POMCP Lemma 1 Correction

Pedro Sandoval-Segura

October 2019

1 Introduction

In Monte-Carlo Planning in Large POMDPs by Silver et al. [1] Lemma 1 intends to show that given a POMDP, the value function $\tilde{V}^\pi(h)$ of the derived MDP, which uses every history as a state, is equal to the value function $V^\pi(h)$ of the POMDP for all policies π. We find the claim is true, but that the proof requires a correction.

1.1 Notation

Silver et al. define the set of states as $\mathcal{S}$, set of actions as $\mathcal{A}$, transition probabilities as $\mathcal{P}^a_{s,s'}$, return/reward as $\mathcal{R}^a_s$, and observation probabilities as $\mathcal{Z}^a_{s',o}$.

1.2 Verbatim Proof

Given a POMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{Z})$, consider the derived MDP with histories as states, $\tilde{\mathcal{M}} = (\mathcal{H}, \mathcal{A}, \tilde{\mathcal{P}}, \tilde{\mathcal{R}})$ where

$$\tilde{\mathcal{P}}^a_{h,hao} = \sum_{s\in\mathcal{S}} \sum_{s'\in\mathcal{S}} \mathcal{B}(s,h)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}$$

$$\tilde{\mathcal{R}}^a_h = \sum_{s\in\mathcal{S}} \mathcal{B}(s,h)\mathcal{R}^a_s$$

Then the value function $\tilde{V}^\pi(h)$ of the derived MDP is equal to the value function $V^\pi(h)$ of the POMDP, $\forall\pi\ \tilde{V}^\pi(h) = V^\pi(h)$.

Proof. By backward induction on the Bellman equation, starting from the horizon,

$$\begin{aligned} V^\pi(h) &= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \sum_{o\in\mathcal{O}} \mathcal{B}(s,h)\pi(h,a)(\mathcal{R}^a_s + \gamma\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}V^\pi(hao)) \\ &= \sum_{a\in\mathcal{A}} \sum_{o\in\mathcal{O}} \pi(h,a)(\tilde{\mathcal{R}}^a_h + \gamma\tilde{\mathcal{P}}^a_{h,hao}\tilde{V}^\pi(hao)) \\ &= \tilde{V}^\pi(h) \end{aligned}$$

□

1.3 An Attempt at Verification

Again, consider a POMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{Z})$, consider the derived MDP with histories as states, $\tilde{\mathcal{M}} = (\mathcal{H}, \mathcal{A}, \tilde{\mathcal{P}}, \tilde{\mathcal{R}})$. We compute:

Proof.

$$
\begin{aligned}
V^\pi(h) &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\sum_{o\in\mathcal{O}} \mathcal{B}(s,h)\pi(h,a)(\mathcal{R}^a_s + \gamma\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}V^\pi(hao)) && \text{(by definition)}\\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} \mathcal{B}(s,h)\pi(h,a)(\mathcal{R}^a_s + \gamma\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}V^\pi(hao)) && \text{(rearrange sums)}\\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} (\mathcal{B}(s,h)\pi(h,a)\mathcal{R}^a_s + \mathcal{B}(s,h)\pi(h,a)\gamma\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}V^\pi(hao)) && \text{(distribute)}\\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}(\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} \mathcal{B}(s,h)\pi(h,a)\mathcal{R}^a_s + \sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\mathcal{B}(s,h)\pi(h,a)\gamma\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}V^\pi(hao)) && \text{(distribute sums)}\\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}(\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} \mathcal{B}(s,h)\pi(h,a)\mathcal{R}^a_s + \gamma V^\pi(hao)\pi(h,a)\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}}\mathcal{B}(s,h)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}) && \text{(factor)}\\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}(\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} \mathcal{B}(s,h)\pi(h,a)\mathcal{R}^a_s + \gamma V^\pi(hao)\pi(h,a)\tilde{\mathcal{P}}^a_{h,hao}) && \text{(by definition)}\\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}(\sum_{s'\in\mathcal{S}}\sum_{s\in\mathcal{S}} \mathcal{B}(s,h)\pi(h,a)\mathcal{R}^a_s + \gamma V^\pi(hao)\pi(h,a)\tilde{\mathcal{P}}^a_{h,hao}) && \text{(rearrange sums)}\\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}(\sum_{s'\in\mathcal{S}} \pi(h,a)\sum_{s\in\mathcal{S}}\mathcal{B}(s,h)\mathcal{R}^a_s + \gamma V^\pi(hao)\pi(h,a)\tilde{\mathcal{P}}^a_{h,hao}) && \text{(factor)}\\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}(\sum_{s'\in\mathcal{S}} \pi(h,a)\tilde{\mathcal{R}}^a_h + \gamma V^\pi(hao)\pi(h,a)\tilde{\mathcal{P}}^a_{h,hao}) && \text{(by definition)}\\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}(\pi(h,a)\tilde{\mathcal{R}}^a_h\sum_{s'\in\mathcal{S}} 1 + \gamma V^\pi(hao)\pi(h,a)\tilde{\mathcal{P}}^a_{h,hao}) && \text{(factor)}\\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}(\pi(h,a)\tilde{\mathcal{R}}^a_h|\mathcal{S}| + \gamma V^\pi(hao)\pi(h,a)\tilde{\mathcal{P}}^a_{h,hao}) && \text{(simplify)}\\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}\pi(h,a)(\tilde{\mathcal{R}}^a_h|\mathcal{S}| + \gamma V^\pi(hao)\tilde{\mathcal{P}}^a_{h,hao}) && \text{(factor)}\\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}\pi(h,a)(\tilde{\mathcal{R}}^a_h|\mathcal{S}| + \gamma\tilde{\mathcal{P}}^a_{h,hao}V^\pi(hao)) && \text{(rearrange term)}
\end{aligned}
$$

□

which is **not equal** to

$$\sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}}\pi(h,a)(\tilde{\mathcal{R}}^a_h + \gamma\tilde{\mathcal{P}}^a_{h,hao}\tilde{V}^\pi(hao))$$

which would be the definition of $\tilde{V}^\pi(h)$. It's a similar expression, but there is not a clear path to removing the factor of $|\mathcal{S}|$.

1.4 Correction Explanation

It seems like the original expression for $V^\pi(h)$ presented in the POMCP paper is incorrect. The Bellman equation for evaluating a state $s \in \mathcal{S}$, given policy π, action set $\mathcal{A}$, transition function

T, and reward function R is

$$V^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} T(s,a,s')(R(s,a,s') + \gamma V^\pi(s'))$$

Now, let's adapt this equation to work for histories and partially-observable domains (by using our observation probabilities). First, we know $T(s,a,s') = Pr(s'|s,a) = P^a_{s,s'}$ and $R(s,a,s') = R^a_s$. In our partially observable setting, we don't know the true state, but we do have a belief state $\mathcal{B}(s,h)$ and so we can use this distribution to compute an expectation over s. Finally, because we are using histories which include observations, we change $\pi(a|s)$ to $\pi(h,a)$ and we need to perform an expectation over observations for the term being multiplied by γ. Thus, we now have:

$$\begin{aligned}
V^\pi(h) &= \sum_{s\in\mathcal{S}}\sum_{a\in A} \mathcal{B}(s,h)\pi(h,a) \sum_{s'\in S} \mathcal{P}^a_{s,s'}(\mathcal{R}^a_s + \gamma \sum_{o\in\mathcal{O}} \mathcal{Z}^a_{s',o} V^\pi(hao)) \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in A} \mathcal{B}(s,h)\pi(h,a)(\mathcal{R}^a_s \sum_{s'\in S} \mathcal{P}^a_{s,s'} + \gamma \sum_{s'\in S} \mathcal{P}^a_{s,s'} \sum_{o\in\mathcal{O}} \mathcal{Z}^a_{s',o} V^\pi(hao)) \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in A} \mathcal{B}(s,h)\pi(h,a)(\mathcal{R}^a_s + \gamma \sum_{s'\in S}\sum_{o\in\mathcal{O}} \mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o} V^\pi(hao))
\end{aligned}$$

We can make this expression more similar to what's written in the publication by the series of steps below:

$$\begin{aligned}
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mathcal{B}(s,h)\pi(h,a)(\sum_{s'\in\mathcal{S}}\sum_{o\in\mathcal{O}} \mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}\mathcal{R}^a_s + \gamma \sum_{s'\in\mathcal{S}}\sum_{o\in\mathcal{O}} \mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o} V^\pi(hao)) && \text{(multiply by 1)} \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mathcal{B}(s,h)\pi(h,a) \sum_{s'\in\mathcal{S}}\sum_{o\in\mathcal{O}} (\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}\mathcal{R}^a_s + \gamma \mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o} V^\pi(hao)) && \text{(rearrange sums)} \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mathcal{B}(s,h)\pi(h,a)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o} \sum_{s'\in\mathcal{S}}\sum_{o\in\mathcal{O}} (\mathcal{R}^a_s + \gamma V^\pi(hao)) && \text{(factor)} \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\sum_{o\in\mathcal{O}} \mathcal{B}(s,h)\pi(h,a)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o} (\mathcal{R}^a_s + \gamma V^\pi(hao)) && \text{(rearrange sums)}
\end{aligned}$$

which looks more similar to the expression that begins the proof of Lemma 1 in the current instance of the paper.

Similarly, the expression for $\tilde{V}^\pi(h)$ in the paper should change. We need to move in expectation over o as follows:

$$\tilde{V}^\pi(h) = \sum_{a\in\mathcal{A}} \pi(h,a)(\tilde{\mathcal{R}}^a_h + \gamma \sum_{o\in\mathcal{O}} \tilde{\mathcal{P}}^a_{h,hao} \tilde{V}^\pi(hao))$$

1.5 Corrected Proof of Lemma 1

Given a POMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{Z})$, consider the derived MDP with histories as states, $\tilde{\mathcal{M}} = (\mathcal{H}, \mathcal{A}, \tilde{\mathcal{P}}, \tilde{\mathcal{R}})$ where

$$\tilde{\mathcal{P}}^a_{h,hao} = \sum_{s\in\mathcal{S}} \sum_{s'\in\mathcal{S}} \mathcal{B}(s,h)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}$$

$$\tilde{\mathcal{R}}^a_h = \sum_{s\in\mathcal{S}} \mathcal{B}(s,h)\mathcal{R}^a_s$$

Then the value function $\tilde{V}^\pi(h)$ of the derived MDP is equal to the value function $V^\pi(h)$ of the POMDP, $\forall\pi\ \tilde{V}^\pi(h) = V^\pi(h)$.

Proof. By backward induction on the Bellman equation, starting from the horizon,

$$
\begin{aligned}
V^\pi(h) &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\sum_{o\in\mathcal{O}} \mathcal{B}(s,h)\pi(h,a)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}(\mathcal{R}^a_s + \gamma V^\pi(hao)) \\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}} \pi(h,a) \sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} (\mathcal{B}(s,h)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}\mathcal{R}^a_s + \gamma\mathcal{B}(s,h)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}V^\pi(hao)) && \text{(factor)} \\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}} \pi(h,a)((\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} \mathcal{B}(s,h)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}\mathcal{R}^a_s) + \gamma(\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} \mathcal{B}(s,h)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}V^\pi(hao))) && \text{(distribute)} \\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}} \pi(h,a)((\sum_{s'\in\mathcal{S}}\sum_{s\in\mathcal{S}} \mathcal{B}(s,h)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}\mathcal{R}^a_s) + \gamma V^\pi(hao)(\sum_{s\in\mathcal{S}}\sum_{s'\in\mathcal{S}} \mathcal{B}(s,h)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o})) && \text{(factor)} \\
&= \sum_{a\in\mathcal{A}}\sum_{o\in\mathcal{O}} \pi(h,a)((\sum_{s'\in\mathcal{S}}\sum_{s\in\mathcal{S}} \mathcal{B}(s,h)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}\mathcal{R}^a_s) + \gamma V^\pi(hao)\tilde{\mathcal{P}}^a_{h,hao}) && \text{(by definition)} \\
&= \sum_{a\in\mathcal{A}} \pi(h,a)((\sum_{s'\in\mathcal{S}}\sum_{s\in\mathcal{S}}\sum_{o\in\mathcal{O}} \mathcal{B}(s,h)\mathcal{P}^a_{s,s'}\mathcal{Z}^a_{s',o}\mathcal{R}^a_s) + \gamma(\sum_{o\in\mathcal{O}} V^\pi(hao)\tilde{\mathcal{P}}^a_{h,hao})) && \text{(factor)} \\
&= \sum_{a\in\mathcal{A}} \pi(h,a)((\sum_{s\in\mathcal{S}} \mathcal{B}(s,h)\mathcal{R}^a_s \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{s,s'} \sum_{o\in\mathcal{O}} \mathcal{Z}^a_{s',o}) + \gamma(\sum_{o\in\mathcal{O}} V^\pi(hao)\tilde{\mathcal{P}}^a_{h,hao})) && \text{(factor)} \\
&= \sum_{a\in\mathcal{A}} \pi(h,a)((\sum_{s\in\mathcal{S}} \mathcal{B}(s,h)\mathcal{R}^a_s \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{s,s'}) + \gamma(\sum_{o\in\mathcal{O}} V^\pi(hao)\tilde{\mathcal{P}}^a_{h,hao})) && \text{(sum of distribution)} \\
&= \sum_{a\in\mathcal{A}} \pi(h,a)((\sum_{s\in\mathcal{S}} \mathcal{B}(s,h)\mathcal{R}^a_s) + \gamma(\sum_{o\in\mathcal{O}} V^\pi(hao)\tilde{\mathcal{P}}^a_{h,hao})) && \text{(sum of distribution)} \\
&= \sum_{a\in\mathcal{A}} \pi(h,a)(\tilde{\mathcal{R}}^a_h + \gamma \sum_{o\in\mathcal{O}} V^\pi(hao)\tilde{\mathcal{P}}^a_{h,hao}) && \text{(by definition)} \\
&= \sum_{a\in\mathcal{A}} \pi(h,a)(\tilde{\mathcal{R}}^a_h + \gamma \sum_{o\in\mathcal{O}} \tilde{\mathcal{P}}^a_{h,hao}\tilde{V}^\pi(hao)) && \text{(inductive hypothesis)} \\
&= \tilde{V}^\pi(h)
\end{aligned}
$$

□

as desired. Thus, the suggested changes to the paper are the highlighted first and final lines of this proof. The original claim still stands.

1.6 Acknowledgements

Special thanks to Olivia Watkins for insights on the Bellman equation.

References

[1] David Silver and Joel Veness. Monte-carlo planning in large pomdps. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, pages 2164–2172, USA, 2010. Curran Associates Inc.