Variational Regret Bounds for Reinforcement Learning

Ronald Ortner, Pratik Gajane, Peter Auer

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

Montanuniversität Leoben

FWF Der Wissenschaftsfonds.

Introduction

- In a standard RL problem, the state-transition dynamics and the reward functions are time-invariant.

➜ **Our setting**: Both the transition dynamics and the reward functions are dependent on the current time step.

Problem setting

- For $t = 1, \ldots, T$, the learner chooses an **action a_t** in the current **state s_t**,
  - receives a reward r_t with **mean $\bar{r}_t(s_t, a_t)$**,
  - and observes a transition to the next state s_{t+1} according to $p_t(s_{t+1}|s_t, a_t)$.
- For $t = 1, \ldots, T$, let **MDP M_t** $= (\mathcal{S}, \mathcal{A}, \bar{r}_t, p_t)$ denote the true MDP at time t. Further, let $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$.
- Assumption: For each $M_{t\{1 \leqslant t \leqslant T\}}$, the **diameter** (minimal expected time it takes to get from any state to any other state [1]) is upper bounded by D.

➜ **Variation:** For time horizon T,

$$\boldsymbol{V_T^r} := \sum_{t=1}^{T-1} \max_{s,a} |\bar{r}_{t+1}(s,a) - \bar{r}_t(s,a)|$$

$$\boldsymbol{V_T^p} := \sum_{t=1}^{T-1} \max_{s,a} \|p_{t+1}(\cdot|s,a) - p_t(\cdot|s,a)\|_1.$$

- Goal: Minimize regret

$$R_T := v_T^*(s_1) - \sum_{t=1}^{T} r_t$$

where $v_T^*(s_1)$ is the optimal expected T-step reward achievable by any policy starting in the initial state s_1.

Main result : Regret Bound

With probability $1-\delta$, the regret of variation-aware UCRL with restarts (Algorithm 2) after any T steps is bounded as

$$R_T \leqslant 74 \cdot DS(V_T^r + V_T^p)^{1/3} T^{2/3} \sqrt{A \log\left(\tfrac{16S^2AT^5}{\delta}\right)}.$$

➜ **Optimal wrt time and variation parameters.**
For (the simpler) bandit setting, a lower bound on the variational regret given by Besbes et al. (2014) [3] shows that our bound is optimal with respect to time and the variation.

Algorithm 1: Variation-aware UCRL

1: **Input:** $\mathcal{S}$, $\mathcal{A}$, δ, variation parameters $\tilde{V}^r$, $\tilde{V}^p$.

2: **Initialization:** Set current time step $t := 1$.

3: **for** episode $k = 1, \ldots$ **do**

4: Set episode start $t_k := t$. Let $v_k(s,a)$ = state-action counts for visits in k, and $N_k(s,a)$ = counts for visits before episode k.

5: For $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$, compute estimates

$$\hat{r}_k(s,a) := \frac{\sum_{\tau=1}^{t_k-1} r_\tau \cdot \mathbb{1}_{s_\tau=s, a_\tau=a}}{\max(1, N_k(s,a))},$$

$$\hat{p}_k(s'|s,a) := \frac{\#\{\tau < t_k : s_\tau = s, a_\tau = a, s_{\tau+1} = s'\}}{\max(1, N_k(s,a))}.$$

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

6: Let $\mathcal{M}_k$ be the set of plausible MDPs $\tilde{M}$ with rewards $\tilde{r}(s,a)$ and transition probabilities $\tilde{p}(\cdot|s,a)$ satisfying

$$\tilde{r}(s,a) - \hat{r}_k(s,a)| \leqslant \tilde{V}^r + \sqrt{\frac{8\log(8SAt_k^3/\delta)}{\max(1, N_k(s,a))}}, \quad (1)$$

$$\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \leqslant \tilde{V}^p + \sqrt{\frac{8S\log(8SAt_k^3/\delta)}{\max(1, N_k(s,a))}}. \quad (2)$$

7: Use extended value iteration [1] to find an optimal policy $\tilde{\pi}_k$ for an optimistic MDP $\tilde{M}_k \in \mathcal{M}_k$ such that

$$\rho(\tilde{M}_k, \tilde{\pi}_k) = \max_{M' \in \mathcal{M}_k} \rho^*(M'),$$

where $\rho^*(M')$ is the optimal average reward of M'.

Execute policy $\tilde{\pi}_k$**:**

8: **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, and observe s_{t+1}.

Set $t = t + 1$.

end while

9: **end for**

Algorithm 2 : Variation-aware UCRL with restarts

1: **Input:** $\mathcal{S}$, $\mathcal{A}$, δ, variation V_T^r and V_T^p.

2: **Initialization:** Set current time step $\tau := 1$.

3: **for** phase $i = 1, \ldots$ **do**

4: Perform variation-aware UCRL with confidence parameter $\delta/2\tau^2$ for $\theta_i := \left\lceil \frac{i^2}{(V_T^r + V_T^p)^2} \right\rceil$ steps.

5: Set $\tau = \tau + \theta_i$.

6: **end for**

Solution sketch

- Devise variation-aware UCRL (Algorithm 1) by adapting confidence intervals (eq. (1) and eq. (2)) to account for the variation in mean rewards and transition probabilities respectively.
- Restart variation-aware UCRL according to a suitable scheme (cf. line 4 in Algorithm 2). For this, the algorithm needs to know the variation V_T^r and V_T^p.

Analysis sketch

- **Optimism:** With high probability, the set of plausible MDPs (line 6 in Algorithm 1) computed at any time t contains the true MDP M_t.
- **Perturbation bound:** For any two MDPs M and M' whose mean rewards differ by at most Δ^r and whose L1-norm of the transition probabilities is at most Δ^p, it holds that

$$|\rho^*(M) - \rho^*(M')| \leqslant \Delta^r + D\Delta^p,$$

where D = maximum of the diameters of M and M'.

- **Regret of variation-aware UCRL:**
With probability $1-\delta$, the regret of variation-aware UCRL is bounded by $32DS\sqrt{AT\log\left(\frac{8SAT^3}{\delta}\right)} + 2T(DV_T^p + V_T^r)$.
- **Regret of variation-aware UCRL with restarts:** We sum up the regret over all the phases of variation-aware UCRL to arrive at the main result.

Conclusion and Further Directions

- Performance guarantees that are optimal in time and variation demonstrate that our algorithm is a competent solution for the considered problem setting.
- Recently, variational bounds for the (contextual) bandit setting have been derived when the variation is unknown [2]. Achieving such bounds in RL is a worthwhile direction to pursue.

Key references

[1] Thomas Jaksch, Ronald Ortner and Peter Auer: Near-optimal Regret Bounds for Reinforcement Learning, JMLR 2010.

[2] Peter Auer, Yifang Chen, Pratik Gajane, Chung-Wei Lee, Haipeng Luo, Ronald Ortner, Chen-Yu Wei: Achieving Optimal Dynamic Regret for Non-stationary Bandits without Prior Information, COLT 2019.

[3] Omar Besbes, Yonatan Gur, and Assaf Zeevi: Stochastic multi-armed-bandit problem with non-stationary rewards. NIPS 2014.