Reinforcement Learning

Lecture 6: RL algorithms 2.0

Alexandre Proutiere, Sadegh Talebi, Jungseul Ok

KTH, The Royal Institute of Technology

Objectives of this lecture

Present and analyse two online algorithms based on the "optimism in front of uncertainty" principle, and compare their regret to algorithms with random exploration

- UCB-VI for episodic RL problems
- UCRL2 for ergodic RL problems

Lecture 6: Outline

1. Minimal exploration in RL
2. UCB-VI
3. UCRL2

Lecture 6: Outline

1. **Minimal exploration in RL**
2. UCB-VI
3. UCRL2

Towards minimal exploration

The MDP model is unknown and has to be learnt. Solutions for on-policy algorithms:

1. Estimate the model then optimise: poor regret and premature exploitation
2. ϵ greedy exploration: undirected exploration (explores too much (state, action) pairs with low values)
3. **Bandit-like optimal exploration-exploitation trade-off**

But how much should a (state,action) pair be explored?

Regret lower bounds

In the case of ergodic RL problems:

- Problem-specific lower bound (Burnetas - Katehakis 1997)

$$\lim \inf_{T \to \infty} \frac{\mathbb{E}[N_{(s,a)}(T)]}{\log(T)} \geq \frac{1}{\mathcal{K}_M(s,a)}$$

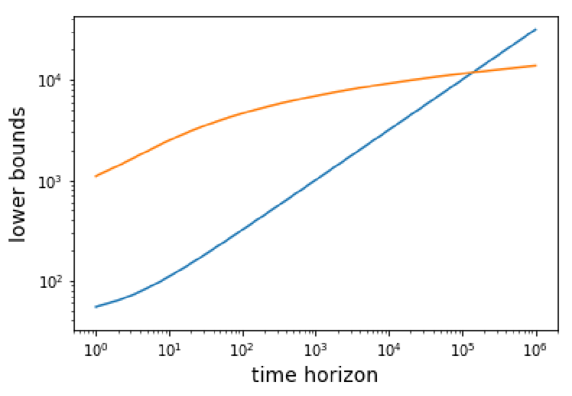
  Leading to an **asymptotic** regret lower bound scaling as $SA\log(T)$
- Minimax lower bound $\Theta(\sqrt{SAT})$

We don't know when the asymptotic problem-specific regret lower bound is representative, often for very large T!

Read for bandit optimisation: "Explore First, Exploit Next: The True Shape of Regret in Bandit Problems", Garivier et al. , https://arxiv.org/abs/1602.07182

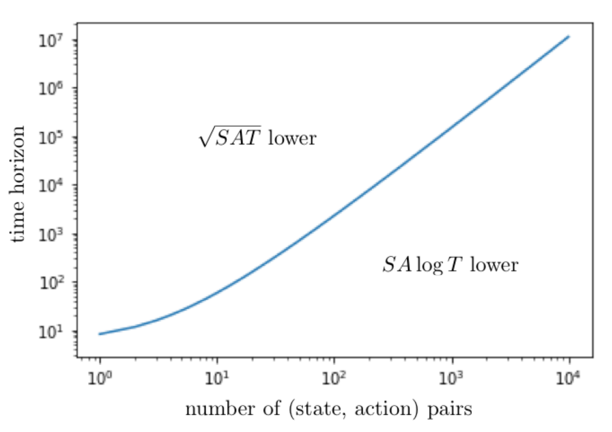
Which regret lower bound should we target?

Example: $SA = 1000$, comparison of $\sqrt{SAT}$ and $SA\log(T)$



Which regret lower bound should we target?

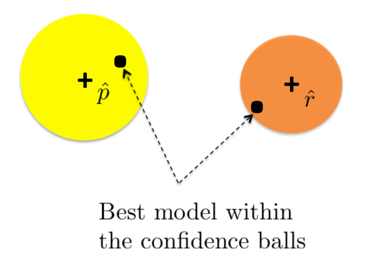
Boundary: $SA = \frac{T}{\log(T)^2}$



Optimism in front of uncertainty

Estimate the unknown system parameters (here $p(\cdot|\cdot,\cdot)$ and $r(\cdot,\cdot)$) and build an optimistic reward estimate to trigger exploration.

Estimate: find confidence balls containing the true model w.h.p.
Optimistic reward estimate: find the model within the confidence balls leading to the highest value.



Optimism in front of uncertainty: generic algorithm

Algorithm. (for Infinite horizon RL problems)
Initialise $\hat{p}$, $\hat{r}$, and $N(s,a)$ For $t = 1, 2, \ldots$

1. Build an optimistic reward model $(\bar{Q}(s,a))_{s,a}$ from $\hat{p}$, $\hat{r}$, and $N(s,a)$
2. Select action $a(t)$ maximising $\bar{Q}(s(t),a)$ over $\mathcal{A}_{s(t)}$
3. Observe the transition to $s(t+1)$ and collect reward $r(s(t),a(t))$
4. Update $\hat{p}$, $\hat{r}$, and $N(s,a)$

Examples

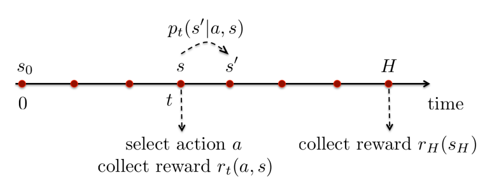
UCB-VI: directly build a confidence ball for the Q function based on the empirical estimates of the model.

UCRL2: first build confidence balls for the reward and transition probabilities, and then identify $\bar{Q}$.

Lecture 6: Outline

1. Minimal exploration in RL
2. **UCB-VI**
3. UCRL2

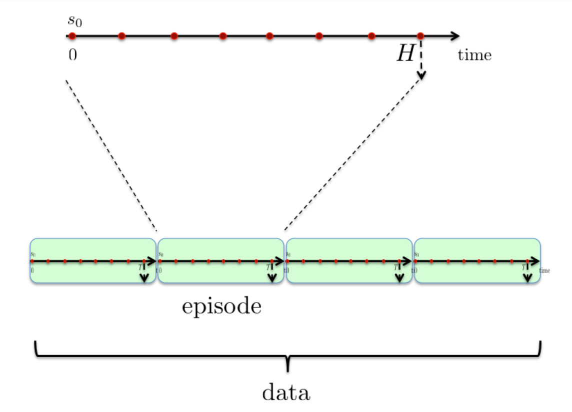
Finite-horizon MDP to episodic RL problems



- Initial state s_0 (could be a r.v.)
- Transition probabilities at time t: $p(s'|s, a)$
- Reward at time t: $r(s, a)$ and at time H: $r_H(s)$
- **Unknown** transition probabilities and reward function
- Objective: *quickly* learn a policy $\pi^\star$ maximising over $\pi_0 \in MD$

$$V_H^{\pi_0} := \mathbb{E}\left[\sum_{u=0}^{H-1} r(s_u^{\pi_0}, a_u^{\pi_0}) + r_H(s_H^{\pi_0})\right].$$

Finite-horizon MDP to episodic RL problems



- Data: K episodes of length H (actions, states, rewards)
- Learning algorithm π : data $\mapsto \pi_K \in MD$
- Performance of π: how close π_K is from the optimal policy $\pi^\star$

UCB-VI

UCBVI is an extension of Value Iteration, guaranteeing that the resulting value function is a (high-probability) upper confidence bound (UCB) on the optimal value function.

At the beginning of episode k, it computes state-action values using empirical transition kernel and reward function. In step h of backward induction (to update $Q_{k,h}(s,a)$ for any (s,a)), it adds a bonus $b_{k,h}(s,a)$ to the value, and ensures that $Q_{k,h}$ never exceeds $Q_{k,h-1}$.

Two variants of UCBVI, depending on the choice of bonus $b_{k,h}$:

- UCBVI-CH
- UCBVI-FB

UCB-VI algorithm

Variables to be maintained by the algorithm: for known reward function

- $\hat{p} = (\hat{p}(s'|s,a), s, s' \in \mathcal{S}, a \in \mathcal{A}_s)$: estimated transition probabilities
- $Q = (Q_h(s,a), h \leq H, s \in \mathcal{S}, a \in \mathcal{A}_s)$: estimated Q-function
- $b = (b_h(s,a), h \leq H, s \in \mathcal{S}, a \in \mathcal{A}_s)$: Q-value bonus
- $N = (N(s,a), s \in \mathcal{S}, a \in \mathcal{A}_s)$: number of visits to (s,a) so far
- $N' = (N_h(s,a), h \leq H, s \in \mathcal{S}, a \in \mathcal{A}_s)$: number of visits in the h-step of episodes to (s,a) so far

UCB-VI algorithm

Algorithm. UCB-VI

Input: Initial state distribution ν_0, precision δ

Initialise the variables $\hat{p}$, N, and N'

For episode $k = 1, 2, \ldots$

1. Optimistic reward:
   a. Compute the bonus: $b \leftarrow \text{bonus}(N, N', \hat{p}, Q, \delta)$
   b. Estimate the Q-function: $Q \leftarrow \text{bellmanOpt}(Q, b, \hat{p})$
2. Initialise the state $s(0) \sim \nu_0$
3. for $h = 1, \ldots, H$, select action
   $a \in \arg\max_{a' \in \mathcal{A}_{s(h-1)}} Q_h(s(h-1), a')$
4. Observe the transition and update $\hat{p}$, N, and N'

UCB-VI algorithm: bonus

UCBVI-CH:

$$b_h(s,a) = \frac{7H}{\sqrt{N(s,a)}} \log(5SAT/\delta)$$

UCBVI-BF:

$$\begin{aligned} b_h(s,a) = & \sqrt{\frac{8L}{N(s,a)} \mathrm{Var}_{\widehat{p}(\cdot|s,a)}(V_{h+1}(Y))} + \frac{14HL}{3N(s,a)} \\ & + \sqrt{\frac{8}{N(s,a)} \sum_y \widehat{p}(y|s,a) \min\left\{ \frac{10^4 H^3 S^2 A L^2}{N'_{h+1}(y)}, H^2 \right\}} \end{aligned}$$

where $L = \log(5SAT/\delta)$.

UCB-VI algorithm: Optimistic Bellman operator

bellmanOpt$(Q, b, \hat{p})$ applies Dynamic Programming with a bonus.

Initialisation: $Q_H(s,a) = r_H(s)$ for all (s,a)
For step $h = H-1, \dots, 1$**:** for all (s,a) visited at least once so far:
$Q_h(s,a) \leftarrow \min\left(Q_h(s,a), H, r(s,a) + \sum_y \hat{p}(y|s,a) V_{h+1}(s) + b_h(s,a)\right)$

UCB-VI: Regret guarantees

Regret up to time $T = KH$:
$R^{UCBVI}(T) = \sum_{k=1}^{K}(V^\star(x_{k,1}) - V^{\pi_k}(x_{k,1}))$.

Theorem *For any $\delta > 0$, the regret of UCB-VI-CH(δ) is bounded w.p. at least $1-\delta$ by:*

$$R^{UCBVI-CH}(T) \leq 20HL\sqrt{SAT} + 250H^2S^2AL^2,$$

with $L = \log(5HSAT/\delta)$.

For $T \geq HS^3A$ and $SA \geq H$, the regret upper bound scales as $\tilde{\mathcal{O}}(H\sqrt{SAT})$ (!?)

Sketch of proof

Notations:

- π_k is the policy applied by UCBVI in the k-th episode
- $V_{k,h}$ is the optimistic value function computed by UCBVI in the h-step of the k-th episode
- V_h^π is the value function from step h under π
- $P^\pi = (p(s'|s,\pi(s)))_{s,s'}$
- $\hat{P}_k^\pi = (\hat{p}_k(s'|s,\pi(s)))_{s,s'}$ where $\hat{p}_k$ is the estimated transitions in episode k

Claim 1: by construction with high probability, $V_{k,h} \geq V_h^\star$. Then:

$$R^{UCBVI}(T) \leq \tilde{R}(T) = \sum_{k=1}^{K} (V_{k,1}(x_{k,1}) - V^{\pi_k}(x_{k,1}))$$

Sketch of proof

Let $\tilde{\Delta}_{k,h} = V_{k,h} - V_h^{\pi_k}$, so that $\tilde{R}(T) = \sum_{k=1}^{K} \tilde{\Delta}_{k,1}(x_{k,1})$.
Backward induction on h to bound $\tilde{\Delta}_{k,1}$: introduce $\tilde{\delta}_{k,h} = \tilde{\Delta}_{k,h}(x_{k,h})$, then

$$\tilde{\delta}_{k,h} \leq (\hat{P}_k^{\pi_k} - P^{\pi_k})\tilde{\Delta}_{k,h+1}(x_{k,h}) + \tilde{\delta}_{k,h+1} + \epsilon_{k,h} + b_{k,h} + e_{k,h}$$

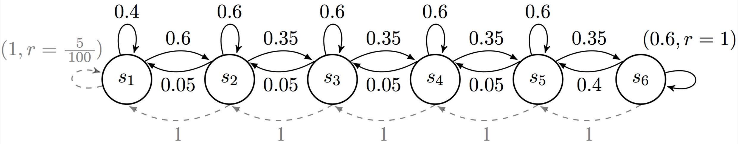
where

$$\begin{cases} \epsilon_{k,h} = P^{\pi_k}\tilde{\Delta}_{k,h+1}(x_{k,h}) - \tilde{\Delta}_{k,h+1}(x_{k,h+1}) \\ e_{k,h} = (\hat{P}_k^{\pi_k} - P^{\pi_k})V_{h+1}^{\star}(x_{k,h}) \end{cases}$$

Concentration + Martingale (Azuma) + bounding bonus

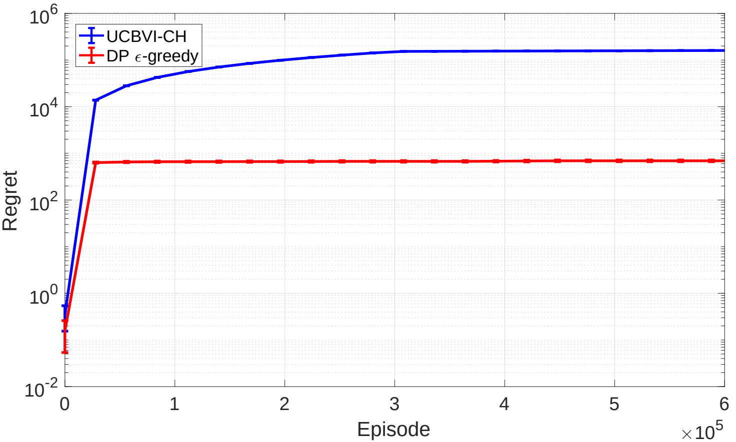
Numerical experiments

The river-swim example ...



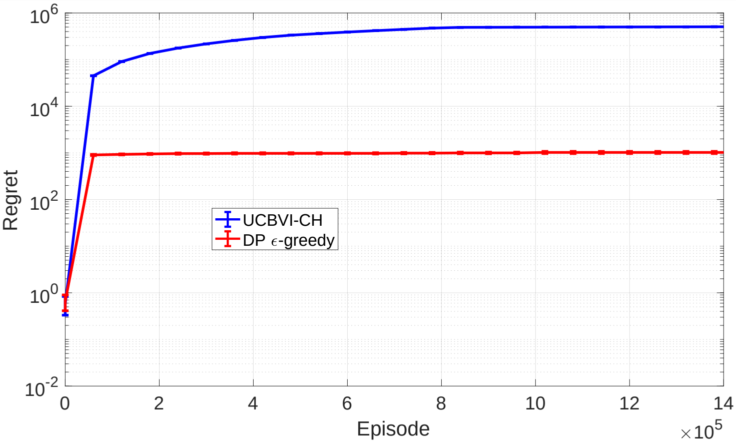
Regret

4 states, $H = 2$, $\delta = 0.05$ (for UCBVI), ϵ-greedy: $\epsilon_t = \min(1, 1000/t)$



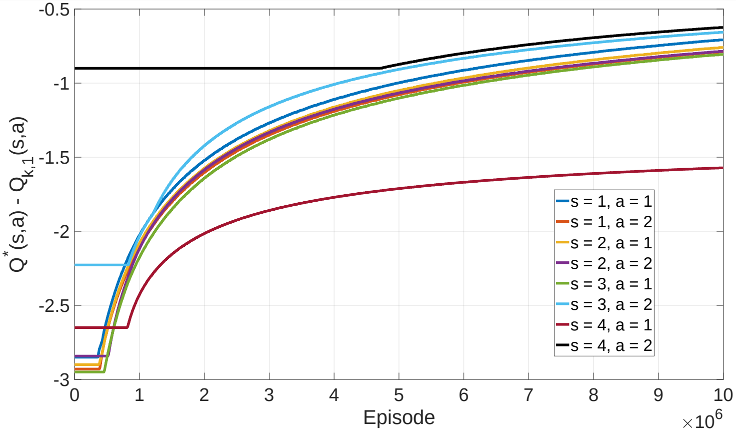
Regret

4 states, $H = 3$, $\delta = 0.05$ (for UCBVI), ϵ-greedy: $\epsilon_t = \min(1, 1000/t)$



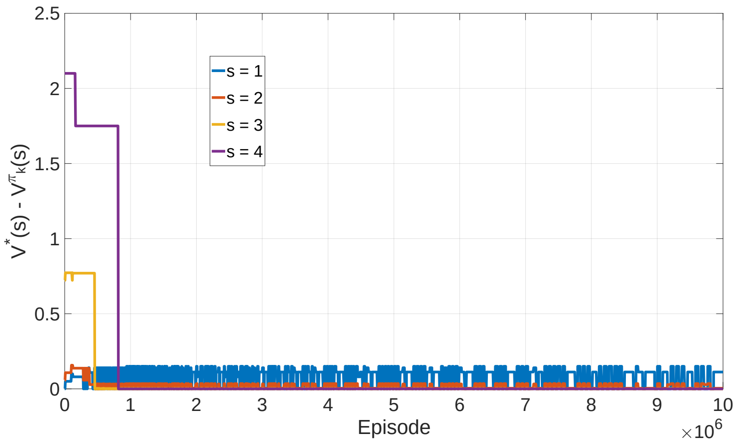
Optimistic Q-values

4 states, $H = 3$, $\delta = 0.05$ (for UCBVI)



Value function convergence under UCBVI

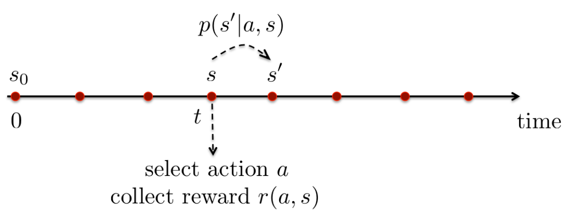
4 states, $H = 3$, $\delta = 0.05$ (for UCBVI)



Lecture 6: Outline

1. Minimal exploration in RL
2. UCB-VI
3. **UCRL2**

Expected average reward MDP to ergodic RL problems



- Stationary transition probabilities $p(s'|s,a)$ and rewards $r(s,a)$, uniformly bounded: $\forall a, s,\ |r(s,a)| \leq 1$
- Objective: learn from data a policy $\pi \in MD$ maximising (over all possible policies)

$$g^\pi = V^\pi(s_0) := \lim \inf_{T\to\infty} \frac{1}{T}\mathbb{E}_{s_0}\left[\sum_{u=0}^{T-1} r(s_u^\pi, a_u^\pi,)\right]$$

Ergodic RL problems: Preliminaries

Optimal policy

Recall Bellman's equation

$$g^\star + h^\star(s) = \max_{a\in\mathcal{A}}\Big(r(s,a) + h^\top p(\cdot|s,a)\Big), \quad \forall s$$

where $g^\star$ is the maximal gain, and $h^\star$ is the *bias* function ($h^\star$ is uniquely determined up to an additive constant). Note: $g^\star$ does not depend on the initial state for communicating MDPs.

Let $a^\star(s)$ denote any optimal action for state s (i.e., a maximizer in the above). Define the gap for sub-optimal action a at state s:

$$\phi(s,a) := \big(r(s,a^\star(s)) - r(s,a)\big) + {h^\star}^\top\big(p(\cdot|s,a^\star(s)) - p(\cdot|s,a)\big)$$

Ergodic RL problems: Preliminaries

Diameter D: defined as

$$D := \max_{s \neq s'} \min_{\pi} \mathbb{E}[T^{\pi}_{s,s'}]$$

where $T^{\pi}_{s,s'}$ denotes the first time step in which s' is reached under π staring from initial state s.

Remark: all communicating MDPs have a finite diameter.

Important parameters impacting performance

- Diameter D
- Gap $\Phi := \min_{s,a \neq a^\star(s)} \phi(s,a)$
- Gap $\Delta := \min_{\pi}(g^\star - g^{\pi})$

Ergodic RL problems: Regret lower bounds

- **Problem-specific regret lower bound:** (Burnetas-Katehakis)
  For any algorithm π,

$$\liminf_{T\to\infty} \frac{R^\pi(T)}{\log(T)} \geq c_{\text{bk}} := \sum_{s,a} \frac{\phi(s,a)}{\inf\{\mathrm{KL}(p(\cdot|s,a), q) : q \in \Theta_{s,a}\}}$$

  where $\Theta_{s,a}$ is the set of distributions q s.t. replacing (only) $p(\cdot|s,a)$ by q makes a the unique optimal action in state s.
  - asymptotic (valid as $T \to \infty$)
  - valid for any ergodic MDP
  - scales as $\Omega(\frac{DSA}{\Phi}\log(T))$ for specific MDPs

- **Minimax regret lower bound:** $\Omega(\sqrt{DSAT})$
  - non-asymptotic (valid for all $T \geq DSA$)
  - derived for a specific family of *hard-to-learn* communicating MDPs

Ergodic RL problems: State-of-the-art

Two types of algorithms targeting different regret guarantees:

- Problem-specific guarantees
  - MDP-specific regret bound scaling as $\mathcal{O}(\log(T))$
  - Algorithms: B-K (Burnetas & Katehakis, 1997), OLP (Tewari & Bartlett, 2007), UCRL2 (Jaksch et al. 2009), KL-UCRL (Filippi et al. 2010)
- Minimax guarantees
  - Valid for a class of MDPs with S states and A actions, and (typically) diameter D
  - Scaling as $\widetilde{\Omega}(\sqrt{T})$
  - Algorithms: UCRL2 (Jaksch et al. 2009), KL-UCRL (Filippi et al. 2010), REGAL (Bartlett & Tewari, 2009), A-J (Agrawal & Jia, 2010)

Ergodic RL problems: State-of-the-art

Algorithm	Setup	Regret
B-K	ergodic MDPs, known rewards	$\mathcal{O}\left(c_{\text{bk}}\log(T)\right)$ – asympt.
OLP	ergodic MDPs, known rewards	$\mathcal{O}\left(\frac{D^2SA}{\Phi}\log(T)\right)$ – asympt.
UCRL	unichain MDPs	$\mathcal{O}\left(\frac{S^5A}{\Delta^2}\log(T)\right)$
UCRL2, KL-UCRL	communicating MDPs	$\mathcal{O}\left(\frac{D^2S^2A}{\Delta}\log(T)\right)$
Lower Bound	ergodic MDPs, known rewards	$\Omega\left(c_{\text{bk}}\log(T)\right),\ \Omega\left(\frac{DSA}{\Phi}\log(T)\right)$

Algorithm	Setup	Regret
UCRL2	communicating MDPs	$\widetilde{\mathcal{O}}\left(DS\sqrt{AT}\right)$
KL-UCRL	communicating MDPs	$\widetilde{\mathcal{O}}\left(DS\sqrt{AT}\right)$
REGAL	weakly comm. MDPs, known rewards	$\widetilde{\mathcal{O}}\left(BS\sqrt{AT}\right)^*$
A-J	communicating MDPs, known rewards	$\widetilde{\mathcal{O}}\left(D\sqrt{SAT}\right),\ T \geq S^5A$
Lower Bound	known rewards	$\Omega\left(\sqrt{DSAT}\right),\ T \geq DSA$

*B denotes the span of bias function of true MDP, and $B \leq D$

UCRL2

UCRL2 is an optimistic algorithm that works in episodes of increasing lengths.

- At the beginning of each episode k, it maintains a set of plausible MDPs $\mathcal{M}_k$ (which contains the true MDP w.h.p.)
- It then computes an optimal policy π_k, which has the largest gain over all MDPs in $\mathcal{M}_k$ ($\pi_k \in \mathrm{argmax}_{M' \in \mathcal{M}_k, \pi}\, g^\pi(M')$).
  - For computational efficiency, UCRL2 computes an $\frac{1}{\sqrt{t_k}}$-optimal policy, where t_k is the starting step of episode k
  - To find a near-optimal policy, UCRL2 uses Extended Value Iteration
- It then follows π_k within episode k until the number of visits for some pair (s, a) is doubled (and so, a new episode starts).

UCRL2

Notations:

- $k \in \mathbb{N}$: index of an episode
- $N_k(s,a)$: total no. visits of pairs (s,a) before episode k
- $\hat{p}_k(\cdot|s,a)$: empirical transition probability of (s,a) made by observations up to episode k
- $\hat{r}_k(s,a)$: empirical reward distribution of (s,a) made by observations up to episode k
- π_k: policy followed in episode k
- $\mathcal{M}_k$: set of models for episode k (defined next)
- $\nu_k(s,a)$: no. of visits of pairs (s,a) seen so far in episode k

UCRL2: Main ingredients

- **The set of plausible MDPs** $\mathcal{M}_k$: for confidence parameter δ, define

$$\mathcal{M}_k = \Bigg\{ M' = (\mathcal{S}, \mathcal{A}, \tilde{r}, \tilde{p}) : \forall (s,a), |\tilde{r}(s,a) - \hat{r}_k(s,a)| \leq \sqrt{\frac{3.5 \log(2SAt/\delta)}{N_k(s,a)^+}}$$
$$\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \leq \sqrt{\frac{14S \log(2At/\delta)}{N_k(s,a)^+}} \Bigg\}$$

- **Optimistic gain**: find in $\mathcal{M}_k$ the MDP that leads to the highest gain. We need to solve for episode k:

$$\text{maximise over } (M, \pi) \quad g^\pi(M)$$
$$\text{subject to } M \in \mathcal{M}_k$$

UCRL2 pseudo-code

Algorithm. UCRL2

Input: Initial state s_0, precision δ, $t = 1$

For each episode $k \geq 1$:

1. Initialisation. $t_k = t$ (start time of the episode)
   Update $N_k(s,a)$, $\hat{r}_k(s,a)$, and $\hat{p}_k(s,a)$ for all (s,a)
2. Compute the set of possible MDPs $\mathcal{M}_k$ (using δ)
3. Compute the policy
   $\pi_k \leftarrow$ ExtendedValueIteration$(\mathcal{M}_k, 1/\sqrt{t_k})$
4. Execute π_k and end the episode:
   While $[\nu_k(s_t, \pi_k(s_t)) < \max(1, N_k(s_t, \pi_k(s_t)))]$
    - Play $\pi_k(s_t)$, observe the reward and the next state
    - Update $\nu_k(s_t, \pi_k(s_t)) \leftarrow \nu_k(s_t, \pi_k(s_t)) + 1$ and $t \leftarrow t + 1$

Extended value iteration

Set of plausible MDPs $\mathcal{M}_k$:

$$\mathcal{M}_k = \Bigg\{ M' = (\mathcal{S}, \mathcal{A}, \tilde{r}, \tilde{p}) : \forall (s,a), |\tilde{r}(s,a) - \hat{r}_k(s,a)| \leq d(s,a)$$

$$\|\tilde{p}(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \leq d'(s,a) \Bigg\}$$

We wish to find $M' \in \mathcal{M}_k$ and a policy π_k maximising $g^{\pi}(M')$ over all possible $M' \in \mathcal{M}_k$ and policy π.

Ideas:

a. we can fix the reward to its maximum: $\bar{r}(s,a) = \hat{r}(s,a) + d(s,a)$

b. solve a large MDP whose set of actions is $\mathcal{A}'_s$ where $(a,q) \in \mathcal{A}'_s$ if and only if $q \in \mathcal{P}_k(s,a)$ with:

$$\mathcal{P}_k(s,a) = \{q : \|q(\cdot) - \hat{p}_k(\cdot|s,a)\|_1 \leq d'(s,a)\}$$

Extended value iteration

Solution: apply one of the known algorithms to find an optimal policy in MDPs, i.e., value iteration algorithm.

Extended Value Iteration: For all $s \in \mathcal{S}$, starting from $u_0(s) = 0$:

$$u_{i+1}(s) = \max_{a \in \mathcal{A}} \Big\{ \bar{r}(s,a) + \max_{q \in \mathcal{P}_k(s,a)} u_i^\top q \Big\}$$

- $\mathcal{P}_k(s,a)$ is a polytope, and the inner maximisation can be done in $\mathcal{O}(S)$ operations.
- To obtain an ε-optimal policy, the update is stopped when $\max_s(u_{i+1}(s) - u_i(s)) - \min_s(u_{i+1}(s) - u_i(s)) \leq \varepsilon$

UCRL2: Regret guarantees

Let $\pi =$UCRL2 Regret up to time T: $\mathcal{R}^\pi(T) = Tg^\star - \sum_{t=1}^T r(s_t^\pi, a_t^\pi)$, a random variable capturing the learning cost and the mixing time problems.

Theorem *W.p. at least $1-\delta$, the regret of UCRL2 satisfies, for any initial state, for any $T > 1$,*

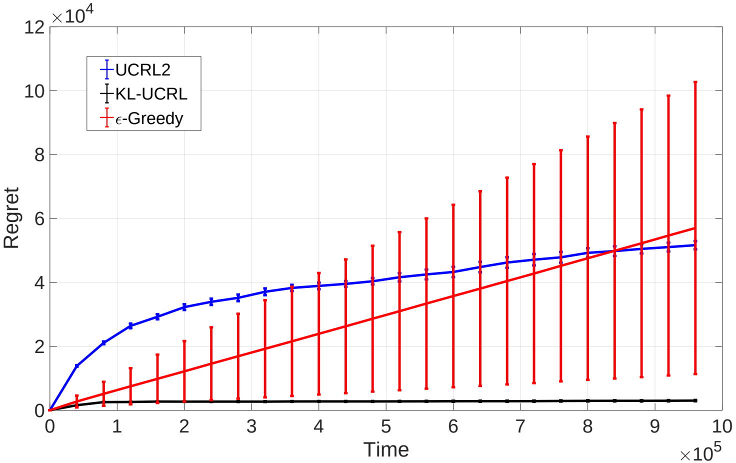
$$\mathcal{R}^\pi(T) \leq 34DS\sqrt{AT\log(\frac{T}{\delta})}.$$

For any initial state, and any $T \geq 1$, we have w.p. at least $1-3\delta$,

$$\mathcal{R}^\pi(T) \leq 34^2\frac{D^2S^2A\log(\frac{T}{\delta})}{\epsilon} + \epsilon T.$$

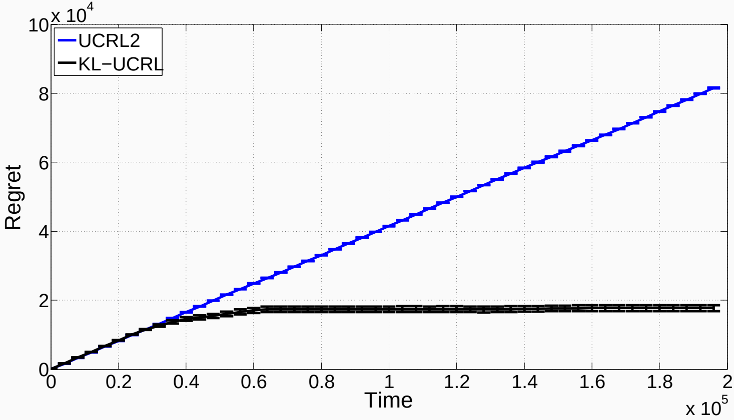
Regret

6 states, $\delta = 0.05$ (for UCRL2), ϵ-greedy: $\epsilon_t = \min(1, 1000/t)$



Regret

12 states, $\delta = 0.05$ (for UCRL2)



References

Episodic RL

- **UCBVI algorithm:**
  M. Gheshlaghi Azar, I. Osband, and R. Munos, "Minimax regret bounds for reinforcement learning," *Proc. ICML*, 2017.

Ergodic RL

- **UCRL algorithm:**
  P. Auer & R. Ortner, "Logarithmic online regret bounds for undiscounted reinforcement learning," *Proc. NIPS*, 2006.
- **UCRL2 algorithm and minimax LB:**
  P. Auer, T. Jaksch, and R. Ortner, "Near-optimal regret bounds for reinforcement learning," *J. Machine Learning Research*, 2010.