Online Supplement: Improving UCT Planning via Approximate Homomorphisms

Nan Jiang[1], Satinder Singh[1], and Richard Lewis[2]

[1]Computer Science and Engineering , University of Michigan
[2]Department of Psychology, University of Michigan

1 Theoretical Analysis

The main idea of our algorithm is to build empirical local layered MDPs based on the trajectories sampled by UCT and then to find approximate homomorphisms in them. This construction is lossy in two ways:

1. The empirical MDP is different from the original MDP.
2. The abstraction is built hence has approximation errors from the empirical MDP.

We will show how to combine the loss from both sources. First, we define the notion of loss in value along with some notation useful for the analysis.

Notation & Objective of Analysis:
For the current state of interest, let the true local layered MDP with depth $d_{\max}$ be M. In the first step that introduces value-loss, UCT samples n trajectories in M and builds an empirical MDP $\hat{M}$. In a second step that also introduces value-loss, an approximate homomorphism h maps $\hat{M}$ to an abstract MDP $\hat{M}_h$ constructed by applying Algorithm 1 in the main paper to $\hat{M}$ with parameter $(\frac{\epsilon'_T}{2}, \frac{\epsilon'_R}{2})$ (hence the approximation error of the constructed abstraction is at most $(\epsilon'_T, \epsilon'_R)$) [1]. Our *analytical objective* is to bound the loss of the abstraction, i.e., to bound $\|V_M^{\pi^*} - V_M^{\pi_h^*}\|_\infty$, where V_M^π is the expected value function of policy π evaluated in MDP M, and π^* is the optimal policy in M, while π_h^* is the optimal policy in $\hat{M}_h$ lifted to true local MDP M.

Theorem 1. *(Main Result)* $\forall \eta_T, \eta_R > 0$, $0 < \delta < 1$,

$$\|V_M^{\pi^*} - V_M^{\pi_h^*}\|_\infty \leq \frac{2(\epsilon'_R + \eta_R)}{1-\gamma} + \frac{\gamma(R_{\max} - R_{\min})(\epsilon'_T + \eta_T)}{(1-\gamma)^2} \tag{1}$$

holds with probability at least $1 - \delta$ *if*

$$n > \max\{\exp^{(d_{\max})}\left[\log\left(a(KB)^{d_{\max}}/\delta\right)/b\right],\ N\}$$

where

1. K *is the number of actions available in state,*
2. B *is the maximal number of possible next states from a state-action pair,*
3. $p \overset{def}{=} \min_{(s,a,s_1,d):P(s,a,s_1,d))>0} P(s,a,s_1,d))/2$,
4. $[R_{\min}, R_{\max}]$ *is the range of reward in* $M, \hat{M}$ *and* $\hat{M}_h$,

[1]We use P, $\hat{P}$ and $\hat{P}_h$ to distinguish transition probabilities of M, $\hat{M}$ and $\hat{M}_h$, and similarly for reward function.

5. N, c *are positive constants that do not depend on the choice of* η_T, η_R,

6. $a \overset{def}{=} \max\{3d_{\max},\ 6B\}$,

7. $b \overset{def}{=} \min\{2cp^2,\ 2c\eta_R^2/(R_{\max}-R_{\min})^2,\ 2c\eta_T^2/B^2\}$.

We prove Theorem 1 using the following three lemmas:

Lemma 2. *The probability that*

$$\begin{aligned} \max_{s,a,d} \sum_{s_1} \left|\hat{P}(s,a,s_1,d)) - P(s,a,s_1,d))\right| \le \eta_T \\ \max_{s,a,d} |\hat{R}(s,a,d) - R(s,a,d)| \le \eta_R \end{aligned} \tag{2}$$

holds is at least

$$\begin{aligned} 1 - (KB)^{d_{\max}} \Big(& d_{\max} \exp(-2p^2 c \log^{(d_{\max})}(n)) + 2\exp(-2c\log^{(d_{\max})}(n)\eta_R^2/(R_{\max}-R_{\min})^2) \\ & +2B\exp(-2\eta_T^2 c \log^{(d_{\max})}(n)/B^2)\Big). \end{aligned} \tag{3}$$

Lemma 3. *Let the approximation parameter for* $h : M \mapsto \hat{M}_h$ *be* (ϵ_T, ϵ_R). *If Eq.(2) holds,* $\epsilon_T \le \epsilon'_T + \eta_T$ *and* $\epsilon_R \le \epsilon'_R + \eta_R$.

Lemma 4. *(Ravindran and Barto, 2004 [1])*

$$\|V_M^{\pi^*} - V_M^{\pi_h^*}\|_\infty \le \frac{2\epsilon_R}{1-\gamma} + \frac{\gamma(R_{\max}-R_{\min})\epsilon_T}{(1-\gamma)^2}.$$

The key idea in the proof of Lemma 2 is to consider each (s,a,d) and bound the probability that $\hat{R}(s,a,d)$ and $\hat{P}(s,a,\cdot,d)$ are not accurate, and then bound the probability that inaccurate estimates do not occur at any (s,a,d) by union bound. To obtained the former result, we first need to bound the number of times an (s,a,d) tuple is visited, which is given in the following lemma.

Lemma 5. $\exists c > 0, N > 0$ *s.t. when* $n > N$,

$$\mathbb{P}\Big\{n_{s,a,d} \ge c\log^{(d+1)}(n)\Big\} \ge 1 - \exp(-2p^2 c\log^{(d_{\max})}(n)))^d.$$

Proof. (By induction.) At $d = 0$, which is the root, the state is visited exactly n times. According to Theorem 3 in [2], $\exists \rho > 0$ s.t. $n_{s,a,d} \ge \rho\log(n_{s,d})$. Therefore, $n_{s,a,0} \ge \rho\log(n_{s,0}) > c\log(n)$ with probability 1 as long as $c < \rho$.

Now consider arbitrary $d < d_{\max}$. Let the state-action pair at the previous level that leads to s be $(s', a', d-1)$. According to the induction assumption,

$$\mathbb{P}\Big\{n_{s',a',d-1} \ge c\log^{(d)}(n)\Big\} \ge (1 - \exp(-2p^2 c\log^{(d_{\max})}(n)))^{d-1}. \tag{4}$$

What we need to bound is $\mathbb{P}\Big\{n_{s,a,d} \ge c\log^{(d+1)}(n)\Big\}$, which can be decomposed in the following way

$$\begin{aligned} & \mathbb{P}\Big\{n_{s,a,d} \ge c\log^{(d+1)}(n)\Big\} \\ \ge\ & \mathbb{P}\Big\{n_{s,a,d} \ge c\log^{(d+1)}(n),\ n_{s',a',d-1} \ge c\log^{(d)}(n)\Big\} \\ =\ & \mathbb{P}\Big\{n_{s,a,d} \ge c\log^{(d+1)}(n) \ \Big|\ n_{s',a',d-1} \ge c\log^{(d)}(n)\Big\} \cdot \mathbb{P}\Big\{n_{s',a',d-1} \ge c\log^{(d)}(n)\Big\}. \end{aligned}$$

The second term has already been bounded in Eq.(4). We will bound the first term in two steps: first, we show that

$$\mathbb{P}\Big\{n_{s,a,d} \geq c\log^{(d+1)}(n) \;\Big|\; n_{s',a',d-1} \geq c\log^{(d)}(n)\Big\} \geq \mathbb{P}\Big\{n_{s,d}/n_{s',a',d-1} \geq p \;\Big|\; n_{s',a',d-1} \geq c\log^{(d)}(n)\Big\}. \quad (5)$$

This is because when $n_{s,d}/n_{s',a',d-1} \geq p$ holds,

$$n_{s,a,d} \geq \rho\log(n_{s,d}) \geq \rho\log(pc\log^{(d)}(n))) = \rho\log^{(d+1)}(n) + \rho\log(pc).$$

Note that the second term is a constant, thus for any $0 < c < \rho$, as long as $\log^{(d_{\max})}(n) > \log(pc)/(c-\rho)$ (solving this inequality yields N, which does not depend on η_T and η_R) we have $n_{s,a,d} \geq c\log^{(d+1)}(n)$. This shows that the right side event of Eq.(5) is a sub-event of the left side, thus the inequality holds.

Second, we bound the right side of Eq.(5). For any fixed $n_{s',a',d-1}$, $n_{s',a',d-1,s_1}/n_{s',a',d-1}$ is the average of Bernoulli random variables with expected value $P(s',a',s,d-1)$. According to Hoeffding bound,

$$\begin{aligned}
&\mathbb{P}\Big\{n_{s,d}/n_{s',a',d-1} \geq p\Big\} \geq \mathbb{P}\Big\{n_{s',a',s,d-1}/n_{s',a',d-1} \geq p\Big\} \\
&= \mathbb{P}\Big\{n_{s',a',s,d-1}/n_{s',a',d-1} - P(s',a',s,d-1) \geq p - P(s',a',s,d-1)\Big\} \\
&\geq 1 - \exp(-2(P(s',a',s,d-1)-p)^2 n_{s',a',d-1}).
\end{aligned}$$

According to the definition of p in Eq.(3), we always have $P(s',a',s,d-1) - p > p$. With $n_{s',a',d-1} \geq c\log^{(d)}(n)$, we have

$$\mathbb{P}\Big\{n_{s,d}/n_{s',a',d-1} \geq p\Big\} \geq 1 - \exp(-2p^2 c\log^{(d)}(n))$$

Hence

$$\begin{aligned}
\mathbb{P}\Big\{n_{s,a,d} \geq c\log^{(d+1)}(n)\Big\} &\geq \mathbb{P}\Big\{n_{s,a,d} \geq c\log^{(d+1)}(n) \;\Big|\; n_{s',a',d-1} \geq c\log^{(d)}(n)\Big\} \cdot \mathbb{P}\Big\{n_{s',a',d-1} \geq c\log^{(d)}(n)\Big\} \\
&\geq \mathbb{P}\Big\{n_{s,d}/n_{s',a',d-1} \geq p \;\Big|\; n_{s',a',d-1} \geq c\log^{(d)}(n)\Big\} \cdot \mathbb{P}\Big\{n_{s',a',d-1} \geq c\log^{(d)}(n)\Big\} \\
&\geq (1 - \exp(-2p^2 c\log^{(d)}(n)))(1 - \exp(-2p^2 c\log^{(d_{\max})}(n)))^{d-1} \\
&\geq (1 - \exp(-2p^2 c\log^{(d_{\max})}(n)))^{d}.
\end{aligned}$$

So the lemma follows. ■

Proof of Lemma 2. Consider a state-action-depth tuple (s,a,d). For the empirical reward and transition probabilities to be accurate at (s,a,d), we first require that (s,a,d) is visited sufficiently. By relaxing Lemma 5 we have a universal bound for $n_{s,a,d}$ that is independent of d: $\forall(s,a,d), \exists c, N'$, s.t. $\forall n > N'$,

$$\mathbb{P}\Big\{n_{s,a,d} \geq c\log^{(d_{\max})}(n)\Big\} \;\geq\; (1 - \exp(-2p^2 c\log^{(d_{\max})}(n)))^{d_{\max}}$$

where c and N' are the constants specified in Lemma 5.

Now we can bound the probability that reward and transition estimates are inaccurate separately. The empirical reward $\hat{R}(s,a)$ is the average of at least $c\log^{(d_{\max})}(n)$ i.i.d. samples of random variables that lie in $[R_{\min}, R_{\max}]$, with expected value $R(s,a)$, hence by Hoeffding bound

$$\mathbb{P}\Big\{|\hat{R}(s,a) - R(s,a)| > \eta_R \;\Big|\; n_{s,a,d} \geq c\log^{(d_{\max})}(n)\Big\} \leq 2\exp(-2c\log^{(d_{\max})}(n)\eta_R^2/(R_{\max}-R_{\min})^2).$$

Similarly for transition probabilities,

$$\begin{aligned}
&\mathbb{P}\Big\{\sum_{s_1}|\hat{P}(s,a,s_1,d) - P(s,a,s_1,d)| > \eta_T \;\Big|\; n_{s,a,d} \geq c\log^{(d_{\max})}(n)\Big\} \\
\leq\;& \mathbb{P}\Big\{\bigcup_{s_1}|\hat{P}(s,a,s_1,d) - P(s,a,s_1,d)| > \eta_T/B \;\Big|\; n_{s,a,d} \geq c\log^{(d_{\max})}(n)\Big\} \\
\leq\;& \sum_{s_1}\mathbb{P}\Big\{|\hat{P}(s,a,s_1,d) - P(s,a,s_1,d)| > \eta_T/B \;\Big|\; n_{s,a,d} \geq c\log^{(d_{\max})}(n)\Big\}.
\end{aligned}$$

Consider a particular possible next state s_1,

$$\mathbb{P}\Big\{|\hat{P}(s,a,s_1,d) - P(s,a,s_1,d)| > \eta_T/B \;\Big|\; n_{s,a,d} \geq c\log^{(d_{\max})}(n)\Big\} \leq 2\exp(-2\eta_T^2 c\log^{(d_{\max})}(n)/B^2).$$

As (s,a) has at most B possible next states,

$$\mathbb{P}\Big\{\sum_{s_1}|\hat{P}(s,a,s_1,d) - P(s,a,s_1,d)| > \eta_T \;\Big|\; n_{s,a,d} \geq c\log^{(d_{\max})}(n)\Big\} \leq 2B\exp(-2\eta_T^2 c\log^{(d_{\max})}(n)/B^2).$$

By observing that empirical reward and transition distribution are independent of each other when fixing $n_{s,a,d}$, we have the following result: $\forall(s,a,d)$,

$$\begin{aligned}
&\mathbb{P}\Big\{|\hat{R}(s,a,d) - R(s,a,d)| \leq \eta_R,\, \sum_{s_1}|\hat{P}(s,a,s_1,d) - P(s,a,s_1,d)| \leq \eta_T\Big\} \\
\geq\;& \mathbb{P}\Big\{|\hat{R}(s,a,d) - R(s,a,d)| \leq \eta_R,\, \sum_{s_1}|\hat{P}(s,a,s_1,d) - P(s,a,s_1,d)| \leq \eta_T,\, n_{s,a,d} \geq c\log^{(d_{\max})}(n)\Big\} \\
=\;& \mathbb{P}\Big\{n_{s,a,d} \geq c\log^{(d_{\max})}(n)\Big\}\cdot\mathbb{P}\Big\{|\hat{R}(s,a,d) - R(s,a,d)| \leq \eta_R \;\Big|\; n_{s,a,d} \geq c\log^{(d_{\max})}(n)\Big\} \\
&\cdot\mathbb{P}\Big\{\sum_{s_1}|\hat{P}(s,a,s_1,d) - P(s,a,s_1,d)| \leq \eta_T \;\Big|\; n_{s,a,d} \geq c\log^{(d_{\max})}(n)\Big\} \\
\geq\;& (1-\exp(-2p^2 c\log^{(d_{\max})}(n)))^{d_{\max}}(1-2\exp(-2c\log^{d+1}(n)\eta_R^2/(R_{\max}-R_{\min})^2)) \\
&(1-2B\exp(-2\eta_T^2 c\log^{(d_{\max})}(n)/B^2)).
\end{aligned} \tag{6}$$

The final step is to bound the probability that the estimate is accurate everywhere. With union bound,

$$\begin{aligned}
&\mathbb{P}\Big\{M' \text{ is not } (\epsilon_T,\epsilon_R) \text{ accurate}\Big\} \\
=\;& \mathbb{P}\Big\{\bigcup_{(s,a,d)} M' \text{ is not } (\epsilon_T,\epsilon_R) \text{ accurate at } (s,a,d)\Big\} \\
\leq\;& \sum_{(s,a,d)}\mathbb{P}\Big\{M' \text{ is not } (\epsilon_T,\epsilon_R) \text{ accurate at } (s,a,d)\Big\} \\
\leq\;& \#(s,a,d)\cdot\left(1-\mathbb{P}\Big\{|\hat{R}(s,a,d) - R(s,a,d)| \leq \eta_R, \sum_{s_1}|\hat{P}(s,a,s_1,d) - P(s,a,s_1,d)| \leq \eta_T\Big\}\right).
\end{aligned}$$

And the bound in Lemma 2 is obtained by plugging in Eq.(6) and noticing that $\#(s,a,d) \leq (KB)^{d_{\max}}$, and can be simplified by its first order approximation (which is strictly smaller). Finally, from Theorem 2 to our main result, we only have to require that each term in the outmost parenthesis in Eq.(3) is less than $\delta/3(KB)^{d_{\max}}$, and find the satisfying n. ■

Proof of Lemma 3.

$$
\begin{aligned}
\epsilon_T &\stackrel{\text{def}}{=} \max_{s,a,d} \sum_x \Big| \hat{P}_h(h(s),a,x,d) - \sum_{s_1:h(s_1)=x} P(s,a,s_1,d) \Big| \\
&\leq \max_{s,a,d} \sum_x \Big| \hat{P}_h(h(s),a,x,d) - \sum_{s_1:h(s_1)=x} \hat{P}(s,a,s_1,d) \Big| \\
&+ \max_{s,a,d} \sum_x \Big| \sum_{s_1:h(s_1)=x} \hat{P}(s,a,s_1,d) - \sum_{s_1:h(s_1)=x} P(s,a,s_1,d) \Big| \\
&= \epsilon'_T + \max_{s,a,d} \sum_x \Big| \sum_{s_1:h(s_1)=x} \big(\hat{P}(s,a,s_1,d) - P(s,a,s_1,d) \big) \Big| \\
&\leq \epsilon'_T + \max_{s,a,d} \sum_x \sum_{s_1:h(s_1)=x} \big| \hat{P}(s,a,s_1,d) - P(s,a,s_1,d) \big| \\
&= \epsilon'_T + \max_{s,a,d} \sum_{s_1} \big| \hat{P}(s,a,s_1,d) - P(s,a,s_1,d) \big| \\
&= \epsilon'_T + \eta_T .
\end{aligned}
$$

The proof of $\epsilon_R \leq \epsilon'_R + \eta_R$ is very similar hence omitted. ∎

References

[1] Balaraman Ravindran and A Barto. Approximate homomorphisms: A framework for nonexact minimization in Markov decision processes. In *5th International Conference on Knowledge-Based Computer Systems*, 2004.

[2] Levente Kocsis and Csaba Szepesvári. Bandit based Monte-Carlo planning. In *15th European Conference on Machine Learning*, pages 282–293, 2006.