



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

Team theory studies decision makers that wish collaborate to accomplish a common task.

Salient feature of Teams:

- Multiple decision makers.
- Decentralized information.
- Common objective.

Team theory
accomplish

laborate to

Team in various applications

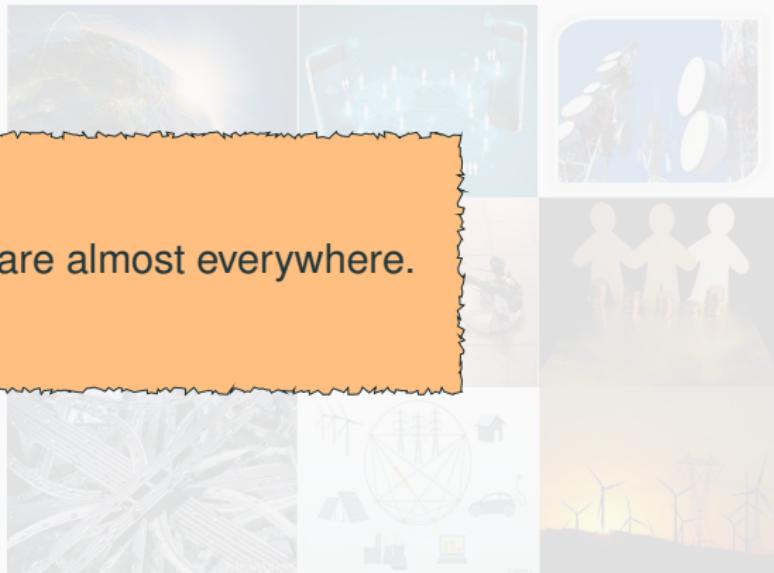
- Networked control
- Robotics
- Communication
- Transportation
- Sensor networks
- Smart grids
- Economics
- etc.



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



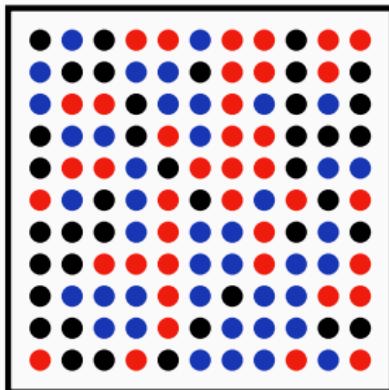
Swarm robotics



Social networks

Notation

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



For entire population:

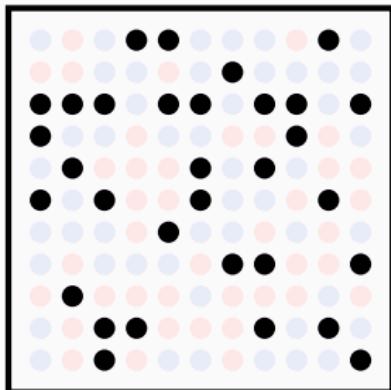
- \mathbf{x}_t : joint state at time t
- \mathbf{u}_t : joint action at time t

For agent i of sub-population $k \in \mathcal{K}$:

- \mathcal{N}^k : entire sub-population of type $k \in \mathcal{K}$
- $x_t^i \in \mathcal{X}^k$: 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



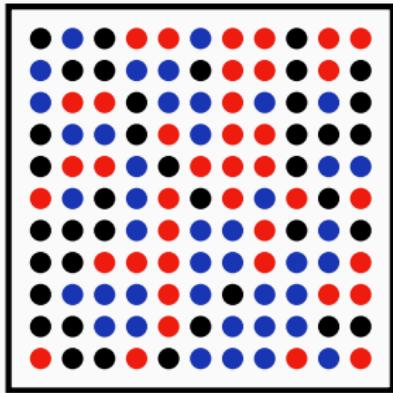
For entire population:

- \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 $k \in \mathcal{K}$
- $x_t^i \in \mathcal{X}^k$: state of agent i at time t
- $u_t^i \in \mathcal{U}^k$: action of agent i at time t

Partially exchangeable agents



Definition (Exchangeable agents)

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

- 1) For any t , and any \mathbf{x} , \mathbf{u} , and \mathbf{w} ,

$$\sigma_{i,j}(f_t(\mathbf{x}, \mathbf{u}, \mathbf{w})) = f_t(\sigma_{i,j}\mathbf{x}, \sigma_{i,j}\mathbf{u}, \sigma_{i,j}\mathbf{w}),$$

- 2) For any t , and any \mathbf{x} and \mathbf{u} ,

$$c_t(\mathbf{x}, \mathbf{u}) = c_t(\sigma_{i,j}\mathbf{x}, \sigma_{i,j}\mathbf{u}),$$

Partially exchangeable agents

Definition (Exchangeable agents)

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



Exchangeable agents $\not\iff$ Exchangeable initial states & noises

Navigation icons: back, forward, search, etc.

$$c_t(\mathbf{x}, \mathbf{u}) = c_t(\sigma_{i,j}\mathbf{x}, \sigma_{i,j}\mathbf{u}),$$

Partially exchangeable agents

Definition (Exchangeable agents)

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

Partially exchangeable agents \equiv Mean-field coupled agents
(Irrespective of information structure)

$$c_t(\mathbf{x}, \mathbf{u}) = c_t(\sigma_{i,j}\mathbf{x}, \sigma_{i,j}\mathbf{u}),$$

Mean field models: controlled Markov chain

Suppose the dynamics $\mathbf{x}_{t+1} = f_t(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t)$.

The per-step cost is $c_t(\mathbf{x}_t, \mathbf{u}_t)$.

Proposition 2.2

There exist functions $\{f_t^k\}_{k \in \mathcal{K}}, \ell_t$ such that for agent $i \in \mathcal{N}^k$

$$x_{t+1}^i = f_t^k(x_t^i, u_t^i, \xi_t, w_t^i),$$

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

$$\ell_t(\xi_t).$$

$$\mathbf{m}_t = \text{vec}(m_t^1, \dots, m_t^K), \quad \xi_t = \text{vec}(\xi_t^1, \dots, \xi_t^K),$$

$$m_t^k = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} \delta_{x_t^i}, \quad \xi_t^k = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} \delta_{x_t^i, u_t^i}.$$

Mean-field models: linear quadratic

Suppose the dynamics are linear, i.e., $\mathbf{x}_{t+1} = A_t \mathbf{x}_t + B_t \mathbf{u}_t + \mathbf{w}_t$.

The per-step cost is quadratic, i.e., $c_t(\mathbf{x}_t, \mathbf{u}_t) = \mathbf{x}_t^\top Q_t \mathbf{x}_t + \mathbf{u}_t^\top R_t \mathbf{u}_t$.

Proposition 2.1

There exist matrices $\{A_t^k, B_t^k, D_t^k, E_t^k, Q_t^k, R_t^k\}_{k \in \mathcal{K}}$ and P_t^x and P_t^u such that

$$x_{t+1}^i = A_t^k x_t^i + B_t^k u_t^i + D_t^k \bar{\mathbf{x}}_t + E_t^k \bar{\mathbf{u}}_t + w_t^i.$$

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

$$\bar{\mathbf{x}}_t^\top P_t^x \bar{\mathbf{x}}_t + \bar{\mathbf{u}}_t^\top P_t^u \bar{\mathbf{u}}_t + \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{N}^k} \frac{1}{|\mathcal{N}^k|} \left[(x_t^i)^\top Q_t^k x_t^i + (u_t^i)^\top R_t^k u_t^i \right].$$

$$\bar{\mathbf{x}}_t = \text{vec}(\bar{x}_t^1, \dots, \bar{x}_t^K),$$

$$\bar{\mathbf{u}}_t = \text{vec}(\bar{u}_t^1, \dots, \bar{u}_t^K),$$

$$\bar{x}_t^k = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} x_t^i,$$

$$\bar{u}_t^k = \frac{1}{|\mathcal{N}^k|} \sum_{i \in \mathcal{N}^k} u_t^i.$$

Mean-field teams: problem formulation

Controlled Markov Chain

- Dynamics: $x_{t+1}^i = f_t^k(x_t^i, u_t^i, \xi_t, w_t^i)$
- Per-step cost: $\ell_t(\xi_t)$
- Information structure: $u_t^i = g_t^i(x_t^i, \mathbf{m}_{1:t})$
- Objective:
$$J^* = \min_{\mathbf{g}} \left(\mathbb{E}^{\mathbf{g}} \left[\sum_{t=1}^T \ell_t(\xi_t) \right] \right)$$

Linear Quadratic

- $x_{t+1}^i = A_t^k x_t^i + B_t^k u_t^i + D_t^k \bar{\mathbf{x}}_t + E_t^k \bar{\mathbf{u}}_t + w_t^i$
- $\ell_t(\mathbf{x}_t, \mathbf{u}_t) = \bar{\mathbf{x}}_t^T P_t^x \bar{\mathbf{x}}_t + \bar{\mathbf{u}}_t^T P_t^u \bar{\mathbf{u}}_t + \sum_{k=1}^K \sum_{i \in \mathcal{N}^k} \frac{1}{|\mathcal{N}^k|} \left[(x_t^i)^T Q_t^k x_t^i + (u_t^i)^T R_t^k u_t^i \right]$
- $u_t^i = g_t^i(x_t^i, \bar{\mathbf{x}}_{1:t})$
- $$J^* = \inf_{\mathbf{g}} \left(\mathbb{E}^{\mathbf{g}} \left[\sum_{t=1}^T \ell_t(\mathbf{x}_t, \mathbf{u}_t) \right] \right).$$

Mean-field teams: key assumptions

Controlled Markov Chain

A 4.1 The control laws are exchangeable i.e. $g_t^i = g_t^j$ for any $i, j \in \mathcal{N}^k$.

It is a standard assumption in large scale systems for reasons:
simplicity, fairness, & robustness.

Linear Quadratic

Not needed.

Mean-field teams: key assumptions

Controlled Markov Chain

Linear Quadratic

No assumptions on the probability distributions across agents.

- Gaussian or non-Gaussian,
- Independent or highly correlated,
- Exchangeable or non-exchangeable.

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.,
 $\mathbb{P}(x_t^1, \dots, x_t^N | \mathbf{m}_{1:t})$.

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., $\mathbb{P}(x_t^1, \dots, x_t^N | \mathbf{m}_{1:t})$.

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



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

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

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**.

Mean-field teams: main challenges

Controlled Markov Chain

Linear Quadratic



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



Lipsa and matrins, Optimal memoryless control in Gaussian noise: A simple **counterexample**, 2008.

- Common information approach, state space of dynamic program increases **exponentially** in number of agents and time, i.e., $\mathbb{P}(x_t^1, \dots, x_t^N | \mathbf{m}_{1:t})$.

- 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

Linear Quadratic



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



Lipsa and matrins, Optimal memoryless control in Gaussian noise: A simple **counterexample**, 2008.



Yuksel and Tatikonda, A **counterexample** in distributed optimal sensing, 2009.

Set of agents and time, i.e.,

$$\mathbb{P}(x_t^1, \dots, x_t^N \mid \mathbf{m}_{1:t}).$$

Mean-field teams: main challenges

Controlled Markov Chain

Linear Quadratic



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



Lipsa and matrins, Optimal memoryless control in Gaussian noise: A simple **counterexample**, 2008.



Yuksel and Tatikonda, A **counterexample** in distributed optimal sensing, 2009.



Whittle and Rudge, The optimal linear solution of a symmetric team control problem, 1974.

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., $\mathbb{P}(x_t^1, \dots, x_t^N | \mathbf{m}_{1:t})$.

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

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

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

- Common information approach, state space of dynamic program increases **exponentially** in number of agents and time, i.e., $\mathbb{P}(x_t^1, \dots, x_t^N | \mathbf{m}_{1:t})$.

Linear Quadratic

- LQG with non-classical information structure is difficult.

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

Mean-field teams: main theorems

Theorem 4.1

Define recursively value functions:

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

and for $t = T, \dots, 1$, for $\mathbf{m} \in \mathcal{M}_n$,

$$V_t(\mathbf{m}) = \min_{\boldsymbol{\gamma}} \mathbb{E} \left[\ell_t(\phi(\mathbf{m}_t, \boldsymbol{\gamma}_t)) + V_{t+1}(\mathbf{m}_{t+1}) \mid \mathbf{m}_t = \mathbf{m}, \boldsymbol{\gamma}_t = \boldsymbol{\gamma} \right],$$

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

$$\phi(\mathbf{m}, \boldsymbol{\gamma})(\mathbf{x}, \mathbf{u}) = \mathbf{m}(\mathbf{x}) \prod_{k=1}^K \mathbb{1}(u^k = \gamma^k(x^k)), \quad \mathbf{x} \in \prod_{k=1}^K \mathcal{X}^k, \mathbf{u} \in \prod_{k=1}^K \mathcal{U}^k, x^k \in \mathcal{X}^k, u^k \in \mathcal{U}^k$$

Let ψ_t^* denote any argmin of the right hand side. Then, optimal solution is

$$g_t^{*,k}(\mathbf{m}, x) := \psi_t^{*,k}(\mathbf{m})(x), \quad \mathbf{m} \in \mathcal{M}_n, x \in \mathcal{X}^k, k \in \mathcal{K}.$$

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,

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

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

$$\check{\Sigma}_t^k := \frac{\sum_{i \in \mathcal{N}^k} \text{var}(w_t^i - \bar{w}_t^k)}{|\mathcal{N}^k|}, \bar{\Sigma}_t := \text{var}(\bar{w}_t), \check{\Xi}^k := \frac{\sum_{i \in \mathcal{N}^k} \text{var}(x_1^i - \bar{x}_1^k)}{|\mathcal{N}^k|}, \bar{\Xi} := \text{var}(\bar{x}_1)$$

Then, the optimal cost is given by

$$J^* = \sum_{k \in \mathcal{K}} \text{Tr}(\check{\Xi}^k \check{M}_1^k) + \text{Tr}(\bar{\Xi} \bar{M}_1) + \sum_{t=1}^{T-1} \left[\sum_{k \in \mathcal{K}} \text{Tr}(\check{\Sigma}_t^k \check{M}_{t+1}^k) + \text{Tr}(\bar{\Sigma}_t \bar{M}_{t+1}) \right].$$

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,

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

For agent $i \in \mathcal{N}^k$ in sub-population $k \in \mathcal{K} = \{1, \dots, K\}$,

$$u_t^i = g_t^{*,k}(\mathbf{m}_t, x_t^i), \quad u_t^i = \check{L}_t^k(x_t^i - \bar{x}_t^k) + \bar{L}_t^k \bar{x}_t.$$

\mathcal{N}^k

\mathcal{N}^k

\mathcal{N}^k

Then, the optimal cost is given by

$$J^* = \sum_{k \in \mathcal{K}} \text{Tr}(\check{\Sigma}^k \check{M}_1^k) + \text{Tr}(\bar{\Sigma} \bar{M}_1) + \sum_{t=1}^{T-1} \left[\sum_{k \in \mathcal{K}} \text{Tr}(\check{\Sigma}_t^k \check{M}_{t+1}^k) + \text{Tr}(\bar{\Sigma}_t \bar{M}_{t+1}) \right].$$

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

Mean-field teams: generalizations

Controlled Markov Chain

Linear Quadratic

Within the same sub-population, each agent is allowed to have different tracking reference and **weights**:

$$u_t^i = \check{L}_t^k(x_t^i - \lambda^i \bar{x}_t^{k,\lambda}) + \lambda^i \bar{L}_t^k \bar{\mathbf{x}}_t^\lambda + \check{F}_t^k v_t^{i,\lambda^i} + \lambda^i \bar{F}_t^k \bar{v}_t^\lambda$$

- Randomized strategies

- Tracking problem

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 $\mathcal{S} \in \mathcal{K}$ are observed:

$$u_t^i = \check{L}_t^k(x_t^i - z_t^k) + \bar{L}_t^k \mathbf{z}_t,$$

where

$$z_{t+1}^k = \begin{cases} \bar{x}_{t+1}^k, & k \in \mathcal{S}, \\ A_t^k z_t^k + (B_t^k \bar{L}_t^k + D_t^k + E_t^k \bar{L}_t) \mathbf{z}_t, & k \in \mathcal{S}^c. \end{cases}$$

The approximation error

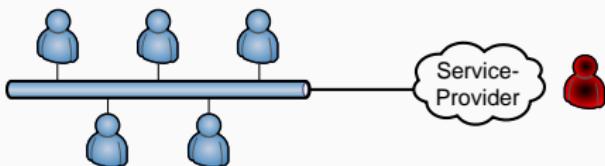
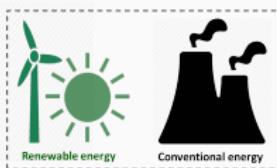
$$\Delta J = \text{Tr}(\tilde{X}_1 \tilde{M}_1) + \sum_{t=1}^{T-1} \text{Tr}(\tilde{W}_t \tilde{M}_{t+1}),$$

where $\tilde{M}_{1:T}$ is the solution of a **Lyapunov equation**. It is bounded as

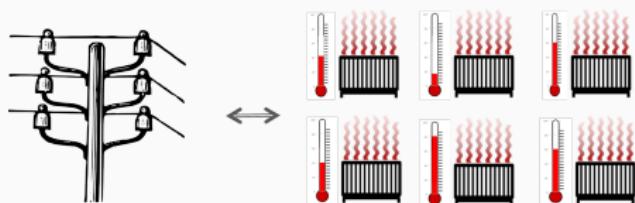
$$\boxed{\Delta J \in \mathcal{O}\left(\frac{T}{n}\right)}.$$

Mean-field teams: numerical examples

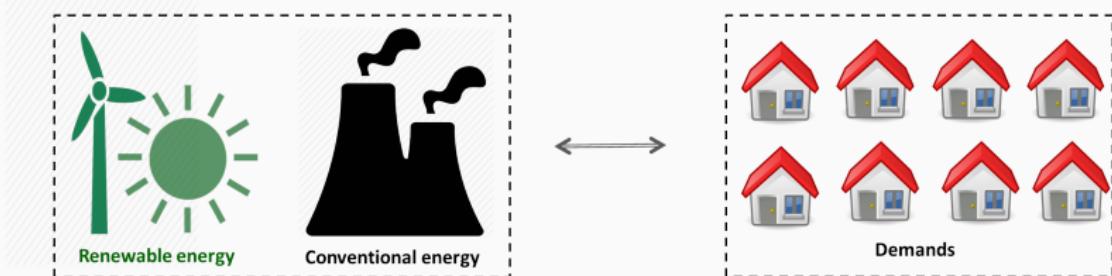
Controlled Markov Chain



Linear Quadratic



Numerical example 1: demand response



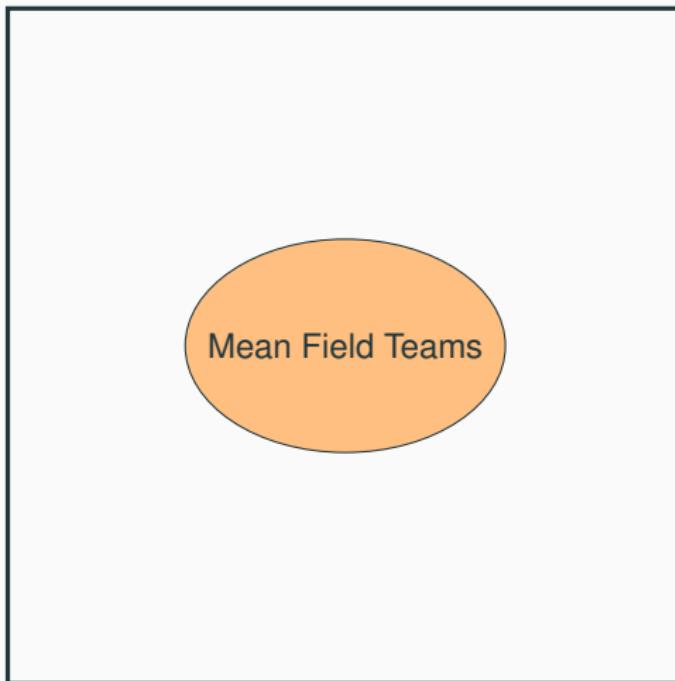
- $x_t^i \in \mathcal{X} = \{OFF, ON\}$, $\textcolor{red}{m}_t = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(x_t^i = OFF)$
- Dynamics: $\mathbb{P}(x_{t+1}^i | x_t^i, u_t^i) =: [P(u_t^i)]_{x_t^i x_{t+1}^i}$
- Actions: $u_t^i \in \mathcal{U} = \{FREE, OFF, ON\}$, Cost of action: $C(u_t^i)$
- Objective: $\min_{\mathbf{g}} \mathbb{E}^{\mathbf{g}} \left[\sum_{t=1}^{\infty} \beta^t \left(\frac{1}{n} \sum_{i=1}^n C(u_t^i) + D(\textcolor{red}{m}_t \| \zeta_t) \right) \right]$.

Numerical example 1: demand response

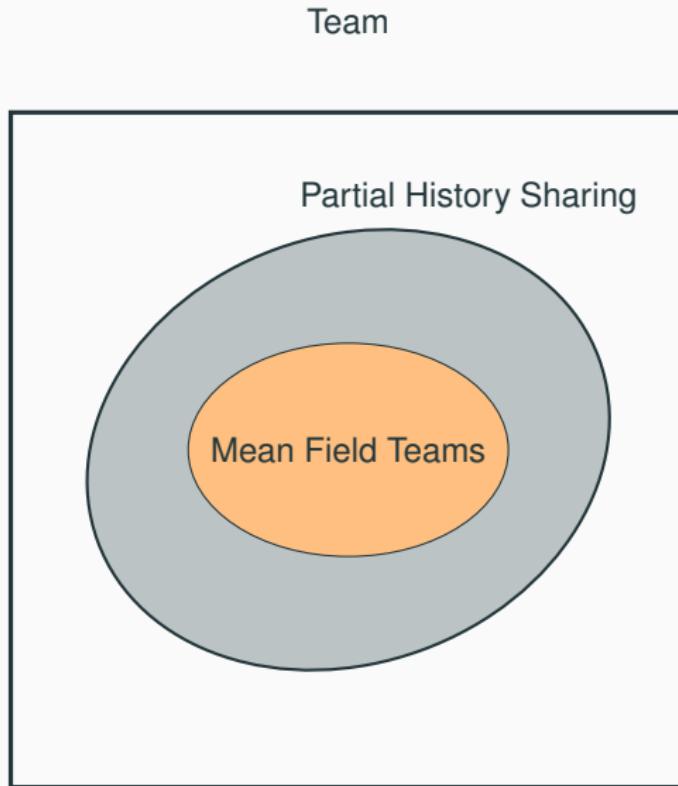
Reinforcement Learning with Partial History Sharing

Reinforcement learning with partial history sharing

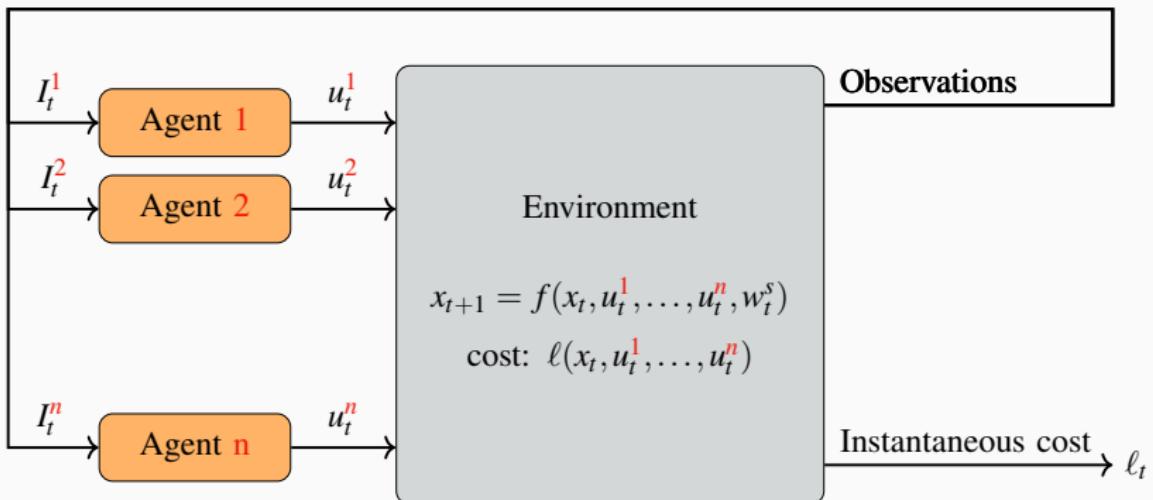
Team



Reinforcement learning with partial history sharing



Reinforcement learning in general team



State: $x_t \in \mathcal{X}$.

Observation: $y_t^i = h(x_t, u_{t-1}^1, \dots, u_{t-1}^n, w_t^{i,o})$.

Control law: $u_t^i = g_t^i(I_t^i)$. Information: $I_t^i \subseteq \{y_{1:t}^1, \dots, y_{1:t}^n, u_{1:t-1}^1, \dots, u_{1:t-1}^n\}$.

System cost: Given $\beta \in (0, 1)$, $J(\mathbf{g}) = \mathbb{E}^{\mathbf{g}} \left[\sum_{t=1}^{\infty} \beta^{t-1} \ell(x_t, u_t^1, \dots, u_t^n) \right]$.

Reinforcement learning in general team

Suppose the system dynamics (f, h) , cost structure ℓ , and probability mass functions are not completely known.

Objective: Given $\epsilon > 0$, find strategy \mathbf{g}_ϵ^* such that

$$J(\mathbf{g}_\epsilon^*) \leq J^* + \epsilon.$$

Reinforcement learning with Partial History Sharing (PHS)

Definition (Partial History Sharing, Nayyer et al. 2013)

Split the information at each agent into two parts:

- *Common information*: $c_t = \bigcap_{i=1}^n I_t^i$ i.e. shared between all agents.
- *Local information*: $m_t^i = I_t^i \setminus c_t$ 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

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

- 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.

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).

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.

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.



[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:

No existing approach to solve decentralized reinforcement learning.

Etc. (active since 80's).

Pre-learning stage

STEP 1: Common Information Approach

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

$$\gamma_t^i := g_t^i(z_{1:t-1}, \cdot) \quad \text{s.t.} \quad u_t^i = \gamma_t^i(m_t^i).$$

Let ψ denote the coordinator's strategy:

$$(\gamma_t^1, \dots, \gamma_t^n) = \psi_t(z_{1:t-1}).$$

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

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

A dynamic program is identified to characterize the optimal strategy based on the information state π .

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

Let \mathcal{R} denote the reachable set of the information state π .

STEP 2: An Approximate POMDP RL Algorithm

Definition (Incrementally Expanding Representation)

Let $\{\mathcal{S}_N\}_{N=1}^{\infty}$ be a sequence of finite sets such that $\mathcal{S}_1 \subsetneq \mathcal{S}_2 \subsetneq \dots \subsetneq \mathcal{S}_N \subsetneq \dots$. Let $\mathcal{S} = \lim_{N \rightarrow \infty} \mathcal{S}_N$ be the countable union of above finite sets. The tuple $\langle \{\mathcal{S}_N\}_{N=1}^{\infty}, B, \tilde{f} \rangle$ is called an *Incrementally Expanding Representation*, if

Incremental Expansion: For any $\gamma \in \mathcal{G}$, $z \in \mathcal{Z}$, and $s \in \mathcal{S}_N$,

$$\tilde{f}(s, \gamma, z) \in \mathcal{S}_{N+1}.$$

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

$$\pi_t = B(s_t).$$

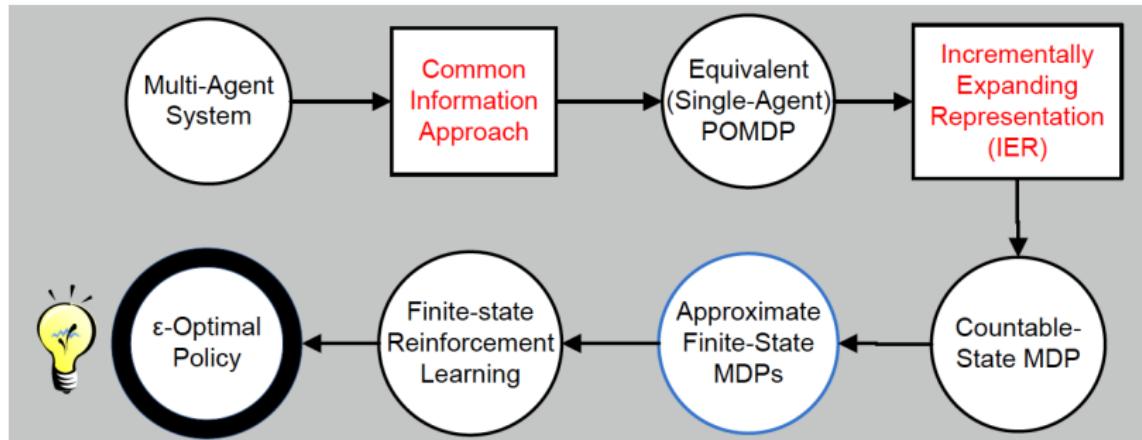
STEP 2: An Approximate POMDP RL Algorithm

Lemma

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

- Construct countable-state MDP Δ with state space \mathcal{S} , action space \mathcal{G} , dynamics \tilde{f} , and cost $\tilde{\ell}(B(s_t), \gamma_t) := \mathbb{E}[\ell(x_t, u_t^1, \dots, u_t^n) | \pi_t, \gamma_t]$.
- Construct an augmented type approximation sequence $\{\Delta_N\}_{N=1}^{\infty}$ of Δ , with state space \mathcal{S}_N , action space \mathcal{G} , dynamics \tilde{f} , and cost $\tilde{\ell}(B(s_t), \gamma_t)$.
- Apply a finite-state RL algorithm \mathcal{T} (such as TD(λ) and Q-learning) to learn optimal strategy of Δ_N . We assume \mathcal{T} converges to optimal strategy of Δ_N .

A Block Diagram



Proposed decentralized RL algorithm

- (1) Given $\epsilon > 0$, choose N such that $\frac{2\beta^N}{1-\beta}(\ell_{max} - \ell_{min}) \leq \epsilon$. Then, construct Δ_N ; particularly, state space \mathcal{S}_N and dynamics \tilde{f} .
- (2) At iteration k , ζ chooses prescriptions $\gamma_k = (\gamma_k^1, \dots, \gamma_k^n)$. (Agents have access to a common random generator to explore consistently). Agent i takes action u_k^i based on prescription γ_k^i and local information m_k^i :

$$u_k^i = \gamma_k^i(m_k^i), \forall i.$$

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

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

- (4) \mathcal{T} learns (updates) the coordinated strategy according to observed cost ℓ_k by performing prescriptions γ_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 $\frac{2\beta^N}{1-\beta}(\ell_{\max} - \ell_{\min}) \leq \epsilon$. Then, construct Δ_N ; particularly, state space \mathcal{S}_N and dynamics \tilde{f} .
- (2) At iteration k , ζ chooses prescriptions $\gamma_k = (\gamma_k^1, \dots, \gamma_k^n)$. (Agents have access to a common random generator to explore consistently). Agent i

The structure of the learned strategy:

$$u_t^i = g_t^i(s_t, m_t^i), \quad i \in \{1, \dots, n\},$$

where s_t is the internal state that changes every time.

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

- (4) \mathcal{T} learns (updates) the coordinated strategy according to observed cost ℓ_k by performing prescriptions γ_k at state s_k and transition to state s_{k+1} .
- (5) $k \leftarrow k + 1$, and go to step 2 until termination.

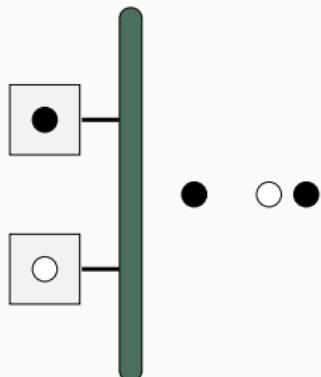
Theorem 6.3

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

Numerical example 2: multi Access Broadcast Channel (MABC)



- $x_t^i \in \{0, 1\}$ with independent arrival probability p^i , $i = 1, 2$.
- $I_t^i = (x_t^i, u_{1:t-1}^1, u_{1:t-1}^2)$.
- $u_t^i \leq x_t^i \in \{0, 1\}$.
- 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 **polynomial** (rather than exponential) in # of agents and **linear** (rather than exponential) in time.
2. Linear quadratic:
 - The optimal solution is **linear**.
 - The solution complexity is **independent of N** and it depends only K .
 - No need to share anything beyond mean field.
 - Each agent solves only two Riccati equations (**distributed computation**).
3. When population is infinite, mean-field is **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.

- **Arbitrarily coupled cost**
- **Infinite horizon**
- **Noisy observation**
- **Major-minor**
- **Randomized strategies**
- **Weighted mean field**
- **Infinite horizon**
- **Partial mean-field sharing**
- **Major-minor**
- **Tracking problem**

teams are robust to node failure.

- Different generalizations.

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 ϵ -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 **guar-**

Three features of designed learning space \mathcal{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 exiting powerful finite state-action RL algorithms.
 - The proposed approach is also **novel in centralized POMDP**.
 - Develop **decentralized Q-learning** for two-user MABC.

Main contributions: Reinforcement Learning with PHS

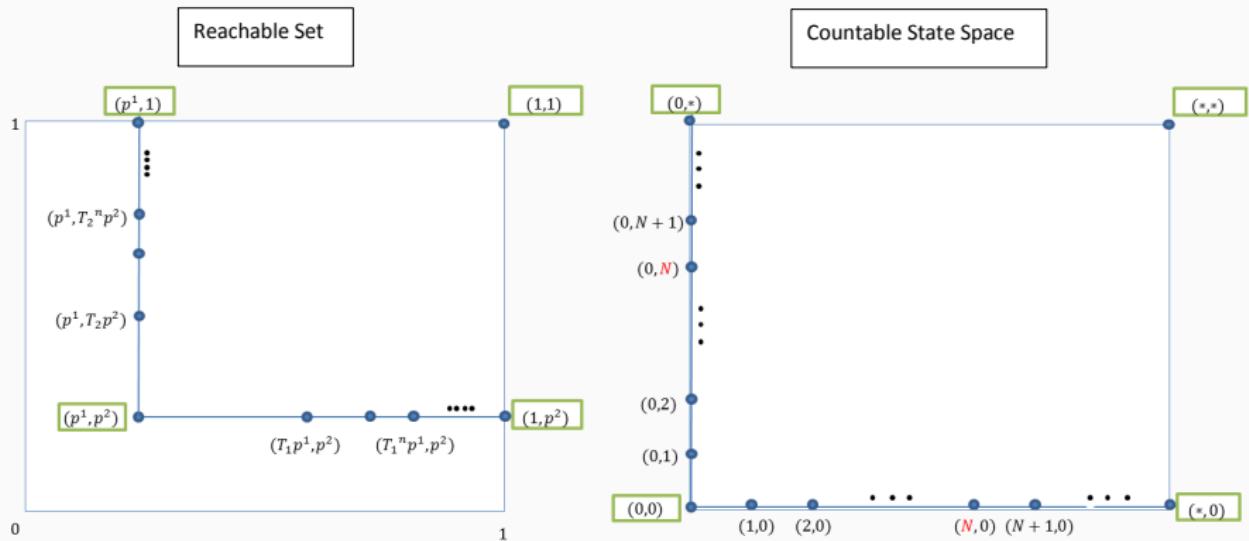
- There is **no existing RL** in team that guarantees optimality .
- Introduce a novel decentralized RL for partial history sharing that **guarantees ϵ -optimal** solution.
- Use **common information approach** and our **proposed approach** to design the learning space.
- Introduce the notion of **Incrementally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

Mean-field team: temperature control

Numerical example 2: temperature control

Reinforcement Learning: Multi-Access Broadcast Channel

Numerical example 3: MABC



-  Both users transmit
-  Only user 1 transmits
-  Only user 2 transmits