Policy Certificates: Towards Accountable Reinforcement Learning

Christoph Dann, CMU

Lihong Li, Google Research

Wei Wei, Stanford University

Emma Brunskill

Minimax-Optimal PAC Bounds

Key contribution: new algorithm for **episodic tabular MDPs** with

PAC Bound

$$\tilde{O}\left(\frac{SAH^2}{\epsilon^2}+\frac{S^2AH^3}{\epsilon}\right)$$

High-prob. Regret Bound

$$\tilde{O}\left(\sqrt{SAH^2T}+S^2AH^3\right)$$

S: #states, A: #actions, H: episode length, T: #episodes, ϵ: accuracy

Minimax-Optimal PAC Bounds

Key contribution: new algorithm for **episodic tabular MDPs** with

PAC Bound

$$\tilde{O}\left(\frac{SAH^2}{\epsilon^2}+\frac{S^2AH^3}{\epsilon}\right)$$

High-prob. Regret Bound

$$\tilde{O}\left(\sqrt{SAH^2T}+S^2AH^3\right)$$

First minimax-optimal! (for small ϵ)

S: #states, A: #actions, H: episode length, T: #episodes, ϵ: accuracy

Prior work:

$$\frac{S^2AH^2}{\epsilon^2}$$

[DLB '17]

$$\frac{SAH^3}{\epsilon^2}$$

[DB '15]

Minimax-Optimal PAC Bounds

Key contribution: new algorithm for **episodic tabular MDPs** with

PAC Bound	High-prob. Regret Bound
$\tilde{O}\left(\frac{SAH^2}{\epsilon^2}+\frac{S^2AH^3}{\epsilon}\right)$	$\tilde{O}\left(\sqrt{SAH^2T}+S^2AH^3\right)$
First minimax-optimal! (for small ϵ)	Matches existing + improves for large H

S: #states, A: #actions, H: episode length, T: #episodes, ϵ: accuracy

Prior work:

$$\frac{S^2AH^2}{\epsilon^2}$$

[DLB '17]

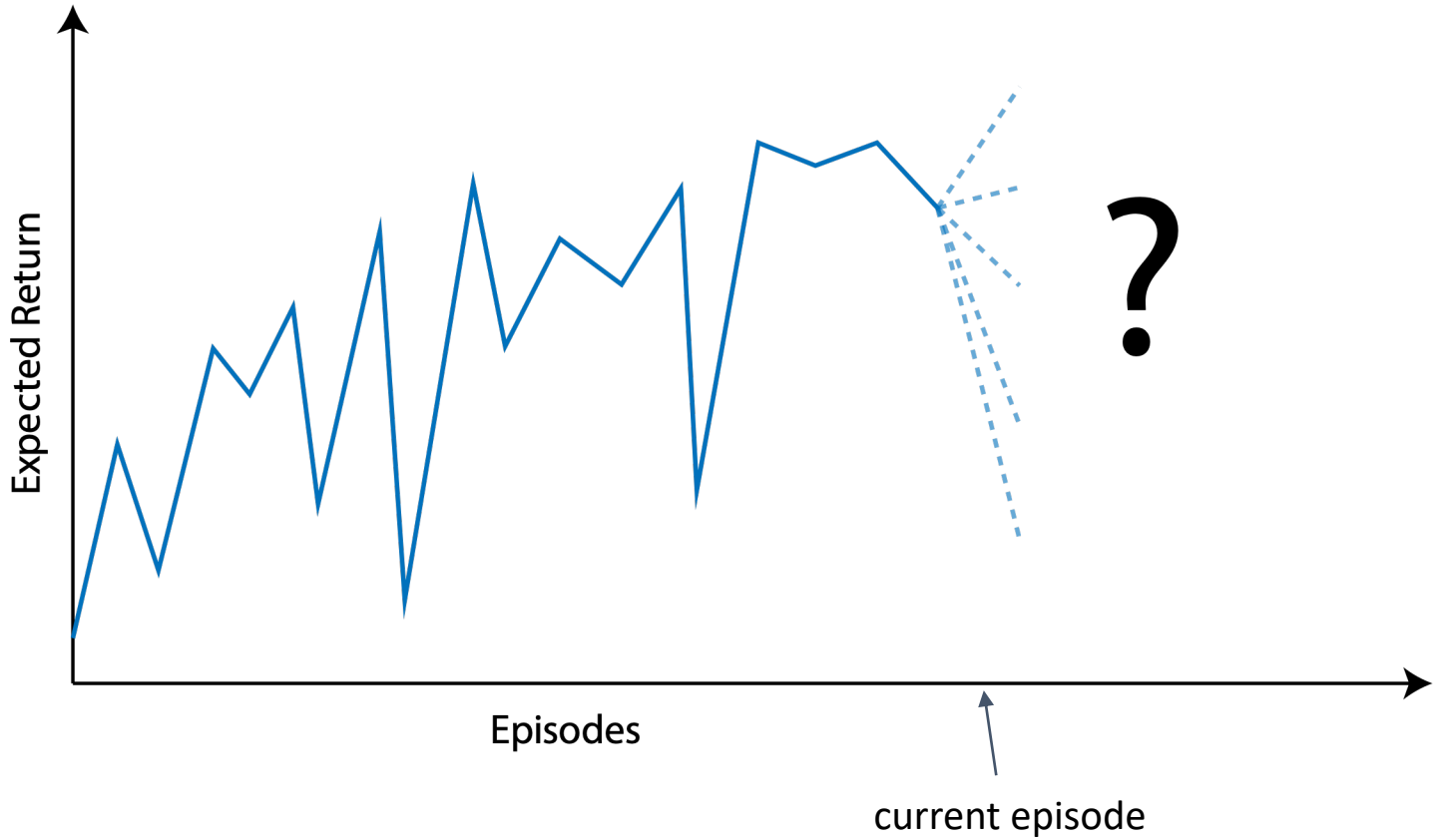
$$\frac{SAH^3}{\epsilon^2}$$

[DB '15]

$$\sqrt{SAH^2T}+\sqrt{H^3T}+S^2AH^2$$

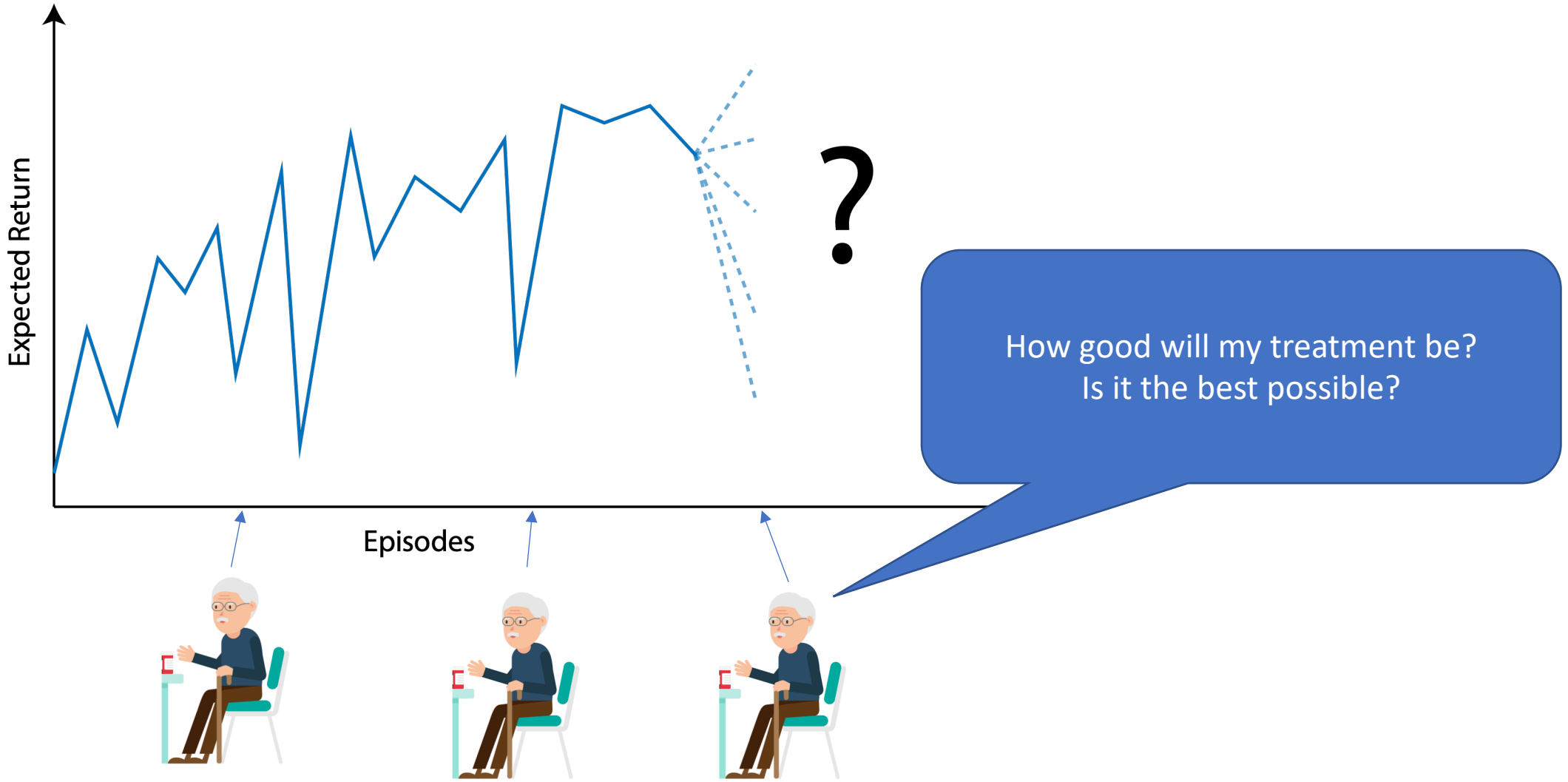
[AOM '17]

Motivation: Need for Accountability in Online RL

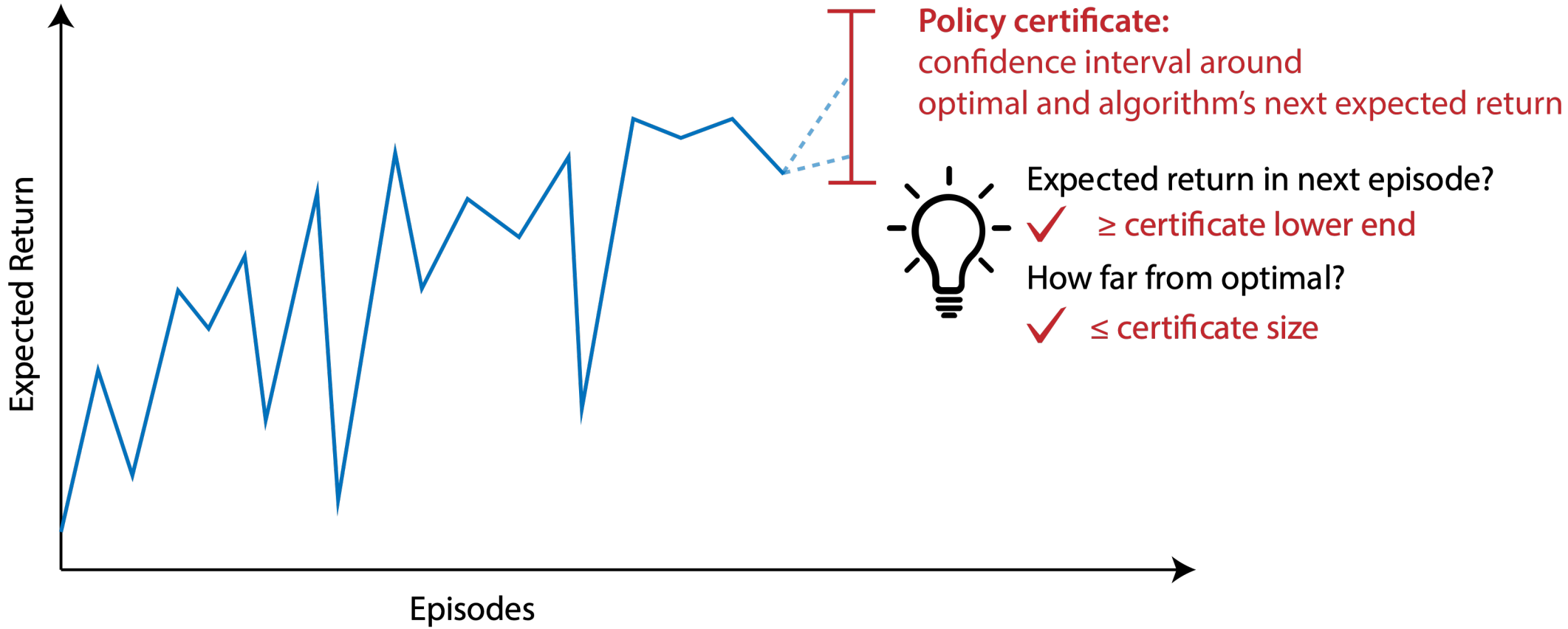


Even with PAC + regret bounds: expected return in next episode during learning unknown

Motivation: Need for Accountability in Online RL



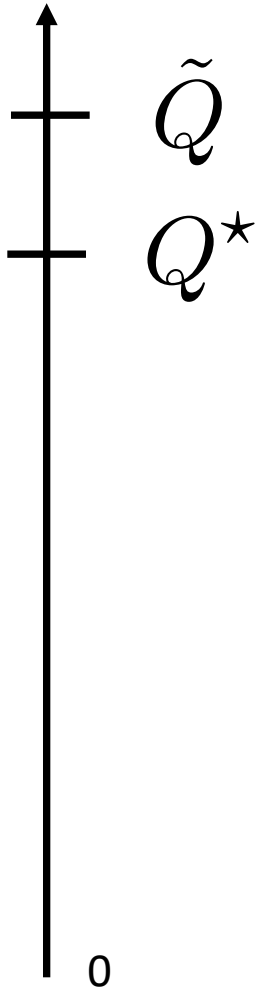
Our Proposal: Algorithms output policy certificates before each episode



Algorithms with policy certificates

Natural extension of model-based optimistic algorithms

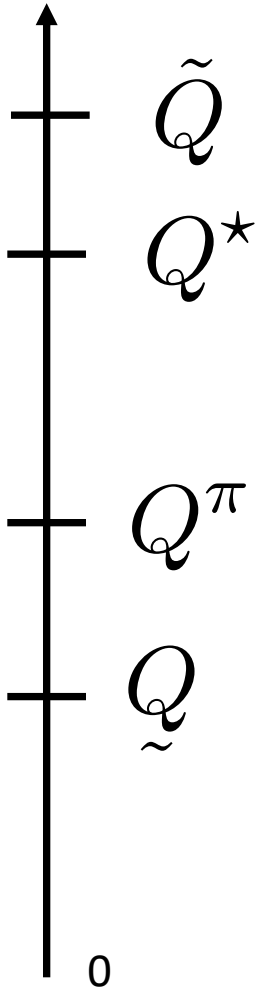
1. UCB on optimal value function
$$\tilde{Q}(s,a) \geq Q^\star(s,a)$$
2. Greedy Policy $\pi = \mathrm{greedy}(\tilde{Q})$



Algorithms with Policy Certificates

Natural extension of model-based optimistic algorithms

1. UCB on optimal value function
$$\tilde{Q}(s,a) \geq Q^\star(s,a)$$
2. Greedy Policy $\pi = \mathrm{greedy}(\tilde{Q})$
3. LCB on value function of current policy
$$\underset{\sim}{Q}(s,a) \leq Q^\pi(s,a)$$
4. Output certificate $[\underset{\sim}{Q}(s_0,a_0), \tilde{Q}(s_0,a_0)]$



Symbiosis of Optimism and Certificates

Certificates:

- Challenge: Q^π random

- Insight from optimism:
  $Q^\pi \to Q^\star$ at known rate

Symbiosis of Optimism and Certificates

Certificates:

- Challenge: Q^π random

- Insight from optimism: $Q^\pi \rightarrow Q^\star$ at known rate

Optimism:

- Challenge: exploration bonus depends on $\tilde{Q}(s', a') - Q^\star(s', a')$

- Insight from certificates: bound by $\tilde{Q}(s', a') - \underset{\sim}{Q}(s', a')$

Symbiosis of Optimism and Certificates

Certificates:

- Challenge: Q^π random

- Insight from optimism: $Q^\pi \to Q^\star$ at known rate

Optimism:

- Challenge: exploration bonus depends on $\tilde{Q}(s', a') - Q^\star(s', a')$

- Insight from certificates: bound by $\tilde{Q}(s', a') - \underset{\sim}{Q}(s', a')$

More accountable algorithms through accurate policy certificates

Better exploration bonuses yield minimax-optimal PAC & regret bounds