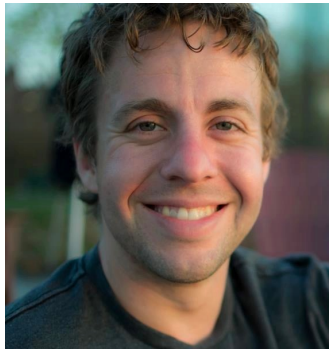
Towards Tight Bounds on the Sample Complexity of Average-reward MDPs

Yujia Jin, Aaron Sidford



ICML 2021, online

Markov Decision Process (MDP)

MDP – modeling 'uncertainty'

given a state, have different actions that leads to different transitions and thus different rewards for each step.

MDP - basic RL model, simple in theory

Elements: state space $s \in S$,

distinct states

action space a $\in A$,

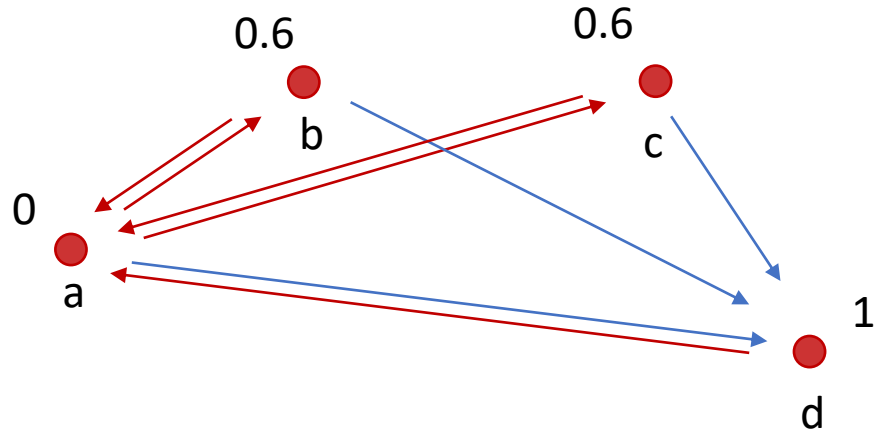
two actions per state

probability transition $p_{s,a} \in P$

evenly distributed through same-color edges

reward $r_{s,a,s'} \in R$.

depending on s' solely



Markov Decision Process (MDP)

MDP – modeling 'uncertainty'

given a state, have different actions that leads to different transitions and thus different rewards for each step.

MDP - reward

There are two classical types of infinite-horizon MDP:

γ **-discounted MDP (DMDP):** discount rewards in future steps

Average-reward MDP (AMDP): don't discount, consider average

Given policy π, induce transition probability P^π, reward r^π

$$\text{value} = \begin{cases} \mathbb{E}^\pi_{\gamma,q}\left[\sum_{t\geq 1}\gamma^{t-1} r_{s_t,a_t} | s_1 \sim q\right] & \gamma\text{-discounted} \\ \lim_{T\to\infty}\frac{1}{T}\mathbb{E}^\pi_q\left[\sum_{t\in[T]} r_{s_t,a_t} | s_1 \sim q\right] & \text{average reward} \end{cases}$$

GOAL: find π that maximizes the value.

MDP - basic RL model, simple in theory

Elements: state space $s \in S$,

distinct states

action space a $\in A$,

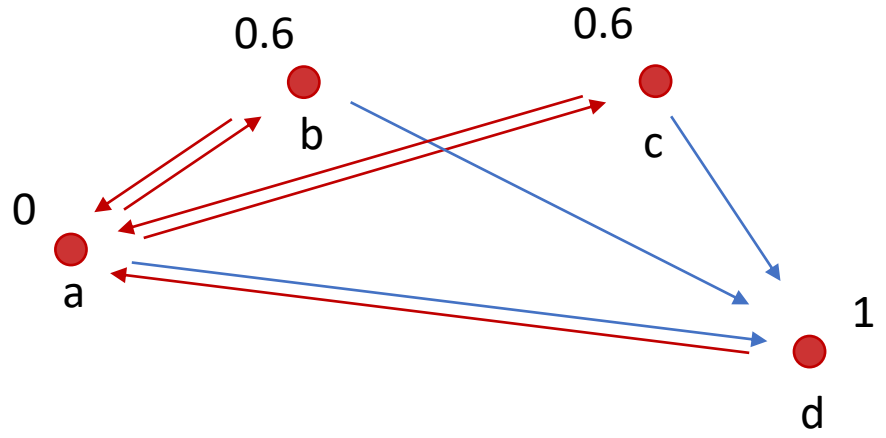
two actions per state

probability transition $p_{s,a} \in P$

evenly distributed through same-color edges

reward $r_{s,a,s'} \in R$.

depending on s' solely



Markov Decision Process (MDP)

MDP
state space $s \in S$,
action space a $\in A$,
probability transition $p_{s,a} \in P$,
reward $r_{s,a} \in R$.

MDP – modeling 'uncertainty'

given a state, have different actions that leads to different transitions and thus different rewards for each step.

MDP - reward

There are two classical types of infinite-horizon MDP:
γ **-discounted MDP (DMDP):** discount rewards in future steps
Average-reward MDP (AMDP): don't discount, consider average

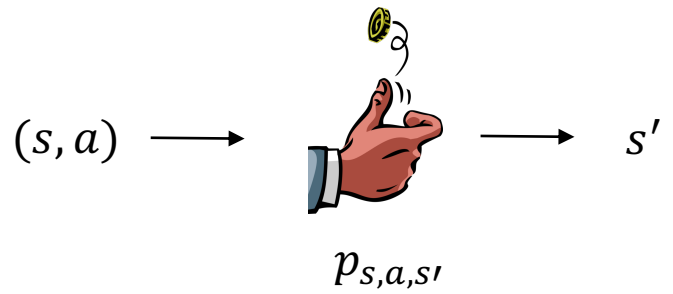
Given policy π, induce transition probability P^π, reward r^π

$$\text{value} = \begin{cases} \mathbb{E}^\pi_{\gamma,q}\left[\sum_{t\geq 1}\gamma^{t-1} r_{s_t,a_t} | s_1 \sim q\right] & \gamma\text{-discounted} \\ \lim_{T\to\infty}\frac{1}{T}\mathbb{E}^\pi_q\left[\sum_{t\in[T]} r_{s_t,a_t} | s_1 \sim q\right] & \text{average reward} \end{cases}$$

GOAL: find π that maximizes the value.

MDP – generative model access

KEY MEASUREMENT: sample complexity
i.e. how many samples to collect in order to find such an approximately optimal policy?



Solving AMDPs (prior art)

MDP
A_{tot}: total state action pair,
γ: discount factor in DMDP
t_{mix}: upper bound on mixing time of probability transition under any given policy , i.e. $(P^{\pi})^{t_{mix}} \approx 1 \cdot (\nu^{\pi})^{T}$, for stationary ν^{π}

Problem: Given an MDP, find an ϵ-approximately optimal policy assuming a generative model access.

	Upper Bound	Lower Bound
γ-discounted	$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [Li20]	$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [Azar13]

[Li20] Li, Gen, et al. "Breaking the sample size barrier in model-based reinforcement learning with a generative model." Advances in Neural Information Processing Systems 33 (2020).
[Azar13] Azar, Mohammad Gheshlaghi, Rémi Munos, and Hilbert J. Kappen. "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model." *Machine learning* 91.3 (2013): 325-349.
[Jin20] Jin, Yujia, and Aaron Sidford. "Efficiently Solving MDPs with Stochastic Mirror Descent." International Conference on Machine Learning. PMLR, 2020.

Solving AMDPs (prior art)

MDP
A_{tot}: total state action pair,
γ: discount factor in DMDP
t_{mix}: upper bound on mixing time of probability transition under any given policy , i.e. $(P^{\pi})^{t_{mix}} \approx 1 \cdot (\nu^{\pi})^{T}$, for stationary ν^{π}

Problem: Given an MDP, find an ϵ- approximately optimal policy assuming a generative model access.

	Upper Bound	Lower Bound
γ-discounted	$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [Li20]	$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [Azar13]
average-reward	$\frac{A_{\text{tot}}t_{\text{mix}}^2}{\epsilon^2}$ [Jin20]	??

[Li20] Li, Gen, et al. "Breaking the sample size barrier in model-based reinforcement learning with a generative model." Advances in Neural Information Processing Systems 33 (2020).
[Azar13] Azar, Mohammad Gheshlaghi, Rémi Munos, and Hilbert J. Kappen. "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model." *Machine learning* 91.3 (2013): 325-349.
[Jin20] Jin, Yujia, and Aaron Sidford. "Efficiently Solving MDPs with Stochastic Mirror Descent." International Conference on Machine Learning. PMLR, 2020.

Solving AMDPs (our results)

MDP

A_{tot}: total state action pair,
γ: discount factor in DMDP
t_{mix}: upper bound on mixing time of probability transition under any given policy , i.e. $(P^{\pi})^{t_{mix}} \approx 1 \cdot (\nu^{\pi})^{T}$, for stationary ν^{π}

Problem: Given an MDP, find an ϵ- approximately optimal policy assuming a generative model access.

	Upper Bound		Lower Bound
average-reward	$\frac{A_{\text{tot}} t_{\text{mix}}^2}{\epsilon^2}$ [Jin20]	$\frac{A_{\text{tot}} t_{\text{mix}}}{\epsilon^3}$	$\frac{A_{\text{tot}} t_{\text{mix}}}{\epsilon^2}$

Contributions:

First characterization of the problem hardness w.r.t mixing time t_{mix}

Solving AMDPs (our results)

MDP
A_{tot}: total state action pair,
γ: discount factor in DMDP
t_{mix}: upper bound on mixing time of probability transition under any given policy , i.e. $(P^{\pi})^{t_{mix}} \approx 1 \cdot (\nu^{\pi})^{T}$, for stationary ν^{π}

Problem: Given an MDP, find an ϵ- approximately optimal policy assuming a generative model access.

	Upper Bound		Lower Bound
average-reward	$\frac{A_{\text{tot}} t_{\text{mix}}^2}{\epsilon^2}$ [Jin20]	$\frac{A_{\text{tot}} t_{\text{mix}}}{\epsilon^3}$	$\frac{A_{\text{tot}} t_{\text{mix}}}{\epsilon^2}$

Contributions:

1. Match lower bound nearly-tightly for constant accuracy ϵ
2. In practice, t_{mix} usually scales with S, while ϵ might not
3. Use deterministic oblivious samples, enables parallel computing
4. Weaker mixing assumption: deterministic policy vs. randomized policy
5. (our technique) Build direct connection with DMDPs

Techniques: upper bound

MDP
state space $s \in S$, action space a $\in A$,
probability transition $p_{s,a} \in P$, reward $r_{s,a} \in R$,
policy π, probability transition P^{π},
reward vector r^{π}, unique stationary distribution ν^{π}

Problem: Given an MDP, find an eps-approximately optimal policy assuming a generative model access.

$$\text{Recall value} \quad \begin{cases} V^{\pi}_{\gamma,q} = \sum_{t \geq 0} \gamma^t q^{\top} (P^{\pi})^t r^{\pi} \in [0, 1/(1-\gamma)] & \gamma\text{-discounted} \\ V^{\pi} = \langle r^{\pi}, \nu^{\pi} \rangle \in [0, 1] & \text{average reward} \end{cases}$$

Techniques: upper bound

MDP
state space $s \in S$, action space a $\in A$,
probability transition $p_{s,a} \in P$, reward $r_{s,a} \in R$,
policy π, probability transition P^{π},
reward vector r^{π}, unique stationary distribution ν^{π}

Problem: Given an MDP, find an eps-approximately optimal policy assuming a generative model access.

$$\text{Recall value} \quad \begin{cases} V^{\pi}_{\gamma,q} = \sum_{t \geq 0} \gamma^t q^{\top} (P^{\pi})^t r^{\pi} \in [0, 1/(1-\gamma)] & \gamma\text{-discounted} \\ V^{\pi} = \langle r^{\pi}, \nu^{\pi} \rangle \in [0,1] & \text{average reward} \end{cases}$$

KEY Lemma : $|V^{\pi} - (1-\gamma) V^{\pi}_{\gamma,q}| \leq 3(1-\gamma) t_{\text{mix}}$

Techniques: upper bound

MDP
state space $s \in S$, action space a $\in A$,
probability transition $p_{s,a} \in P$, reward $r_{s,a} \in R$,
policy π, probability transition P^{π},
reward vector r^{π}, unique stationary distribution ν^{π}

Problem: Given an MDP, find an eps-approximately optimal policy assuming a generative model access.

$$\text{Recall value} \begin{cases} V_{\gamma,q}^{\pi} = \sum_{t \geq 0} \gamma^t q^{\top} (P^{\pi})^t r^{\pi} \in [0, 1/(1-\gamma)] & \gamma\text{-discounted} \\ V^{\pi} = \langle r^{\pi}, \nu^{\pi} \rangle \in [0,1] & \text{average reward} \end{cases}$$

KEY Lemma : $|V^{\pi} - (1-\gamma) V_{\gamma,q}^{\pi}| \leq 3(1-\gamma) t_{\text{mix}}$

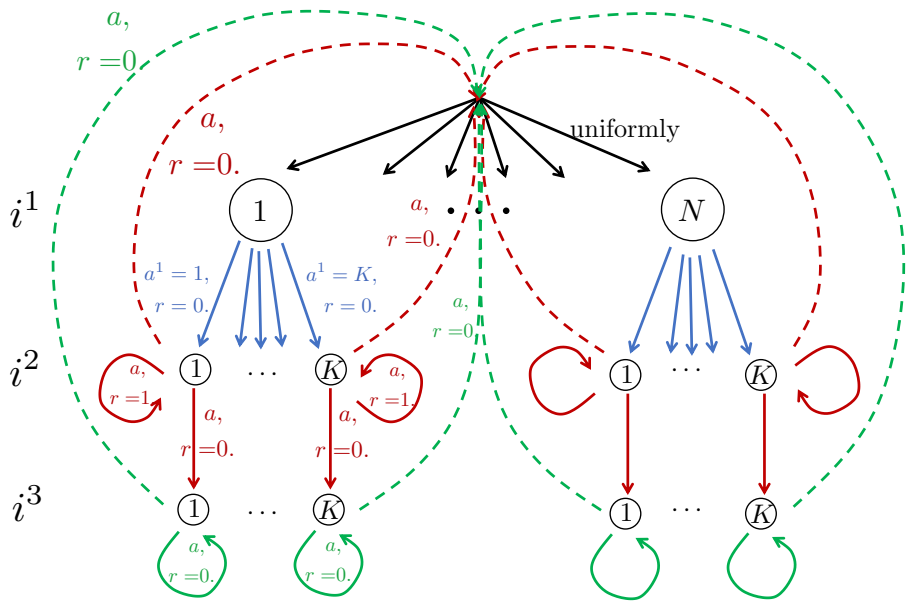
Implications: reduce solving AMDPs to solving DMDPs with $\gamma = 1 - \Theta(\epsilon / t_{\text{mix}})$ to accuracy $\varepsilon = \Theta(\frac{\epsilon}{1-\gamma})$

SOTA DMDP solver $\Longrightarrow$ $\frac{SA}{(1-\gamma)^3 \varepsilon^2}$ Choice of γ, ε $\Longrightarrow$ $\frac{SAt_{\text{mix}}}{\epsilon^3}$

Techniques: lower bound

MDP
state space $s \in S$, action space a $\in A$,
probability transition $p_{s,a} \in P$, reward $r_{s,a} \in R$,
policy π, probability transition P^{π},
reward vector r^{π}, unique stationary distribution ν^{π}

Hard Instance Construction: modified from the hard instance in DMDP [Azar13]



$p(\circlearrowright) = \gamma p_{(i^1,a^1)},$
$p(\downarrow) = \gamma(1 - p_{(i^1,a^1)}),$
$p(\searrow) = 1 - \gamma.$

$p(\circlearrowright) = \gamma,$
$p(\searrow) = 1 - \gamma.$

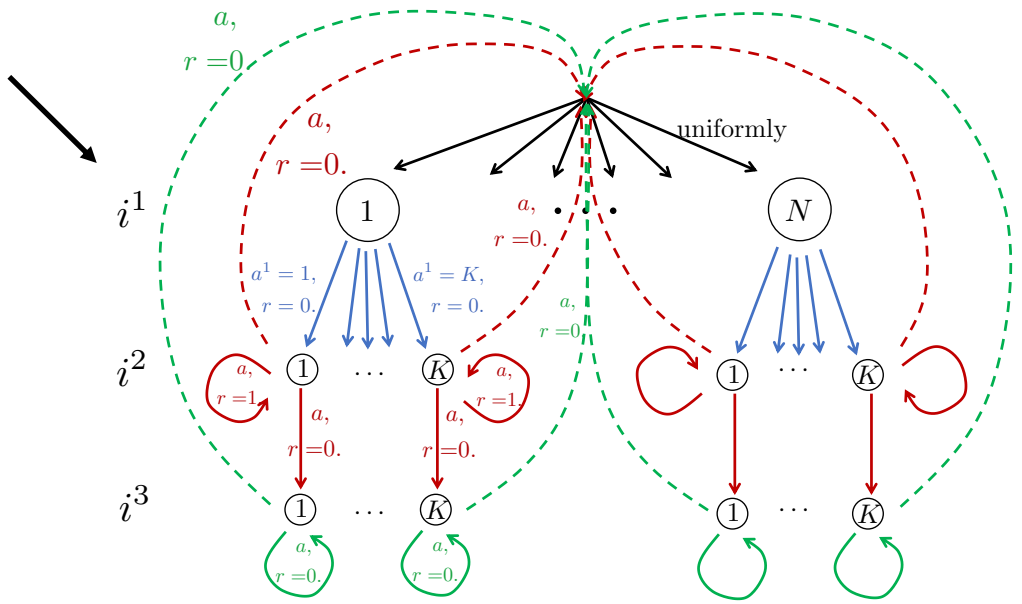
[Azar13] Azar, Mohammad Gheshlaghi, Rémi Munos, and Hilbert J. Kappen. "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model." *Machine learning* 91.3 (2013): 325-349.

Techniques: lower bound

MDP
state space $s \in S$, action space a $\in A$,
probability transition $p_{s,a} \in P$, reward $r_{s,a} \in R$,
policy π, probability transition P^{π},
reward vector r^{π}, unique stationary distribution ν^{π}

Hard Instance Construction: modified from the hard instance in DMDP [Azar13]

level 1: N states, each has K actions that transit to different states at level 2



$p(\circlearrowright) = \gamma p_{(i^1,a^1)},$
$p(\downarrow) = \gamma(1 - p_{(i^1,a^1)}),$
$p(\searrow) = 1 - \gamma.$

$p(\circlearrowright) = \gamma,$
$p(\searrow) = 1 - \gamma.$

[Azar13] Azar, Mohammad Gheshlaghi, Rémi Munos, and Hilbert J. Kappen. "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model." *Machine learning* 91.3 (2013): 325-349.

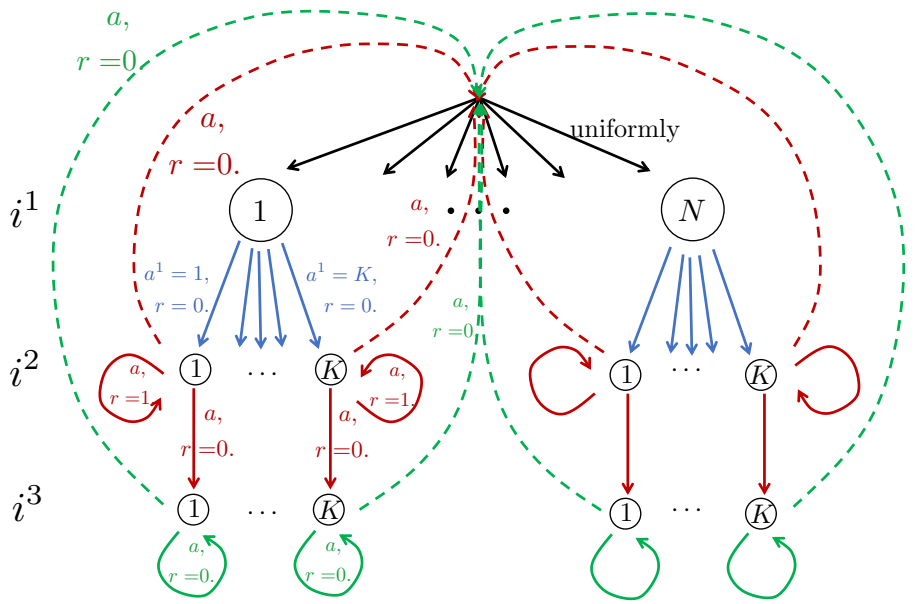
Techniques: lower bound

MDP
state space $s \in S$, action space a $\in A$,
probability transition $p_{s,a} \in P$, reward $r_{s,a} \in R$,
policy π, probability transition P^{π},
reward vector r^{π}, unique stationary distribution v^{π}

Hard Instance Construction: modified from the hard instance in DMDP [Azar13]

level 2: each state has $1-\gamma$ refreshing probability,
γp probability staying,
$\gamma(1-p)$ probability to level 3

level 3: each state has $1-\gamma$ refreshing probability,
γ probability staying



$p(\circlearrowright) = \gamma p_{(i^1,a^1)}$,
$p(\downarrow) = \gamma(1-p_{(i^1,a^1)})$,
$p(\searrow) = 1-\gamma$.

$p(\circlearrowright) = \gamma$,
$p(\searrow) = 1-\gamma$.

[Azar13] Azar, Mohammad Gheshlaghi, Rémi Munos, and Hilbert J. Kappen. "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model." *Machine learning* 91.3 (2013): 325-349.

Techniques: lower bound

MDP
state space $s \in S$, action space a $\in A$,
probability transition $p_{s,a} \in P$, reward $r_{s,a} \in R$,
policy π, probability transition P^{π},
reward vector r^{π}, unique stationary distribution v^{π}

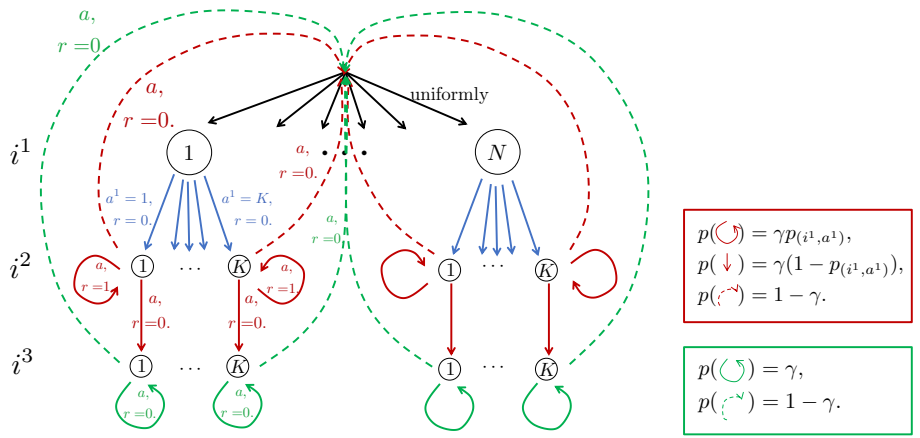
Hard Instance Construction: modified from the hard instance in DMDP [Azar13]

KEY structure:

level 1: N states, each has K actions that transit to different states at level 2

only transition with instant reward 1
—> GOAL: hide a best p* among p in each K actions

level 2: each state has $1-\gamma$ refreshing probability,
γp probability staying,
$\gamma(1-p)$ probability to level 3



level 3: each state has $1-\gamma$ refreshing probability,
γ probability staying

[Azar13] Azar, Mohammad Gheshlaghi, Rémi Munos, and Hilbert J. Kappen. "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model." *Machine learning* 91.3 (2013): 325-349.

Techniques: lower bound

MDP
state space $s \in S$, action space a $\in A$,
probability transition $p_{s,a} \in P$, reward $r_{s,a} \in R$,
policy π, probability transition P^{π},
reward vector r^{π}, unique stationary distribution v^{π}

Hard Instance Construction: modified from the hard instance in DMDP [Azar13]

KEY structure:

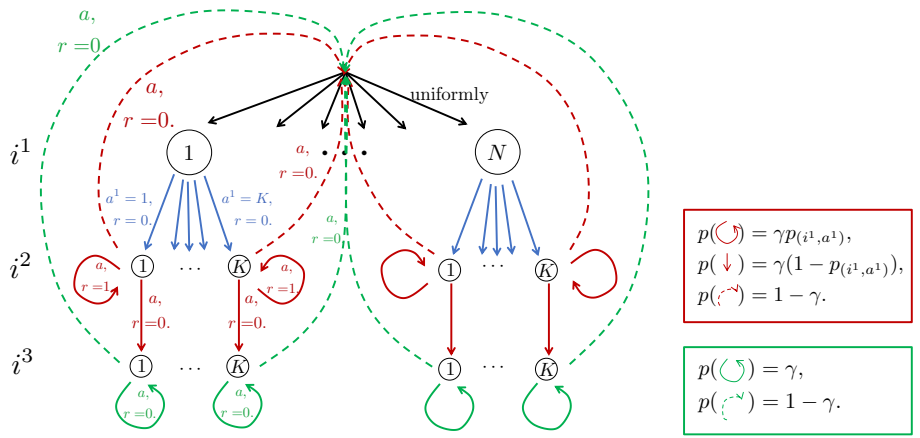
level 1: N states, each has K actions that transit to different states at level 2

only transition with instant reward 1
—> GOAL: hide a best p* among p in each K actions

level 2: each state has $1-\gamma$ refreshing probability,
γp probability staying,
$\gamma(1-p)$ probability to level 3

refreshing probability
—> mixing time $O(1/(1-\gamma))$

level 3: each state has $1-\gamma$ refreshing probability,
γ probability staying



[Azar13] Azar, Mohammad Gheshlaghi, Rémi Munos, and Hilbert J. Kappen. "Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model." *Machine learning* 91.3 (2013): 325-349.

Techniques: lower bound

MDP
state space $s \in S$, action space a $\in A$,
probability transition $p_{s,a} \in P$, reward $r_{s,a} \in R$,
policy π, probability transition P^{π},
reward vector r^{π}, unique stationary distribution ν^{π}

Hard Instance Construction: modified from the hard instance in DMDP [Azar13]

KEY structure:

only action with instant reward 1
—> GOAL: hide a best p* among p in each K actions

$$p = \gamma, p^{\star} = \gamma + \epsilon(1-\gamma)$$

Multi-armed bandit lower bound for "best-arm identification"

$$\Longrightarrow \quad K \cdot \frac{1}{(1-\gamma)\epsilon^2} \text{ samples to find one best action}$$

Techniques: lower bound

MDP
state space $s \in S$, action space a $\in A$,
probability transition $p_{s,a} \in P$, reward $r_{s,a} \in R$,
policy π, probability transition P^{π},
reward vector r^{π}, unique stationary distribution ν^{π}

Hard Instance Construction: modified from the hard instance in DMDP [Azar13]

KEY structure:

only action with instant reward 1
—> GOAL: hide a best p* among p in each K actions

$$p = \gamma, p^{\star} = \gamma + \epsilon(1-\gamma)$$

Multi-armed bandit lower bound
for "best-arm identification"

$$\Longrightarrow \quad K \cdot \frac{1}{(1-\gamma)\epsilon^2} \text{ samples to find one best action}$$

refreshing probability
—> mixing time $O(1/(1-\gamma))$

$$\Longrightarrow \quad \Omega(K \cdot \frac{t_{\text{mix}}}{\epsilon^2}) \text{ samples to find one best action}$$

$$\text{find } \epsilon \text{ -optimal policy} \Longrightarrow \text{find } \Omega(N) \text{ best actions for level-1 states} \Longrightarrow \text{require samples } \Omega(NK\frac{t_{\text{mix}}}{\epsilon^2})$$

Solving AMDPs (prior art)

MDP

A_{tot}: total state action pair,
γ: discount factor in DMDP
t_{mix}: upper bound on mixing time of probability transition under any given policy , i.e. $(P^{\pi})^{t_{mix}} \approx 1 \cdot (\nu^{\pi})^{T}$, for stationary ν^{π}

Problem: Given an MDP, find an ϵ- approximately optimal policy assuming a generative model access.

	Upper Bound		Lower Bound	
γ-discounted	$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [Li20]		$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [Azar13]	
average-reward	$\frac{A_{\text{tot}}t_{\text{mix}}^2}{\epsilon^2}$ [Jin20]	$\frac{A_{\text{tot}}t_{\text{mix}}}{\epsilon^3}$	??	$\frac{A_{\text{tot}}t_{\text{mix}}}{\epsilon^2}$

Solving AMDPs (prior art)

MDP
A_{tot}: total state action pair,
γ: discount factor in DMDP
t_{mix}: upper bound on mixing time of probability transition under any given policy , i.e. $(P^{\pi})^{t_{mix}} \approx 1 \cdot (\nu^{\pi})^{T}$, for stationary ν^{π}

Problem: Given an MDP, find an ϵ- approximately optimal policy assuming a generative model access.

	Upper Bound		Lower Bound	
γ-discounted	$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [Li20]		$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [Azar13]	
average-reward	$\frac{A_{\text{tot}}t_{\text{mix}}^2}{\epsilon^2}$ [Jin20]	$\frac{A_{\text{tot}}t_{\text{mix}}}{\epsilon^3}$	??	$\frac{A_{\text{tot}}t_{\text{mix}}}{\epsilon^2}$

OPEN PROBLEMs: Can we obtain tight upper bounds?
Can we characterize AMDP hardness with measurements other than mixing time t_{mix}?

Thank You

Towards Tight Bounds on the Sample Complexity of Average-reward MDPs

Yujia Jin, Aaron Sidford