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

Setting: episodic tabular RL
Goal: automatically inherit instance-dependent regret bounds

Exploration in RL
Learn quickly how to play near optimally
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

\[ \tilde{O}(H\sqrt{SA}) \quad \tilde{O}(T) \]

Efficient Exploration

No Intelligent Exploration

\( \tilde{O}(HS\sqrt{AT}) \) (naive greedy)

(UCRL2, Jaksch 2010)
State of the Art Regret Bounds for Episodic Tabular MDPs

Efficient Exploration

\( \tilde{O}(H\sqrt{SAT}) \)
(Dann 2015)

\( \tilde{O}(HS\sqrt{AT}) \)
(UCRL2, Jaksch 2010)

No Intelligent Exploration

\( \tilde{O}(T) \)
(naive greedy)
State of the Art Regret Bounds for Episodic Tabular MDPs

- Efficient Exploration
  - $\tilde{O}(S\sqrt{HAT})$ (Dann 2017)
  - $\tilde{O}(H\sqrt{SAT})$ (Dann 2015)
  - $\tilde{O}(HS\sqrt{AT})$ (UCRL2, Jaksch 2010)

- No Intelligent Exploration
  - $\tilde{O}(T)$ (naive greedy)
State of the Art Regret Bounds for Episodic Tabular MDPs

Efficient Exploration

\[ \tilde{O}(\sqrt{SA/T}) \]
(Azar 2017)

\[ \tilde{O}(S\sqrt{HAT}) \]
(Dann 2017)

\[ \tilde{O}(H\sqrt{SAT}) \]
(Dann 2015)

\[ \tilde{O}(HS\sqrt{AT}) \]
(UCRL2, Jaksch 2010)

No Intelligent Exploration

\[ \tilde{O}(T) \]
(naive greedy)
State of the Art Regret Bounds for Episodic Tabular MDPs

Lower Bound  Efficient Exploration  No Intelligent Exploration

$\tilde{O}(\sqrt{HSAT})$ (Azar 2017)  $\tilde{O}(H\sqrt{SAT})$ (Dann 2015)  $\tilde{O}(T)$ (naive greedy)

$\Omega(\sqrt{HSAT})$ (Lower Bound)  $\tilde{O}(S\sqrt{HAT})$ (Dann 2017)  (UCRL2, Jaksch 2010)
State of the Art Regret Bounds for Episodic Tabular MDPs

Problem Dependent Analysis | Lower Bound | Efficient Exploration | No Intelligent Exploration
---|---|---|---
\(\tilde{O}(\sqrt{Q^{*}SAT})\) (Our work) | \(\Omega(\sqrt{HSAT})\) (Lower Bound) | \(\tilde{O}(\sqrt{HSAT})\) (Dann 2017) | \(\tilde{O}(T)\) (naive greedy)
\(\tilde{O}(\sqrt{HSAT})\) (Azar 2017) | \(\tilde{O}(S\sqrt{HAT})\) (Dann 2015) | \(\tilde{O}(HS\sqrt{AT})\) (UCRL2, Jaksch 2010)
Main Result
Main Result

\((s, a)\)
Main Result

\[(s, a)\]
Main Result

\[ V^\star(s_i^+) \]

\[(s, a) \rightarrow t \rightarrow t+1 \]

\[ V^\star(s_1^+) \]
\[ V^\star(s_2^+) \]
\[ V^\star(s_3^+) \]
\[ Q^* = \max_{s,a} \text{Var}_{s^* \sim p(s,a)} V^*(s^*) \]

The main result is illustrated with a diagram showing transitions from state-action pairs to their corresponding values at each time step. The diagram visualizes the Q-learning process, where the optimal Q-value function is the maximum over all state-action pairs, considering the expected value of the next state under the policy distribution.
Main Result

\[ Q^* = \max_{s, a} \mathbb{V} \mathbb{a}_{s^+ \sim p(s, a)} V^*(s^+) \]

\[ \mathbb{Q} \star = \mathbb{m}_a x s, a + \sim p(s, a) V^*(s^+) \]

\[ t \rightarrow t+1 \]

\((s, a)\)

\[ V^*(s^+_1) \]

\[ V^*(s^+_2) \]

\[ V^*(s^+_3) \]

\[ r_1 \rightarrow r_2 \rightarrow \ldots \rightarrow r_H \leq \mathcal{G} \]
Main Result

\[ \mathbb{Q}^* = \max_{s,a} \text{Var}_{s \sim p(s,a)} V^*(s^+) \]

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

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

\[ r_1 + r_2 + \ldots + r_H \leq G \]
Main Result: An algorithm with a (high probability) regret bound:

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

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* \( r \geq 0, \sum_{t=1}^{H} r_{t} \leq 1 \)

* this is a more general setting
Long Horizon MDPs

Standard Setting \( r \in [0,1] \)

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

* this is a more general setting

COLT Conjecture of Jiang & Agarwal, 2018:

Any algorithm must suffer \( \sim H \) 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 \( \sim H \) 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

Bandit Like Structure

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

\( \tilde{O}(\sqrt{SAT} + [\ldots]) \)
Effect of MDP Stochasticity

Stochasticity in the Transition Dynamics

Deterministic MDP

\[ \tilde{O}(SAH^2) \]

Hard Instances of the Lower Bound

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

Bandit Like Structure

\[ \tilde{O}(\sqrt{SAT} + [\ldots]) \]
Effect of MDP Stochasticity

Stochasticity in the Transition Dynamics

Deterministic MDP

Hard Instances of the Lower Bound

Bandit Like Structure

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

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

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

Our algorithm matches in dominant terms the best performance for each setting.
Related Work (infinite horizon)
## Related Work (infinite horizon)

<table>
<thead>
<tr>
<th>In mixing domains:</th>
<th>May not improve over worst-case:</th>
<th>With domain knowledge:</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Ortner, 2018)</td>
<td></td>
<td>[SCAL] (Fruit et al, 2018)</td>
</tr>
</tbody>
</table>
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