Reinforcement Learning in Multi-Agent Systems with Partial History Sharing

Jalal Arabneydi and Aditya Mahajan

Electrical and Computer Engineering Department, McGill University

Website: www.cim.mcgill.ca/jarabney

Date: June 9th, 2015
Motivation

- We are interested in systems with multiple agents (decision makers) that wish to cooperate in order to accomplish a common task while
  - agents have different information (decentralized information)
  - agents do not know the complete model of the system i.e., they may only know the partial model or may not know the model at all.

- Multi-agent systems arise in various applications: Networked control systems, Robotics, Communication networks, Transportation networks, Sensor networks, Smart grids, Economics, etc.

- Advantages of multi-agent (decentralized) over single-agent (centralized) systems:
  - distributes computational resources and capacities.
  - provides robustness, maintainability, and flexibility.
  - implements the solution efficiently (physically and economically).
Challenges

- The **discrepancy in perspectives** makes establishing cooperation among agents conceptually challenging.

- In general, these problems belong to **NEXP complexity** class.

- Finding team-optimal solution is more challenging when agents have only **partial knowledge** or **no knowledge** of system model.
Problem Formulation

- Consider a system with finite-valued variables that consists of $n \in \mathbb{N}$ agents.

- State of system $S_t \in S$ and action of agent $i$: $A_t^i \in \mathcal{A}^i$, where $t \in \mathbb{N}$ denotes time.

- Observation of agent $i$: $O_t^i = h^i(S_t, A_1^{t-1}, \ldots, A_{n-1}^{t-1}, V_t)$

- Information of agent $i$: $I_t^i \subseteq \{O_1^{1:t}, \ldots, O_n^{1:t}, A_1^{1:t-1}, \ldots, A_{n-1}^{1:t-1}\}$.
Problem Formulation

- Control law of agent $i$: $A_t^i = g_t^i(I_t^i)$.

- Control strategy $g := (g_1, g_2, \ldots)$, where $g_t := (g_t^1, \ldots, g_t^n)$.

- Reward given control strategy $g$ : $J(g) = \mathbb{E}^g \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r(S_t, A_t^1, \ldots, A_t^n) \right]$.

- Agents observe the immediate reward.

- Objective: Develop a (model-based or model-free) reinforcement learning algorithms that guarantees an $\epsilon$-optimal strategy $g^*$ i.e. $J^* - J(g^*) \leq \epsilon$. 

(J. Arabneydi, www.cim.mcgill.ca/ jarabney)
Partial History Sharing

Split $l_t^i = \{C_t, M_t^i\}$, where $C_t = \bigcap_i l_t^i$ is common information and $M_t^i = l_t^i \setminus C_t$ is local information.

\[\text{(A1) Common information is nested: } C_{t+1} = \{C_t, Z_t\}, \text{ where } Z_t := C_{t+1} \setminus C_t \text{ is common observation such that } C_{t+1} = Z_{1:t}.\]

\[\text{(A2) The update of local information } M_{t+1}^i \subseteq \{M_t^i, A_t^i, O_{t+1}^i\}.\]

\[\text{(A3) The size of } Z_t \text{ and the size of } M_t^i, \forall i, \text{ are uniformly bounded in time } t.\]

- (A1), (A2), and (A3) are mild conditions. Also, $C_t$ is allowed to be empty set.
- A large class of multi-agent systems have partial history sharing such as: delayed sharing, control sharing, mean-field sharing, etc.
Methodology

Our approach has two main steps:

- **Step 1)** Common Information Approach
- **Step 2)** Approximate RL algorithm for centralized (single-agent) POMDPs
This approach guarantees \( \epsilon \)-optimality performance.

It encompasses a large class of multi-agent systems.

Various POMDP RL algorithms may be used in step 2 to obtain different approaches.

The approach used in Step 2 is a novel POMDP RL algorithm.
Step 1) Common Information Approach [Nayyar, Mahajan, Teneketzis 2013]

Define partial function $\beta^i_t : \mathcal{M}^i \rightarrow \mathcal{A}^i$ as follows:

$$\beta^i_t(\cdot) := g^i_t(Z_{1:t}, \cdot) \text{ such that } \mathcal{A}^i_t = \beta^i_t(\mathcal{M}^i_t).$$

Define coordinator’s strategy as follows:

$$\psi_t(Z_{1:t}) := g_t(Z_{1:t}, \cdot) \text{ such that } (\beta^1_t, \ldots, \beta^n_t) = \psi_t(Z_{1:t}).$$

Virtual coordinator observes $C_t$ and prescribes $\beta_t =: (\beta^1_t, \ldots, \beta^n_t) \in \mathcal{G}$.

**An Equivalent Centralized POMDP**

The total expected reward in coordinated system is as follows:

$$\hat{J}(\psi) = \mathbb{E}^{\psi} \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r_t(S_t, \beta^1_t(M^1_t), \ldots, \beta^n_t(M^n_t)) \right].$$
Step 1) Common Information Approach [Nayyar, Mahajan, Teneketzis 2013]

- $\Pi_t = \mathbb{P}(S_t, M^n_1, \ldots, M^n_t | Z_{1:t-1}, \beta_{1:t-1})$ is an information state.

- Let $\mathcal{R}$ be the reachable set of the obtained POMDP with action $\beta \in \mathcal{G}$ and observation $Z \in \mathcal{Z}$.

- Given fixed initial distribution $\pi_1$, reachable set $\mathcal{R}$ is at most countable.
Step 2: An Approximate POMDP RL Algorithm

Incrementally Expanding Representation (IER)

IER is a 3-tuple $\langle \{\mathcal{X}\}_N^{\infty}, \tilde{f}, B \rangle$ such that

- $\{\mathcal{X}\}_N^{\infty}$ is a sequence of finite sets such that $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \ldots \mathcal{X}_N \subset \ldots$, and $\mathcal{X}_1$ is singleton say $\mathcal{X}_1 = \{x^*\}$. Let $\mathcal{X} = \lim_{N \to \infty} \mathcal{X}_N$.

- For any $\beta$ and $z$, and $x \in \mathcal{X}_N$, we have that $\tilde{f}(x, \beta, z) \in \mathcal{X}_{N+1}$.

- $B$ is surjective function that maps $\mathcal{X}$ to the reachable set s.t. $\Pi_t = B(\mathcal{X}_t)$.

Lemma 1

For every multi-agent system with partial history sharing information structure, there exists at least one IER such that $\mathcal{X}$ and $\tilde{f}$ do not depend on unknowns.

Note that $B$ may depend on unknowns.
Step 2: An Approximate POMDP RL Algorithm

- Construct countable-state MDP $\Delta$ with state space $\mathcal{X}$, action space $\mathcal{G}$, dynamics $\tilde{f}$, and reward $\tilde{r}(B(X_t), \beta_t) := \hat{r}_t(\Pi_t, \beta_t)$.

- Approximate $\Delta$ by finite-state MDPs $\{\Delta_N\}_{N=1}^{\infty}$ where state space is $\mathcal{X}_N$, action space $\mathcal{G}$, dynamics $\tilde{f}$, and reward $\tilde{r}(B(X_t), \beta_t)$.

- Apply a generic finite-state RL algorithm $\zeta$ to learn optimal strategy of $\Delta_N$. We assume $\zeta$ converges to an optimal strategy of $\Delta_N$.

- Translate the strategies in $\Delta_N$ to strategies in the original multi-agent system.

Main Theorem

Let $J^*$ be the optimal performance (reward) of the original MAS system and $\tilde{J}$ be the performance under the learned strategy. Then,

$$J - \tilde{J} \leq \epsilon_N,$$

where $\epsilon_N = \frac{2\gamma^{\tau_N}}{1-\gamma} (r_{\text{max}} - r_{\text{min}}) \leq \frac{2\gamma^N}{1-\gamma} (r_{\text{max}} - r_{\text{min}})$ and $\tau_N$ is a model dependent parameter that $\tau_N \geq N$. Note that error goes to zero exponentially in $N$. 

(J. Arabneydi, www.cim.mcgill.ca/jarabney)
Multi-Agent RL Algorithm

(1) Given $\epsilon > 0$, choose $N$ such that $\frac{2\gamma^N}{1-\gamma} (r_{\text{max}} - r_{\text{min}}) \leq \epsilon$. Then, construct $\Delta_N$; particularly, state space $\mathcal{X}_N$ and dynamics $\tilde{f}$.

(2) At iteration $k$, $\zeta$ chooses prescriptions $\beta_k = (\beta^1_k, \ldots, \beta^n_k)$. (Agents have access to a common random generator to explore consistently). Agent $i$ takes action $a^i_k$ based on prescription $\beta^i_k$ and local information $m^i_k$:

$$a^i_k = \beta^i_k(m^i_k), \forall i.$$

(3) Based on taken actions, system incurs reward $r_k$, evolves, and generates common observation $z_k$ that is observable to every agent. Agents consistently compute next state as follows

$$x_{k+1} = \tilde{f}(x_k, \beta_k, z_k) \in \mathcal{X}_N.$$

(4) $\zeta$ learns (updates) the coordinated strategy according to observed reward $r_k$ by performing prescriptions $\beta_k$ at state $x_k$ and transiting to state $x_{k+1}$.

(5) $k \leftarrow k + 1$, and got step 2 until termination.
Example: Multi-Access Broadcasting Channel (MABC)

- \( S_t = (S^1_t, S^2_t) \in \{0, 1\}^2, \quad A^i_t \in \{0, 1\} \).

- Packets arrive at user \( i \) according to independent Bernoulli process with rate \( p^i \in (0, 1) \) and are unknown.

- Each user transmits if it has a packet i.e. \( A^i_t \leq S^i_t \).

- Information at each agent \( I^i_t = \{S^i_t, A^{1:1:t-1}_t, A^{2:1:t-1}_t\} \).

- The objective is to maximize the throughput; hence, reward function \( r(S_t, A^1_t, A^2_t) = A^1_t + A^2_t - 2A^1_t A^2_t \).
Example: Multi-Access Broadcasting Channel (MABC)
Example: Multi-Access Broadcasting Channel (MABC)

Figure: This figure shows the learning process of MDP $\Delta_N$ in a few snapshots. Numerical values: $b_1 = 0.25$, $b_2 = 0.83$, $N = 50$, $\gamma = 0.99$, $p^1 = 0.3$, $p^2 = 0.6$. 
Given $\epsilon > 0$, we presented a (model-based or model-free) RL algorithm that guarantees $\epsilon$-optimality for a large class of multi-agent systems with partial history sharing.

Our approach has two main steps: Common Information Approach + POMDP RL.

We provided a novel approach for approximate solution of POMDPs (model known and unknown model).

We developed a multi-agent Q-learning algorithm for MABC problem that converges to optimal policy.

The obtained error bound is conservative and in practice, the actual error is less.
Thank You

Poster number: T47

"Reinforcement Learning in Decentralized Stochastic Control Systems with Partial History Sharing", Accepted in IEEE American Control Conference (ACC), Jul., 2015.
Main Property

When state $x_t$ steps out of $\Delta_N$, it will come back to $\mathcal{X}_N$ after a finite time.

This property is required for a model to have every pair of state and action in $\Delta_N$ visited infinitely often. In the literature, different versions of this property have been considered.

- There exists an oracle that provides the agent with exact information about the current state, upon request; however, using the oracle is expensive and reserved for the learning phase.
- In sensor networks, where the communication is sensing is cheap but communication is expensive.
- Agents have access to "reset" or "off-line" simulation.
- $\Delta$ is such that after a finite time, the state will come back to $\Delta_N$, so it is better to wait until the state comes back.