CS 4/5789 23 Feb 2021 Prof Sarah Dean

Lecture 10: Model Based RL

MDP model $\mathcal{M} = \{\underbrace{\mathcal{S}}_{\text{states}}, \underbrace{\mathcal{A}}_{\text{actions}}, \underbrace{P}_{\text{transition}}, \underbrace{r}_{\text{reward}}, \underbrace{\gamma}_{\text{discount}}\}$ infinite horizon tabular

or $\mathcal{M} = \{\mathbb{R}^{n_s}, \mathbb{R}^{n_a}, \underbrace{f}_{\text{dynamics}}, \underbrace{c}_{\text{cost}}, \underbrace{H}_{\text{Horizon}}, \underbrace{\mu_0}_{\text{initial distribution}}\}$ finite horizon continuous.

But now transitions/dynamics are unknown!

1) MBRL Algorithm with query model

The query model (also called generative model):

For any s, a we can query the transition/dynamics model to sample the next state.

$$s' \sim P(s,a) \quad \left(\text{equivalently, } s' \sim f(s,a,w) \text{ s.t. } w \sim \mathcal{D}\right)$$

Black-box sampling access.

Applicable to games + physics simulators.

Also simple, so it is a good starting point to understand sample complexity: How many samples are required for good performance?

Alg: MBRL with query model

1) For $i = 1, \dots, N$:
   Sample $s_i' \sim P(s_i, a_i)$ and record (s_i', s_i, a_i)
2) Fit transition model $\hat{P}$ from data $\{(s_i', s_i, a_i)\}_{i=1}^N$
3) Design $\hat{\pi}$ using $\hat{P}$

Today we will investigate the sample complexity of this method in two specific settings: tabular & LQR.

2) Tabular Setting

Specializing the algorithm to this setting:

1) sample all (s,a) evenly: $\frac{N}{SA}$ times each

2) Fit transition model by counting

$$\hat{P}(s'|s,a) = \frac{\sum_{i=1}^{N} \mathbb{1}\{s_i = s \,\&\, a_i = a\} \mathbb{1}\{s_i' = s'\}}{\sum_{i=1}^{N} \mathbb{1}\{s_i = s \,\&\, a_i = a\}}$$

3) Design $\hat{\pi}$ with Policy Iteration: $\hat{\pi} = PI(\hat{P}, r)$

Recall: $PI(P, r)$

Initialize π^0

For $t = 1, \to T$:

$$Q^{\pi^t} = \text{PolicyEval}(\pi^t; P, r)$$

$$\left\{ \begin{array}{l} V^{\pi} = (I - \gamma P^{\pi})^{-1} R \\ Q^{\pi^t}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P}(V^{\pi}(s')) \end{array} \right.$$

$$\pi^t(s) = \underset{a}{\text{argmax}}\ Q^{\pi^t}(s,a) \quad \forall s$$

Goal: Compare performance of π_* vs. $\hat{\pi}$

Strategy:

i) Compare P vs. $\hat{P}$

ii) Translate P vs. $\hat{P}$ into difference between value functions

iii) Translate difference in value functions to PI

1) P vs. $\hat{P}$: similar to last lectures discussion

<u>Lemma</u>: with probability $1-\delta$, for all s,a

$$\underbrace{\sum_{s' \in S} |\hat{P}(s'|s,a) - P(s'|s,a)|}_{\|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1} \leq \sqrt{\frac{S^2 A \log(2SA/\delta)}{N}}$$

Proof is out of scope

11) Value Functions: effect of model error

Given a policy π, what is the difference between the value function defined by P compared to the value function defined by $\hat{P}$?

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \begin{array}{l} s_0 = s \\ s_{t+1} \sim P(s_t, a_t) \\ a_t = \pi(s_t) \end{array}\right] \qquad \hat{V}^\pi(s) = \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \begin{array}{l} s_0 = s \\ s_{t+1} \sim \hat{P}(s_t, a_t) \\ a_t = \pi(s_t) \end{array}\right]$$

Recall: Discounted State-action distribution

$$d^\pi_{s_0}(s,a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \underbrace{\mathbb{P}^\pi_t(s,a;s_0)}$$

probability of visiting s,a at step t starting at initial state s_0

Simulation Lemma:

$$\hat{V}^\pi(s_0) - V^\pi(s_0) \leq \frac{1}{(1-\gamma)^2} \mathbb{E}_{s,a \sim d^\pi_{s_0}} \left[\| \hat{P}(\cdot|s,a) - P(\cdot|s,a) \|_1 \right]$$

distribution under true P

disagreement $\hat{P}$ vs. P

Proof: First, we claim that

$$\hat{V}^\pi(s_0) - V^\pi(s_0) = \gamma \mathbb{E}_{a_0 \sim \pi(s_0)} \left[\mathbb{E}_{s_1 \sim \hat{P}(s_0,a_0)} [\hat{V}^\pi(s_1)] - \mathbb{E}_{s_1 \sim P(s_0,a_0)} [\hat{V}^\pi(s_1)] \right]$$
$$+ \gamma \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_1 \sim P(s_0,a_0)}} \left[\hat{V}^\pi(s_1) - V^\pi(s_1) \right]$$

By iterating this expression k times,

$$\hat{V}^\pi(s_0) - V^\pi(s_0) = \sum_{t=1}^{k} \gamma^t \mathbb{E}_{s_{t-1},a_{t-1}} \left[\mathbb{E}_{s_t \sim \hat{P}(s_{t-1},a_{t-1})} [\hat{V}^\pi(s_t)] - \mathbb{E}_{s_t \sim P(s_{t-1},a_{t-1})} [\hat{V}^\pi(s_t)] \right]$$
$$+ \gamma^k \mathbb{E}_{s_k \sim P(s_{k-1},a_{k-1})} \left[\hat{V}^\pi(s_k) - V^\pi(s_k) \right]$$

Letting $k \to \infty$,

$$\hat{V}^\pi(s_0) - V^\pi(s_0) = \frac{\gamma}{1-\gamma} \mathbb{E}_{s,a \sim d^\pi_{s_0}} \left[\mathbb{E}_{s' \sim \hat{P}(s,a)}[\hat{V}^\pi(s')] - \mathbb{E}_{s' \sim P(s,a)}[\hat{V}^\pi(s')] \right]$$

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

since $r(s,a) \leq 1$, $\hat{V}^\pi(s') \leq \frac{1}{1-\gamma}$

$$\leq \sum_{s' \in S} \left| \hat{P}(s'|s,a) - P(s'|s,a) \right| \frac{1}{1-\gamma}$$

$$= \frac{1}{1-\gamma} \| \hat{P}(\cdot|s,a) - P(\cdot|s,a) \|_1$$

Then all that's left is to prove the initial claim.

$$\hat{V}^\pi(s_0) - V^\pi(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, \hat{P} \right] - \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, \hat{P} \right]$$

(t=0 term is equal)

$$= \gamma \, \mathbb{E}_{a_0 = \pi(s_0)} \left[\mathbb{E}_{s_1 \sim \hat{P}(s_0,a_0)}[\hat{V}^\pi(s_1)] - \mathbb{E}_{s_1 \sim P(s_0,a_0)}[V^\pi(s_1)] \right]$$

$$= \gamma \, \mathbb{E}_{a_0 = \pi(s_0)} \left[\underbrace{\mathbb{E}_{s_1 \sim \hat{P}}[\hat{V}^\pi(s_1)] - \mathbb{E}_{s_1 \sim P}[\hat{V}^\pi(s_1)]} + \underbrace{\mathbb{E}_{s_1 \sim P}[\hat{V}^\pi(s_1)] - \mathbb{E}_{s_1 \sim P}[V^\pi(s_1)]} \right]$$

✓ $\square$

III) Policy Iteration

Let $\hat{\pi}^* = PI(\hat{P}, r)$ ← ignore iteration approximation for now — assume $T > SA$ (HW1)

comparing to true optimal value.

$$V^*(s_0) - V^{\hat{\pi}^*}(s_0) \leq V^*(s_0) - \underbrace{\hat{V}^{\pi^*}(s_0) + \hat{V}^{\hat{\pi}^*}(s_0)}_{\hat{\pi}^* \text{ is optimal on } \hat{P} \text{ so } \hat{V}^{\hat{\pi}^*}(s_0) \geq \hat{V}^{\pi}(s_0)\ \forall \pi.} - V^{\hat{\pi}^*}(s_0)$$

(simulation lemma 2x)
$$\leq \frac{1}{(1-\gamma)^2}\left(\mathbb{E}_{s,a\sim d_{s_0}^{\pi^*}} \|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 + \mathbb{E}_{s,a\sim d_0^{\hat{\pi}^*}} \|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1\right)$$

(model error bound)
$$\leq \frac{2}{(1-\gamma)^2}\sqrt{\frac{S\log(2SA/\delta)}{N}} \quad \text{w.p. } 1-\delta$$

<u>Theorem:</u> (Sample complexity)

For $0 \leq \delta \leq 1$, $0 \leq \varepsilon \leq \frac{1}{1-\gamma}$, Let $N = \frac{4S^2 A \log\left(\frac{2SA}{\delta}\right)}{\varepsilon^2 (1-\gamma)^4}$

Then with probability at least $1-\delta$,

$$V^*(s_0) - V^{\hat{\pi}^*}(s_0) \leq \varepsilon.$$

3) LQR

MBRL in this setting:

1) generate i.i.d. samples $s_i \sim \mathcal{N}(0, \sigma^2 I)$, $a_i \sim \mathcal{N}(0, \sigma^2 I)$

2) estimate parameters by least squares

$$(\hat{A}, \hat{B}) = \operatorname{argmin} \sum_{i=1}^{N} (s_i' - As_i - Ba_i)^2$$

3) compute $K_* = LQR(\hat{A}, \hat{B}, Q, R)$

We won't derive results in detail for this setting. But at a high level,

i) parameter estimation

$$\left\| \begin{bmatrix} \hat{A} - A \\ \hat{B} - B \end{bmatrix} \right\|_2 \lesssim \sqrt{\frac{(n_s + n_a)\log(1/\delta)}{N}}$$

matrix norm

ii) Difference in value $(V_t^*(s) = s^T P_t s + p_t)$

$$\| P_t - \hat{P}_t \|_2 \lesssim \left\| \begin{bmatrix} \hat{A} - A \\ \hat{B} - B \end{bmatrix} \right\|_2$$

iii) Difference in performance

$$\hat{V}_0^*(s_0) - V_0^*(s_0) \lesssim \| P_t - \hat{P}_t \|_2 \lesssim \sqrt{\frac{(n_s + n_a)\log(1/\delta)}{N}}$$

Sample complexity: ε-optimal policy after $N \gtrsim \frac{(n_s + n_a)^2}{\varepsilon^2}$ sampes