why is posterior sampling better than optimism for reinforcement learning?

Ian Osband\textsuperscript{1}  Benjamin Van Roy\textsuperscript{1}

\textsuperscript{1}Stanford University

ICML, 2017
Presenter: Beilun Wang
1 Introduction
   - Motivation
   - Previous Solutions
   - Contributions

2 Background
   - Random finite-horizon MDP

3 Main conclusion

4 Summary
Outline

1. Introduction
   - Motivation
   - Previous Solutions
   - Contributions

2. Background
   - Random finite-horizon MDP

3. Main conclusion

4. Summary
Motivation:

- Computational results demonstrate that posterior sampling for reinforcement learning (PSRL) dramatically outperforms existing algorithms driven by optimism.
- Need theoretical proofs about this result
- Regret bounds comparison
Problem Setting:

- **Input**: A reinforcement learning algorithm
- **Target**: finite-horizon episodic Markov decision processes
- **Output**: A regret bound
Outline

1 Introduction
   - Motivation
   - Previous Solutions
   - Contributions

2 Background
   - Random finite-horizon MDP

3 Main conclusion

4 Summary
optimism in the face of uncertainty (OFU)

Old bound \(\tilde{O}(HS\sqrt{AT})\)

\(H\): Horizen, the number of steps within an episode
\(S\): the number of states
\(A\): the number of actions
\(T\): the number of steps

The authors want to improve the bound to \(\tilde{O}(\sqrt{HSAT})\)
Outline

1 Introduction
   - Motivation
   - Previous Solutions
   - Contributions

2 Background
   - Random finite-horizon MDP

3 Main conclusion

4 Summary
Contributions

- PSRL is no worse than OFU
- PSRL achieves the better Bayesian regret bound $\tilde{O}(H\sqrt{SAT})$
Outline

1 Introduction
   - Motivation
   - Previous Solutions
   - Contributions

2 Background
   - Random finite-horizon MDP

3 Main conclusion

4 Summary
Definition: Random finite-horizon MDP / Bayesian reinforcement learning

\[ M = (S, A, R^*, P^*, H, \rho) \]
- The state space \( S \)
- The Action space \( A \)
- \( H \) is the number of steps within an episode
- \( \rho \) is the initial state distribution
- A new state reward \( r_h \sim R^*(s_h, a_h) \)
- A new transition \( s_{h+1} \sim P^*(s_h, a_h) \)
Value function and policy function in Bayesian reinforcement learning

- state-action value function for each period $h$:

  \[
  Q_{\mu,h}^M(s, a) := \mathbb{E}_{M,\mu}\left[\sum_{j=h}^{H} \bar{r}_M^{M}(s_j, a_j)s_h = s, a_h = a\right]
  \]  

  (1)

  where $\bar{r}_M^{M}(s, a) = \mathbb{E}[r | r \sim R^M(s, a)]$

  - $\mu$ is a policy
Value function and policy function in Bayesian reinforcement learning

- $V_{\mu,h}^M(s) := Q_{\mu,h}^M(s, \mu(s, h))$
- Optimal policy for MDP $M$: $\mu^M \in \arg \max_{\mu} V_{\mu,h}^M(s)$
- History prior to time $t$: $H_t$
- $s_{kh} = s_t$, where $t = (k1)H + h$.
- $H_{kh} = H_t$.
- An RL algorithm $\{\pi_k|k = 1, 2, \ldots\}$
Regret Bound

- Regret:
  
  \[
  \text{Regret}(T, \pi, M^*) := \sum_{k=1}^{[T/H]} \Delta_k
  \]
  
  where
  
  \[
  \Delta_k := \sum_S \rho(s)(V_{\mu^*,1}^M(s) - V_{\mu_k,1}^M(s))
  \]

- true MDP \( M^* \)

- \( \mu^* = \mu^M \)
Bayes Regret

\[ \text{BayesRegret}(T, \pi, \phi) := \mathbb{E}[\text{Regret}(T, \pi, M^*) | M^* \sim \phi] \]  

(4)
Algorithm 1 OFU RL

**Input:** confidence set constructor $\Phi$

1: for episode $k = 1, 2, \ldots$ do
2: Construct confidence set $\mathcal{M}_k = \Phi(\mathcal{H}_{k,1})$
3: Compute $\mu_k \in \arg\max_{\mu, M \in \mathcal{M}_k} V_{\mu,1}^M$
4: for timestep $h = 1, \ldots, H$ do
5: take action $a_{kh} = \mu_k(s_{kh}, h)$
6: update $H_{kh+1} = \mathcal{H}_{kh} \cup (s_{kh}, a_{kh}, r_{kh}, s_{kh+1})$
7: end for
8: end for
Algorithm 2 PSRL

Input: prior distribution $\phi$

1: for episode $k = 1, 2, \ldots$ do
2: Sample MDP $M_k \sim \phi(\cdot | \mathcal{H}_{k1})$
3: Compute $\mu_k \in \arg \max_{\mu} V_{\mu, 1}^{M_k}$
4: for timestep $h = 1, \ldots, H$ do
5: take action $a_{kh} = \mu_k(s_{kh}, h)$
6: update $H_{kh+1} = \mathcal{H}_{kh} \cup (s_{kh}, a_{kh}, r_{kh}, s_{kh+1})$
7: end for
8: end for
PSRL matches OFU-RL in BayesRegret

- If OFU-RL has the regret
  \[ \text{Regret}(T, \pi^{\text{opt}}, M^*) \leq f(S, A, H, T, \delta) \]
- Then PSRL has the Bayes regret
  \[ \text{BayesRegret}(T, \pi^{\text{PSRL}}, \phi) \leq 2f(S, A, H, T, \delta) + 2 \]
Regret bound improvement

- PSRL achieves the better Bayesian regret bound $\tilde{O}(H\sqrt{SAT})$
- It is possible to have bound $\tilde{O}(\sqrt{HSAT})$ with additional assumptions
- This bound cannot be improved.
Summary

- PSRL is no worse than OFU
- PSRL achieves the better Bayesian regret bound $\tilde{O}(H\sqrt{SAT})$