New Concepts in Team Theory:
Mean Field Teams & Reinforcement Learning

JALAL ARABNEYDI

PhD Oral Defense: December 8, 2016
Electrical and Computer Engineering department, McGill University
Introduction to Team Theory
Team theory studies decision makers that wish collaborate to accomplish a common task.
Team theory

Salient feature of Teams:

- Multiple decision makers.
- Decentralized information.
- Common objective.
Team in various applications

- Networked control
- Robotics
- Communication
- Transportation
- Sensor networks
- Smart grids
- Economics
- etc.
Team in various applications

- Networked control
- Robotics
- Communication
- Transportation
- Sensor networks
- Smart grids
- Economics
- etc.

Teams are almost everywhere.
Background of team theory

Static team (Radner 1962, Marschack and Radner 1972)

Dynamic team (Witsenhausen 1971, Witsenhausen 1973)

Specific information structure

- Partially nested (Ho and Chu 1972)
- One-step delayed sharing (Witsenhausen 1971, Yoshikawa 1978)
- n-step delayed sharing (Witsenhausen 1971, Varaiya 1978, Nayyar 2011)
- Common past sharing (Aicardi 1978)
- Periodic sharing (Ooi 1997)
- Belief sharing (Yuksel 2009)
- Partial history sharing (Nayyar 2013)
Motivation

- Explicit optimal solutions typically for 2-3 agents: big gap between theory and application.

- When the model is not known completely: no optimal result even for 2-3 agents.
Motivation

- Explicit optimal solutions typically for 2-3 agents: big gap between theory and application.

- When the model is not known completely: no optimal result even for 2-3 agents.

- Mean Field Teams.

- Reinforcement Learning w.t. partial history sharing.
Mean Field Teams
Partially exchangeable agents

Smart grids

Swarms robotics

Social networks
Notation

- $\mathcal{N}$: set of heterogeneous agents
- $\mathcal{K}$: set of sub-populations

**For entire population:**
- $x_t$: joint state at time $t$
- $u_t$: joint action at time $t$

**For agent $i$ of sub-population $k \in \mathcal{K}$:**
- $\mathcal{N}^k$: entire sub-population of type $k \in \mathcal{K}$
- $x_t^i \in \mathcal{X}^k$: state of agent $i$ at time $t$
- $u_t^i \in \mathcal{U}^k$: action of agent $i$ at time $t$
Notation

• $\mathcal{N}$: set of heterogeneous agents
• $\mathcal{K}$: set of sub-populations

For entire population:
• $\mathbf{x}_t$: joint state at time $t$
• $\mathbf{u}_t$: joint action at time $t$

For agent $i$ of sub-population $k \in \mathcal{K}$:
• $\mathcal{N}^k$: entire sub-population of type $k \in \mathcal{K}$
• $x^i_t \in \mathcal{X}^k$: state of agent $i$ at time $t$
• $u^i_t \in \mathcal{U}^k$: action of agent $i$ at time $t$
Partially exchangeable agents

**Definition (Exchangeable agents)**

A pair \((i, j)\) of agents is exchangeable if:

1) For any \(t\), and any \(x, u, \text{ and } w\),

\[
\sigma_{i,j}(f_t(x, u, w)) = f_t(\sigma_{i,j}x, \sigma_{i,j}u, \sigma_{i,j}w),
\]

2) For any \(t\), and any \(x\) and \(u\),

\[
c_t(x, u) = c_t(\sigma_{i,j}x, \sigma_{i,j}u),
\]
Partially exchangeable agents

Definition (Exchangeable agents)

A pair \((i, j)\) of agents is exchangeable if:

1) For any \(t\), and any \(x\), \(u\), and \(w\),

\[
\sigma_{i,j}(f_t(x, u, w)) = f_t(\sigma_{i,j}x, \sigma_{i,j}u, \sigma_{i,j}w),
\]

2) For any \(t\), and any \(x\) and \(u\),

\[
c_t(x, u) = c_t(\sigma_{i,j}x, \sigma_{i,j}u),
\]

Exchangeable agents \(\not\iff\) Exchangeable initial states & noises
Partially exchangeable agents

Definition (Exchangeable agents)
A pair \((i, j)\) of agents is exchangeable if:

1) For any \(t\), and any \(x, u, w\),
   \[
   \sigma_{i,j}(f_t(x, u, w)) = f_t(\sigma_{i,j}x, \sigma_{i,j}u, \sigma_{i,j}w),
   \]

2) For any \(t\), and any \(x, u\),
   \[
   c_t(x, u) = c_t(\sigma_{i,j}x, \sigma_{i,j}u),
   \]

Partially exchangeable agents \(\equiv\) Mean-field coupled agents
(Irrespective of information structure)
Mean field models: controlled Markov chain

Suppose the dynamics \( x_{t+1} = f_t(x_t, u_t, w_t) \).

The per-step cost is \( c_t(x_t, u_t) \).

**Proposition 2.2**

There exist functions \( \{ \{ f^k_t \} \}_{k \in \mathcal{K}}, \ell_t \) such that for agent \( i \in \mathcal{N}^k \)

\[
x^i_{t+1} = f^k_t(x^i_t, u^i_t, \xi^i_t, w^i_t),
\]

and the per-step cost at time \( t \), may be written as

\[
\ell_t(\xi_t).
\]

\[
m_t = \text{vec}(m^1_t, \ldots, m^K_t),
\]

\[
m^k_t = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} \delta_{x^i_t},
\]

\[
\xi_t = \text{vec}(\xi^1_t, \ldots, \xi^K_t),
\]

\[
\xi^k_t = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} \delta_{x^i_t, u^i_t}.
\]
Suppose the dynamics are linear, i.e.,

\[ x_{t+1} = A_t x_t + B_t u_t + w_t. \]

The per-step cost is quadratic, i.e.,

\[ c_t(x_t, u_t) = x_t^T Q_t x_t + u_t^T R_t u_t. \]

**Proposition 2.1**

There exist matrices \( \{A^k, B^k, D^k, E^k, Q^k, R^k\}_{k \in K} \) and \( P^x_t \) and \( P^u_t \) such that

\[ x^i_{t+1} = A^k x^i_t + B^k u^i_t + D^k \bar{x}_t + E^k \bar{u}_t + w^i_t. \]

and the per-step cost at time \( t \), may be written as

\[
\bar{x}_t^T P^x_t \bar{x}_t + \bar{u}_t^T P^u_t \bar{u}_t + \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \frac{1}{|\mathcal{N}^k|} \left[ (x^i_t)^T Q^k x^i_t + (u^i_t)^T R^k u^i_t \right].
\]

\[
\bar{x}_t = \text{vec}(\bar{x}_t^1, \ldots, \bar{x}_t^K), \quad \bar{u}_t = \text{vec}(\bar{u}_t^1, \ldots, \bar{u}_t^K),
\]

\[
\bar{x}_t^k = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} x^i_t, \quad \bar{u}_t^k = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} u^i_t.
\]
Mean-field teams: problem formulation

Controlled Markov Chain

- Dynamics: \( x_{t+1}^i = f_k^k(x_t^i, u_t^i, \xi_t^i, w_t^i) \)

- Per-step cost: \( \ell_t(\xi_t) \)

- Information structure: \( u_t^i = g_t^i(x_t^i, m_{1:t}) \)

- Objective:
  \[
  J^* = \min_g \left( \mathbb{E}^g \left[ \sum_{t=1}^T \ell_t(\xi_t) \right] \right)
  \]

Linear Quadratic

- \( x_{t+1}^i = A_t^k x_t^i + B_t^k u_t^i + D_t^k \bar{x}_t + E_t^k \bar{u}_t + w_t \)

- \( \ell_t(x_t, u_t) = \bar{x}_t^T P_t^x \bar{x}_t + \bar{u}_t^T P_t^u \bar{u}_t + \sum_{k=1}^K \sum_{i \in \mathcal{N}_k} \frac{1}{|\mathcal{N}_k|} \left[ (x_t^i)^T Q_t^k x_t^i + (u_t^i)^T R_t^k u_t^i \right] \)

- \( u_t^i = g_t^i(x_t^i, \bar{x}_{1:t}) \)

- \( J^* = \inf_g \left( \mathbb{E}^g \left[ \sum_{t=1}^T \ell_t(x_t, u_t) \right] \right) \)
<table>
<thead>
<tr>
<th>Controlled Markov Chain</th>
<th>Linear Quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td>A 4.1 The control laws are exchangeable i.e. $g^i_t = g^j_t$ for any $i, j \in \mathbb{N}^k$.</td>
<td>Not needed.</td>
</tr>
<tr>
<td>It is a standard assumption in large scale systems for reasons: simplicity, fairness, &amp; robustness.</td>
<td></td>
</tr>
</tbody>
</table>
### Mean-field teams: key assumptions

#### Controlled Markov Chain

No assumptions on the probability distributions across agents.
- Gaussian or non-Gaussian,
- Independent or highly correlated,
- Exchangeable or non-exchangeable.

#### Linear Quadratic

A 4.1. The control laws are exchangeable.

It is a standard assumption in large scale systems for reasons:
- simplicity, fairness, & robustness.
**Mean-field teams: main challenges**

**Controlled Markov Chain**

- Coupling in dynamic and cost with non-classical information structure. This belongs to NEXP.
- Designer’s approach, impractical dynamic program.
- Common information approach, state space of dynamic program increases exponentially in number of agents and time, i.e., \( P(x_t^1, \ldots, x_t^N | m_{1:t}) \).

**Linear Quadratic**
Mean-field teams: main challenges

**Controlled Markov Chain**

- Coupling in dynamic and cost with non-classical information structure. This belongs to **NEXP**.
  
- Designer’s approach, **impractical** dynamic program.
  
- Common information approach, state space of dynamic program increases **exponentially** in number of agents and time, i.e.,
  \[ \mathbb{P}(x_1^t, \ldots, x_N^t | m_{1:t}) \].

**Linear Quadratic**

- LQG with non-classical information structure is difficult.
  
- Linear strategies are optimal only for Gaussian and partially nested.
  
- The mean field sharing is not partially nested and the noises are allowed to be non-Gaussian.
### Mean-field teams: main challenges

**Controlled Markov Chain**

- Designer’s approach, **impractical** dynamic program.
- Common information approach, state space of dynamic program increases **exponentially** in number of agents and time, i.e., $\mathbb{P}(x_t^1, \ldots, x_t^N | m_{1:t})$.

---

**Linear Quadratic**

- Linear strategies are optimal only for **Gaussian** and partially nested.
- The **mean field sharing** is not partially nested and the noises are allowed to be non-Gaussian.

---

Witsenhausen’s **counterexample** is still an open problem after 48 years!
## Mean-field teams: main challenges

### Controlled Markov Chain

- Witsenhausen's **counterexample** is still an open problem after 48 years!
- Common information approach, state space of dynamic program increases exponentially in number of agents and time, i.e., \( P(x_t^1, \ldots, x_t^N \mid m_{1:t}) \).

### Linear Quadratic

- The mean field sharing is not partially nested and the noises are allowed to be non-Gaussian.
## Mean-field teams: main challenges

### Controlled Markov Chain

- Witsenhausen’s **counterexample** is still an open problem after 48 years!


### Linear Quadratic

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbb{P}(x_t^1, \ldots, x_t^N \mid m_{1:t})$.</td>
</tr>
</tbody>
</table>
## Mean-field teams: main challenges

<table>
<thead>
<tr>
<th>Controlled Markov Chain</th>
<th>Linear Quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Witsenhausen’s <strong>counterexample</strong> is still an open problem after 48 years!</td>
<td></td>
</tr>
<tr>
<td>Yuksel and Tatikonda, A <strong>counterexample</strong> in distributed optimal sensing, 2009.</td>
<td></td>
</tr>
<tr>
<td>Whittle and Rudge, The optimal linear solution of a symmetric team control problem, 1974.</td>
<td></td>
</tr>
</tbody>
</table>
Mean-field teams: main challenges

Controlled Markov Chain

- Coupling in dynamic and cost with non-classical information structure. This belongs to NEXP.

- Designer’s approach, impractical dynamic program.

- Common information approach, state space of dynamic program increases exponentially in number of agents and time, i.e., \( \mathbb{P}(x_t^1, \ldots, x_t^N \mid m_{1:t}) \).

Linear Quadratic

- LQG with non-classical information structure is difficult.

- Linear strategies are optimal only for Gaussian and partially nested.

- The mean field sharing is not partially nested and the noises are allowed to be non-Gaussian.
## Mean-field teams: main challenges

<table>
<thead>
<tr>
<th>Controlled Markov Chain</th>
<th>Linear Quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Coupling in dynamic and cost with non-classical information structure. This belongs to NEXP.</td>
<td>• LQG with non-classical information structure is difficult.</td>
</tr>
</tbody>
</table>
| • Common information approach, state space of dynamic program increases exponentially in number of agents and time, i.e., 
\[ P(x_1^t, \ldots, x_N^t | m_{1:t}) \] | • The mean field sharing is not partially nested and the noises are allowed to be non-Gaussian. |

There is no existing approach to solve mean-field teams.
Theorem 4.1

Define recursively value functions:

\[ V_{T+1}(m) = 0, \quad m \in \mathcal{M}_n, \]

and for \( t = T, \ldots, 1 \), for \( m \in \mathcal{M}_n \),

\[
V_t(m) = \min_{\gamma} \mathbb{E} \left[ \ell_t(\phi(m_t, \gamma_t)) + V_{t+1}(m_{t+1}) \mid m_t = m, \gamma_t = \gamma \right],
\]

where \( \gamma = (\gamma^1, \ldots, \gamma^K) \), \( \gamma^k : \mathcal{X}^k \rightarrow \mathcal{U}^k \), and

\[
\phi(m, \gamma)(x, u) = m(x) \prod_{k=1}^{K} 1(u^k = \gamma^k(x^k)), \quad x \in \prod_{k=1}^{K} \mathcal{X}^k, \quad u \in \prod_{k=1}^{K} \mathcal{U}^k, \quad x^k \in \mathcal{X}^k, \quad u^k \in \mathcal{U}^k.
\]

Let \( \psi_t^* \) denote any argmin of the right hand side. Then, optimal solution is

\[
g_{t}^{*, k}(m, x) := \psi_{t}^{*, k}(m)(x), \quad m \in \mathcal{M}_n, x \in \mathcal{X}^k, \quad k \in K.
\]
Mean-field teams: main theorems

**Theorem 3.1**

The optimal strategy is unique, identical across sub-populations, and is linear in local state and the mean-field of the system. In particular,

$$u_i^t = \tilde{L}_t^k (x_i^t - \bar{x}_t^k) + \bar{L}_t^k \bar{x}_t, \quad i \in \mathcal{N}^k, \ k \in \mathcal{K},$$

where the above gains are obtained by the solution of $K + 1$ Riccati equations: one for computing each $\tilde{L}_t^k, \ k \in \mathcal{K}$, and one for $\bar{L}_t := \text{vec}(\tilde{L}_1^T, \ldots, \tilde{L}_K^T)$. Let $\tilde{M}_t^k$ and $\bar{M}^k_t$ denote the solution of the above Riccati equations and

$$\tilde{\Sigma}_t^k := \sum_{i \in \mathcal{N}^k} \text{var}(w_i^t - \bar{w}_t^k) / |\mathcal{N}^k|, \quad \bar{\Sigma}_t := \text{var}(\bar{w}_t), \quad \tilde{\Xi}_t^k := \sum_{i \in \mathcal{N}^k} \text{var}(x_i^t - \bar{x}_t^k) / |\mathcal{N}^k|, \quad \bar{\Xi} := \text{var}(\bar{x}_t).$$

Then, the optimal cost is given by

$$J^* = \sum_{k \in \mathcal{K}} \text{Tr}(\tilde{\Xi}_t^k \tilde{M}_t^k) + \text{Tr}(\bar{\Xi} \bar{M}_t) + \sum_{t=1}^{T-1} \left[ \sum_{k \in \mathcal{K}} \text{Tr}(\tilde{\Sigma}_t^k \tilde{M}_{t+1}^k) + \text{Tr}(\bar{\Sigma}_t \bar{M}_{t+1}) \right].$$
Mean-field teams: main theorems

Theorem 3.1

The optimal strategy is unique, identical across sub-populations, and is linear in local state and the mean-field of the system. In particular,

\[ u_t^i = \tilde{L}_t^k(x_t^i - \bar{x}_t^k) + \bar{L}_t^k\bar{x}_t, \quad i \in \mathcal{N}^k, k \in \mathcal{K}, \]

For agent \( i \in \mathcal{N}^k \) in sub-population \( k \in \mathcal{K} = \{1, \ldots, K\} \),

\[ u_t^i = g_t^{*, k}(m_t, x_t^i), \quad u_t^i = \tilde{L}_t^k(x_t^i - \bar{x}_t^k) + \bar{L}_t^k\bar{x}_t. \]

Then, the optimal cost is given by

\[ J^* = \sum_{k \in \mathcal{K}} \text{Tr}(\tilde{\Sigma}_t^k \tilde{M}_t^k) + \text{Tr}(\bar{\Sigma}_t \bar{M}_t) + \sum_{t=1}^{T-1} \left[ \sum_{k \in \mathcal{K}} \text{Tr}(\tilde{\Sigma}_t^k \tilde{M}_{t+1}^k) + \text{Tr}(\bar{\Sigma}_t \bar{M}_{t+1}) \right]. \]
Salient features

Controlled Markov Chain

- The solution complexity is polynomial in number of agents (rather than exponential) and linear in time (rather than exponential.)

Linear Quadratic

- No need to share anything beyond mean field.
- The solution complexity depends on the number of sub-populations, i.e., $K$ but not on the number of agents in each sub-population, i.e., $N^k$.
- Each agent needs to solve only two Riccati equations (distributed computation).
Mean-field teams: generalizations

**Controlled Markov Chain**
- Arbitrarily coupled cost
- Infinite horizon
- Noisy observation
- Major-minor
- Randomized strategies

**Linear Quadratic**
- Weighted mean field
- Infinite horizon
- Partial mean field sharing
- Major-minor
- Tracking problem
Within the same sub-population, each agent is allowed to have different tracking reference and weights:

\[ u_t^i = \bar{L}_t^k(x_t^i - \lambda^i \bar{x}_t^k, \lambda) + \lambda^i \bar{L}_t^k \bar{x}_t^\lambda + \bar{F}_t^k v_t^i, \lambda^i + \lambda^i \bar{F}_t^k \bar{v}_t^\lambda \]
Mean-field teams: generalizations

**Controlled Markov Chain**
- Arbitrarily coupled cost
- Infinite horizon
- Noisy observation
- Major-minor
- Randomized strategies

**Linear Quadratic**
- Weighted mean field
- Infinite horizon
- Partial mean field sharing
- Major-minor
- Tracking problem
Mean-field teams: generalizations

When mean field of only sub-populations $S \in \mathcal{K}$ are observed:

$$u^i_t = \bar{L}^k_t (x^i_t - z^k_t) + \bar{L}^k_t z_t,$$

where

$$z^k_{t+1} = \begin{cases} \bar{x}^k_{t+1}, & k \in S, \\ A^k_t z^k_t + (B^k_t \bar{L}^k_t + D^k_t + E^k_t \bar{L}_t)z_t, & k \in S^c. \end{cases}$$

The approximation error

$$\Delta J = \text{Tr}(\tilde{X}_1 \tilde{M}_1) + \sum_{t=1}^{T-1} \text{Tr}(\tilde{W}_t \tilde{M}_{t+1}),$$

where $\tilde{M}_1:T$ is the solution of a Lyapunov equation. It is bounded as

$$\Delta J \in O \left(\frac{T}{n}\right).$$
Mean-field teams: numerical examples

Controlled Markov Chain

Linear Quadratic

Renewable energy
Conventional energy

Dynamics

Service-Provider

Service-Provider
Numerical example 1: demand response

- $x^i_t \in \mathcal{X} = \{OFF, ON\}$, \quad $m_t = \frac{1}{n} \sum_{i=1}^{n} 1(x^i_t = OFF)$

- Dynamics: $\mathbb{P}(x^i_{t+1}|x^i_t, u^i_t) =: [P(u^i_t)]_{x^i_t x^i_{t+1}}$

- Actions: $u^i_t \in \mathcal{U} = \{FREE, OFF, ON\}$, \quad Cost of action: $C(u^i_t)$

- Objective: $\min_{g} \mathbb{E}^g \left[ \sum_{t=1}^{\infty} \beta^t \left( \frac{1}{n} \sum_{i=1}^{n} C(u^i_t) + D(m_t \| \zeta_t) \right) \right]$. 
Numerical example 1: demand response
Reinforcement Learning with Partial History Sharing
Reinforcement learning with partial history sharing

Team

Mean Field Teams
Reinforcement learning with partial history sharing

Team

Partial History Sharing

Mean Field Teams
Reinforcement learning in general team

State: $x_t \in \mathcal{X}$.  
Observation: $y^i_t = h(x_t, u^1_{t-1}, \ldots, u^n_{t-1}, w_t^{i,o})$.

Control law: $u^i_t = g^i_t(l^i_t)$.  
Information: $l^i_t \subseteq \{y^1_{1:t}, \ldots, y^n_{1:t}, u^1_{1:t-1}, \ldots, u^n_{1:t-1}\}$.

System cost: Given $\beta \in (0, 1)$, 
$$J(g) = \mathbb{E}^g \left[ \sum_{t=1}^{\infty} \beta^{t-1} \ell(x_t, u^1_t, \ldots, u^n_t) \right].$$
Suppose the system dynamics \((f, h)\), cost structure \(\ell\), and probability mass functions are not completely known.

**Objective:** Given \(\epsilon > 0\), find strategy \(g^*_\epsilon\) such that

\[
J(g^*_\epsilon) \leq J^* + \epsilon.
\]
**Definition (Partial History Sharing, Nayyer et al. 2013)**

Split the information at each agent into two parts:

- **Common information**: $c_t = \bigcap_{i=1}^{n} l_t^i$ i.e. shared between all agents.
- **Local information**: $m_t^i = l_t^i \setminus c_t$ that is the local information of agent $i$.

Define $z_t := c_{t+1} \setminus c_t$ as common observation, hence $c_{t+1} = z_{1:t}$. Then,

a) The update of local information

$$m_{t+1}^i \subseteq \{m_t^i, u_t^i, y_{t+1}^i\} \setminus z_t, \quad i \in \{1, \ldots, n\}.$$

b) For every agent $i$, $|m_t^i|$ and $|z_t|$ are uniformly bounded in time $t$.

PHS encompasses: **delayed sharing**, **mean-field sharing**, **periodic sharing**, **control sharing**, etc.
Given centralized MDP, there are two ways to learn the optimal solution:

- **Indirect**: supervised learning and dynamic program.

- **Direct (Reinforcement Learning)**: Barto, Sutton, Watkins, Dayan, Singh, etc. (active since 80’s).
Reinforcement learning in team: main challenges

Most of existing RL methods are developed for finite state-action MDPs. However, decentralized systems are not MDP in general.

- **Indirect**: supervised learning and dynamic program.

- **Direct (Reinforcement Learning)**: Barto, Sutton, Watkins, Dayan, Singh, etc. (active since 80’s).
Reinforcement learning in team: main challenges

Most of existing RL methods are developed for finite state-action MDPs. However, decentralized systems are not MDP in general.

The **indirect method may not be feasible** due to the incomplete information i.e. dynamics and cost may not be fully identified.
Most of existing RL methods are developed for finite state-action MDPs. However, decentralized systems are not MDP in general.

The indirect method may not be feasible due to the incomplete information i.e. dynamics and cost may not be fully identified.

[Nayyer et al. 2013] identifies a dynamic program for PHS; however,

The state space is an infinite set.

The state space depends on the model.

There is no RL algorithm for POMDP that guarantees optimality.
Reinforcement learning in team: main challenges

Given centralized MDP, there are two ways to learn the optimal solution:

• Indirect: supervised learning and dynamic program.
• Direct (Reinforcement Learning): Barto, Sutton, Watkins, Dayan, Singh, etc. (active since 80’s).

No existing approach to solve decentralized reinforcement learning.
STEP 1: Common Information Approach

Define partial function $\gamma^i_t : \mathcal{M}^i \rightarrow \mathcal{U}^i$:

$$\gamma^i_t := g^i_t(z_{1:t-1}, \cdot) \quad \text{s.t.} \quad u^i_t = \gamma^i_t(m^i_t).$$

Let $\psi$ denote the coordinator's strategy:

$$(\gamma^1_t, \ldots, \gamma^n_t) = \psi_t(z_{1:t-1}).$$

Virtual coordinator observes $z_{1:t-1}$ and prescribes $\gamma_t := (\gamma^1_t, \ldots, \gamma^n_t) \in \mathcal{G}$.

An equivalent centralized POMDP [Nayyer et al., 2013]

A dynamic program is identified to characterize the optimal strategy based on the information state $\pi$.

$$V(\pi) = \min_{\gamma \in \mathcal{G}} \mathbb{E} \left[ \ell(x_t, u_t) + V(\pi_{t+1}) | \pi_t = \pi, \gamma_t = \gamma \right].$$

Let $\mathcal{R}$ denote the reachable set of the information state $\pi$. 
## Definition (Incrementally Expanding Representation)

Let \( \{S_N\}_{N=1}^{\infty} \) be a sequence of finite sets such that \( S_1 \subset S_2 \subset \ldots \subset S_N \subset \ldots \). Let \( S = \lim_{N \to \infty} S_N \) be the countable union of above finite sets. The tuple \( \langle \{S_N\}_{N=1}^{\infty}, B, \tilde{f} \rangle \) is called an **Incrementally Expanding Representation**, if

### Incremental Expansion

For any \( \gamma \in \mathcal{G}, z \in \mathcal{Z}, \) and \( s \in S_N \),

\[
\tilde{f}(s, \gamma, z) \in S_{N+1}.
\]

### Consistency

For any \( (\gamma_{1:t-1}, z_{1:t-1}) \), let \( \pi_t \) and \( s_t \) be the corresponding states at time \( t \). Then,

\[
\pi_t = B(s_t).
\]
Pre-learning stage

STEP 2: An Approximate POMDP RL Algorithm

**Lemma**

Every decentralized systems with PHS has at least one IER such that $S$ and $\tilde{f}$ do not depend on unknowns.

- Construct countable-state MDP $\Delta$ with state space $S$, action space $G$, dynamics $\tilde{f}$, and cost $\tilde{\ell}(B(s_t), \gamma_t) := \mathbb{E}[\ell(x_t, u^1_t, \ldots, u^n_t)|\pi_t, \gamma_t]$.

- Construct an augmented type approximation sequence $\{\Delta_N\}_{N=1}^{\infty}$ of $\Delta$, with state space $S_N$, action space $G$, dynamics $\tilde{f}$, and cost $\tilde{\ell}(B(s_t), \gamma_t)$.

- Apply a finite-state RL algorithm $\mathcal{T}$ (such as TD($\lambda$) and Q-learning) to learn optimal strategy of $\Delta_N$. We assume $\mathcal{T}$ converges to optimal strategy of $\Delta_N$. 
A Block Diagram

Multi-Agent System -> Common Information Approach -> Equivalent (Single-Agent) POMDP -> Incrementally Expanding Representation (IER)

ε-Optimal Policy -> Finite-state Reinforcement Learning -> Approximate Finite-State MDPs -> Countable-State MDP
Proposed decentralized RL algorithm

(1) Given $\epsilon > 0$, choose $N$ such that $\frac{2\beta N}{1-\beta} (\ell_{\text{max}} - \ell_{\text{min}}) \leq \epsilon$. Then, construct $\Delta_N$; particularly, state space $S_N$ and dynamics $\tilde{f}$.

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

$$u^i_k = \gamma^i_k(m^i_k), \forall i.$$ 

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

$$s_{k+1} = \tilde{f}(s_k, \gamma_k, z_k) \in S_N.$$ 

(4) $T$ learns (updates) the coordinated strategy according to observed cost $\ell_k$ by performing prescriptions $\gamma_k$ at state $s_k$ and transition to state $s_{k+1}$.

(5) $k \leftarrow k + 1$, and go to step 2 until termination.
Proposed decentralized RL algorithm

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

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

$$u^i_k = \gamma^i_k(m^i_k), \quad \forall i.$$

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

$$s_{k+1} = \tilde{f}(s_k, \gamma_k, z_k) \in S_N.$$

(4) $T$ learns (updates) the coordinated strategy according to observed cost $\ell_k$ by performing prescriptions $\gamma_k$ at state $s_k$ and transition to state $s_{k+1}$.

(5) $k \leftarrow k + 1$, and go to step 2 until termination.

The structure of the learned strategy:

$$u^i_t = g^i_t(s^i_t, m^i_t), \quad i \in (1, \ldots, n),$$

where $s^i_t$ is the internal state that changes every time.
Theorem 6.3
Let $J^*$ be the optimal performance of the original decentralized system and $\tilde{J}$ be the performance under the learned strategy. Then,

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

where $\epsilon_N = \frac{2\beta \tau_N}{1-\beta} (\ell_{\text{max}} - \ell_{\text{min}}) \leq \frac{2\beta^N}{1-\beta} (\ell_{\text{max}} - \ell_{\text{min}})$ and $\tau_N$ is a model dependent parameter that $\tau_N \geq N$. 

Learning stage: main theorem
Numerical example 2: multi Access Broadcast Channel (MABC)

- $x^i_t \in \{0, 1\}$ with independent arrival probability $p^i$, $i = 1, 2$.
- $I^i_t = (x^i_t, u^1_{1:t-1}, u^2_{1:t-1})$.
- $u^i_t \leq x^i_t \in \{0, 1\}$.
- In case of collision, packets remain in buffers.

- **Objective:** maximize the throughput.
  - State of other agent is unknown.  
    (decentralized information)
  - Arrival probabilities are unknown.  
    (incomplete model)
Future Work
Future work

- Game theory
- Markov chain
- **Reinforcement learning**: Specific teams such as mean-field teams.
- Mean-field teams and consensus algorithms
- Various approximations in mean-field teams: Information & model.
- New model of mean-field teams
- Various applications: Smart grids, communications, economics, robotics, social networks, etc.
Thank you.
Contributions
Main contributions: Mean Field Teams

- Introduce partially exchangeable agents and **mean-field teams**.
- Allow agents to be coupled in **dynamics** and cost under mild assumptions.
- Mean field sharing is **non-classical**. (difficult problems)
- We use novel approaches to find the **global** optimal solution.
  - Solution approach works for **arbitrary** # of agents. (not necessarily large)
  - Mean field can be computed and communicated easily or by local interactions using **consensus** algorithms.
  - In large sub-populations, mean-field is **predictable**. Also, mean-field teams are **robust to node failure**.
  - Different generalizations.
Main contributions: Mean Field Teams

• Introduce partially exchangeable agents and mean-field teams.

Salient features of mean-field teams:

1. Controlled Markov chain: solution complexity is \textit{polynomial} (rather than exponential) in # of agents and \textit{linear} (rather than exponential) in time.

2. Linear quadratic:
   • The optimal solution is \textit{linear}.
   • The solution complexity is \textit{independent of} $N$ and it depends only on $K$.
   • No need to share anything beyond mean field.
   • Each agent solves only two Riccati equations (\textit{distributed computation}).

3. When population is infinite, mean-field is \textit{deterministic and computable}. 
Main contributions: Mean Field Teams

- Introduce partially exchangeable agents and **mean-field teams**.
- Allow agents to be coupled in **dynamics** and cost under mild assumptions.
- Mean field sharing is **non-classical**. (difficult problems)
- We use novel approaches to find the **global** optimal solution.
- Solution approach works for **arbitrary** # of agents. (not necessarily large)
- Mean field can be computed and communicated easily or by local interactions using **consensus** algorithms.
- In large sub-populations, mean-field is **predictable**. Also, mean-field teams are **robust to node failure**.
- Different generalizations.
Main contributions: Mean Field Teams

- Introduce partially exchangeable agents and mean-field teams.
- Allow agents to be coupled in dynamics and cost under mild assumptions.
- Mean field sharing is non-classical (difficult problems).
- We use novel approaches to find the global optimal solution.
- Solution approach works for arbitrary # of agents. (not necessarily large)
- Mean field can be computed and communicated easily or by local interactions using consensus algorithms.
- In large sub-populations, mean-field is predictable. Also, mean-field teams are robust to node failure.
- Different generalizations.

- Arbitrarily coupled cost
- Infinite horizon
- Noisy observation
- Major-minor
- Randomized strategies

- Weighted mean field
- Infinite horizon
- Partial mean-field sharing
- Major-minor
- Tracking problem
Main contributions: Reinforcement Learning with PHS

- There is no existing RL in team that guarantees optimality.
- Introduce a novel decentralized RL for partial history sharing that guarantees $\epsilon$-optimal solution.
- Use common information approach and our proposed approach to design the learning space.
- Introduce the notion of Incrementally Expanding Representation.
- The proposed approach is also novel in centralized POMDP.
- Develop decentralized Q-learning for two-user MABC.
Main contributions: Reinforcement Learning with PHS

- There is no existing RL in team that guarantees optimality.
- Introduce a novel decentralized RL for partial history sharing that guarantees $\epsilon$-optimal solution.
- Use common information approach and our proposed approach to design the learning space.
- Introduce the notion of Incrementally Expanding Representation.
- The proposed approach is also novel in centralized POMDP.
- Develop decentralized Q-learning for two-user MABC.

Three features of designed learning space $S_N$:
- It is implementable by every agent based on common knowledge.
- It takes into account of the model and cost (not a prefixed space).
- It adapts to the existing powerful finite state-action RL algorithms.
Main contributions: Reinforcement Learning with PHS

- There is **no existing RL** in team that guarantees optimality.

- Introduce a novel decentralized RL for partial history sharing that **guarantees** $\epsilon$-**optimal** solution.

- Use **common information approach** and our **proposed approach** to design the learning space.

- Introduce the notion of **Incrementally Expanding Representation**.

- The proposed approach is also **novel in centralized POMDP**.

- Develop **decentralized Q-learning** for two-user MABC.
Reinforcement Learning: Multi-Access Broadcast Channel
Numerical example 2: MABC

Reachable Set

Countable State Space
Mean-field team: temperature control
Numerical example 3: temperature control