Long Version of Proof of Theorem 3

Supplement to ICML submission "Learning in POMDPs with Monte Carlo Tree Search"

While RS-BA-POMCP is potentially more efficient, it is not directly whether it still converges to an ϵ-optimal value function. Here we show the main steps in proving that it is sound. These main steps are similar to the proof in POMCP. We point out however, that the technicalities of proving the components are far more involved.

Notation

We will use the following notation.

Action-observation histories.

- h_d is an action-observation history at depth d of a simulation,
- $h_d = (a_0, z_1, \ldots, a_{d-1}, z_d)$.

'Full' histories. In addition to actions and observations, full histories also include the states.

- H_0 the (unknown) full history (of *real* experience) at the root of the simulation: i.e., if there have been k steps of 'real' experience $H_0 = (s_{-k}, a_{-k}, s_{-k+1}, z_{-k-1}, \ldots, a_{-1}, s_0, z_0)$
- H_d is a full history (of *simulated* experience) at depth d in the lookahead tree: $H_d = (H_0, a_0, s_1, z_1, a_1, s_2, z_2, \ldots, a_{d-1}, s_d, z_d)$ $(H_{d-1}, a_{d-1}, s_d, z_d) = \langle s_{0:d}, h_d \rangle$.
- $H_d^{(i)}$ the full history at depth d corresponding to simulation i.
- In our proof, we will also need to indicate if a particular full history H_d is consistent with a full history at the root of simulation:

$$\text{Cons}(H_0, H_d) = \begin{cases} 1 & \text{if } H_d \text{ is consistent with the full history at the root } H_0 \ , \\ 0 & \text{otherwise.} \end{cases}$$

Dynamics Function. We fold transition and observations function into one:

- D denotes the dynamics model
- $D^{s_t z_t}_{s_{t-1} a_{t-1}} = D^{s' z}_{sa} = D_{s_{t-1}, a_{t-1}}(s_t, z_t) = D(s_t, z_t | s_{t-1}, a_{t-1}) = \Pr(s_t, z_t | s_{t-1}, a_{t-1})$
- D_{sa} denotes the vector: $\left\langle D^{s^1 z^1}_{sa}, \ldots, D^{s^{|\mathcal{S}|} z^{|\mathcal{Z}|}}_{sa} \right\rangle$

Counts

- $\chi^{s' z}_{sa}$ denotes how often $\langle s', z \rangle$ occurred after $\langle s, a \rangle$
- χ_{sa} is the vector of counts for $\langle s, a \rangle$.
- $\chi = \left\langle \chi_{s_1 a_1}, \ldots, \chi_{s_{|\mathcal{S}|} a_{|\mathcal{A}|}} \right\rangle$ is the total collection of all such count vectors.
- $\chi(H_d)$ denotes the vector of counts at simulated full history H_d.
- If χ_0 is the count vector at the root of simulation, we have that $\chi(H_d) = \chi_0 + \Delta(H_d)$, with $\Delta(H_d)$ the vector of counts of all (s, a, s', z) quadruples occurring in H_d.

Dirichlet distributions

- let $x = \langle x_1 \ldots x_K \rangle \in \Delta^K$ and $\alpha = \langle \alpha_1 \ldots \alpha_K \rangle$ be a count vector
- $Dir(x|\alpha) = \Pr(x; \alpha) = B(\alpha) \prod_{i=1}^K x_i^{\alpha_i - 1}$
- with $B(\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}$ the Dirichlet normalization constant, with Γ the gamma function.
- So, in translated in terms of dynamics function and counts, we have:
  - for a particular s, a: $Dir(D_{sa}|\chi_{sa}) = \Pr(D_{sa}; \chi_{sa}) = B(\chi_{sa}) \prod_{\langle s', z \rangle \in \mathcal{S} \times \mathcal{Z}} \left(D^{s' z}_{sa} \right)^{\chi^{s' z}_{sa}}$

– we will also abuse notation and write $Dir(D|\chi) = \prod_{\langle s,a \rangle} Dir(D_{sa}|\chi_{sa})$.

Var.

- $\dot{x}$ denotes a root sampled quantity x.
- $\mathbb{I}_{\{condition\}}$ is the indicator function which is 1 iff condition is true and 0 otherwise.

Definitions

Definition 1. The ***expected*** *full-history* ***expected transition*** *BA-POMDP rollout distribution* is the distribution over full histories of a BA-POMDP, when performing Monte-Carlo simulations according to a policy π. It is given by

$$P^{\pi}(H_{d+1}) = D_{\chi(H_d)}(s_{d+1}, z_{d+1}|a_s, s_d)\pi(a_d|h_d)P^{\pi}(H_d) \tag{1}$$

with $P^{\pi}(H_0) = b_0(\langle s_0, \chi_0 \rangle)$ the belief 'now' (at the root of the online planning).

Definition 2. The *full-history* ***root-sampling (RS)*** *BA-POMDP rollout distribution* is the distribution over full histories of a BA-POMDP, when performing Monte-Carlo simulations according to a policy π **in combination with root sampling** of the dynamics model D. This distribution, for a particular stage d, is given by

$$\tilde{P}^{\pi}_{K}(H_d) \triangleq \frac{1}{K_d} \sum_{i=1}^{K_d} \mathbb{I}_{\left\{H_d = H_d^{(i)}\right\}},$$

where

- K is the number of simulations that comprise the empirical distribution,
- K_d is the number of simulations that reach depth d (not all simulations might be equally long),
- $H_d^{(i)}$ is the history specified by the i-th particle at stage d.

Remark: throughout this proof we assume that there is only 1 initial count vector at the root. Or put better: we assume that there is one unique H_0 at which all simulations start. However, for 'real' steps $t > 0$ we could be in different H_t^{real} all corresponding to the same observed real history h_t^{real}. In this case, root sampling from the belief can be thought of root sampling the initial full history $H_0 \sim b(H_t^{real})$. As such, our proof shows convergence in probability of

$$\forall_{H_0} \forall_{H_d} \quad \tilde{P}^{\pi}_{K_d}(H_d|H_0) \xrightarrow{p} P^{\pi}(H_d|H_0).$$

for each such sampled H_0. It is clear that that directly implies that

$$\forall_{H_d} \quad \tilde{P}^{\pi}_{K_d}(H_d) = \mathbf{E}_{H_0}\left[\tilde{P}^{\pi}_{K_d}(H_d|H_0)\right] \xrightarrow{p} \mathbf{E}_{H_0}\left[P^{\pi}(H_d|H_0)\right] = P^{\pi}(H_d).$$

In the below, we omit the explicit conditioning on H_0.

Proof of Main Theorem

The proof depends on a lemma that follows below.

Theorem 3. *The full-history RS-BA-POMDP rollout distribution (Def. 2) converges in probability to full-history BA-POMDP rollout distribution (Def. 1):*

$$\forall_{H_d} \quad \tilde{P}^{\pi}_{K_d}(H_d) \xrightarrow{p} P^{\pi}(H_d). \tag{2}$$

Proof. For ease of notation we prove this for stage $d+1$. Note that a history $H_{d+1} = (H_d, a_d, s_{d+1}, z_{d+1})$, *only* differs from H_d in that it has one extra transition for the $(s_d, a_d, s_{d+1}, z_{d+1})$ quadruple, implying that $\chi(H_{d+1})$ only differs from $\chi(H_d)$ in the counts $\chi_{s_d a_d}$ for $s_d a_d$. Therefore, the expression for $\tilde{P}^{\pi}_{K_d}(H_d)$ derived in Lemma 4 below (cf. equation (20)) can be written in recursive form as

$$
\begin{aligned}
\tilde{P}^{\pi}(H_{d+1}) &= \text{Cons}(H_0, H_d) \prod_{t=0}^{d} \pi(a_t|h_t) \prod_{\langle s,a \rangle} \frac{B(\chi_{sa}(H_0))}{B(\chi_{sa}(H_{d+1}))} \\
&= \text{Cons}(H_0, H_d) \prod_{t=0}^{d-1} \pi(a_t|h_t)\pi(a_d|h_d) \prod_{\langle s,a \rangle} \frac{B(\chi_{sa}(H_0))}{B(\chi_{sa}(H_d))} \frac{B(\chi_{sa}(H_d))}{B(\chi_{sa}(H_{d+1}))} \\
&= \text{Cons}(H_0, H_d) \prod_{t=0}^{d-1} \pi(a_t|h_t)\pi(a_d|h_d) \left[\prod_{\langle s,a \rangle} \frac{B(\chi_{sa}(H_0))}{B(\chi_{sa}(H_d))} \right] \left[\prod_{\langle s,a \rangle} \frac{B(\chi_{sa}(H_d))}{B(\chi_{sa}(H_{d+1}))} \right] \\
&= \left[\text{Cons}(H_0, H_d) \prod_{t=0}^{d-1} \pi(a_t|h_t) \prod_{\langle s,a \rangle} \frac{B(\chi_{sa}(H_0))}{B(\chi_{sa}(H_d))} \right] \pi(a_d|h_d) \frac{B(\chi_{s_d a_d}(H_d))}{B(\chi_{s_d a_d}(H_{d+1}))} \\
&= \tilde{P}^{\pi}(H_d)\pi(a_d|h_d) \frac{B(\chi_{s_d a_d}(H_d))}{B(\chi_{s_d a_d}(H_{d+1}))}
\end{aligned}
$$

with base case $\tilde{P}^{\pi}(H_0) = 1$, and

$$
\frac{B(\chi_{s_d a_d}(H_d))}{B(\chi_{s_d a_d}(H_{d+1}))} = \frac{B(\chi_{s_d a_d}(H_0))}{B(\chi_{s_d a_d}(H_{d+1}))} \cdot \frac{B(\chi_{s_d a_d}(H_d))}{B(\chi_{s_d a_d}(H_0))} = \frac{B(\chi_{s_d a_d}(H_0))/B(\chi_{s_d a_d}(H_{d+1}))}{B(\chi_{s_d a_d}(H_0))/B(\chi_{s_d a_d}(H_d))} \tag{3}
$$

the result of dividing out the contribution of the old counts for $s_d a_d$ and multiplying in the new contribution, and similar for the observations probabilities. Now, we investigate these terms more closely.

Again remember that the sole difference between $H_{d+1} = (H_d, a_d, s_{d+1}, z_{d+1})$ andH_d is that it has one extra transition for the $(s_d, a_d, s_{d+1}, z_{d+1})$ quadruple. Let us write $T = \sum_{(s',z)} \chi^{s'z}_{s_d a_d}(H_d)$ for the total of the counts for s_d, a_d and $N = \chi^{s_{d+1} z_{d+1}}_{s_d a_d}(H_d)$ for the number of counts for that such a transition was to $(s_{d+1} z_{d+1})$. Because H_{d+1} only has 1 extra transition, we also know that for this history, the total counts is one higher: $\sum_{(s',z)} \chi^{s'z}_{s_d a_d}(H_{d+1}) = T + 1$ and since that transition was to $(s_{d+1} z_{d+1})$ the counts $\chi^{s_{d+1} z_{d+1}}_{s_d a_d}(H_{d+1}) = N + 1$. Now let us expand the term from (3):

$$
\begin{aligned}
\frac{B(\chi_{s_d a_d}(H_d))}{B(\chi_{s_d a_d}(H_{d+1}))} &= \frac{\Gamma(T)/\prod_{s'z} \Gamma(\chi^{s'z}_{s_d a_d}(H_d))}{\Gamma(T+1)/\prod_{s'z} \Gamma(\chi^{s'z}_{s_d a_d}(H_{d+1}))} \\
&= \frac{\Gamma(T)}{\Gamma(T+1)} \frac{\prod_{s'z} \Gamma(\chi^{s'z}_{s_d a_d}(H_{d+1}))}{\prod_{s'z} \Gamma(\chi^{s'}_{s_d a_d}(H_d))} \\
&= \frac{\Gamma(T)}{\Gamma(T+1)} \frac{\Gamma(\chi^{s_{d+1} z_{d+1}}_{s_d a_d}(H_{d+1})) \prod_{s'z \neq (s_{d+1} z_{d+1})} \Gamma(\chi^{s'z}_{s_d a_d}(H_{d+1}))}{\Gamma(\chi^{s_{d+1} z_{d+1}}_{s_d a_d}(H_d)) \prod_{s'z \neq (s_{d+1} z_{d+1})} \Gamma(\chi^{s'z}_{s_d a_d}(H_d))} \\
&= \frac{\Gamma(T)}{\Gamma(T+1)} \frac{\Gamma(\chi^{s_{d+1} z_{d+1}}_{s_d a_d}(H_{d+1}))}{\Gamma(\chi^{s_{d+1} z_{d+1}}_{s_d a_d}(H_d))} = \frac{\Gamma(T)}{\Gamma(T+1)} \frac{\Gamma(N+1)}{\Gamma(N)}
\end{aligned}
$$

Now, the gamma function has the property that $\Gamma(x+1) = x\Gamma(x)$ [DeGroot, 2004], which means that we get

$$
= \frac{\Gamma(T)}{T\Gamma(T)} \frac{N\Gamma(N)}{\Gamma(N)} = \frac{N}{T}.
$$

Therefore we get

$$
\frac{B(\chi_{s_d a_d}(H_d))}{B(\chi_{s_d a_d}(H_{d+1}))} = \frac{\chi^{s_{d+1} z_{d+1}}_{s_d a_d}(H_d)}{\sum_{(s',z)} \chi^{s'z}_{s_d a_d}(H_d)}
$$

and thus

$$
\tilde{P}^{\pi}(H_{d+1}) = \tilde{P}^{\pi}(H_d)\pi(a_d|h_d) \frac{\chi^{s_{d+1} z_{d+1}}_{s_d a_d}(H_d)}{\sum_{(s',z)} \chi^{s'z}_{s_d a_d}(H_d)}. \tag{4}
$$

the r.h.s. of this equation is identical to (1) except for the difference in between $\tilde{P}^{\pi}(H_d)$ and $P^{\pi}(H_d)$. This can be resolved by forward induction with base step: $\tilde{P}^{\pi}(H_0) = b_0(\langle s_0, \chi_0, \psi_0 \rangle) = P^{\pi}(H_0)$, and the induction step (show $\tilde{P}^{\pi}(H_{d+1}) = P^{\pi}(H_{d+1})$ given $\tilde{P}^{\pi}(H_d) = P^{\pi}(H_d)$) directly following from (1) and (4). Therefore we can conclude that $\forall_d \quad \tilde{P}^{\pi}(H_d) = P^{\pi}(H_d)$.

Since Lemma 4 already established that $\forall_{H_d} \quad \tilde{P}^{\pi}_{K_d}(H_d) \xrightarrow{p} \tilde{P}^{\pi}(H_d)$, we directly have

$$\forall_{H_d} \quad \tilde{P}^{\pi}_{K_d}(H_d) \xrightarrow{p} P^{\pi}(H_d),$$

thus proving the result. □

The proof depends on the following lemma:

Lemma 4. *The full-history RS-BA-POMDP rollout distribution converges in probability to the following quantity:*

$$\forall_{H_d} \quad \tilde{P}^{\pi}_{K_d}(H_d) \xrightarrow{p} b_0(s_0) \left[\prod_{t=1}^{d} \pi(a_{t-1}|h_{t-0})\right] \left[\prod_{\langle s,a\rangle} \frac{B(\chi_{sa}(H_0))}{B(\chi_{sa}(H_d)}\right] \tag{5}$$

with $B(\alpha) = \frac{\Gamma(\alpha_1+\ldots+\alpha_k)}{\Gamma(\alpha_1)\cdot\ldots\cdot\Gamma(\alpha_k)}$ *the normalization term of a Dirichlet distribution with parametric vector* α*.*

Proof. Via the weak law of large numbers, we have that the empirical mean of a random variable converges in probability to its expectation.

$$\forall H_d \qquad \frac{1}{K_d}\sum_{i=1}^{K_d} \mathbb{I}_{\left\{H_d=H_d^{(i)}\right\}} \xrightarrow{p} \mathbf{E}\left[\mathbb{I}_{\left\{H_d=H_d^{(i)}\right\}}\right]$$

This expectation can be rewritten as follows

$$\mathbf{E}\left[\mathbb{I}_{\left\{H_d=H_d^{(i)}\right\}}\right] = \sum_{H_d^{(i)}} \tilde{P}^{\pi}\left(H_d^{(i)}\right) \mathbb{I}_{\left\{H_d=H_d^{(i)}\right\}} = \tilde{P}^{\pi}\left(H_d\right) \tag{6}$$

where $\tilde{P}^{\pi}(H_d)$ denotes the (true, non-empirical) probability that the RS-BA-POMDP rollout generates full history H_d. This is an expectation over the root sampled model $\dot{D}$:

$$\begin{aligned}
\tilde{P}^{\pi}(H_d) &= \int \tilde{P}^{\pi}\left(H_d|\dot{D}\right) Dir(\dot{D}|\dot{\chi}) d\dot{D} && (7)\\
&= \int \left[\mathrm{Cons}(H_0, H_d) \prod_{t=1}^{d} \dot{D}(s_t z_t|s_{t-1}, a_{t-1}) \pi(a_{t-1}|h_{t-1})\right] Dir(\dot{D}|\dot{\chi}) d\dot{D} && (8)\\
&= \mathrm{Cons}(H_0, H_d) \left[\prod_{t=1}^{d} \pi(a_{t-1}|h_{t-1})\right] \left(\int \left[\prod_{t=1}^{d} \dot{D}(s_t z_t|s_{t-1}, a_{t-1})\right] Dir(\dot{D}|\dot{\chi}) d\dot{D}\right) && (9)
\end{aligned}$$

Where $\mathrm{Cons}(H_0, H_d)$ is a term that indicates whether (takes value 1 if) H_d is consistent with the full history at the root H_0.[1]

[1] An earlier version of this proof [anonymous] contained a term $b_0(s_0)$ instead of $\mathrm{Cons}(H_0, H_d)$. This was incorrect since it failed to recognize that this proof assumes H_0 to be fixed. See also the remark on page 2.

Now we can exploit the fact that only the Dirichlet for the transitions specified by H_d matter.

$$\int \left[\prod_{t=1}^{d} \dot{D}(s_t z_t | s_{t-1}, a_{t-1})\right] Dir(\dot{D}|\chi_0) d\dot{D} \tag{10}$$

$=$\{split up the integral over one big vector into integrals over smaller vectors\}

$$\int \dots \int \left[\prod_{t=1}^{d} \dot{D}^{s_t,z_t}_{s_{t-1},a_{t-1}}\right] \left[\prod_{\langle s,a\rangle} Dir(\dot{D}_{sa}|\chi_{sa}(H_0))\right] d\dot{D}_{s^1a^1} \dots d\dot{D}_{s^{|\mathcal{S}|}a^{|\mathcal{A}|}} \tag{11}$$

$=$\{reorder the transition probabilities: $\Delta_\chi^{sas'z}(H_d)$is the number of occurences of (s,a,s',z)in H_d\}

$$\int \dots \int \left[\prod_{\langle s,a\rangle} \prod_{\langle s',z\rangle} \left(\dot{D}^{s'z}_{sa}\right)^{\Delta_\chi^{sas'z}(H_d)}\right] \left[\prod_{\langle s,a\rangle} Dir(\dot{D}_{sa}|\chi_{sa}(H_0))\right] d\dot{D}_{s^1a^1} \dots d\dot{D}_{s^{|\mathcal{S}|}a^{|\mathcal{A}|}} \tag{12}$$

$$= \int \dots \int \left[\prod_{\langle s,a\rangle} \prod_{\langle s',z\rangle} \left(\dot{D}^{s'z}_{sa}\right)^{\Delta_\chi^{sas'z}(H_d)}\right] \left[\prod_{\langle s,a\rangle} B(\dot{\chi}_{sa}) \prod_{\langle s',z\rangle} \left(\dot{D}^{s'z}_{sa}\right)^{\chi_0^{sas'z}-1}\right] d\dot{D}_{s^1a^1} \dots d\dot{D}_{s^{|\mathcal{S}|}a^{|\mathcal{A}|}} \tag{13}$$

$$= \int \dots \int \left[\prod_{\langle s,a\rangle} \left(\left[\prod_{\langle s',z\rangle} \left(\dot{D}^{s'z}_{sa}\right)^{\Delta_\chi^{sas'z}(H_d)}\right] \left[B(\dot{\chi}_{sa}) \prod_{\langle s',z\rangle} \left(\dot{D}^{s'z}_{sa}\right)^{\chi_0^{sas'z}-1}\right]\right)\right] d\dot{D}_{s^1a^1} \dots d\dot{D}_{s^{|\mathcal{S}|}a^{|\mathcal{A}|}} \tag{14}$$

$$= \int \dots \int \left[\prod_{\langle s,a\rangle} \left(B(\dot{\chi}_{sa}) \left[\prod_{\langle s',z\rangle} \left(\dot{D}^{s'z}_{sa}\right)^{\Delta_\chi^{sas'z}(H_d)}\right] \left[\prod_{\langle s',z\rangle} \left(\dot{D}^{s'z}_{sa}\right)^{\chi_0^{sas'z}-1}\right]\right)\right] d\dot{D}_{s^1a^1} \dots d\dot{D}_{s^{|\mathcal{S}|}a^{|\mathcal{A}|}} \tag{15}$$

$$= \int \dots \int \left[\prod_{\langle s,a\rangle} B(\dot{\chi}_{sa}) \prod_{\langle s',z\rangle} \left(\dot{D}^{s'z}_{sa}\right)^{\chi_0^{sas'z}-1+\Delta_\chi^{sas'z}(H_d)}\right] d\dot{D}_{s^1a^1} \dots d\dot{D}_{s^{|\mathcal{S}|}a^{|\mathcal{A}|}} \tag{16}$$

Now we reverse the order of integration and multiplication, which is possible since the different s,a pairs over which we integrate are disjoint.[2] We obtain:

$$= \prod_{\langle s,a\rangle} B(\chi_{sa}(H_0)) \int \prod_{\langle s',z\rangle} \left(\dot{D}_{sa}(s',z)\right)^{\chi_0^{sas'z}+\Delta_\chi^{sas'z}(H_d)-1} d\dot{D}_{sa} \tag{17}$$

$=$\{since we integrate over the entire vector $\dot{D}_{sa}$, the integral equals $1/B(\chi_{sa}(H_0)+\Delta_\chi^{sa}(H_d))$\}

$$\prod_{\langle s,a\rangle} B(\chi_{sa}(H_0)) \frac{1}{B(\chi_{sa}(H_0)+\Delta_\chi^{sa}(H_d))} \tag{18}$$

$$= \prod_{\langle s,a\rangle} \frac{B(\chi_{sa}(H_0))}{B(\chi_{sa}(H_d))} \tag{19}$$

Therefore

$$\tilde{P}^\pi(H_d) = \text{Cons}(H_0, H_d) \left[\prod_{t=0}^{d-1} \pi(a_t|h_t)\right] \left[\prod_{\langle s,a\rangle} \frac{B(\chi_{sa}(H_0))}{B(\chi_{sa}(H_d))}\right], \tag{20}$$

proving (5). □

[2] E.g, consider two sets $A_1 = \left\{a_1^{(1)}, a_1^{(2)}\right\}$ and $A_2 = \left\{a_2^{(1)}, a_2^{(2)}, a_2^{(3)}\right\}$. Equation (16) is of the same form as

$$\begin{aligned}\sum_{a_1\in A_1}\sum_{a_2\in A_2}\prod_{i=1}^{2} a_i &= \sum_{a_1\in A_1}\sum_{a_2\in A_2} a_1a_2 = a_1^{(1)}a_2^{(1)} + a_1^{(1)}a_2^{(2)} + a_1^{(1)}a_2^{(3)} + a_1^{(2)}a_2^{(1)} + a_1^{(2)}a_2^{(2)} + a_1^{(2)}a_2^{(3)} \\ &= a_1^{(1)}\left(a_2^{(1)} + a_2^{(2)} + a_2^{(3)}\right) + a_1^{(2)}\left(a_2^{(1)} + a_2^{(2)} + a_2^{(3)}\right) = \left(a_1^{(1)} + a_1^{(2)}\right)\left(a_2^{(1)} + a_2^{(2)} + a_2^{(3)}\right) \\ &= \left[\sum_{a_1\in A_1} a_1\right]\left[\sum_{a_2\in A_2} a_2\right] = \prod_{i=1}^{2}\sum_{a_i\in A_i} a_i\end{aligned}$$

References

Morris H. DeGroot. *Optimal Statistical Decisions*. Wiley-Interscience, April 2004.