The Online Discovery Problem and Its Application to Lifelong Reinforcement Learning

Emma Brunskill
Carnegie Mellon University

Lihong Li
Microsoft Research

Multi-disciplinary Conference on Reinforcement Learning and Decision Making 2015
Edmonton, AB, Canada

Full version to be available on arXiv

Lifelong Learning Example: Intelligent Tutoring Systems

Alice

Lifelong Learning Example: Intelligent Tutoring Systems

Alice

State = (courses taken, skills mastered, grades, …)

Lifelong Learning Example: Intelligent Tutoring Systems

Action ∈ { test skill,
teach new concept,
review old lectures,
… }

Alice

State = (courses taken, skills mastered, grades, …)

Lifelong Learning Example: Intelligent Tutoring Systems

Action ∈ { test skill,
teach new concept,
review old lectures,
… }

Alice

State = (courses taken, skills mastered, grades, …)

Lifelong Learning Example: Intelligent Tutoring Systems

Action ∈ { test skill,
teach new concept,
review old lectures,
… }

Alice

State = (courses taken, skills mastered, grades, …)

Lifelong Learning Example: Intelligent Tutoring Systems

Action ∈ { test skill,
teach new concept,
review old lectures,
… }

Alice

State = (courses taken, skills mastered, grades, …)

How to benefit from past teaching experience?

Lifelong Learning Example: Intelligent Tutoring Systems

Action ∈ { test skill,
teach new concept,
review old lectures,
… }

How to teach Alice to benefit future students?

Alice

State = (courses taken, skills mastered, grades, …)

How to benefit from past teaching experience?

Task as Finite Markov Decision Process (MDP)

$$M = \langle S, A, P, R, \gamma \rangle$$

State space

Action space

Transition probabilities

Reward function

Discount factor in (0,1)

$$\dashrightarrow s_t \to a_t \to r_t \to s_{t+1} \dashrightarrow$$

$$\mathbf{E}[r_t] = R(s_t, a_t)$$

$$s_{t+1} \sim P(\cdot \mid s_t, a_t)$$

A Class of Lifelong RL Problems

- Given (known): S (finite), A (finite), $\gamma \in (0,1)$
- Unknown: $\mathbf{M} = \{M^1, M^2, \dots, M^C\}$

  $\forall M \in \mathbf{M}, M = \langle S, A, P_M, R_M, \gamma \rangle$

For $t = 1,2,\dots,T$

- Environment chooses an unknown $M_t \in \mathbf{M}$
- Agent acts in M_t for H steps

Note: Many previous works on LLRL

with different setups

A Class of Lifelong RL Problems

- Given (known): S (finite), A (finite), $\gamma \in (0,1)$
- Unknown: $\mathbf{M} = \{M^1, M^2, \dots, M^C\}$

  $\forall M \in \mathbf{M}, M = \langle S, A, P_M, R_M, \gamma \rangle$

Finite S and A

Finitely many MDPs with "large" model differences

For $t = 1,2, \dots, T$
- Environment chooses an unknown $M_t \in \mathbf{M}$
- Agent acts in M_t for H steps

Note: Many previous works on LLRL
with different setups

A Class of Lifelong RL Problems

- Given (known): S (finite), A (finite), $\gamma \in (0,1)$
- Unknown: $\mathbf{M} = \{M^1, M^2, \dots, M^C\}$

  $\forall M \in \mathbf{M}, M = \langle S, A, P_M, R_M, \gamma \rangle$

For $t = 1,2, \dots, T$

- Environment chooses an unknown $M_t \in \mathbf{M}$
- Agent acts in M_t for H steps

Note: Many previous works on LLRL
with different setups

Finite S and A

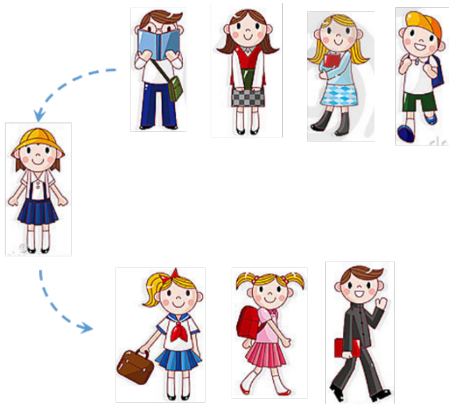
⇩

Finitely many MDPs with "large" model differences

Examples

- Student types w/ varying learning rates [Liu&Koedinger]
- User types in human robot interaction [Nikolaidis et al.]
- User goal recognition for task assistance [Fern et al.]

Two Kinds of Exploration

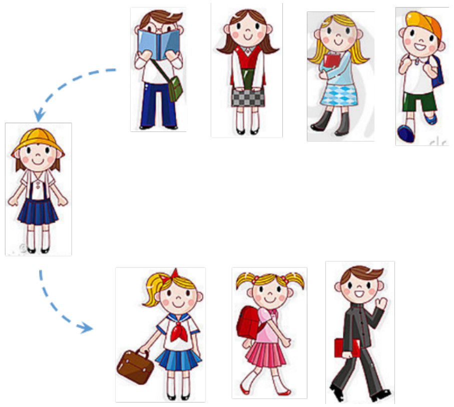


$\widehat{\mathbf{M}}$: set of discovered types before t

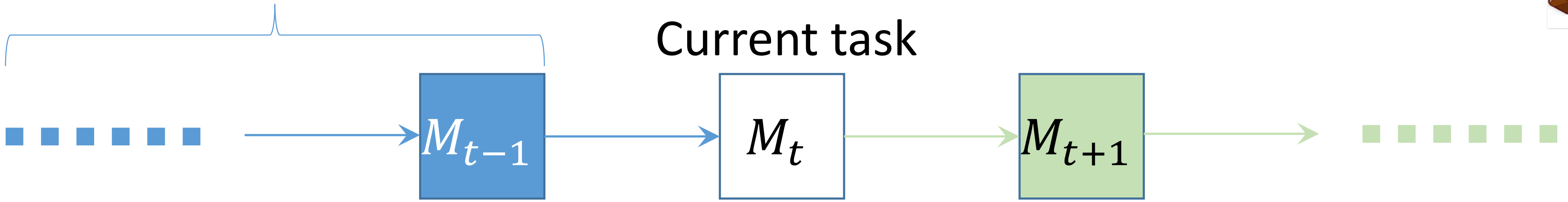
Current task

$$\overbrace{\cdots \rightarrow M_{t-1}} \rightarrow M_t \rightarrow M_{t+1} \rightarrow \cdots$$

Two Kinds of Exploration



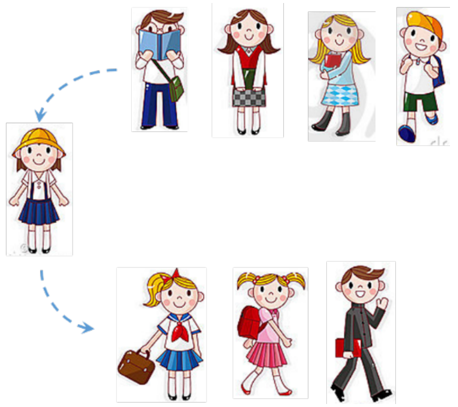
$\widehat{\mathbf{M}}$: set of discovered types before t



When in M_t…

- Within-task learning:
- Cross-task knowledge transfer:

Two Kinds of Exploration



$\hat{\mathbf{M}}$: set of discovered types before t

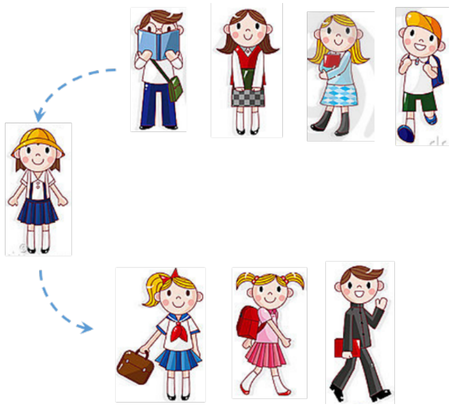
Current task

$$\overbrace{\cdots\cdots \longrightarrow \boxed{M_{t-1}}} \longrightarrow \boxed{M_t} \longrightarrow \boxed{M_{t+1}} \longrightarrow \cdots\cdots$$

When in M_t...

- o Within-task learning:
  - Goal: maximize reward in M_t
  - Explore promising states in M_t until policy is ϵ-optimal
- o Cross-task knowledge transfer:

Two Kinds of Exploration



$\widehat{\mathbf{M}}$: set of discovered types before t

Current task

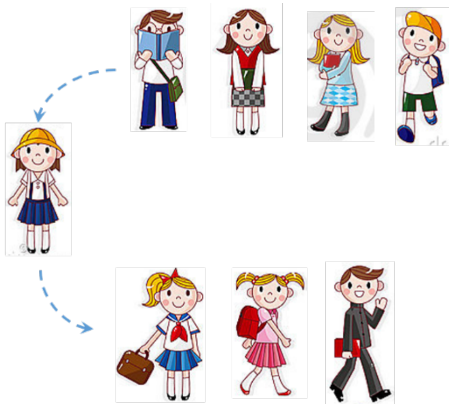
$$\overbrace{\cdots\cdots \longrightarrow M_{t-1}} \longrightarrow M_t \longrightarrow M_{t+1} \longrightarrow \cdots\cdots$$

When in M_t…

- Within-task learning:
  - Goal: maximize reward in M_t
  - Explore promising states in M_t until policy is ϵ-optimal

- Cross-task knowledge transfer:
  - Goal: maximize reward in M_{t+1}, … w/ transferable info.
  - Explore possibly all states in M_t to discover novel types

Two Kinds of Exploration



$\widehat{\mathbf{M}}$: set of discovered types before t

Current task

$\cdots \rightarrow M_{t-1} \rightarrow M_t \rightarrow M_{t+1} \rightarrow \cdots$

When in M_t…

- Within-task learning:
  - Goal: maximize reward in M_t
  - Explore promising states in M_t until policy is ϵ-optimal
- Cross-task knowledge transfer:
  - Goal: maximize reward in M_{t+1}, … w/ transferable info.
  - Explore possibly all states in M_t to discover novel types

Cross-task E/E tradeoff
over
within-task E/E tradeoff

The *Online Discovery* Problem: Abstraction of Cross-task Exploration

Environment has an unknown set $\mathbf{M} = \{M^1, M^2, \dots, M^C\}$

Agent starts with $\widehat{\mathbf{M}} = \emptyset$

For $t = 1,2, \dots, T$

- $\in \mathbf{M}$
- **Agent** chooses to explore ($A_t = 1$) or exploit ($A_t = 0$)
  - If $A_t = 1$, $\widehat{\mathbf{M}} \leftarrow \widehat{\mathbf{M}} \cup \{M_t\}$
- Loss to agent

Agent aims to minimize total loss

The *Online Discovery* Problem: Abstraction of Cross-task Exploration

$\mathbf{MM}$

Environment has an unknown set $\mathbf{M} = \{M^1, M^2, \ldots, M^C\}$

Agent starts with $\widehat{\mathbf{M}} = \emptyset$

For $t = 1,2, \ldots, T$

- **Environment** selects $M\ t \in\ t\ t\ t\ \in \mathbf{M}$
- **Agent** chooses to explore ($A_t = 1$) or exploit ($A_t = 0$)
  - If $A_t = 1$, $\widehat{\mathbf{M}} \leftarrow \widehat{\mathbf{M}} \cup \{M_t\}$
- Loss to agent

Agent aims to minimize total loss

The *Online Discovery* Problem: Abstraction of Cross-task Exploration

1, $\mathbf{M}\ \mathbf{MM}\ \mathbf{M} \leftarrow \mathbf{M}\ \mathbf{MM}\ \mathbf{M} \cup \{ M\ t\ MM\ M\ t\ tt\ M\ t \}$

1**) or exploit (** $\boldsymbol{A\ t}$ *AA A t tt A t =0)*

$\boldsymbol{MM}$

Environment has an unknown set $\mathbf{M} = \{M^1, M^2, \dots, M^C\}$

Agent starts with $\widehat{\mathbf{M}} = \emptyset$

For $t = 1,2, \dots, T$

- If $A\ t = \ t\ t\ t \ = 1, \widehat{\mathbf{M}} \leftarrow \widehat{\mathbf{M}} \cup \{M_t\}$
- **Agent** chooses to explore ($A_t = 1$) or exploit ($A_t = 0$)
  - If $A_t = 1, \widehat{\mathbf{M}} \leftarrow \widehat{\mathbf{M}} \cup \{M_t\}$
- Loss to agent

The *Online Discovery* Problem: Abstraction of Cross-task Exploration

1, $\mathbf{M}\ \mathbf{MM}\ \mathbf{M} \leftarrow \mathbf{M}\ \mathbf{MM}\ \mathbf{M} \cup \{ M\ t\ MM\ M\ t\ tt\ M\ t \}$

1) **or exploit (** $\boldsymbol{A\ t}\ AA\ A\ t\ tt\ A\ t = 0)$

$\boldsymbol{MM}$

Environment has an unknown set $\mathbf{M} = \{M^1, M^2, \dots, M^C\}$

Agent starts with $\widehat{\mathbf{M}} = \emptyset$

For $t = 1,2,\dots,T$

- Loss to agent

	$M_t \in \widehat{\mathbf{M}}$	$M_t \notin \widehat{\mathbf{M}}$
$A_t = 0$	ρ_0	ρ_3
$A_t = 1$	ρ_1	ρ_2

$(\rho_0 \ll \rho_1 \leq \rho_2 \ll \rho_3)$

- **Agent** c 1) or exploit ($A_t = 0$)
    - If $A_t = 1$, $\mathbf{M} \leftarrow \mathbf{M} \cup \{M_t\}$
- Loss to agent

The *Online Discovery* Problem: Abstraction of Cross-task Exploration

1, **M MM M** ← **M MM M** ∪{ *M t MM M t tt M t* }

1) **or exploit (** ***A t*** *AA A t tt A t =0)*

MM

Environment has an unknown set $\mathbf{M} = \{M^1, M^2, \dots, M^C\}$

Agent starts with $\widehat{\mathbf{M}} = \emptyset$

For $t = 1,2,\dots,T$

- Loss to agent
- **Agent** c 1) or exploit ($A_t = 0$)
  - If $A_t = 1$, $\mathbf{M} \leftarrow \mathbf{M} \cup \{M_t\}$

	$M_t \in \widehat{\mathbf{M}}$	$M_t \notin \widehat{\mathbf{M}}$
$A_t = 0$	ρ_0	ρ_3
$A_t = 1$	ρ_1	ρ_2

$(\rho_0 \ll \rho_1 \le \rho_2 \ll \rho_3)$

Agent aims to minimize total loss

The *Online Discovery* Problem: Abstraction of Cross-task Exploration

1, **M MM M** ← **M MM M** ∪{ M t MM M t tt M t }

1) **or exploit (** ***A t*** *AA A t tt A t =0)*

MM

Environment has an unknown set $\mathbf{M} = \{M^1, M^2, \dots, M^C\}$

Agent starts with $\widehat{\mathbf{M}} = \emptyset$

For $t = 1,2,\dots,T$

- Loss to agent
- **Agent** c 1) or exploit ($A_t = 0$)
  - If $A_t = 1$, $\mathbf{M} \leftarrow \mathbf{M} \cup \{M_t\}$

	$M_t \in \widehat{\mathbf{M}}$	$M_t \notin \widehat{\mathbf{M}}$
$A_t = 0$	ρ_0	ρ_3
$A_t = 1$	ρ_1	ρ_2

$(\rho_0 \ll \rho_1 \leq \rho_2 \ll \rho_3)$

ρ_0: successful transfer

ρ_3: negative transfer

Agent aims to minimize total loss

Explore-First Algorithm

Stochastic assumptions:

$M_t \sim \mu$ *i.i.d.* over $\mathbf{M}$, and $\mu_m \coloneqq \min_{M \in \mathbf{M}} \mu(M)$

- Action selection

$$A_t = \begin{cases} 1 & \text{if } t \leq E \qquad \text{(Exploration phase)} \\ 0 & \text{otherwise} \qquad \text{(Exploitation phase)} \end{cases}$$

- ____ $= O(\mu_m^{-1} \log(C\mu_m T))$, then

$$\text{AverageLoss} \leq \text{OptLoss} + \frac{1}{T \cdot \mu_\mathrm{m}} \log\left(\frac{TC\mu_m\rho_3}{\rho_1}\right)$$

Explore-First Algorithm

≤OptLoss+ 1 T· μ m 1 1 T· μ m TT· μ m μμ μ m m μ m 1 T· μ m log TC μ m ρ 3 ρ 1 TC μ m ρ 3 ρ 1 TTCC μ m μμ μ m mm μ m ρ 3 ρρ ρ 3 3 ρ 3 TC μ m ρ 3 ρ 1 ρ 1 ρρ ρ 1 1 ρ 1 TC μ m ρ 3 ρ 1 TC μ m ρ 3 ρ 1

OO μ m −1 log (C μ m T) μ m −1 μμ μ m −1 mm μ m −1 −1 μ m −1 log (C μ m T) log log (C μ m T) (CC μ m μμ μ m mm μ m TT) log (C μ m T) μ m −1 log (C μ m T) , then

Stochastic assumptions:

$$M_t \sim \mu \ \textit{i.i.d.} \text{ over } \mathbf{M}, \text{ and } \mu_m := \min_{M \in \mathbf{M}} \mu(M)$$

- Action selection

$$A_t = \begin{cases} 1 & \text{if } t \leq E \qquad \text{(Exploration phase)} \\ 0 & \text{otherwise} \qquad \text{(Exploitation phase)} \end{cases}$$

$$\text{AverageLoss} \leq \text{OptLoss} + \frac{1}{T \cdot \mu_{\mathrm{m}}} \log\left(\frac{TC\mu_m \rho_3}{\rho_1}\right)$$

$$\text{AverageLoss} \leq \text{OptLoss} + \frac{1}{T \cdot \mu_{\mathrm{m}}} \log\left(\frac{TC\mu_m \rho_3}{\rho_1}\right)$$

Forced-Exploration Algorithm

No stochastic assumption (M_t can even be generated **adversarially**!)

- $\geq \eta_2 \geq \cdots \geq \eta_T > 0$
- Algorithm chooses action

$$A_t \sim \mathrm{Bernoulli}(\eta_t)$$

- **Theorem**: If choose $\eta_t = 1/\sqrt{t}$, then

$$\mathrm{AverageLoss} \leq \mathrm{OptLoss} + \frac{1}{\sqrt{T}}(2\rho_1 + C\rho_3)$$

Forced-Exploration Algorithm

Bernoulli $\eta\ t \quad \eta\ t\ \eta\eta\ \eta\ t\ tt\ \eta\ t \quad \eta\ t$

η 1 ≥ η 2 $\eta\eta$ η 2 $2\ \eta\ 2 \geq \ldots \geq \eta\ T\ \eta\eta\ \eta\ T\ TT\ \eta\ T > 0$

No stochastic assumption (M_t can even be generated **adversarially**!)

- Algorithm chooses action

$$A\ t \sim \ t\ t\ t\ \sim \mathrm{Bernoulli}(\eta_t)$$
$$A_t \sim \mathrm{Bernoulli}(\eta_t)$$

- **Theorem**: If choose $\eta_t = 1/\sqrt{t}$, then

$$\mathrm{AverageLoss} \leq \mathrm{OptLoss} + \frac{1}{\sqrt{T}}(2\rho_1 + C\rho_3)$$

Forced-Exploration Algorithm

OptLoss+ 1 T 1 1 T T TT T 1 T 2 ρ 1 +C ρ 3 2 ρ 1 ρρ ρ 1 1 ρ 1 +CC ρ 3 ρρ ρ 3 3 ρ 3 2 ρ 1 +C ρ 3

1/ t t tt t , then

Bernoulli *η t η t ηη η t tt η t η t*

η 1 ≥ *η* 2 *ηη η* 2 *2 η 2 ≥…≥ η T ηη η T TT η T >0*

No stochastic assumption (M_t can even be generated **adversarially**!)

- Algorithm chooses action

$$A\ t \sim\ t\ t\ t\ \sim \mathrm{Bernoulli}(\eta_t)$$

$$\mathrm{AverageLoss} \leq \mathrm{OptLoss} + \frac{1}{\sqrt{T}}(2\rho_1 + C\rho_3)$$

- **<u>Theorem</u>**: If choose $\eta_t = 1/\sqrt{t}$, then

$$\mathrm{AverageLoss} \leq \mathrm{OptLoss} + \frac{1}{\sqrt{T}}(2\rho_1 + C\rho_3)$$

Forced-Exploration Algorithm

OptLoss+ 1 T 1 1 T T T TT T 1 T 2 ρ 1 +C ρ 3 2 ρ 1 ρρ ρ 1 1 ρ 1 +CC ρ 3 ρρ ρ 3 3 ρ 3 2 ρ 1 +C ρ 3

1/ t t tt t , then

Bernoulli *η t η t ηη η t tt η t η t*

η 1 ≥ *η* 2 *ηη η* 2 *2 η 2 ≥…≥ η T ηη η T TT η T >0*

No stochastic assumption (M_t can even be generated **adversarially**!)

- Algorithm chooses action

$$A\ t \sim\ t\ t\ t\ \sim \mathrm{Bernoulli}(\eta_t)$$

$$\mathrm{AverageLoss} \leq \mathrm{OptLoss} + \frac{1}{\sqrt{T}}(2\rho_1 + C\rho_3)$$

- **Theorem**: If choose $\eta_t = 1/\sqrt{t}$, then

$$\mathrm{AverageLoss} \leq \mathrm{OptLoss} + \underbrace{\frac{1}{\sqrt{T}}(2\rho_1 + C\rho_3)}$$

We have an $\Omega\left(\frac{1}{\sqrt{T}}\right)$ lower bound

➔ **Forced-Exploration is essentially optimal**

A Lifelong RL Algorithm based on FE

Input: S, A, γ

Initia

A Lifelong RL Algorithm based on FE

Input: S, A, γ

Initia

A Lifelong RL Algorithm based on FE

Input: S, A, γ

Initia

A Lifelong RL Algorithm based on FE

Input: S, A, γ

Initia

$= \ \ t \ t \ t \ = 1$ if data shows M_t is novel

Sample Complexity of Exploration

Sample complexity of algorithm **A** (given ϵ) [Kakade]

Number of steps where $Q^{\mathbf{A}_t}(s_t, a_t) \leq Q^*(s_t, a_t) - \epsilon$

Measures number of ϵ-mistakes made by the algorithm

- ___ long enough, with high prob.

$$\text{SampleComplexity(Our Algorithm)} = \tilde{O}\left(\frac{CD}{\Gamma^2}T + SAN\sqrt{T}\right)$$

In contrast, single-task RL's Sample Complexity is $\Omega(SAT)$

Sample Complexity of Exploration

ur Algorithm Our Algorithm $= O\ OO\ O\ \ CD\ \ \Gamma\ 2\ \ T{+}SAN\ \ T\ \ \ CD\ \ \Gamma\ 2\ \ CC$
$DD\ CD\ \ \Gamma\ 2\ \ \ \Gamma\ 2\ \Gamma\ \Gamma\ 2\ 2\ \Gamma\ 2\ \ CD\ \ \Gamma\ 2\ \ TT{+}SSAANN\ \ T\ \ \ T\ TT\ \ T\ \ \ CD\ \ \Gamma\ 2\ \ T$
$+SAN\ \ T$

Sample complexity of algorithm **A** (given ϵ) [Kakade]

Number of steps where $Q^{\mathbf{A}_t}(s_t, a_t) \le Q^*(s_t, a_t) - \epsilon$

Measures number of ϵ-mistakes made by the algorithm

- **<u>Theorem</u>**: For H long enough, with high prob.

$$\text{SampleComplexity}\ O\text{ur Algorithm} = \tilde{O}\left(\frac{CD}{\Gamma^2} T + SAN\sqrt{T}\right)$$

In contrast, single-task RL's Sample Complexity is $\Omega(SAT)$

Sample Complexity of Exploration

ur Algorithm Our Algorithm $= O\ OO\ O\ \ CD\ \ \Gamma\ 2\ \ T{+}SAN\ \ T\ \ \ CD\ \ \Gamma\ 2\ \ CC$
$DD\ CD\ \ \Gamma\ 2\ \ \ \Gamma\ 2\ \Gamma\ \Gamma\ 2\ 2\ \Gamma\ 2\ \ CD\ \ \Gamma\ 2\ \ TT{+}SSAANN\ \ T\ \ \ T\ TT\ \ T\ \ \ CD\ \ \Gamma\ 2\ \ T$
$+SAN\ \ T$

Sample complexity of algorithm **A** (given ϵ) [Kakade]

Number of steps where $Q^{\mathbf{A}_t}(s_t, a_t) \leq Q^*(s_t, a_t) - \epsilon$

Measures number of ϵ-mistakes made by the algorithm

- **<u>Theorem</u>**: For H long enough, with high prob.

$$\text{SampleComplexity}\ O\text{ur Algorithm} = \tilde{O}\left(\frac{CD}{\Gamma^2}T + SAN\sqrt{T}\right)$$

Asymptotic performance

In contrast, single-task RL's Sample Complexity is $\Omega(SAT)$

Sample Complexity of Exploration

$SAT\ SSAATT\ SAT$

ur Algorithm Our Algorithm $= O\ OO\ O\ \ CD\ \Gamma\ 2\ \ T+SAN\ \ T\ \ CD\ \Gamma\ 2$
$CCDD\ CD\ \Gamma\ 2\ \ \Gamma\ 2\ \Gamma\ \Gamma\ 2\ 2\ \Gamma\ 2\ \ CD\ \Gamma\ 2\ \ TT+SSAANN\ \ T\ \ T\ TT\ \ T\ \ C$
$D\ \Gamma\ 2\ \ T+SAN\ \ T$

Sample complexity of algorithm **A** (given ϵ) [Kakade]

Number of steps where $Q^{\mathbf{A}_t}(s_t, a_t) \leq Q^*(s_t, a_t) - \epsilon$

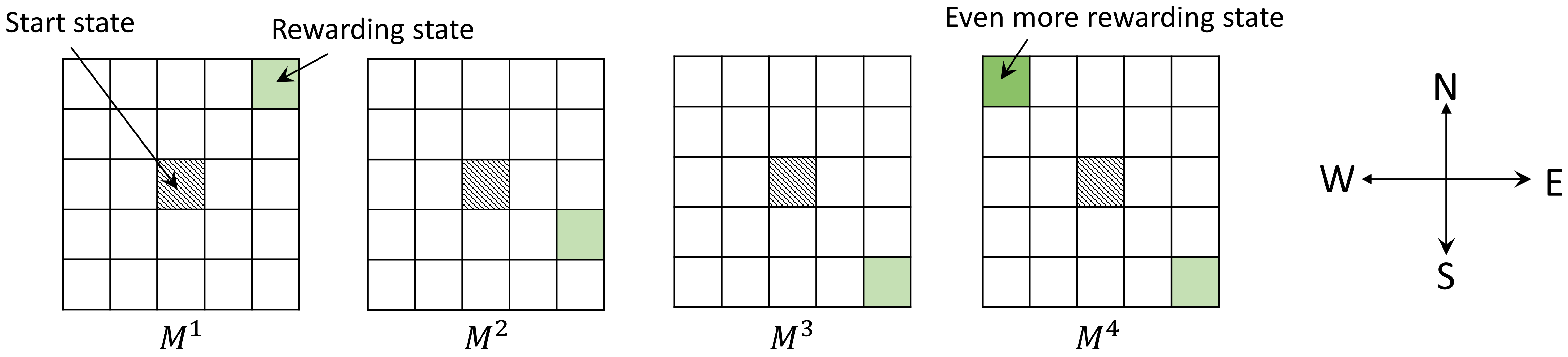
Measures number of ϵ-mistakes made by the algorithm

Asymptotic performance

- **<u>Theorem</u>**: For H long enough, with high prob.

$$\text{SampleComplexity}\ O\text{ur Algorithm} = \tilde{O}\left(\frac{CD}{\Gamma^2}T + SAN\sqrt{T}\right)$$

Experiment



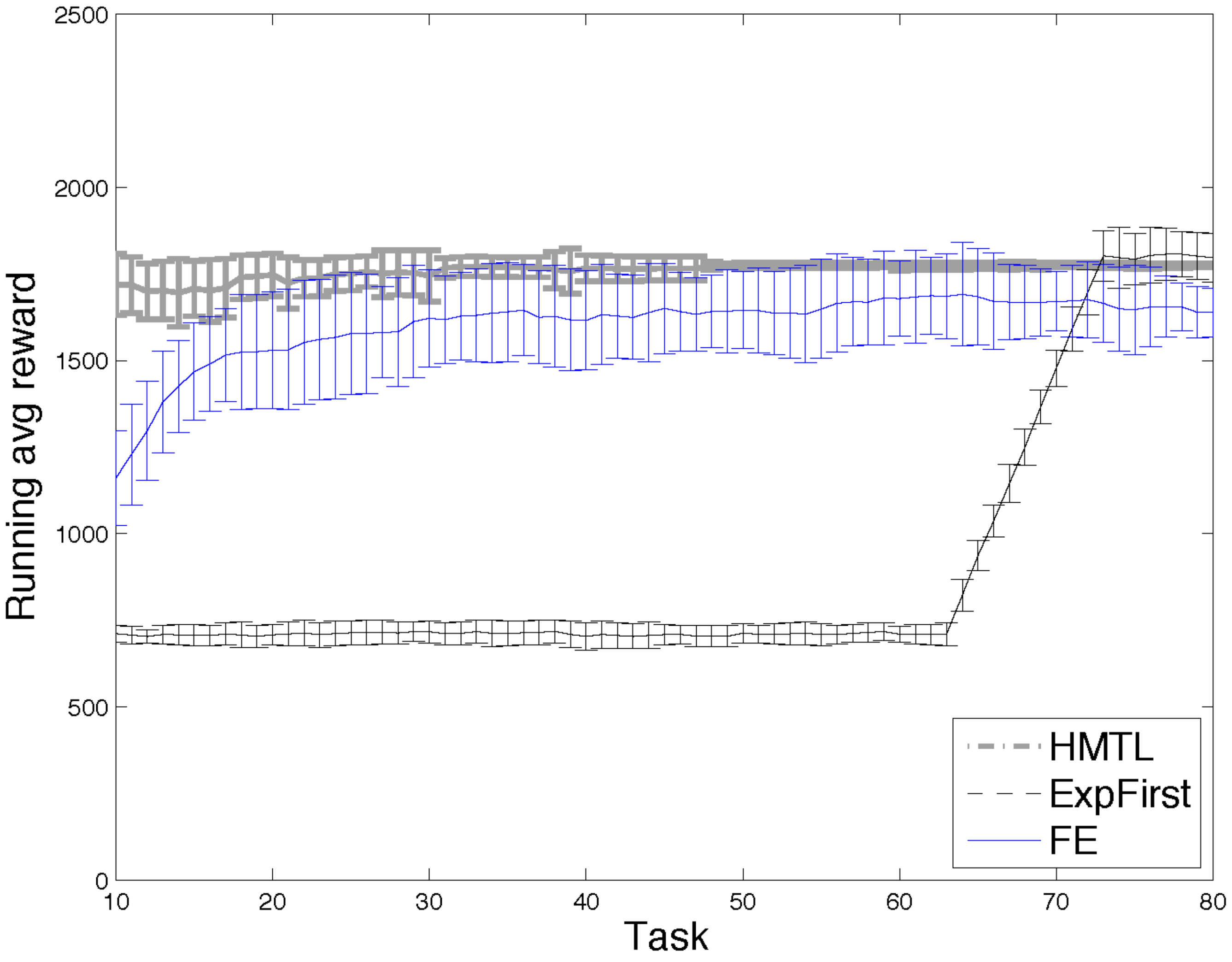
4 possible MDPs with

- noisy state transitions
- different rewarding states

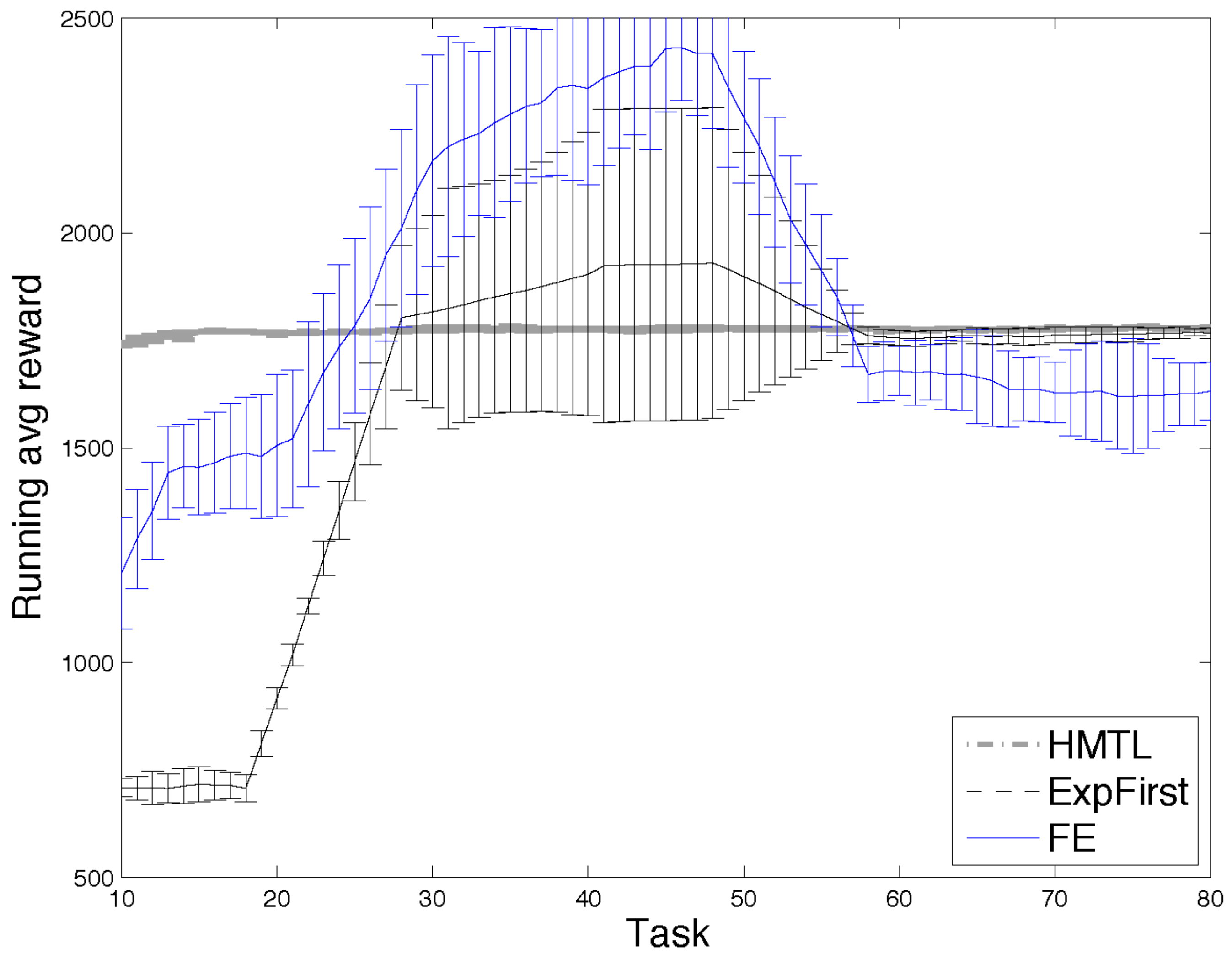
Algorithms for comparison

- Forced-exploration [this work]
- Explore-first [Brunskill-Li]
- Hierarchical Multi-task Learning [Wilson et al.]

Stochastic Setting with small μ_m



Adversarial Setting with Changing Distribution



Conclusions

- Two kinds of exploration needed in LLRL
- Online discovery problem as abstraction for cross-task exploration
- A new lifelong RL algorithm based on optimal ODP algorithm
  - Provably sample complexity better than single-task RL
  - Proof-of-concept experiments demonstrating desired behavior

Future work

- Function approximation
- Use of prior information