**Introduction**

**What do we do?**
- Exploration in continuous MDPs
- With theoretical guarantees

**How?**
- Using exploration bonus and discretization

**Online Learning in MDPs**
- Markov Decision Process $M = \{S, A, r, p\}$
- Optimality criterion: average reward

For any policy $\pi$ starting from $s \in S$, the expected time $\lim_{t \to \infty} \mathbb{E}_{T} [T(s, a)]$ is unknown.

**Diameter:** $D = \min_{s, s' \in S} \mathbb{E}_{T} [T(s, a)]$

**Prior Knowledge on the Bias Span**
- Provides a sense of what is realizable in the true MDP
- Averts over-optimism
- Necessary to define the exploration bonus

**UCRL2-like Exploration**

For $k = 1, 2, \ldots$

1. Estimation of model and uncertainty
   \[ M_k = \{ M = (S, A, \tilde{r}, \tilde{p}) : \tilde{p}(s, a, \sigma) \in B_{\alpha}(s, a), \tilde{r}(s, a, \sigma) \in B_{H}(s, a) \} \]

2. Planning for optimistic policy
   \[ (M_k, \pi_k) = \arg \max_{M \in M_k} \max_{\pi} \mathbb{E}_{g^*}(M) \]

3. Execution of policy $\pi_k$
   - execute action $a_t \sim \pi_k$
   - observe reward $r_t$ and next state $s_{t+1}$

**SCAL+: tabular MDP**

- Exploration bonus: Used in deep RL [Bellemare et al. 2016, Tang et al. 2017] and/or when the intrinsic horizon is known [Auer et al. 2017, Jin et al. 2018]

For $k = 1, 2, \ldots$

1. Estimation of empirical model
   \[ M_k = (S, A, \tilde{p}, \tilde{r}) \]

2. Planning for optimistic policy
   \[ \pi_k = \arg \max_{g^* \in \Gamma} \mathbb{E}_{g^*}(M_k) \]

**Numerical Results**

- **GARNET:** tabular MDP
- **SHIP STEERING:** tabular MDP
- **RIVER SWIM:** continuous MDP

**SCAL+: regret**

For any MDP such $\mathbb{E}_{g}(h^*) \leq c$, w.p. $1 - \delta$

\[ R(\text{SCAL}^+, T) \approx \tilde{O}(c\sqrt{\Delta T}) \]

- Continuous MDPs –
- First implementable algorithm with guarantees in cont. MDPs (lots details are missing here, see paper)
- More stable than model-free version