CS 4/5789 Lecture 21 Prof Sarah Dean

1) MBRL with Exploration

Let's consider a finite horizon tabular MDP:

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, H, s_0\}$$

where $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$

transition probability P unknown.

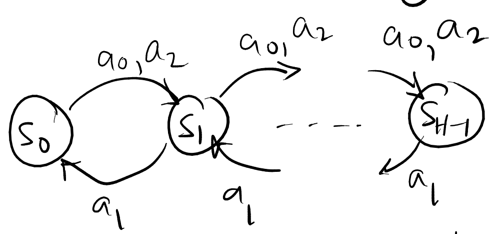
(for simplicity we assume reward is known)

This is different from the generative model that we studied in Lecture 10. We can't just pick a state s and action a and query

$$s' \sim P(s,a).$$

<u>Example</u>: Need for strategic exploration

chain:
(length $= H$)



$$r(s,a) = \begin{cases} 0 & s \neq s_{H-1} \\ 1 & s = s_{H-1} \end{cases}$$

The probability of a random walk hitting s_{H-1} starting from s_0 is $(1/3)^{-H}$

(Recall SARSA, Q-Learning, policy search require observed rewards to update!)

Naive idea: MDP as MAB:

Can we directly convert this MDP to a multi-armed bandit problem?

MAB: find the best of K actions.
MDP: find the best policy

Q: How many policies are there?

(Recall the finite contexts from Lecture 19)

This approach drops the shared information between rollouts from different policies. (E.g. transitions, rewards)

2) Upper-Confidence Bound Value Iteration

This is optimistic model-based learning

Alg: UCB-VI

initialize transition probability $\hat{P}_0$, reward bonus $b_0(s,a)$
for $i = 0, \ldots, T$
    optimistically plan: $\pi^i = VI(\hat{P}_i, r + b_i)$
    collect new trajectory with π^i
    update $\hat{P}_{i+1}$ and b_{i+1}

Model Estimation

Estimate $\hat{P}_i$ using Dataset $\{\{s_t^k, a_t^k\}_{t=0}^{H-1}\}_{k=0}^{i}$

Counts:

$$N_i(s,a) = \sum_{k=1}^{i-1} \sum_{t=0}^{H-1} \mathbb{1}\{s_t^k, a_t^k = s, a\}$$

# of times we take action a in state s.

$$N_i(s,a,s') = \sum_{k=1}^{i-1} \sum_{t=0}^{H-1} \mathbb{1}\{s_t^k, a_t^k, s_{t+1}^k = s, a, s'\}$$

# of times we transition to s' from s, a

Then $$\hat{P}_i(s'|s,a) = \frac{N_i(s,a,s')}{N_i(s,a)}$$

Reward Bonus

Encourage exploration of new state-action pairs

$$b_i(s,a) = H\sqrt{\frac{\alpha}{N_i(s,a)}}$$

Generate Policy:

In this case, VI reduces to Dynamic Programming

$\hat{V}_H^i(s) = 0.$

For $t = H-1, H-2, \ldots 0$:

$$\hat{Q}_t^i(s,a) = r(s,a) + b_i(s,a) + \mathbb{E}_{s' \sim \hat{P}(s,a)}\left[\hat{V}_{t+1}^i(s')\right]$$

$$\pi_t^i(s) = \operatorname{argmax} \hat{Q}_t^i(s,a)$$

$$\hat{V}_t^i(s) = \hat{Q}_t^i(s, \pi_t^i(s))$$

3) Analysis of UCB-VI

Two key facts about UCB-VI:

1) The exploration bonus bounds the difference $\left| \mathbb{E}_{s' \sim \hat{P}}[V(s')] - \mathbb{E}_{s' \sim P}[V(s')] \right|$ with high probability

(similar to confidence intervals $|\mu - \hat{\mu}|$ in MAB setting)

2) The exploration bonus yeilds <u>optimism</u>

$$\hat{V}_t^i(s) \geq V_t^*(s)$$

(similar to upper confidence bound in MAB setting)

These two facts are key in proving a <u>regret bound</u>, where we can define regret for this RL setting analogously to in the MAB setting: replace <u>reward</u> with <u>cumulative reward</u> (i.e. value)

$$R(T) = \mathbb{E}\left[\sum_{i=1}^{T} V_0^*(s_0) - V_0^{\pi^i}(s_0) \right]$$

The argument is very similar to the UCB proof.

1) use optimism: $V_0^*(s_0) - V_0^{\pi^i}(s_0) \leq \hat{V}_0^i(s_0) - V_0^{\pi^i}(s_0)$

2) Simulation Lemma to compare $\hat{V}_0^i(s_0)$ & $V_0^{\pi^i}(s_0)$.

Regret bound is out of scope for this class (you'd see in 6000 level)
But we will prove 2 key facts.

Lemma (Exploration Bonus): for any fixed function $V: \mathcal{S} \to [0, H]$, with high probability,

$$\left| \mathbb{E}_{s' \sim \hat{P}_i(s,a)}[V(s')] - \mathbb{E}_{s' \sim P(s,a)}[V(s')] \right| \leq H\sqrt{\frac{\alpha}{N_i(s,a)}} = b_i(s,a)$$

where α is dependent on S, A, H, and probability.

Proof:

$$\left| \mathbb{E}_{s' \sim \hat{P}_i(s,a)}[V(s')] - \mathbb{E}_{s' \sim P(s,a)}[V(s')] \right| = \left| \sum_{s' \in \mathcal{S}} [\hat{P}_i(s'|s,a) - P(s'|s,a)] V(s') \right|$$

$$\leq \sum_{s' \in \mathcal{S}} |\hat{P}_i(s'|s,a) - P(s'|s,a)| \, |V(s')|$$

(using result from Lecture 10, details out of scope) $$\leq \underbrace{\max_{s'} |V(s')|}_{\leq H \text{ since reward bounded}} \cdot \sqrt{\frac{\alpha}{N_i(s,a)}}$$

Lemma: (optimism) as long as $r(s,a) \in [0,1]$,

$$\hat{V}^i_t \geq V^*_t(s) \quad \forall n, i, s.$$

Proof: We show by induction. $\hat{V}^i_H = 0 = V^*_H$.

Suppose $\hat{V}^i_{t+1}(s) \geq V^*_{t+1}(s) \quad \forall s.$

Then: for any s, a:

$$\hat{Q}^i_t(s,a) - Q^*_t(s,a) = \cancel{r(s,a)} + b(s,a) + \mathbb{E}_{s' \sim \hat{P}(s,a)}\left[\hat{V}^i_{t+1}(s')\right] - \cancel{r(s,a)} - \mathbb{E}_{s' \sim P(s,a)}\left[V^*_{t+1}(s')\right]$$

(by inductive assumption) $$\geq b_i(s,a) + \mathbb{E}_{s' \sim \hat{P}_i(s,a)}\left[V^*_{t+1}(s')\right] - \mathbb{E}_{s' \sim P(s,a)}\left[V^*_{t+1}(s')\right]$$

(by bonus Lemma) $$\geq b_i(s,a) - b_i(s,a) = 0.$$

Therefore, $\hat{Q}^i_t(s,a) \geq Q^*_t(s,a) \quad \forall s, a.$

This implies that $\hat{V}^i_t(s) \geq V^*_t(s) \quad \forall s.$

(Exercise: argue why second to last line implies last line.) $\square$