A Sliding-Window Approach for RL in MDPs with Arbitrarily Changing Rewards and Transitions

Pratik Gajane, Ronald Ortner, Peter Auer

pratik.gajane@unileoben.ac.at, ronald.ortner@unileoben.ac.at, auer@unileoben.ac.at

Montanuniversitat Leoben

Introduction and motivation

- MDP : standard model for problems in decision making with uncertainty like RL.
- In classical MDPs, stochastic state-transition dynamics and reward functions remain fixed.
- ➜ Our setting (**Switching-MDP**): transition dynamics and reward functions change a certain number of times.

Problem Setting

- Classical MDP $M(\mathcal{S}, \mathcal{A}, p, F)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition probability p, reward function F.
- Learner selects action a in state s at time $t = 1, \ldots, T$
  - learner receives reward r_t drawn from dist. with mean $\bar{r}(s, a)$.
  - environment transitions into next state $s' \in \mathcal{S}$ according to $p(s' \mid s, a)$.
- Diameter $D(M_i) = \max_{s_1, s_2 \in \mathcal{S}, s_1 \neq s_2} \min_{\pi \in \Pi} \mathbb{E}[\tau(s_1, s_2, M_i, \pi)]$.
- **Switching-MDP** $\mathbf{M} = (\mathbb{S} = (M_0, \ldots, M_l), c = (c_1, \ldots, c_l))$
- At $t < c_1$, $\mathbf{M}$ is in its initial configuration $M_0(\mathcal{S}, \mathcal{A}, p_0, F_0)$.
- At time step $c_i \leqslant t < c_{i+1}$, $\mathbf{M}$ is in configuration $M_i(\mathcal{S}, \mathcal{A}, p_i, F_i)$.
- ➜ Goal of algorithm $\mathfrak{A}$ starting from an initial state s

  Minimize regret $\Delta(\mathbf{M}, \mathfrak{A}, s, T) = \sum_{t-1}^{T} (\rho^*_{\mathbf{M}}(t) - r_t)$

  $\rho^*_{\mathbf{M}}(t) :=$ Optimal average reward of M which is active at time t.

Analysis of SW-UCRL

Theorem 1. *Given a switching-MDP with l changes, the regret of* SW-UCRL *using window size W is upper-bounded with probability at least $1 - \delta$ by*

$$2lW + 66.12 \left[\frac{T}{\sqrt{W}}\right] DS \sqrt{A \log\left(\frac{T}{\delta}\right)},$$

where D = max of diameters of constituent MDPs.

Corollary 1. *Given a switching-MDP with l changes, the regret of* SW-UCRL *using $W^* = \left(\frac{16.53}{l} TDS\sqrt{A \log\left(\frac{T}{\delta}\right)}\right)^{2/3}$ is upper-bounded with probability at least $1 - \delta$ by*

$$38.94 \cdot l^{1/3} T^{2/3} D^{2/3} S^{2/3} \left(A \log\left(\frac{T}{\delta}\right)\right)^{1/3}.$$

Contribution: Improves upon the regret bound for UCRL2 with restarts (Jaksch et al.(2010)) in terms of D, S and A.

Further Directions

- Variation-dependent regret bound.
- Link between allowable variation in rewards and state-transition probabilities and minimal achievable regret.
- Refine the episode-stopping criterion.

Key reference

[Jaksch, Thomas and Ortner, Ronald and Auer, Peter(2010)] Near-optimal Regret Bounds for Reinforcement Learning

Proposed algorithm: SW-UCRL

- **Key idea:** Modify UCRL2 to use only the last W samples for computing the estimates.

1: **Input:** A confidence parameter $\delta \in (0, 1)$, $\mathcal{S}$, $\mathcal{A}$ and window size W.
2: **Initialization:** Set $t := 1$, and observe the initial state s_1.
For episodes $k = 1, 2, \ldots$ **do**
3: **Initialize episode** k:

i Set the start time of episode k, $t_k := t$.

ii For all (s, a) in $\mathcal{S} \times \mathcal{A}$ initialize the state-action counts for episode k, $v_k(s, a) := 0$. Further, set the the number of times any action action a was executed in state s in W time steps prior to episode k for all the states $s \in \mathcal{S}$ and actions $a \in \mathcal{A}$,

$$N_k(s, a) := \#\{t_k - W \leqslant \tau < t_k : s_\tau = s, a_\tau = a\}$$

iii For all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, set the observed cumulative rewards when action a was executed in state s and the number of times that resulted into the next state being s' during W time steps prior to episode k,

$$R_k(s, a) := \sum_{\tau = t_k - W}^{t_k - 1} r_\tau \mathbb{1}\{s_\tau = s, a_\tau = a\} \qquad P_k(s, a, s') := \#\{t_k - W \leqslant \tau < t_k : s_\tau = s, a_\tau = a, s_{\tau+1} = s'\}$$

Compute estimates $\hat{r}_k(s, a) := \frac{R_k(s,a)}{\max\{1, N_k(s,a)\}}$, $\hat{p}_k(s'|s, a) := \frac{P_k(s,a,s')}{\max\{1, N_k(s,a)\}}$

Compute policy $\tilde{\pi}_k$:

4. Let $\mathcal{M}_k$ be the set of all MDPs with state space $\mathcal{S}$ and action space $\mathcal{A}$, and with transition probabilities $\tilde{p}(\cdot|s, a)$ close to $\hat{p}_k(\cdot|s, a)$, and rewards $\tilde{r}(s, a) \in [0, 1]$ close to $\hat{r}_k(s, a)$, that is,

$$|\tilde{r}(s, a) - \hat{r}_k(s, a)| \leqslant \sqrt{\frac{7 \log(2SAt_k/\delta)}{\cdots}} \quad \text{and} \qquad (1)$$