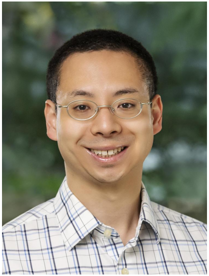
Settling the sample complexity of online reinforcement learning

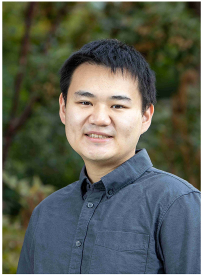
Yuxin Chen

Statistics & Data Science, Wharton, UPenn

Zihan Zhang
Princeton

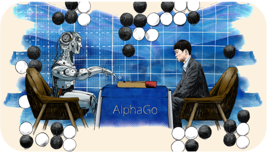


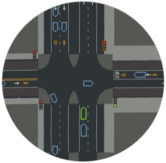
Jason Lee
Princeton

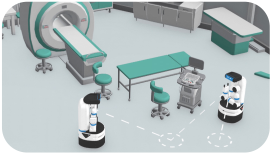


Simon Du
UWashington

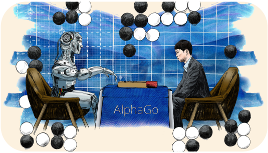
"Settling the sample complexity of online reinforcement learning," Z. Zhang, Y. Chen, J. Lee, S. Du, arXiv:2307.13586, 2023

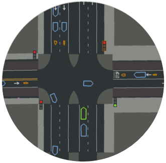


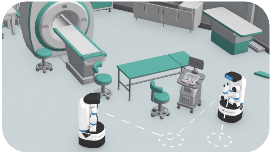




In RL, agent(s) often learn by probing the environment





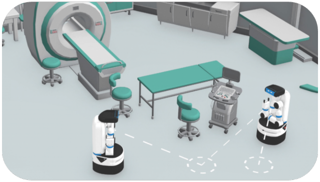


In RL, agent(s) often learn by probing the environment

- unknown environment
- explosion of dimensionality
- delayed feedback
- nonconvexity

Data efficiency

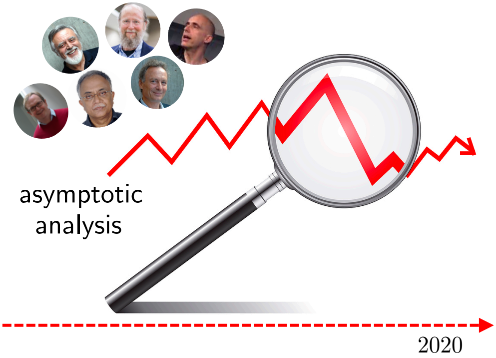
Data collection might be expensive, time-consuming, or high-stakes

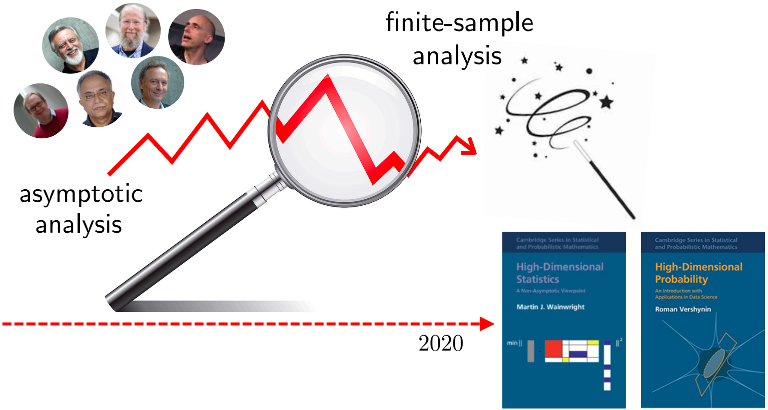


clinical trials

self-driving cars

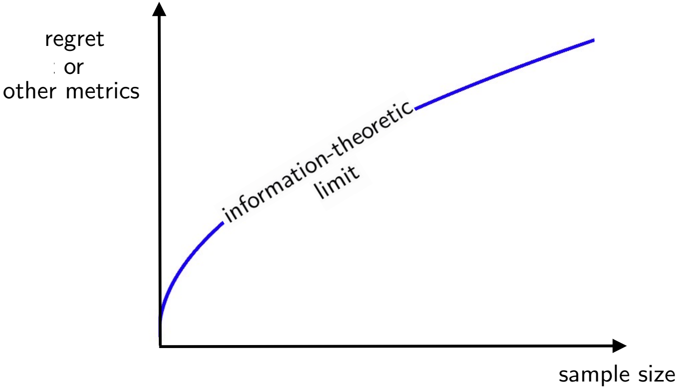
Calls for design of sample-efficient RL algorithms!


asymptotic
analysis
2020

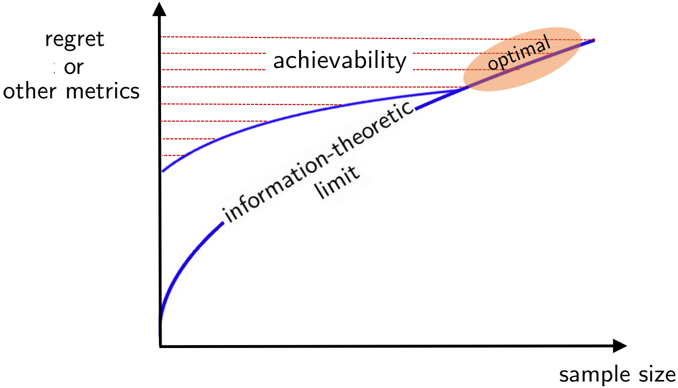


Understanding efficiency of contemporary RL requires a modern suite of non-asymptotic analysis

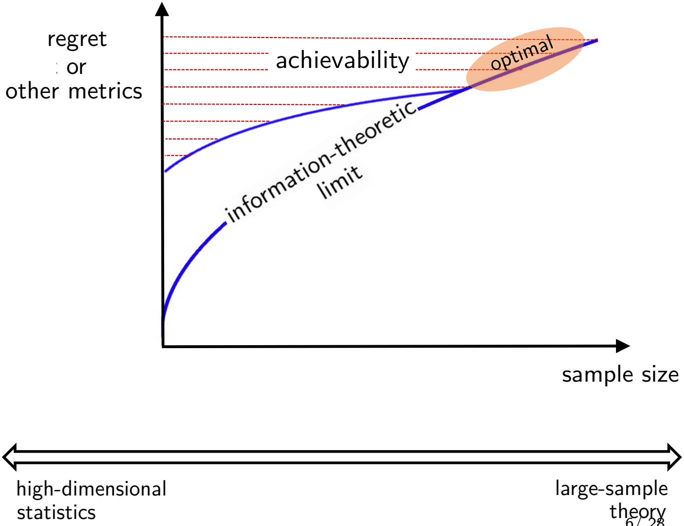
Sample complexity issues that permeate state-of-the-art RL theory



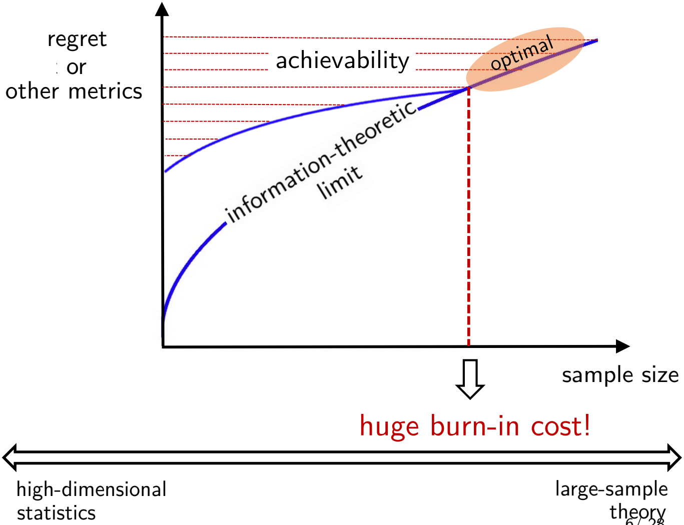
Sample complexity issues that permeate state-of-the-art RL theory



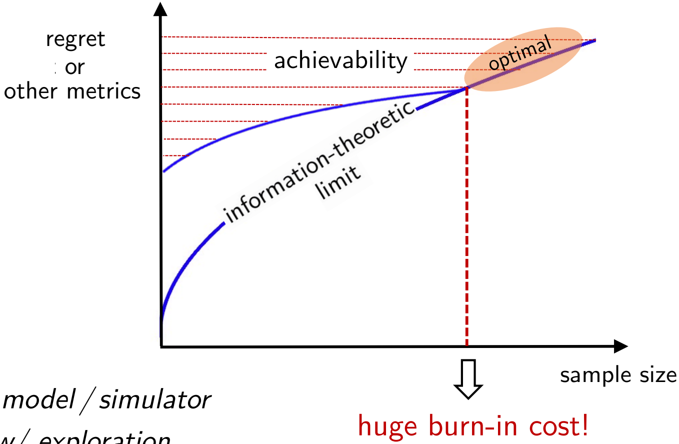
Sample complexity issues that permeate state-of-the-art RL theory



Sample complexity issues that permeate state-of-the-art RL theory

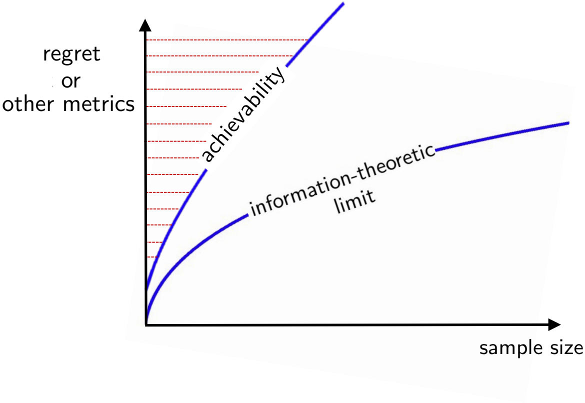


Sample complexity issues that permeate state-of-the-art RL theory



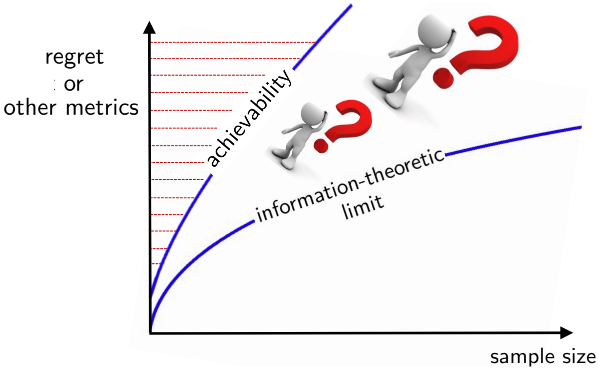
- *generative model / simulator*
- *online RL w/ exploration*
- *offline / batch RL*
- ...

Sample complexity issues that permeate state-of-the-art RL theory

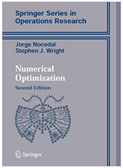


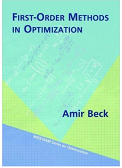
- *multi-agent RL*
- *partially observable MDPs*
- ...

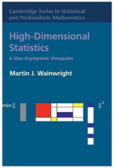
Sample complexity issues that permeate state-of-the-art RL theory

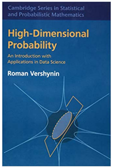


- *multi-agent RL*
- *partially observable MDPs*
- ...









(large-scale) optimization

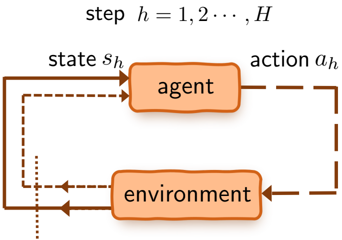
(high-dimensional) statistics

This talk: breaking sample size barrier in **online RL**

— *accomplished by a model-based approach!*

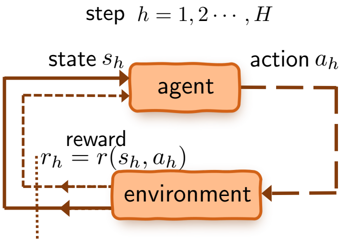
Background: Markov decision process (MDP)

Finite-horizon Markov decision process (MDP)



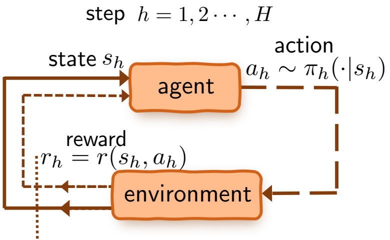
- H: horizon length (large)
- $\mathcal{S} = \{1, \dots, S\}$: state space (large)
- $\mathcal{A} = \{1, \dots, A\}$: action space (large)

Finite-horizon Markov decision process (MDP)



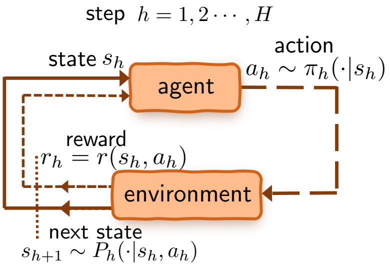
- H: horizon length (large)
- $\mathcal{S} = \{1, \ldots, S\}$: state space (large)
- $\mathcal{A} = \{1, \ldots, A\}$: action space (large)
- $r_h(s_h, a_h) \in [0, 1]$: immediate reward in step h

Finite-horizon Markov decision process (MDP)

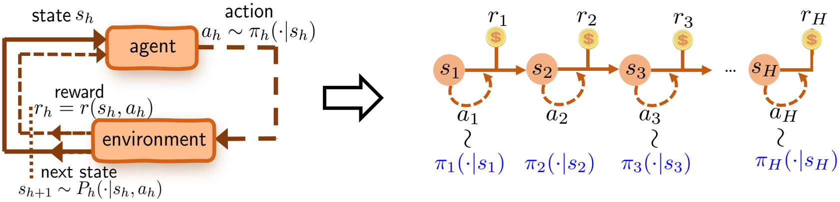


- H: horizon length (large)
- $\mathcal{S} = \{1, \ldots, S\}$: state space (large)
- $\mathcal{A} = \{1, \ldots, A\}$: action space (large)
- $r_h(s_h, a_h) \in [0, 1]$: immediate reward in step h
- $\pi = \{\pi_h\}_{1 \le h \le H}$: policy

Finite-horizon Markov decision process (MDP)



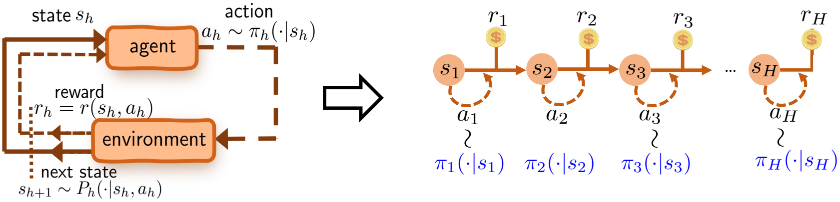
- H: horizon length (large)
- $\mathcal{S} = \{1, \ldots, S\}$: state space (large)
- $\mathcal{A} = \{1, \ldots, A\}$: action space (large)
- $r_h(s_h, a_h) \in [0, 1]$: immediate reward in step h
- $\pi = \{\pi_h\}_{1 \leq h \leq H}$: policy
- $P_h(\cdot \mid s, a)$: transition probability in step h



execute policy π to generate a trajectory $\{(s_t, a_t)\}_{1\le t\le H}$

value function of π :

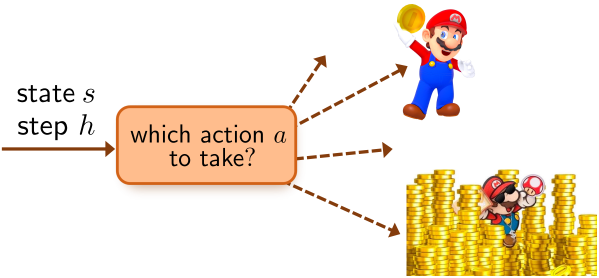
$$V_h^\pi(s) := \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\Big|\, s_h = s\right]$$



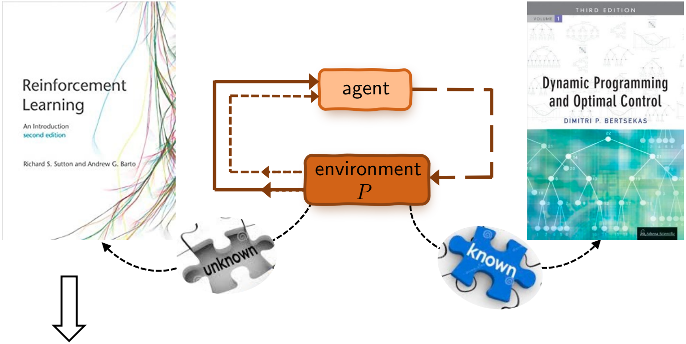
execute policy π to generate a trajectory $\{(s_t, a_t)\}_{1\le t\le H}$

value function of π : $$V_h^\pi(s) := \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\Big|\, s_h = s\right]$$

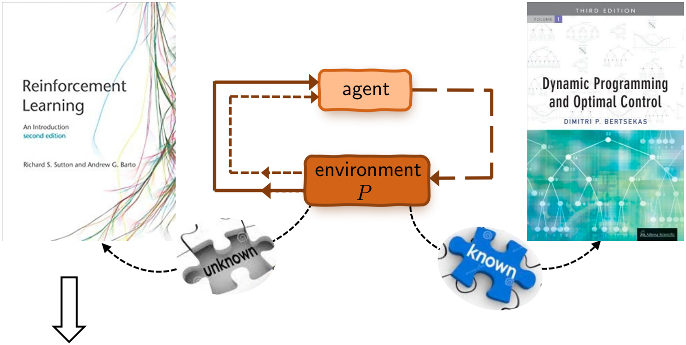
Q-function of π : $$Q_h^\pi(s,a) := \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\Big|\, s_h = s, a_h = a\right]$$



- **Optimal policy** $\pi^{\star}$: maximizing the value function
- Optimal values: $V^{\star} := V^{\pi^{\star}}$



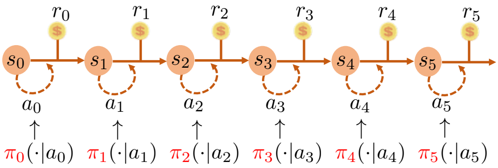
Need to collect data to learn unknown environments



Need to collect data to learn unknown environments

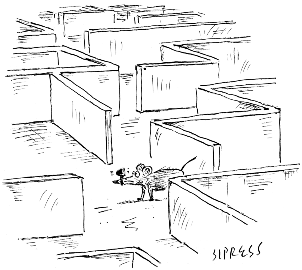
1. simulator (Li, Wei, Chi, Chen '24, Operations Research)
2. offline RL (Li, Shi, Chen, Chi, Wei '24, Annals. Stats)
3. **online exploratory RL** **(this talk)**

Online RL: interacting with real environment



exploration via adaptive sampling

- trial-and-error
- sequential and online
- adaptive learning from data

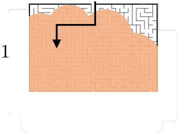


"Recalculating ... recalculating ..."

Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps

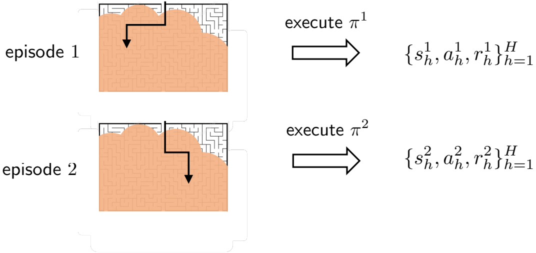
episode 1



execute π^1 $\Longrightarrow$ $\{s_h^1, a_h^1, r_h^1\}_{h=1}^H$

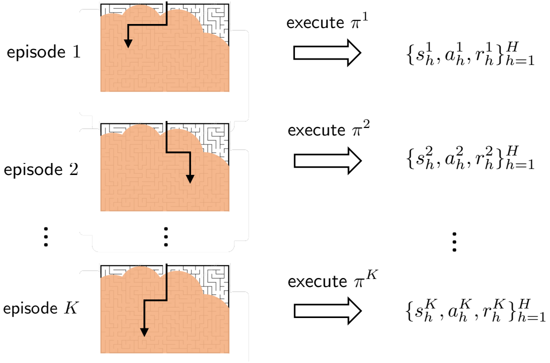
Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps



Online episodic RL

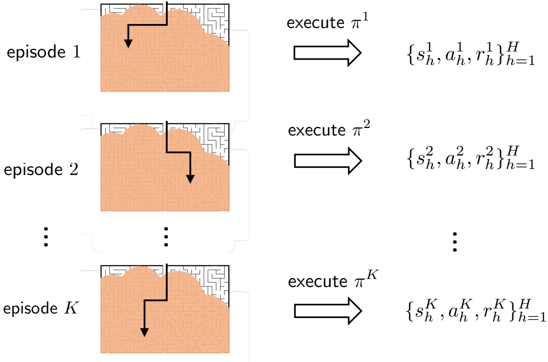
Sequentially execute MDP for K episodes, each consisting of H steps



Online episodic RL

Sequentially execute MDP for K episodes, each consisting of H steps

— *sample size:* $T = KH$



exploration (exploring unknowns) vs. **exploitation** (exploiting learned info)

Regret: gap between learned policy & optimal policy

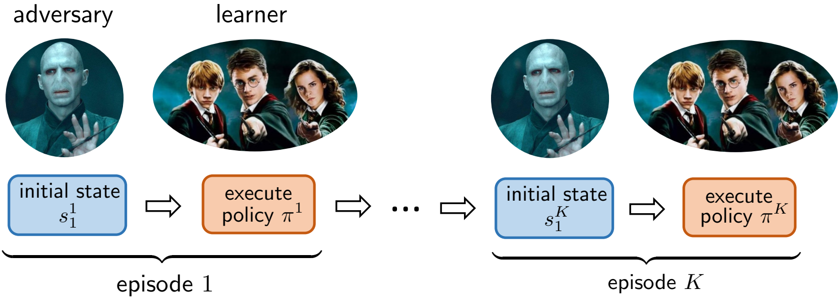
adversary

learner

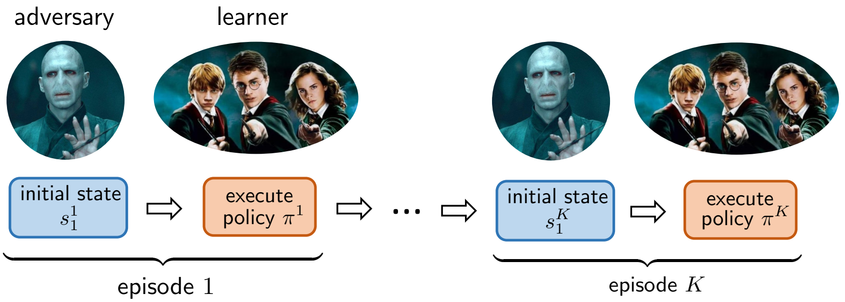
initial state s_1^1 $\Rightarrow$ execute policy π^1

episode 1

Regret: gap between learned policy & optimal policy



Regret: gap between learned policy & optimal policy



Performance metric: given initial states $\{s_1^k\}_{k=1}^K$, define

$$\text{Regret}(T) \quad := \quad \sum_{k=1}^{K} \Big(V_1^{\star}(s_1^k) - V_1^{\pi^k}(s_1^k) \Big)$$

Lower bound

(Domingues et al. '21)

$$\text{Regret}(T) \gtrsim \sqrt{H^2SAT}$$

Existing algorithms

- UCB-VI: Azar et al. '17
- UBEV: Dann et al. '17
- UCB-Q-Hoeffding: Jin et al. '18
- UCB-Q-Bernstein: Jin et al. '18
- UCB2-Q-Bernstein: Bai et al. '19
- EULER: Zanette et al. '19
- UCB-Q-Advantage: Zhang et al. '20
- MVP: Zhang et al. '20
- UCB-M-Q: Menard et al. '21
- Q-EarlySettled-Advantage: Li et al. '21
- (modified) MVP: Zhang et al. '23

Lower bound

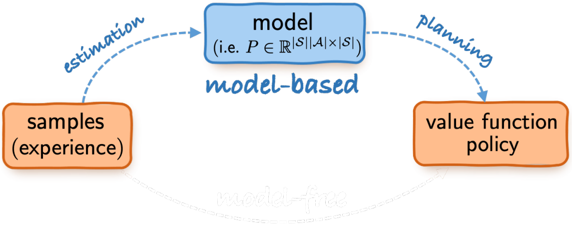
(Domingues et al. '21)

$$\text{Regret}(T) \gtrsim \sqrt{H^2SAT}$$

Existing algorithms

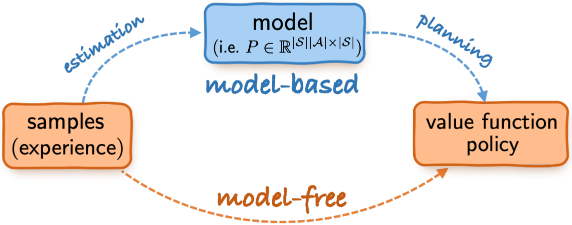
- UCB-VI: Azar et al. '17
- UBEV: Dann et al. '17
- UCB-Q-Hoeffding: Jin et al. '18
- UCB-Q-Bernstein: Jin et al. '18
- UCB2-Q-Bernstein: Bai et al. '19
- EULER: Zanette et al. '19
- UCB-Q-Advantage: Zhang et al. '20
- MVP: Zhang et al. '20
- UCB-M-Q: Menard et al. '21
- Q-EarlySettled-Advantage: Li et al. '21
- (modified) MVP: Zhang et al. '23

Which online RL algorithms achieve near-minimal regret?



Model-based approach ("plug-in")

1. build an empirical estimate $\widehat{P}$ for P
2. planning based on the empirical $\widehat{P}$

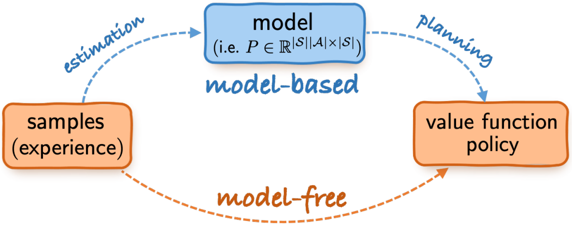


Model-based approach ("plug-in")

1. build an empirical estimate $\widehat{P}$ for P
2. planning based on the empirical $\widehat{P}$

Model-free approach (e.g. Q-learning)

— learning w/o estimating the model explicitly



Model-based approach ("plug-in")

1. build an empirical estimate $\widehat{P}$ for P
2. planning based on the empirical $\widehat{P}$

Model-free approach (e.g. Q-learning)

— learning w/o estimating the model explicitly

T. L. Lai

H. Robbins

Optimism in the face of uncertainty:

- explores based on the best optimistic estimates associated with the actions!
- a common framework: utilize $\underbrace{\text{upper confidence bounds (UCB)}}_{\text{accounts for estimates + uncertainty level}}$

T. L. Lai

H. Robbins

Optimism in the face of uncertainty:

- explores based on the best optimistic estimates associated with the actions!
- a common framework: utilize $\underbrace{\text{upper confidence bounds (UCB)}}_{\text{accounts for estimates + uncertainty level}}$

Optimistic model-based approach: incorporates UCB framework into model-based approach

UCB-VI (Azar et al. '17)

For each episode:

1. Backtrack $h = H, H-1, \ldots, 1$: run **value iteration**

$$Q_h(s_h, a_h) \leftarrow r_h(s_h, a_h) + \underbrace{\widehat{P}_{h,s_h,a_h}}_{\text{model estimate}} V_{h+1}$$

$$V_h(s_h) \leftarrow \max_{a \in \mathcal{A}} Q_h(s_h, a)$$

UCB-VI (Azar et al. '17)

For each episode:

1. Backtrack $h = H, H-1, \ldots, 1$: run **optimistic value iteration**

$$Q_h(s_h, a_h) \leftarrow r_h(s_h, a_h) + \underbrace{\widehat{P}_{h,s_h,a_h}}_{\text{model estimate}} V_{h+1} + \underbrace{b_h(s_h, a_h)}_{\text{bonus}}$$

$$V_h(s_h) \leftarrow \max_{a \in \mathcal{A}} Q_h(s_h, a)$$

UCB-VI (Azar et al. '17)

For each episode:

1. Backtrack $h = H, H-1, \dots, 1$: run **optimistic value iteration**

$$Q_h(s_h, a_h) \leftarrow r_h(s_h, a_h) + \underbrace{\widehat{P}_{h,s_h,a_h}}_{\text{model estimate}} V_{h+1} + \underbrace{b_h(s_h, a_h)}_{\text{bonus}}$$

$$V_h(s_h) \leftarrow \max_{a \in \mathcal{A}} Q_h(s_h, a)$$

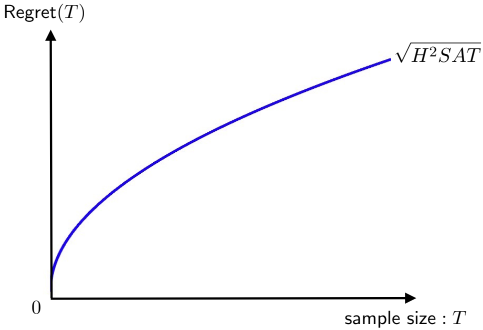
2. Forward $h = 1, \dots, H$: take actions according to **greedy policy**

$$\pi_h(s) \leftarrow \text{argmax}_{a \in \mathcal{A}} Q_h(s, a)$$

to collect a new episode $\{s_h, a_h, r_h\}_{h=1}^{H}$

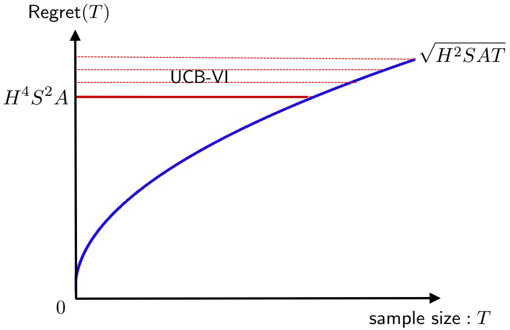
UCB-VI is asymptotically regret-optimal

— *Azar, Osband, Munos '17*



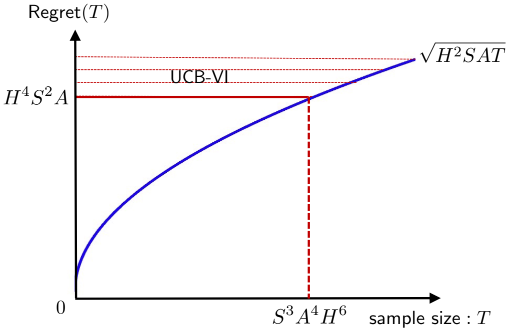
UCB-VI is asymptotically regret-optimal

— *Azar, Osband, Munos '17*



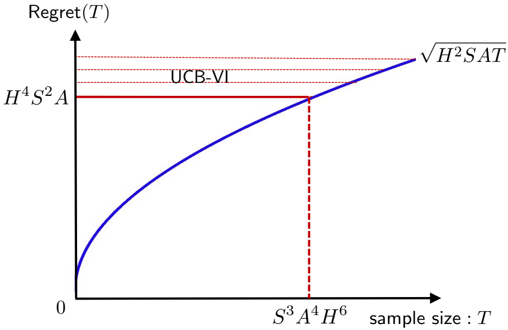
UCB-VI is asymptotically regret-optimal

— Azar, Osband, Munos '17



UCB-VI is asymptotically regret-optimal

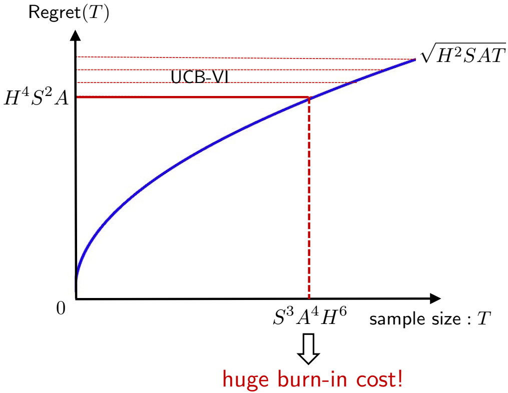
— *Azar, Osband, Munos '17*



huge burn-in cost!

UCB-VI is asymptotically regret-optimal

— *Azar, Osband, Munos '17*



Issues: large burn-in cost

Other asymptotically regret-optimal algorithms

Algorithm	Regret upper bound	Range of K that attains optimal regret
UCBVI (Azar et al. 17)	$\sqrt{SAH^2T} + S^2AH^3$	$[S^3AH^3, \infty)$
ORLC (Dann et al. '19)	$\sqrt{SAH^2T} + S^2AH^4$	$[S^3AH^5, \infty)$
EULER (Zanette et al. '19)	$\sqrt{SAH^2T} + S^{3/2}AH^3(\sqrt{S} + \sqrt{H})$	$[S^2AH^3(\sqrt{S} + \sqrt{H}), \infty)$
UCB-Adv (Zhang et al. '20)	$\sqrt{SAH^2T} + S^2A^{3/2}H^{33/4}K^{1/4}$	$[S^6A^4H^{27}, \infty)$
MVP (Zhang et al. '20)	$\sqrt{SAH^2T} + S^2AH^2$	$[S^3AH, \infty)$
UCB-M-Q (Menard et al. '21)	$\sqrt{SAH^2T} + SAH^4$	$[SAH^5, \infty)$
Q-Earlysettled-Adv (Li et al. '21)	$\sqrt{SAH^2T} + SAH^6$	$[SAH^9, \infty)$

Other asymptotically regret-optimal algorithms

Algorithm	Regret upper bound	Range of K that attains optimal regret
UCBVI (Azar et al. 17)	$\sqrt{SAH^2T} + S^2AH^3$	$[S^3AH^3, \infty)$
ORLC (Dann et al. '19)	$\sqrt{SAH^2T} + S^2AH^4$	$[S^3AH^5, \infty)$
EULER (Zanette et al. '19)	$\sqrt{SAH^2T} + S^{3/2}AH^3(\sqrt{S} + \sqrt{H})$	$[S^2AH^3(\sqrt{S} + \sqrt{H}), \infty)$
UCB-Adv (Zhang et al. '20)	$\sqrt{SAH^2T} + S^2A^{3/2}H^{33/4}K^{1/4}$	$[S^6A^4H^{27}, \infty)$
MVP (Zhang et al. '20)	$\sqrt{SAH^2T} + S^2AH^2$	$[S^3AH, \infty)$
UCB-M-Q (Menard et al. '21)	$\sqrt{SAH^2T} + SAH^4$	$[SAH^5, \infty)$
Q-Earlysettled-Adv (Li et al. '21)	$\sqrt{SAH^2T} + SAH^6$	$[SAH^9, \infty)$

Can we find a regre-optimal algorithm with no burn-in cost?

Monotonic Value Propagation (Zhang et al. '21)

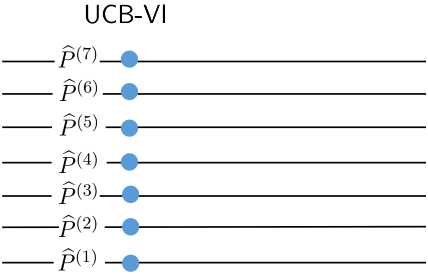
UCB-VI with doubling update rules and variance-aware bonus

- (s, a, h) is updated only when visited the $\{1, 3, 7, 15, \cdots\}$-th time

Monotonic Value Propagation (Zhang et al. '21)

UCB-VI with doubling update rules and variance-aware bonus

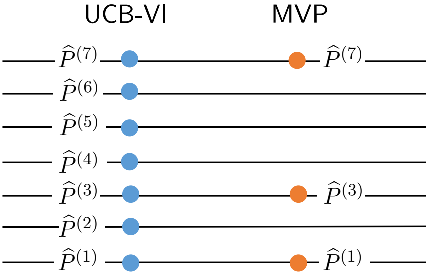
- (s, a, h) is updated only when visited the $\{1, 3, 7, 15, \cdots\}$-th time



Monotonic Value Propagation (Zhang et al. '21)

UCB-VI with doubling update rules and variance-aware bonus

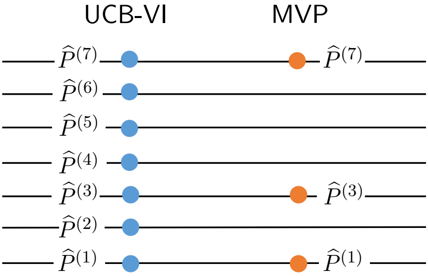
- (s, a, h) is updated only when visited the $\{1, 3, 7, 15, \cdots\}$-th time



Monotonic Value Propagation (Zhang et al. '21)

UCB-VI with doubling update rules and variance-aware bonus

- (s, a, h) is updated only when visited the $\{1, 3, 7, 15, \cdots\}$-th time

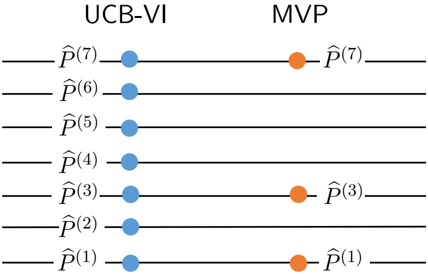


  - visitation counts change much less frequently
    $\longrightarrow$ reduces covering number dramatically

Monotonic Value Propagation (Zhang et al. '21)

UCB-VI with doubling update rules and variance-aware bonus

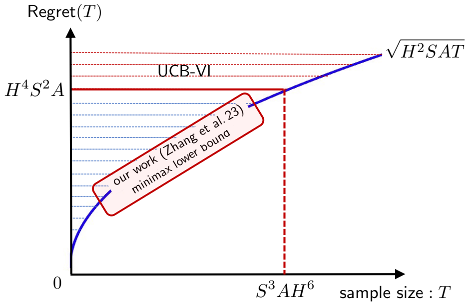
- (s, a, h) is updated only when visited the $\{1, 3, 7, 15, \cdots\}$-th time



  - visitation counts change much less frequently
    $\longrightarrow$ reduces covering number dramatically

- data-driven bonus terms (chosen based on empirical variances)

Regret-optimal algorithm w/o burn-in cost

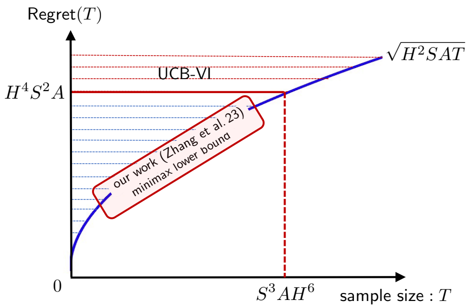


Theorem 1 (Zhang, Chen, Lee, Du '23)

The model-based algorithm Monotonic Value Propagation achieves

$$\text{Regret}(T) \lesssim \widetilde{O}(\sqrt{H^2SAT})$$

Regret-optimal algorithm w/o burn-in cost



Theorem 1 (Zhang, Chen, Lee, Du '23)

The model-based algorithm Monotonic Value Propagation achieves

$$\textit{Regret}(T) \lesssim \widetilde{O}(\sqrt{H^2SAT})$$

- the only algorithm so far that is regret-optimal w/o burn-ins

Key technical innovation

online data collection
w/ sample reuse

samples drawn
independently from simulator

Decoupling complicated statistical dependency during online learning

Key technical innovation

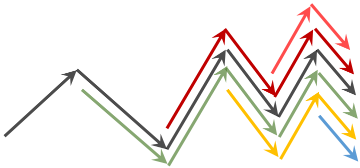
online data collection
w/ sample reuse

samples drawn
independently from simulator

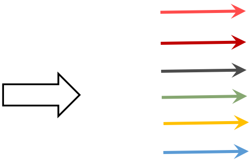
Decoupling complicated statistical dependency during online learning

- couples online data collection with i.i.d. sampling

Key technical innovation



online data collection
w/ sample reuse



samples drawn
independently from simulator

Decoupling complicated statistical dependency during online learning

- couples online data collection with i.i.d. sampling
- exploit *compressibility* of visitation counts
    - w/ the aid of doubling algorithmic trick

Summary for online RL

- model-based approach is regret-optimal w/ no burn-in cost

Summary for online RL

- model-based approach is regret-optimal w/ no burn-in cost

open problems:

- how to design model-free algorithms w/o burn-in cost (i.e., w/ optimal H-dependency too)?

Summary for online RL

- model-based approach is regret-optimal w/ no burn-in cost

open problems:

- how to design model-free algorithms w/o burn-in cost (i.e., w/ optimal H-dependency too)?
- how to achieve full-range regret-optimal algorithms for:
  - discounted infinite-horizon MDPs?
  - finite-horizon stationary MDPs?
  - ...

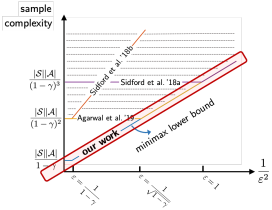
Concluding remarks

Model-based alg. remains the only solution that achieves optimal sample complexity w/o burn-ins for these scenarios *and beyond*

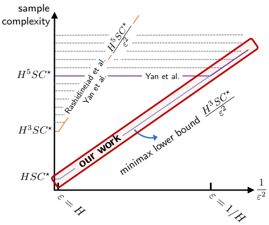
Concluding remarks

Model-based alg. remains the only solution that achieves optimal sample complexity w/o burn-ins for these scenarios *and beyond*

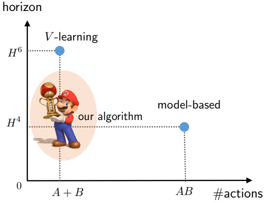
Model-based approach is also optimal w/o burn-ins for



RL w/ simulator



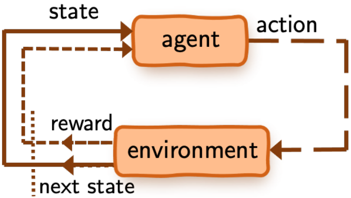
Offline RL

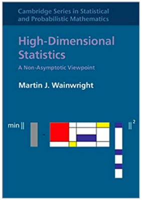


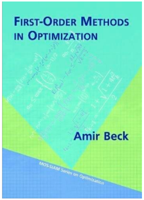
2-player zero-sum Markov games

Concluding remarks

Understanding RL requires modern statistics and optimization







"Settling the sample complexity of online reinforcement learning," Z. Zhang, Y. Chen, J. Lee, S. Du, arXiv:2307.13586, 2023

"Breaking the sample size barrier in model-based reinforcement learning with a generative model," G. Li, Y. Wei, Y. Chi, Y. Chen, *Operations Research*, 2024

"Settling the sample complexity of model-based offline reinforcement learning," G. Li, L. Shi, Y. Chen, Y. Chi, Y. Wei, *Annals of Statistics*, 2024