Lecture 8: Exploration - Model-based, model-free, deep exploration

SUMS 707 - Basic Reinforcement Learning

Gabriela Moisescu-Pareja and Viet Nguyen

McGill University, Mila

March 18, 2021

Recap: Finite-Horizon MDPs

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

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

$T = KH$

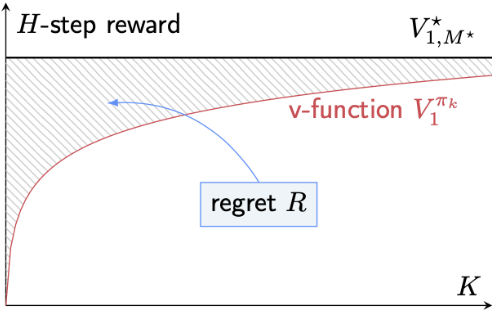
Recap: Online Learning

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



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

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

Regret lower-bound in the tabular setting

Jaksch et. al., 2010:

- stationary transitions ($p_1 = \cdots = p_H$): $\Omega\left(\sqrt{H\mathcal{S}\mathcal{A}T}\right)$
- non-stationary (effective number of states is $S' = HS$): $\Omega\left(H\sqrt{\mathcal{S}\mathcal{A}T}\right)$

Recap: Model-based exploration, OFU

- Hallucinate the empirical MDP and do value iteration on there
- Confidence bounds captures EXACTLY this feeling of uncertainty!
- UCB: Just take the MDP in our confidence set that predicts the highest value function!

Our empirical MDP:

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

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

Confidence region: MDPs we care about

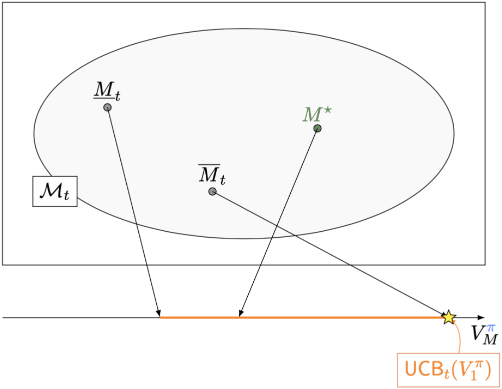
$$M_k = \left\{M : \forall h \in [H], r_h(s,a) \in B^r_{kh}(s,a), p_h(\cdot|s,a) \in B^p_{kh}(s,a), \forall (s,a)\right\}$$

where

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

β^r and β^p are chosen by magic.

A drawing of the confidence region



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

[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 $\Omega\left(\sqrt{H\mathcal{S}\mathcal{A}T}\right)$*)*

We also saw UCBVI (Azar et. al., 2017) which achieved

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

UCBVI turns the problem back into a standard value iteration with some bonus terms in the computation of Q_{kh}.

Interesting things with UCBVI

Instead of using Chernoff-Hoeffding inequalities, use Bernstein-Freedman-type inequalities to obtain:

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

which actually matches the lower bound $\Omega\left(\sqrt{H\mathcal{S}\mathcal{A}T}\right)$, but has a longer warm-up phase.

Model-based posterior sampling (PS/TS) algorithms

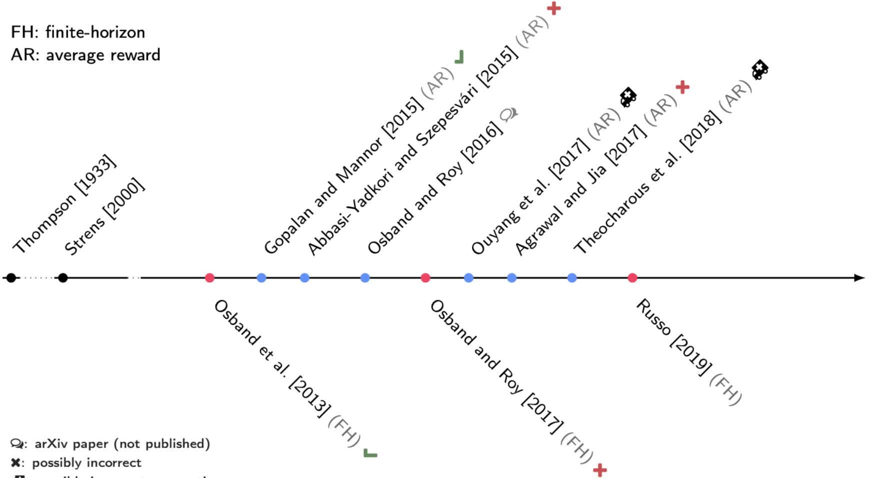
Big idea: to maintain a Bayesian posterior for the unknown MDP.

- Every timestep t, posterior distribution μ_t, sample some $M_t \sim \mu_t$
-
  - Few samples $\implies$ uncertainty in the estimates
  - More samples $\implies$ posterior μ_t concentrates on the true MDP

A brief history of PS

Tabular MDPs

FH: finite-horizon
AR: average reward



: arXiv paper (not published)
✖: possibly incorrect
: possibly incorrect assumptions

Posterior sampling

(First proposed in 2000 by Strens, but important work by Osband et. al. in 2013) We are given some prior μ_1, we initialize the dataset $\mathcal{D}_1 = \emptyset$.

For episode $k \in [K]$,

- Observe the initial state s_{k1}
- Sample $M_k \sim \mu_k(\cdot|\mathcal{D}_k)$
- Compute $\pi_k \in \arg\max_\pi V^\pi_{1,,M_k}$
- for $h = 1, \ldots, H$
  - $a_k h = \pi_{kh}(s_{kh})$
  - observe r_{kh} and $s_{k,h+1}$
- Add trajectory $\{(s_{kh}, a_{kh}, r_{kh})\}$ to $\mathcal{D}_{k+1}$

Priors, posteriors

A prior distribution would satisfy the following:

$$\forall \Theta, \mathbb{P}\left(M^* \in \Theta\right) = \mu_1(\Theta)$$

We can formulate the posterior distribution:

$$\forall \Theta, \mathbb{P}\left(M^* \in \Theta | \mathcal{D}_k, \mu_1\right) = \mu_k(\Theta)$$

Examples of priors that are commonly used:

- Dirichlet for transitions
- Beta, Normal-Gamma for rewards

Transition model updates with Dirichlet priors

Assuming r is known, at timestep t, given μ_t and the transition s_t, a_t, s_{t+1}, we want to compute μ_{t+1}.

- μ_1 is a Dirichlet distribution
- $\mu_t(s,a) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_{\mathcal{S}})$ on $p(\cdot|s,a)$ is also a Dirichlet distribution
- We observe $s_{t+1} \sim p(\cdot|s,a)$ (outcome of a multivariate Bernoulli such that $s_{t+1} = i$. The Bayesian posterior is updated as follows:

$$\mu_{t+1}(s,a) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_i + 1, \ldots, \alpha_{\mathcal{S}})$$

  - Posterior mean vector $\hat{p}_{t+1}(s_i|s,a) = \frac{\alpha_i}{n}$ where $n = \sum_{i=1}^{\mathcal{S}} \alpha_i$
  - Variance bounded by $\frac{1}{n}$

Posterior sampling RL (PSRL): performance, regret

It turns out that in many settings, PSRL outperforms UCB, even though the latter has (in general) better

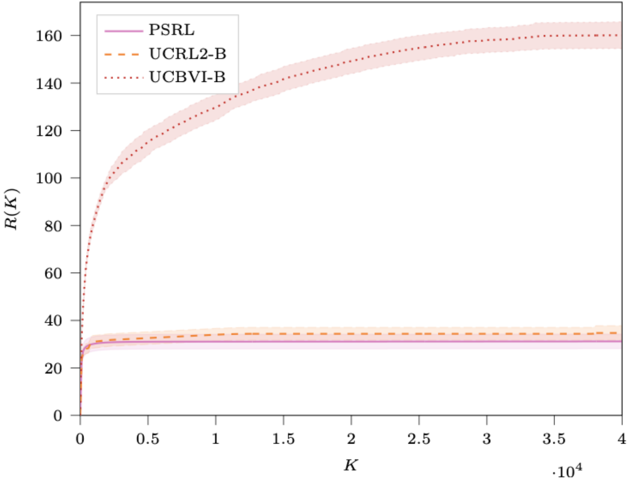
Theorem

For any prior μ_1 with any independent Dirichlet prior over stationary transitions, the Bayesian regret (expectation of frequentist regret) of PSRL is bounded as

$$\text{BayesianRegret}(K, \mu_1, PSRL) = \tilde{\mathcal{O}}\left(H\mathcal{S}\sqrt{\mathcal{A}T}\right)$$

Compared to the theoretical lower bound, this expected regret bound has an additional factor of $\sqrt{\mathcal{A}T}$.

PSRL on Riverswim



From PSRL to randomized value functions

- PSRL samples an MDP from the space of MDPs according to some posterior distribution
- This sampling process can be very costly, even intractable at times...
- On the other hand, when we do value iteration, most of the business happens at the value function level

The 200 IQ move: randomly sampling MDPs makes value functions random, why not add randomness directly to the value functions themselves?

Big brain idea: as long as you sample from a distribution with enough concentration and anti-concentration properties (hint: normal distribution!), you are "approximately sampling" from the posterior! Can you see why I am so hype about this?

Randomized Least-Squares Value Iteration (RLSVI)

(Osband, Van Roy, Wen, 2016) At every episode, get the empirical MDP $\hat{M}_k = (\mathcal{S}, \mathcal{A}, \hat{p}_h, \hat{r}_h, H)$.

For $h = H, \ldots, 1$

- Sample $\xi_{kh} \sim N(0, \sigma_{kh}^2 I)$
- Compute

$$\forall (s,a), \hat{Q}_{kh}(s,a) = \hat{r}_{kh}(s,a) + \xi_{kh}(s,a) + \sum_{s' \in \mathcal{S}} \hat{p}_{kh}(s'|s,a) \hat{V}_{k,h+1}(s')$$

Finally, return $\hat{Q}_{kh}, h \in [H]$.

RLSVI: Frequentist regret

Theorem

For any tabular MDP with non-stationary transitions, RLSVI with

$$\sigma_{kh}(s,a) = \tilde{\mathcal{O}}\left(\sqrt{\frac{\mathcal{S}H^3}{N_{kh}(s,a)+1}}\right)$$

w/ high probability, suffers a frequentist regret of

$$\text{Regret}(K, M^*, RLSVI) = \tilde{\mathcal{O}}(H^{5/2}\mathcal{S}^{3/2}\sqrt{\mathcal{A}T})$$

Looks $H^{3/2}\mathcal{S}$ worse than the theoretical lower bound for *non-stationary* $\Omega\left(H\sqrt{\mathcal{S}\mathcal{A}T}\right)$.

Model-based issues

- Space $\mathcal{O}\left(H\mathcal{S}^2\right)$
- Time $\mathcal{O}\left(KH\mathcal{S}^2\mathcal{A}\right)$, where $H\mathcal{S}^2\mathcal{A}$ comes from planning by value iteration.

You can sort of solve the time complexity issue by methods mased on incremental planning (Opt-RTDP, maybe I said something about this algo last lecture)
To solve space complexity, we can think of methods that *avoid estimating rewards and transitions*, basically just not compute the empirical model at all. This takes us to model-free methods.

Optimistic Q-learning

We maintain an estimate $\hat{Q}_h$. We learn on the fly:

For $k = 1, \ldots, K$

- Observe s_{k1}
- For $h = 1, \ldots, H$
  - do $a_{kh} = \arg\max_a \hat{Q}_h(s_{kh}, a)$
  - Observe $r_{kh}, s_{k,h+1}$
  - $N_h(s_{kh}, a_{kh}) += 1$
  - Update

$$Q_h(s_{kh}, a_{kh}) = (1 - \alpha_t) Q_h(s_{kh}, a_{kh}) + \alpha_t \left(r_{kh} \hat{V}_{h+1}(s_{k,h+1}) + b_t \right)$$

  - Set $\hat{V}_h(s_{kh}) \min \left\{ H - (h-1), \max_{a \in \mathcal{A}} Q_h(s_{kh}, a) \right\}$

Step size α_t of order $\mathcal{O}(1/t)$ or $\mathcal{O}(1/\sqrt{t})$, with $t = N_{kh}(s, a)$ for Q-learning, for Opt-QL, we use $\alpha_t = (H+1)/(H+t)$.

Optimistic Q-learning: regret

Theorem

(Jin et. al., 2018) FOr any tabular MDP with non-stationary transitions, Opt-QL with Hoeffding inequalities $(b_t = \tilde{\mathcal{O}}\left(\sqrt{H^3/t}\right))$ *with high probability suffers a regret*

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

The bound does *not* improve in stationary MDPs :(, but can get tighter bounds with Bernstein-Freedman inequalities.

Open Questions (from 2019)

- Prove a frequentist regret bound for PSRL
- Whether the gap between the regret of model-based and model-free should exist?
- Which algorithms are better in practice?

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