Lecture 7: Exploration - Finite Horizon MDPs, Regret Minimization in Tabular MDPs

SUMS 707 - Basic Reinforcement Learning

Gabriela Moisescu-Pareja and Viet Nguyen

McGill University, Mila

March 18, 2021

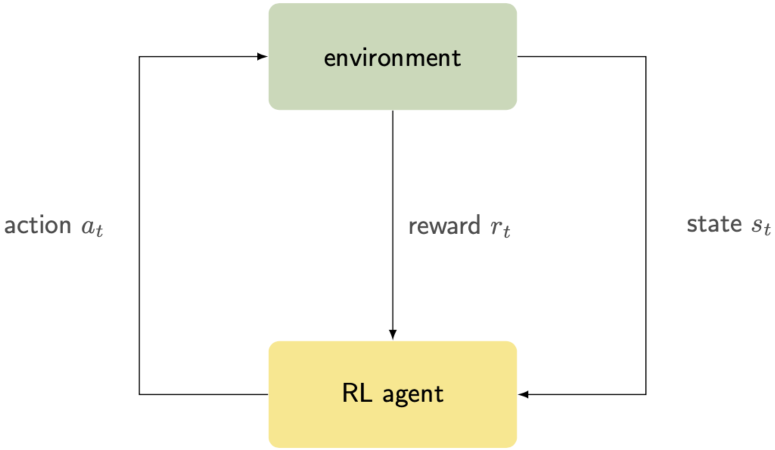
A big problem

RL studies the problem of **sequential** decision-making when the environment is unknown, *but can be learned through direct ineraction*.
Problem:

- Need a huge amount of data to learn a satisfactory policy
- Cannot be used in domains where sampling is expensive, long, or simulations are not possible

We want to do RL in a *sample-efficient* way! This very often means that we might not explore ALL the possibilities, but we can be *confident* that what we come up with given what we see is *good*.

Agent-Environment Interaction



Outline

The need for directed exploration:

- Finite-horizon MDPs, value iteration
- Regret
- Why ϵ-greedy might not be the ideal thing to do

Two paradigms (there's a 3rd one that's cooler but we'll stick with these two for now)

- Optimism in the Face of Uncertainty (OFU)
- Posterior Sampling (or Thompson Sampling, or randomized algorithms, ...)

Finite-Horizon MDP

A *finite-horizon* MDP is a tuple $M = (\mathcal{S}, \mathcal{A}, r_h, p_h, H)$ where:

- $\mathcal{S}$ is the state space
- $\mathcal{A}$ is the action space, *finite*
- H is the *horizon*, or episode length
- p_h is the transition distribution, $p_h(\cdot|s,a) \in \Delta(\mathcal{S}), h \in [H]$
- r_h is the *expectation* of the random rewards, $r_h(s,a) \in [0,1], h \in [H]$

An agent acts according to a *time-variant* policy

$$\pi_h : \mathcal{S} \to \mathcal{A}, \quad h \in [H]$$

$T = KH$

(aside) Bandits

Value functions, optimality

Analogous to what we studied previously, our value functions look like:

$$Q_h^\pi(s,a) = r_h(s,a) + \mathbb{E}\left[\sum_{l=h+1}^{H} r_l(s_l, \pi(s_l))\right]$$
$$V_h^\pi(s) = Q_h^\pi(s, \pi_h(s))$$

and our optimality equations:

$$Q_h^*(s,a) = \sup_\pi Q_h^\pi(s,a)$$
$$\pi_h^*(s) = \arg\max_{a \in \mathcal{A}} Q_h^*(s,a)$$

We immediately have that $Q_h, V_h \in [0, H-(h-1)]$ (why?).

Bellman Equations

Our Bellman *expectation* equations:

$$\begin{aligned} Q_h^\pi(s,a) &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[Q_{h+1}^\pi(s', \pi_{h+1}(s'))\right] \\ &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[V_{h+1}^\pi(s')\right] \end{aligned}$$

and Bellman optimality equations:

$$\begin{aligned} Q_h^*(s,a) &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[\max_{a' \in \mathcal{A}} Q_{h+1}^*(s', a')\right] \\ &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)}\left[V_{h+1}^*(s')\right] \end{aligned}$$

Value Iteration

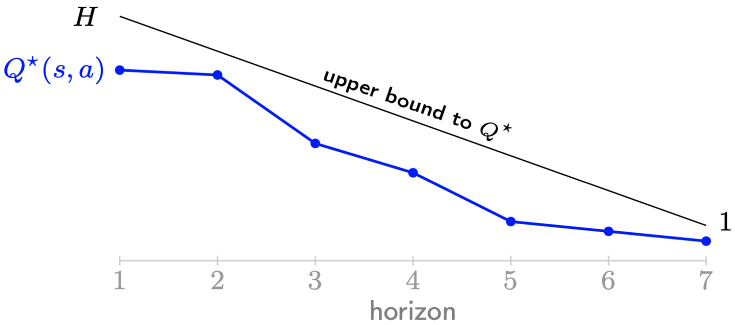
We previously stated the value iteration algorithm in the discounted MDP case. Adapting what we know to the episodic MDP setting is not very hard: Set $Q^*_{H+1}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

- for $h = H, \ldots, 1$,
  - for $(s,a) \in \mathcal{S} \times \mathcal{A}$,
    - Compute

$$\begin{aligned} Q^*_h(s,a) &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q^*_{h+1}(s',a') \right] \\ &= r_h(s,a) + \mathbb{E}_{s' \sim p_h(\cdot|s,a)} \left[V^*_{h+1}(s') \right] \end{aligned}$$

  - return $\pi^*_h(s) = \arg\max_{a \in \mathcal{A}} Q^*_h(s,a)$

Value Iteration (but more visual)



Online Learning

This process simulates an RL agent learning as it does things in an environment.

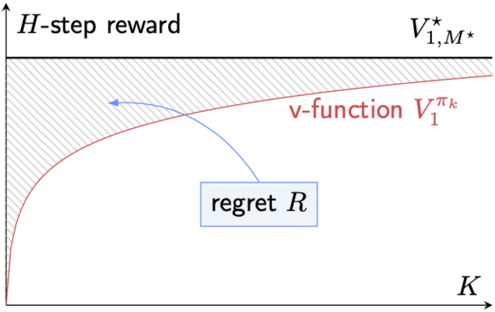
We are running over K episodes. Initialize Q_h to zero.

- for $k \in [K]$
  - Define $\pi_k = \{\pi_{kh}, h \in [H]\}$ based on $\{Q_{kh}\}_{h=1}^{H}$ (run value iter.)
  - Observe the initial state s_{k1}
  - for $h = 1, \ldots, H$
    - Choose action $a_{kh} = \pi_{kh}(s_{kkh})$
    - Observe reward r_{kh} and $s_{k,h+1}$
  - Remember $(s_{kh}, a_{kh}, r_{kh})_{h=1}^{H}$
  - Compute $(Q_{k+1,h})_{h=1}^{H}$

Regret

A lot

Regret



We define the regret of an algorithm $\mathfrak{U} = \{\pi_k\}_{k=1}^{K}$ on an unknown true MDP M^* after K episodes as:

$$\text{Regret}(K, M^*, \mathfrak{U}) = \sum_{k=1}^{K} \left(V^*(s_{k1}) - V^{\pi_k}(s_{k1})\right)$$

Q-learning + ϵ-greedy

Recall the ϵ-greedy action-selection protocol:

$$a_{kh} = \begin{cases} \arg\max_{a \in \mathcal{A}} Q_{kh}(s_{kh}, a) & \text{w.p. } 1 - \epsilon_{kh} \\ U(\mathcal{A}) & \text{otherwise} \end{cases}$$

Where α_t is some learning rate (assuming you scheduled them properly, RM go brrr), we have the Q-learning updates:

$$\begin{aligned} Q_{k+1,h}(s_{kh}, a_{kh}) &\leftarrow Q_{kh}(s_{kh}, a_{kh}) + \alpha_t \left(r_{kh} + \max_{a' \in \mathcal{A}} Q_{k,h+1}(s_{k,h+1}, a') - Q_{kh}(s_{kh}, a \right. \\ &= (1 - \alpha_t) Q_{kh}(s_{kh}, a_{kh}) + \alpha_t \left(r_{kh} + \max_{a' \in \mathcal{A}} Q_{k,h+1}(s_{k,h+1}, a') \right) \end{aligned}$$

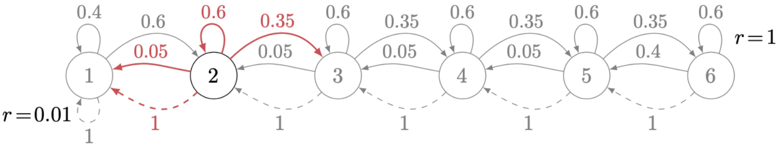
Q-learning + ϵ-greedy: problems

- Exploration strategy relies on **biased** estimates of Q_{kh}
- Samples are used **once**
- **Dithering effect**: exploration not effective in covering the state space
- **Policy shift**: policy changes at each step (problems on stability)

Regret of $\Omega\left(\min\left\{T, A^{H/2}\right\}\right)$ (Jin et. al., 2018).

Pathological Environments: Riverswim (1)

(Strehl, Littman, 2008)

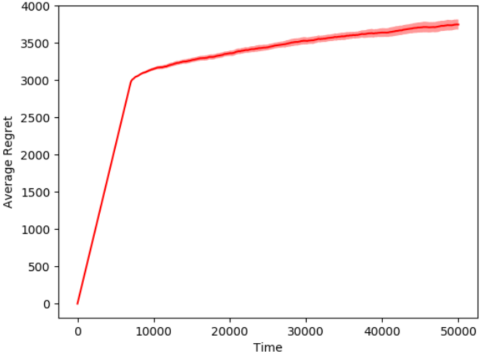


Notation: $\pi_L(s) = L, \pi_R(s) = R$ In Riverswim, if you set ϵ_t to something funny like 1.0, you have linear regret.

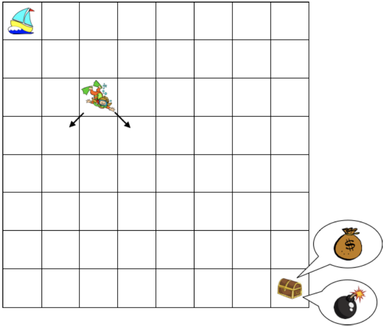
Pathological Environments: Riverswim (2)

Set:

$$\epsilon_t = \begin{cases} 1.0, & t < 6000 \\ \frac{\epsilon_0}{N(s_t)^{1/2}}, & \text{otherwise} \end{cases}$$



Pathological Environments: Deep Sea (1)



$\mathcal{A} = \{0, 1\}$, whether 0 is the left or the right action depends on the depth. A left action gets rewarded with 0, and a right action gets rewarded with a small negative number. What happens at the bottom of the sea stays at the bottom of the sea.

Pathological Environments: Deep Sea (2)

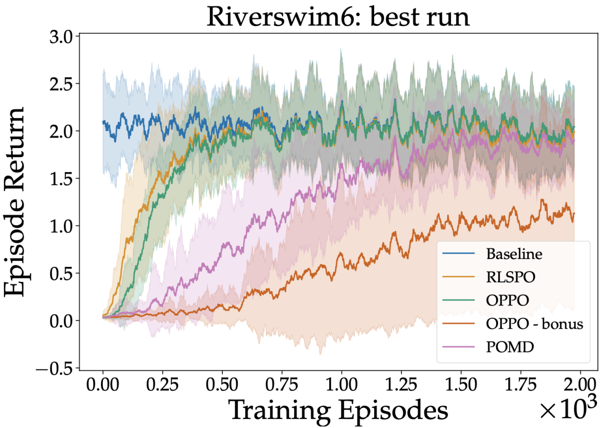
You might see that there is a $1/2^N$ probability that you actually get the treasure if you play the ϵ-greedy policy.

For cool pathological environments to stress-test your exploration algorithms (and to see why explorations is a big yikes), check out `bsuite`.

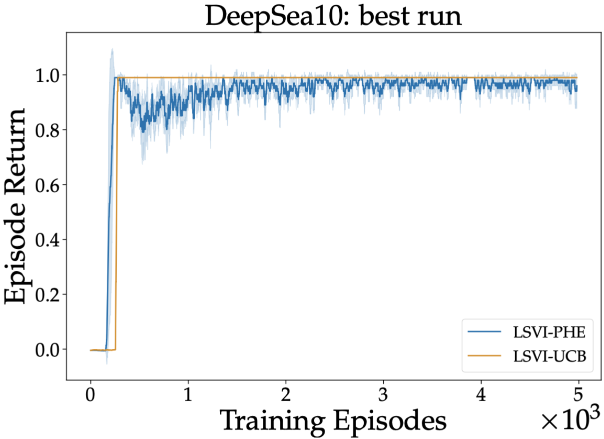
ϵ-greedy

- Undirected exploration
- Inefficient use of samples
- $\Omega\left(\min\left\{T, A^{H/2}\right\}\right)$ (yikes)

Directed Exploration: SotA



Directed Exploration: SotA



Directed Exploration: Tabular Setting

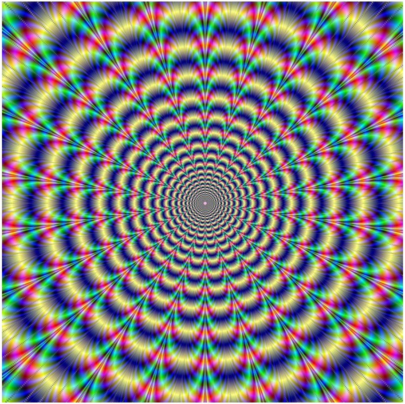
Jaksch et. al. proved in 2010 in their *seminal* paper that:

- stationary transitions ($p_1 = \cdots = p_H$): $\Omega\left(\sqrt{HSAT}\right)$
- non-stationary (effective number of states is $S' = HS$): $\Omega\left(H\sqrt{SAT}\right)$

Two paradigms for exploration:

- OFU
- PS

Model-Based Exploration



We first look at how we can explore by hallucinating some MDP from our experience and act accordingly.

Optimism in the Face of Uncertainty

The canonical intuition that is stated everytime this topic is touched upon, but I don't really find the intuition useful.

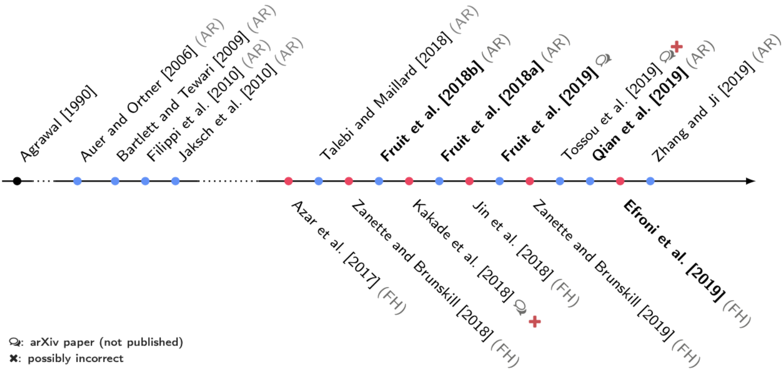
Central idea: When you are uncertain, consider the *best possible world* (reward-wise)

- (Exploitation) If the best possible world is correct, you have *no regret*
- (Exploration) IF the best possible world is wrong, you learn some useful information at least

When we are optimistic, what we mean is that our *estimates* for the value functions are optimistic.

History: OFU for Regret Minimization

FH: finite-horizon
AR: average reward



: arXiv paper (not published)
✖: possibly incorrect

Online Learning Problem

We are running over K episodes. Initialize Q_h to zero. Imagine that we have $\mathcal{D}_k$ which stores our history.

- for $k \in [K]$
  - **Compute** Q_{kh} from $\mathcal{D}_k$
  - **Define** $\pi_k = \{\pi_{kh}, h \in [H]\}$ **based on** $\{Q_{kh}\}_{h=1}^{H}$
  - **Observe the initial state** s_{k1}
  - for $h = 1, \ldots, H$
    - Choose action $a_{kh} = \pi_{kh}(s_{kkh})$
    - Observe reward r_{kh} and $s_{k,h+1}$
  - Remember $(s_{kh}, a_{kh}, r_{kh})_{h=1}^{H}$
  - Compute $(Q_{k+1,h})_{h=1}^{H}$

This is when we know what the underlying MDP is. When we don't know, we make one up (next slide).

Model-Based Learning

To compute Q_{kh} from $\mathcal{D}_k$ to define π_{kh}, we write the empirical MDP $\hat{M}_k$ where the transitions and rewards are now:

$$\hat{p}_{kh}(s'|s,a) = \frac{\sum_{\tau=1}^{k-1} \mathbb{1}\left((s_{\tau h}, a_{\tau h}, s_{\tau,h+1}) = (s,a,s')\right)}{N_{kh}(s,a)}$$

$$\hat{r}_{kh}(s,a) = \frac{\sum_{\tau=1}^{k-1} r_{\tau h} \cdot \mathbb{1}\left((s_{\tau h}, a_{\tau h}) = (s,a)\right)}{N_{kh}(s,a)}$$

With these, we can now compute Q_{kh} and subsequently, π_{kh}, in the online learning algorithm.

Uncertainty in our Empirical Model

What is the uncertainty around our empirical MDP? Consider:

$$M_k = \left\{M : \forall h \in [H], r_h(s,a) \in B^r_{kh}(s,a), p_h(\cdot|s,a) \in B^p_{kh}(s,a), \forall (s,a)\right\}$$

the space of "interesting" MDPs, where our balls are the confidence sets:

$$\begin{aligned}
B^r_{kh}(s,a) &:= [\hat{r}_{kh}(s,a) \pm \beta^r_{kh}(s,a)] \\
B^p_{kh}(s,a) &:= \left\{p(\cdot|s,a) \in \Delta(\mathcal{S}) : \|p(\cdot|s,a) - \hat{p}_{kh}(\cdot|s,a)\|_1 \leq \beta^p_{kh}(s,a)\right\}
\end{aligned}$$

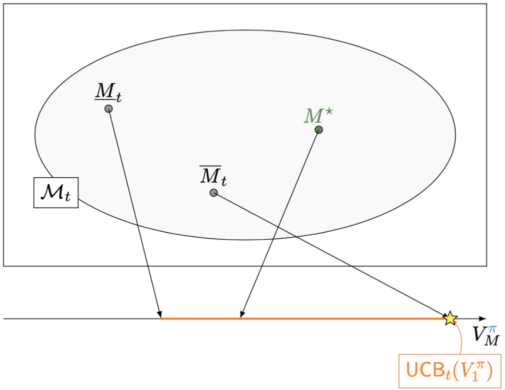
where $\|\cdot\|_1$ is the *total variation*.

Confidence bounds

Hoeffding b like

$$\beta^r_{kh}(s,a) \propto \sqrt{\frac{\log\left(N_{kh}(s,a)/\delta\right)}{N_{kh}(s,a)}}, \quad \beta^p_{kh}(s,a) \propto \sqrt{\frac{\mathcal{S}\log\left(N_{kh}(s,a)/\delta\right)}{N_{kh}(s,a)}}$$

Optimism but more visual



w.p.a.l. $1-\delta$, the true MDP M^* is within the confidence region!

Extended Value Iteration

[Jaksch et. al., 2010]
Set $Q_{k,H+1}(s,a) = 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

- for $h = H, \ldots, 1$,
  - for $(s,a) \in \mathcal{S} \times \mathcal{A}$,
    - Compute

$$
\begin{aligned}
Q_{kh}(s,a) &= \max_{r_h \in B^r_{kh}(s,a)} r_h(s,a) + \max_{p_h \in B^p_{kh}(s,a)} \mathbb{E}_{s' \sim p_h(\cdot|s,a)} \left[V^*_{h+1}(s') \right] \\
&= \hat{r}_{kh}(s,a) + \beta^r_{kh}(s,a) + \max_{p_h \in B^p_{kh}(s,a)} \mathbb{E}_{s' \sim p_h(\cdot|s,a)} \left[V^*_{h+1}(s') \right]
\end{aligned}
$$

$$
V_{kh}(s) = \min \left\{ H - (h-1), \max_{a \in \mathcal{A}} Q_{kh}(s,a) \right\}
$$

  - return $\pi_{kh}(s) = \arg\max_{a \in \mathcal{A}} Q_{kh}(s,a)$

With very high probability, $Q_{kh}(s,a) \geq Q^*_h(s,a)$.

UCRL2-CH for Finite-Horizon

(Jaksch et. al., 2010)

Theorem

For any tabular MDP with stationary transitions, the UCRL2 algorithm with Chernoff-Hoeffding bounds, with high prob., suffers a regret

$$\text{Regret}(K, M^*, \text{UCRL2-CH}) = \tilde{\mathcal{O}}\left(H\mathcal{S}\sqrt{\mathcal{A}T} + H^2\mathcal{S}\mathcal{A}\right)$$

(recall the lower bound $\Omega\left(\sqrt{H\mathcal{S}\mathcal{A}T}\right)$*)*

Azar et. al., 2017, had the idea of playing with the Q_{kh} terms and get rid of the maximizations over B^p and B^r, thus now we don't need to do EVI, we can just add a bonus term. They showed:

$$Q_{kh}(s,a) \leq \hat{r}_{kh}(s,a) + \beta^r_{kh}(s,a) + H\beta^p_{kh}(s,a) + \mathbb{E}_{s' \sim \hat{p}_{kh}(\cdot|s,a)}\left[V_{k,h+1}(s')\right]$$

They essentially **combined uncertainties in rewards and transitions** into one chonky term:

$$b_{kh}(s,a) = \beta^r_{kh}(s,a) + H\beta^r_{kh}(s,a)$$

essentially reduces the problem to doing *value iteration* on this MDP:

$$M = (\mathcal{S}, \mathcal{A}, \hat{r}_{kh}, \hat{p}_{kh}, H)$$

Big brainy things for tighter bounds

They showed (with Chernoff-Hoeffding) that by setting:

$$b_{kh}(s,a) = (H+1)\sqrt{\frac{\log\left(N_{kh}(s,a)/\delta\right)}{N_{kh}(s,a)}} < \beta^r_{kh} + H\beta^p_{kh}$$

one gets: [Azar et. al., 2017]

Theorem

For any tabular MDP with stationary transitions, UCBVI-CH, with high probability, suffers a regret:

$$\text{Regret}(K, M^*, \text{UCBVI-CH}) = \tilde{\mathcal{O}}\left(H\sqrt{\mathcal{SAT}} + H^2\mathcal{S}^2\mathcal{A}\right)$$

Refining confidence bounds

- Use Bernstein-Freedman concentration inequalities for Bernstein-type bounds
- Work has been done in this fashion (e.g. Zanette and Brunskill, 2019)
- Still does not match the theoretical lower bound
- More interesting works actually involves playing around with Q_{kh}, directly making it optimistic (e.g. Opt-RTDP by Efroni et. al., 2019), etc...
- Still a challenge to match the theoretical lower bound

Posterior Sampling (PS)

Next time!

Lecture recap

- We study finite-horizon MDPs
- Value iteration to solve for π_h^*
- Regret, and why ϵ-greedy straight up fails sometimes
- Model-based exploration:
  - Hallucinate the empirical MDP and do value iteration on there
  - But how good is our imagination?
  - Confidence bounds captures EXACTLY this feeling of uncertainty!
  - UCB: Just take the MDP in our confidence set that predicts the highest value function!
  - PS: We sample from the confidence set

Next time

- Model-based PS
- Model-free UCB and PS
- Graduate from the tabular setting

References

- Exploration - Exploitation in Reinforcement Learning, RLGammaZero (Ghavamzadeh, Lazaric, Pirotta)
- Reinforcement Learning: Theory and Algorithms by Alekh Agarwal, Nan Jiang, Sham M. Kakade
- Algorithms for Reinforcement Learning by Csaba Szepesvári
- Reinforcement Learning: An Introduction by Andrew Barto and Richard S. Sutton
- "Introduction to Reinforcement Learning" Lectures by David Silver
- "CS 598 - Statistical Reinforcement Learning" Notes by Nan Jiang