Sample Complexity of Episodic Fixed-Horizon Reinforcement Learning

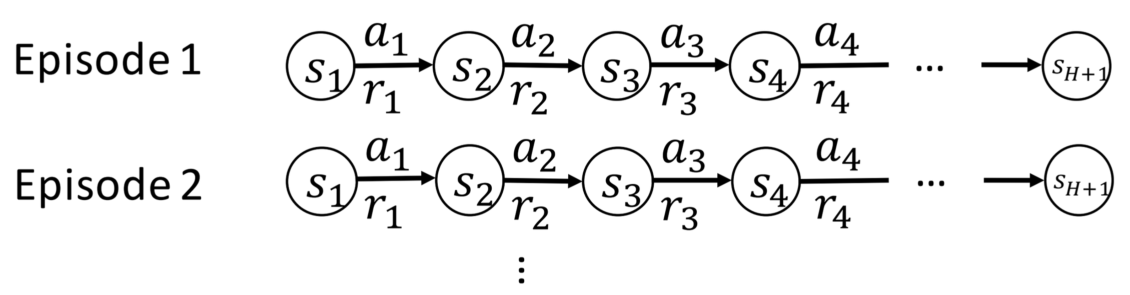
Christoph Dann, Emma Brunskill

cdann@cdann.net ebrun@cs.cmu.edu

Summary

- Episodic Reinforcement Learning: Agent interacts in episodes of length H with finite MDP of S states and A actions
- Episodic RL is a good model for many important applications, e.g., drug treatment optimization, automated tutoring of students for exams.
- We provide an algorithm and show that it achieves near-optimal expected return in all but
$$\tilde{O}\left(\frac{S^2AH^2}{\epsilon^2}\log\frac{1}{\delta}\right)$$
episodes with high probability
- We show that no algorithm can achieve the same guarantee in number of episodes less than
$$\tilde{\Omega}\left(\frac{SAH^2}{\epsilon^2}\log\frac{1}{\delta+c}\right).$$
- These bounds match up to log-terms and a factor of S and are tighter than prior bounds by at least H.

Episodic Reinforcement Learning



We assume a finite Markov decision process (MDP) with states $\mathcal{S}$, actions $\mathcal{A}$, state transitions p, initial state distribution p_0 and reward function r

$$\begin{aligned} s_1 &\sim p_0 & \\ a_t &\sim \pi(s_t) & \text{for } t = 1,\dots,H \\ s_{t+1} &\sim p(\cdot|s_t,a_t) & \text{for } t = 1,\dots,H \\ r_t &= r(s_t,a_t) & \text{for } t = 1,\dots,H \end{aligned}$$

Agent interacts with MDP in episodes of H time-steps to produce a policy π that maximizes the total expected return per episode

$$\mathbb{E}\left[\sum_{t=1}^{H} r_t\right] = \mathbb{E}[V^\pi_{1:H}(s_1)]$$

Example: Call Center Help Support
- Each call is an episode
- Bounded number of interactions per call (fixed horizon)
- Goal: try to best help caller, i.e., maximize caller satisfaction = maximize total return per episode

Our UCFH Algorithm

Model-based algorithm UCFH (Upper Confidence Fixed-Horizon episodic RL) with optimism under uncertainty inspired by UCRL-γ [Lattimore & Hutter 2012] and others

Algorithm 1: `UCFH`: **U**pper-**C**onfidence **F**ixed-**H**orizon episodic reinforcement learning algorithm

Input: desired accuracy $\epsilon \in (0,1]$, failure tolerance $\delta \in (0,1]$, fixed-horizon MDP M
Result: with probability at least $1-\delta$: ϵ-optimal policy

$k := 1, \quad w_{\min} := \frac{\epsilon}{4H|\mathcal{S}|}, \quad \delta_1 := \frac{\delta}{2U_{\max}C}, \quad U_{\max} := |\mathcal{S}\times\mathcal{A}|\log_2\frac{|\mathcal{S}|H}{w_{\min}}$;
$m := 512(\log_2\log_2 H)^2\frac{CH^2}{\epsilon^2}\log^2\left(\frac{8H^2|\mathcal{S}|^2}{\epsilon}\right)\ln\frac{6|\mathcal{S}\times\mathcal{A}|C\log_2^2(4|\mathcal{S}|^2H^2/\epsilon)}{\delta}$;
$n(s,a) = v(s,a) = n(s,a,s') := 0 \quad \forall, s\in\mathcal{S}, a\in\mathcal{A}, s'\in\mathcal{S}(s,a)$;
while do
  /* Optimistic planning */
  $\hat{p}(s'|s,a) := n(s,a,s')/n(s,a)$, for all (s,a) with $n(s,a) > 0$ and $s'\in\mathcal{S}(s,a)$;
  $\mathcal{M}_k := \{\tilde{M}\in\mathcal{M}_{\text{nonst.}} : \forall(s,a)\in\mathcal{S}\times\mathcal{A}, t = 1\dots H, s'\in\mathcal{S}(s,a)$
    $\tilde{p}_t(s'|s,a) \in \texttt{ConfidenceSet}(\hat{p}(s'|s,a), n(s,a))\}$;
  $\tilde{M}_k, \pi^k := \texttt{FixedHorizonEVI}(\mathcal{M}_k)$;
  /* Execute policy */
  repeat
    `SampleEpisode`(π^k) ; // from M using π^k
  until *there is a* $(s,a)\in\mathcal{S}\times\mathcal{A}$ *with* $v(s,a) \ge \max\{mw_{\min}, n(s,a)\}$ *and* $n(s,a) < |\mathcal{S}|mH$;
  /* Update model statistics for one (s,a)-pair with condition above */
  $n(s,a) := n(s,a) + v(s,a)$;
  $n(s,a,s') := n(s,a,s') + v(s,a,s') \quad \forall s'\in\mathcal{S}(s,a)$;
  $v(s,a) := v(s,a,s') := 0 \quad \forall s'\in\mathcal{S}(s,a); k := k+1$

Procedure `SampleEpisode`(π)
  $s_0 \sim p_0$;
  for $t = 0$ **to** $H-1$ **do**
    $a_t := \pi_{t+1}(s_t)$ and $s_{t+1} \sim p(\cdot|s_t,a_t)$;
    $v(s_t,a_t) := v(s_t,a_t) + 1$ and $v(s_t,a_t,s_{t+1}) := v(s_t,a_t,s_{t+1}) + 1$;

Function `ConfidenceSet`(p, n)

$$\mathcal{P} := \left\{p'\in[0,1] : \text{if } n>1 : |p'(1-p') - p(1-p)| \le \frac{2\ln(6/\delta_1)}{n-1}, \right. \tag{1}$$

$$\left. |p-p'| \le \min\left(\sqrt{\frac{\ln(6/\delta_1)}{2n}}, \sqrt{\frac{2p(1-p)}{n}\ln(6/\delta_1)} + \frac{2}{3n}\ln\frac{6}{\delta_1}\right)\right\} \tag{2}$$

  return $\mathcal{P}$

Comparison to Existing Results

Previous work mostly considers bounds on the suboptimal time-steps instead of episodes → less meaningful in episodic RL

Existing Episode Bounds:

Translation of UCRL2 regret bounds [Jaksch et al 2010]	$\tilde{O}\left(\frac{S^2AH^3}{\epsilon^2}\log\frac{1}{\delta}\right)$
Translation of episodic discounted infinite horizon return bound by Fiechter [1994] with $H = 1/(1-\gamma)$	$\tilde{O}\left(\frac{S^2AH^7}{\epsilon^2}\log\frac{1}{\delta}\right)$
Bound for acyclic MDPs with min. state reachability probabilities q by Reveliotis et. al. [2007]	$\tilde{O}\left(\frac{SAH^4}{\epsilon^2 q}\log\frac{1}{\delta}\right)$
Translation of regret bound for UCB-type algorithm by Auer & Ortner [2005]	$\tilde{O}\left(\frac{S^{10}AH^7}{\epsilon^3}\log\frac{1}{\delta}\right)$

There are no prior lower bounds for this setting.

Upper PAC Bound

Theorem: Upper Bound on Sample Complexity of Episodic RL

With probability at least $1-\delta$, our algorithm follows a policy which has expected return per episode at most ϵ worse than optimal in all but

$$\tilde{O}\left(\frac{CSAH^2}{\epsilon^2}\log\frac{1}{\delta}\right) \le \tilde{O}\left(\frac{S^2AH^2}{\epsilon^2}\log\frac{1}{\delta}\right)$$

episodes where each state has at most $C \le S$ successor states

1. Variance-Sensitive Concentration Inequalities (Bernstein)

Similar to recent tight analyses of other settings, our analysis builds on Bernstein's concentration inequality. Tighter bounds can be achieved by bounding the next state value variances of an episode

$$\sigma^2_{t:H}(s) = \mathrm{Var}\left[V^\pi_{t+1:H}(s_{t+1})|s_t = s\right]$$

2. Non-Trivial Variance Bound for State Values

We show that the variance of the value function

$$\mathcal{V}_{t:H}(s) = \mathbb{E}\left[\left(\sum_{i=t}^{H} r_i - V^\pi_{t:H}(s_t)\right)^2 \middle| s_t = s\right]$$

satisfies a Bellman-style equation

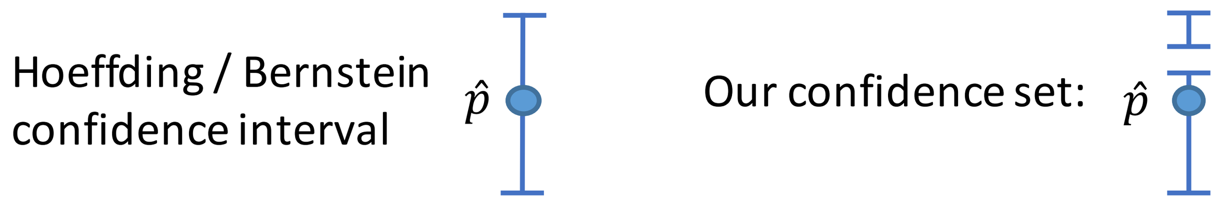
$$\mathcal{V}_{t:H}(s) = \sigma^2_{t:H}(s) + \mathbb{E}\left[\mathcal{V}_{t+1:H}(s_{t+1})|s_t = s\right]$$

which let us bound

$$\sum_{i=t}^{H}\mathbb{E}\left[\sigma^2_{i:H}(s_i)|s_t = s\right] \le H^2 \leftarrow \boxed{\text{better than trivial } H^3!}$$

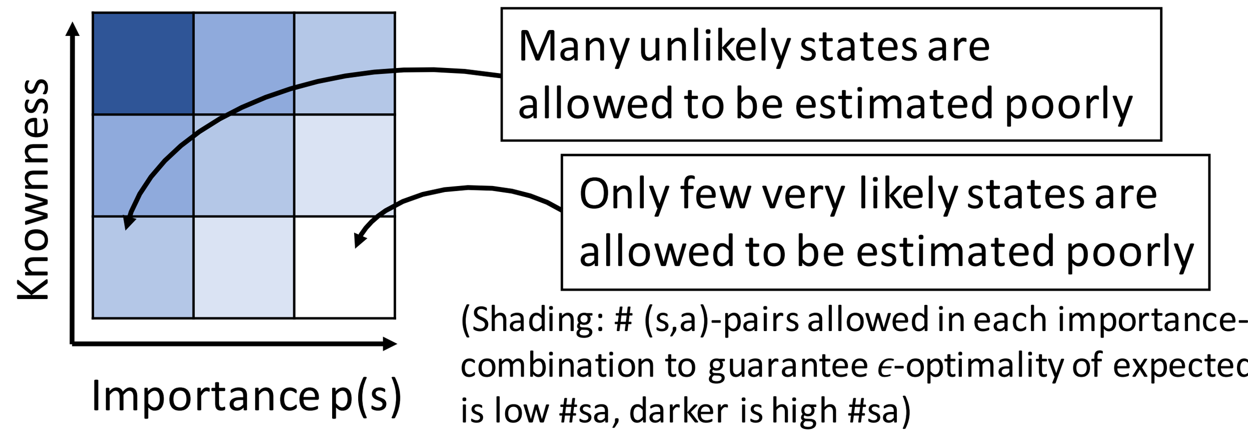
3. Specific Confidence Sets for Tight Bounds in non-sparse MDPs

We use specific confidence sets for transition probabilities instead of simple Hoeffding / Bernstein-based intervals. Can lead to disconnected confidence sets:



4. Fine-Grained Categorization of State-Action-Pairs

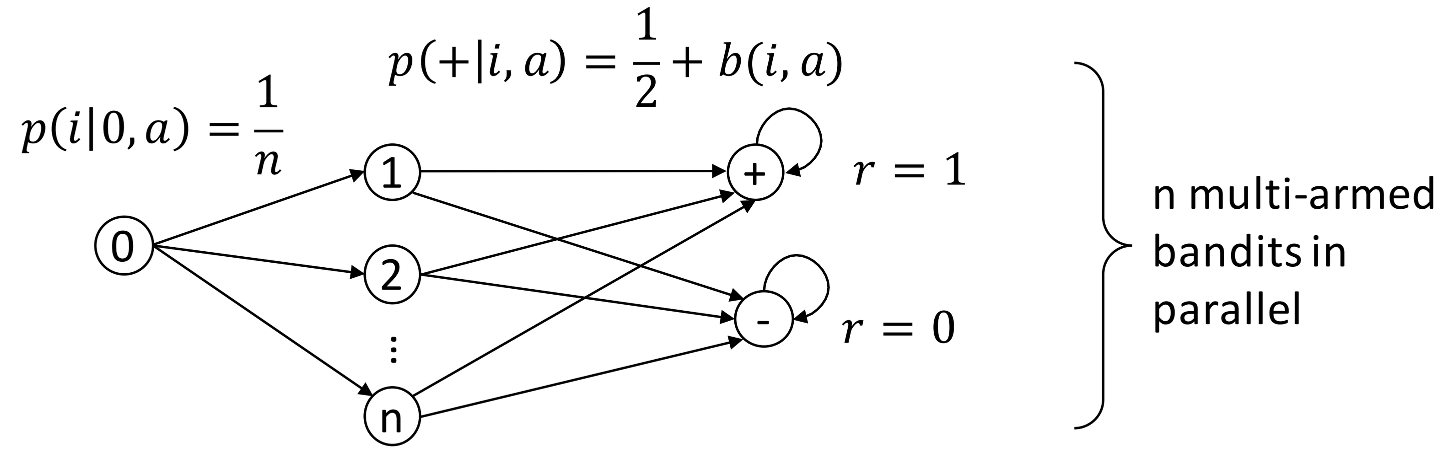
Similar to Lattimore & Hutter [2012] we distinguish not only between *known* and *unknown* (s,a)-pairs but distinguish between several levels of *knownness* (= confidence about transition probability estimates) and *importance* (=probability to encounter a state under current policy).



(Shading: # (s,a)-pairs allowed in each importance-knownness combination to guarantee ϵ-optimality of expected return, lighter is low #sa, darker is high #sa)

Lower PAC Bound

Difficult episodic Markov Decision Process:



- Transition to one of $n = O(S)$ different A-armed bandits
- Pulled arm a determines bias $b(i,a)$ of bandit i
- Coin flip with bias $b(i,a)$ decides whether agent transitions to good state (+) with total return $O(H)$ or bad state (–) with total return 0

Main Steps of Analysis:

1. ϵ-optimal expected return per episode ≙ solve at least a fraction of the multi-armed bandits
2. Best strategy for agent: try to solve each bandit with the same effort / confidence
3. Slightly biased multi-armed bandits hard to learn [Mannor & Tsitsiklis 2005]

Theorem: Lower Bound on Sample Complexity of Episodic RL

For any algorithm which outputs a deterministic policy with a PAC guarantee for precision $\epsilon \le \epsilon_0$ and failure probability $\delta \le \delta_0$, there is an episodic fixed-horizon MDP so that the algorithm requires at least

$$\Omega\left(\frac{SAH^2}{\epsilon^2}\log\frac{1}{\delta+c}\right)$$

episodes.

References & Acknowledgements

We thank Tor Lattimore for the helpful discussion. This work has been supported by an NSF CAREER award and the ONR Young Investigator program.

- C.-N. Fiechter. Efficient reinforcement learning (COLT 1994)
- S. Mannor & J. Tsitsiklis. The Sample Complexity of Exploration in the Multi-Armed Bandit Problem (JMLR 2004)
- P. Auer & R. Ortner. Online Regret Bounds for a New Reinforcement Learning Algorithm. (1st Austrian Cognitive Vision Workshop 2005)
- S. Reveliotis & T. Bountourelis. Efficient PAC learning for episodic tasks with acyclic state spaces. (Discrete Event Dynamic Systems: Theory and Applications 2007)
- T. Jaksch, R. Ortner & P. Auer. Near-optimal Regret Bounds for Reinforcement Learning (JMLR 2010)
- T. Lattimore & M. Hutter. PAC Bounds for discounted MDPs (ALT 2012)