IEOR 8100-001: Learning and Optimization for Sequential Decision Making 04/06/16

Lecture 21: Learning and optimization for sequential decision making

Instructor: Shipra Agrawal Scribed by: Ji Xu

1 Problem definition:

- At every time $t = 1, 2, \cdots$:

  1 Observe state $s_t \in S$

  2 Take action $a_t \in A$, observe reward $r_{a_t}(s_t)$.

  3 Observe transition s_{t+1} from s_t. s_{t+1} follows (unknown) probability distribution $p_{a_t}(s_t)$ over states in S.

Assumptions:

1 Bounded rewards: $r_a(s) \in [0, 1]$ (not a random variable)

2 MDP is communicating, i.e, there exists finite diameter D defined as

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

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

Regret definition

Given starting state $s_1 = s$ and any finite time period T, our natural goal is to maximize the finite horizon reward

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

Then Regret is defined as the following:

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

where ρ^* is considered as optimal infinite horizon average reward. It is achieved by a Markovian stationary policy π^* and ρ^* is defined as the following:

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

Notice that ρ^* doesn't depend on the initial state s_1. This is because we can always spend finite expected steps to any state required which is guaranteed by the Assumption 2 that MDP is communicating, and finite steps of different rewards won't effect the limit of the average reward. For any given algorithm, we define the Regret of the algorithm as the following:

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

where $M = (S, A, r, p, s)$ with finite $|S|$, finite $|A|$ and Assumption 1 and 2 hold.

Heuristic argument:

Comparing to the model we have learned, there are now $|A||S|$ arms instead of $|A|$ to represent any pair of state s and action a. Moreover, we can only determine a(action) instead of directly choosing an arm (s, a) to play. A direct bad consequences is that even if we are using the optimal policy π^* at time t, there is still a chance s_{t+1} is bad. Therefore, we need communicating assumption in Assumption 2, otherwise we could get stuck on a bad state forever. With finite diameter D assumption, if we stuck in bad state, instead of immediately switching to good state in one step, it will cost us D steps to get out of bad state. Therefore, if earlier we needed to explore every of $|A|$ arms $\log(T)/gap$ times, we should need to explore every of $|S||A|$ arms at least $D\log(T)/gap$ times.

2 Main results:

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

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

Theorem 2. *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*

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

Algorithm

Similar as UCB algorithm, at any given time, we will use sample transitions observed so far to construct estimate $\hat{p}_a(s, s')$ of $p_a(s, s')$ for every pair of s, a. Then we build high probability confidence intervals around these estimates. Then we find a MDP $\tilde{M}$ with highest gain $\tilde{\rho}$ among all plausible MDPs whose transition probability lies in these confidence intervals. Then find the optimal policy $\tilde{\pi}$ based on $\tilde{M}$. Since even if $\tilde{\pi}$ is in fact the best policy π^*, the next reward might still be bad, we run the policy $\tilde{\pi}$ for some τ time steps instead of updating the policy in each step. We call the time period of using the same policy as one episode. The algorithm is designed as following:
Initialize $t = 1$, observe state s_1.
For episodes $k = 1, 2, \ldots$
Step 1: Initialize episode 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.
- $\hat{p}_{k-1}(s, a, s') = \frac{P_{k-1}(s,a,s')}{\max\{1, n_{k-1}(s,a)\}}$.

Step 2: Find optimistic MDP

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

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

- Find (near) optimal policy for the best MDP in M_k,

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

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

- 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.

3 Interpretation

Step 1 is initializing the algorithm and is quite straightforward. Step 3 executes the candidate policy $\tilde{\pi}_k$ over the steps of episode k, and it controls the length of the current episode k – the condition in the last line is to ensure that current episode is long enough to get good estimate of gain of the candidate policy, but also short enough so that there aren't very long gaps between updates of candidate policy. We will now mainly explain the intuition for Step 2.
For Step 2, a simple heuristic argument for the derivation of Eq.(1) is the following:
Using Hoefding Inequality, we have

$$\mathbf{Pr}(|\hat{p}_{k-1}(s,a,s') - p(s,a,s')| \geq \epsilon) \leq 2e^{-2n\epsilon^2},$$

where from step 1, we know $n = \max\{1, n_{k-1}(s,a)\}$.
By choosing $\epsilon = \sqrt{\frac{1}{2\max\{1, n_{k-1}(s,a)\}} \log(\frac{|A||S|^2 t}{\delta})}$, we have

$$\sum_{s,a,s'} \mathbf{Pr}(|\hat{p}_{k-1}(s,a,s') - p(s,a,s')| \geq \epsilon) \leq 2 \sum_{s,a,s'} \frac{\delta}{|A||S|^2 t} = \frac{2\delta}{t}.$$

Using this ϵ, with at least $1 - \frac{2\delta}{t}$ probability, we have

$$\|p_a(s,\cdot) - \hat{p}_{k-1}(s,a,\cdot)\|_1 \leq |S|\epsilon \leq |S|\sqrt{\frac{\log(|A||S|^2 t/\delta)}{2\max\{1, n_{k-1}(s,a)\}}}, \tag{2}$$

having extra factor $\sqrt{|S|}$ comparing to Eq.(1). This is because we are not using the fact that both $p_a(s,\cdot)$ and $\hat{p}_{k-1}(s,a,\cdot)$ are probability measures that sum up to 1. If we are using this fact, from the Theorem 2.1 in [2], we have

$$\mathbf{Pr}(\|\hat{p}_{k-1}(s,a,\cdot) - p_a(s,\cdot)\|_1 \geq \epsilon) \leq (2^{|S|} - 2)e^{-\frac{n\epsilon^2}{2}}.$$

By choosing $\epsilon = \sqrt{\frac{14|S|\log(2|A|t/\delta)}{\max\{1,n_{k-1}(s,a)\}}}$, we have

$$\sum_{s,a} \mathbf{Pr}(\|\hat{p}_{k-1}(s,a,\cdot) - p_a(s,\cdot)\|_1 \geq \epsilon) \leq |S||A|(2^{|S|}-2)(\frac{\delta}{2|A|t})^{7|S|} < \frac{\delta}{t}.$$

Therefore with probability at least $1 - \frac{\delta}{t}$, we have

$$\|\hat{p}_{k-1}(s,a,\cdot) - p_a(s,\cdot)\|_1 \leq \epsilon = \sqrt{\frac{14|S|\log(2|A|t/\delta)}{\max\{1,n_{k-1}(s,a)\}}}, \forall a, s.$$

In class, we define M_k based on all plausible $\tilde{p}_a(s)$ since the reward are deterministic. Yet in [1], the author consider the reward $r_a(s)$ is a random variable. Therefore they used estimates $\hat{r}_a(s)$ for the reward and construct a confidence interval with high probability. In addition, they define M_k not only based on all plausible $\tilde{p}_a(s)$, but also on all plausible $\tilde{r}_a(s)$ which lie in the confidence interval for $\hat{r}_a(s)$.
For finding the optimal policy, the second part of step 2, we first define a single "large" MDP and then find most optimistic policy by solving this "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 $\tilde{p}_a(s)$ in the confidence interval.
- 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.

As we showed in the first part of Step 2, at the beginning of episode $k+1$, for $\forall a, s$, $p_a(s)$ is one of the plausible distribution lies in the confidence interval with high probability. Moreover, we have

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

Before we give an algorithm to solve the best policy for M^+, we present value iteration algorithm for solving a finite state and finite action known MDP.
Let $J^n(s)$ denote the optimal finite horizon reward achievable in n steps, starting at state s. Then,

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

Using the idea of dynamic programming, we can induct that

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

Now letting steps n goes to infinity, we have

$$J^n(s) - J^{n-1}(s) \to \rho^* + V^*(s), \forall s.$$

More rigorously, we have the following theorem:

Theorem 3. *(Puterman 1994)*
For aperiodic communicating MDPs

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

In fact,

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

Therefore, using this fact,we have the following algorithm for infinite horizon average reward:

1 Initialize $J(s) = 0, \forall s, n = 0$.

2 For $n = 1, 2, \ldots$

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

If $\max_s\{J^{n+1}(s) - J^n(s)\} - \min_s\{J^{n+1}(s) - J^n(s)\} < \epsilon$, goto step 3, otherwise continue.

3 Output policy π_ϵ:

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

Though $\pi_\epsilon(s)$ is not exactly the optimal policy, but the following theorem guarantees that it is almost the best policy:

Theorem 4. *(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,

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

Now we move to our own problem of solving optimal policy for M^+ where the difference is that each action a in state s is replaced by new actions corresponding to all plausible $\tilde{p}_a(s)$, and for each of these actions, $p_a(s)$ is replaced by $\tilde{p}_a(s)$. The algorithm will be the following:

1 Initialize $J(s) = 0, \forall s, n = 0$.

2 For $n = 1, 2, \ldots$

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

If $\max_s\{J^{n+1}(s) - J^n(s)\} - \min_s\{J^{n+1}(s) - J^n(s)\} < \epsilon$, goto step 3, otherwise continue.

3 Output policy $\tilde{\pi}_\epsilon$:

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

As a final summary of Step 2, let $\tilde{\rho}_k(\tilde{s}_k)$ be the gain of the most optimistic MDP, when starting state is $\tilde{s}_k$. Then,

Theorem 5. *With probability* $1 - \frac{\delta}{20t^6}$,

$$M \in M_k$$

such that

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

In fact, $\tilde{\rho}_k(s) = \tilde{\rho}_k \geq \rho^*$ for all s

Proof. First claim ($\tilde{\rho}_k(\tilde{s}_k) \geq \rho^*$) follows from the construction, which ensures that the original MDP is one of the plausible MDPs, with the given probability. The second claim ($\tilde{\rho}_k(s) = \tilde{\rho}_k$) follows directly from the observation that $\tilde{\rho}_k(\tilde{s}_k)$ is equal to the gain of corresponding extended MDP, which is a communicating MDP. □

And using the algorithm provided above, we can find an almost optimal policy $\tilde{\pi}_\epsilon(s)$.

References

[1] ***Near-optimal Regret Bounds for Reinforcement Learning*** by Thomas Jaksch, Ronald Ortner, Peter Auer, 2010

[2] ***Inequalities for the L_1 Deviation of the Empirical Distribution*** by Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, Marcelo J. Weinberger, 2003