

Exploration in RL

Traditional Exploration Strategies for MDPs

Marius Lindauer



Leibniz
Universität
Hannover



Winter Term 2021

Recap: Bandit Exploration

- ▶ Optimistic initialization
- ▶ Optimism in the face of uncertainty (Upper Confidence bounds)
- ▶ Probability matching (Thompson Sampling)

Optimistic Initialization: Model-free RL

- ▶ Initialize action-value function $Q(s, a)$ to $\frac{r_{max}}{1-\gamma}$
- ▶ Run favorite model-free RL algorithm
 - ▶ Monte-carlo method
 - ▶ Sarsa
 - ▶ Q-Learning
- ▶ Encourages systematic exploration of states and actions

Upper Confidence Bounds: Model-free RL

- ▶ Maximize UCB on action-value function $Q^\pi(s, a)$

$$a_t \in \arg \max_{a \in A} Q(s_t, a) + U(s_t, a)$$

- ▶ Estimate uncertainty in policy evaluation (easy)
 - ▶ Ignores uncertainty from policy improvement
- ▶ Maximize UCB on optimal action-value function $Q^*(s, a)$

$$a_t \in \arg \max_{a \in A} Q(s_t, a) + U_1(s_t, a) + U_2(s_t, a)$$

- ▶ Estimate uncertainty in policy evaluation (easy)
 - ▶ plus uncertainty from policy improvement (hard)

Bayesian Model-based RL

- ▶ Maintain posterior distribution over MDP models
- ▶ Estimate both transitions and rewards $\mathbb{P}[P, R \mid h_t]$
 - ▶ where $h_t = s_1, a_1, r_2, \dots, s_t$ is the history
- ▶ Use posterior to guide exploration
 - ▶ Upper confidence bounds (Bayesian UCB)
 - ▶ Probability matching (Thompson sampling)

Thompson Sampling: Model-based RL

- ▶ Thompson sampling implements probability matching

$$\begin{aligned}\pi(s, a \mid h_t) &= \mathbb{P}[Q^*(s, a) > Q^*(s, a'), \forall a' \neq a \mid h_t] \\ &= \mathbb{E}_{P, R \mid h_t} \left[\mathbf{1}(a \in \arg \max_{a \in A} Q^*(s, a)) \right]\end{aligned}$$

1. Use Bayes law to compute posterior $\mathbb{P}[P, R \mid h_t]$
2. Sample an MDP P, R from posterior
3. Solve MDP using favorite planning algorithm to get $Q^*(s, a)$
4. Select optimal action for sample MDP: $a_t \in \arg \max_{a \in A} Q^*(s_t, a)$