}\mathbb{P}\left(o_k \mid H^{-}\right) log\left(\frac{\bar{\mathbb{P}}\left(o_k \mid H^{-}\right)}{\mathbb{P}\left(o_k \mid H^{-}\right)}\right)$$
$$\leq log(K)\sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{Kc}\mathbb{P}\left(o_k \mid H^{-}\right)\int_s bds + 0 = log(K).$$

The inequality is due to positiveness of the denominator in the first term and Jensen's inequality in the second term. we end up with,

$$0 \leq \bar{\mathbb{E}}_o\left[\mathcal{H}\left(\bar{b}\right)\right] - \mathbb{E}_o\left[\mathcal{H}\left(b\right)\right] \leq log(K). \tag{21}$$

□

A.4 Corollary 1.1

Proof. From Lemma 2 it is clear that the expected state-dependent reward is unaffected by the abstraction, and thus will not affect the value function. For the sake of conciseness and clarity, we prove the case that the value function depends only on the entropy. The general case derived similarly by applying Lemma 2 instead of the expected state-dependent reward.

$$V^{\pi}(b_t) = \sum_{n=1}^{N_o}\mathbb{P}\left(o_{t+1}^{n} \mid H_{t+1}^{-}\right)\left[-\mathcal{H}(b_{t+1}) + V^{\pi}(b_{t+1})\right]$$

expanding the value function,

$$-\sum_{n=1}^{N_o}\mathbb{P}\left(o_{t+1}^{n} \mid H_{t+1}^{-}\right)\left[\mathcal{H}(b_{t+1})\right. \tag{22}$$
$$\left.+\sum_{n'=1}^{N_o}\mathbb{P}\left(o_{t+2}^{n'} \mid H_{t+2}^{-}\right)\mathcal{H}(b_{t+2}) + \cdots\right]\Big]$$

by linearity of expectation,

$$-\mathbb{E}_{o_{t+1}}\left[\mathcal{H}(b_{t+1})\right] + \mathbb{E}_{o_{t+1}}\left[\mathbb{E}_{o_{t+2}}\left[\mathcal{H}(b_{t+2})\right]\right] + \cdots \tag{23}$$

using Theorem 1 for each of the expected entropy terms separately until time-step $\mathcal{T}-1$,

$$V^{\pi}(b_t) \geq -\left[\bar{\mathbb{E}}[\mathcal{H}(\bar{b}_{t+1})] + log(K)\right] \tag{24}$$
$$-\mathbb{E}_{o_{t+1}}\left[\bar{\mathbb{E}}_{o_{t+2}}[\mathcal{H}(\bar{b}_{t+2})] + log(K)\right]\cdots$$
$$= -\bar{\mathbb{E}}_{o_{t+1}}\left[\mathcal{H}(\bar{b}_{t+1})\right] \tag{25}$$
$$-\mathbb{E}_{o_{t+1}}\bar{\mathbb{E}}_{o_{t+2}}\left[\mathcal{H}(\bar{b}_{t+2})\right]\cdots + \mathcal{T}\cdot log(K)$$

applying similar steps in reverse order yields the abstract value function,

$$\bar{V}^{\pi}(b_t) + \mathcal{T}\cdot log(K)$$
$$\Longrightarrow \bar{V}^{\pi}(b_t) - V^{\pi}(b_t) \leq \mathcal{T}\cdot log(K).$$

Following the same derivation and applying the other side of the inequality of Theorem 1, completes the derivation for the entropy as reward. Using the more general reward, (1), and applying Lemma 2, yields the proof for corollary 1.1,

$$0 \leq \bar{V}^{\pi}(b_t) - V^{\pi}(b_t) \leq \mathcal{T}\cdot\omega_2 log(K).$$

□

A.5 Expected Entropy Estimation

We derive an estimator to the expected differential entropy with continuous observation space. The discrete state or observation spaces follows similar derivation by replacing integrals with summations.

$$\mathbb{E}[\mathcal{H}(b_t)] = -\int_{o_t} p(o_t \mid H_t^-) \int_{s_t} p(s_t \mid o_t, H_t^-) \cdot log(p(s_t \mid o_t, H_t^-)) \tag{26}$$

applying Bayes' rule,

$$\begin{aligned}\mathbb{E}[\mathcal{H}(b_t)] = &-\int_{o_t}\int_{s_t} p(s_t, o_t \mid H_t^-) \\ &\cdot log\left(Z(o_t \mid s_t)\int_{s_{t-1}} T(s_t \mid s_{t-1}, a_{t-1}) b(s_{t-1})\right) \\ &+\int_{o_t}\int_{s_t} p(s_t, o_t \mid H_t^-) \cdot log(p(o_t \mid H_t^-))\end{aligned} \tag{27}$$

by chain rule and marginalization,

$$\begin{aligned}\mathbb{E}[\mathcal{H}(b_t)] = &-\int_{o_t}\int_{s_t} Z(o_t \mid s_t) b^-(s_t) \\ &\cdot log\left(Z(o_t \mid s_t)\int_{s_{t-1}} T(s_t \mid s_{t-1}, a_{t-1}) b(s_{t-1})\right) \\ &+\int_{o_t}\int_{s_t} Z(o_t \mid s_t) b^-(s_t) \\ &\cdot log\left(\int_{s_t} Z(o_t \mid s_t) b^-(s_t)\right)\end{aligned} \tag{28}$$

using particle filter, the belief represented as a set of weighted particles, $\{(s^1, q^1), \ldots, (s^i, q^i), \ldots, (s^n, q^n)\}$. Where q^i denotes the weight of particle i.

$$\begin{aligned}\mathbb{E}[\mathcal{H}(b_t)] \approx &-\int_{o_t} \eta_t \sum_{i=1}^{n} Z\left(o_t \mid s_t^i\right) q_{t-1}^i \cdot \\ &\cdot log\left(Z\left(o_t \mid s_t^i\right) \sum_{j=1}^{n} p\left(s_t^i \mid s_{t-1}^j, a_{t-1}\right) q_{t-1}^j\right) \\ &+\int_{o_t} \eta_t \sum_{i=1}^{n} Z\left(o_t \mid s_t^i\right) q_{t-1}^i \cdot \\ &\cdot log\left(\sum_{i} Z\left(o_t \mid s_t^i\right) q_{t-1}^i\right)\end{aligned} \tag{29}$$

where $\eta_t = \int_{o_t} \sum_{i=1}^{n} Z\left(o_t \mid s_t^i\right) q_{t-1}^i$ normalizes the estimator for the probability function so that it sums to 1. Then, we approximate expectation over the observation space using observation samples, and query the likelihood model conditioned on the state samples, $Z\left(o^m \mid s^i\right) \quad \forall o^m \in \{o^1, \ldots, o^M\}$,

$$\begin{aligned}\hat{\mathbb{E}}[\mathcal{H}(\hat{b}_t)] = &-\overline{\eta}_t \sum_{m=1}^{M} \sum_{i=1}^{n} Z\left(o_t^m \mid s_t^i\right) q_{t-1}^i \cdot \\ &\cdot log\left(Z\left(o_t^m \mid s_t^i\right) \sum_{j=1}^{n} T\left(s_t^i \mid s_{t-1}^j, a_{t-1}\right) q_{t-1}^j\right) \\ &+\overline{\eta}_t \sum_{m=1}^{M}\left[\sum_{i=1}^{n} Z\left(o_t^m \mid s_t^i\right) q_{t-1}^i\right] \\ &\cdot log\left(\sum_{i'=1}^{n} Z\left(o_t^m \mid s_t^{i'}\right) q_{t-1}^{i'}\right)\end{aligned} \tag{30}$$

$$\overline{\eta}_t = \frac{1}{\sum_{m=1}^{M}\sum_{i=1}^{n} Z\left(o_t^m \mid s_t^i\right) q_{t-1}^i} \tag{31}$$

which concludes the derivation.

A.6 Theorem 2

Note that the reward function, (1), is built of two terms, state dependent reward and entropy,

$$R(b,a,b') = \omega_1 \mathbb{E}_{s\sim b'}\left[r(s,a)\right] + \omega_2 \mathcal{H}(b'). \tag{32}$$

For clarity, we divide the proof into two parts,

$$\hat{\bar{\mathbb{E}}}_o\left[R\left(\hat{b},a,\hat{\bar{b}}'\right)\right] - \hat{\mathbb{E}}_o\left[R\left(\hat{b},a,\hat{b}'\right)\right] = \tag{33}$$

$$\omega_1\left(\hat{\bar{\mathbb{E}}}_o\left[\mathbb{E}_{s\sim\hat{\bar{b}}'}\left[r(s,a)\right]\right] - \hat{\mathbb{E}}_o\left[\mathbb{E}_{s\sim\hat{b}'}\left[r(s,a)\right]\right]\right) \tag{34}$$

$$+\omega_2\left(\hat{\bar{\mathbb{E}}}_o\left[\mathcal{H}\left(\hat{\bar{b}}\right)\right] - \hat{\mathbb{E}}_o\left[\mathcal{H}\left(\hat{b}\right)\right]\right).$$

The first is about the difference in expected entropy, which is similar in spirit to the proof of Theorem 1. The second follows the claim and proof of Lemma (2). We begin with the difference of the expected entropy. For clarity, we derive the upper and lower bounds separately. For the upper bound,

Proof. In the following we directly plug-in the expected entropy estimator, with both the abstract observation model and the original observation model. For clarity, we split the expression into two parts and deal with each separately.

$$\begin{aligned}
&\hat{\bar{\mathbb{E}}}_o\left[\mathcal{H}\left(\hat{\bar{b}}\right)\right] - \hat{\mathbb{E}}_o\left[\mathcal{H}\left(\hat{b}\right)\right] \\
&= \underbrace{-\overline{\eta}_t \sum_{m=1}^{M}\sum_{i=1}^{n} \overline{Z}\left(o_t^m \mid s_t^i\right) q_{t-1}^i}_{(a)} \\
&\underbrace{\cdot log\left(\overline{Z}\left(o_t^m \mid s_t^i\right)\sum_{j=1}^{n} T\left(s_t^i \mid s_{t-1}^j, a_{t-1}\right) q_{t-1}^j\right)}_{(a)} \\
&\underbrace{+\overline{\eta}_t \sum_{m=1}^{M}\sum_{i=1}^{n} \overline{Z}\left(o_t^m \mid s_t^i\right) q_{t-1}^i \cdot log\left(\sum_{i=1}^{n} \overline{Z}\left(o_t^m \mid s_t^i\right) q_{t-1}^i\right)}_{(b)} \\
&\underbrace{+\overline{\eta}_t \sum_{m=1}^{M}\sum_{i=1}^{n} Z\left(o_t^m \mid s_t^i\right) q_{t-1}^i}_{(a)} \\
&\underbrace{\cdot log\left(Z\left(o_t^m \mid s_t^i\right)\sum_{j=1}^{n} T\left(s_t^i \mid s_{t-1}^j, a_{t-1}\right) q_{t-1}^j\right)}_{(a)} \\
&\underbrace{-\overline{\eta}_t \sum_{m=1}^{M}\sum_{i=1}^{n} Z\left(o_t^m \mid s_t^i\right) q_{t-1}^i \cdot log\left(\sum_{i=1}^{n} Z\left(o_t^m \mid s_t^i\right) q_{t-1}^i\right)}_{(b)}
\end{aligned} \tag{35}$$

In the first expression we start by splitting the summation to sum over its clusters and sum over the components of each cluster,

$$(a) = \overline{\eta}_t \sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{K\cdot c}\sum_{i=1}^{n} Z\left(o_t^k \mid s_t^i\right) q_{t-1}^i \tag{36}$$

$$\cdot log\left(\frac{Z\left(o_t^k \mid s_t^i\right)\sum_{j=1}^{n} p\left(s_t^i \mid s_{t-1}^j, a_{t-1}\right) q_{t-1}^j}{\overline{Z}\left(o_t^k \mid s_t^i\right)\sum_{j=1}^{n} p\left(s_t^i \mid s_{t-1}^j, a_{t-1}\right) q_{t-1}^j}\right)$$

$$(a) = \overline{\eta}_t \sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{K\cdot c}\sum_{i=1}^{n} Z\left(o_t^k \mid s_t^i\right) q_{t-1}^i \tag{37}$$

$$\cdot log\left(\frac{Z\left(o_t^k \mid s_t^i\right)}{\overline{Z}\left(o_t^k \mid s_t^i\right)}\right)$$

using the abstract model, (4),

$$(a) = \overline{\eta}_t \sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{K\cdot c}\sum_{i=1}^{n} Z\left(o_t^k \mid s_t^i\right) q_{t-1}^i \tag{38}$$

$$\cdot log\left(\frac{K\cdot Z\left(o_t^k \mid s_t^i\right)}{\sum_{k=K(c-1)+1}^{K\cdot c} Z\left(o_t^k \mid s_t^i\right)}\right). \tag{39}$$

since the denominator within the log is a sum of positive values, the following clearly holds,

$$(a) \leq \overline{\eta}_t \sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{K\cdot c}\sum_{i=1}^{n} Z\left(o_t^k \mid s_t^i\right) q_{t-1}^i \cdot log(K) \tag{40}$$

by taking the constant $log(K)$ out of the summation, the rest sums to one, so $(a) \leq log(K)$. Next we bound the second expression from above,

$$(b) = \overline{\eta}_t \sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{K\cdot c}\sum_{i=1}^{n} Z\left(o_t^k \mid s_t^i\right) q_{t-1}^i \tag{41}$$

$$\cdot log\left(\frac{\sum_{i=1}^{n} \overline{Z}\left(o_t^k \mid s_t^i\right) q_{t-1}^i/\overline{\eta}_t}{\sum_{i=1}^{n} Z\left(o_t^k \mid s_t^i\right) q_{t-1}^i/\overline{\eta}_t}\right)$$

applying Jensen's inequality,

$$(b) \leq log\left(\overline{\eta}_t \sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{K\cdot c}\sum_{i=1}^{n} \overline{Z}\left(o_t^k \mid s_t^i\right) q_{t-1}^i\right) \tag{42}$$

by recalling the definition of the normalizer, we end up with $log(1) = 0$ □

Last, we provide a proof for the lower bound,

Proof.

$$\hat{\bar{\mathbb{E}}}_o\left[\mathcal{H}\left(\hat{\bar{b}}\right)\right]-\hat{\mathbb{E}}_o\left[\mathcal{H}\left(\hat{b}\right)\right]= \tag{43}$$
$$-\overline{\eta}_t\sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{K\cdot c}\sum_{i=1}^{n}Z\left(o_t^k\mid s_t^i\right)q_{t-1}^i$$
$$\cdot log\Big[\frac{\overline{Z}\left(o_t^k\mid s_t^i\right)\sum_{j=1}^{n}T\left(s_t^i\mid s_{t-1}^j,a_{t-1}\right)q_{t-1}^j}{\sum_{i=1}^{n}\overline{Z}\left(o_t^k\mid s_t^i\right)q_{t-1}^i}$$
$$\cdot\frac{\sum_{i=1}^{n}Z\left(o_t^k\mid s_t^i\right)q_{t-1}^i}{Z\left(o_t^k\mid s_t^i\right)\sum_{j=1}^{n}T\left(s_t^i\mid s_{t-1}^j,a_{t-1}\right)q_{t-1}^j}\Big]$$

since $log(x)\leq x-1,\ \forall x>0$,

$$\hat{\bar{\mathbb{E}}}_o\left[\mathcal{H}\left(\hat{\bar{b}}\right)\right]-\hat{\mathbb{E}}_o\left[\mathcal{H}\left(\hat{b}\right)\right]\geq \tag{44}$$
$$\overline{\eta}_t\sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{K\cdot c}\sum_{i=1}^{n}Z\left(o_t^k\mid s_t^i\right)q_{t-1}^i$$
$$\cdot\left(1-\frac{\overline{Z}\left(o_t^k\mid s_t^i\right)}{\sum_{i=1}^{n}\overline{Z}\left(o_t^k\mid s_t^i\right)q_{t-1}^i}\cdot\frac{\sum_{i=1}^{n}Z\left(o_t^k\mid s_t^i\right)q_{t-1}^i}{Z\left(o_t^k\mid s_t^i\right)}\right)$$

rearranging terms,

$$1-\overline{\eta}_t\sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{K\cdot c}\sum_{i=1}^{n}Z\left(o_t^k\mid s_t^i\right)q_{t-1}^i \tag{45}$$
$$\cdot\frac{\overline{Z}\left(o_t^k\mid s_t^i\right)}{\sum_{i=1}^{n}\overline{Z}\left(o_t^k\mid s_t^i\right)q_{t-1}^i}\cdot\frac{\sum_{i=1}^{n}Z\left(o_t^k\mid s_t^i\right)q_{t-1}^i}{Z\left(o_t^k\mid s_t^i\right)} \tag{46}$$

we conclude with,

$$1-\overline{\eta}_t\sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{K\cdot c}\sum_{i=1}^{n}Z\left(o_t^k\mid s_t^i\right)q_{t-1}^i=0 \tag{47}$$

□

We now derive the second part of Theorem 2, i.e. for the difference of expected state-dependent reward.

Lemma 1. *The value of the estimated expected state-dependent reward is not affected by the abstraction shown in (4), i.e.,*

$$\hat{\bar{\mathbb{E}}}_o\left[\mathbb{E}_{s\sim\hat{\bar{b}}'}\left[r(s,a)\right]\right]=\hat{\mathbb{E}}_o\left[\mathbb{E}_{s\sim\hat{b}'}\left[r(s,a)\right]\right] \tag{48}$$

Proof.

$$\hat{\bar{\mathbb{E}}}_o\left[\mathbb{E}_{s\sim\hat{\bar{b}}_t'}\left[r(s_t,a_t)\right]\right]= \tag{49}$$
$$\sum_{m=1}^{M}\bar{\mathbb{P}}(o_t^m\mid H_t^-)\sum_{i=1}^{n}\bar{\mathbb{P}}(s_t^i\mid o_t^m,H_t^-)r(s_t^i,a_t) \tag{50}$$

applying chain rule,

$$\sum_{m=1}^{M}\sum_{i=1}^{n}\bar{\mathbb{P}}(s_t^i,o_t^m\mid H_t^-)r(s_t^i,a_t), \tag{51}$$

then applying chain-rule from the other direction and using the markovian assumption of the observation model,

$$\sum_{m=1}^{M}\sum_{i=1}^{n}\bar{Z}(o_t^m\mid s_t^i)b_t^-\cdot r(s_t^i,a_t). \tag{52}$$

Applying the transition function on particles from b_{t-1}, does not alter their weights, therefore we receive the following expression,

$$\sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{K\cdot c}\sum_{i=1}^{n}\bar{Z}\left(o_t^k\mid s_t^i\right)q_{t-1}^i r(s_t^i,a_t) \tag{53}$$

Using (1),

$$\sum_{c=1}^{C}\sum_{k=K(c-1)+1}^{K\cdot c}\sum_{i=1}^{n}\frac{\sum_{\bar{k}=K(c-1)+1}^{K\cdot c}Z\left(o_t^{\bar{k}}\mid s_t^i\right)}{K}q_{t-1}^i r(s_t^i,a_t), \tag{54}$$

followed by canceling the summation over k with K in the denominator,

$$\sum_{m=1}^{M}\sum_{i=1}^{n}Z(o_t^m\mid s_t^i)b_t^-\cdot r(s_t^i,a_t). \tag{55}$$

We then end up with the desired result,

$$\hat{\bar{\mathbb{E}}}_o\left[\mathbb{E}_{s\sim\hat{\bar{b}}'}\left[r(s,a)\right]\right]=\hat{\mathbb{E}}_o\left[\mathbb{E}_{s\sim\hat{b}'}\left[r(s,a)\right]\right] \tag{56}$$

□

To conclude the proofs of Theorem 2, note that,

$$0\leq\omega_1\left(\hat{\bar{\mathbb{E}}}_o\left[\mathbb{E}_{s\sim\hat{\bar{b}}'}\left[r(s,a)\right]\right]-\hat{\mathbb{E}}_o\left[\mathbb{E}_{s\sim\hat{b}'}\left[r(s,a)\right]\right]\right) \tag{57}$$
$$+\omega_2\left(\hat{\bar{\mathbb{E}}}_o\left[\mathcal{H}\left(\hat{\bar{b}}\right)\right]-\hat{\mathbb{E}}_o\left[\mathcal{H}\left(\hat{b}\right)\right]\right)\leq\omega_2 log(K) \tag{58}$$

A.7 Corollary 2.1

The proof of 2.1 follows closely to the proof in A.4. Replacing the exact value function with its estimated counterpart from A.6 yields the desired result.

B Implementation Details

B.1 Domain

We compared the different algorithms on a two-dimensional Light Dark environment. In this domain, the unobserved state of the agent is its pose, (X, Y), defined relative to a global coordinate frame, located at $(0, 0)$. There are 9 possible actions, eight of which has one unit of translation, and they differ from each other by the direction which is equally spaced on a circle, the ninth action has zero translation. We denote the transition model as $x' = f(x, a, w)$. At each time step, the agent receives a noisy estimate of its position as an observation, denoted by $o = h(x, v)$. In our experiments we chose w and v to be distributed according to a Gaussian noise, although in general they may be arbitrary. The reward function defined as the negative weighted sum of distance to goal and entropy,

$$r(b, a) = -\mathrm{E}_b[\|x - x_g\|] - \mathcal{H}(b), \tag{59}$$

The prior belief assumed to be Gaussian, $b_0 = \mathcal{N}([0, 0], \Sigma_0)$. In all our experiments, we employ a receding horizon approach. At each iteration we calculate a solution from scratch and share no information across different time steps.

B.2 Domain - Total Return Evaluation

We performed the experiments on a modification of Light Dark 2D and added forbidden regions to the environment. Whenever the agent crosses to a forbidden region, a -10 reward was added to its immediate reward. Also, we added +10 reward whenever the agent reached the goal and stayed there until the episode terminated.

B.3 Hyperparameters

Here we present the hyperparameters used to evaluate the total return performance.

AI-FSSS			
n	C	K	
20	4	4	
FFFS			
n	C		
20	4		
PFT-DPW			
n	c[1]	k_o	α_o
20	1	4	0.014

Table 1: Hyperparameters used in the experiments.

[2] c controls the bonus of the UCB function, which is different from the observation branching factor in FSSS and AI-FSSS, C.