Reinforcement learning in decentralized stochastic control

Jalal Arabneydi and Aditya Mahajan
McGill University

American Control Conference (ACC)
3 July, 2015

Learning (or adaptation) in dynamical systems

Learning in centralized systems

- Adaptive control
- Model predictive control
- Reinforcement learning

Various techniques
Relatively well understood.

Learning (or adaptation) in dynamical systems

Learning in centralized systems

- Adaptive control
- Model predictive control
- Reinforcement learning

Various techniques
Relatively well understood.

Learning in decentralized systems

- Learning in games
- Reinforcement learning in teams

Few techniques
Not as well understood

Learning (or adaptation) in dynamical systems

Learning in centralized systems

- Adaptive control
- Model predictive control
- Reinforcement learning

Various techniques
Relatively well understood.

Learning in decentralized systems

- Learning in games
- Reinforcement learning in teams

Few techniques
Not as well understood

We present a new RL algorithm for decentralized systems

Basic Idea

Many reinforcement learning algorithms are based on dynamic programming

- Q-learning
- TD(λ)
- SARSA
- REINFORCE

Basic Idea

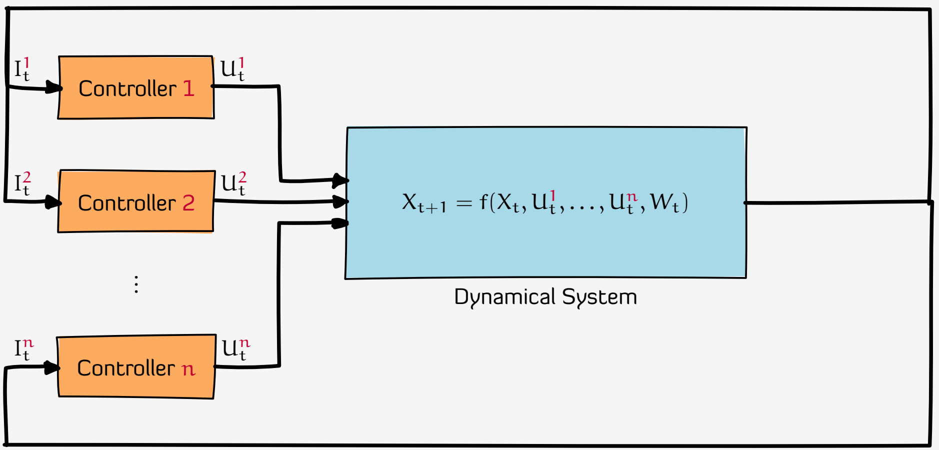
Many reinforcement learning algorithms are based on dynamic programming

- Q-learning
- TD(λ)
- SARSA
- REINFORCE

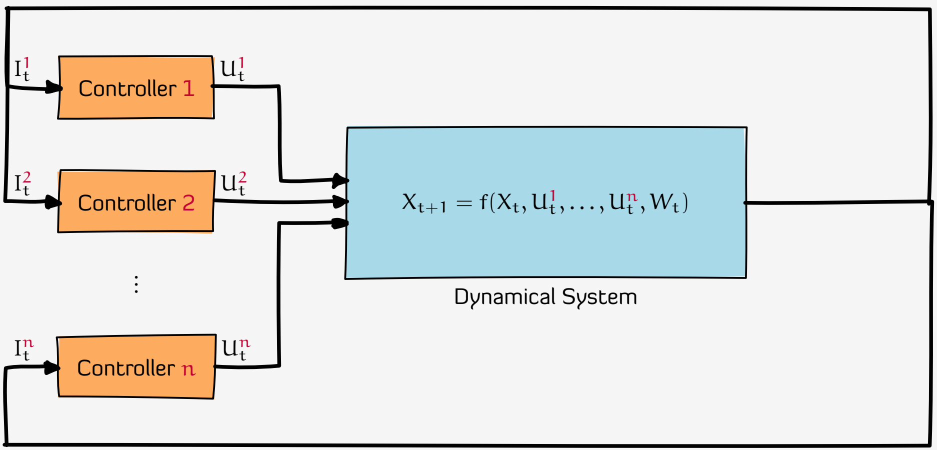
The common-information approach [NMT13] provides dynamic program for a large class of decentralized control systems.

- Nayyar, Mahajan and Teneketzis, "Decentralized stochastic control with partial history sharing: A common information approach," IEEE TAC 2013.

The main result



The main result



Construct a countable state MDP Δ, and a seq. of finite-state MDP approximations Δ_m s.t.

- For any $\varepsilon > 0$, there exists $m(\varepsilon) \leqslant \log(2(\ell_{\max} - \ell_{\min})/\varepsilon(1-\beta))/\log(1/\beta)$ such that running reinforcement learning on MDP $\Delta_{m(\varepsilon)}$ converges to an ε-team-optimal strategy.
- In the worst case, the state space of Δ_m is $O(|\mathcal{Y}|^m|\mathcal{U}|^{m-1})$; but for some models it is $\Theta(m)$.

Outline

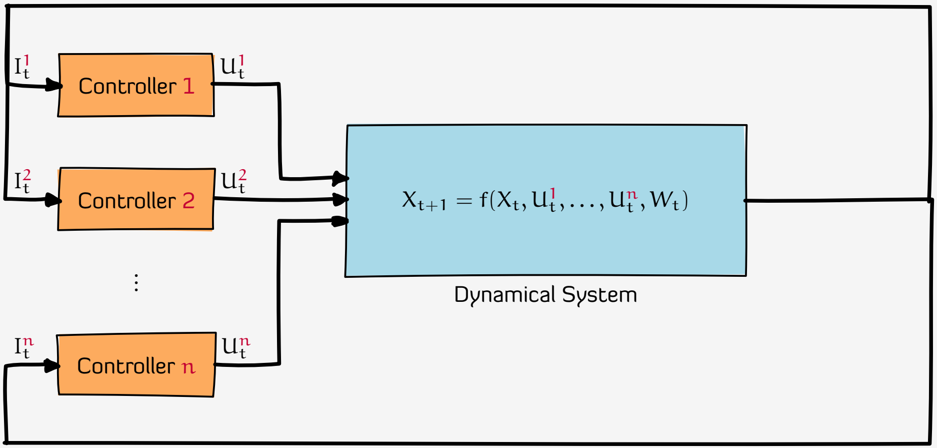
Problem formulation

Solution methodology

- Step 1: Common information approach
- Step 2: Reinforcement learning for POMDPs

Numerical example

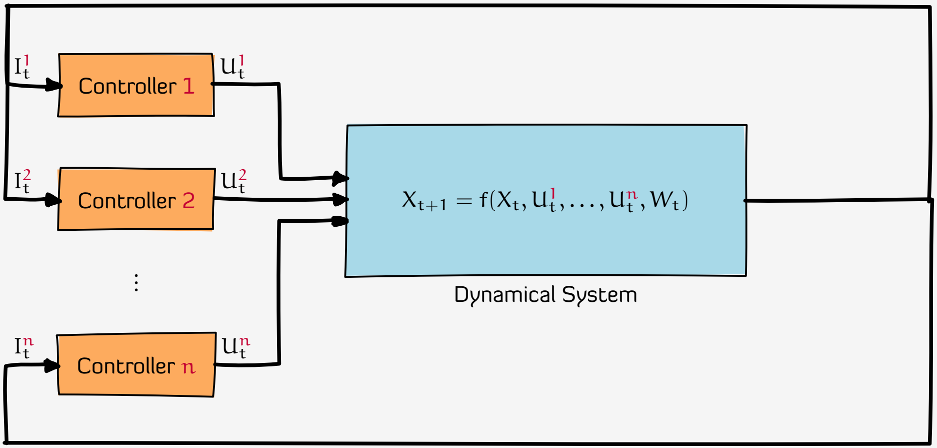
Problem formulation



State X_t Observations $Y_t^i = h(X_t, W_t^i)$ Control $U_t^i = g_t^i(I_t^i)$.

Total cost $J(\mathbf{g}) = \mathbb{E}^{\mathbf{g}}\Big[\sum_{t=1}^{\infty} \beta^{t-1} \ell(X_t, U_t^1, \dots, U_t^n)\Big].$

Problem formulation



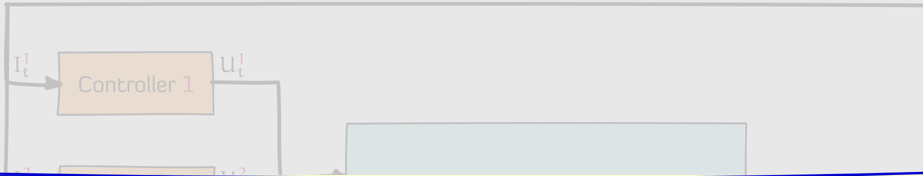
State X_t **Observations** $Y_t^i = h(X_t, W_t^i)$ **Control** $U_t^i = g_t^i(I_t^i)$.

Total cost $J(\mathbf{g}) = \mathbb{E}^{\mathbf{g}}\left[\sum_{t=1}^{\infty} \beta^{t-1} \ell(X_t, U_t^1, \dots, U_t^n)\right].$

Information structure $I_t^i \subseteq \{Y_{1:t}^1, \dots, Y_{1:t}^n, U_{1:t-1}^1, \dots, U_{1:t-1}^n\}$

Assumed to satisfy partial-history sharing model.

Problem formulation



Objective: Given $\varepsilon > 0$, develop a (model-based or model-free) reinforcement learning algorithm that finds an ε-optimal strategy $\mathbf{g}_\varepsilon$ such that

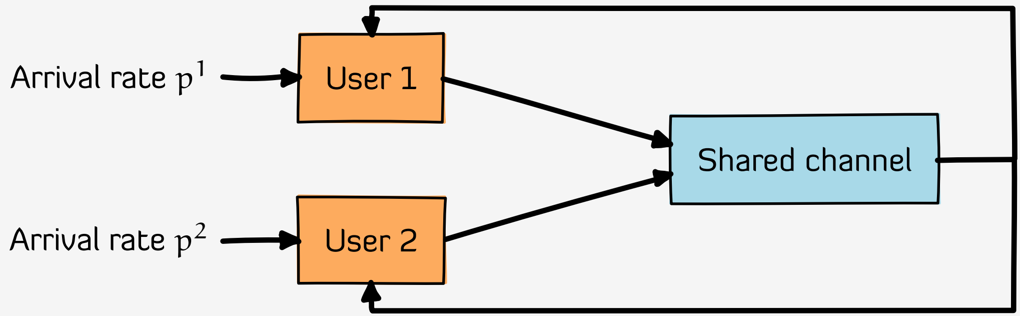
$$J(\mathbf{g}_\varepsilon) - J(\mathbf{g}) \leqslant \varepsilon$$

Total cost $\quad J(\mathbf{g}) = \mathbb{E}^{\mathbf{g}}\Big[\sum_{t=1}^{\infty} \beta^{t-1} \ell(X_t, U_t^1, \dots, U_t^n)\Big].$

Information structure $\quad I_t^i \subseteq \{Y_{1:t}^1, \dots, Y_{1:t}^n, U_{1:t-1}^1, \dots, U_{1:t-1}^n\}$

Assumed to satisfy partial-history sharing model.

Multi-access broadcast example



Arrival process

- Bernoulli process with rate p^i.
- Arrival rates (p^1, p^2) are unknown.

Control actions

- $U_t^i \in \{0, 1\} = \{$ Don't transmit, transmit $\}$
- If only one user transmits, packet goes through.
- If both users transmit, packets collide and don't go through.

Information structure

$I_t^i = \{X_t^i, U_{1:t-1}^1, U_{1:t-1}^2\}$

Objective

Maximize throughput (# of successfully transmissions)

Solution methodology

The basic idea

Step 1 Follow the common information approach [NMT13] to convert the decentralized control problem into a centralized (partially-observed) control problem

Step 2 Use a Reinforcement-learning algorithm for POMDPs to learn the optimal strategy when the model is unknown

Solution methodology

The basic idea

Step 1 Follow the common information approach [NMT13] to convert the decentralized control problem into a centralized (partially-observed) control problem

Step 2 Use a Reinforcement-learning algorithm for POMDPs to learn the optimal strategy when the model is unknown

We propose a new reinforcement-learning algorithm for POMDPs

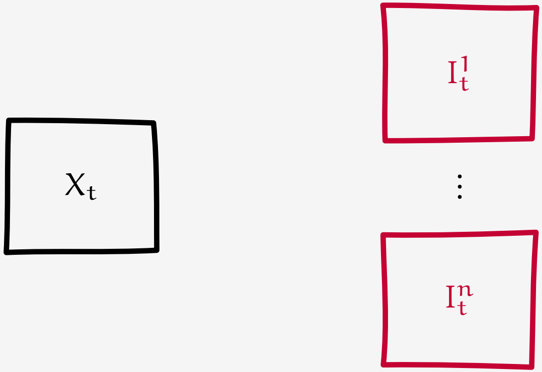
- Given a belief state, the reachable set of belief states (under all strategies) is countable. Therefore,

$$\text{POMDP} \equiv \text{Countable state MDP}$$

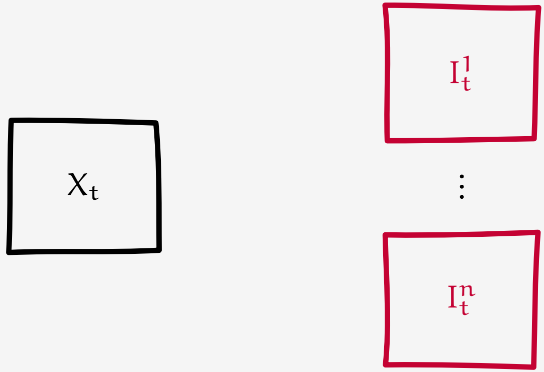
- Countable state MDPs can be approximated by a sequence of finite-state MDPs.
- Guarantees convergence to ε-optimal strategy.

Step 1: Converting the decentralized system into an equivalent centralized system

An overview of the common-information approach [NMT13]



An overview of the common-information approach [NMT13]

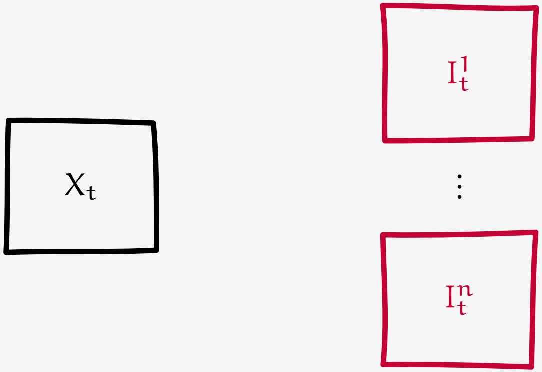


Common infomation $\quad C_t = \bigcap_{\tau \geqslant t} \bigcap_{i=1}^{n} I_\tau^i, \qquad Z_t = C_t \setminus C_{t-1}.$

Local information $\quad L_t^i = I_t^i \setminus C_t.$

Prescriptions $\quad \gamma_t^i \colon L_t^i \mapsto U_t^i.$

An overview of the common-information approach [NMT13]



Common infomation $$C_t = \bigcap_{\tau \geqslant t} \bigcap_{i=1}^{n} I_\tau^i, \qquad Z_t = C_t \setminus C_{t-1}.$$

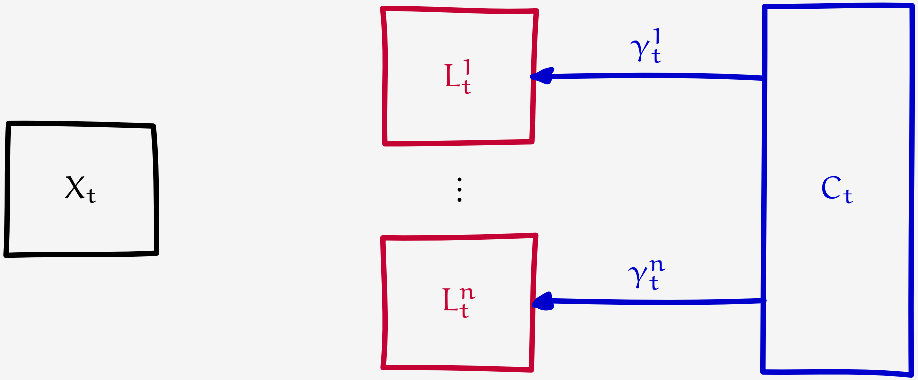
Local information $$L_t^i = I_t^i \setminus C_t.$$

Prescriptions $$\gamma_t^i \colon L_t^i \mapsto U_t^i.$$

Partial history sharing

- $|\mathcal{L}_t^i|$ is uniformly bounded
- $L_{t+1}^i \subseteq \{L_t^i, U_t^i Y_{t+1}^i\}$

An overview of the common-information approach [NMT13]



Common infomation $$C_t = \bigcap_{\tau \geqslant t} \bigcap_{i=1}^{n} I_\tau^i, \qquad Z_t = C_t \setminus C_{t-1}.$$

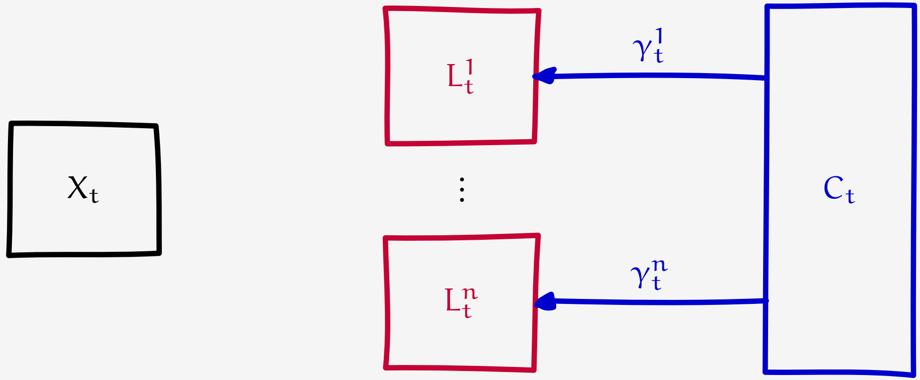
Local information $$L_t^i = I_t^i \setminus C_t.$$

Prescriptions $$\gamma_t^i \colon L_t^i \mapsto U_t^i.$$

Partial history sharing

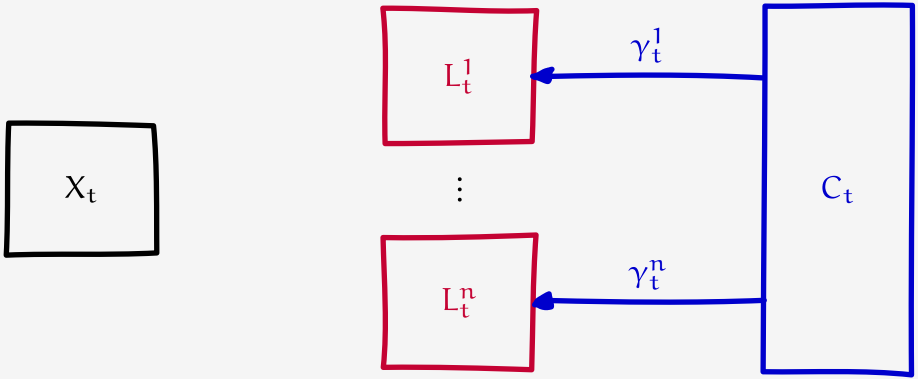
- $|\mathcal{L}_t^i|$ is uniformly bounded
- $L_{t+1}^i \subseteq \{L_t^i, U_t^i Y_{t+1}^i\}$

An overview of the common-information approach [NMT13]



	Original System	**Coordinated System**
Information structure	I_t^i (Note: $I_t^i \not\subseteq I_{t+1}^j$)	C_t (Note: $C_t \subseteq C_{t+1}$)
Control action	$U_t^i = g_t^i(C_t, L_t^i)$	$\Gamma_t^i = \psi_t^i(C_t)$, where $\gamma_t^i \colon L_t^i \mapsto U_t^i$

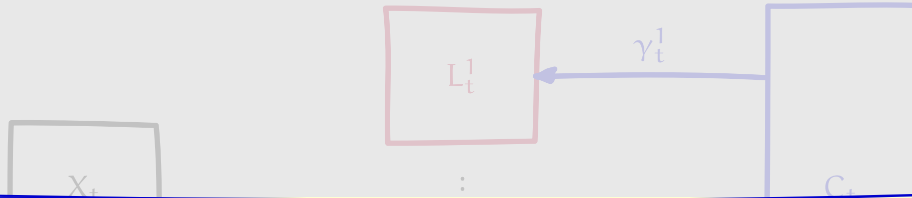
An overview of the common-information approach [NMT13]



	Original System	**Coordinated System**
Information structure	I_t^i (Note: $I_t^i \not\subseteq I_{t+1}^j$)	C_t (Note: $C_t \subseteq C_{t+1}$)
Control action	$U_t^i = g_t^i(C_t, L_t^i)$	$\Gamma_t^i = \psi_t^i(C_t)$, where $\gamma_t^i \colon L_t^i \mapsto U_t^i$

If we choose $g_t^i(C_t, L_t^i) = \psi_t^i(C_t)(L_t^i)$, the both systems have identical realization of system variables. Hence, the systems are equivalent.

An overview of the common-information approach [NMT13]



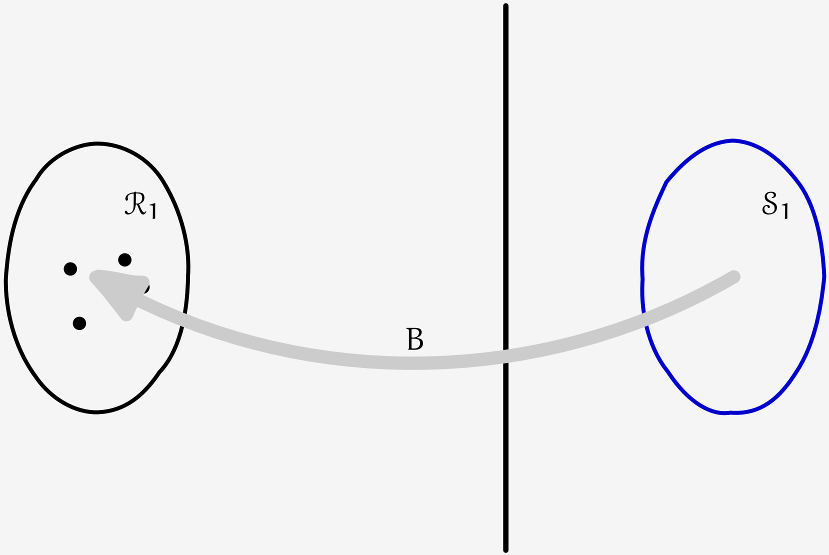
The coordinated system is a centralized system (POMDP Π). We can use any standard method to identify an information state and obtain a dynamic program!

Information structure	I^i_t (Note: $I^i_t \not\subseteq I^j_{t+1}$)	C_t (Note: $C_t \subseteq C_{t+1}$)
Control action	$U^i_t = g^i_t(C_t, L^i_t)$	$\Gamma^i_t = \psi^i_t(C_t)$, where $\gamma^i_t \colon L^i_t \mapsto U^i_t$

If we choose $g^i_t(C_t, L^i_t) = \psi^i_t(C_t)(L^i_t)$, the both systems have identical realization of system variables. Hence, the systems are equivalent.

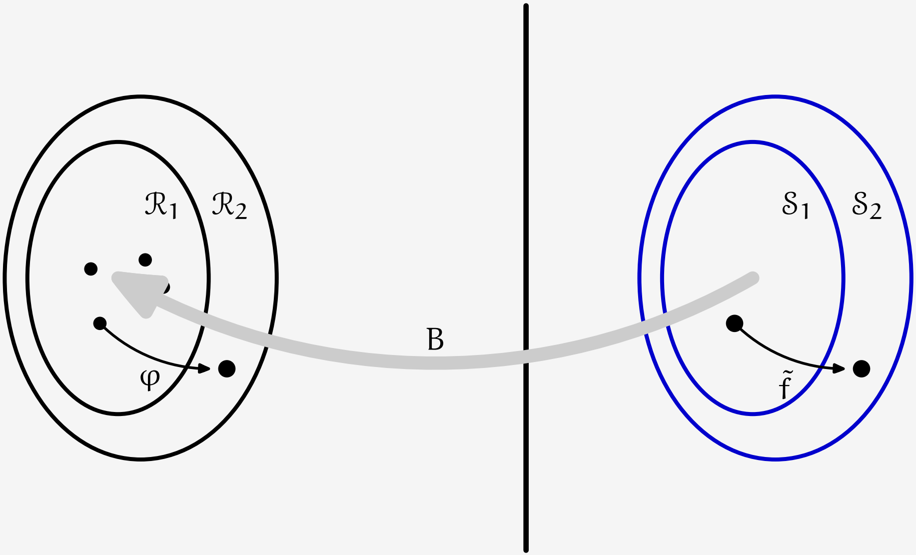
Step 2: Reinforcement learning for POMDPs

Incrementally expanding representation



- $\mathcal{R}_1 =$ (Finite) set of initial information states
- $\mathcal{S}_1 =$ A set isomorphic to $\mathcal{R}_1$ that does not depend on the unknowns.
- Surjection B between $\mathcal{R}_1$ and $\mathcal{S}_1$

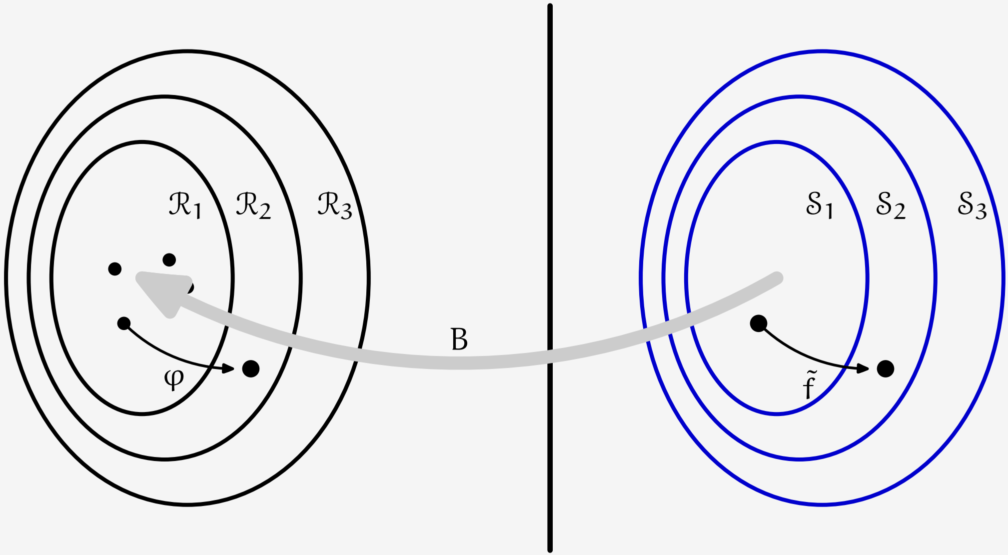
Incrementally expanding representation



- $\mathcal{R}_2 = \{\varphi(\pi, z, \gamma) : \pi \in \mathcal{R}_1, z \in \mathcal{Z}, \gamma \in \Gamma\}$.
- There exists a function $\tilde{f}$ (that does not depend on unknowns) such that for every $s \in \mathcal{S}_1$, $z \in \mathcal{Z}$, $\gamma \in \Gamma$

$$B(\tilde{f}(s, z, \gamma)) = \varphi(B(s), z, \gamma), \qquad \mathcal{S}_2 = \{\tilde{f}(s, z, \gamma) : s \in \mathcal{S}_1, z \in \mathcal{Z}, \gamma \in \Gamma\}$$

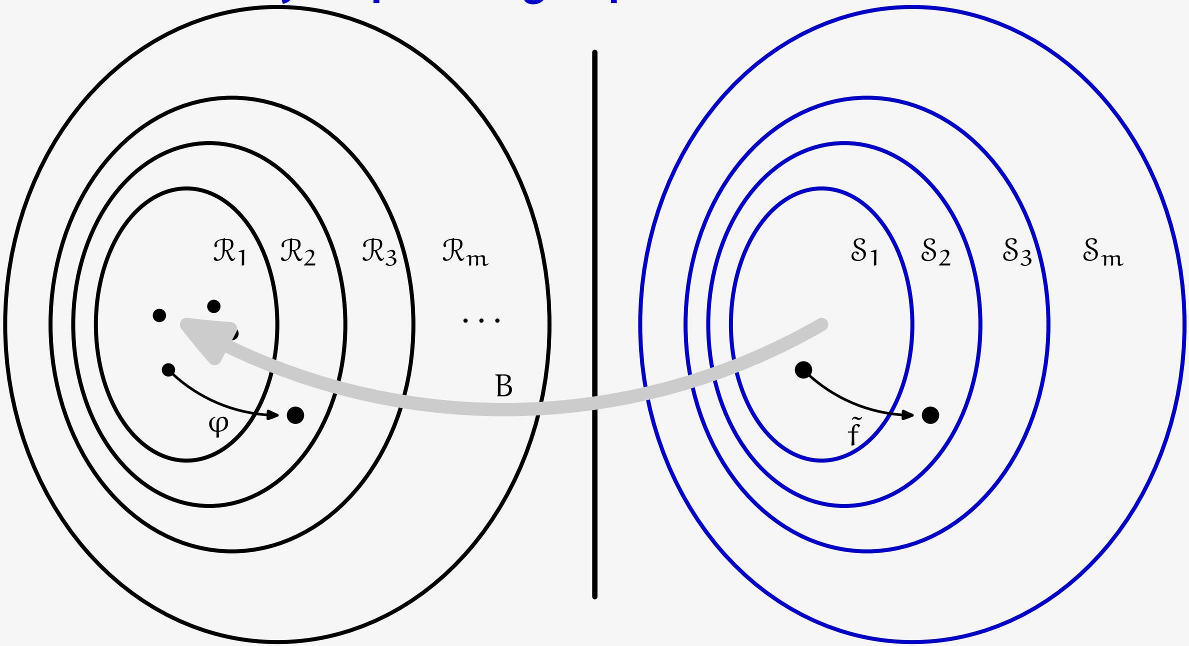
Incrementally expanding representation



- $\mathcal{R}_2 = \{\varphi(\pi, z, \gamma) : \pi \in \mathcal{R}_1, z \in \mathcal{Z}, \gamma \in \Gamma\}$.
- There exists a function $\tilde{f}$ (that does not depend on unknowns) such that for every $s \in \mathcal{S}_1$, $z \in \mathcal{Z}$, $\gamma \in \Gamma$

$$B(\tilde{f}(s, z, \gamma)) = \varphi(B(s), z, \gamma), \qquad \mathcal{S}_2 = \{\tilde{f}(s, z, \gamma) : s \in \mathcal{S}_1, z \in \mathcal{Z}, \gamma \in \Gamma\}$$

Incrementally expanding representation



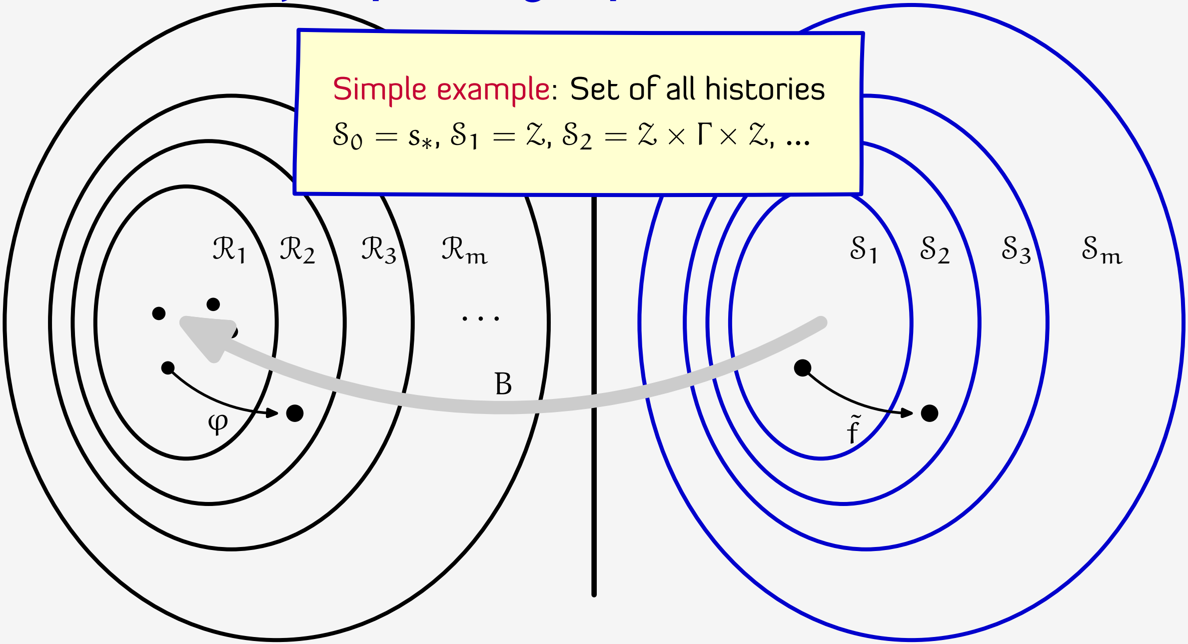
- $\mathcal{R}_2 = \{\varphi(\pi, z, \gamma) : \pi \in \mathcal{R}_1, z \in \mathcal{Z}, \gamma \in \Gamma\}$.
- There exists a function $\tilde{f}$ (that does not depend on unknowns) such that for every $s \in \mathcal{S}_1$, $z \in \mathcal{Z}$, $\gamma \in \Gamma$

$$B(\tilde{f}(s, z, \gamma)) = \varphi(B(s), z, \gamma), \qquad \mathcal{S}_2 = \{\tilde{f}(s, z, \gamma) : s \in \mathcal{S}_1, z \in \mathcal{Z}, \gamma \in \Gamma\}$$

Incrementally expanding representation

Simple example: Set of all histories

$\mathcal{S}_0 = s_*$, $\mathcal{S}_1 = \mathcal{Z}$, $\mathcal{S}_2 = \mathcal{Z} \times \Gamma \times \mathcal{Z}$, ...



- $\mathcal{R}_2 = \{\varphi(\pi, z, \gamma) : \pi \in \mathcal{R}_1, z \in \mathcal{Z}, \gamma \in \Gamma\}$.
- There exists a function $\tilde{f}$ (that does not depend on unknowns) such that for every $s \in \mathcal{S}_1$, $z \in \mathcal{Z}$, $\gamma \in \Gamma$

$$B(\tilde{f}(s, z, \gamma)) = \varphi(B(s), z, \gamma), \qquad \mathcal{S}_2 = \{\tilde{f}(s, z, \gamma) : s \in \mathcal{S}_1, z \in \mathcal{Z}, \gamma \in \Gamma\}$$

Incrementally expanding representation

Formal definition

An IER is a tuple $\langle \{\mathcal{S}_m\}_{m=1}^{\infty}, \tilde{f}, B \rangle$ such that $\{\mathcal{S}_m\}_{m=1}^{\infty}$ and $\tilde{f}$ do not depend on the unknowns and

- $\mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \cdots \subseteq \mathcal{S}_m \cdots$
- For any $s \in \mathcal{S}_m$, $z \in \mathcal{Z}$, $\gamma \in \Gamma$, $\quad \tilde{f}(s, z, \gamma) \in \mathcal{S}_{m+1}$.
- Let $\mathcal{S} = \lim_{m \to \infty} \mathcal{S}_m$. Then B is a surjective map from $\mathcal{S}$ to Π such that $\pi_t = B(s_t)$

Note

The surjection B may depend on the unknowns.

Lemma

A decentralized control system with partial history sharing has at least one IER.

An equivalent (countable-state) MDP

Countable state MDP Δ

State Space $\mathcal{S}$, Dynamics $\tilde{f}$, Action Space Γ

Cost function $\tilde{\ell}(s, \gamma) = \mathbb{E}[\ell(X, \mathbf{U}) | \pi = B(s), \gamma]$

An equivalent (countable-state) MDP

Countable state MDP Δ

State Space $\mathcal{S}$, Dynamics $\tilde{f}$, Action Space Γ

Cost function $\tilde{\ell}(s, \gamma) = \mathbb{E}[\ell(X, \mathbf{U}) | \pi = B(s), \gamma]$

Theorem

The optimal strategy of MDP Δ is equivalent to the optimal strategy of POMDP Π.

An equivalent (countable-state) MDP

Countable state MDP Δ

State Space $\mathcal{S}$, Dynamics $\tilde{f}$, Action Space Γ

Cost function $\tilde{\ell}(s, \gamma) = \mathbb{E}[\ell(X, \mathbf{U}) | \pi = B(s), \gamma]$

Theorem

The optimal strategy of MDP Δ is equivalent to the optimal strategy of POMDP Π.

Finite state MDP Δ_m

Consider the following truncated dynamics $\tilde{f}_m$ on $\mathcal{S}_m$.
Pick a set $D^\circ \in \mathcal{S}_m$ such that for all $s \in \mathcal{S}_m$, $z \in \mathcal{Z}$, $\gamma \in \Gamma$, set $\tilde{f}_m(s, z, \gamma) \in D^\circ$.

For RL, this is only possible if there exists a reset action or a homing strategy.

An equivalent (countable-state) MDP

Countable state MDP Δ

State Space $\mathcal{S}$, Dynamics $\tilde{f}$, Action Space Γ
Cost function $\tilde{\ell}(s, \gamma) = \mathbb{E}[\ell(X, \mathbf{U}) | \pi = B(s), \gamma]$

Theorem

The optimal strategy of MDP Δ is equivalent to the optimal strategy of POMDP Π.

Finite state MDP Δ_m

Consider the following truncated dynamics $\tilde{f}_m$ on $\mathcal{S}_m$. Pick a set $D^\circ \in \mathcal{S}_m$ such that for all $s \in \mathcal{S}_m$, $z \in \mathcal{Z}$, $\gamma \in \Gamma$, set $\tilde{f}_m(s, z, \gamma) \in D^\circ$.

For RL, this is only possible if there exists a reset action or a homing strategy.

Theorem

MDPs $\{\Delta_m\}_{m=1}^{\infty}$ is an augmentation type approximation sequence of MDP Δ [Sennott99].

Therefore, $V_m^* \to V^*$ and any limit point of the sequence $\{\psi_m^*\}$ is optimal for Δ.

An IER converts the POMDP to an equivalent countable state MDP whose state space and dynamics do not depend on the unknowns.

The countable state MDP may be approximated by a sequence of finite state MDPs

Approximation error and RL algorithm

Theorem

The difference in performance between MDP Δ and MDP Δ_m is bounded.

$$|J(\psi^*) - J_m(\psi^*_m)| \leqslant 2(\ell_{\text{max}} - \ell_{\text{min}}) \frac{\beta^{\tau_m}}{1-\beta},$$

where $\tau_m \geqslant m$ is a model-dependent parameter.

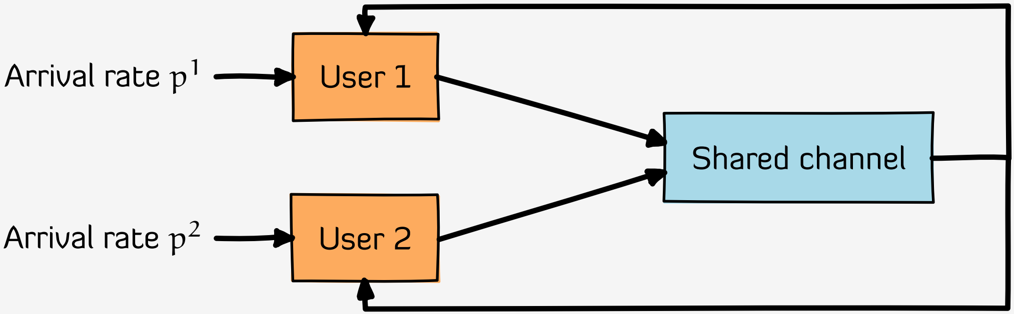
ε-optimal RL

- Given an ε, pick m such that

$$2(\ell_{\text{max}} - \ell_{\text{min}}) \frac{\beta^{\tau_m}}{1-\beta} < \varepsilon.$$

- Use any RL algorithm for the finite-state MDP Δ_m.

Multi-access broadcast example



Information structure $\quad I_t^i = \{X_t^i, U_{1:t-1}^1, U_{1:t-1}^2\}$

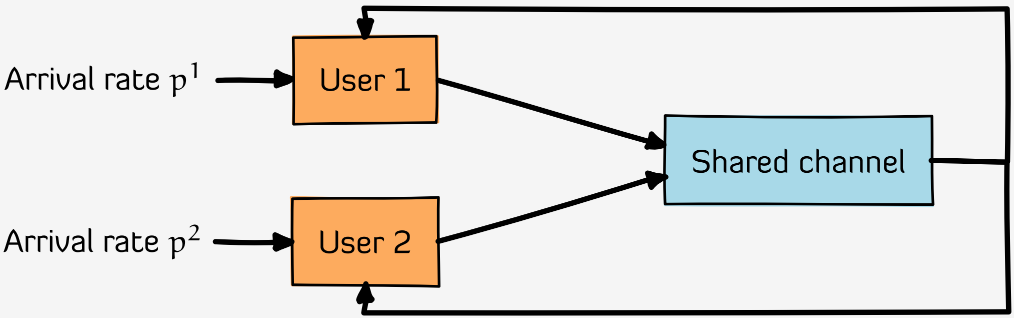
Common information $= U_{1:t-1}^1, U_{1:t-1}^2$ $\quad$ Local information $= X_t^i$

Prescriptions $\quad \gamma_t^i : \{0, 1\} \to \{0, 1\}$

For ease of notation, let $d_t^i = \gamma_t^i(1)$. Then

$$U_t^i = d_t^i X_t^i$$

Multi-access broadcast example



Information structure $\quad I_t^i = \{X_t^i, U_{1:t-1}^1, U_{1:t-1}^2\}$

Common information $= U_{1:t-1}^1, U_{1:t-1}^2$ $\quad$ Local information $= X_t^i$

Prescriptions $\quad \gamma_t^i : \{0, 1\} \to \{0, 1\}$

For ease of notation, let $d_t^i = \gamma_t^i(1)$. Then

$$U_t^i = d_t^i X_t^i$$

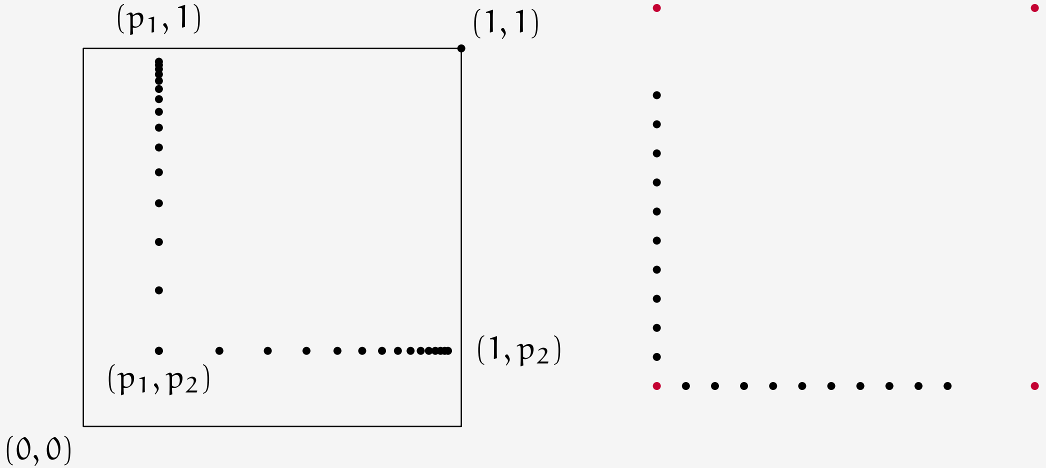
POMDP Π

Info. state $\quad \mathbb{P}(X_t^1, X_t^2 \mid U_{1:t-1}^1, U_{1:t-1}^2)$

$$\equiv \big(\mathbb{P}(X_t^1 \mid \mathbf{U}_{1:t-1}), \mathbb{P}(X_t^2 \mid \mathbf{U}_{1:t-1}) \big)$$

Action space $\quad \{(d^1, d^2)\} = \{(0, 0), (1, 0), (0, 1), (1, 1)\}$

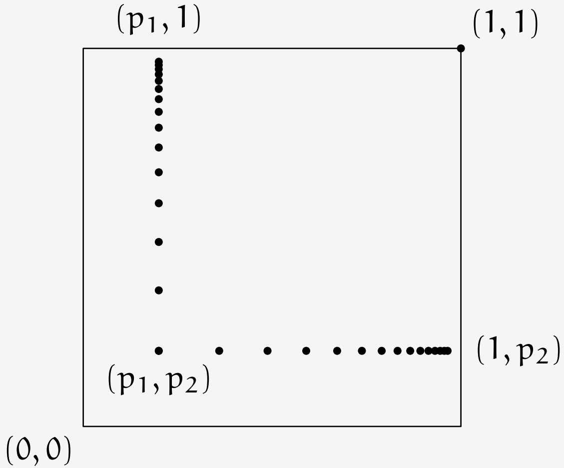
IER for multi-access broadcast

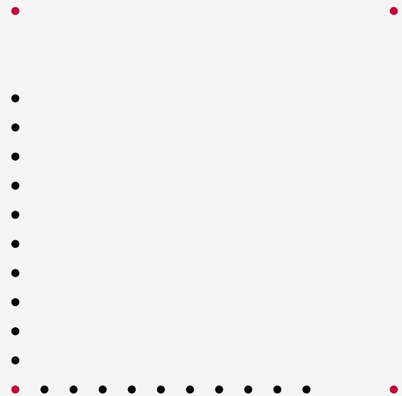


Parameters (p^1, p^2) are unknown.

Reachable set $\mathcal{R} = \{(p^1, p^1), (p^1, 1), (1, p^2), ((T^1)^m p^1, p^2), (p^1, (T^2)^m p^2)\}$.

IER for multi-access broadcast





Parameters (p^1, p^2) are unknown.
Reachable set $\mathcal{R} = \{(p^1, p^1), (p^1, 1), (1, p^2), ((T^1)^m p^1, p^2), (p^1, (T^2)^m p^2)\}$.

IER Space $\mathcal{S} = \{(0, 0), (0, *), (*, 0), (*, *), (m, 0), (0, m)\}$

Numerical Examples

- Parameters: $p^1 = 0.3$, $p^2 = 0.6$, $\beta = 0.99$, $m = 20$
- $(d^1, d^2) = (1, 0)$, $\quad (d^1, d^2) = (0, 1)$, $\quad (d^1, d^2) = (1, 1)$.
- Reachable set under optimal strategy $\{(0, 1), (1, 0), (2, 0), (3, 0)\}$

Numerical Examples

- Parameters: $p^1 = 0.1$, $p^2 = 0.3$, $\beta = 0.99$, $m = 20$
- $(d^1, d^2) = (1, 0)$, $\quad (d^1, d^2) = (0, 1)$, $\quad (d^1, d^2) = (1, 1)$.

- Reachable set under optimal strategy $\{(0, 0), (0, 1), (1, 0), (*, 0), (*, *)\}$

Summary

A (model-based or model-free) reinforcement learning algorithm

Guarantees ε-optimality for a large class of decentralized systems control systems with partial history sharing.

Two steps: Common information approach and POMDP reinforcement learning

Developed a new approximate RL algorithm for POMDPs

Salient features

- The algorithm is based on information commonly known to all controllers. Therefore, it can be executed in a distributed manner
- All controllers need access to a shared random number generator for exploring the system consistently.
- The cost function should be known, otherwise all controllers need to observe the per-step cost.
- In practice, the actual error is much less than the obtained error bound.