Leveraging structure in online and collaborative learning problems

Mahsa Asadi

Supervisors:
Marc Tommasi
Aurelien Bellet

In collaboration with:
Odalric-Ambrym Maillard

November 25, 2022

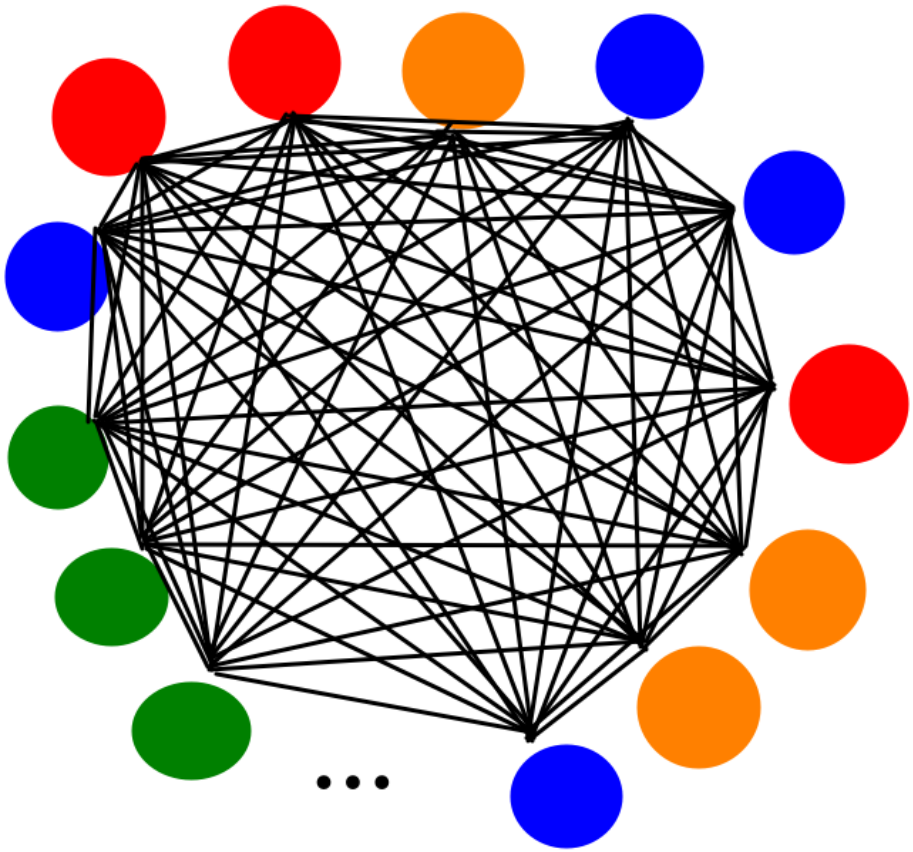
Table of Contents

Introduction

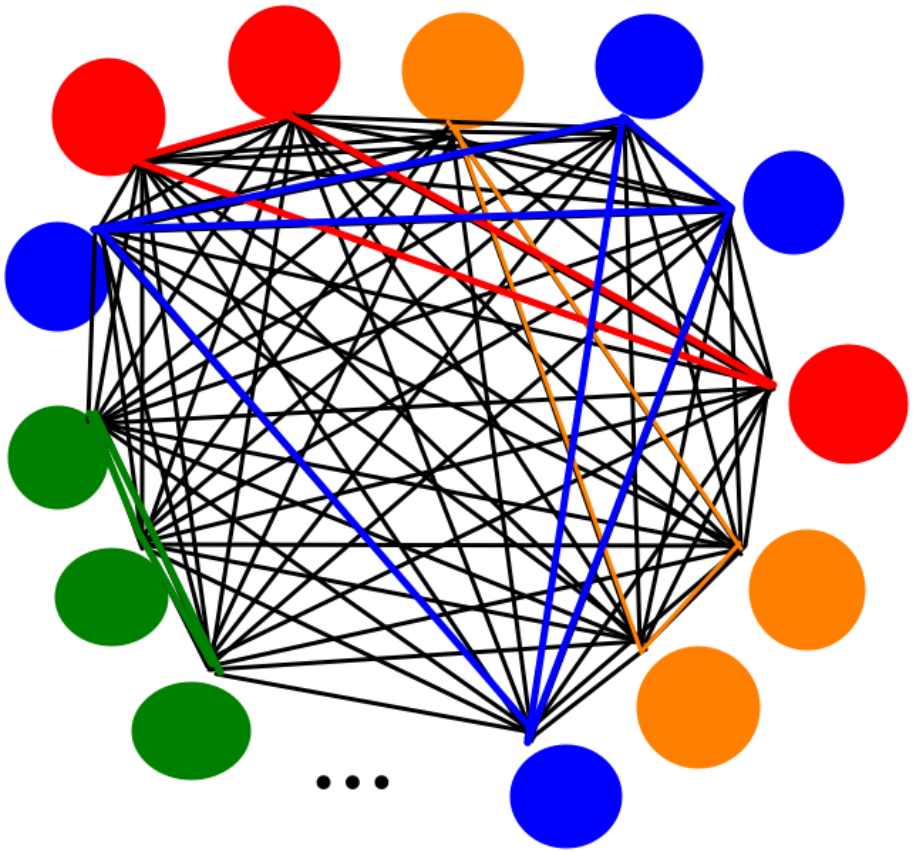
House Navigation Problem

House Navigation Problem

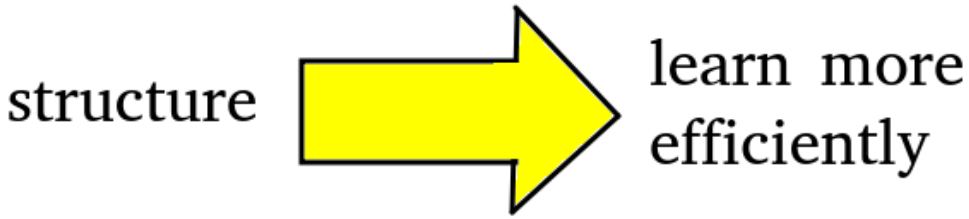
Online Plant Distribution Learning Network



Online Plant Distribution Learning Network



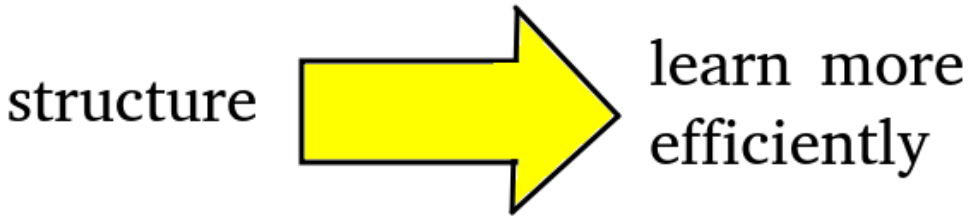
This thesis leverages structure in an online way!



Aggregate data and reuse of information:

- Less samples
- Less time to collect the samples

This thesis leverages structure in an online way!



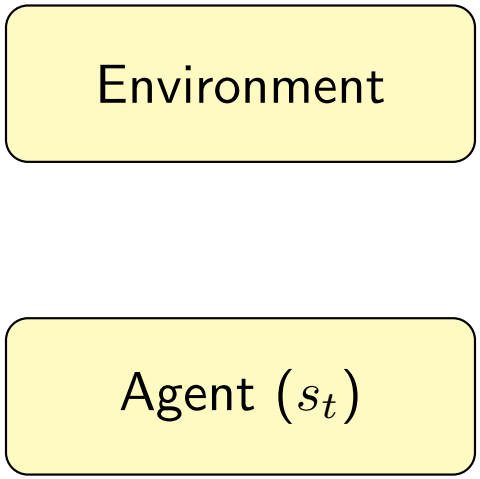
Aggregate data and reuse of information:

- Less samples
- Less time to collect the samples

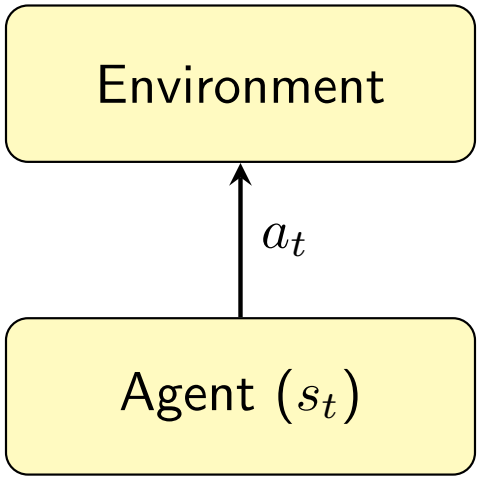
How do we identify the structure
when it is unknown?

Exploiting Structure in Model-based Reinforcement Learning

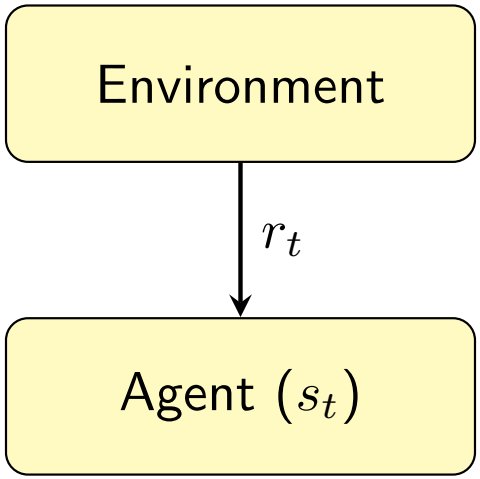
What is Reinforcement Learning(RL)?



What is Reinforcement Learning(RL)?



What is Reinforcement Learning(RL)?



What is Reinforcement Learning(RL)?

Environment

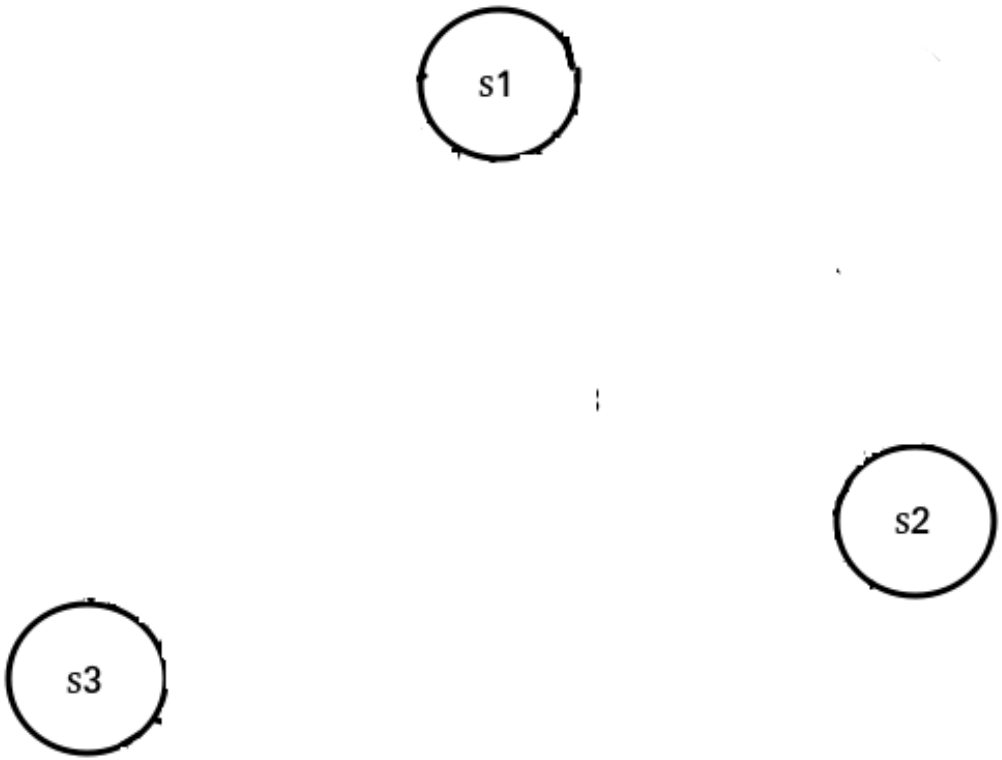
Agent(s_{t+1})

Reinforcement Learning

Definition (Markov Decision Process)

An Markov Decision Process (MDP) is described by four components:

- $\mathcal{S}$,

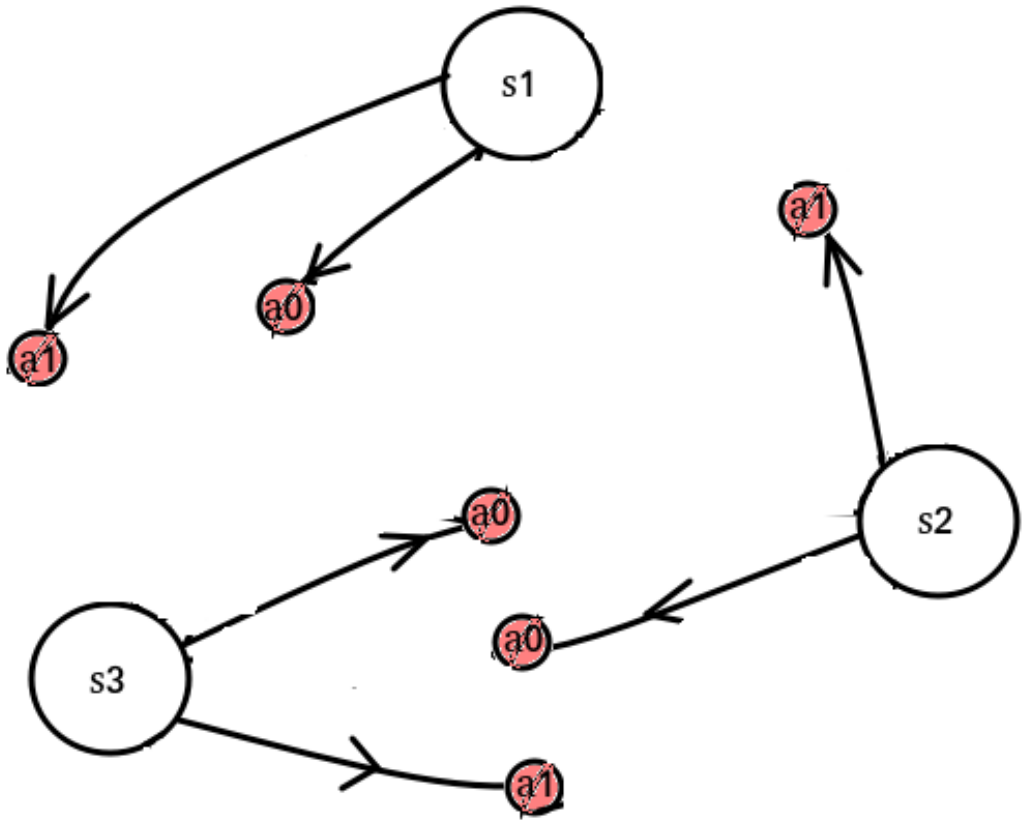


Reinforcement Learning

Definition (Markov Decision Process)

An Markov Decision Process (MDP) is described by four components:

- $\mathcal{S}$, $\mathcal{A}$,

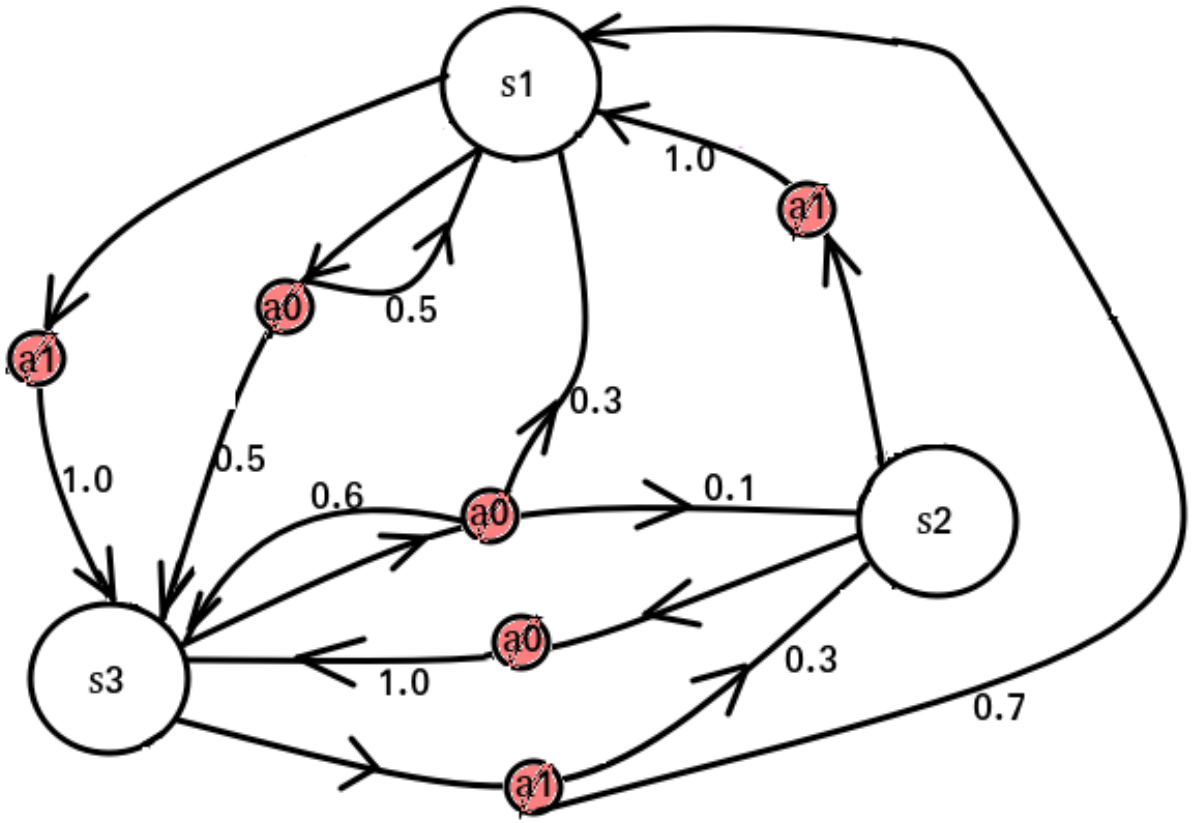


Reinforcement Learning

Definition (Markov Decision Process)

An Markov Decision Process (MDP) is described by four components:

- $\mathcal{S}$, $\mathcal{A}$, $p(s_{t+1}|s_t, a_t)$,

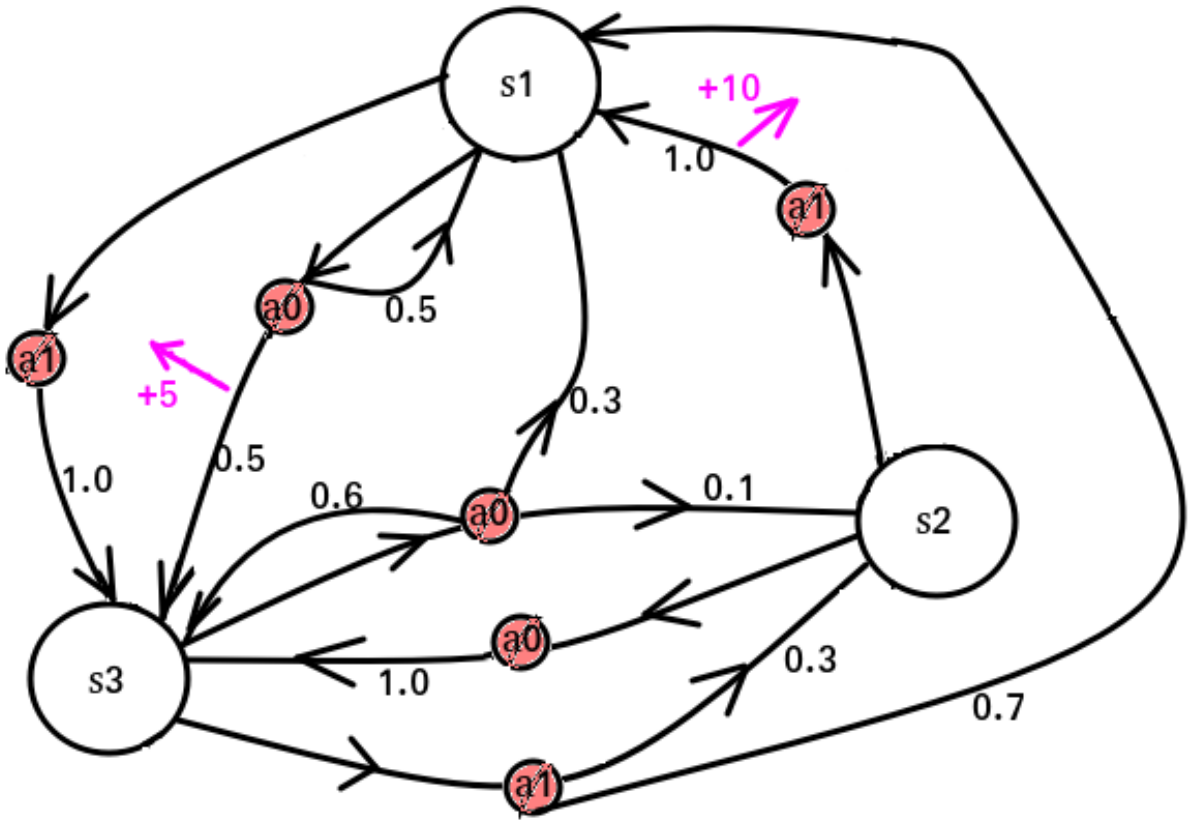


Reinforcement Learning

Definition (Markov Decision Process)

An Markov Decision Process (MDP) is described by four components:

- $\mathcal{S}$, $\mathcal{A}$, $p(s_{t+1}|s_t, a_t)$, $\nu(s_t, a_t)$.



Reinforcement Learning

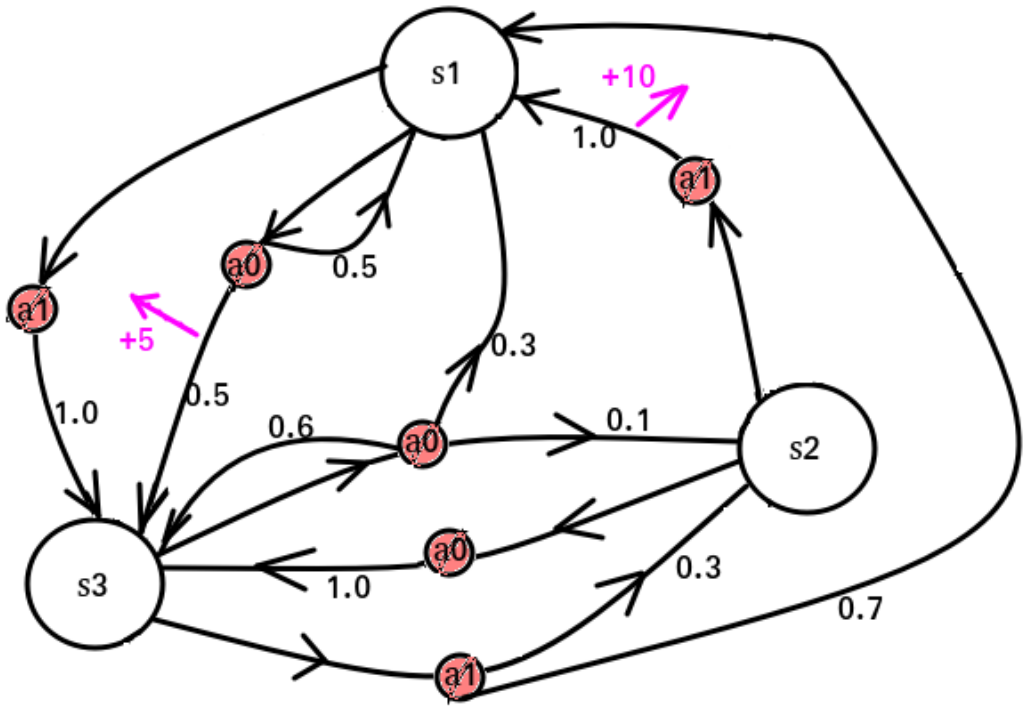
Definition (Markov Decision Process)

An Markov Decision Process (MDP) is described by four components:

- $\mathcal{S}$, $\mathcal{A}$, $p(s_{t+1}|s_t, a_t)$, $\nu(s_t, a_t)$.

$|\mathcal{S}| = S, |\mathcal{A}| = A$

μ is mean of ν



Reinforcement Learning

Goal:

- Learn a *good* policy.
- Minimize regret:

$$\mathbb{R}(T) = \sum_{t=1}^{T} g_\star - \sum_{t=1}^{T} r_t(s_t, a_t)$$

where $g_\star$ is the average reward of optimal policy.

Model-based Reinforcement Learning

- Set of probable MDPs:

$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) :$$

Model-based Reinforcement Learning

- Set of probable MDPs:

$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) : p \in \text{ConfidenceSet}$$

Model-based Reinforcement Learning

- Set of probable MDPs:

$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) : p \in \text{ConfidenceSet} \wedge \mu \in \text{ConfidenceSet}'\}$$

Model-based Reinforcement Learning

- Set of probable MDPs:

$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) : p \in \text{ConfidenceSet} \wedge \mu \in \text{ConfidenceSet}'\}$$

where:

Model-based Reinforcement Learning

- Set of probable MDPs:

$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) : p \in \text{ConfidenceSet} \\ \wedge \mu \in \text{ConfidenceSet}'\}$$

where:

$$\text{ConfidenceSet} = \{p' : \| \quad - \quad \|_1 \leq \quad \}$$

Model-based Reinforcement Learning

- Set of probable MDPs:

$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) : p \in \text{ConfidenceSet} \\ \wedge \mu \in \text{ConfidenceSet}'\}$$

where:

$$\text{ConfidenceSet} = \{p' : \| \qquad\qquad - p'(\cdot|s,a)\|_1 \leq \qquad\qquad \}$$

Model-based Reinforcement Learning

- Set of probable MDPs:

$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) : p \in \text{ConfidenceSet} \\ \wedge \mu \in \text{ConfidenceSet}'\}$$

where:

$$\text{ConfidenceSet} = \left\{p' : \|\widehat{p}_t(\cdot|s,a) - p'(\cdot|s,a)\|_1 \leq \right\}$$

$\widehat{p}_t(s'|s,a) =$ Empirical transition probability given the observations at time t.

Model-based Reinforcement Learning

- Set of probable MDPs:

$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) : p \in \text{ConfidenceSet} \\ \wedge \mu \in \text{ConfidenceSet}'\}$$

where:

$$\text{ConfidenceSet} = \left\{p' : \|\widehat{p}_t(\cdot|s,a) - p'(\cdot|s,a)\|_1 \leq \beta \qquad \left(\tfrac{\delta}{SA}\right), \qquad \right\}$$

$\widehat{p}_t(s'|s,a) =$ Empirical transition probability given the observations at time t.

Model-based Reinforcement Learning

- Set of probable MDPs:

$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) : p \in \text{ConfidenceSet} \wedge \mu \in \text{ConfidenceSet}'\}$$

where:

$$\text{ConfidenceSet} = \left\{p' : \|\widehat{p}_t(\cdot|s,a) - p'(\cdot|s,a)\|_1 \leq \beta_{N_t(s,a)}\left(\tfrac{\delta}{SA}\right), \qquad \right\}$$

$\widehat{p}_t(s'|s,a) =$ Empirical transition probability given the observations at time t.

Model-based Reinforcement Learning

- Set of probable MDPs:

$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) : p \in \text{ConfidenceSet} \\ \wedge \mu \in \text{ConfidenceSet}'\}$$

where:

$$\text{ConfidenceSet} = \left\{p' : \|\widehat{p}_t(\cdot|s,a) - p'(\cdot|s,a)\|_1 \leq \beta_{N_t(s,a)}\left(\tfrac{\delta}{SA}\right), \ \forall s, a\right\}$$

$\widehat{p}_t(s'|s,a) =$ Empirical transition probability given the observations at time t.

Model-based Reinforcement Learning

- Set of probable MDPs:

$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) : p \in \text{ConfidenceSet} \\ \wedge \mu \in \text{ConfidenceSet}'\}$$

where:

$$\text{ConfidenceSet} = \left\{p' : \|\widehat{p}_t(\cdot|s,a) - p'(\cdot|s,a)\|_1 \leq \beta_{N_t(s,a)}\left(\tfrac{\delta}{SA}\right), \ \forall s, a\right\}$$

$\widehat{p}_t(s'|s,a) =$ Empirical transition probability given the observations at time t.

Model-based Reinforcement Learning

- Set of probable MDPs:

$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) : p \in \text{ConfidenceSet} \\ \wedge \mu \in \text{ConfidenceSet}'\}$$

where:

$$\text{ConfidenceSet} = \left\{p' : \|\widehat{p}_t(\cdot|s,a) - p'(\cdot|s,a)\|_1 \leq \beta_{N_t(s,a)}\left(\tfrac{\delta}{SA}\right), \forall s, a\right\}$$
$$\text{ConfidenceSet}' = \left\{\mu' : |\widehat{\mu}_t(s,a) - \mu'(s,a)| \leq \beta'\left(\tfrac{\delta}{SA}\right), \forall s, a\right\}$$

$\widehat{p}_t(s'|s,a) =$ Empirical transition probability given the observations at time t.

Model-based Reinforcement Learning

- Set of probable MDPs:

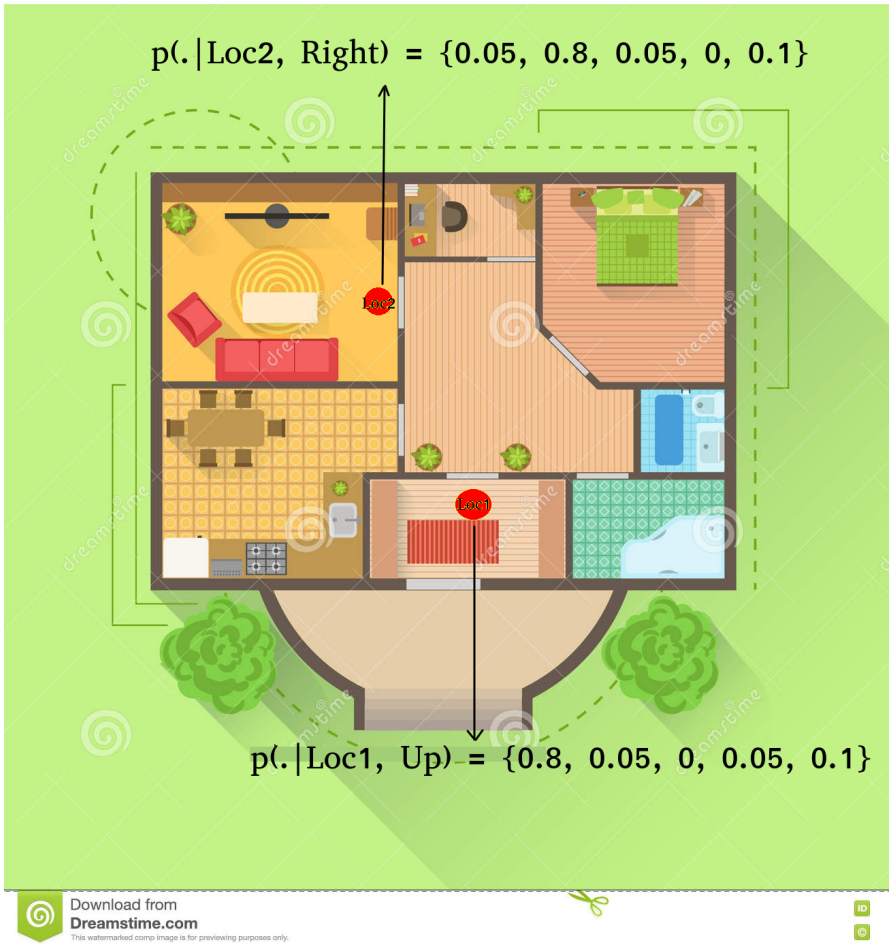
$$\mathcal{M} \in \{(\mathcal{S}, \mathcal{A}, p, \nu) : p \in \text{ConfidenceSet} \\ \wedge \mu \in \text{ConfidenceSet}'\}$$

where:

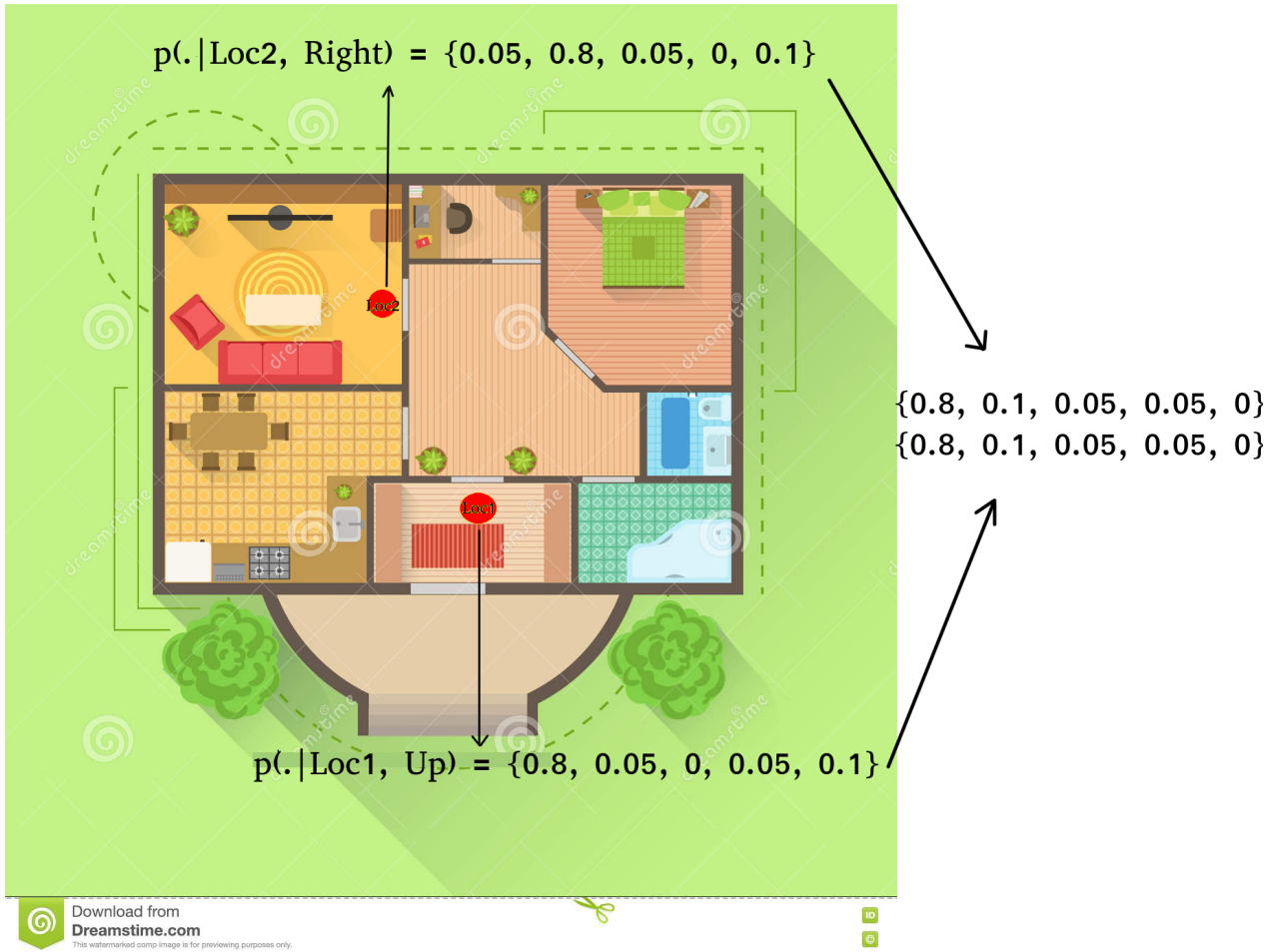
$$\text{ConfidenceSet} = \left\{p' : \|\widehat{p}_t(\cdot|s,a) - p'(\cdot|s,a)\|_1 \leq \beta_{N_t(s,a)}\left(\tfrac{\delta}{SA}\right), \ \forall s, a\right\}$$

$\widehat{p}_t(s'|s,a) =$ Empirical transition probability given the observations at time t.

Equivalence Structure or Class



Equivalence Structure or Class



Equivalence Structure or Class

- Profile Mapping $\sigma_{s,a}$:
  Introduces a permutation of states such that:

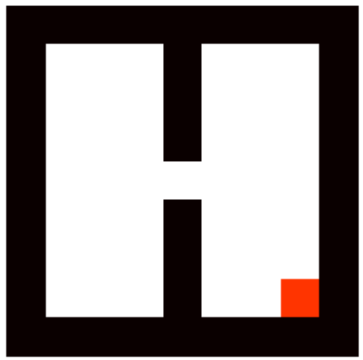
$$p(\sigma_{s,a}(1)|s,a) \geq p(\sigma_{s,a}(2)|s,a) \geq \cdots \geq p(\sigma_{s,a}(S)|s,a).$$

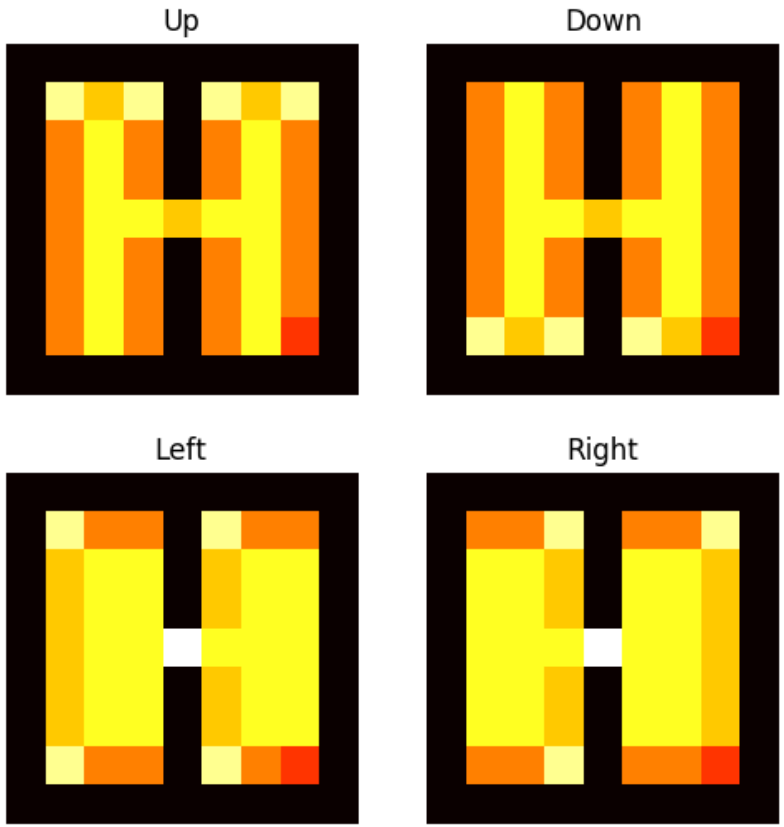
Definition (Equivalent state-action pairs or class)

The pair (s', a') is said to be equivalent to the pair (s, a), if

$$p(\sigma_{s,a}(\cdot)|s,a) = p(\sigma_{s',a'}(\cdot)|s',a') \quad \text{and} \quad \mu(s,a) = \mu(s',a').$$

Equivalence Structure or Class





$SA = 81 \times 4 = 324$

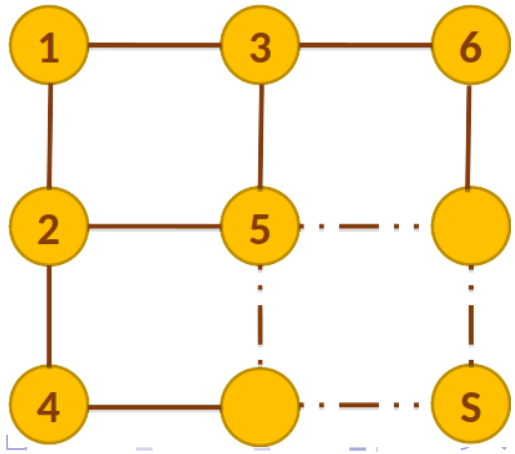
$C = 5$ shown in colors

State of the Art

- ASAP (Anand, Grover, and Singla, 2015)
  - state-action pair aggregation
  - no ordering
  - no regret analysis
- UCAgg (Ortner, 2013)
  - state aggregation(no pairs)
  - no ordering

Two state action pairs (s, a) and (s', a) are in same class under conditions that:

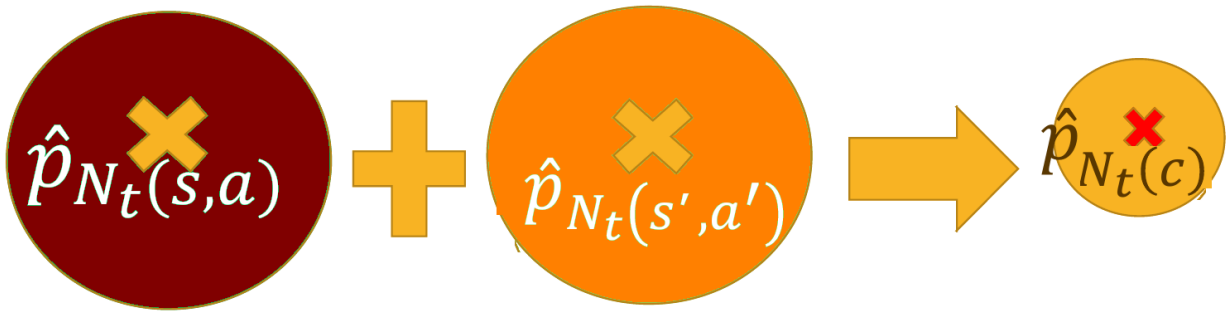
1. The **sum of the transition probabilities**, to each abstract state that these state action pairs transition to, match.
2. The **reward** of applying corresponding state action pairs is identical.



Do we know $\mathcal{C}$ and σ?

1. Known $\mathcal{C}$ and σ:
   Straightforward Extension,
2. Unknown $\mathcal{C}$ and σ:
   Approximate $\mathcal{C}_t$.

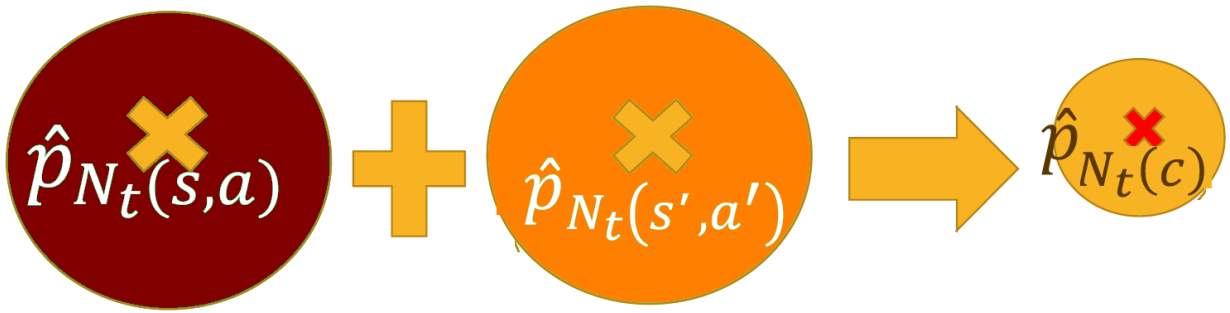
Known $\mathcal{C}, \sigma$



Confidence bound is proportional to:

$$\beta_\delta(N(u)) \propto \sqrt{\frac{1}{N(u)}}$$

Known $\mathcal{C}, \sigma$



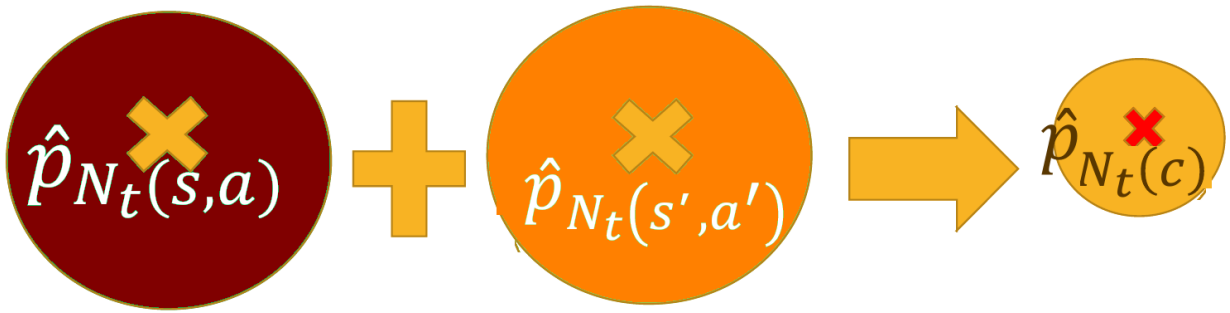
Confidence bound is proportional to:

$$\beta_\delta(N(u)) \propto \sqrt{\frac{1}{N(u)}}$$

$$\text{ConfidenceSet} = \{ \qquad\qquad \leq \qquad\qquad , \}$$

.

Known $\mathcal{C}, \sigma$



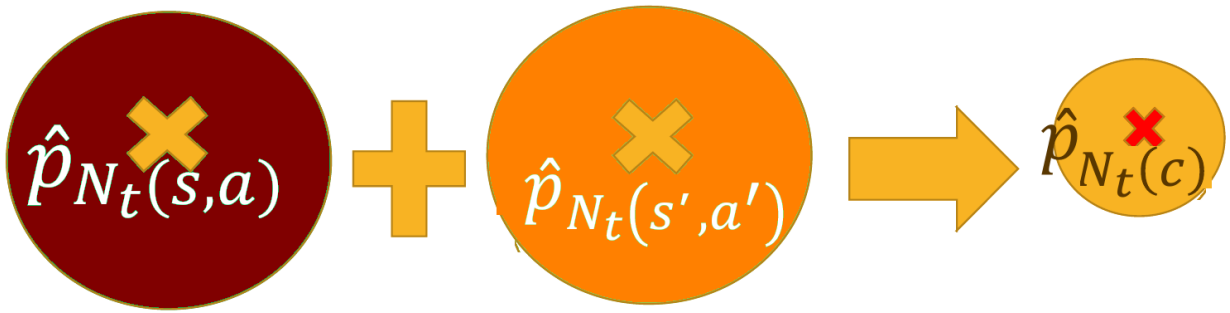
Confidence bound is proportional to:

$$\beta_\delta(N(u)) \propto \sqrt{\frac{1}{N(u)}}$$

$$\text{ConfidenceSet} = \{p' : \|\widehat{p}_t(\cdot|c) - p'(\cdot|s,a)\|_1 \leq \quad , \}$$

.

Known $\mathcal{C}, \sigma$



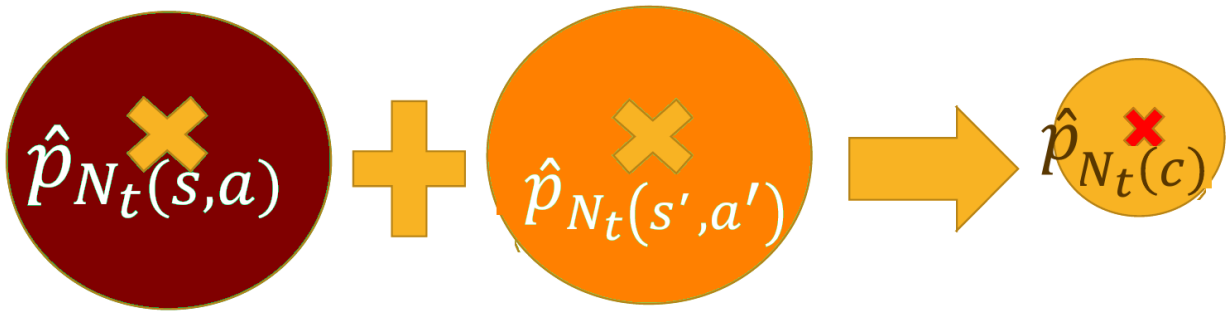
Confidence bound is proportional to:

$$\beta_\delta(N(u)) \propto \sqrt{\frac{1}{N(u)}}$$

$$\text{ConfidenceSet} = \left\{p' : \|\widehat{p}_t(\cdot|c) - p'(\cdot|s,a)\|_1 \leq \beta_{N_t(c)}\left(\frac{\delta}{C}\right), \right\}$$

.

Known $\mathcal{C}, \sigma$



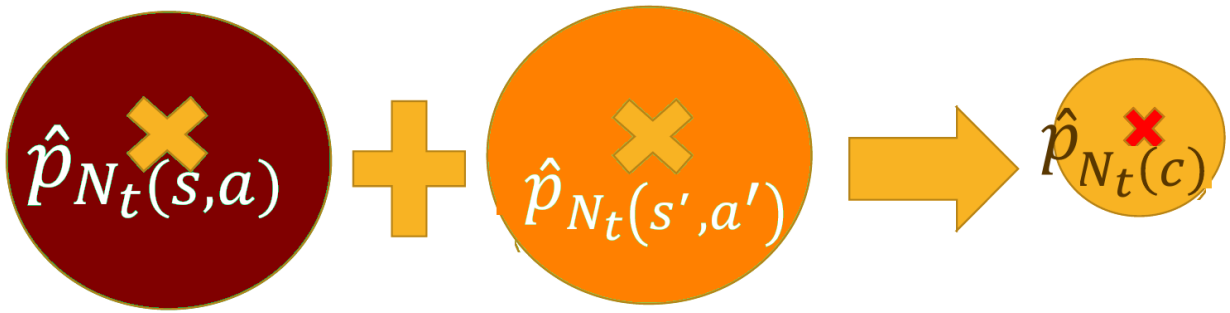
Confidence bound is proportional to:

$$\beta_\delta(N(u)) \propto \sqrt{\frac{1}{N(u)}}$$

$$\text{ConfidenceSet} = \left\{p' : \|\widehat{p}_t(\cdot|c) - p'(\cdot|s,a)\|_1 \leq \beta_{N_t(c)}\left(\frac{\delta}{C}\right), \right\}$$
$$\forall c \in \mathcal{C},$$

.

Known $\mathcal{C}, \sigma$



Confidence bound is proportional to:

$$\beta_\delta(N(u)) \propto \sqrt{\frac{1}{N(u)}}$$

$$\text{ConfidenceSet} = \left\{p' : \|\widehat{p}_t(\cdot|c) - p'(\cdot|s,a)\|_1 \leq \beta_{N_t(c)}\left(\frac{\delta}{C}\right), \right\}$$
$$\forall c \in \mathcal{C}, \forall (s,a) \in c.$$

Unknown $\mathcal{C}, \sigma$

- Unknown Profile?
- Unknown Class?

Unknown $\mathcal{C}, \sigma$

- Unknown Profile?
  Empirical Profile $\boldsymbol{\sigma}_t$
- Unknown Class?
  Approximate $\mathcal{C}_t$

- Empirical Profile Mapping $\sigma_{s,a,t}$:
  Introduces a permutation of states such that:

$$\widehat{p}_t(\sigma_{s,a,t}(1)|s,a) \geq \widehat{p}_t(\sigma_{s,a,t}(2)|s,a) \geq \cdots \geq \widehat{p}_t(\sigma_{s,a,t}(S)|s,a).$$

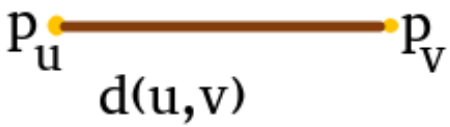
Unknown $\mathcal{C}, \sigma$

Center: set of state-action pairs

Unknown $\mathcal{C}, \sigma$

Center: set of state-action pairs
Distance of centers $u, v \subseteq \mathcal{S} \times \mathcal{A}$:

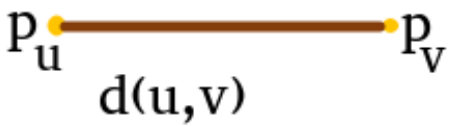
$$d(u, v) = \|p\big(\boldsymbol{\sigma}_u(\cdot)|u\big) - p\big(\boldsymbol{\sigma}_v(\cdot)|v\big)\|_1 \,,$$



Unknown $\mathcal{C}, \sigma$

Center: set of state-action pairs
Distance of centers $u, v \subseteq \mathcal{S} \times \mathcal{A}$:

$$d(u, v) = \|p\big(\boldsymbol{\sigma}_u(\cdot)|u\big) - p\big(\boldsymbol{\sigma}_v(\cdot)|v\big)\|_1 \,,$$

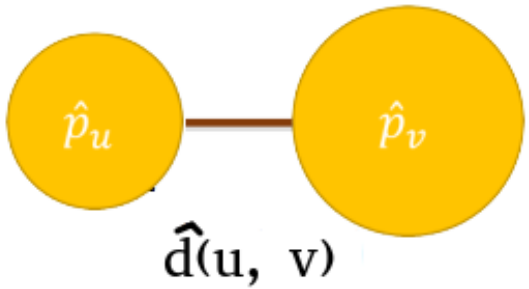


- But d is unknown!

Unknown $\mathcal{C}, \sigma$

Optimistic distance of centers $u, v \subseteq \mathcal{S} \times \mathcal{A}$:

$$\widehat{d}(u,v) = \left\| \widehat{p}_t\left(\boldsymbol{\sigma}_{u,t}(\cdot)|u\right) - \widehat{p}_t\left(\boldsymbol{\sigma}_{v,t}(\cdot)|v\right) \right\|_1 - \epsilon_{u,t} - \epsilon_{v,t} \,,$$



Unknown $\mathcal{C}, \sigma$

Definition (PAC Neighbor)

For a given equivalence structure $\mathcal{C}'$, and given $u \in \mathcal{C}'$, we say that $v \in \mathcal{C}'$ is a *PAC Neighbor* of u if it satisfies:

$$\widehat{d}(u,v) = \left\| \widehat{p}_t\big(\boldsymbol{\sigma}_{u,t}(\cdot)|u\big) - \widehat{p}_t\big(\boldsymbol{\sigma}_{v,t}(\cdot)|v\big) \right\|_1 - \epsilon_{u,t} - \epsilon_{v,t} \leq 0.$$

We further define $\mathcal{N}(u) = \big\{v \in \mathcal{C}' \setminus \{u\} \colon \widehat{d}(u,v) \leq 0\big\}$ as the set of all PAC Neighbors of u.

Unknown $\mathcal{C}, \sigma$

Definition (PAC Neighbor)

For a given equivalence structure $\mathcal{C}'$, and given $u \in \mathcal{C}'$, we say that $v \in \mathcal{C}'$ is a *PAC Neighbor* of u if it satisfies:

$$\widehat{d}(u,v) = \left\| \widehat{p}_t\big(\boldsymbol{\sigma}_{u,t}(\cdot)|u\big) - \widehat{p}_t\big(\boldsymbol{\sigma}_{v,t}(\cdot)|v\big) \right\|_1 - \epsilon_{u,t} - \epsilon_{v,t} \leq 0.$$

We further define $\mathcal{N}(u) = \big\{v \in \mathcal{C}' \setminus \{u\} \colon \widehat{d}(u,v) \leq 0\big\}$ as the set of all PAC Neighbors of u.

Definition (PAC Nearest Neighbor)

For a given equivalence structure $\mathcal{C}'$ and $u \in \mathcal{C}'$, we define the *PAC Nearest Neighbor* of u (when it exists) as:

$$\mathit{Near}(u, \mathcal{C}') \in \operatorname*{argmin}_{z \in \mathcal{N}(u)} \widehat{d}(u,z)\,.$$

Algorithm ApproxEquivalence

Input N

Output $\mathcal{C}^k$

Algorithm ApproxEquivalence

Input N

Output $\mathcal{C}^k$

1: **Initialization:** $\mathcal{C}^0 \leftarrow \{\{1\}, \{2\}, \ldots, \{SA\}\}$, $k = 0$;

Algorithm ApproxEquivalence

Input N

Output $\mathcal{C}^k$

1: **Initialization:** $\mathcal{C}^0 \leftarrow \{\{1\}, \{2\}, \ldots, \{SA\}\}$, $k = 0$;

2: **repeat**

12: **until** $\mathcal{C}^k = \mathcal{C}^{k-1}$

Algorithm ApproxEquivalence

Input N

Output $\mathcal{C}^k$

1: **Initialization:** $\mathcal{C}^0 \leftarrow \{\{1\}, \{2\}, \ldots, \{SA\}\}$, $k = 0$;
2: **repeat**
3: $\quad k \leftarrow k + 1$
4: $\quad \mathcal{C}^k \leftarrow \mathcal{C}^{k-1}$

12: **until** $\mathcal{C}^k = \mathcal{C}^{k-1}$

Algorithm ApproxEquivalence

Input N

Output $\mathcal{C}^k$

1: **Initialization:** $\mathcal{C}^0 \leftarrow \{\{1\}, \{2\}, \ldots, \{SA\}\}$, $k = 0$;
2: **repeat**
3: $k \leftarrow k + 1$
4: $\mathcal{C}^k \leftarrow \mathcal{C}^{k-1}$
5: **for all** $u \in \mathcal{C}^{k-1}$ in a non decreasing order of their $N(u)$ **do**

11: **end for**
12: **until** $\mathcal{C}^k = \mathcal{C}^{k-1}$

Algorithm ApproxEquivalence

Input N

Output $\mathcal{C}^k$

Initialization: $\mathcal{C}^0 \leftarrow \{\{1\}, \{2\}, \ldots, \{SA\}\}$, $k = 0$;
repeat
  $k \leftarrow k + 1$
  $\mathcal{C}^k \leftarrow \mathcal{C}^{k-1}$
  for all $u \in \mathcal{C}^{k-1}$ in a non decreasing order of their $N(u)$ **do**
    $v \leftarrow \text{Near}(u, \mathcal{C}^{k-1})$

  end for
until $\mathcal{C}^k = \mathcal{C}^{k-1}$

Algorithm ApproxEquivalence

Input N
Output $\mathcal{C}^k$

Initialization: $\mathcal{C}^0 \leftarrow \{\{1\}, \{2\}, \ldots, \{SA\}\}$, $k = 0$;
repeat
  $k \leftarrow k + 1$
  $\mathcal{C}^k \leftarrow \mathcal{C}^{k-1}$
  for all $u \in \mathcal{C}^{k-1}$ in a non decreasing order of their $N(u)$ **do**
    $v \leftarrow \text{Near}(u, \mathcal{C}^{k-1})$
    if $v \neq \emptyset$ **then**

    end if
  end for
until $\mathcal{C}^k = \mathcal{C}^{k-1}$

Algorithm ApproxEquivalence

Input N
Output $\mathcal{C}^k$

Initialization: $\mathcal{C}^0 \leftarrow \{\{1\}, \{2\}, \ldots, \{SA\}\}$, $k = 0$;
repeat
  $k \leftarrow k + 1$
  $\mathcal{C}^k \leftarrow \mathcal{C}^{k-1}$
  for all $u \in \mathcal{C}^{k-1}$ in a non decreasing order of their $N(u)$ **do**
    $v \leftarrow \text{Near}(u, \mathcal{C}^{k-1})$
    if $v \neq \emptyset$ **then**
      Merge u, v
      Remove u, v from $\mathcal{C}^k$
    end if
  end for
until $\mathcal{C}^k = \mathcal{C}^{k-1}$

Algorithm ApproxEquivalence

Input N
Output $\mathcal{C}^k$

Initialization: $\mathcal{C}^0 \leftarrow \{\{1\}, \{2\}, \ldots, \{SA\}\}$, $k = 0$;
repeat
  $k \leftarrow k + 1$
  $\mathcal{C}^k \leftarrow \mathcal{C}^{k-1}$
  for all $u \in \mathcal{C}^{k-1}$ in a non decreasing order of their $N(u)$ **do**
    $v \leftarrow \text{Near}(u, \mathcal{C}^{k-1})$
    if $v \neq \emptyset$ **then**
      Merge u, v
      Remove u, v from $\mathcal{C}^k$
    end if
  end for
until $\mathcal{C}^k = \mathcal{C}^{k-1}$

Remark

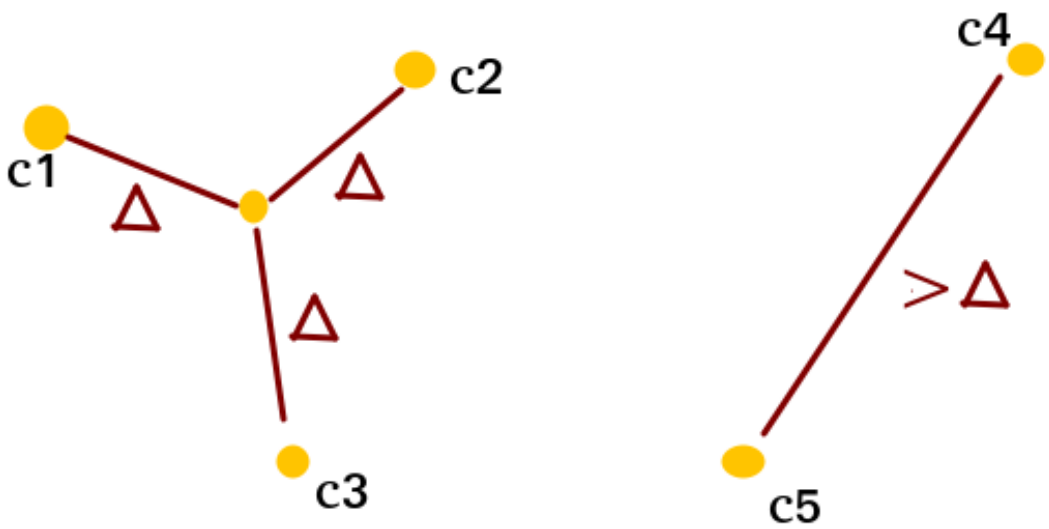
The algorithm ApproxEquivalence stops after, at most, $SA - 1$ steps.

Unknown $\mathcal{C}, \sigma$

Assumption (Separability)

There exists some $\Delta > 0$ *such that*

$$\forall c \neq c' \in \mathcal{C}, \forall \ell \in c, \forall \ell' \in c', \quad d(\{\ell\}, \{\ell'\}) \geq \Delta \,.$$



Unknown $\mathcal{C}, \sigma$

Theorem (Asadi et al., 2019)

Let N be such that $\Delta > 4\beta_N(\delta/SA)$, provided that $\min_{s,a} N_t(s,a) > N$, under Separability Assumption, ApproxEquivalence outputs the correct equivalence structure $\mathcal{C}$ of state-action pairs with probability at least $1-\delta$.

Unknown $\mathcal{C}, \sigma$

Theorem (Asadi et al., 2019)

Let N be such that $\Delta > 4\beta_N(\delta/SA)$, provided that $\min_{s,a} N_t(s,a) > N$, under Separability Assumption, ApproxEquivalence outputs the correct equivalence structure $\mathcal{C}$ of state-action pairs with probability at least $1-\delta$.

Proof.

Unknown $\mathcal{C}, \sigma$

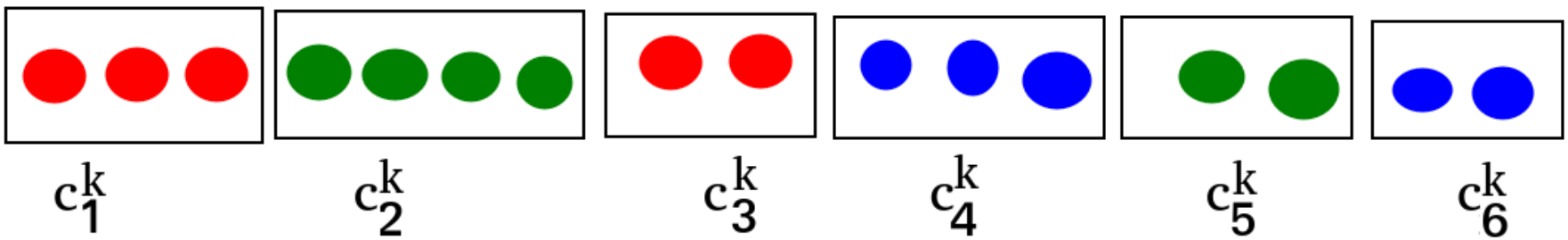
Theorem (Asadi et al., 2019)

Let N be such that $\Delta > 4\beta_N(\delta/SA)$, provided that $\min_{s,a} N_t(s,a) > N$, under Separability Assumption, ApproxEquivalence outputs the correct equivalence structure $\mathcal{C}$ of state-action pairs with probability at least $1-\delta$.

Proof.

1. (Induction): $\forall u \in \mathcal{C}^k$ and $\forall (s_1,a_1),(s_2,a_2) \in u$, $\exists v \in \mathcal{C}$ such that $(s_1,a_1),(s_2,a_2) \in v$: we say that $\mathcal{C}^k$ is a *valid partition*.

□



Unknown $\mathcal{C}, \sigma$

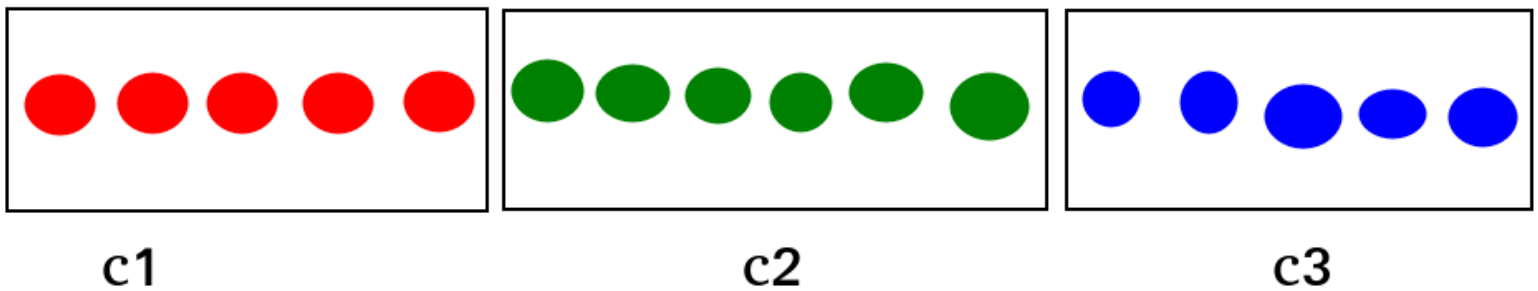
Theorem (Asadi et al., 2019)

Let N be such that $\Delta > 4\beta_N(\delta/SA)$, provided that $\min_{s,a} N_t(s,a) > N$, under Separability Assumption, ApproxEquivalence outputs the correct equivalence structure $\mathcal{C}$ of state-action pairs with probability at least $1-\delta$.

Proof.

1. (Induction): $\forall u \in \mathcal{C}^k$ and $\forall (s_1,a_1),(s_2,a_2) \in u$, $\exists v \in \mathcal{C}$ such that $(s_1,a_1),(s_2,a_2) \in v$: we say that $\mathcal{C}^k$ is a *valid partition*.
2. When the algorithm stops, no two states in the same class of $\mathcal{C}$ are that have not been merged.

□

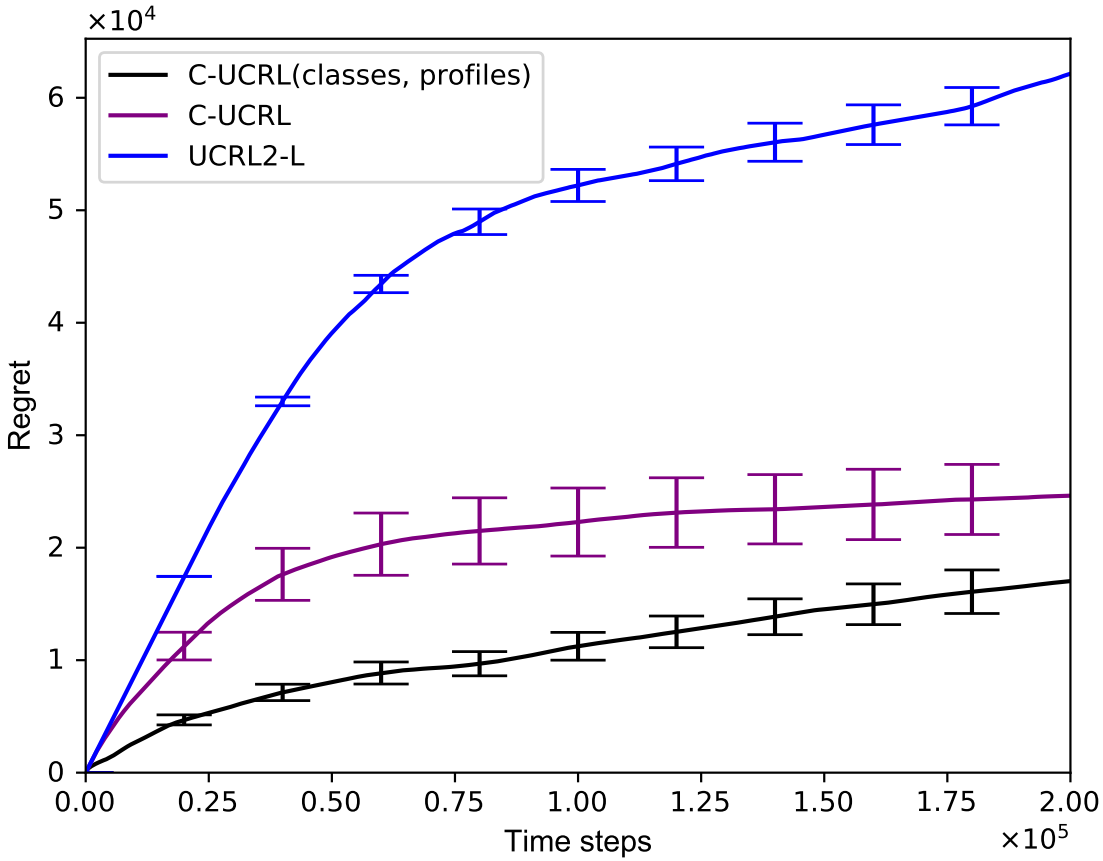


Application: UCRL2 (Jaksch, Ortner, and Auer, 2010)

- Confidence-based Model-based RL approach
- Known $\mathcal{C}, \boldsymbol{\sigma}$:
  - C-UCRL$(\mathcal{C}, \sigma)$: Improved the regret by $\sqrt{\frac{SA}{C}}$
- Unknown $\mathcal{C}, \boldsymbol{\sigma}$:
  - C-UCRL: No theoretical guarantees but significant improvements on ergodic MDPs in practice.

25-state Ergodic RiverSwim

Regret analysis

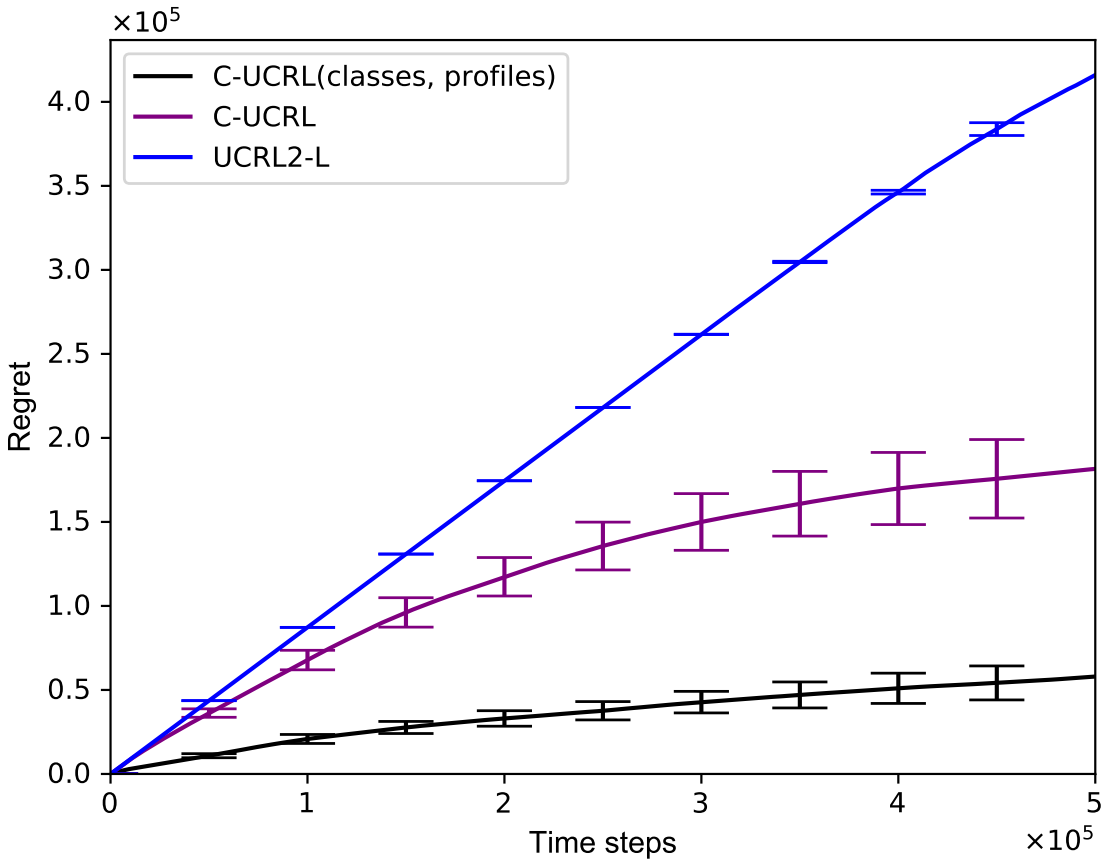


runs $= 100$, $\delta = 0.05$, $SA = 25 \times 2 = 50$, $C = 6$

$\sqrt{\frac{SA}{C}} \approx 2.9$

50-state Ergodic RiverSwim

Regret analysis



$SA = 50 \times 2 = 100,\ C = 6$

Contributions

- Introducing a class structure for state-action pairs
- For known $\mathcal{C}, \sigma$ setting:
  - Tighter confidence sets
  - Fewer parameters to estimate
- Unknown $\mathcal{C}, \sigma$ setting:
  - Approximating Equivalent class structure $\mathcal{C}$ using ApproxEquivalence
- Extension on UCRL2:
  - C-UCRL$(\mathcal{C}, \sigma)$: Improved the regret of UCRL2 by factor of $\sqrt{\frac{SA}{C}}$.
  - C-UCRL: No theoretical guarantees but significant improvements on ergodic MDPs in practice

Exploiting Structure in Online Personalized Mean Estimation Network

Decentralized Learning

- Existence of a network of agents
- Learning collaboratively
- No centralized server
- Data is kept locally
- Aggregated quantities are shared

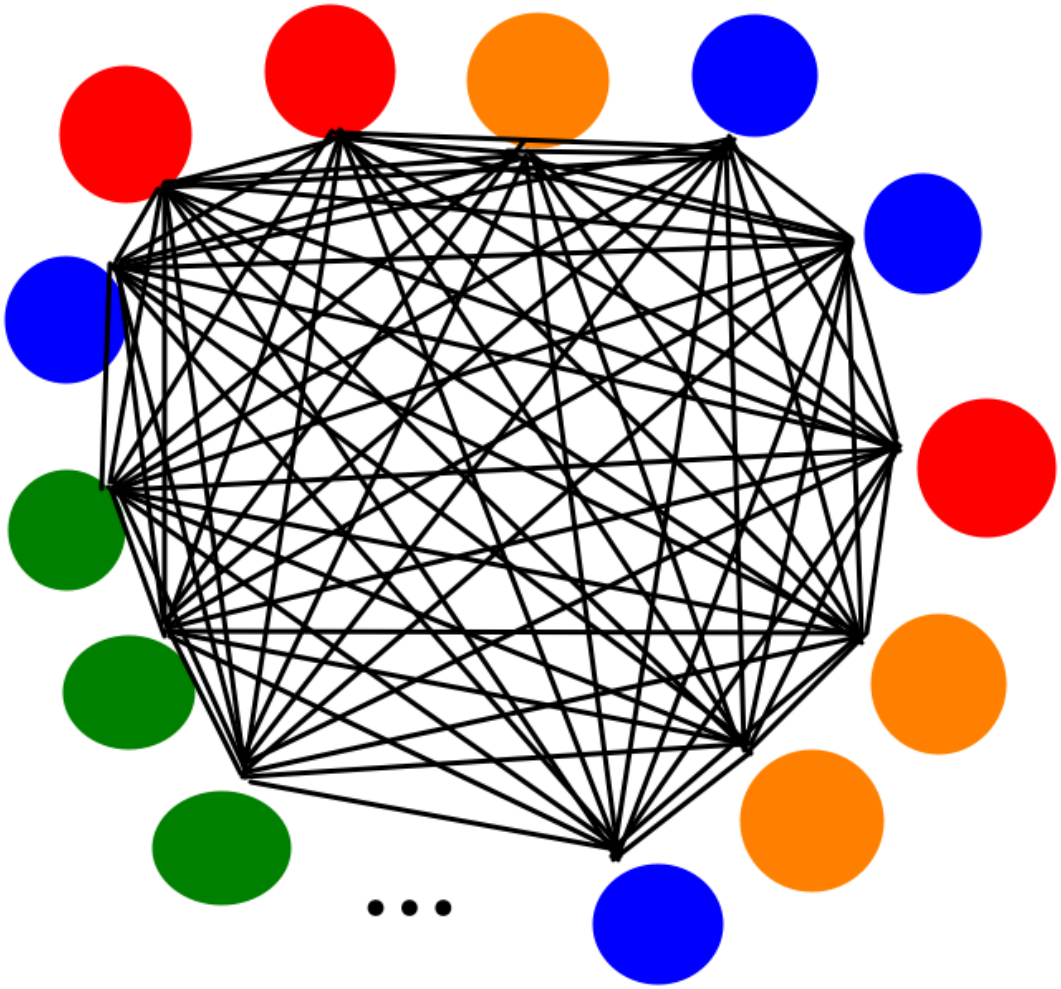
Decentralized Learning

- Existence of a network of agents
- **Learning collaboratively**
- No centralized server
- Data is kept locally
- Aggregated quantities are shared

Decentralized Learning

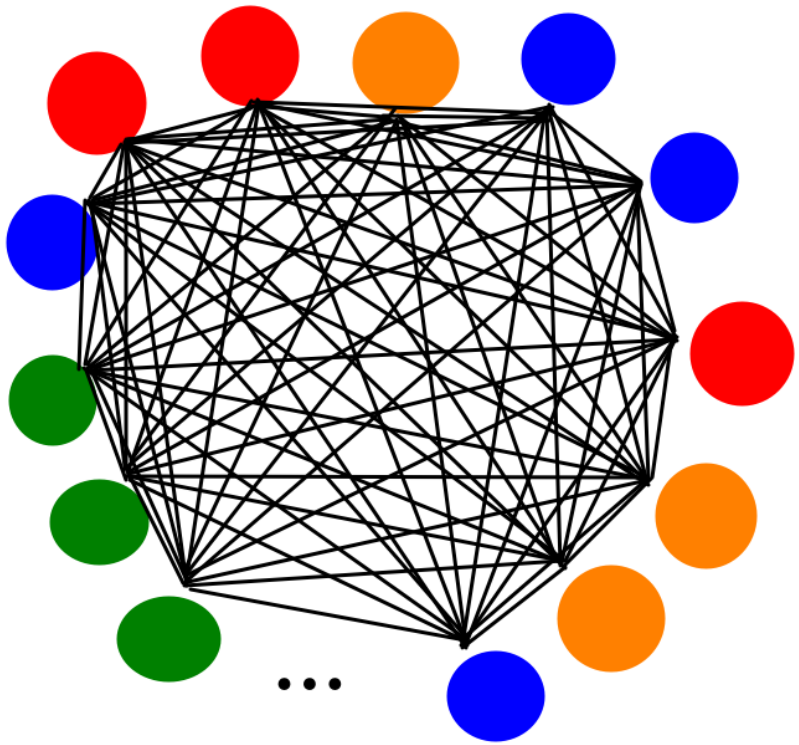
- Existence of a network of agents
- **Learning collaboratively**
  - working in a group of two or more
  - achieve a common goal
- No centralized server
- Data is kept locally
- Aggregated quantities are shared

Online Personalized Mean Estimation Network



Agents = {1, 2, ... , A}

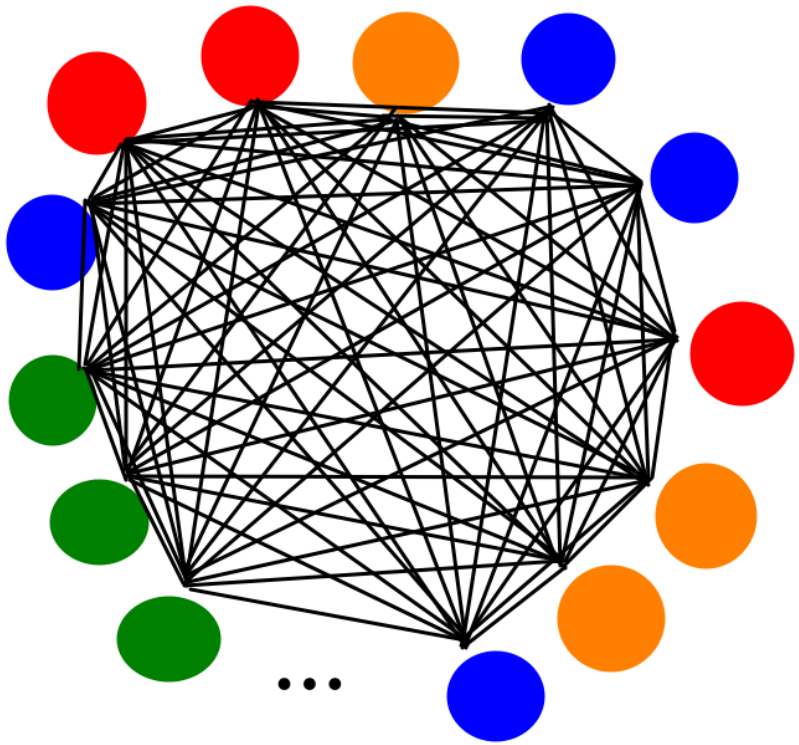
Online Personalized Mean Estimation Network



At each time step t:

- Receives new **sample** $x_a^t \sim \nu_a$
- Chooses agent l to **query**
- Keeps a memory of local estimates $[\bar{x}_{a,1}^t, \bar{x}_{a,2}^t, \ldots, \bar{x}_{a,A}^t], [n_{a,1}^t, n_{a,2}^t, \ldots, n_{a,A}^t]$

Online Personalized Mean Estimation Network



- From the perspective of agent a:
  - Goal: Find a good mean estimate μ_a^t of personal distribution ν_a with mean μ_a

Related Work

- Multiplayer Multi-armed bandit with a single global problem
- Personalization
  1. Personalized Federated Multi-armed Bandit (Shi, Shen, and Yang, 2021)
     - combination of local and global objectives
     - regret minimization
  2. Weighted Collaborative Best Arm Identification Bandit (Réda, Vakili, and Kaufmann, 2022)
     - introduces a weighting (known)
     - best arm identification
     - But we estimate the weight (unknown) and to enable collaboration we need to discover structure

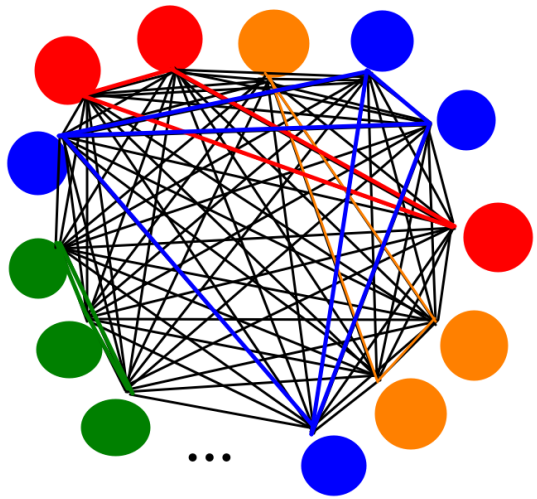
Equivalence Structure or Class

Definition (Similarity class)

The *similarity class* of agent a is given by:

$$\mathcal{C}_a = \{l \in [A] : \Delta_{a,l} = 0\},$$

where $\Delta_{a,l} = |\mu_a - \mu_l|$ is the gap between the means of agent a and agent l.



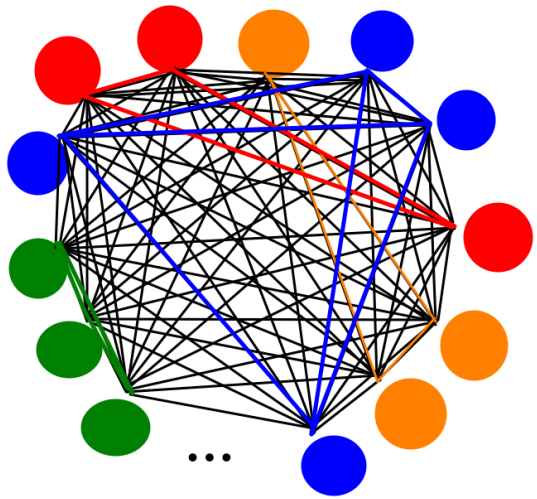
Equivalence Structure or Class

Definition (Similarity class)

The *similarity class* of agent a is given by:

$$\mathcal{C}_a = \{l \in [A] : \Delta_{a,l} = 0\},$$

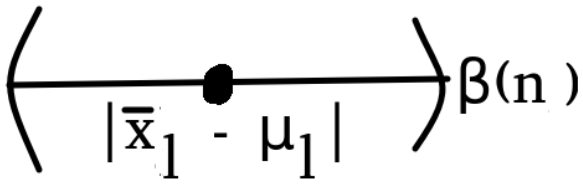
where $\Delta_{a,l} = |\mu_a - \mu_l|$ is the gap between the means of agent a and agent l.



But $|\mu_a - \mu_l| = \Delta_{a,l}$ is unknown!

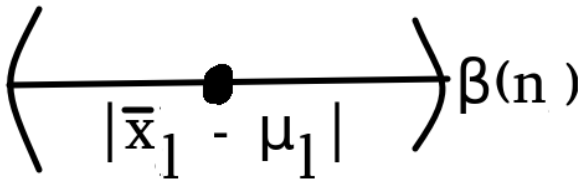
Equivalence Structure or Class

But $|\mu_a - \mu_l| = \Delta_{a,l}$ is unknown!



Equivalence Structure or Class

But $|\mu_a - \mu_l| = \Delta_{a,l}$ is unknown!



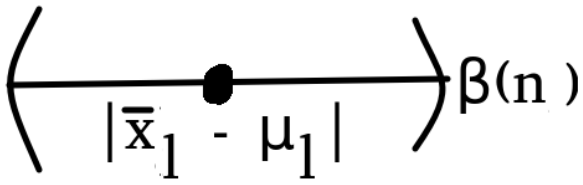
Definition (Optimistic distance)

The *optimistic distance* with agent l from the perspective of agent a is defined as:

$$d_a^t(l) = |\bar{x}_{a,a}^t - \bar{x}_{a,l}^t| - \beta(n_{a,a}^t) - \beta(n_{a,l}^t).$$

Equivalence Structure or Class

But $|\mu_a - \mu_l| = \Delta_{a,l}$ is unknown!



Definition (Optimistic distance)

The *optimistic distance* with agent l from the perspective of agent a is defined as:

$$d_a^t(l) = |\bar{x}_{a,a}^t - \bar{x}_{a,l}^t| - \beta(n_{a,a}^t) - \beta(n_{a,l}^t).$$

Definition (optimistic similarity class)

The *optimistic similarity class* from the perspective of agent a at time t is defined as:

$$\mathcal{C}_a^t = \{l \in [A] : d_a^t(l) \leq 0\}.$$

Algorithm CoIME

At each time step:

- Perceive
- Query
- Estimate

Algorithm CoIME

At each time step:

- **Perceive**
- Query
- Estimate

Algorithm CoIME

At each time step:

- **Perceive**
  1. receives a sample $x_a^t \sim \nu_a$
  2. updates the local average $\bar{x}_{a,a}^t$
- Query
- Estimate

Algorithm CoIME

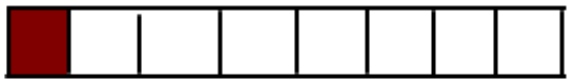
At each time step:

- Perceive
- **Query**
- Estimate

Algorithm CoIME

At each time step:

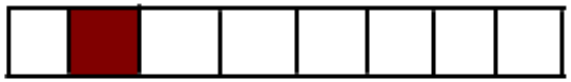
- Perceive
- **Query**
  1. Round-Robin
- Estimate



Algorithm CoIME

At each time step:

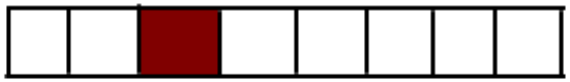
- Perceive
- **Query**
  1. Round-Robin
- Estimate



Algorithm CoIME

At each time step:

- Perceive
- **Query**
  1. Round-Robin
- Estimate



Algorithm CoIME

At each time step:

- Perceive
- **Query**
  1. Round-Robin
- Estimate

Algorithm CoIME

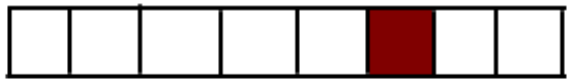
At each time step:

- Perceive
- **Query**
  1. Round-Robin
- Estimate

Algorithm CoIME

At each time step:

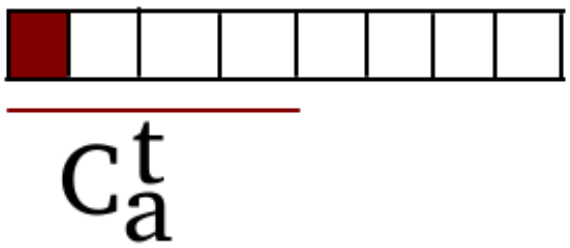
- Perceive
- **Query**
  1. Round-Robin
- Estimate



Algorithm ColME

At each time step:

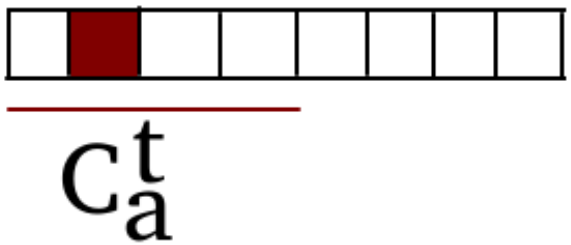
- Perceive
- **Query**
  1. Round-Robin
  2. Restricted-Round-Robin
- Estimate



Algorithm ColME

At each time step:

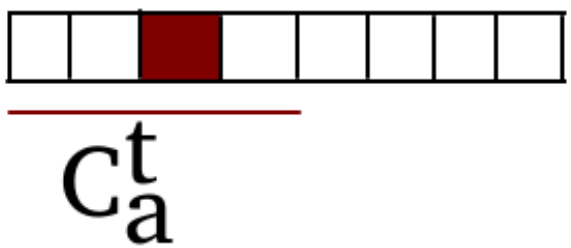
- Perceive
- **Query**
  1. Round-Robin
  2. Restricted-Round-Robin
- Estimate



Algorithm CoIME

At each time step:

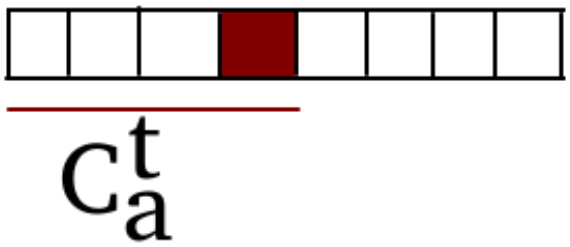
- Perceive
- **Query**
  1. Round-Robin
  2. Restricted-Round-Robin
- Estimate



Algorithm ColME

At each time step:

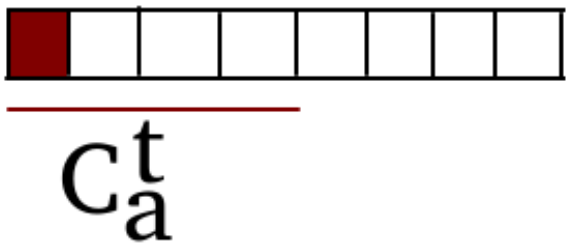
- Perceive
- **Query**
  1. Round-Robin
  2. Restricted-Round-Robin
- Estimate



Algorithm ColME

At each time step:

- Perceive
- **Query**
  1. Round-Robin
  2. Restricted-Round-Robin
- Estimate



Algorithm ColME

At each time step:

- Perceive
- Query
- **Estimate**

$$\mu_a^t \leftarrow \sum_{l \in \mathcal{C}_a^t} \alpha_{a,l}^t \times \bar{x}_{a,l}^t$$

Algorithm ColME

At each time step:

- Perceive
- Query
- **Estimate**
  $\mu_a^t \leftarrow \sum_{l \in \mathcal{C}_a^t} \alpha_{a,l}^t \times \bar{x}_{a,l}^t$
  - Simple-weighting

Algorithm ColME

At each time step:

- Perceive
- Query
- **Estimate**
  $\mu_a^t \leftarrow \sum_{l \in \mathcal{C}_a^t} \alpha_{a,l}^t \times \bar{x}_{a,l}^t$
  - Simple-weighting

$$\alpha_{a,l}^t = \frac{n_{a,l}^t}{\sum_{l \in \mathcal{C}_a^t} n_{a,l}^t}.$$

Algorithm CoIME

At each time step:

- Perceive
- Query
- **Estimate**

  $\mu_a^t \leftarrow \sum_{l \in \mathcal{C}_a^t} \alpha_{a,l}^t \times \bar{x}_{a,l}^t$

  - Simple-weighting
  - Soft-weighting

$$\alpha_{a,l}^t = n_{a,l}^t \frac{|I_{a,a} \cap I_{a,l}|}{|I_{a,a} \cup I_{a,l}|} \times \frac{1}{Z_{\text{soft}}},$$

Algorithm CoIME

At each time step:

- Perceive
- Query
- **Estimate**

  $\mu_a^t \leftarrow \sum_{l \in \mathcal{C}_a^t} \alpha_{a,l}^t \times \bar{x}_{a,l}^t$

  - Simple-weighting
  - Soft-weighting
  - Aggressive-weighting(Extension of Soft-weighting)

$$E_{a,l} = \mathbb{1}_{\{|I_{a,a} \cap I_{a,l}| > \min\{\beta(n_{a,l}^t), \beta(n_{a,a}^t)\}\}}$$

Analysis

Baselines:

- Local estimation
- Oracle weighting

Analysis

Baselines:

- **Local estimation**
  No collaboration
- Oracle weighting

Analysis

Baselines:

- Local estimation
- **Oracle weighting**
  Known true class $\mathcal{C}_a$

Theoretical Results

Theorem (CoIME class estimation time complexity (Asadi et al., 2022))

For any $\delta \in (0, 1)$, employing Restricted-Round-Robin query strategy and simple weighting, we have:

(1)

Theoretical Results

Theorem (CoIME class estimation time complexity (Asadi et al., 2022))

For any $\delta \in (0,1)$, employing Restricted-Round-Robin query strategy and simple weighting, we have:

$$\mathbb{P}\left(\exists t > \zeta_a : \mathcal{C}_a^t \neq \mathcal{C}_a\right) \leq \frac{\delta}{8}, \tag{1}$$

Theoretical Results

Theorem (CoIME class estimation time complexity (Asadi et al., 2022))

For any $\delta \in (0,1)$, employing Restricted-Round-Robin query strategy and simple weighting, we have:

$$\mathbb{P}\left(\exists t > \zeta_a : \mathcal{C}_a^t \neq \mathcal{C}_a\right) \leq \frac{\delta}{8}, \quad \textit{with} \quad \zeta_a = \tag{1}$$

Theoretical Results

Theorem (CoIME class estimation time complexity (Asadi et al., 2022))

For any $\delta \in (0,1)$, employing Restricted-Round-Robin query strategy and simple weighting, we have:

$$\mathbb{P}\left(\exists t > \zeta_a : \mathcal{C}_a^t \neq \mathcal{C}_a\right) \leq \frac{\delta}{8}, \quad \textit{with} \;\; \zeta_a = n^{\star}_{a,a} + A - 1 \tag{1}$$

$n^{\star}_{a,l}$ = number of observations required to distinct class a and l.

Theoretical Results

Theorem (CoIME class estimation time complexity (Asadi et al., 2022))

For any $\delta \in (0,1)$, employing Restricted-Round-Robin query strategy and simple weighting, we have:

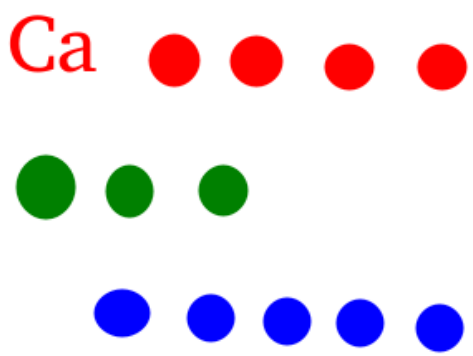
$$\mathbb{P}\left(\exists t > \zeta_a : \mathcal{C}_a^t \neq \mathcal{C}_a\right) \leq \frac{\delta}{8}, \quad \textit{with} \quad \zeta_a = n^\star_{a,a} + A - 1 - \sum_{l \in [A] \setminus \mathcal{C}_a} \mathbb{1}_{\{n^\star_{a,a} > n^\star_{a,l} + A - 1\}}. \tag{1}$$

$n^\star_{a,l}$ = number of observations required to distinct class a and l.

Theoretical Results

Proof.

1. Knowing $\mathcal{C}_a^t = \{l \in [A] : d_a^t(l) \leq 0\}$:

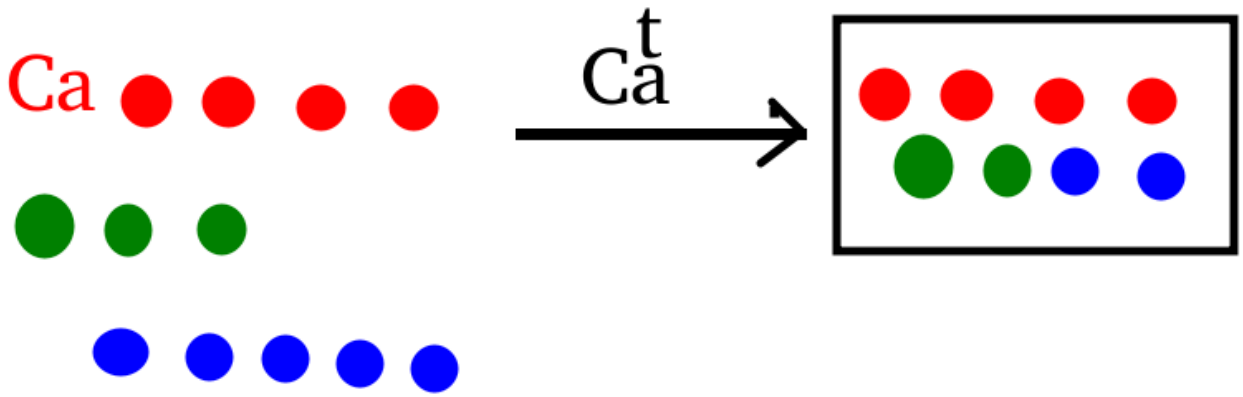


Theoretical Results

Proof.

1. Knowing $\mathcal{C}_a^t = \{l \in [A] : d_a^t(l) \leq 0\}$:
   - $l \in \mathcal{C}_a$

□

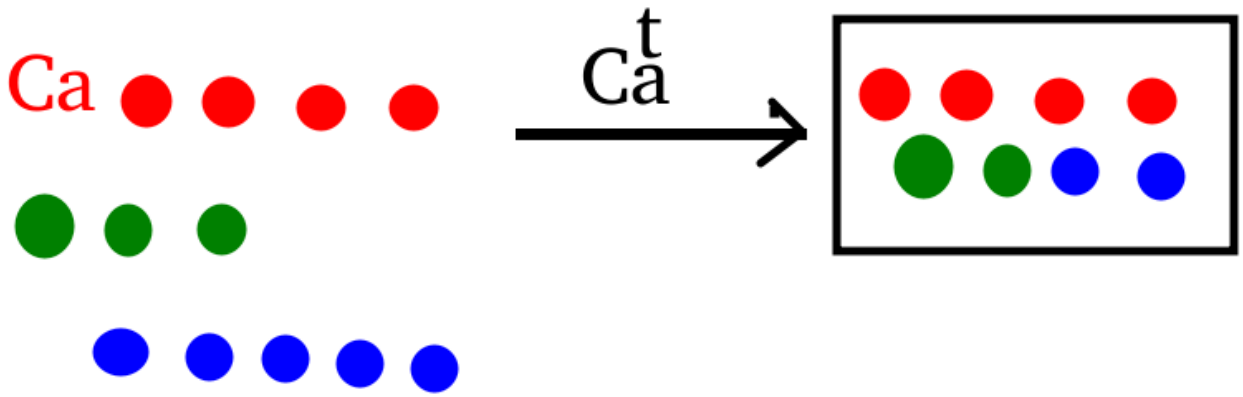


Theoretical Results

Proof.

1. Knowing $\mathcal{C}_a^t = \{l \in [A] : d_a^t(l) \leq 0\}$:
   - $l \in \mathcal{C}_a$
   - $l \in [A] \backslash \mathcal{C}_a$

□



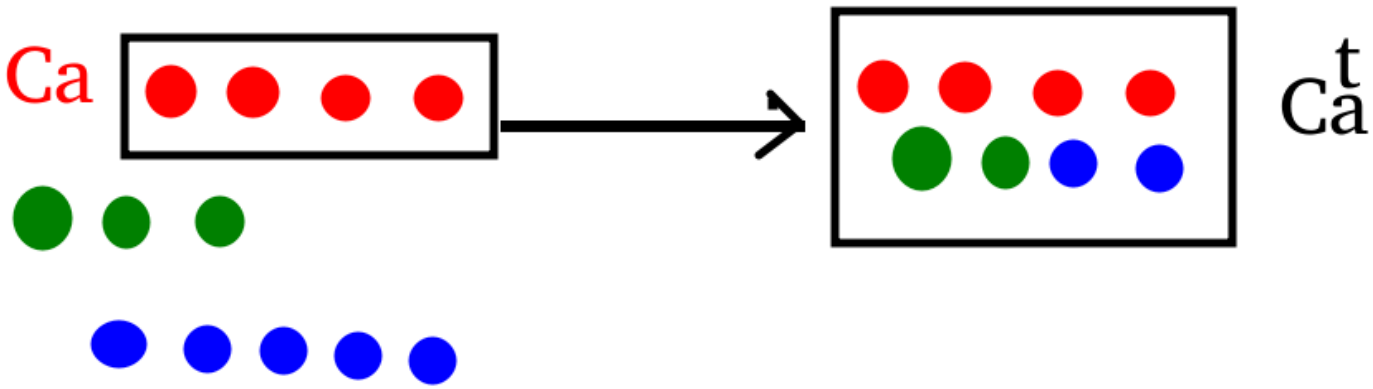
Theoretical Results

Proof.

1. Knowing $\mathcal{C}_a^t = \{l \in [A] : d_a^t(l) \leq 0\}$:

2. at all time, if $l \in \mathcal{C}_a$, then $d_a^t(l) \leq 0$.

□

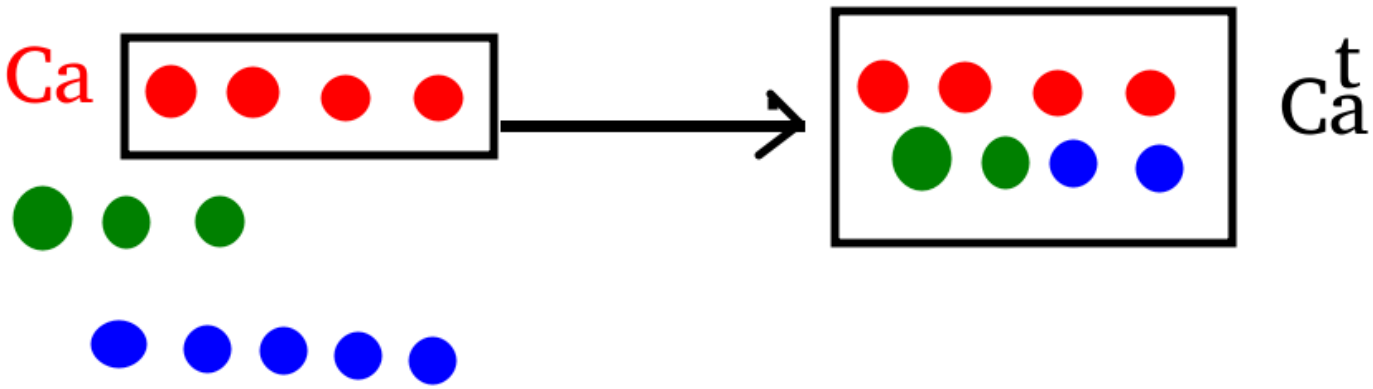


Theoretical Results

Proof.

1. Knowing $\mathcal{C}_a^t = \{l \in [A] : d_a^t(l) \leq 0\}$:
2. at all time, if $l \in \mathcal{C}_a$, then $d_a^t(l) \leq 0$.
3. when agents have enough samples($n_{a,l}^\star$), if $l \in [A] \backslash \mathcal{C}_a \iff d_a^t(l) > 0$

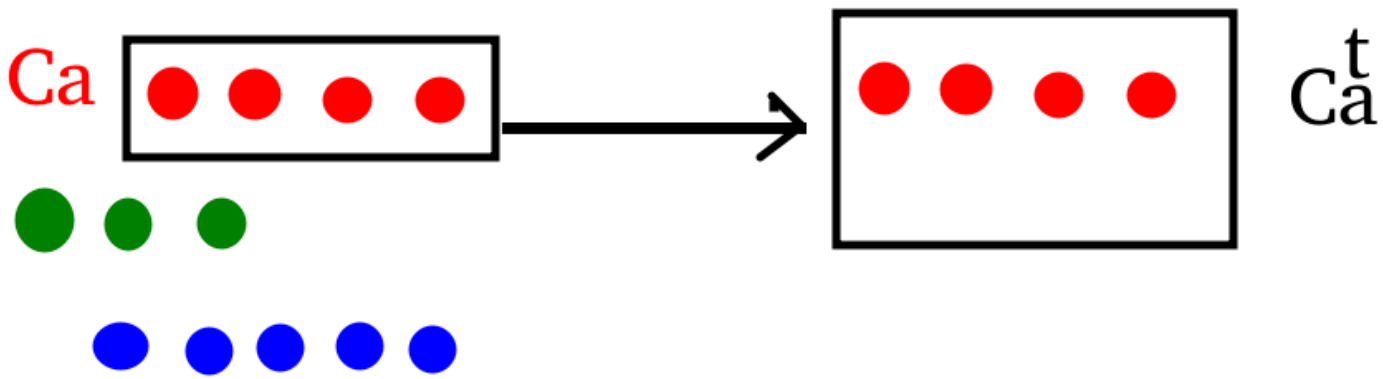
□



Theoretical Results

Proof.

1. Knowing $\mathcal{C}_a^t = \{l \in [A] : d_a^t(l) \leq 0\}$:
2. at all time, if $l \in \mathcal{C}_a$, then $d_a^t(l) \leq 0$.
3. when agents have enough samples($n_{a,l}^\star$), if $l \in [A] \backslash \mathcal{C}_a \iff d_a^t(l) > 0$



Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

(2)

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \tag{2}$$

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \tag{2}$$

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \max \tag{2}$$

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \max\left(\zeta_a, \right. \tag{2}$$

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \max\left(\zeta_a, \frac{\lceil \beta_\delta^{-1}(\epsilon) \rceil}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a| - 1}{2}\right). \tag{2}$$

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \max\left(\zeta_a, \frac{\lceil \beta_\delta^{-1}(\epsilon) \rceil}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a| - 1}{2}\right). \tag{2}$$

- CoIME outperforms Local by a factor of $|\mathcal{C}_a|$.
- CoIME performs as good as Oracle in some regimes.

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \max\left(\zeta_a, \frac{\lceil \beta_\delta^{-1}(\epsilon) \rceil}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a| - 1}{2}\right). \tag{2}$$

- CoIME outperforms LocalI by a factor of $|\mathcal{C}_a|$.
- CoIME performs as good as Oracle in some regimes.

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \max\left(\zeta_a, \frac{\lceil \beta_\delta^{-1}(\epsilon) \rceil}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a| - 1}{2}\right). \tag{2}$$

- To have $|\mu_a^t - \mu_a| < \epsilon$

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \max\left(\zeta_a, \frac{\lceil \beta_\delta^{-1}(\epsilon) \rceil}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a| - 1}{2}\right). \tag{2}$$

- To have $|\mu_a^t - \mu_a| < \epsilon$, knowing that $\mathcal{C}_a = \mathcal{C}_a^t \rightarrow$

.

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \max\left(\zeta_a, \frac{\lceil \beta_\delta^{-1}(\epsilon) \rceil}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a| - 1}{2}\right). \tag{2}$$

- To have $|\mu_a^t - \mu_a| < \epsilon$, knowing that $\mathcal{C}_a = \mathcal{C}_a^t \rightarrow \ \beta_\delta(\sum_{l \in \mathcal{C}_a} n_{a,l}^t) < \epsilon$.

.

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \max\left(\zeta_a, \frac{\lceil \beta_\delta^{-1}(\epsilon) \rceil}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a| - 1}{2}\right). \tag{2}$$

- To have $|\mu_a^t - \mu_a| < \epsilon$, knowing that $\mathcal{C}_a = \mathcal{C}_a^t \rightarrow \ \beta_\delta(\sum_{l \in \mathcal{C}_a} n_{a,l}^t) < \epsilon$.
- Summing of samples: .

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \max\left(\zeta_a, \frac{\lceil \beta_\delta^{-1}(\epsilon) \rceil}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a| - 1}{2}\right). \tag{2}$$

- To have $|\mu_a^t - \mu_a| < \epsilon$, knowing that $\mathcal{C}_a = \mathcal{C}_a^t \rightarrow \ \beta_\delta(\sum_{l \in \mathcal{C}_a} n_{a,l}^t) < \epsilon$.
- Summing of samples: ($n_{\epsilon,a}$: Maximum number of samples for an agent).

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \max\left(\zeta_a, \frac{\lceil \beta_\delta^{-1}(\epsilon) \rceil}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a| - 1}{2}\right). \tag{2}$$

- To have $|\mu_a^t - \mu_a| < \epsilon$, knowing that $\mathcal{C}_a = \mathcal{C}_a^t \rightarrow \ \beta_\delta(\sum_{l \in \mathcal{C}_a} n_{a,l}^t) < \epsilon$.
- Summing of samples: ($n_{\epsilon,a}$: Maximum number of samples for an agent).

$$n_{\epsilon,a} + (n_{\epsilon,a} - 1) + \cdots + (n_{\epsilon,a} - |\mathcal{C}_a| + 1) = |\mathcal{C}_a| n_{\epsilon,a} - \frac{|\mathcal{C}_a| - 1}{2} |\mathcal{C}_a|,$$

Theoretical Results

Theorem (CoIME mean estimation time complexity (Asadi et al., 2022))

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and simple weighting, the mean estimator μ_a^t of agent a:

$$\mathbb{P}\left(\forall t > \tau_a : |\mu_a^t - \mu_a| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau_a = \max\left(\zeta_a, \frac{\lceil \beta_\delta^{-1}(\epsilon) \rceil}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a| - 1}{2}\right). \tag{2}$$

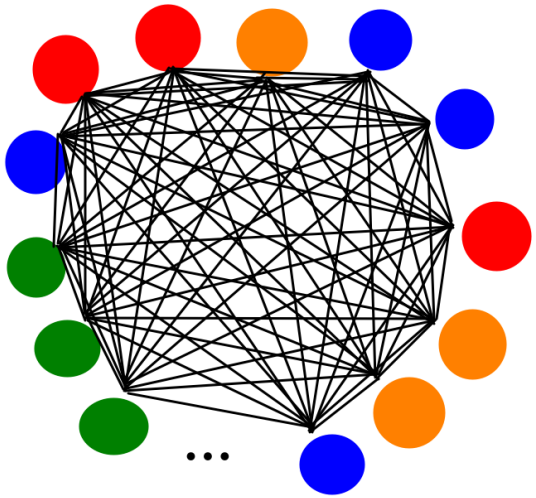
- To have $|\mu_a^t - \mu_a| < \epsilon$, knowing that $\mathcal{C}_a = \mathcal{C}_a^t \rightarrow \ \beta_\delta(\sum_{l \in \mathcal{C}_a} n_{a,l}^t) < \epsilon$.
- Summing of samples: ($n_{\epsilon,a}$: Maximum number of samples for an agent).

$$n_{\epsilon,a} + (n_{\epsilon,a} - 1) + \cdots + (n_{\epsilon,a} - |\mathcal{C}_a| + 1) = |\mathcal{C}_a| n_{\epsilon,a} - \frac{|\mathcal{C}_a| - 1}{2} |\mathcal{C}_a|,$$

$$n_{\epsilon,a} = \frac{\lceil \beta^{-1}(\epsilon) \rceil}{|\mathcal{C}_a|} + \frac{|\mathcal{C}_a| - 1}{2}.$$

Experiments

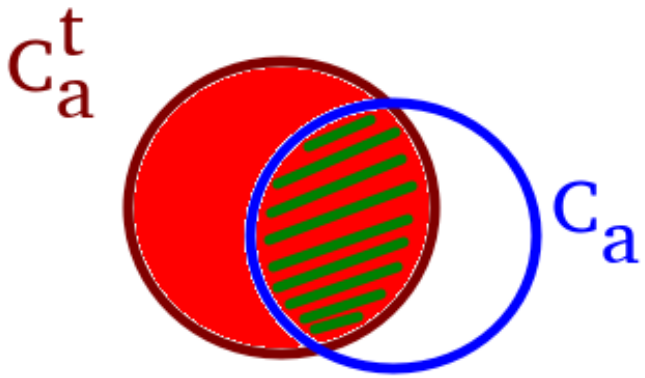
Setup



$A = 200$
Time horizon $= 2500$
risk parameter $\delta = 0.001$
$\mu = \{0.2, 0.4, 0.8\}$
$\textsc{Gaussian}(\mu_a, \sigma^2 = 0.25)$
runs $= 20$ (total of 2500*200*20*200 iterations)

Experiments

Class Identification:



$$\text{precision}_{\mathcal{C}_a^t} = \frac{|\mathcal{C}_a^t \cap \mathcal{C}_a|}{|\mathcal{C}_a^t|}.$$

Experiments

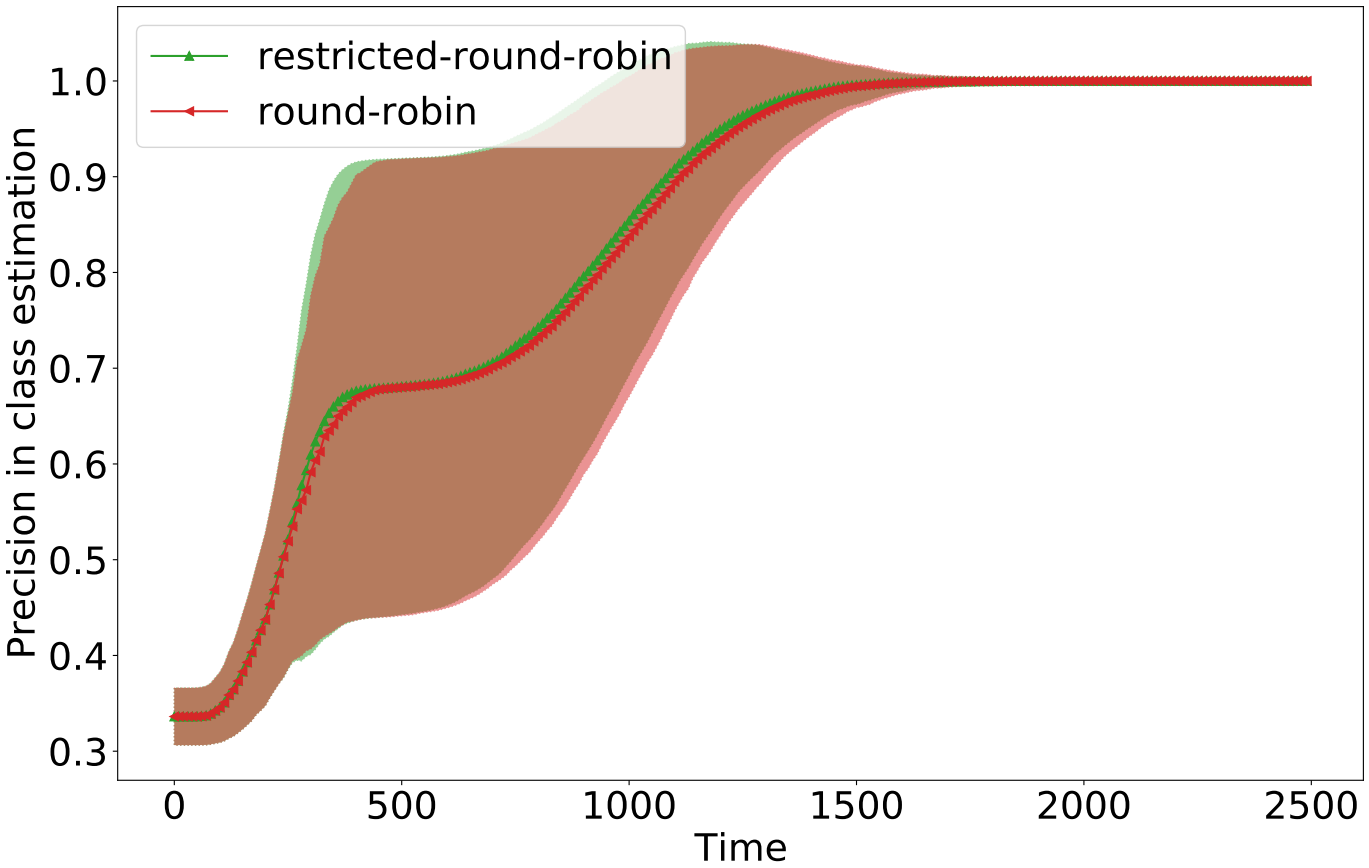


Figure: Class estimation precision over all agents

Experiments

Error of an agent a at time t:

$$\text{error}_a^t = |\mu_a^t - \mu_a|.$$

Experiments

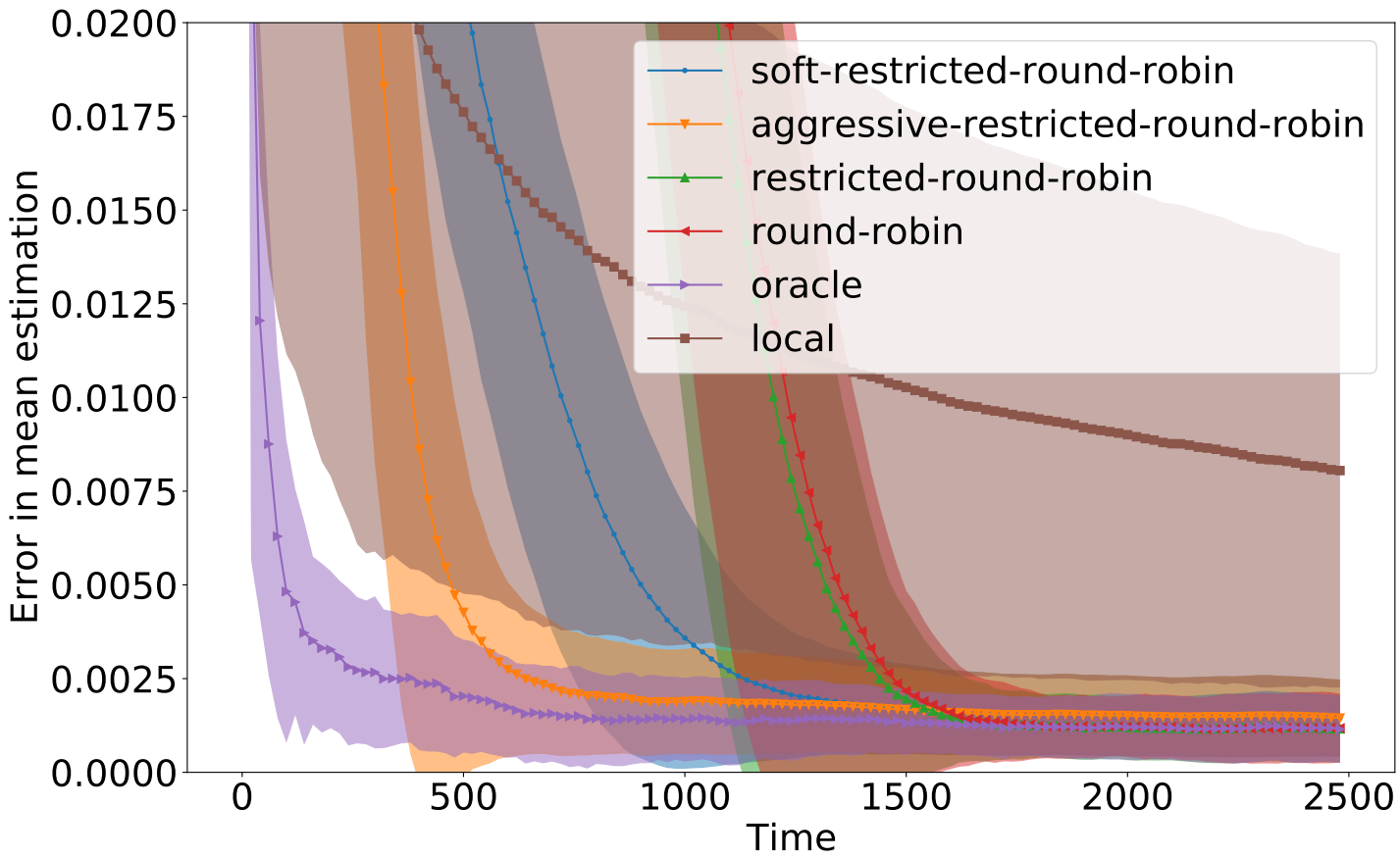


Figure: Error in mean estimation over all agents

Extension: Imperfect classes

Definition (η-similarity class)

Given $\eta > 0$, the *η-similarity class* of agent a is given by:

$$\mathcal{C}_{\eta,a} = \{l \in [A] : \Delta_{a,l} \leq \eta\},$$

where $\Delta_{a,l} = |\mu_a - \mu_l|$ is the gap between the means of agent a and agent l.

Estimating mean of class $\mathcal{C}_{\eta,a}$:

$$\mu_{\eta,a} = \frac{1}{|\mathcal{C}_{\eta,a}|} \sum_{l \in \mathcal{C}_{\eta,a}} \mu_l.$$

Contributions

- Addressing the challenging problem of **online personalized** mean estimation
- Presented collaborative learning algorithms by exploiting similarity structure between distributions
- PAC-style guarantees for class and mean estimation
  - Mean estimation time complexity is improvement by $|\mathcal{C}_a|$ times in comparison to Local
  - Mean estimation time complexity as good as Oracle in some regimes
- Heuristic weighting mechanisms that empirically shown to decrease bias in early rounds
- Extension of the ideas to imperfect classes
  - Providing class and mean estimation time complexity

Perspectives

Future Work

Model-based Reinforcement Learning

- Regret analysis of the unknown $\mathcal{C}$ case
- Extension to multiagent setting
  - What to share? observations?
  - How to coordinate and learn policy?
  - How to minimize communication?
- Extension to multiagent setting with personalized problems.

Future Work

Online Personalized Mean Estimation

- Provide theoretical guarantees for Soft-Restricted-Round-Robin and Aggressive-Restricted-Round-Robin.
- Extend to incomplete communication graph networks
  - Considering similarity between neighbour of neighbours
- Solve machine learning tasks beyond mean estimation

Thank you!

Regret bound C-UCRL

Lemma (Time-uniform Azuma-Hoeffding)

Let $(X_t)_{t\geq 1}$ be a martingale difference sequence bounded by b for some $b > 0$ (that is, $|X_t| \leq b$ for all t). Then, for all $\delta \in (0, 1)$,

$$\mathbb{P}\left(\exists T \in \mathbb{N} : \sum_{t=1}^{T} X_t \geq b\sqrt{\frac{1}{2}(T+1)\log\left(\sqrt{T+1}/\delta\right)}\right) \leq \delta\,.$$

Lemma 7

Lemma (Non-expansive ordering (Asadi et al., 2019))

Let p and q be two discrete distributions, defined on the same alphabet $\mathcal{S}$, with respective profile mappings σ_p and σ_q. Then,

$$\|p(\sigma_p(\cdot)) - q(\sigma_q(\cdot))\|_1 \leq \|p - q\|_1 .$$

Lemma 7

Proof.

- $\|p - q\|_1 = \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1 \rightarrow$

Lemma 7

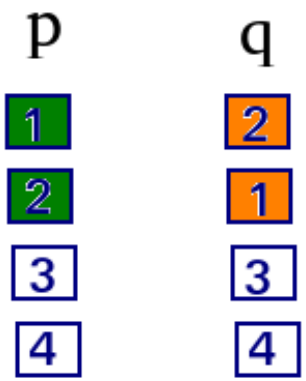
Proof.

- $\|p - q\|_1 = \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1 \rightarrow$
  $\|p(\sigma_p(\cdot)) - q(\sigma_q(\cdot))\|_1 \leq \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1$

Lemma 7

Proof.

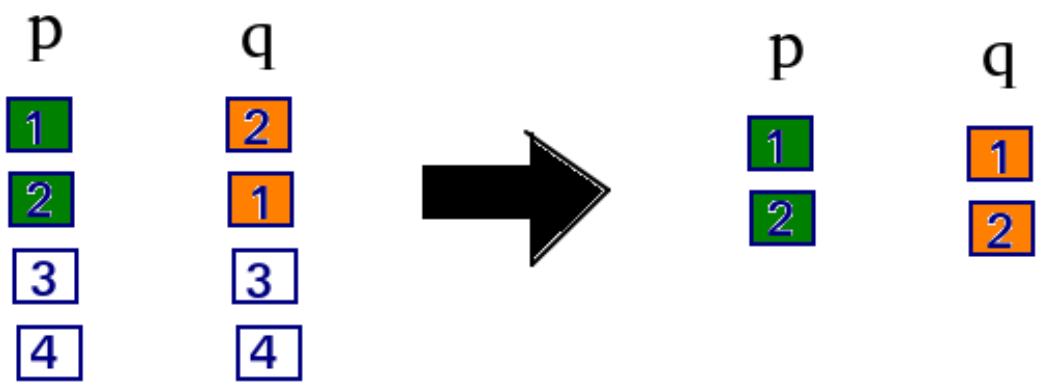
- $\|p - q\|_1 = \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1 \rightarrow$
  $\|p(\sigma_p(\cdot)) - q(\sigma_q(\cdot))\|_1 \leq \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1$
- Lets say index 1 and 2 are switched:
  - Therefore, $\sigma_p(1) = \sigma_q(2)$ and $\sigma_p(2) = \sigma_q(1)$



Lemma 7

Proof.

- $\|p - q\|_1 = \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1 \rightarrow$
  $\|p(\sigma_p(\cdot)) - q(\sigma_q(\cdot))\|_1 \leq \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1$
- Lets say index 1 and 2 are switched:
  - Therefore, $\sigma_p(1) = \sigma_q(2)$ and $\sigma_p(2) = \sigma_q(1)$
  - Therefore, $p(\sigma_p(1)) > p(\sigma_p(2))$ but $p(\sigma_q(1)) < p(\sigma_q(2))$.



Lemma 7

Proof.

- $\|p - q\|_1 = \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1 \rightarrow$
  $\|p(\sigma_p(\cdot)) - q(\sigma_q(\cdot))\|_1 \leq \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1$
- Lets say index 1 and 2 are switched:
  - Therefore, $\sigma_p(1) = \sigma_q(2)$ and $\sigma_p(2) = \sigma_q(1)$
  - Therefore, $p(\sigma_p(1)) > p(\sigma_p(2))$ but $p(\sigma_q(1)) < p(\sigma_q(2))$.
  - Then, we study $\sum_{i=1,2} |p(\sigma_p(i)) - q(\sigma_q(i))|$:
    - if $q(\sigma_q(1)) < p(\sigma_p(1))$ and $q(\sigma_q(2)) < p(\sigma_p(2))$ then: The sum equals $|p(\sigma_p(1)) - q(\sigma_p(1))| + |p(\sigma_p(2)) - q(\sigma_p(2))|$.

Lemma 7

Proof.

- $\|p - q\|_1 = \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1 \rightarrow$
  $\|p(\sigma_p(\cdot)) - q(\sigma_q(\cdot))\|_1 \leq \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1$
- Lets say index 1 and 2 are switched:
  - Therefore, $\sigma_p(1) = \sigma_q(2)$ and $\sigma_p(2) = \sigma_q(1)$
  - Therefore, $p(\sigma_p(1)) > p(\sigma_p(2))$ but $p(\sigma_q(1)) < p(\sigma_q(2))$.
  - Then, we study $\sum_{i=1,2} |p(\sigma_p(i)) - q(\sigma_q(i))|$:
    - if $q(\sigma_q(1)) < p(\sigma_p(1))$ and $q(\sigma_q(2)) < p(\sigma_p(2))$ then: The sum equals $|p(\sigma_p(1)) - q(\sigma_p(1))| + |p(\sigma_p(2)) - q(\sigma_p(2))|$.
    - The same for $q(\sigma_q(1)) > p(\sigma_p(1))$ and $q(\sigma_q(2)) > p(\sigma_p(2))$.

Lemma 7

Proof.

- $\|p-q\|_1 = \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1 \rightarrow$
  $\|p(\sigma_p(\cdot)) - q(\sigma_q(\cdot))\|_1 \leq \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1$
- Lets say index 1 and 2 are switched:
  - Therefore, $\sigma_p(1) = \sigma_q(2)$ and $\sigma_p(2) = \sigma_q(1)$
  - Therefore, $p(\sigma_p(1)) > p(\sigma_p(2))$ but $p(\sigma_q(1)) < p(\sigma_q(2))$.
  - Then, we study $\sum_{i=1,2} |p(\sigma_p(i)) - q(\sigma_q(i))|$:
    - if $q(\sigma_q(1)) < p(\sigma_p(1))$ and $q(\sigma_q(2)) < p(\sigma_p(2))$ then: The sum equals $|p(\sigma_p(1)) - q(\sigma_p(1))| + |p(\sigma_p(2)) - q(\sigma_p(2))|$.
    - The same for $q(\sigma_q(1)) > p(\sigma_p(1))$ and $q(\sigma_q(2)) > p(\sigma_p(2))$.
    - For the remaining cases also: The sum $\leq |p(\sigma_p(1)) - q(\sigma_p(1))| + |p(\sigma_p(2)) - q(\sigma_p(2))|$

Lemma 7

Proof.

- $\|p-q\|_1 = \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1 \rightarrow$
  $\|p(\sigma_p(\cdot)) - q(\sigma_q(\cdot))\|_1 \leq \|p(\sigma_p(\cdot)) - q(\sigma_p(\cdot))\|_1$
- Lets say index 1 and 2 are switched:
  - Therefore, $\sigma_p(1) = \sigma_q(2)$ and $\sigma_p(2) = \sigma_q(1)$
  - Therefore, $p(\sigma_p(1)) > p(\sigma_p(2))$ but $p(\sigma_q(1)) < p(\sigma_q(2))$.
  - Then, we study $\sum_{i=1,2} |p(\sigma_p(i)) - q(\sigma_q(i))|$:
    - if $q(\sigma_q(1)) < p(\sigma_p(1))$ and $q(\sigma_q(2)) < p(\sigma_p(2))$ then: The sum equals $|p(\sigma_p(1)) - q(\sigma_p(1))| + |p(\sigma_p(2)) - q(\sigma_p(2))|$.
    - The same for $q(\sigma_q(1)) > p(\sigma_p(1))$ and $q(\sigma_q(2)) > p(\sigma_p(2))$.
    - For the remaining cases also: The sum $\leq |p(\sigma_p(1)) - q(\sigma_p(1))| + |p(\sigma_p(2)) - q(\sigma_p(2))|$
- For more switches, this iteratively holds.

Known $\mathcal{C}$, Unknown σ

Corollary

The modified confidence set $\mathsf{CB}_{t,\delta}(\mathcal{C})$ *contains the true transition function* p *with probability at least* $1-\delta$*, uniformly over all time* t*, i.e.:*

$$\mathbb{P}\left(\bigcap_t p \in \mathsf{CB}_{t,\delta}(\mathcal{C})\right) > 1-\delta\,.$$

where

$$\mathsf{CB}_{t,\delta} = \{p' : \|\widehat{p}_t^{\boldsymbol{\sigma}^t}(\cdot|c) - p'(\sigma_{s,a}(\cdot)|s,a)\|_1 \leq \frac{1}{N_t(c)} \sum_{(s',a')\in c} N_t(s',a')\beta_{N_t(s',a')}\left(\tfrac{\delta}{C}\right)$$
$$\forall c \in \mathcal{C}, \forall (s,a) \in c\}$$

Theoretical Results

C-UCRL Regret bound

Theorem (Regret of C-UCRL)

Given $\mathcal{C}$, $\boldsymbol{\sigma}$ and for any $\delta \in (0,1)$ such that with probability higher than $1-\delta$, uniformly over all time horizon T,

$$\mathbb{R}(\textit{C-UCRL}(\mathcal{C}, \boldsymbol{\sigma}, \delta'), T) \leq 20 D \sqrt{CT\big(S + \log(C\sqrt{T+1}/\delta')\big)} + DC \log_2(\tfrac{8T}{C}) \,.$$

where $\delta' = \delta/4$, C is the number of classes and D is the diameter of the MDP.

Proof.

$$\mathbb{R}(T) = \sum_{t=1}^{T} g_\star - \sum_{t=1}^{T} r_t(s_t, a_t)$$
$$= \sum_{t=1}^{T} (g_\star - \mu(s_t, a_t)) + \sum_{t=1}^{T} (\mu(s_t, a_t) - r_t(s_t, a_t)) \,.$$

Hence with probability $1 - \delta'$:

$$\mathbb{R}(T) \leq \sum_{k=1}^{m(T)} \Delta_k + \sqrt{\tfrac{1}{2}(T+1)\log(\sqrt{T+1}/\delta')} \,.$$

where $\Delta_k = \sum_{c \in \mathcal{C}} V_k(c)\big(g_\star - \mu(c)\big)$. □

Proof.

- Bad episodes($M \notin \mathcal{M}_{k,\delta'}$):

$$\mathbb{P}\left(\bigcap_t p \in \mathsf{CB}_{t,\delta'}(\mathcal{C},\boldsymbol{\sigma}) \wedge \mu \in \mathsf{CB}'_{t,\delta'}(\mathcal{C},\boldsymbol{\sigma})\right) \geq 1-2\delta'.$$

  Therefore, with probability at least $1-2\delta'$ and

$$\sum_{k=1}^{m(T)} \Delta_k \mathbb{1}_{\{M\notin\mathcal{M}_{k,\delta'}\}} = 0\,,$$

- Good episodes($M \in \mathcal{M}_{k,\delta'}$):

$$\mathbb{R}(T) \leq 20D\sqrt{CT\big(S+\log(C\sqrt{T+1}/\delta')\big)} + DC\log_2(\tfrac{8T}{C})\,,$$

River Swim MDP

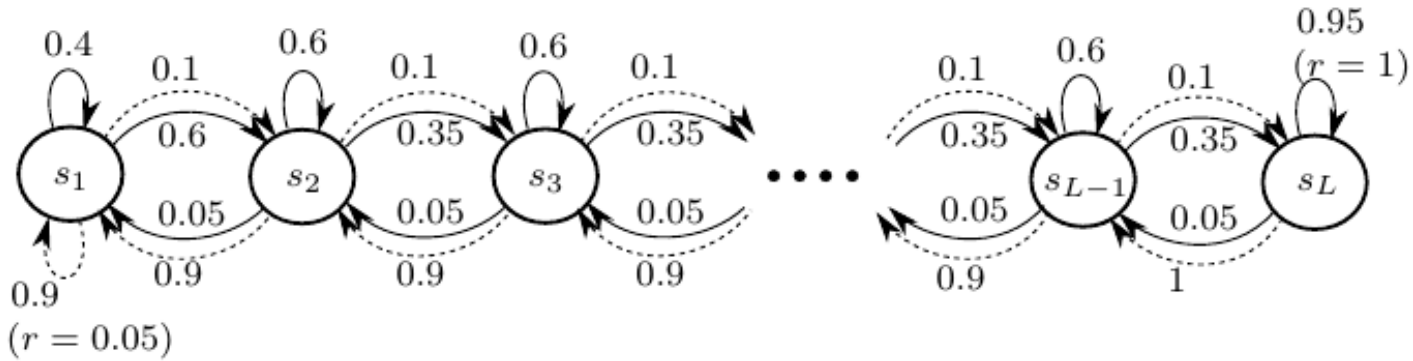


Figure : The L-state *Ergodic RiverSwim* MDP

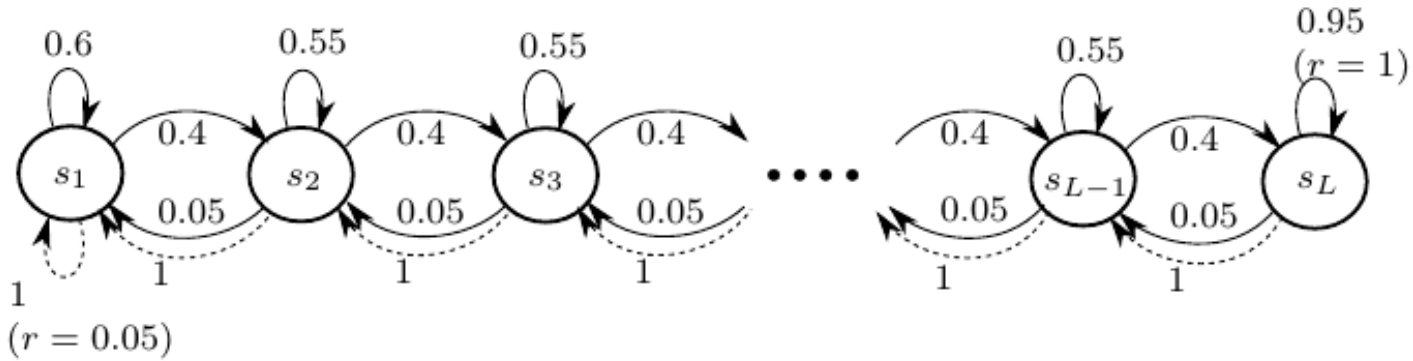
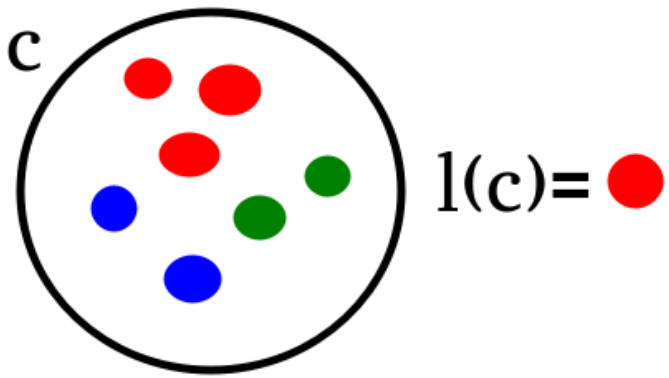


Figure : The L-state *RiverSwim* MDP

Numerical Results

Label of class

$$\ell(c) \in \underset{x \in \mathcal{C}|c}{\operatorname{argmax}} |x|$$



Numerical Results

Performance metric

$$\text{mis-clustering ratio at time } t = \frac{1}{SA} \sum_{c \in \mathcal{C}_t} \left| \{\text{pairs in c that are mis-clustered}\} \right| .$$

Numerical Results

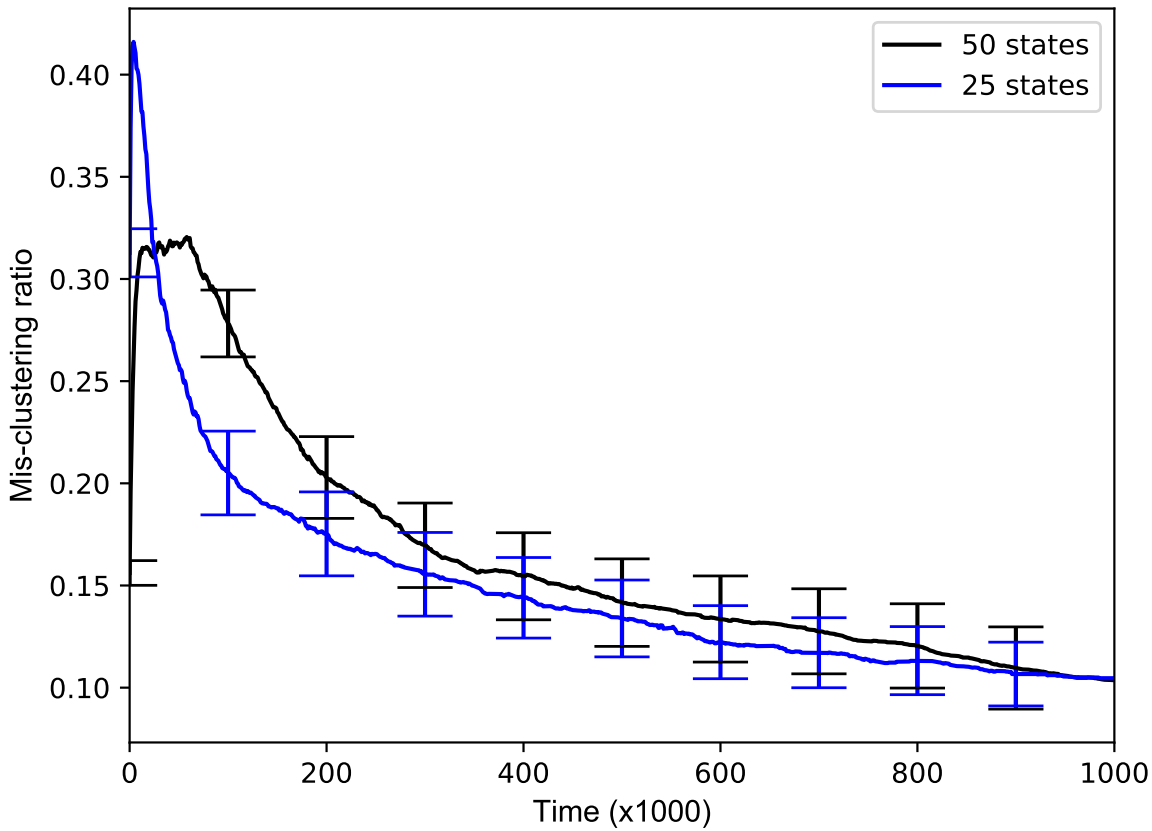


Figure: Mis-clustering ratio - Ergodic RiverSwim(25 and 50 states)

Numerical Results

Performance metric

mis-clustering bias at time $t =$

Numerical Results

Performance metric

$$\text{mis-clustering bias at time } t = \sum_{c \in \mathcal{C}_t}$$

Numerical Results

Performance metric

$$\text{mis-clustering bias at time } t = \sum_{c \in \mathcal{C}_t} \sum_{e \notin \ell(c)}$$

Numerical Results

Performance metric

$$\text{mis-clustering bias at time } t = \sum_{c \in \mathcal{C}_t} \sum_{e \notin \ell(c)} \| \qquad - \qquad \|_1 \, .$$

Numerical Results

Performance metric

$$\text{mis-clustering bias at time } t = \sum_{c \in \mathcal{C}_t} \sum_{e \notin \ell(c)} \|\hat{p}_t(\cdot|c_e) - \qquad \|_1 \,.$$

Numerical Results

Performance metric

$$\text{mis-clustering bias at time } t = \sum_{c \in \mathcal{C}_t} \sum_{e \notin \ell(c)} \|\hat{p}_t(\cdot|c_e) - \hat{p}_t(\cdot|c_e \setminus \{e\})\|_1 \,.$$

Numerical Results

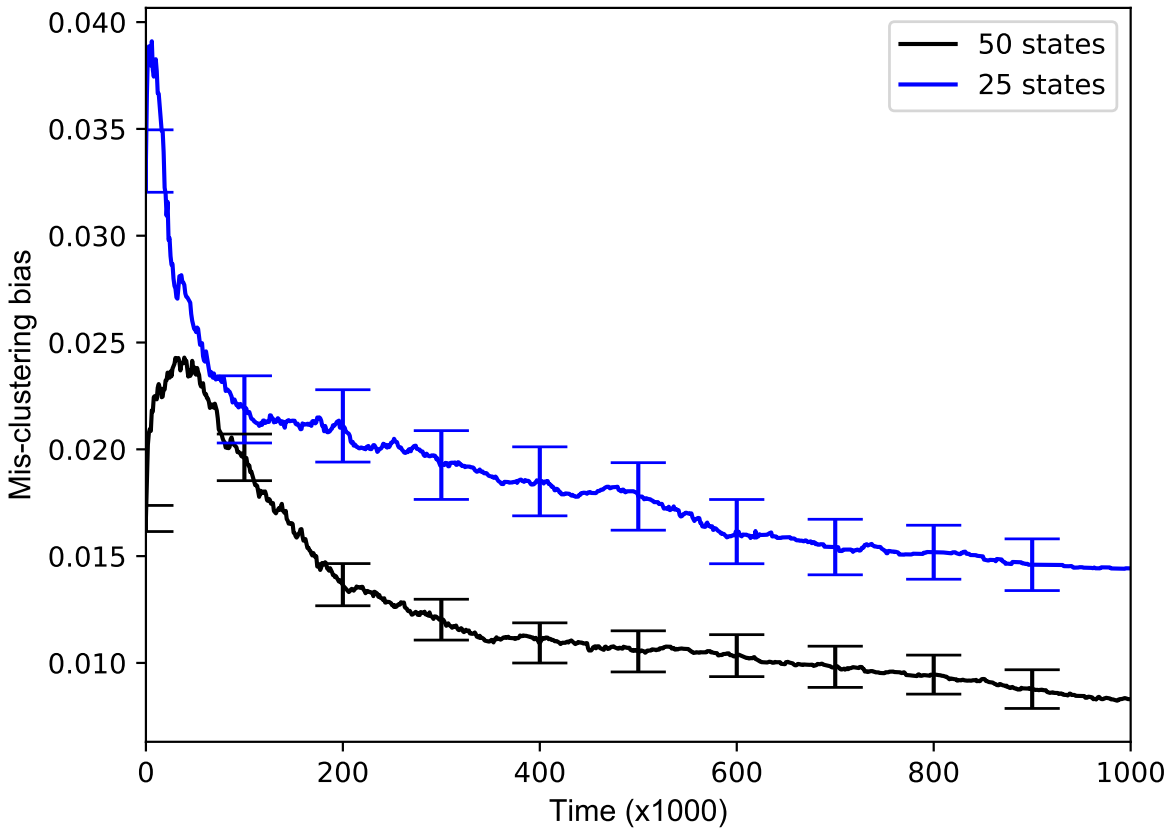


Figure: Mis-clustering bias - Ergodic RiverSwim(25 and 50 states)

Numerical Results

Communicating MDP

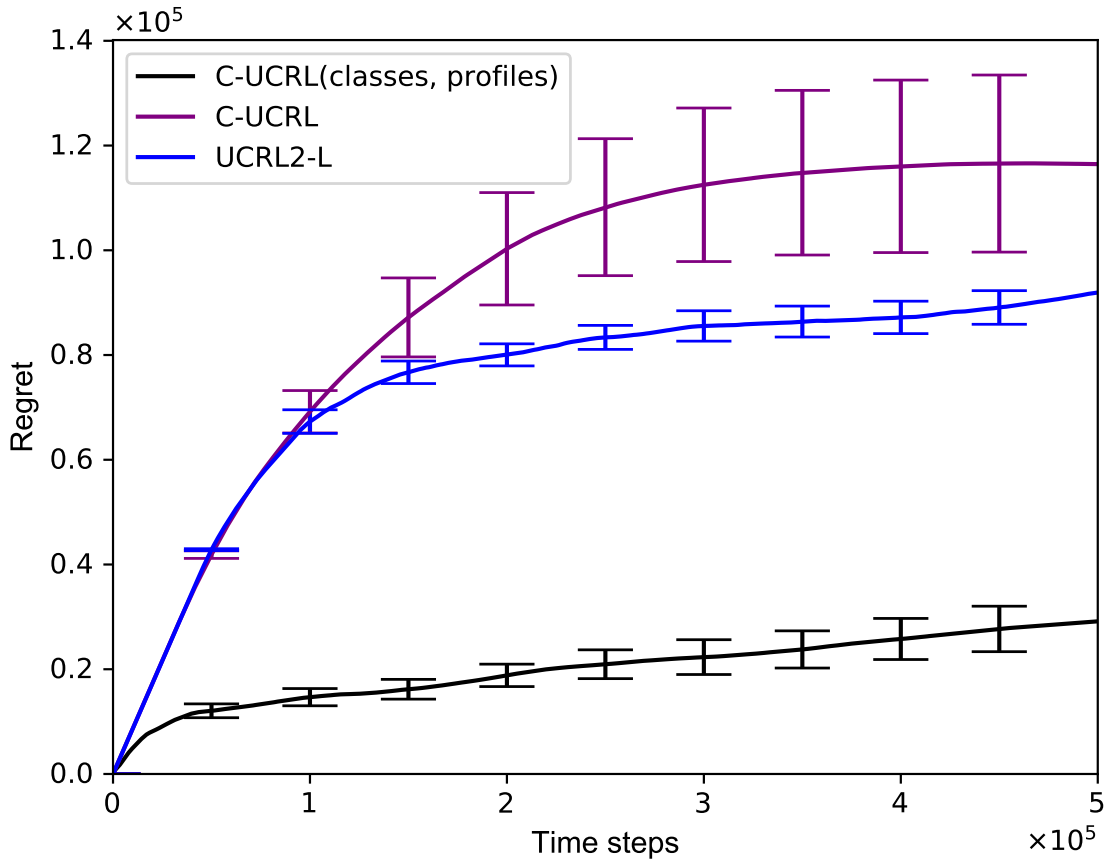


Figure: 25-state RiverSwim

Algorithm 10 ColME

Parameters: agent a, time horizon H, risk δ, weighting scheme α, and query strategy choose_agent

$\mathcal{C}_a^0 \leftarrow \{l \in [A] : d_{a,\delta}^t(l) \leq 0\} = [A]$
for $t = 1, \ldots, H$ **do**
  $\forall l \in [A]$: $\bar{x}_{a,l}^t \leftarrow \bar{x}_{a,l}^{t-1}$, $n_{a,l}^t \leftarrow n_{a,l}^{t-1}$
  Perceive:
    Receive sample $x_a^t \sim \nu_a$
    $\bar{x}_{a,a}^t \leftarrow \bar{x}_{a,a}^{t-1} \times \frac{t-1}{t} + x_a^t \times \frac{1}{t}$, $n_{a,a}^t \leftarrow t$
  Query:
    $\mathcal{C}_a^t \leftarrow \{l \in [A] : d_{a,\delta}^t(l) \leq 0\}$
    Query agent $l =$ choose_agent$(\mathcal{C}_a^t)$ to get $\bar{x}_{l,l}^t$
    $\bar{x}_{a,l}^t \leftarrow \bar{x}_{l,l}^t$, $n_{a,l}^t \leftarrow t$
  Estimate:
    $\mathcal{C}_a^t \leftarrow \{l \in [A] : d_{a,\delta}^t(l) \leq 0\}$
    $\mu_a^t \leftarrow \sum_{l \in \mathcal{C}_a^t} \alpha_{a,l}^t \times \bar{x}_{a,l}^t$
end for

Output: μ_a^H

Definition

We define the following events:

$$E_a^t = \bigcap_{l \in [A]} |\bar{x}_{a,l}^t - \mu_l| \leq \beta_\delta(n_{a,l}^t),$$

$$E_a = \bigcap_{t \in \mathbb{N}} E_a^t.$$

Lemma

Let $\delta \in (0,1)$, $a \in [A]$. *Setting* $\beta_\delta(n) = \sigma\sqrt{2\frac{1}{n} \times (1+\frac{1}{n}) \ln(\sqrt{n+1}/\gamma(\delta))}$ *with* $\gamma(\delta) = \frac{\delta}{8 \times A}$, *we have:*

$$\mathbb{P}\left(E_a\right) \geq 1 - \frac{\delta}{8}. \tag{3}$$

Definition

From the perspective of agent a and at time t, event G_a^t is defined as:

$$G_a^t = \bigcap_{l \in [A]} n_{a,l}^t > n_{a,l}^\star,$$

where
$$n_{a,l}^\star = \begin{cases} \lceil \beta_\delta^{-1}(\frac{\Delta_{a,l}}{4}) \rceil & \text{if } l \notin \mathcal{C}_a, \\ \lceil \beta_\delta^{-1}(\frac{\Delta_a}{4}) \rceil & \text{otherwise,} \end{cases}$$

with $\Delta_a = \min_{l \in [A] \setminus \mathcal{C}_a} \Delta_{a,l}$.

Empirical convergence time of an agent:

$$\mathrm{conv}_a(\epsilon) = \min\{\tau \in \mathbb{N} : \forall t \geq \tau, \mathsf{error}_a^t \leq \epsilon\}.$$

Analysis

Numerical

Algorithm	conv(0.1)		conv(0.01)	
	avg/std	max	avg/std	max
Round-Robin (RR)	417 ± 230	1239	916 ± 427	2088
Restricted-RR	405 ± 227	1175	894 ± 438	1925
Soft-Restricted-RR	82 ± 48	340	608 ± 290	1432
Aggressive-Restricted-RR	56 ± 39	340	$\mathbf{335 \pm 127}$	**958**
Local	$\mathbf{41 \pm 39}$	**289**	4494 ± 3945	28616
Oracle	*5 ± 4*	*19*	*98 ± 63*	*303*
Theoretical Restricted-RR (τ_a)	3878		33406	
Theoretical Local (τ_a)	885		100216	

Imperfect classes

Definition

The η-*optimistic similarity class* from the perspective of agent a at time t is defined as:

$$\mathcal{C}^t_{\eta,a} = \{l \in [A] : d^t_{a,\delta}(l) \leq \eta\}.$$

Imperfect classes

Definition

From the perspective of agent a and at time t, event $G^t_{\eta,a}$ is defined as:

$$G^t_{\eta,a} = \bigcap_{l \in [A]} n^t_{a,l} > n^{\eta}_{a,l}, \tag{4}$$

where $$n^{\eta}_{a,l} = \begin{cases} \lceil \beta_\delta^{-1}(\frac{\Delta_{a,l}-\eta}{4}) \rceil & \text{if } l \notin \mathcal{C}_a, \\ \lceil \beta_\delta^{-1}(\frac{\Delta_{\eta,a}-\eta}{4}) \rceil & \text{otherwise,} \end{cases}$$

with $\Delta_{\eta,a} = \min_{l \in [A] \setminus \mathcal{C}_{\eta,a}} \Delta_{a,l}$.

Imperfect classes

Theorem (η-CoIME mean estimation time complexity)

Given the risk parameter δ, using the Restricted-Round-Robin query strategy and class-uniform weighting (while employing $\mathcal{C}_{\eta,a}$), the mean estimator $\mu^t_{\eta,a}$ of agent a is:

$$\mathbb{P}\left(\forall t > \tau^\eta_a : |\mu^t_{\eta,a} - \mu_{\eta,a}| \leq \epsilon\right) > 1 - \frac{\delta}{4}, \quad \textit{with} \quad \tau^\eta_a = \max(\zeta^\eta_a, \beta^{-1}_\delta(\epsilon) + |\mathcal{C}_{\eta,a}| - 1). \tag{5}$$