IEOR 8100-001: Learning and Optimization for Sequential Decision Making 04/11/16

Lecture 21: UCRL2 Algorithm

Instructor: Shipra Agrawal Scribed by: Tingjun Chen

1 Recap of UCRL2 Algorithm

First, some notations are as follows:

- start time of episode k is τ_k.
- $n_{k-1}(s,a)$ is the number of times (s,a) was visited before τ_k, i.e., in episodes $1,2,\cdots,k-1$.
- $P_{k-1}(s,a,s')$ is the number of transitions to state s', when action 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 the UCRL2 algorithm:

1. Observe state s_1.
2. For episode $k = 1,2,\cdots$:
   - Plausible MDPs: **Find optimistic MDP**: Let $\mathcal{M}_k$ 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 $(s,a,\tilde{p}_a(s))$.
   - **Execute the optimistic policy** $\tilde{\pi}_k$: for $t = \tau_k, t+1, t+2, \cdots$, (1) observe s_t and play $a_t = \tilde{\pi}_k(s_t)$, (2) increment $n_k(s_t,a_t)$ by 1, and (3) break the episode if $n_k(s_t,a_t) \geq 2n_{k-1}(s_t,a_t)$.
   - **Go to episode** $k+1$.

2 Regret Analysis

2.1 Definition

We define the regret of the algorithm in time T as

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

in which (1) ρ^* is the gain (infinite horizon average expected reward) of actual MDP $M = (S,A,r,p)$, (2) s_t, a_t is state reached and action picked by algorithm at time t with s_1 as 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 $p_{a_t}(s_t)$.

We first fix an episode k, and bound the regret in episode k (defined as **regret in an episode**):

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

2.2 Algorithm's workings in an episode

Fix an episode. The algorithm uses the same policy $\tilde{\pi}$ at all time steps in an episode, in which $\tilde{\pi}$ is the optimal policy for best MDP $\tilde{M} = (S, A, r, \tilde{p}, \tilde{s}_0)$ in $\mathcal{M}_{k-1}$, the collection of plausible MDPs in the previous episodes. For $\forall t$ in episode k, the algorithm picks action according to $a_t = \tilde{\pi}(s_t)$. Equivalently, this is also the 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 we have

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

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

2.3 Regret in an episode

Use the above results, assume that for all states, actions, time steps, actual transition distributions lie in corresponding confidence intervals (happens with probability $1 - \delta$), we have

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

Now we proceed to bound the regret in an episode. Note that for any communicating MDP $M = (S, A, r, p)$, optimal gain ρ^*, optimal policy π^*, for all s, it holds that

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

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,

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

Note that above equation is indifferent to constant additive shift in $\tilde{V}$ (this is because $\tilde{p}_a(s)$ sums to 1 for all a, s). So, we can perform the following **normalization**: for all s, replace $\tilde{V}(s)$ by $h_s = \tilde{V}(s) - \min_{s'} \tilde{V}(s')$. Then, $h_s \geq 0$ and $||h||_\infty \leq \max_{s_1, s_2} \tilde{V}(s_1) - \tilde{V}(s_2)$. After normalization, the above equation can be written as

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

Substituting back, we have

$$\begin{aligned}
R_k &= \sum_{t \in \text{episode } k} (\rho^* - r_{a_t}(s_t)) \\
&\leq \sum_{t \in \text{episode } k} (\tilde{\rho} - r_{a_t}(s_t)) \\
&= \sum_{t \in \text{episode } k} \left(\tilde{p}_{a_t}(s_t)^T h - h(s_t)\right) \\
&= \sum_{t \in \text{episode } k} \left(-h(s_t) + p_{a_t}(s_t)^T h + \left(\tilde{p}_{a_t}(s_t) - p_{a_t}(s_t)\right)^T h\right) \\
&\leq \left(\sum_{t \in \text{episode } k} -h(s_t) + p_{a_t}(s_t)^T h\right) + \left(\sum_{t \in \text{episode } k} ||p_{a_t}(s_t) - \tilde{p}_{a_t}(s_t)||_1 ||h||_\infty\right).
\end{aligned}$$

Intuitively, the first term is due to "bias" – it appears because there is gap from optimal gain even if the infinite horizon *optimal policy* is executed for a finite time. And, the second term term is due to estimation error – because we used $\tilde{p}$ instead of the actual p. There is also a trade-off: for small bias term, we want lengthy episodes, but for small estimation error we want episodes to end quickly so that we can update and improve our estimation.

2.3.1 Bounding the first term

We know that $\mathbb{E}[h(s_{t+1})|s_t, h, a_t] = p_{a_t}(s_t)^T h$ and the expected value of the first term is $-h(s_1) + p_{a_t}(s_t)^T h$, where t is the last step of the episode k. In addition, the absolute value of expectation is at most $||h||_\infty$. According to Azuma-Hoeffding bound we have this term deviates from its expectation by at most $O\left(||h||_\infty \sqrt{T_k \log(1/\delta)}\right)$ with probability $1 - \delta$ (T_K is the length of episode k). Further, (as discussed in the following) we can bound $||h||_\infty$ by a constant D.

2.3.2 Bounding the second term

Let $L = \sqrt{\log(AT/\delta)}$, and $\bar{n}_{k-1}(s_t, a_t) = \max\{1, n_{k-1}(s_t, a_t)\}$. Then, by construction, and confidence bounds, we have with probability $1 - \delta$,

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

Therefore the second term

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

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

$$\nu_k(s, a) \leq \bar{n}_{k-1}(s, a), \forall s, a$$

Hence, we have in total

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

And we bound $||h||_\infty$ by D, which is the diameter of the communicating MDP M, and of M^+.

2.3.3 Bounding $||h||_\infty$

Recall that $h(s)$ was defined as normalization of bias $\tilde{V}(s)$ of optimal policy $\tilde{\pi}$, i.e., $h(s) = \tilde{V}(s) - \min_s \tilde{V}(s)$, and the bias of policy $\tilde{\pi}$ is given by:

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

We will show that $||h||_\infty = \max_{s_1 \neq s_2} \tilde{V}(s_1) - \tilde{V}(s_2) \leq D$. Note that sometimes $||h||_\infty$ is called the *span of bias*.

In previous lecture, we saw that for any optimal policy of communicating MDP with diameter D, value-iteration algorithm converges for (aperiodic) communicating MDP, i.e., let for all s,

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

Then, for all s

$$J^{n+1}(s) - J^n(s) \to \tilde{\rho}$$

And, 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, the time to go from s_2 to s_1 is at most $\mathbb{E}[\tau] \leq D$. Therefore, for optimality of $J^n(s)$ for n steps, we have

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

and

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

Since $\max_{s_1,s_2} \tilde{V}(s_1) - \tilde{V}(s_2) = \|h\|_\infty$, we finish the proof of the bound of $||h||_\infty$.

2.4 Combining per-episode regret into total regret

Now we bound $R(T)$ based on the bounds of per-episode regret we got. Let $\mathcal{E}$ be the number of episodes. The way we defined episodes gives $\nu_k(s,a) \leq n_{k-1}(s,a)$, $n_k(s,a) = n_{k-1}(s,a) + \nu_k(s,a)$. Therefore,

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

where we used that $\sum_{k=1}^{|\mathcal{E}|} T_k = T$. The number of episodes $\mathcal{E}$ is bounded by $SA\log_2(T)$.

$$|\mathcal{E}| \leq SA\log_2(T) \tag{1}$$

This is because every episode doubles the number of plays of at least one s,a, so per s,a, there can be at most $\log_2(T)$ episodes. **Remark**: Slightly more careful analysis gives $SA\log_s\left(\frac{8T}{SA}\right)$ bound of the number of episodes, $\mathcal{E}$.

Also, following bound can be derived to bound the second term:

$$\sum_{s,a}\sum_k \frac{\nu_k(s,a)}{\sqrt{\bar{n}_{k-1}(s,a)}} \leq O(\sqrt{SAT}), \tag{2}$$

The complete derivation of (2) is provided in Appendix C.3 of Jaksch, Ortner, Auer JMLR 2010), an outline is provided in next section. Notably, this result does not uses any property of the algorithm, but is a worst-case upper bound on above expression under the constraint that for all k,s,a,

- $\nu_k(s,a) \leq \bar{n}_{k-1}(s,a)$,
- $n_k(s,a) = \nu_k(s,a) + n_{k-1}(s,a)$, $\bar{n}_k(s,a) = \max\{1, n_k(s,a)\}$, and that
- $\sum_{k,s,a} \nu_k(s,a) = T = \sum_{s,a} n_{|\mathcal{E}|}(s,a)$

Substituting these bounds, we obtain desired regret bound of $O(LS\sqrt{AT})$ on $R(T)$.

2.5 Discussion: on derivation of (2)

Consider numbers $z_1, z_2, \ldots, z_m$, $z_1 + \ldots + z_m = Z$. **Important:** $z_k \leq Z_{k-1} := \max\{1, z_1 + \cdots + z_{k-1}\}$ Then, we bound

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

If $z_k = 1$ for all k, then clearly, above is $\sum_{k=1}^{T} \frac{1}{\sqrt{k}} \leq 2\sqrt{Z}$. For general values of z_k, we can prove by induction. Let

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

(for base case, check for any i such that $Z_k = 1, k = 1, \ldots, i-1$, use $z_i \leq Z_{i-1} = 1$ in that case). Now, we need to show

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

We use

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

To show above, square the RHS, and use $z_{i+1} \leq Z_i$.

For any fixed s, a, using above result with $z_k = \nu_k(s,a), k = 1, \ldots, T$, $Z = \bar{n}_{|\mathcal{E}|}(s,a)$, we get

$$\sum_k \frac{\nu_k(s,a)}{\sqrt{\bar{n}_{k-1}(s,a)}} \leq (\sqrt{2}+1)\sqrt{\bar{n}_{\mathcal{E}}(s,a)}$$

so that the sum over all s, a is bounded as

$$\sum_{s,a}\sum_k \frac{\nu_k(s,a)}{\sqrt{\bar{n}_{k-1}(s,a)}} \leq (\sqrt{2}+1)\sum_{s,a}\sqrt{\bar{n}_{\mathcal{E}}(s,a)}$$

where $\sum_{s,a} n_{\mathcal{E}}(s,a) \leq T$. The worst case is (above sum is maximum) when $n_{\mathcal{E}}(s,a)$ are all equal (i.e., equal to $T/(SA)$), so that

$$(\sqrt{2}+1)\sum_{s,a}\sqrt{\bar{n}_{\mathcal{E}}(s,a)} \leq (\sqrt{2}+1)\sum_{s,a}\sqrt{\frac{T}{SA}} \leq (\sqrt{2}+1)\sqrt{SAT}$$