1) MBRL with Exploration

Let's consider a finite horizon tabular MDP:

\[ M = \{ S, A, P, r, H, s_0 \} \]

where \( |S| = S \) and \( |A| = A \)

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

**Example:** Need for strategic exploration

\[ r(s,a) = \sum_{h=0}^{\infty} (1/3)^h \]

The probability of a random walk hitting \( S_{t+1} \)

starting from \( S_0 \) is \((1/3)^{-H}\).

(Recall SARS, 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: \( \Pi^i = \text{VI}(\hat{P}_i, r + b_i) \)

collect new trajectory with \( \Pi^i \)

update \( \hat{P}_{i+1} \) and \( b_{i+1} \)
Model Estimation

Estimate $\tilde{P}_i$ using Dataset $\{s_k^i, a^k, s^i_{t+1}, s^i_k\}_{k=0}^{H-1}$

Counts:

$$N_i(s,a) = \sum_{k=1}^{H-1} \sum_{t=0}^{i-1} 1 \left[ s_k^i, a_k^i = s, a \right]$$

# of times we take action $a$ in state $s$.

$$N_i(s,a,s') = \sum_{k=1}^{H-1} \sum_{t=0}^{i-1} 1 \left[ s_k^i, a_k^i, s_{t+1}^i = s, a, s' \right]$$

# 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

$$\hat{V}_t^i(s) = 0.$$ For $t = H-1, H-2, \ldots, 0$:

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

$$\Pi_t^i(s) = \arg\max_a \hat{Q}_t^i(s,a)$$

$$\hat{V}_t^i(s) = \hat{Q}_t^i(s, \Pi_t^i(s))$$
3) Analysis of UCB-VI

Two key facts about UCB-VI:

1) The exploration bonus bounds the difference
\[ \left| \mathbb{E}_{s' \sim \mathcal{P}} [V(s')] - \mathbb{E}_{s' \sim \mathcal{P}} [V(s')] \right| \]
with high probability (similar to confidence intervals |\hat{y} - y| in MAB setting)

2) The exploration bonus yields optimism
\[ \hat{V}_t^i (s) \geq V_t^* (s) \]
(similar to upper confidence bound in MAB setting)

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

\[ R(T) = \mathbb{E} \left[ \sum_{t=1}^{T} V_t^* (s_0) - V_t^0 (s_0) \right] \]

The argument is very similar to the UCB proof:

1) Use optimism: \( V_t^0 (s_0) - V_t^0 (s_0) \leq V_t^0 (s_0) - V_t^0 (s_0) \)

2) Simulation Lemma to compare \( \hat{V}_t^i (s_0) \) & \( V_t^0 (s_0) \).

Regret bound is out of scope for this class (you'd see in 6006 level) but we will prove 2 key facts.
Lemma (Exploration Bonus): for any fixed function \( V: S \rightarrow [0,1] \), with high probability,

\[
| \mathbb{E}[V(s')] - \mathbb{E}[V(s'')] | \leq H \sqrt{\frac{\alpha}{N_i(s,a)}} = b_i(s,a)
\]

where \( \alpha \) is dependent on \( S, A, H \), and probability.

Proof:

\[
| \mathbb{E}[V(s')] - \mathbb{E}[V(s'')] | = \left| \sum_{s' \in S} \left[ \hat{P}_i(s'|s,a) - P(s'|s,a) \right] V(s') \right| 
\leq \sum_{s' \in S} \left| \hat{P}_i(s'|s,a) - P(s'|s,a) \right| |V(s')|
\leq \max_{s'} |V(s')| \cdot \sqrt{\frac{\alpha}{N_i(s,a)}}
\leq H \text{ since reward bounded}
\]

(Using result from Lecture 10, details out of scope)
Lemma: (optimism) as long as \( r(s,a) \in [0,1] \),
\[ \hat{V}_t^i \geq V_t^*(s) \quad \forall \; n, i, s. \]

Proof: We show by induction. \( \hat{V}_H^i = 0 = V_H^* \).

Suppose \( \hat{V}_{t+1}^i(s) \geq V_{t+1}^*(s) \) \( \forall s \).

Then, for any \( s, a \):
\[
\hat{Q}_t^i(s,a) - Q_t^*(s,a) = r(s,a) + b(s,a) + \mathbb{E}_{s' \sim P(s,a)} \left[ \hat{V}_{t+1}^i(s') \right] \\
- \mathbb{E}_{s' \sim P(s,a)} \left[ V_{t+1}^*(s') \right] \\
(\text{by inductive assumption}) \geq b_i(s,a) + \mathbb{E}_{s' \sim P_1(s,a)} \left[ V_{t+1}^*(s') \right] - \mathbb{E}_{s' \sim P_2(s,a)} \left[ V_{t+1}^*(s') \right] \\
(\text{by bonus lemma}) \geq b_i(s,a) - b_i(s,a) = 0.
\]

Therefore, \( \hat{Q}_t^i(s,a) \geq Q_t^*(s,a) \) \( \forall s, a \).

This implies that \( \hat{V}_t^i(s) \geq V_t^*(s) \) \( \forall s \).

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