IEOR 8100-001: Learning and optimization for sequential decision making

Instructor: Shipra Agrawal

Industrial Engineering and Operations Research
Columbia University

...

Reinforcement learning: problem definition

"Near optimal bounds regret bounds for reinforcement learning." Jaksch, Ortner, Auer, JMLR 2010.

At every time $t = 1, 2, \ldots,$

- Observe state $s_t \in S$
- Take action $a_t \in A$, observe reward $r_{a_t}(s_t)$.
- Observe transition from s_t to s_{t+1} with (unknown) probability $p_{a_t}(s_t, s_{t+1})$. Or, $s_{t+1} \sim p_{a_t}(s_t)$, where $p_{a_t}(s_t)$ is a distribution over states in S.

Unknown transition distribution $p_a(s)$ for every a, s. You only get to observe a sample from distribution for s_t, a_t at time t.

Assumptions

- Bounded rewards $r_a(s) \in [0, 1]$.
- MDP is *communicating*, i.e., finite diameter D.

$$D = \max_{s \neq s'} \min_{\pi : S \to A} \mathbb{E}[\tau(s, s', \pi)]$$

where $\tau(s, s', \pi)$ is the (random) number of time steps it takes to reach s' from s when using Markovian stationary policy π.

Problem Setting [Jaksch, Ortner, Auer 2010]

- Goal: Given starting state $s_1 = s$, maximize finite horizon reward

$$\sum_{t=1}^{T} r_{a_t}(s_t)$$

  Regret is defined with respect to the optimal expected infinite horizon average reward for the underlying MDP.
- (Theorem for communicating MDP) Optimal infinite horizon average reward is achieved by a Markovian stationary policy , say π^*, and

$$\forall s, \lim_{T\to\infty} \frac{1}{T}\mathbb{E}[\sum_{t=1}^{T} r_{\pi^*(s_t)}(s_t)|s_1 = s] = \rho^* \text{ (optimal gain)}$$

- $$\text{Regret } R(T) = T\rho^* - \sum_{t=1}^{T} r_{a_t}(s_t)$$

  UCB like approach to achieve high probability regret bounds.

Regret definition[Jacksh, Ortner, Auer 2010]

Given MDP $M = (S, A, \nu, p, s)$ with unknown $\{\nu_a(s)\}, \{p_a(s, s')\}$, finite diameter D, finite $|S|, |A|$. Also assume bounded rewards, i.e., suppport of distribution $\nu_a(s)$ is $[0, 1]$.

Regret of an algorithm

$$R(M, s, T) = T\rho^* - \mathbb{E}[\sum_{t=1}^{T} r_t | s_1 = s]$$

where ρ^* is the gain of the MDP M, and $r_t = r_{a_t}(s_t)$ is the reward for actions a_t taken by the algorithm in state s_t at time t.

What to expect?

- Intuitively, there are $|S||A|$ arms now, instead of $|A|$: we need to learn about every state and action pair.
- However, it is worse because algorithm cannot directly decide (**state**, action) to "play". It can only decide on which action to play in current state
- could get stuck on a bad state – that's why finite diameter assumption.
- we can hope to get out of bad states in D steps on average. So if earlier we needed to explore every of $|A|$ arms $\log(T)/gap$ times, we should need to explore every of $|S||A|$ arms, atleast $D\log(T)/gap$ times.
- Can we achieve regret of order $D|S||A|log(T)/gap$ or $\sqrt{D|S||A|T}$?

Main result of [Jacksch, Ortner, Auer 2010]

A UCB based algorithm UCRL2 which achieves regret bound

Theorem

With probability of at least $1-\delta$*, it holds that for any initial state* s *and any* $T > 1$*, the regret of UCRL2 is bounded by*

$$R(M, s, T) \leq 34 D |S| \sqrt{|A| \, T \log(T/\delta)}$$

There are other results in the paper on problem dependent regret bound.

Lower bound[Jacksch, Ortner, Auer 2010]

Theorem

For any algorithm, $S, A \geq 10$, $D \geq 20 \log_A(S)$ and $T \geq DSA$, there exists an MDP M with state action space of size S, A, diameter D, such that regret in time T

$$\mathbb{E}[R(M, s, T)] \geq 0.015\sqrt{D|S||A|T}$$

They show that the diameter is at least $\log_{|A|}(|S|) - 3$.

Algorithm outline

At any given time

- Use sample transitions observed so far to construct estimate $\hat{p}_a(s, s')$ of $p_a(s, s')$ for every s, a.
- Build high probability confidence intervals around these estimates: with high probability, actual transition probability lies in this confidence interval.
- Find most optimistic estimates in the confidence interval??
  - Define a HUGE set of plausible MDPs, corresponding to all possible transition probability values.
  - $\tilde{M}$ with highest gain $\tilde{\rho}$ among all plausible MDPs. Optimal policy $\tilde{\pi}$
- Work in episodes: Run the policy $\tilde{\pi}$ for some τ time steps.

Algorithm

Initialize $t = 1$, observe state s_1.
For episodes $k = 1, 2, \ldots$
Step 1: Initialize epsiode k

- Set start time of episode $\tau_k = t$.
- $n_{k-1}(s, a)$: number of times s, a was visited before τ_k.
- $P_{k-1}(s, a, s')$: number of transitions to s', when a was played in state s before τ_k
- $\hat{p}_{k-1}(s, a, s') = \frac{P_{k-1}(s,a,s')}{\max(1, n_{k-1}(s,a))}$

..algorithm continued

Step 2: Find optimistic MDP

- Plausible MDPs: $\mathcal{M}_k$ be the set of all MDPs $(S, A, r, \tilde{p})$ such that

$$\|\tilde{p}_a(s,\cdot) - \hat{p}_{k-1}(s,a,\cdot)\|_1 \leq \sqrt{\frac{14S\log(2At/\delta)}{\max\{1, n_{k-1}(s,a)\}}}$$

- Find (near) optimal policy for the best MDP in $\mathcal{M}_k$.

$$(\tilde{M}_k, \tilde{\pi}_k, \tilde{s}_k) = \arg \max_{\tilde{M}\in\mathcal{M}_k,\pi,s} \rho^{\pi}(s, \tilde{M})$$

Note: In the paper, above is only approximately solved – a near optimal policy is found.

Aside

Let $\tilde{\rho}_k(\tilde{s}_k)$ be the gain of the most optimistic MDP. Then,

Theorem

With probability $1 - \frac{\delta}{20t^6}$,

$$M \in \mathcal{M}_k$$

so that

$$\tilde{\rho}_k(\tilde{s}_k) \geq \rho^*.$$

In fact, we show later that $\tilde{\rho}_k(s) = \tilde{\rho}_k \geq \rho^*$ for all s.

..algorithm continued

Step 3: Execute the optimistic policy $\tilde{\pi}_k$
For $t = \tau_k, t+1, t+2, ...,$

- Observe s_t, play $a_t = \tilde{\pi}_k(s_t)$
- update $n_k(s_t, a_t)$
- Break if $n_k(s_t, a_t) \geq 2n_{k-1}(s_t, a_t)$.

Go to episode $k+1$

Finding most optimistic policy

by solving a single "large" MDP

Define "extended MDP" M^+, with states as original states S, but many more (continuous space of) actions A^+.
In particular, for every original a, s

- one new action for every plausible distribution $\tilde{p}_a(s)$
- on taking this "new"' action in state s, we see reward $r_a(s)$, but state transition follows $\tilde{p}_a(s)$ distribution.
- on taking this "new"' action in other states s', we see reward $r_a(s')$, but state transition follows $\hat{p}_a(s')$ distribution.

Some observations

Assume at beginning of episode $k+1$, for all a, s, $p_a(s)$ is one of the plausible distribution (which is true with high probability). Then,

- Finding optimal policy for extended M_k^+ is equivalent to finding $\tilde{\pi}_k$: optimal policy for most optimistic MDP in $\mathcal{M}_k$
- Extended MDP M_k^+ is also communicating with diameter D

How to solve extended MDP?

Value iteration algorithm

First, we present value iteration algorithm for solving a finite space finite action known MDP. This will then be extended to solve extended MDP.

The intuition for this algorithm comes from finding optimal policy for finite horizon reward.

Optimal policy for finite horizon

Let $J^n(s)$ denote the optimal finite horizon reward achievable in n steps, starting at state s. Then,

$$J^1(s) = \max_a r_a(s)$$

$$J^n(s) = \max_a r_a(s) + \sum_{s'} J^{n-1}(s') p_a(s, s')$$

(dynamic programming)

Value iteration for inifinte horizon average reward

1. Initialize $J(s) = 0$, for all s, $n = 0$.
2. For $n = 1, 2, ...$
   - $$J^n(s) = \max_a r_a(s) + \sum_{s'} p_a(s, s') J^{n-1}(s')$$
   - If $\max_s\{J^n(s) - J^{n-1}(s)\} - \min_s\{J^n(s) - J^{n-1}(s)\} < \epsilon$, go to Step 3, otherwise continue.
3. Output policy:

$$\pi_\epsilon(s) \in \arg\max_a r_a(s) + \sum_{s'} p_a(s, s') J^{n-1}(s')$$

Convergence of value iteration

Theorem (Puterman 1994)

For aperiodic communicating MDPs

$$\lim_{n\to\infty} \max_s \{J^{n+1}(s) - J^n(s)\} - \min_s \{J^{n+1}(s) - J^n(s)\} \to 0$$

In fact,

$$\lim_{n\to\infty} J^n(s_1) - J^n(s_2) = V^*(s_1) - V^*(s_2)$$

Theorem (Puterman 1994)

For any $\epsilon > 0$, aperiodic communicating MDPs value iteration terminates in finite steps, and the policy π_ϵ, is ϵ-optimal policy, in that for all s,

$$\rho(\pi_\epsilon, s) \geq \rho^* - \epsilon$$

Value iteration for solving M^+

Extended MDP M_k^+ is also communicating with diameter $D \Rightarrow$ value iteration will converge

1. Initialize $J(s) = 0$, for all s, $n = 0$.
2. For $n = 1, 2, ...$
   - $$J^n(s) = \max_a \max_{\text{plausible } \tilde{p}_a(s)} r_a(s) + \sum_{s'} \tilde{p}_a(s, s') J^{n-1}(s')$$
   - If $\max_s\{J^n(s) - J^{n-1}(s)\} - \min_s\{J^n(s) - J^{n-1}(s)\} < \epsilon$, go to Step 3, otherwise continue.
3. Output policy:

$$\tilde{\pi}(s) \in \arg\max_a \max_{\text{plausible} \tilde{p}_a(s)} r_a(s) + \sum_{s'} \tilde{p}_a(s, s') J^{n-1}(s')$$