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.

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.

You may want to think of every state-action pair as an arm, but one big difference is you cannot choose arbitrarily any state-action pair that you want to "pull", you can only choose actions in the state you are in at time t. state is decided by the transition process based on what you did earlierIn fact this implicit choice of state decided by your previous actions is the most difficult part of this problem. Therefore the simple setting with fixed reward is as difficult and will highlight the problems introduced due to unknown transition matrix

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:

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

Regret Analysis of UCRL2

Recap of Algorithm UCRL2

Some definitions

- start time of episode k, τ_k.
- $n_{k-1}(s,a)$: number of times s,a was visited before τ_k, i.e. in episodes $1,\ldots,k-1$.
- $P_{k-1}(s,a,s')$: number of transitions to s', when a was played in state s before τ_k
- Empirical estimate $\hat{p}_{k-1}(s,a,s') = \frac{P_{k-1}(s,a,s')}{\max(1,n_{k-1}(s,a))}$

Recap of Algorithm

Observe state s_1.
For episodes $k = 1, 2, \ldots$
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 $\mathcal{M}_k$.

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

by solving **extended MDP** M^+ with one new action a' for every plausible $(s, a, \tilde{p}_a(s))$.

..algorithm continued

Execute the optimistic policy $\tilde{\pi}_k$

For $t = \tau_k, t+1, t+2, ...,$

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

Go to episode $k+1$

Regret Analysis

Regret of an algorithm in time T

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

- ρ^* is the gain (infinite horizon average expected reward) of actual MDP $M = (S, A, r, p)$.
- s_t, a_t is state reached and action picked by algorithm at time t. s_1 is the starting state.
- Note that actions a_t are picked by algorithm, but state transitions from s_t to s_{t+1} happen by *actual unknown* probability distribution $p_{a_t}(s_t)$.

Regret in an episode: Fix an episode k, we first bound the regret in episode k.

$$R_k = \sum_{t \in \text{episode } k} (\rho^* - r_{a_t}(s_t))$$

Algorithm's workings in an episode

Algorithm uses same policy $\tilde{\pi}$ at all time steps in an episode

- $\tilde{\pi}$ is the optimal policy for best MDP $\tilde{M} = (S, A, r, \tilde{p}, \tilde{s}_0)$ in $\mathcal{M}_{k-1}$. For all t in episode k

$$a_t = \tilde{\pi}(s_t)$$

- equivalently, it is optimal policy π^+ for M^+, the extended MDP formulation of $\mathcal{M}_{k-1}$.

$$(a_t, \tilde{p}_{a_t}(s_t)) = a'_t = \pi^+(s_t)$$

- with probability $1 - \delta$, for all t,

$$\|\tilde{p}_{a_t}(s_t) - p_a(s)\|_1 \leq O(\sqrt{\frac{S \log(At/\delta)}{\max\{1, n_{k-1}(s_t, a_t)\}}})$$

note that actual state transitions happen with distribution p, not $\tilde{p}$.

- Let $\tilde{\rho}$ be gain of MDP $\tilde{M}$ (and M^+), then with probability $1 - \delta$ (i.e., if above event is true),

$$\tilde{\rho} \geq \rho^*$$

Regret in an episode

Assuming for all states, actions, time steps, actual transition distributions lie in corresponding confidence intervals (happens with probability $1 - \delta$),

$$\sum_{t \in \text{episode } k} \rho^* - r_{a_t}(s_t) \quad \leq \quad \sum_{t \in \text{episode } k} \tilde{\rho} - r_{a_t}(s_t)$$

Two reasons for gap

1. average reward at limit will different from initial average reward (bias)
2. $\tilde{p}$ vs. p.

First consideration captured by optimality equations:

- for any communicating MDP $M = (S, A, r, p)$, optimal gain ρ^*, optimal policy π^*, for all s,

$$\rho^* = r_{\pi^*(s)}(s) - V^*(s) + p_{\pi^*(s)}(s)^T V^*$$

  where V^* is the bias of optimal policy π^*.
- Applying to communicating MDP M^+, for all s,

$$\tilde{\rho} = r_{\tilde{\pi}(s)}(s) - \tilde{V}(s) + \tilde{p}_{\tilde{\pi}(s)}(s)^T \tilde{V}$$

- giving for any time t, in the episode

$$\tilde{\rho} = r_{a_t}(s_t) - \tilde{V}(s_t) + \tilde{p}_{a_t}(s_t)^T \tilde{V}$$

- **Normalization:** we can replace $\tilde{V}(s)$ by $h_s = \tilde{V}(s) - \min_{s'} \tilde{V}(s')$, Then, $h_s \geq 0$, $\|h\|_\infty \leq \max_{s_1,s_2} \tilde{V}(s_1) - \tilde{V}(s_2)$

$$\tilde{\rho} = r_{a_t}(s_t) - h_{s_t} + \tilde{p}_{a_t}(s_t)^T h$$

Substituting back,

$$\begin{aligned} R_k &= \sum_{t \in \text{episode } k} \rho^* - r_{a_t}(s_t) \\ &\leq \sum_{t \in \text{episode } k} \tilde{\rho} - r_{a_t}(s_t) \\ &= \sum_{t \in \text{episode } k} \tilde{p}_{a_t}(s_t)^T h - h(s_t) \end{aligned}$$

$$\leq \sum_{t \in \text{episode } k} -h(s_t) + p_{a_t}(s_t)^T h + (\tilde{p}_{a_t}(s_t) - p_{a_t}(s_t))^T h$$

$$\leq \sum_{t \in \text{episode } k} (-h(s_t) + p_{a_t}(s_t)^T h) + \|p_{a_t}(s_t)^T - \tilde{p}_{a_t}(s_t)\|_1 \|h\|_\infty$$

- Intuitively, first term is due to the first reason : bias. And, second term is because we used $\tilde{p}$ instead of actual p.
- For small bias term, we want lengthy episodes, but for small estimation term we want episodes to end quickly so that estimation can improve.

First term

- $\mathbb{E}[h(s_{t+1})|s_t, h, a_t] = p_{a_t}(s_t)^T h$.
- expected value of first term is $-h(s_1) + p_{a_t}(s_t)^T h$, where t is last step of the epsiode k.
- Absolute value of expectation at most $\|h\|_\infty$
- Azuma-Hoeffding bounds deviation of this term from its expectation by $O(\|h\|_\infty \sqrt{T_k \ln(1/\delta)})$, with probability $1 - \delta$
- T_k is length of episode k.
- we need to bound $\|h\|_\infty$, we show it is bounded by D.

Second term

Let $L = \sqrt{\log(AT/\delta)}$, $\bar{n}_{k-1}(s_t, a_t) = \max\{1, n_{k-1}(s_t, a_t)\}$.

$$\|p_{a_t}(s_t)^T - \tilde{p}_{a_t}(s_t)\|_1 \leq O(L\sqrt{\frac{S}{\bar{n}_{k-1}(s_t, a_t)}})$$

$$\begin{aligned} \text{Second term} &\leq L\|h\|_\infty \sum_{t \in \text{episode } k} \sqrt{\frac{S}{\bar{n}_{k-1}(s_t, a_t)}} \\ &= L\|h\|_\infty \sum_{s,a} \nu_k(s, a) \sqrt{\frac{S}{\bar{n}_{k-1}(s, a)}} \end{aligned}$$

where $\nu_k(s, a)$ be number of plays of action a in state s in this episode. The episode will break when $\nu_k(s, a) \geq n_{k-1}(s, a)$ for some s, a.

Regret in an episode

$$R_k \leq O(\|h\|_\infty \sqrt{T_k \ln(1/\delta)} + \|h\|_\infty) + L\|h\|_\infty \sum_{s,a} \nu_k(s,a) \sqrt{\frac{S}{\bar{n}_{k-1}(s,a)}}$$

We bound $\|h\|_\infty$ by D: the diameter of the communicating MDP M, and of M^+.

Bound on $\|h\|_\infty$

Recall $h(s)$ was defined as normalization of bias $\tilde{V}(s)$ of optimal policy $\tilde{\pi}$,

$$h(s) = \tilde{V}(s) - \min_s \tilde{V}(s)$$

Recall bias of policy $\tilde{\pi}$:

$$\tilde{V}(s) = \lim_{T \to \infty} \mathbb{E}[\sum_{t=1}^{T} (r_{a_t}(s_t) - \tilde{\rho})]$$

We show that

$$\|h\|_\infty = \max_{s_1 \neq s_2} \tilde{V}(s_1) - \tilde{V}(s_2) \leq D$$

$\max_{s_1 \neq s_2} \tilde{V}(s_1) - \tilde{V}(s_2)$ is sometimes called span of bias.

Bound on span of bias

for any optimal policy of communicating MDP with diameter D

- We have that value iteration algorithm converges for (aperiodic) communicating MDP, i.e. for all s,

$$J^{n+1}(s) := \max_a r_a(s) + \tilde{p}_a(s, s')J^n(s)$$

- we have $J^n(s)$ gets closer and closer to $n\tilde{\rho} + \tilde{V}(s)$, so that for all s_1, s_2

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

- For communicating MDP with diameter D, there exists a policy π such that τ, time to go from s_2 to s_1 is at most $\mathbb{E}[\tau] \leq D$.
- Now, from optimality of $J^n(s)$for n steps,

$$J^n(s_2) \geq \mathbb{E}[J^{n-\tau}(s_1)]$$

$$\begin{aligned} \lim_{n\to\infty} J^n(s_1) - J^n(s_2) &\leq \lim_{n\to\infty} J^n(s_1) - \mathbb{E}_\tau[J^{n-\tau}(s_1)] \\ &\leq \tilde{\rho}\mathbb{E}[\tau] \leq \tilde{\rho}D \leq D \end{aligned}$$

Combining per-episode regret into total regret

Let $\mathcal{E}$ be number of episodes. The way we defined episodes gives $\nu_k(s,a) \leq n_{k-1}(s,a)$, $n_k(s,a) = n_{k-1}(s,a) + \nu_k(s,a)$.

$$
\begin{aligned}
R(T) &= \sum_k R_k \\
&\leq O(\sum_k D\sqrt{T_k \ln(1/\delta)} + D) + \sum_k LD \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{\bar{n}_{k-1}(s,a)}}\sqrt{S} \\
(*) &\leq O(LD\sqrt{\mathcal{E}T}) + \mathcal{E}D + LD\sqrt{SAT}\sqrt{S}
\end{aligned}
$$

Remains to bound number of episodes $\mathcal{E}$.

Combining per-episode regret into total regret

Let $\mathcal{E}$ be number of episodes. The way we defined episodes gives $\nu_k(s,a) \le n_{k-1}(s,a)$, $n_k(s,a) = n_{k-1}(s,a) + \nu_k(s,a)$.

$$\begin{aligned} R(T) &= \sum_k R_k \\ &\le O(\sum_k D\sqrt{T_k \ln(1/\delta)} + D) + \sum_k LD \sum_{s,a} \frac{\nu_k(s,a)}{\sqrt{\bar{n}_{k-1}(s,a)}}\sqrt{S} \\ (*) &\le O(LD\sqrt{\mathcal{E}T}) + \mathcal{E}D + LD\sqrt{SAT}\sqrt{S} \end{aligned}$$

Remains to bound number of episodes $\mathcal{E}$.

Derivation of (*) in previous slide (Reference: Appendix C.3 of Jaksch, Ortner, Auer JMLR 2010)
Consider numbers $z_1, z_2, \ldots, z_m$, $z_1 + \ldots + z_m = Z$. **Important:** $z_k \le Z_{k-1} := \max\{1, z_1 + \cdots + z_{k-1}\}$
Then, we bound

$$\sum_{k=1}^{m} \frac{z_k}{\sqrt{Z_{m-1}}} \text{ by } (\sqrt{2}+1)\sqrt{Z_m} = (\sqrt{2}+1)\sqrt{Z}.$$

If $z_k = 1$ for all k, then clearly, above is $\sum_{k=1}^{T} \frac{1}{\sqrt{k}} \le 2\sqrt{Z}$. For general values of z, proof by induction, let

$$\sum_{k=1}^{i} \frac{z_k}{\sqrt{Z_{k-1}}} \le (\sqrt{2}+1)\sqrt{Z_i}$$

(check for any i such that $Z_k = 1$, $k = 1, \ldots, i-1$, use $z_i \le Z_{i-1} = 1$ in that case) and show

$$\sum_{k=1}^{i+1} \frac{z_k}{\sqrt{Z_{k-1}}} \le (\sqrt{2}+1)\sqrt{Z_i} + \frac{z_{i+1}}{\sqrt{Z_i}}$$

To show above, square the RHS, and use $z_{i+1} \le Z_i$s.
Now, $Z = \bar{n}_T(s,a)$, for every s, a, so that the sum is:

$$2\sum_{s,a} \sqrt{\bar{n}_{T-1}(s,a)}$$

where $\sum_{s,a} n_{T-1}(s,a) \le T$. The worst case is (sum is maximum) when $n_{T-1}(s,a)$ are all equal, so that

$$2\sum_{s,a} \sqrt{\bar{n}_{T-1}(s,a)} \le 2\sum_{s,a} \sqrt{\frac{T}{SA}} \le 2\sqrt{SAT}$$

Bound on number of episodes $\mathcal{E}$

Number of episodes is at most $SA \log_2(T)$:

- every epsiode doubles the number of plays of at least one s, a, so for every s, a at most $\log_2(T)$ episodes.

Slightly more careful analysis gives $SA \log_2(\frac{8T}{SA})$ bound on $\mathcal{E}$.

Finishing the regret bound

Substituting the bound of $SA\log_2(T)$on $\mathcal{E}$

$$\begin{aligned} R(T) &= \sum_k R_k \\ &\leq O(D\sqrt{\mathcal{E}T\ln(1/\delta)} + \mathcal{E}D) + LD\sqrt{SAT}\sqrt{S} \\ &= O(D\sqrt{SAT\ln(T)\ln(1/\delta)}) + D\sqrt{SAT\ln(TA/\delta)}\sqrt{S} \\ &= O(DS\sqrt{AT}\ln(TA/\delta)) \end{aligned}$$

The logarithmic term can be optimized further by more careful analysis of the first term (applying Azuma-Hoeffding once to sum of all T terms across episodes) to get

$$R(T) = O(DS\sqrt{AT\ln(TA/\delta)})$$

Other related results

Problem dependent bounds: Optimal Adaptive Policies for Markov Decision Processes, Apostolos N. Burnetas and Michael N. Katehakis, Mathematics of Operations Research, 1997

- Bounds Logarithmic in T, instead of $\sqrt{T}$,
- Recall in MAB: optimal regret order of

$$\sum_{i:KL(\nu^*,\nu_i)>0} \frac{\log(T)\Delta_i}{KL(\nu^*,\nu_i)}, \text{close} \sum_{i:\mu_i<\mu^*} \frac{\log(T)}{\Delta_i}$$

  where $\Delta_i = \mu^* - \mu_i$.
- What is Δ_i here?
- $\Delta_i = \mu^* - \mu_i$ is the amount by which μ_i needs to be increased so that i becomes uniquely optimal. Or, $KL(\nu^*, \nu_i)$ is how much distribution of i needs to be changed to make i optimal.

Substitute for Δ

For s, a, let $A^*(s)$ be all optimal actions for s.

$$a^*(s, p) = \arg\max_a r_a(s) + p_a(s)^T V^*(s)$$

Following could be a substitute for Δ: regret suffered if a is taken instead of a^*?

$$\Delta(s, a) = r_{a^*(s)}(s) + p_{a^*(s)}(s)^T V^*(s) - (r_a(s) + p_a(s)^T V^*(s))$$

How many mistakes will be made due to inaccurate transition probability p.

Critical state-action pairs

Define critical state-action pairs $B(p)$ as s, a such that a can become uniquely optimal action for s by changing only $p_a(s)$ in p.

$$\Theta(s,a) = \{q : \text{distribution over states}, A^*(s, p - p_a(s) + q) = a\}$$

Critical state action pairs are actions with $\Theta(s,a) \neq \phi$. And,

$$K(s,a) = \inf_{q \in \Theta(s,a)} KL(p_a(s), q)$$

MDP Regret lower bound

$$\lim_{T\to\infty} \frac{R^\pi(T)}{\log T} \geq \sum_{s,a \in B(p)} \frac{\Delta(s,a)}{K(s,a)}$$

for *uniformly fast convergent policies* π. Compare to MAB lower bound

$$\lim_{T\to\infty} \frac{R(T)}{\log T} \geq \sum_{\mu_i < \mu^*} \frac{\Delta_i}{KL(\nu^*, \nu_i)}$$

Problem dependent upper bound

"Optimistic Linear Programming gives Logarithmic Regret for Irreducible MDPs, Ambuj Tewari, Peter Bartlett, NIPS 2007"

- Uses L1 norm instead of KL-divergence to define

$$K(s,a) = \inf_{q \in \Theta(s,a)} \|p_a(s) - q\|_1^2$$

- Achieves **Regret upper bound** of form

$$R(T) = O(\log T \sum_{s,a \in B(p)} \frac{\Delta(s,a)}{K(s,a)})$$

- In every step, solves empirical MDP to find estimate of $\hat{V}(s), \forall s$, and then plays action

$$\arg\max_a \max_{q \in \text{confidence interval around } \hat{p}_a(s)} r_a(s_t) + q^T \hat{V}$$

Compare to $\sum_{i:\mu_i < \mu^*} \frac{\log(T)}{\Delta_i}$

Thompson Sampling for MDP

"(More) Efficient Reinforcement Learning via Posterior Sampling"
Ian Osband, Daniel Russo, Benjamin Van Roy, NIPS 2013
Main differences:

- Bayesian regret: known prior on MDPs
- Epsiodic set up: finite horizon MDP, runs for known deterministic τ time steps
- (Bayesian) Regret Bound: $O(\tau S\sqrt{AT\log(SAT)})$ (depends on τ instead of D)

Algorithm: At beginning of every episode

- Sample an MDP $\tilde{M}$ from posterior distribution
- Solve $\tilde{M}$ to find optimal policy $\tilde{\pi}$
- Use $\tilde{\pi}$ for all time steps in the episode

Open problems

- Immediate: Single algorithm that achieves optimal problem dependent and problem independent bounds (like UCB and TS did for MAB)
  - $\sqrt{S}$ gap in problem independent bounds
- Immediate: Thompson (aka Posterior) Sampling with unknown priors
  - bounding worst-case regret, communicating MDP
  - non-episodic set up
- Other: Continuous state space but structured, e.g., the inventory problem
- Other: Large state space, feature based approximation (compare to contextual bandits)