Reinforcement Learning in Many-Agent Settings Under Partial Observability: Supplementary File

Keyang He  Prashant Doshi  Bikramjit Banerjee

1 THINC Lab, Department of Computer Science, University of Georgia, Athens, GA, USA
2 School of Computing Sciences and Engineering, University of Southern Mississippi, Hattiesburg, MS, USA

1 DYNAMIC PROGRAMMING ALGORITHM

Algorithm 1 Computing configuration distribution

\[ Pr(C|b_0(M_1), b_0(M_2), \ldots, b_0(M_N)) \]

Require: \( b_0(M_1), b_0(M_2), \ldots, b_0(M_N) \)
Ensure: \( P_N \), which is the distribution \( Pr(C|a^{1-o}) \) represented as a trie.

Initialize \( c_0^{a_i} \leftarrow (0, 0, \ldots, 0) \), \( P_0[c_0^{a_i}] \leftarrow 1.0 \)
for \( k = 1 \) to \( N \) do
   Initialize \( P_k \) to be an empty trie
   for \( c_{k-1}^{a_i} \) from \( P_{k-1} \) do
      for \( a_k^{a_i} \in A^{a_i}_k \) such that \( \pi_k^{a_i}(a_k^{a_i}) > 0 \) do
         \( c_k^{a_i} \leftarrow c_{k-1}^{a_i} \)
         if \( a_k^{a_i} \neq 0 \) then
            \( c_k^{a_i}[(a_k^{a_i})] \leftarrow 1 \)
         end if
      end if
   end if
   if \( P_k[c_k^{a_i}] \) does not exist then
      \( P_k[c_k^{a_i}] \leftarrow 0 \)
   end if
   \( P_k[c_k^{a_i}] \leftarrow P_{k-1}[c_{k-1}^{a_i}] \times \pi_k^{a_i}(a_k^{a_i}) \)
end for
end for
return \( P_N \)

2 PROOF OF PROPOSITION 1

Here we assume a common model of noise, \( Pr(a_j^{o}|a_k^{e}) \), where the subject agent observes action \( a_j^{o} \) from another agent when the latter executed action \( a_k^{e} \), as

\[ Pr(a_j^{o}|a_k^{e}) = \begin{cases} 1 - \delta & \text{if } a_j^{o} = a_k^{e} \\ \delta/|A|-1 & \text{otherwise} \end{cases} \]

for some small \( \delta \). The effect of such noise from the private observation of an individual agent’s action can be aggregated over \( N \) agents in terms of \( \delta \) as follows. Suppose the observed configuration, \( \omega^{o} \), is \( C^{o} = (#a_1^{o}, #a_2^{o}, \ldots, #a_A^{o}) \), and the true configuration is \( C^{e} = (#a_1^{e}, #a_2^{e}, \ldots, #a_A^{e}) \). Then the probability of an error in the observation of a configuration is

\[ P(\text{error}) = \sum_{C^{e}} \sum_{C^{o} \neq C^{e}} Pr(C^{o} \land C^{e}) \]

\[ = \sum_{C^{e}} \sum_{C^{o} \neq C^{e}} Pr(C^{o}|C^{e})Pr(C^{e}) \]

where

\[ Pr(C^{e}) = \prod_{i} \theta_{i}^{#a_i^{e}}, \text{ and} \]

\[ Pr(C^{o}|C^{e}) = \prod_{(j,k) \in A \times A} P(a_j^{o}|a_k^{e})^{n_{jk}} \]

\[ s.t. \left( \sum_{j} n_{jk} = \#a_k^{e} \right) \land \left( \sum_{k} n_{jk} = \#a_j^{o} \right) \]

Let \( m^{o,e} = \min\{\#a_i^{o}, \#a_i^{e}\} \). Then \( Pr(C^{o}|C^{e}) \) can be maximized by setting the diagonal of the matrix \( n_{jk} \) as \( n_{ii} = m^{o,e} \), and distributing the remaining weight \( N - \sum_{i} m^{o,e} \) to the off-diagonal positions while satisfying Eq. \( 2 \). This yields

\[ Pr(C^{o}|C^{e}) \leq (1 - \delta) \sum_{j} m^{o,e} \left( \frac{\delta}{|A|-1} \right)^{N-\sum_{j} m^{o,e}} \]

\[ \leq (1 - \delta)^{N-1} \left( \frac{\delta}{|A|-1} \right) \]

in order to ensure that \( C^{o} \neq C^{e} \). Furthermore, the number of solutions of Eq. \( 2 \) is \( \leq \prod_{i} (m^{o,e} + 1) = O(N^{|A|}) \). Hence

\[ P(\text{error}) \leq N^{|A|}(1 - \delta)^{N-1} \left( \frac{\delta}{|A|-1} \right) \]

The above is a decreasing function of \( N \) when \( N > \frac{|A|}{\log(1/1-\delta)} \).

3 POLICY VALUE WITH RESPECT TO EPISODES

We choose to use time in hours as metric for demonstrating efficiency of tested algorithms. We provide additional plots

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).
Figure 1: Cumulative reward of learned policies in (a) tree structure, (b) star structure, and (c) fully connected structure. (d) Win rate against pre-trained agents in the MAgent battlefield domain.

that use episodes as metric in Fig. [II] QMIX and MF-AC do not converge to optimal policy given same amount of episodes as IA2C-BU, however, it only takes QMIX and MF-AC about one third of the time to finish one episode compared to IA2C-BU.