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

Keyang He[1] **Prashant Doshi**[1] **Bikramjit Banerjee**[2]

[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 $Pr(\mathcal{C}|b_0(M_1), b_0(M_2), \ldots, b_0(M_N))$

```
Require: $\langle b_0(M_1), b_0(M_2), \ldots, b_0(M_N) \rangle$
Ensure: $P_N$, which is the distribution $Pr(\boldsymbol{\mathcal{C}}^{\boldsymbol{a}_{-0}})$ represented as a trie.
  Initialize $c_0^{a_i} \leftarrow (0, \ldots, 0)$, $P_0[c_0^{a_i}] \leftarrow 1.0$
  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 $a_k^{a_i} \in A_k^{a_i}$ such that $\pi_k^{a_i}(a_k^{a_i}) > 0$ do
        $c_k^{a_i} \leftarrow c_{k-1}^{a_i}$
        if $a_k^{a_i} \neq \emptyset$ then
          $c_k^{a_i}(a_k^{a_i}) \overset{+}{\leftarrow} 1$
        end if
        if $P_k[c_k^{a_i}]$ does not exist then
          $P_k[c_k^{a_i}] \leftarrow 0$
        end if
        $P_k[c_k^{a_i}] \overset{+}{\leftarrow} P_{k-1}[c_{k-1}^{a_i}] \times \pi_k^{a_i}(a_k^{a_i})$
      end for
    end for
  end for
  return $P_N$
```

2 PROOF OF PROPOSITION 1

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

$$P(a_j^o|a_k^e) = \begin{cases} 1-\delta & if\ a_j^o = a_k^e \\ \frac{\delta}{|A|-1} & otherwise \end{cases} \qquad (1)$$

for some small δ. The effect of such noise from the private observation of an individual agent's action can be aggregated over N agents in terms of δ as follows. Suppose the observed configuration, ω_0', is $\mathcal{C}^o = (\#a_1^o, \#a_2^o, \ldots, \#a_{|A|}^o)$, and the true configuration is $\mathcal{C}^e = (\#a_1^e, \#a_2^e, \ldots, \#a_{|A|}^e)$. Then the probability of an error in the observation of a configuration is

$$\begin{aligned} P(error) &= \sum_{\mathcal{C}^e} \sum_{\mathcal{C}^o \neq \mathcal{C}^e} P(\mathcal{C}^o \wedge \mathcal{C}^e) \\ &= \sum_{\mathcal{C}^e} \sum_{\mathcal{C}^o \neq \mathcal{C}^e} P(\mathcal{C}^o|\mathcal{C}^e) P(\mathcal{C}^e) \end{aligned}$$

where

$$\begin{aligned} P(\mathcal{C}^e) &= \prod_i \theta_i^{\#a_i^e},\ and \\ P(\mathcal{C}^o|\mathcal{C}^e) &= \prod_{(j,k) \in A \times A} P(a_j^o|a_k^e)^{n_{jk}} \\ & s.t.\ (\sum_j n_{jk} = \#a_k^e) \wedge (\sum_k n_{jk} = \#a_j^o) \qquad (2) \end{aligned}$$

Let $m_i^{oe} = \min\{\#a_i^o, \#a_i^e\}$. Then $P(\mathcal{C}^o|\mathcal{C}^e)$ can be maximized by setting the diagonal of the matrix $[n_{jk}]$ as $n_{ii} = m_i^{oe}$, and distributing the remaining weight $N - \sum_i m_i^{oe}$ to the off-diagonal positions while satisfying Eq. 2. This yields

$$\begin{aligned} P(\mathcal{C}^o|\mathcal{C}^e) &\leq (1-\delta)^{\sum_i m_i^{oe}} \left(\frac{\delta}{|A|-1}\right)^{N - \sum_i m_i^{oe}} \\ &\leq (1-\delta)^{N-1} \left(\frac{\delta}{|A|-1}\right) \end{aligned}$$

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

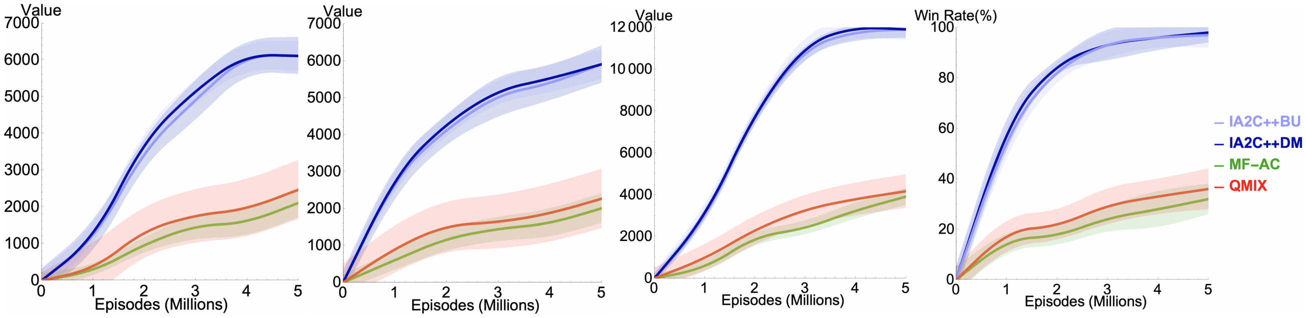
$$P(error) \leq N^{|A|} (1-\delta)^{N-1} \left(\frac{\delta}{|A|-1}\right)$$

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).



(a) Tree (b) Star (c) Fully Connected (d) Battlefield

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. 1. 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.