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

Pratik Gajane

Dec 28, 2018

Presentation at IIT Madras, Department of Computer Science

Formalization

Classical Markov Decision Process (MDP)

- MDP : standard model for problems in decision making with uncertainty like RL.

Classical Markov Decision Process (MDP)

- MDP : standard model for problems in decision making with uncertainty like RL.
- Classical MDP $M(\mathcal{S}, \mathcal{A}, p, F)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition probability p, reward function F.

Classical Markov Decision Process (MDP)

- MDP : standard model for problems in decision making with uncertainty like RL.
- 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 $p(s' \mid s, a)$.

Classical Markov Decision Process (MDP)

- MDP : standard model for problems in decision making with uncertainty like RL.
- 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 $p(s' \mid s, a)$.
- In classical MDPs, stochastic state-transition dynamics and reward functions remain fixed (Bartlett and Tewari [2009], Burnetas and Katehakis [1997], Jaksch et al. [2010]).

Switching-MDP

➜ Our setting (**Switching-MDP**): transition dynamics and reward functions change a certain number of times (abrupt changes)

- **Switching-MDP M** $:= (\mathbb{S} = (M_0, \ldots, M_l), c = (c_1, \ldots, c_l))$

Switching-MDP

➜ Our setting (**Switching-MDP**): transition dynamics and reward functions change a certain number of times (abrupt changes)

- **Switching-MDP M** $:= (\mathbb{S} = (M_0, \ldots, M_l), c = (c_1, \ldots, c_l))$
- At $t < c_1$, **M** is in its initial configuration $M_0(\mathcal{S}, \mathcal{A}, p_0, F_0)$ i.e. M_0 is *active*.

Switching-MDP

➜ Our setting (**Switching-MDP**): transition dynamics and reward functions change a certain number of times (abrupt changes)

- **Switching-MDP M** $:= (\mathbb{S} = (M_0, \ldots, M_l), c = (c_1, \ldots, c_l))$
- At $t < c_1$, **M** is in its initial configuration $M_0(\mathcal{S}, \mathcal{A}, p_0, F_0)$ i.e. M_0 is *active*.
- At time step $c_i \leq t < c_{i+1}$, **M** is in configuration $M_i(\mathcal{S}, \mathcal{A}, p_i, F_i)$ i.e. M_i is *active*.

Switching-MDP

➜ Our setting (**Switching-MDP**): transition dynamics and reward functions change a certain number of times (abrupt changes)

- **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 $M_0(\mathcal{S}, \mathcal{A}, p_0, F_0)$ i.e. M_0 is *active*.
- At time step $c_i \leq t < c_{i+1}$, $\mathbf{M}$ is in configuration $M_i(\mathcal{S}, \mathcal{A}, p_i, F_i)$ i.e. M_i is *active*.

➜ 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 the active MDP.

Related work

- MDPs in which the state-transition probabilities change arbitrarily but the reward functions remain fixed (Nilim and El Ghaoui [2005], Xu and Mannor [2006]).

Related work

- MDPs in which the state-transition probabilities change arbitrarily but the reward functions remain fixed (Nilim and El Ghaoui [2005], Xu and Mannor [2006]).
- MDPs with fixed state-transition probabilities and changing reward functions Even-dar et al. [2005]

Related work

- MDPs in which the state-transition probabilities change arbitrarily but the reward functions remain fixed (Nilim and El Ghaoui [2005], Xu and Mannor [2006]).
- MDPs with fixed state-transition probabilities and changing reward functions Even-dar et al. [2005]
- Yuan Yu and Mannor [2009a] and Yuan Yu and Mannor [2009b] consider arbitrary changes in the reward functions and arbitrary, but bounded, changes in the state-transition probabilities.

Related work

- MDPs in which the state-transition probabilities change arbitrarily but the reward functions remain fixed (Nilim and El Ghaoui [2005], Xu and Mannor [2006]).
- MDPs with fixed state-transition probabilities and changing reward functions Even-dar et al. [2005]
- Yuan Yu and Mannor [2009a] and Yuan Yu and Mannor [2009b] consider arbitrary changes in the reward functions and arbitrary, but bounded, changes in the state-transition probabilities.
- Abbasi et al. [2013] consider MDP problems with (oblivious) adversarial changes in state-transition probabilities and reward functions.

Proposed algorithm: SW-UCRL

Proposed algorithm: SW-UCRL

- **Key idea:** Modify UCRL2 to use only the last W samples for computing the estimates.
- **Input:** A confidence parameter $\delta \in (0, 1)$ and window size W.
- **Initialization:** Set $t := 1$, and observe the initial state s_1.

SW-UCRL: Episode Initialization

1. Set the start time of episode k, $t_k := t$.
2. For all (s, a) in $\mathcal{S} \times \mathcal{A}$, set $v_k(s, a) := 0$

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

SW-UCRL: Episode Initialization

1. Set the start time of episode k, $t_k := t$.
2. For all (s, a) in $\mathcal{S} \times \mathcal{A}$, set $v_k(s, a) := 0$

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

3. For all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$R_k(s, a) := \sum_{\tau = t_k - W}^{t_k - 1} r_\tau \mathbb{1}\{s_\tau = s, a_\tau = a\}$$

$$P_k(s, a, s') := \#\{t_k - W \leq \tau < t_k : s_\tau = s, a_\tau = a, s_{\tau+1} = s'\}$$

SW-UCRL: Episode Initialization

1. Set the start time of episode k, $t_k := t$.
2. For all (s, a) in $\mathcal{S} \times \mathcal{A}$, set $v_k(s, a) := 0$

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

3. For all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$R_k(s, a) := \sum_{\tau = t_k - W}^{t_k - 1} r_\tau \mathbb{1}\{s_\tau = s, a_\tau = a\}$$

$$P_k(s, a, s') := \#\left\{t_k - W \leq \tau < t_k : s_\tau = s, a_\tau = a, s_{\tau+1} = s'\right\}$$

4. Compute estimates

$$\hat{r}_k(s, a) := \frac{R_k(s, a)}{\max\{1, N_k(s, a)\}}$$

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

SW-UCRL: Policy Computation

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

$$\left|\tilde{r}(s,a) - \hat{r}_k\left(s,a\right)\right| \leq \sqrt{\frac{7\log(2SAt_k/\delta)}{2\max\{1,N_k(s,a)\}}} \quad \text{and} \tag{1}$$

$$\left\|\tilde{p}\left(\cdot|s,a\right) - \hat{p}_k\left(\cdot|s,a\right)\right\|_1 \leq \sqrt{\frac{14S\log(2At_k/\delta)}{\max\{1,N_k(s,a)\}}}\,. \tag{2}$$

2. Use extended value iteration to find a near optimal policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$

SW-UCRL: Policy Execution

Episode stopping criterion: number of occurrences of any (s, a) in the episode $(v_k(s, a)) =$ number of occurrences of same (s, a) in W observations before episode start$(N_k(s, a))$

While $v_k(s_t, \tilde{\pi}_k(s_t)) < \max\{1, N_k(s_t, \tilde{\pi}_k(s_t))\}$ **do**

- Choose action $a_t = \tilde{\pi}_k(s_t)$, obtain reward r_t.
- Observe next state s_{t+1}.
- Update $v_k(s_t, a_t) := v_k(s_t, a_t) + 1$.
- Set $t := t + 1$.

Performance Bounds

Theorem (Regret Upper Bound)

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*

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

where D = max of diameters of constituent MDPs.

- Optimal value of W:

$$W^* = \left(\frac{16.53}{l} TDS\sqrt{A \log\left(\frac{T}{\delta}\right)}\right)^{2/3}$$

Performance Bounds

Corollary (Regert Upper Bound using W^*)

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*

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

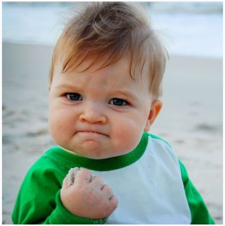
Performance Bounds

Corollary (Regert Upper Bound using W^*)

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) Jaksch et al. [2010]) in terms of D, S and A.



Performance Bounds

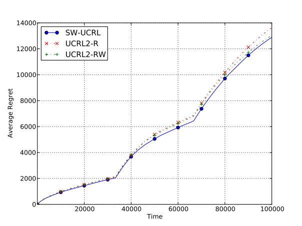
Corollary (Sample Complexity Bound)

Given a switching-MDP problem with l changes, the average per-step regret of SW-UCRL *using W^* is at most ϵ with probability at least $1-\delta$ after any T steps with*

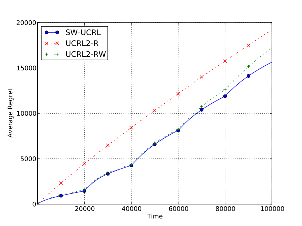
$$T \geq 2 \cdot (38.94)^3 \cdot \frac{lD^2S^2A}{\epsilon^3} \log\left(\frac{(38.94)^3 lD^2S^2A}{\epsilon^3\delta}\right).$$

Experiments

Experiments



(a) Average regret plot for 2 changes



(b) Average regret plot for 4 changes

Figure 1: Average regret plots for switching-MDPs

- Switching-MDPs with $S = 5, A = 3$, and $T = 100000$.
- l changes happen at every $\lceil \frac{T}{l} \rceil$ time steps.
- SW-UCRL with optimum window size W^*
- For comparison : UCRL2 with restarts (UCRL2-R) and UCRL2 with restarts after every W^* time steps (UCRL2-RW)

Summary and Future Directions

Summary and Future Directions

- SW-UCRL: a competent solution for regret-minimization on switching-MDPS.
- Variation-dependent regret bound?
- Link between allowable variation in rewards and transition probabilities and minimal achievable regret? (like Besbes et al. [2014] for bandits)
- Refine episode-stopping criterion?

Thank you all.

References

Yasin Abbasi, Peter L Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvari. Online learning in Markov decision processes with adversarially chosen transition probability distributions. In *Advances in Neural Information Processing Systems 26*, pages 2508–2516. Curran Associates, Inc., 2013.

Peter L. Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 35–42, Arlington, Virginia, United States, 2009. AUAI Press. ISBN 978-0-9749039-5-8.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 199–207. Curran Associates, Inc., 2014.

Apostolos N. Burnetas and Michael N. Katehakis. Optimal adaptive policies for markov decision processes. *Math. Oper. Res.*, 22(1): 222–255, February 1997. ISSN 0364-765X.

Eyal Even-dar, Sham M Kakade, and Yishay Mansour. Experts in a Markov decision process. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 401–408. MIT Press, 2005.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, 11: 1563–1600, August 2010.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Oper. Res.*, 53(5): 780–798, September 2005.

Huan Xu and Shie Mannor. The robustness-performance tradeoff in Markov decision processes. In *NIPS*, pages 1537–1544. MIT Press, 2006.

Jia Yuan Yu and Shie Mannor. Arbitrarily modulated Markov decision processes. In *Proceedings of the IEEE Conference on Decision and Control*, pages 2946–2953, 12 2009a.

Jia Yuan Yu and Shie Mannor. Online learning in Markov decision processes with arbitrarily changing rewards and transitions. In *2009 International Conference on Game Theory for Networks*, pages 314–322, May 2009b.

SW-UCRL: Policy computation

1. 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,

$$\begin{aligned} \left|\tilde{r}(s,a) - \hat{r}_k(s,a)\right| &\leq \sqrt{\frac{7\log(2SAt_k/\delta)}{2\max\{1, N_k(s,a)\}}} \quad \text{and} \qquad (3) \\ \left\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\right\|_1 &\leq \sqrt{\frac{14S\log(2At_k/\delta)}{\max\{1, N_k(s,a)\}}}\,. \qquad (4) \end{aligned}$$

2. Use extended value iteration to find a policy near optimal policy $\tilde{\pi}_k$ and an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ such that

$$\tilde{\rho}_k := \min_s \rho(\tilde{M}_k, \tilde{\pi}_k, s) \geq \max_{M' \in \mathcal{M}_k, \pi, s'} \rho(M', \pi, s') - \frac{1}{\sqrt{t_k}}.$$