why is posterior sampling better than optimism for reinforcement learning?

Ian Osband[1] Benjamin Van Roy[1]

[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

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:

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

Previous Solutions

- optimism in the face of uncertainty (OFU)
- Old bound $\tilde{O}(HS\sqrt{AT})$
- H: Horizen, the number of steps within a 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

- $M = (\mathcal{S}, \mathcal{A}, R^*, P^*, H, \rho)$
- The state space $\mathcal{S}$
- The Action space $\mathcal{A}$
- H is the number of steps within an episode
- ρ 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:

$$Q^M_{\mu,h}(s,a) := \mathbb{E}_{M,\mu}[\sum_{j=h}^{H} \bar{r}^M(s_j,a_j)s_h = s, a_h = a] \tag{1}$$

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

Value function and policy function in Bayesian reinforcement learning

- $V^M_{\mu,h}(s) := Q^M_{\mu,h}(s, \mu(s,h))$
- Optimal policy for MDP M: $\mu^M \in \arg\max_{\mu} V^M_{\mu,h}(s)$
- History prior to time t: $\mathcal{H}_t$
- $s_{kh} = s_t$, where $t = (k1)H + h$.
- $\mathcal{H}_{kh} = \mathcal{H}_t$.
- An RL algorithm $\{\pi_k | k = 1, 2, \dots\}$

Regret Bound

- Regret:

$$\text{Regret}(T, \pi, M^*) := \sum_{k=1}^{[T/H]} \Delta_k \tag{2}$$

- where

$$\Delta_k := \sum_{\mathcal{S}} \rho(s)(V^{M^*}_{\mu^*,1}(s) - V^{M^*}_{\mu_k,1}(s)) \tag{3}$$

- true MDP M^*
- $\mu^* = \mu^{M^*}$

Bayes Regret

$$\text{BayesRegret}(T, \pi, \phi) := \mathbb{E}[\text{Regret}(T, \pi, M^*) | M^* \sim \phi] \tag{4}$$

OFU-RL

Algorithm 1 OFU RL

Input: confidence set constructor Φ

for episode $k = 1, 2, ..$ **do**
  Construct confidence set $\mathcal{M}_k = \Phi(\mathcal{H}_{k1})$
  Compute $\mu_k \in \text{argmax}_{\mu, M \in \mathcal{M}_k} V^M_{\mu,1}$
  for timestep $h = 1, .., H$ **do**
    take action $a_{kh} = \mu_k(s_{kh}, h)$
    update $H_{kh+1} = \mathcal{H}_{kh} \cup (s_{kh}, a_{kh}, r_{kh}, s_{kh+1})$
  end for
end for

PSRL

Algorithm 2 PSRL

Input: prior distribution ϕ

for episode $k = 1, 2, ..$ **do**
  Sample MDP $M_k \sim \phi(\cdot \mid \mathcal{H}_{k1})$
  Compute $\mu_k \in \text{argmax}_\mu V_{\mu,1}^{M_k}$
  for timestep $h = 1, .., H$ **do**
    take action $a_{kh} = \mu_k(s_{kh}, h)$
    update $H_{kh+1} = \mathcal{H}_{kh} \cup (s_{kh}, a_{kh}, r_{kh}, s_{kh+1})$
  end for
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 $\tilde{O}(H\sqrt{SAT})$