Generalized Optimistic Q-Learning with Provable Efficiency
Grigory Neustroev and Mathijs M de Weerdt {g.neustroev, m.m.deweerdt}@tudelft.nl
This is an extended abstract of (Neustroev and de Weerdt, 2020)

1 What is this paper about?
Reinforcement learning
Learns via rewards from interactions with an environment (samples)

Model-free (Q-learning)
Uses less memory, is fast
Generally is not sample-efficient

Model-based
Needs more memory, is slow
Is often sample-efficient

Optimistic methods
Some are provably efficient

What did we do?
We proved that optimistic Q-learning is generally efficient

2 How do we measure efficiency?
value (collected rewards)
optimal value $V^*$

Regret is a measure of efficiency
Efficient methods have sublinear total regret, which means that regret per sample decreases over time

3 What is the main result?
We prove that for optimistic Q-learning in general:

$$ R = O \left( \mu \cdot \left( X + B + E \right) \right) $$

4 What is the intuition behind this result?

**Magnitude** shows how regret scales when the problem values change. For example, $\mu = (1 - \gamma)^{-1} \cdot (V_{\text{max}} - V_{\text{min}})$ for $\gamma$-discounted problems.

Regret is proportionate to the **problem size** $X = |S \times A|$, because we need to explore all of the state-action combinations.

Optimistic methods add special bonuses to Q-values to make them look better (i.e., optimistic). This results in a **bonus effect** $B \sim X \cdot \theta(T/X)$.

**Estimation error** $E \sim \sqrt{T \ln(TX)}$ arises because observations are used instead of expected rewards and state changes in the Bellman equation:

**Bellman equation**

$$ Q^*(s, a) = \mathbb{E}_{p(s', a)} \left[ r(s'|s, a) + \gamma \max_{a' \in A} Q^*(s', a') \right] $$

**Q-learning update**

$$ Q(s_t, a_t) \leftarrow r_t + \gamma \max_{a' \in A} Q(s_{t+1}, a') $$

5 What can we do with this theory?

For UCB-H, Jin et al. (2018) show that $R = O(H^2 \sqrt{TX})$. Using our framework, we find that $\mu = H^2$ and $B = \sqrt{TX}$. Because $X, E = o(B)$, we conclude that $R = O(\mu B) = O(H^2 \sqrt{TX})$. Our proof is shorter and easier to interpret.

We also design a new optimistic method, UCB-H⁺, which outperforms UCB-H in two problems, frozen lake and automobile replacement:

References