Optimizing Average Reward MDPs with a Generative Model

Aaron Sidford

Stanford University
Departments of Management Science
& Engineering and Computer Science

Contact Info:
- email: *sidford@stanford.edu*
- website: *www.aaronsidford.com*

Efficiently Solving MDPs with Stochastic Mirror Descent

joint with Yujia Jin (ICML 2020, arXiv: 2008.12776)

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

joint with Yujia Jin (ICML 2021, arXiv: 2106.07046)

Yujia Jin

Thank you for slide material!

This Talk

Part 1
Problem and Results

Part 2
Approach #1

[Jin**S**20]

Part 3
Approach #2

[Jin**S**21]

Part 4
Lower Bound

[Jin**S**21]

Take-aways

- Algorithmic tools for solving MDPs!
- Open problem!

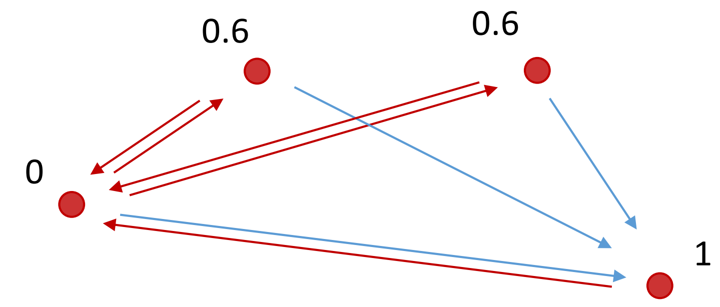
Markov Decision Process (MDPs)

Setup

- **States**: finite set S
- **Actions**: finite set A_s for each $s \in S$
- **Transition probabilities**: $p_{s,a} \in \Delta^S$ for each $s \in S, a \in A_s$
- **Rewards**: $r \in [-1,1]^A$ for $A = \cup_{s \in S} A_s$

Goal: compute an ϵ-optimal policy

- Randomized policy: $\pi(s) \in \Delta^{A_s}$ for all s
- Deterministic policy: $\pi(s) \in A_s$ for all s



What reward function?

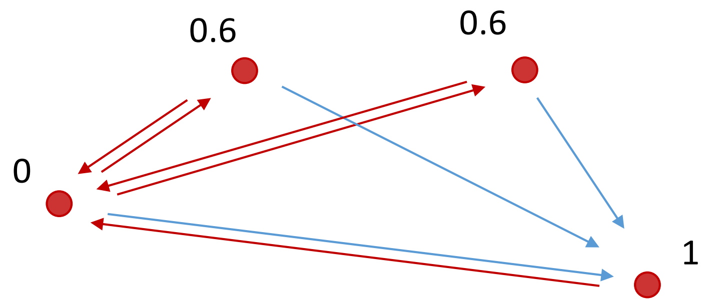
- **Discounted reward (DMDP)**: $\gamma \in (0,1)$ and $q \in \Delta^S$
  $$v_{\gamma,q}^{\pi} \stackrel{\text{def}}{=} \mathbb{E}_{s_t,\pi(s_t)} \sum_t \gamma^t r_{s_t,\pi(s_t)} \text{ for } s_0 \sim q$$
- **Average reward (AMDP)**: $\gamma \to 1$
  $$v^{\pi} \stackrel{\text{def}}{=} \lim_{t \to \infty} \frac{1}{T} \mathbb{E}_{s_t,\pi(s_t)} \sum_{t \in [T]} r_{s_t,\pi(s_t)}$$

Won't always distinguish between the two but may mention open problems.

There are other functions, e.g. finite horizon, see next talk!

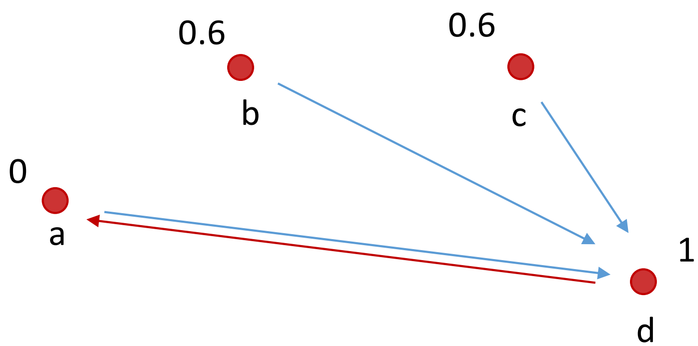
For discounted approximate policies there is a difference between whether reward is for specific q or all $q \in \Delta^S$.

Reward Functions



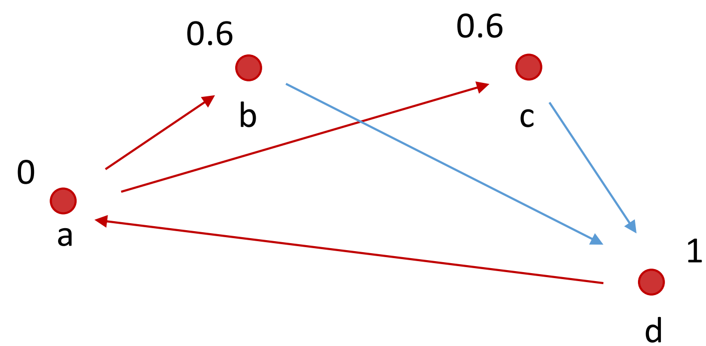
DMDP optimal policy

(small γ)



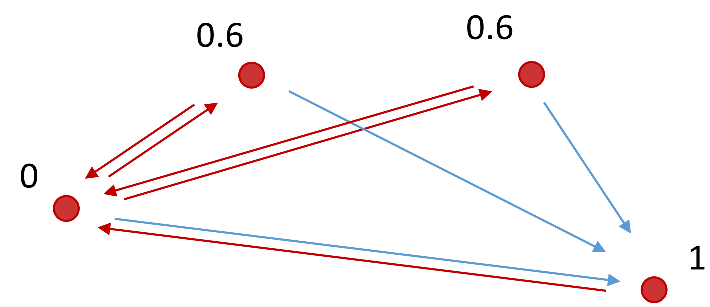
AMDP Optimal Policy

($\gamma \rightarrow 1$)



The Problem

Setup
- **States**: finite set S
- **Rewards**: $r \in [-1,1]^S$
- **Actions**: finite set A_s for each $s \in S$
- **Transition probabilities**: $p_{s,a} \in \Delta^S$ for each $s \in S, a \in A_s$



Goal: compute an ϵ-optimal policy
- Randomized policy: $\pi(s) \in \Delta^{A_s}$ for all s
- Deterministic policy: $\pi(s) \in A_s$ for all s

Reward Function
- Discounted: $v^{\pi}_{\gamma,q} = \mathbb{E}_{s_t,\pi(s_t)} \sum_t \gamma^t r_{s_t,\pi(s_t)}$ for $s_0 \sim q$
- Average: $v^{\pi} = \lim_{t\to\infty} \frac{1}{T} \mathbb{E}_{s_t,\pi(s_t)} \sum_{t\in[T]} r_{s_t,\pi(s_t)}$

Generative Model Sample Complexity
- States, actions, and rewards are known
- Transition probabilities unknown
- Given any s, a can *query* generative model for a sample from $p_{s,a}$
- **Question**: how many samples needed to compute an ϵ-optimal policy?

State-of-the-art

sample complexities

- States S
- A_{tot} total state action pairs
- Discount factor γ
- Max ratio of stationary probability τ
- Largest mixing time of any policy t_{mix}

	Upper Bound		**Lower Bound**
Discounted Reward (DMDP)	$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [AMK13,**S**WWYY18, W19,AKY20,LWCGC20]		$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [AMK13]
Average Reward (AMDP)	$\frac{A_{\text{tot}}t_{\text{mix}}^2\tau^4}{\epsilon^2}$ [W17]		?
Our AMDP Results	$\frac{A_{\text{tot}}t_{\text{mix}}^2}{\epsilon^2}$ [J**S**20]	$\frac{A_{\text{tot}}t_{\text{mix}}}{\epsilon^3}$ [J**S**21]	$\frac{A_{\text{tot}}t_{\text{mix}}}{\epsilon^2}$ [J**S**21]

oblivious samples (refers to the two [J**S**21] results)

Open Problem
What is the optimal sample complexity for $\epsilon = \tilde{O}(1)$?

Optimal sample complexities for $\epsilon = \tilde{\Omega}(1)$!

Ignoring log factors, differences between $|S||A|$ and A_{tot}, the domain of r, etc.

This Talk

Part 1
Problem and Results

Part 2
Approach #1
[Jin**S**20]

Part 3
Approach #2
[Jin**S**21]

Part 4
Lower bound
[Jin**S**21]

Take-aways

- Algorithmic tools for solving MDPs!
- Open problem!

State-of-the-art

sample complexities

- States S
- A_{tot} total state action pairs
- Discount factor γ
- Max ratio of stationary probability τ
- Largest mixing time of any policy t_{mix}

	Upper Bound		**Lower Bound**
Discounted Reward (DMDP)	$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [AMK13,**S**WWYY18, W19,AKY20,LWCGC20]		$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [AMK13]
Average Reward (AMDP)	$\frac{A_{\text{tot}} t_{\text{mix}}^2 \tau^4}{\epsilon^2}$ [W17]		?
Our AMDP Results	$\boldsymbol{\frac{A_{\text{tot}} t_{\text{mix}}^2}{\epsilon^2}}$ [J**S**20]	$\frac{A_{\text{tot}} t_{\text{mix}}}{\epsilon^3}$ [J**S**21]	$\frac{A_{\text{tot}} t_{\text{mix}}}{\epsilon^2}$ [J**S**21]

oblivious samples

Open Problem
What is the optimal sample complexity for $\epsilon = \tilde{O}(1)$?

Optimal sample complexities for $\epsilon = \tilde{\Omega}(1)$!

Ignoring log factors, differences between $|S||A|$ and A_{tot}, the domain of r, etc.

Approach #1: Convex Optimization

MDP
- state space $s \in S$
- actions A_s for $a \in S$ and $A \stackrel{\text{def}}{=} \cup_{s \in S} A_s$
- transition probabilities $p_{s,a} \in \Delta^S$
- rewards $r_{s,a} \in [-1,1]$

Convex formulation

Same / similar to [W17]

$$\min_{v \in t_{\text{mix}} \cdot [-1,1]^S} \max_{\mu \in \Delta^A} \mu^\top [(\boldsymbol{P} - \boldsymbol{E}) v + r]$$

- row $a \in A_s$ of $\boldsymbol{P}$ is $p_{s,a}$
- row $a \in A_s$ of $\boldsymbol{E}$ is e_a
- ℓ_1 norm of each row of $\boldsymbol{P} - \boldsymbol{E}$ is ≤ 2

Box Simplex Game!
[S17,J**S**T19,C**S**T21]

Solver

Related to [W17,CJ**S**T19,CJ**S**T20]

- Stochastic mirror descent with careful local norm analysis
- $\tilde{O}(A_{\text{tot}} t_{\text{mix}}^2 / \epsilon^2)$ steps and $\tilde{O}(1)$ per step
- General result about box simplex games!

Rounding

Similar observation / approach taken in [CCBG20] for DMDP

- Scale μ across each A_s so probability distribution
- Lemma: ϵ-approximate $\mu \Rightarrow O(\epsilon)$-approximate policy
- $\Rightarrow \tilde{O}(A_{\text{tot}} t_{\text{mix}}^2 / \epsilon^2)$ samples to solve an AMDP [J**S**20]

Discussion

Theorem [JS20]: Can compute ϵ-optimal policy to AMDP using $\tilde{O}\left(\frac{A_{\text{tot}} t_{\text{mix}}^2}{\epsilon^2}\right)$ queries

Properties of resulting algorithm

- Queries are dynamic, which state chosen to query depends on algorithm
- Outputs a randomized property

Generalizations

- General sublinear box-simplex solver!
- Recovers $\tilde{O}\left(\frac{A_{\text{tot}}}{(1-\gamma)^4 \epsilon^2}\right)$ sample bound of other convex optimization approach to solving DMDP [CCBG20] (for a fixed initial distribution on vertices)
- Generalizes to constrained MDPs!

Open Problems

- Convex approach completely matching state-of-the-art for DMDPs?

This Talk

Part 1
Problem and Results

Part 2
Approach #1

[Jin**S**20]

Part 3
Approach #2

[Jin**S**21]

Part 4
Lower bound

[Jin**S**21]

Take-aways

- Algorithmic tools for solving MDPs!
- Open problem!

State-of-the-art

sample complexities

- States S
- A_{tot} total state action pairs
- Discount factor γ
- Max ratio of stationary probability τ
- Largest mixing time of any policy t_{mix}

	Upper Bound		Lower Bound	
Discounted Reward (DMDP)	$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [AMK13,**S**WWYY18, W19,AKY20,LWCGC20]		$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [AMK13]	
Average Reward (AMDP)	$\frac{A_{\text{tot}}t_{mix}^2\tau^4}{\epsilon^2}$ [W17]		?	**Open Problem** What is the optimal sample complexity for $\epsilon = \tilde{O}(1)$?
Our AMDP Results	$\frac{A_{\text{tot}}t_{\text{mix}}^2}{\epsilon^2}$ [J**S**20]	$\boldsymbol{\frac{A_{\text{tot}}t_{\text{mix}}}{\epsilon^3}}$ [J**S**21]	$\frac{A_{\text{tot}}t_{\text{mix}}}{\epsilon^2}$ [J**S**21]	Optimal sample complexities for $\epsilon = \tilde{\Omega}(1)$!

oblivious samples

Ignoring log factors, differences between $|S||A|$ and A_{tot}, the domain of r, etc.

Approach #2: Reduction

MDP
- state space $s \in S$
- actions A_s for $a \in S$ and $A \stackrel{\text{def}}{=} \cup_{s \in S} A_s$
- transition probabilities $p_{s,a} \in \Delta^S$
- rewards $r_{s,a} \in [-1,1]$

Problem: Given an MDP, find ϵ-optimal policy given generative model access.

- **Discounted reward (DMDP)**: $\gamma \in (0,1)$ and $q \in \Delta^S$: $v^{\pi}_{\gamma,q} \stackrel{\text{def}}{=} \mathbb{E}_{s_t,\pi(s_t)} \sum_t \gamma^t \, r_{s_t,\pi(s_t)}$ for $s_0 \sim q$
- **Average reward (AMDP)**: $v^{\pi} \stackrel{\text{def}}{=} \lim_{t \to \infty} \frac{1}{T} \mathbb{E}_{s_t,\pi(s_t)} \sum_{t \in [T]} r_{s_t,\pi(s_t)}$

Lemma: $\left| v^{\pi} - (1-\gamma) v^{\pi}_{\gamma,q} \right| \leq 3(1-\gamma) t_{\text{mix}}$ for all $\gamma \in (0,1)$

Implication: suffices to compute $\varepsilon = \Theta\left(\frac{\epsilon}{1-\gamma}\right)$-approximate policy to DMDP with $\gamma = 1 - \Theta\left(\frac{\epsilon}{t_{\text{mix}}}\right)$

DMDP Bound	→	**AMDP Bound**
$\frac{A_{\text{tot}}}{(1-\gamma)^3 \epsilon^2}$		$\frac{A_{\text{tot}} t_{\text{mix}}}{\epsilon^3}$
[AMK13,**S**WWYY18, W19,AKY20,LWCGC20]		[J**S**21]

Discussion

Theorem: any $\varepsilon = \Theta\left(\frac{\epsilon}{1-\gamma}\right)$-optimal policy to DMDP with $\gamma = 1 - \Theta\left(\frac{\epsilon}{t_{\text{mix}}}\right)$ is an ϵ-optimal policy to the AMDP

$\Rightarrow \tilde{O}\left(\frac{A_{\text{tot}} t_{\text{mix}}}{\epsilon^3}\right)$ samples suffice

Note

- To improve on $\tilde{O}\left(A_{\text{tot}} t_{\text{mix}}^2 \epsilon^{-2}\right)$ need to set $\epsilon = \tilde{\Omega}(t_{\text{mix}}^{-1}) = \Theta((1-\gamma)\epsilon^{-1})$
- Consequently, $\varepsilon = \Omega((1-\gamma)^{-1/2})$ and need [LWCGC20] to improve

Properties of Resulting Algorithm

- $\tilde{O}(t_{\text{mix}}\epsilon^{-2})$ *oblivious* samples per state
- Computes deterministic policy
- Only depend on mixing time of deterministic policies

Algorithmic Implication

- Combining with [BLLLS**S**SW21] obtain $\tilde{O}(A_{\text{tot}}|S| + |S|^{2.5})$ time algorithm

This Talk

Part 1
Problem and Results

Part 2
Approach #1

[Jin**S**20]

Part 3
Approach #2

[Jin**S**21]

Part 4
Lower bound

[Jin**S**21]

Take-aways

- Algorithmic tools for solving MDPs!
- Open problem!

State-of-the-art

sample complexities

- States S
- A_{tot} total state action pairs
- Discount factor γ
- Max ratio of stationary probability τ
- Largest mixing time of any policy t_{mix}

	Upper Bound		Lower Bound	
Discounted Reward (DMDP)	$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [AMK13,**S**WWYY18, W19,AKY20,LWCGC20]		$\frac{A_{\text{tot}}}{(1-\gamma)^3\epsilon^2}$ [AMK13]	
Average Reward (AMDP)	$\frac{A_{\text{tot}}t_{\text{mix}}^2\tau^4}{\epsilon^2}$ [W17]		?	**Open Problem** What is the optimal sample complexity for $\epsilon = \tilde{O}(1)$?
Our AMDP Results	$\frac{A_{\text{tot}}t_{\text{mix}}^2}{\epsilon^2}$ [J**S**20]	$\frac{A_{\text{tot}}t_{\text{mix}}}{\epsilon^3}$ [J**S**21]	$\frac{A_{\text{tot}}t_{\text{mix}}}{\epsilon^2}$ [J**S**21]	Optimal sample complexities for $\epsilon = \tilde{\Omega}(1)$!

oblivious samples

Ignoring log factors, differences between $|S||A|$ and A_{tot}, the domain of r, etc.

Lower Bound Approach

Modify the DMDP construction in [AMK13]

Note

- Essentially reducing AMDP lower bound to DMDP lower bound to best arm identification.
- Proved for oblivious queries. Open / TODO: prove for arbitrary dynamic queries.

MDP

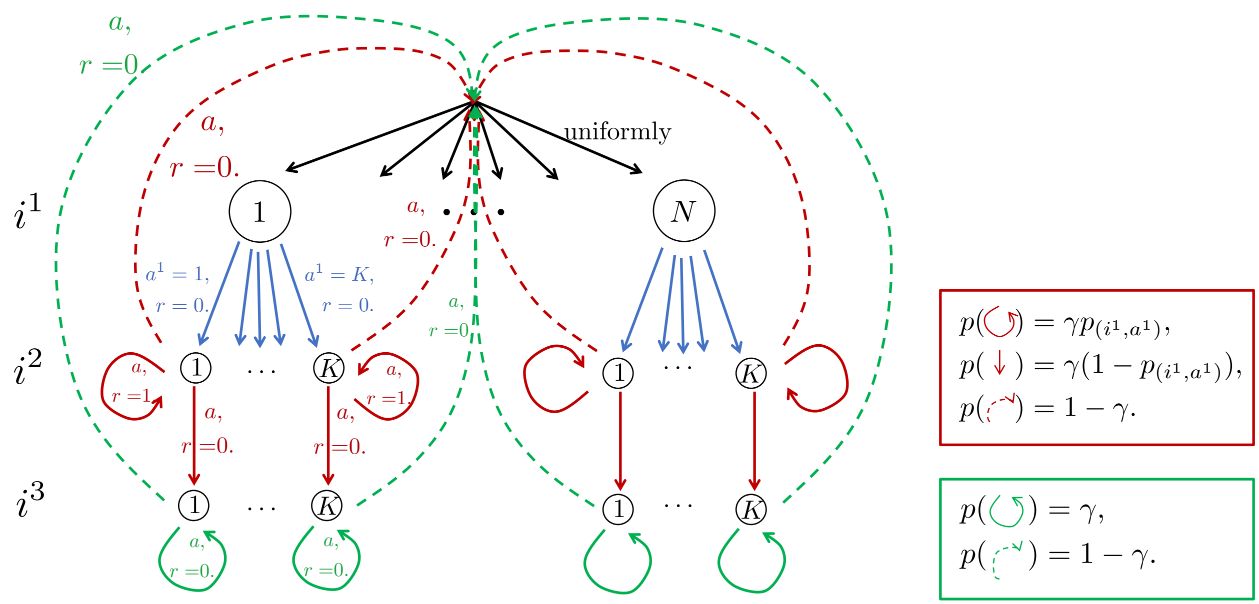
- state space $s \in S$
- actions A_s for $a \in S$ and $A \stackrel{\text{def}}{=} \cup_{s \in S} A_s$
- transition probabilities $p_{s,a} \in \Delta^S$
- rewards $r_{s,a} \in [-1,1]$

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

Level 2: each state s goes uniformly to level 1 with probability $1-\gamma$, stays with probability γp_s, and goes to level 3 with probability $\gamma(1-p_s)$

Level 3: each state goes uniformly to level 1 with probability $1-\gamma$ and stays with probability γ

Rewards: All 0 except at level 1



Lower bound strategy

- Each level 1 state has on action to a level 2 state with a higher γ_s
- Lower bound how many samples need to find enough higher p_s

This Talk

Open Problem
What is the optimal sample complexity for approximately solving MDPs with a generative model?

Part 1
Problem and Results

Part 2
Approach #1
[Jin**S**20]

Part 3
Approach #2
[Jin**S**21]

Part 4
Lower bound
[Jin**S**21]

Take-aways

- Algorithmic tools for solving MDPs!
- Open problem!

Thank you

Questions?

arXiv:2008.12776
arXiv:2106.07046

Yujia Jin

Contact Info:

- email: *sidford@stanford.edu*
- website: *www.aaronsidford.com*