**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

$$R_T \leq 74 \cdot DS(V_T^p + V_T^\beta)^{1/2}/T^{2/3} \sqrt{A \log \left( \frac{18S^2A^2T}{\delta} \right)}.$$  

**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.

---

**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 $M_t = (S, A, \bar{r}_t, p_t)$ denote the true MDP at time $t$. Further, let $S := |S|$ and $A := |A|$.

- Assumption: For each $M_t(1 \leq t \leq T)$, 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 $T$,

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

$$V_T^p := \sum_{t=1}^{T-1} \max_{a,s} |p_{t+1}(s'|s,a) - p_t(s'|s,a)| 1.$$  

**Goal:** Minimize regret

$$R_T := \mathcal{V}^\#_T(s_1) - \sum_{t=1}^{T-1} r_t$$

where $\mathcal{V}^\#_T(s_1)$ is the optimal expected $T$-step reward achievable by any policy starting in the initial state $s_1$.

---

**Algorithm 1: Variation-aware UCRL**

1. **Input:** $S, A, \delta, \text{variation parameters } \bar{V}^r, \bar{V}^p$.
2. **Initialization:** Set current time step $t := 1$.
3. **for episode $k = 1, \ldots, do**
   4. **Set episode start $t_k := t$**. Let $V_k(s,a) = \text{state-action counts for visits in } k$, and $N_k(s,a) = \text{counts for visits before episode } k$.
   5. **for $s, s_t \in S$ and $a \in A$, compute estimates**

$$\hat{r}_k(s,a) := \frac{\sum_{t=1}^{t_k-1} r_t - 1_{s_t = s, a_t = a, s_{t+1} = s'}}{\max(1, N_k(s,a))},$$

$$\hat{p}_k(s'|s,a) := \frac{\#\{t < t_k : s_t = s, a_t = a, s_{t+1} = s'\}}{\max(1, N_k(s,a))}.$$  

**Compute policy $\pi_k$:**

6. **Let $M_k$ be the set of plausible MDPs $\bar{M}$ with rewards $\bar{r}(s,a)$ and transition probabilities $\bar{p}(s'|s,a)$ satisfying**

$$|\bar{r}(s,a) - \hat{r}_k(s,a)| \leq \bar{V}^r + \sqrt{\frac{8 \log(8SA^2T/\delta)}{\max(1, N_k(s,a))}}.$$  

$$|\bar{p}(s'|s,a) - \hat{p}_k(s'|s,a)| \leq \bar{V}^p + \sqrt{\frac{8 \log(8SA^2T/\delta)}{\max(1, N_k(s,a))}}.$$  

7. **Use extended value iteration [1] to find an optimal policy $\pi_k$ for an optimistic MDP $\bar{M}_k \in M_k$ such that**

$$\rho(\bar{M}_k, \pi_k) = \max_{M \in M_k} \rho^*(M')$$

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

**Execute policy $\pi_k$:**

8. **while $V_k(s, \pi_k(s)) < \max(1, N_k(s, \pi_k(s)))$, do**

   - Choose action $a_t = \pi_k(s_t)$.
   - Obtain reward $r_t$, and observe $s_{t+1}$.
   - Set $t = t + 1$.

9. **end while**

---

**Algorithm 2 : Variation-aware UCRL with restarts**

1. **Input:** $S, A, \delta, \text{variation parameters } V_T^p$ and $V_T^\beta$.
2. **Initialization:** Set current time step $t := 1$.
3. **for phase $i = 1, \ldots, do**
   4. **Perform variation-aware UCRL with confidence parameter $\delta/2r^2$ for $\pi_i := \left\lceil \frac{1}{V_T^p + V_T^\beta} \right\rceil$ steps.**
   5. **Set $\tau = \tau + 1$.
   6. **end for**

---

**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**

