Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds

Andrea Zanette*, Emma Brunskill

zanette@stanford.edu ebrun@cs.stanford.edu

Institute for Computational and Mathematical Engineering and Department of Computer Science, Stanford University

Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds

Andrea Zanette*, Emma Brunskill

zanette@stanford.edu ebrun@cs.stanford.edu

Institute for Computational and Mathematical Engineering and Department of Computer Science, Stanford University

Exploration in RL

=

Learn quickly how to play near optimally

Setting: episodic tabular RL

Goal: <u>automatically</u> inherit instance-dependent regret bounds

State of the Art Regret Bounds for Episodic Tabular MDPs

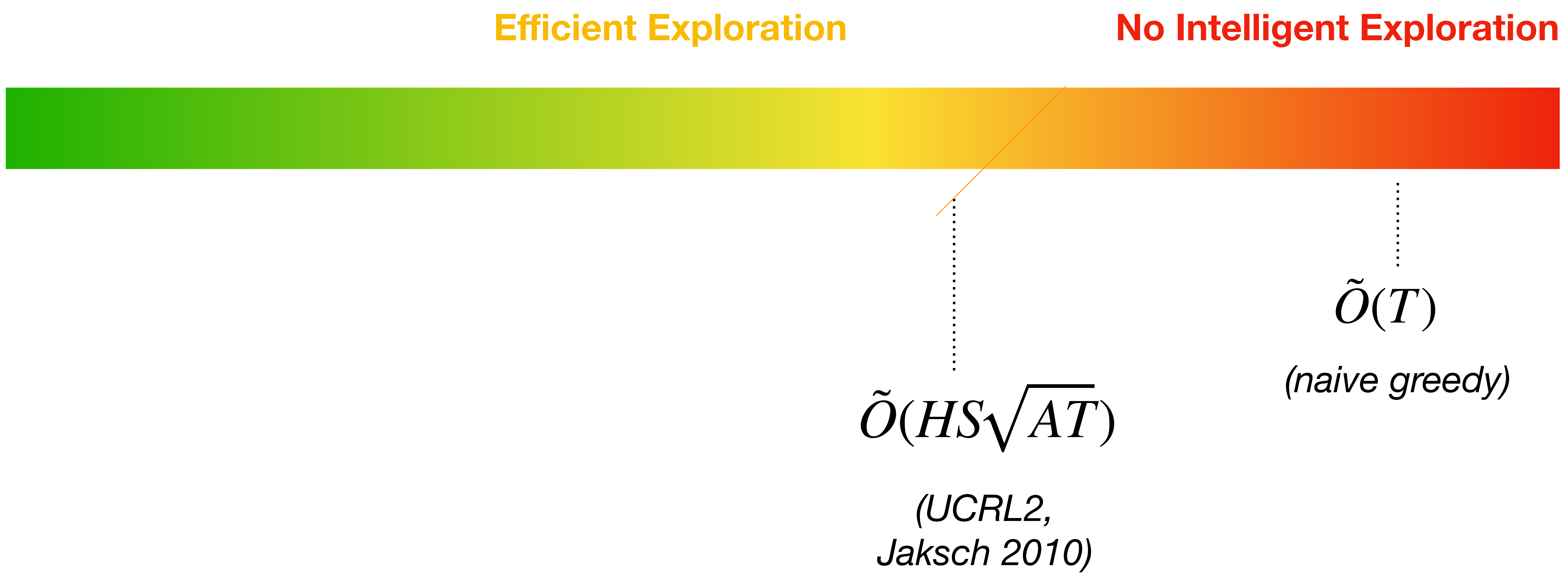
State of the Art Regret Bounds for Episodic Tabular MDPs

No Intelligent Exploration

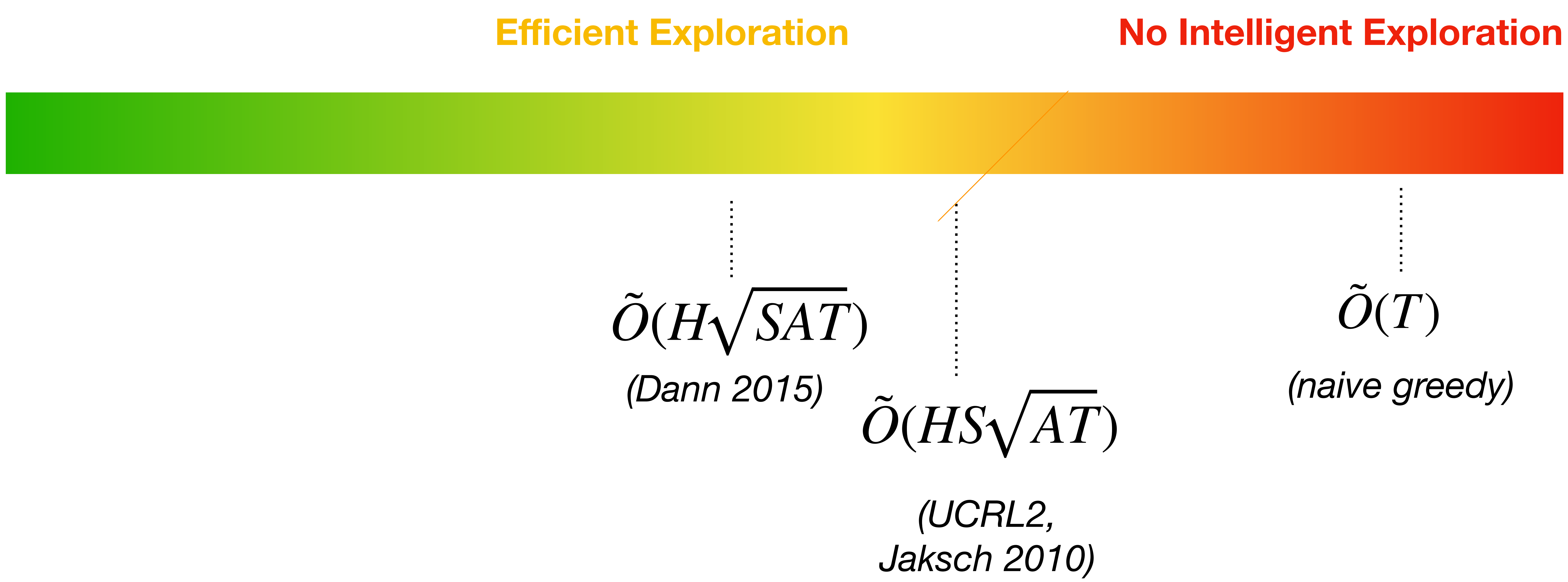
$\tilde{O}(T)$

(naive greedy)

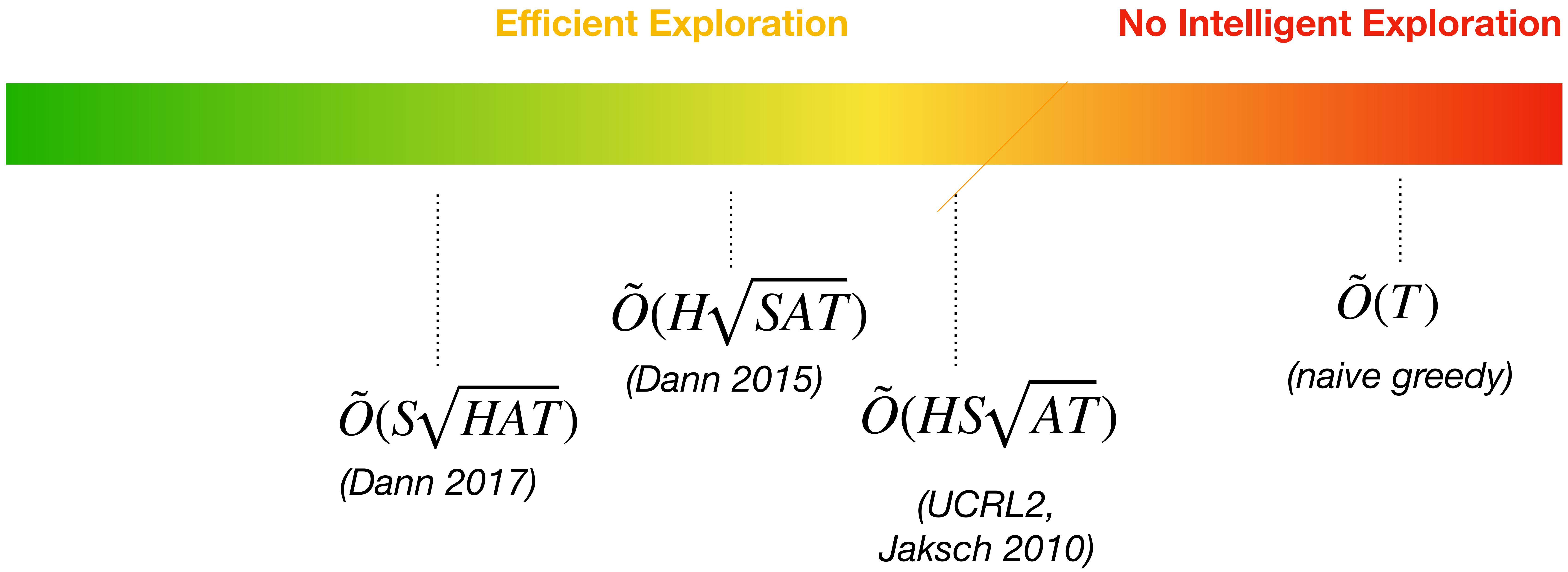
State of the Art Regret Bounds for Episodic Tabular MDPs



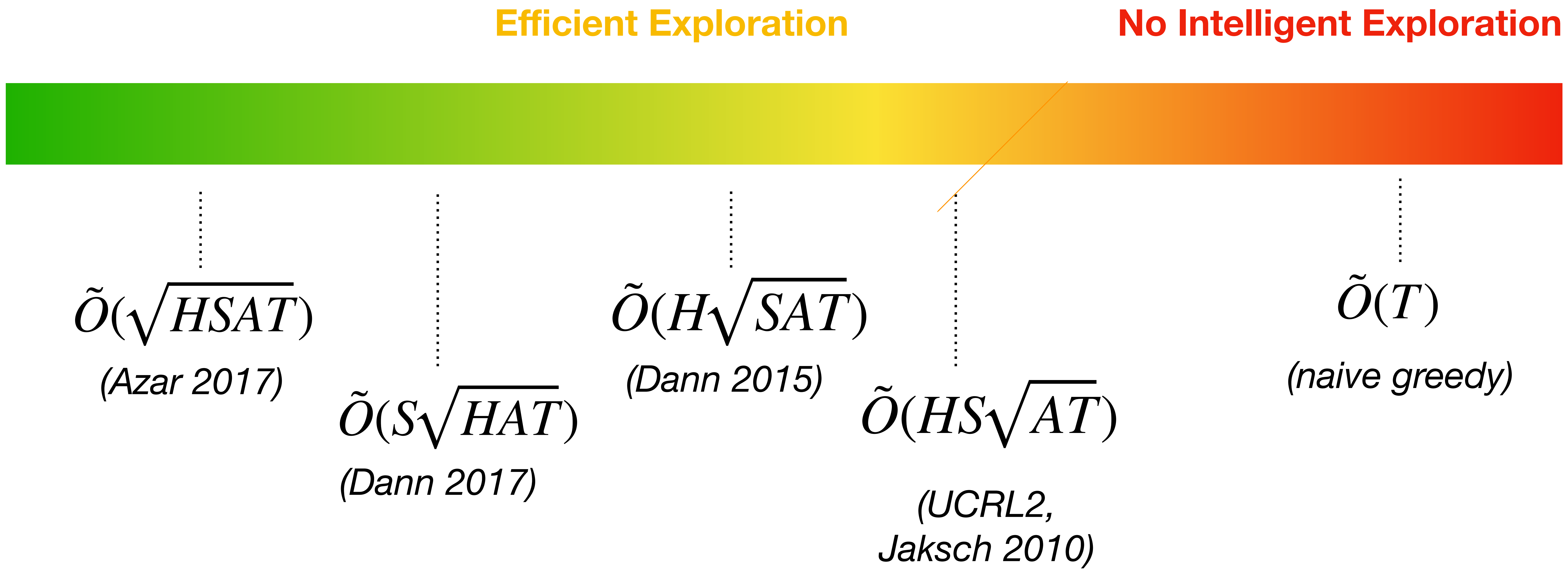
State of the Art Regret Bounds for Episodic Tabular MDPs



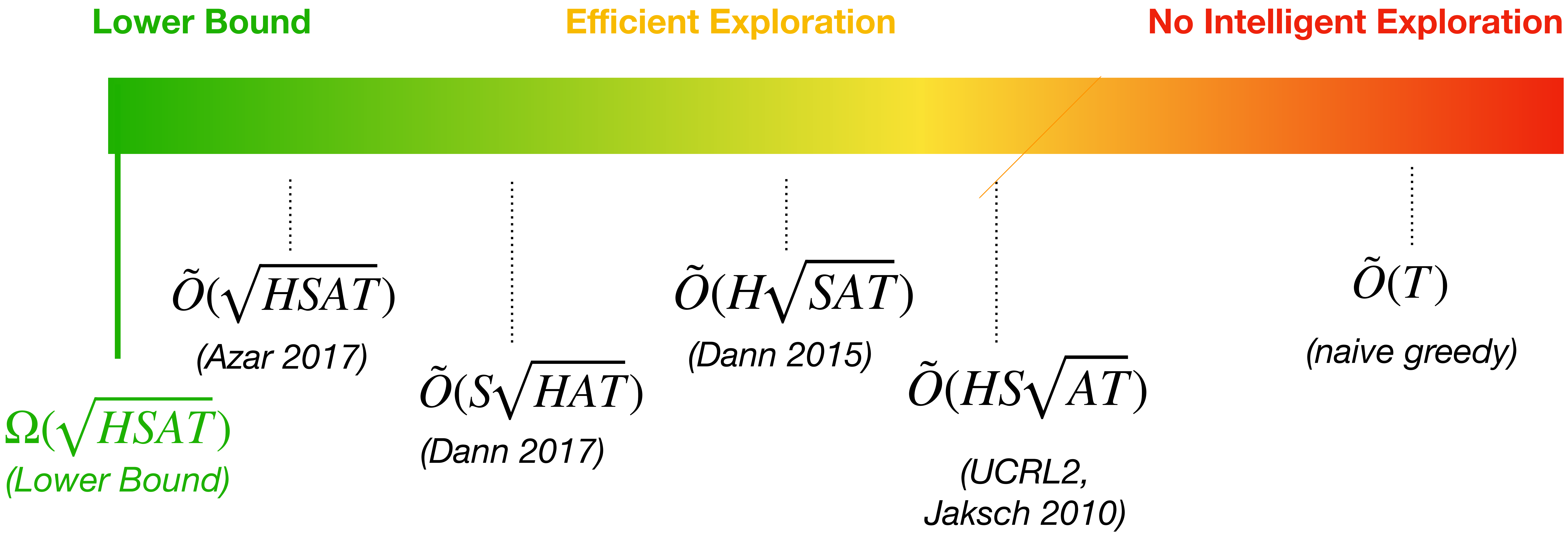
State of the Art Regret Bounds for Episodic Tabular MDPs

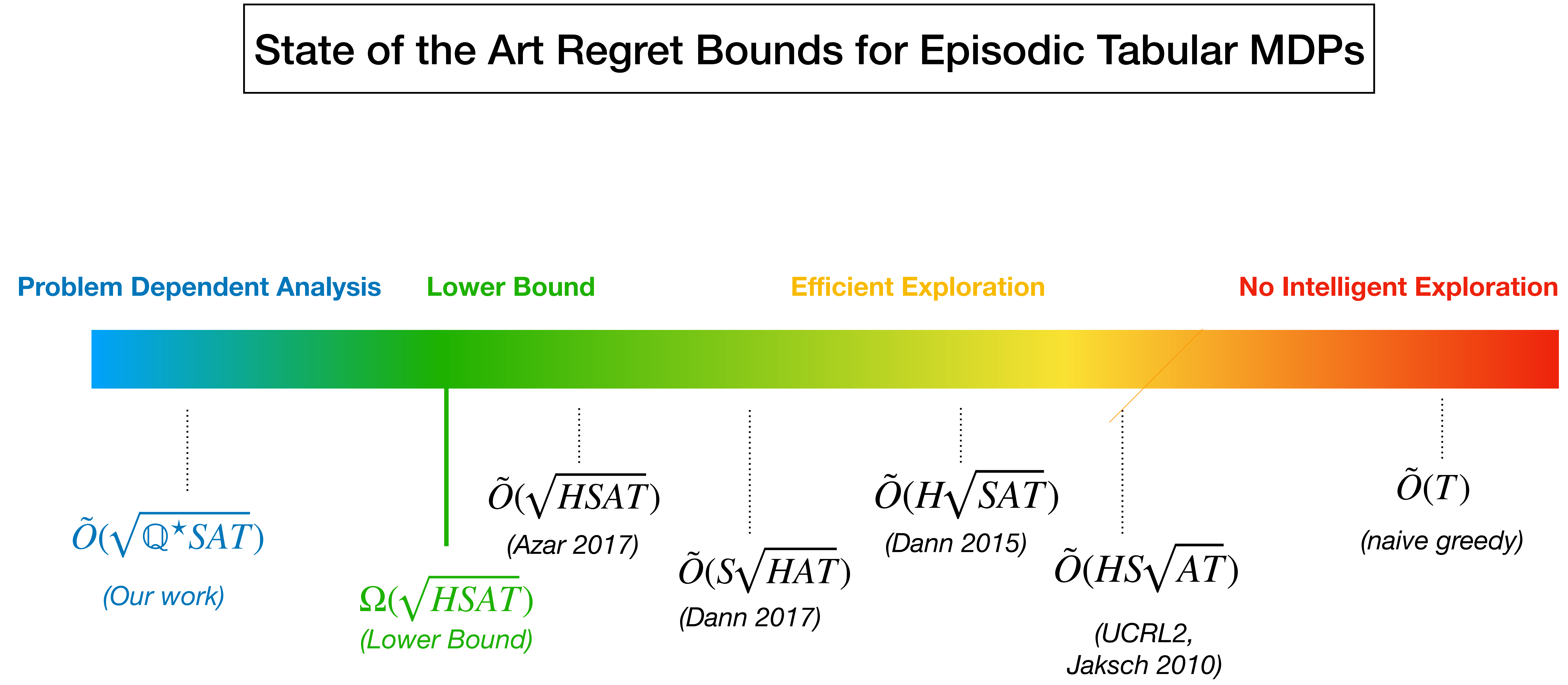


State of the Art Regret Bounds for Episodic Tabular MDPs



State of the Art Regret Bounds for Episodic Tabular MDPs

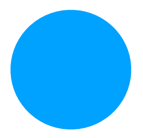




State of the Art Regret Bounds for Episodic Tabular MDPs
Problem Dependent Analysis
Lower Bound
Efficient Exploration
No Intelligent Exploration
$\tilde{O}(\sqrt{\mathbb{Q}^\star SAT})$
(Our work)
$\Omega(\sqrt{HSAT})$
(Lower Bound)
$\tilde{O}(\sqrt{HSAT})$
(Azar 2017)
$\tilde{O}(S\sqrt{HAT})$
(Dann 2017)
$\tilde{O}(H\sqrt{SAT})$
(Dann 2015)
$\tilde{O}(HS\sqrt{AT})$
(UCRL2, Jaksch 2010)
$\tilde{O}(T)$
(naive greedy)

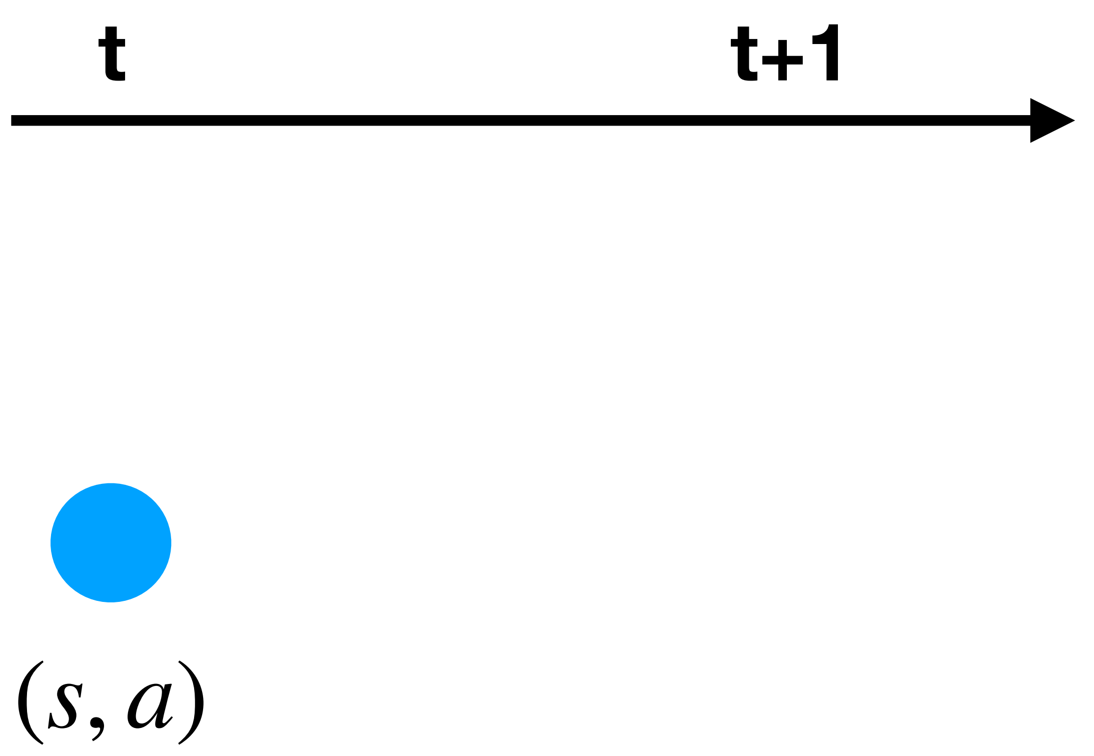
Main Result

Main Result

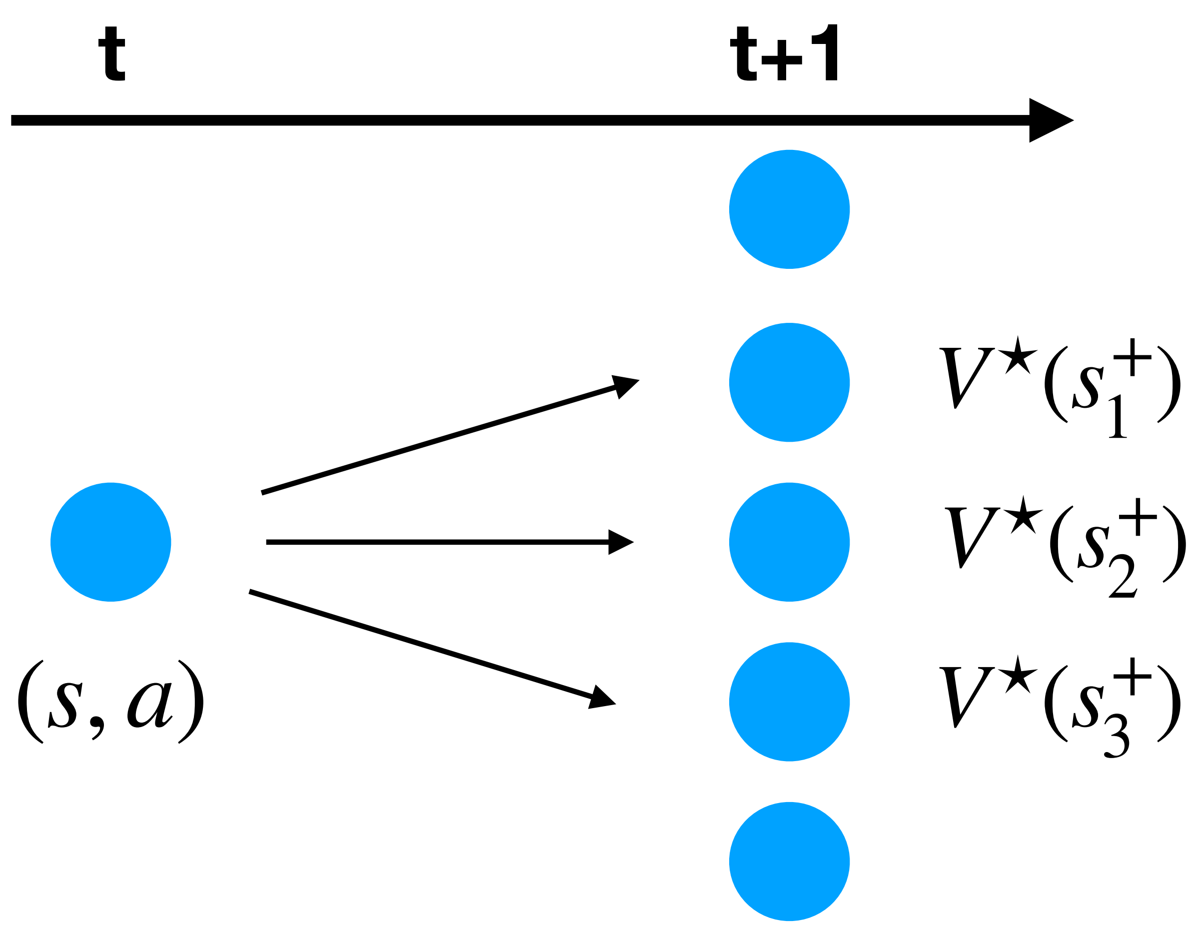


(s, a)

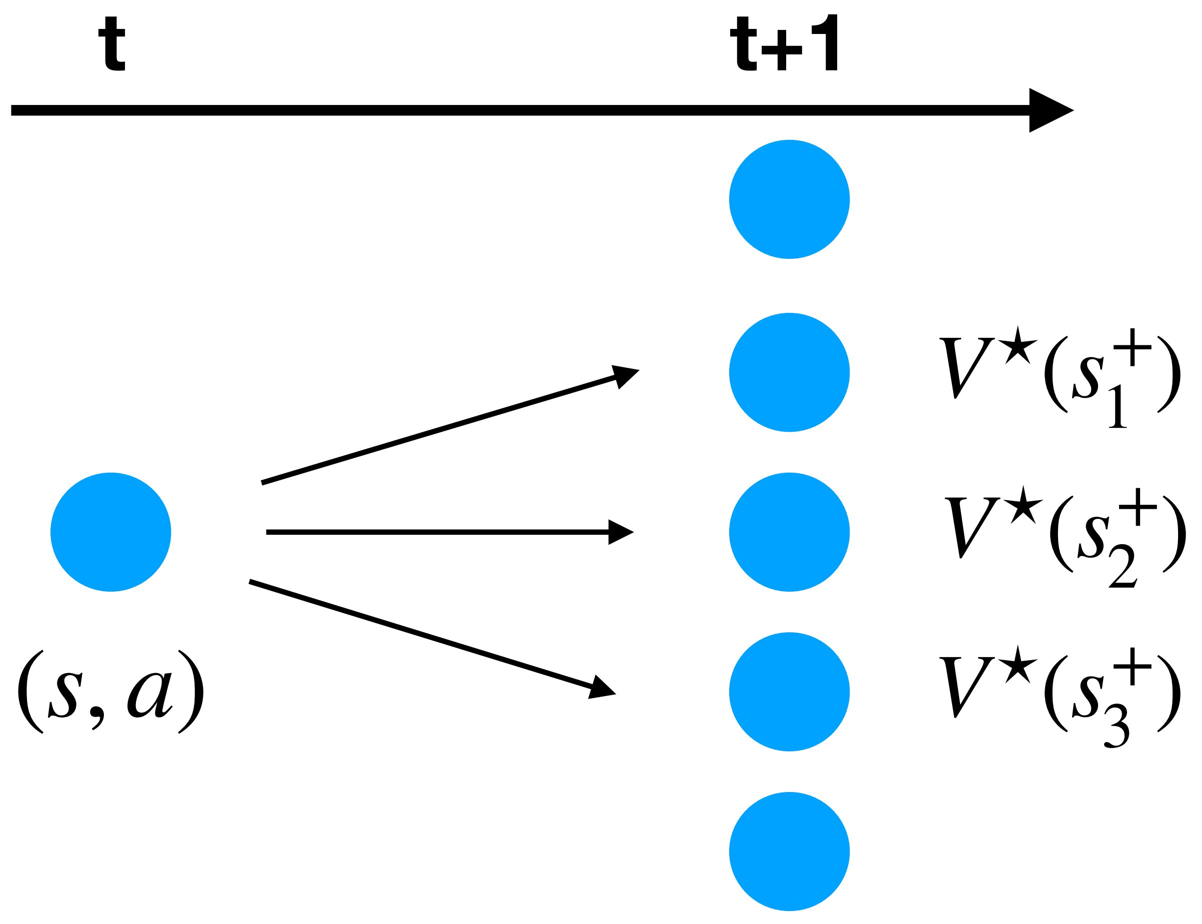
Main Result



Main Result

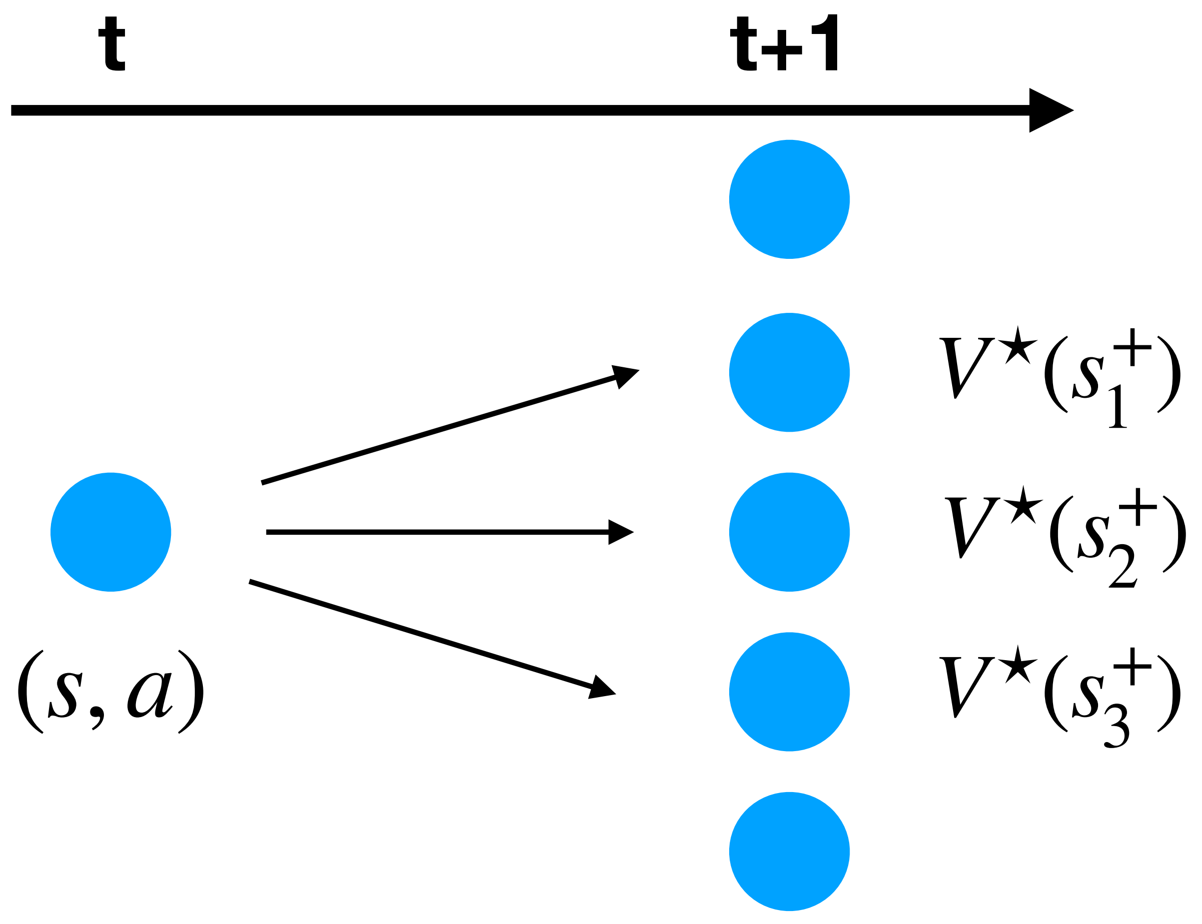


Main Result

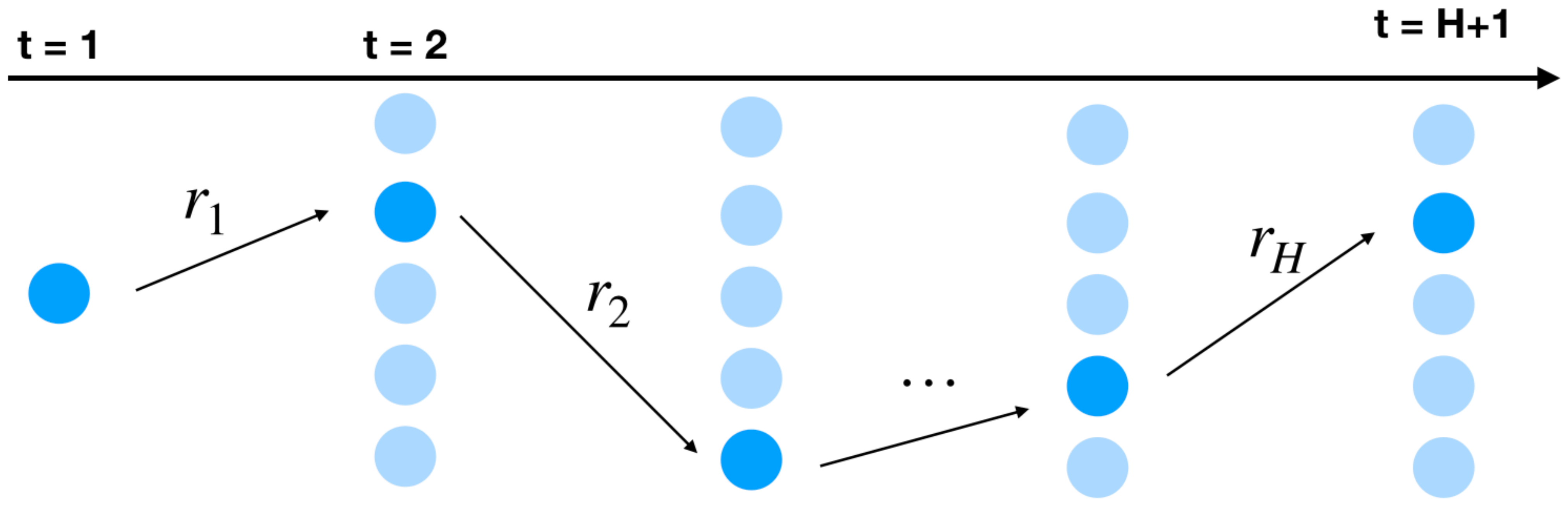


$$\mathbb{Q}^\star = max_{s,a} Var_{s^+ \sim p(s,a)} V^\star(s^+)$$

Main Result

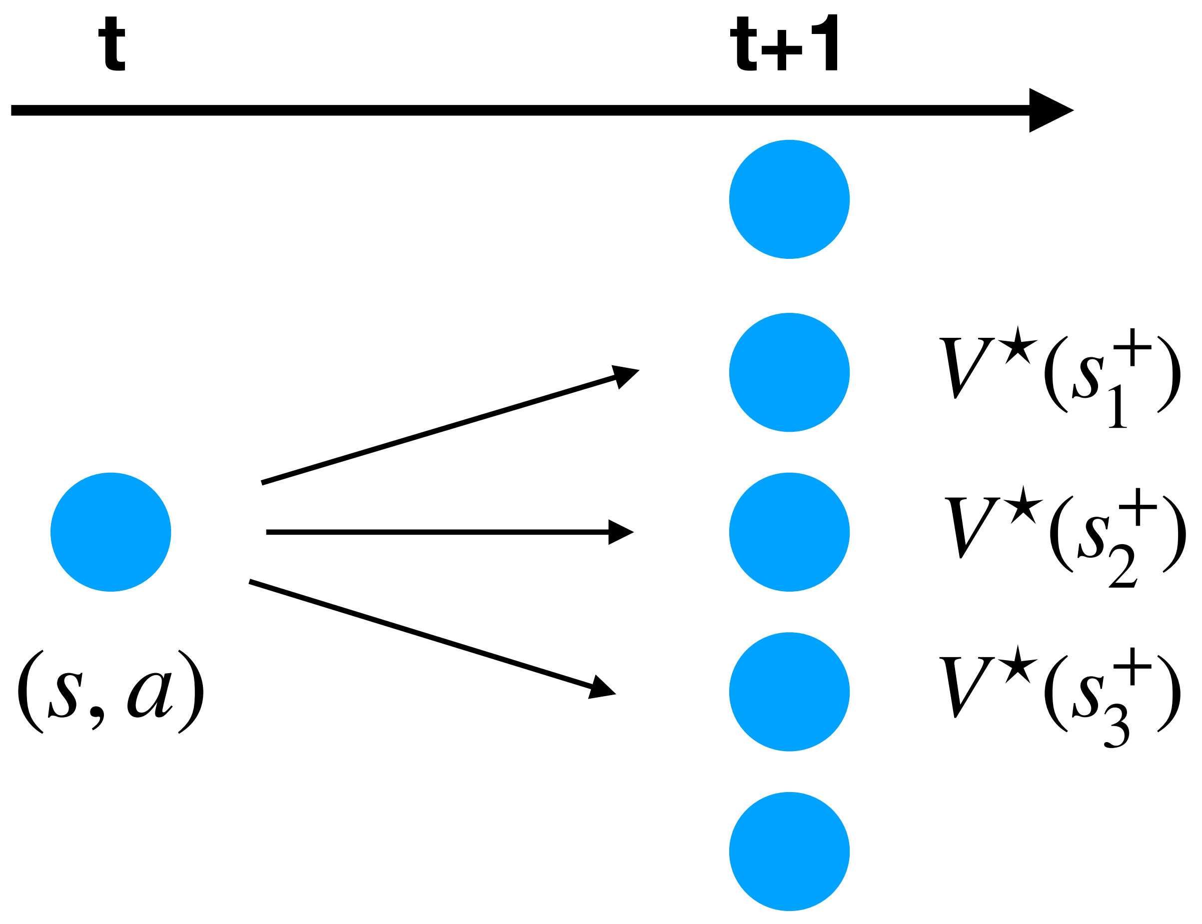


$$\mathbb{Q}^\star = max_{s,a} Var_{s^+\sim p(s,a)} V^\star(s^+)$$

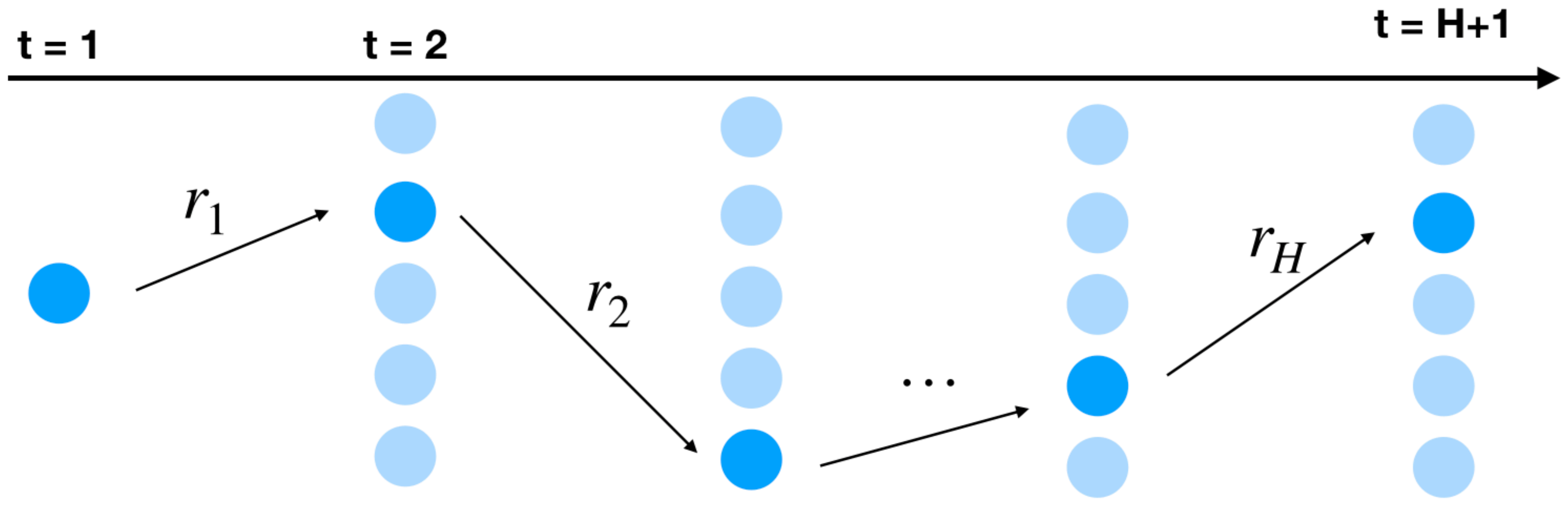


$$r_1 + r_2 + \ldots + r_H \leq \mathcal{G}$$

Main Result



$$\mathbb{Q}^\star = max_{s,a} Var_{s^+ \sim p(s,a)} V^\star(s^+)$$

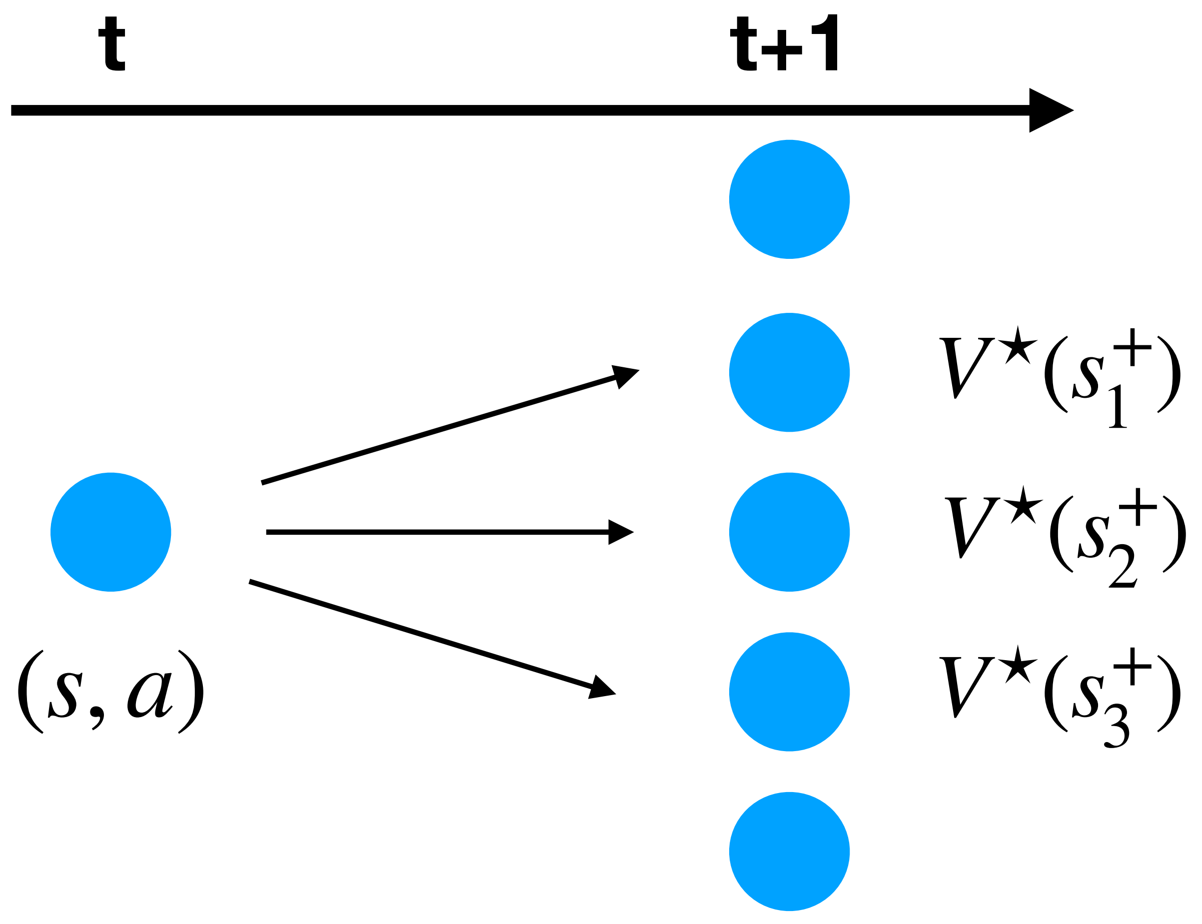


$$r_1 + r_2 + \ldots + r_H \leq \mathcal{G}$$

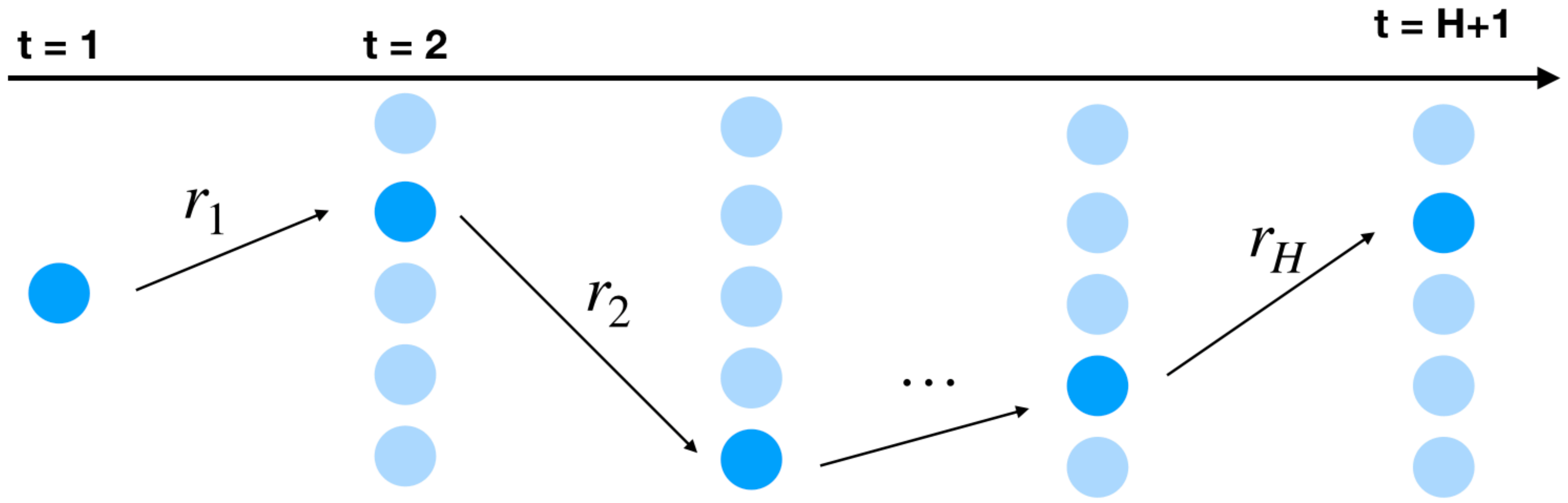
Main Result: An algorithm with a (high probability) regret bound:

$$\min \left\{ \tilde{O}(\sqrt{\mathbb{Q}^\star SAT}) + [const], \quad \tilde{O}\left(\sqrt{\frac{\mathcal{G}^2}{H} SAT}\right) + [const] \right\}$$

Main Result



$$\mathbb{Q}^\star = max_{s,a} Var_{s^+ \sim p(s,a)} V^\star(s^+)$$



$$r_1 + r_2 + \ldots + r_H \leq \mathcal{G}$$

Main Result: An algorithm with a (high probability) regret bound:

$$\min \left\{ \tilde{O}(\sqrt{\mathbb{Q}^\star SAT}) + [const], \quad \tilde{O}\left(\sqrt{\frac{\mathcal{G}^2}{H} SAT} \right) + [const] \right\}$$

Technique: exploration bonus which is adaptively adjusted as a function of the problem difficulty

Long Horizon MDPs

Long Horizon MDPs

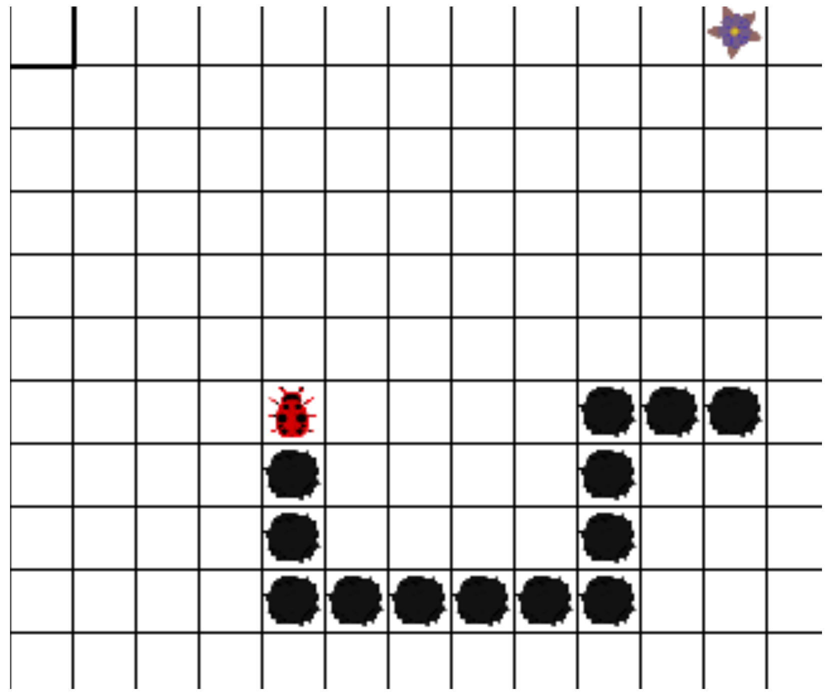
Standard Setting $r \in [0,1]$

Long Horizon MDPs

Standard Setting $r \in [0,1]$

Goal MDP Setting* $r \geq 0, \quad \sum_{t=1}^{H} r_t \leq 1$

** this is a more general setting*

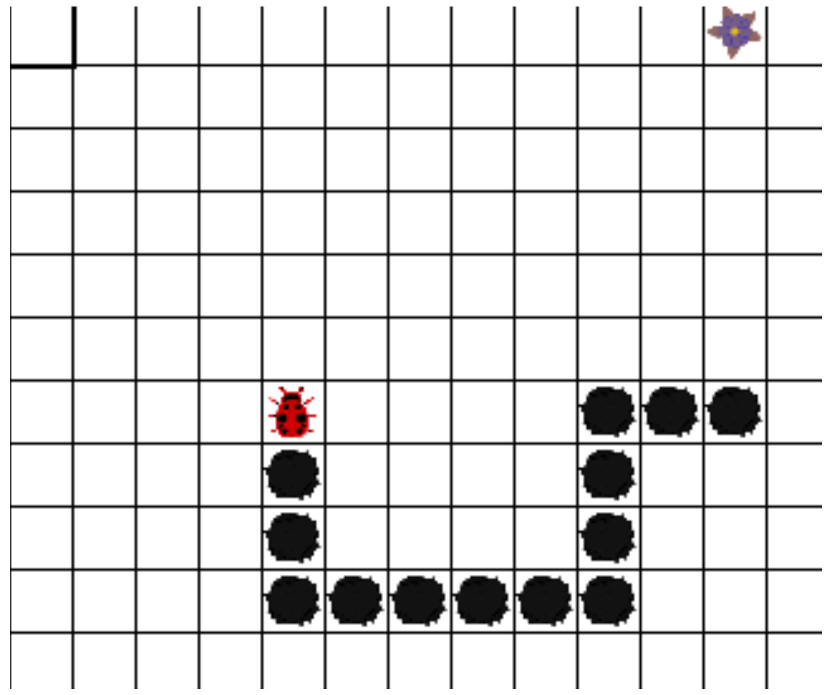


Long Horizon MDPs

Standard Setting $r \in [0,1]$

Goal MDP Setting* $r \geq 0, \quad \sum_{t=1}^{H} r_t \leq 1$

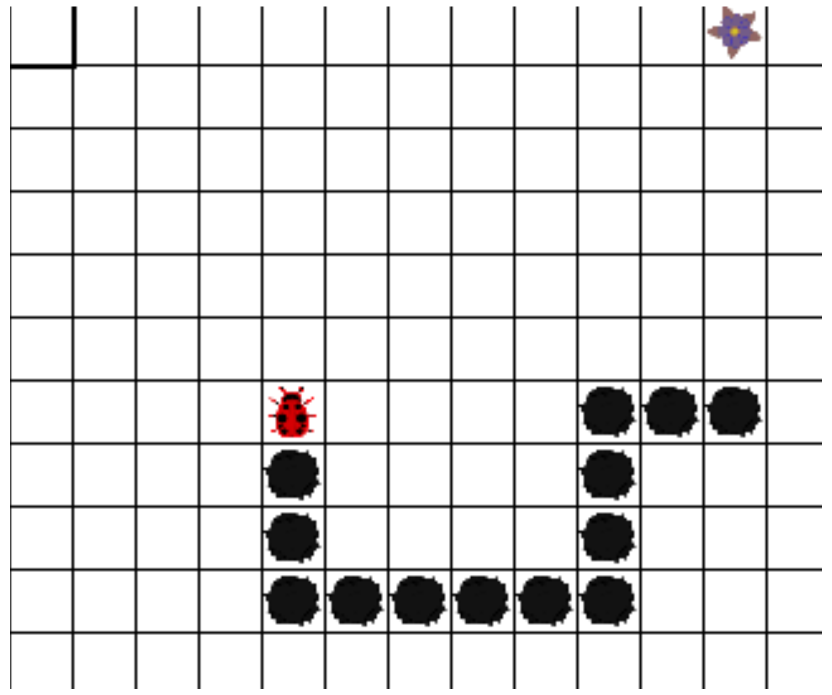
** this is a more general setting*



<u>COLT Conjecture of Jiang & Agarwal, 2018:</u>

Any algorithm must suffer ~H dependence in terms of sample complexity and regret in the Goal MDP setting

Long Horizon MDPs



Standard Setting $r \in [0,1]$

Goal MDP Setting* $r \geq 0, \quad \sum_{t=1}^{H} r_t \leq 1$

** this is a more general setting*

COLT Conjecture of Jiang & Agarwal, 2018:

Any algorithm must suffer ~H dependence in terms of sample complexity and regret in the Goal MDP setting

Our algorithm yields
no horizon dependence in the regret bound for the setting of the COLT conjecture without being informed of the setting.

Effect of MDP Stochasticity

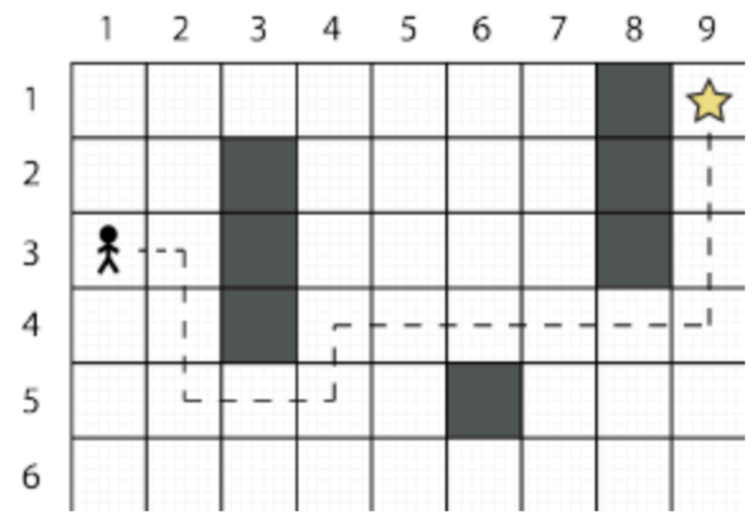
Effect of MDP Stochasticity

Stochasticity in the Transition Dynamics

Effect of MDP Stochasticity

Stochasticity in the Transition Dynamics

Deterministic MDP

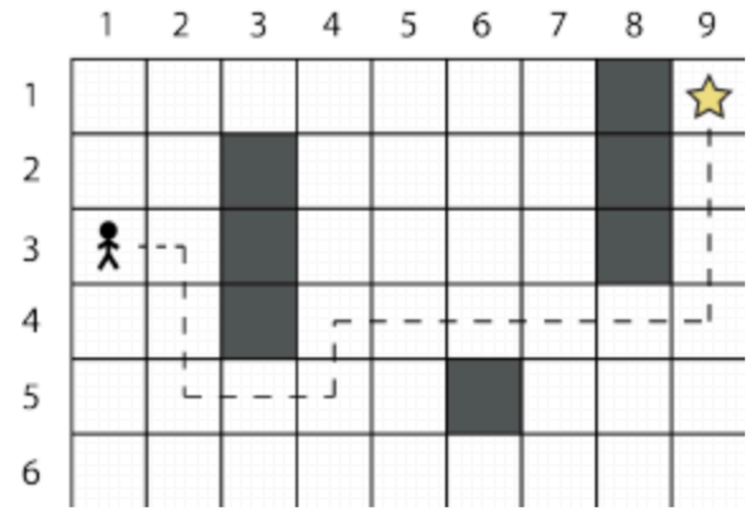


$\tilde{O}(SAH^2)$

Effect of MDP Stochasticity

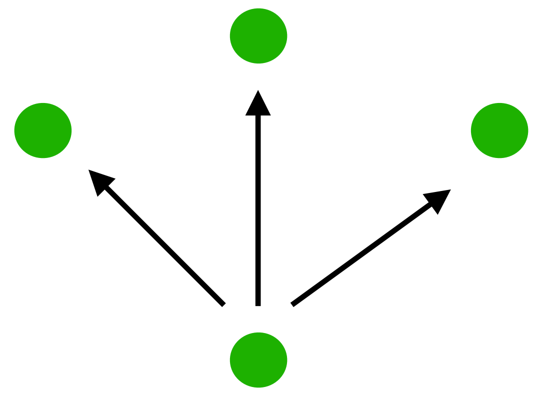
Stochasticity in the Transition Dynamics

Deterministic MDP



$\tilde{O}(SAH^2)$

Bandit Like Structure

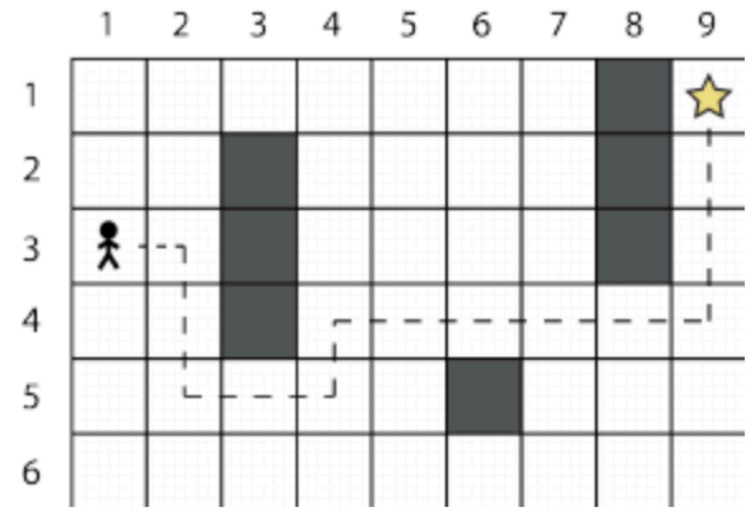


$\tilde{O}(\sqrt{SAT} + [\dots])$

Effect of MDP Stochasticity

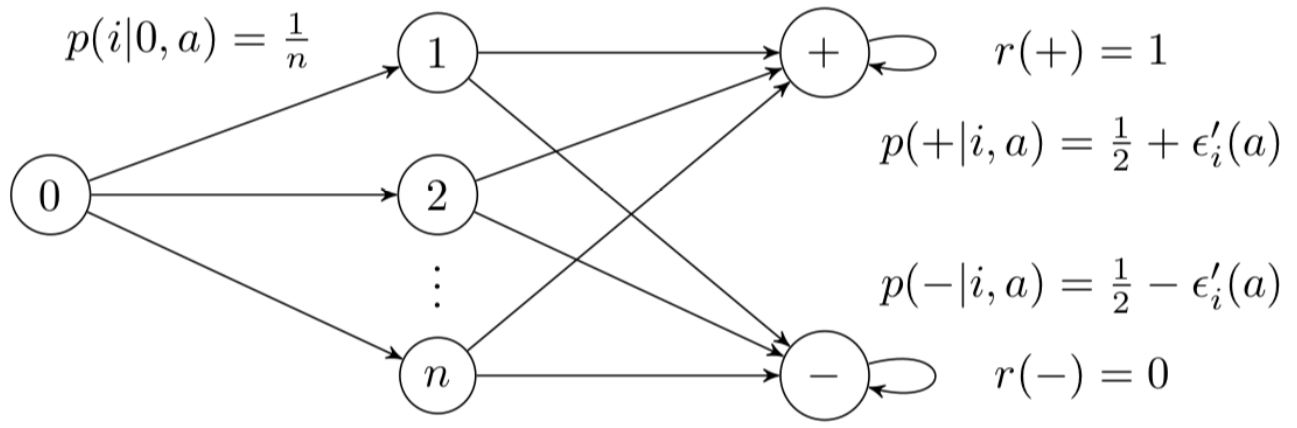
Stochasticity in the Transition Dynamics

Deterministic MDP



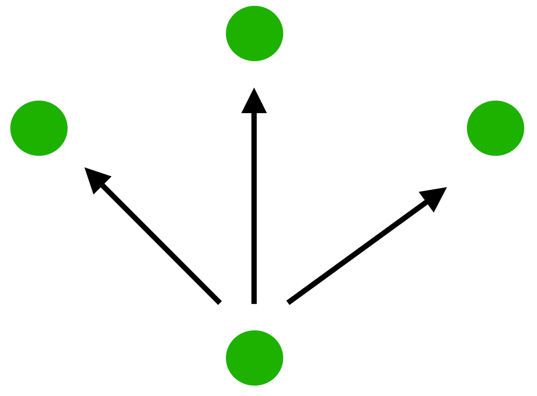
$\tilde{O}(SAH^2)$

Hard Instances of the Lower Bound



$\tilde{O}(\sqrt{HSAT} + [\dots])$

Bandit Like Structure

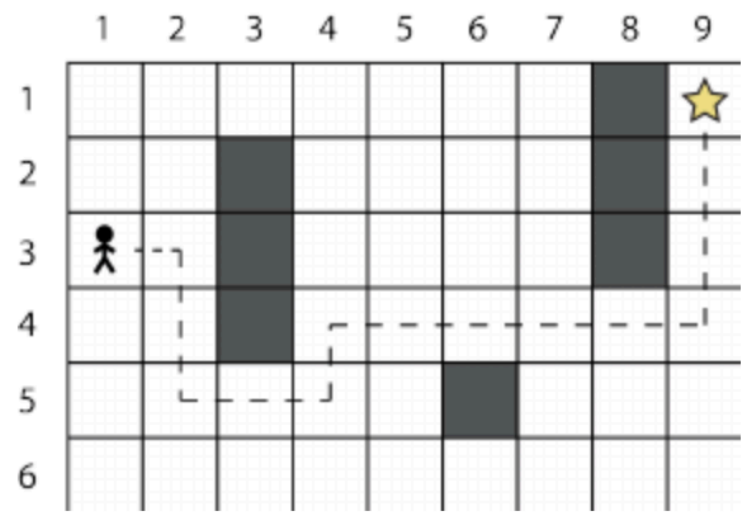


$\tilde{O}(\sqrt{SAT} + [\dots])$

Effect of MDP Stochasticity

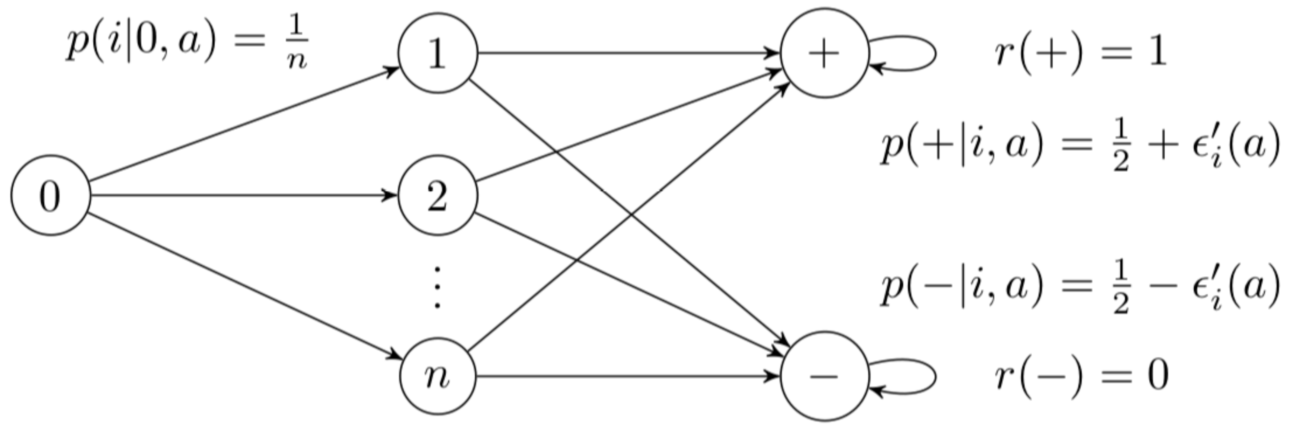
Stochasticity in the Transition Dynamics

Deterministic MDP



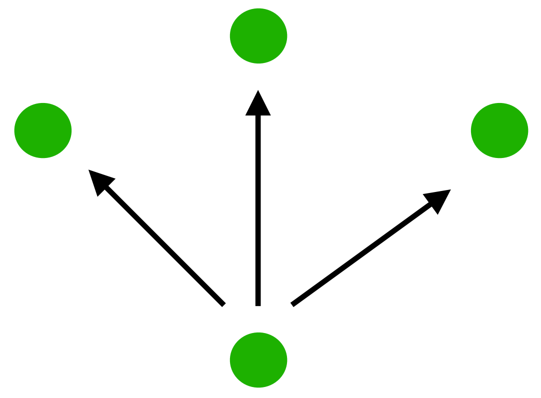
$\tilde{O}(SAH^2)$

Hard Instances of the Lower Bound



$\tilde{O}(\sqrt{HSAT} + [\dots])$

Bandit Like Structure



$\tilde{O}(\sqrt{SAT} + [\dots])$

<u>Our algorithm matches in dominant terms the best performance for each setting</u>

Related Work (infinite horizon)

Related Work (infinite horizon)

In mixing domains:
- (*Talebi et al, 2018)*
- *(Ortner, 2018)*

May not improve over worst-case:
-*(Maillard et al, 2014)*

With domain knowledge:
- [REGAL] *(Bartlett et al, 2010)*
- [SCAL] *(Fruit et al, 2018)*

Related Work (infinite horizon)

In mixing domains:
- (*Talebi et al, 2018)*
- *(Ortner, 2018)*

May not improve over worst-case:
-*(Maillard et al, 2014)*

With domain knowledge:
- [REGAL] *(Bartlett et al, 2010)*
- [SCAL] *(Fruit et al, 2018)*

Conclusion

- Episodic tabular MDP instance dependent bound without knowledge of the environment
- Insights into hardness of RL; provable improvements in many settings of interest

Related Work (infinite horizon)

In mixing domains:
- *(Talebi et al, 2018)*
- *(Ortner, 2018)*

May not improve over worst-case:
-*(Maillard et al, 2014)*

With domain knowledge:
- [REGAL] *(Bartlett et al, 2010)*
- [SCAL] *(Fruit et al, 2018)*

Conclusion

- Episodic tabular MDP instance dependent bound without knowledge of the environment
- Insights into hardness of RL; provable improvements in many settings of interest

near-deterministic MDPs

Bandit-structure

long horizon MDPs

limited range of optimal value function

limited variability in value function among successor states