

Near-optimal Regret Bounds for Reinforcement Learning

ID: T2



MU-Leoben

Peter Auer, Thomas Jaksch, and Ronald Ortner

- New analysis of “**Optimism in the face of uncertainty**”.
- **Regret** =
Sum of missed rewards (also during training!) compared to an optimal policy.
- **New complexity parameter** for MDPs: The **diameter D** .
How long does it take to travel from one state to another state?
- **Best known bounds** for undiscounted RL with finite state/action space $\mathcal{S} \times \mathcal{A}$.

- Regret after T steps is at most

$$\text{const} \cdot DS \sqrt{TA \log(T)} .$$

- PAC-like bound: The average per step regret is at most ϵ after

$$T \geq \text{const} \cdot \frac{D^2 S^2 A}{\epsilon^2} \log \left(\frac{DSA}{\epsilon} \right) \text{ steps.}$$