Lecture 10: Model Based RL

MDP model \( \mathcal{M} = \{ S, A, P, R, \gamma \} \) infinite horizon tabular

or \( \mathcal{M} = \{ \mathbb{R}^n, \mathbb{R}^n, f, c, \gamma, \mu_0 \} \) 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.

\[ s' \sim P(s, a) \] (equivalently, \( s' \sim f(s, a, w) \) s.t. \( w \sim D \))

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 \( i = 1, ..., N \):
   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 \( \sum_{i=1}^{N} (s'_i, s_i, a_i) \)
3) Design \( \hat{r} \) 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

\[
\hat{P}(s' | s, a) = \frac{\sum_{i=1}^{N} \mathbb{I}\{s_i = s \land a_i = a \land s_{i+1} = s'\}}{\sum_{i=1}^{N} \mathbb{I}\{s_i = s \land a_i = a\}}
\]

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

Recall: \(\text{PI}(P, r)\)

Initialize \(\pi^0\)

For \(t = 1, \ldots, T\):

\[Q^\pi_t = \text{Policy Eval}(\pi^t, P, r)\]

\[\pi^t(s) = \arg\max_a Q^\pi_t(s, a) \land s\]

**Goal:** Compare performance of \(\pi^*\) vs. \(\hat{\pi}\)

**Strategy:**

1) Compare \(P\) vs. \(\hat{P}\)

2) Translate \(P\) vs. \(\hat{P}\) into difference between value functions

3) Translate difference in value functions to \(\text{PI}\)

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

**Lemma:** with probability \(1 - \delta\), for all \(s, a\)

\[
\sum_{s' \in S} |\hat{P}(s' | s, a) - P(s' | s, a)| \leq \sqrt{SA \text{log} \left(\frac{2SA}{\delta}\right) \frac{1}{N}}
\]

**Proof is out of scope**
II) Value Functions: effect of model error

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

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

Recall: Discounted state-action distribution

\[
d_\pi(s, a) = (1 - \gamma) \sum_{s_0} \gamma^t \mathbb{P}(s, a, s_0)
\]

probability of visiting \( s, a \) at step \( t \)

starting at initial state \( s_0 \)

**Simulation Lemma:**

\[
\hat{V}^\pi(s_0) - V^\pi(s_0) \leq \frac{\gamma}{(1-\gamma)^2} \mathbb{E} \left[ \| \mathbb{P}(\cdot | s_0, a_0) - \mathbb{P}(\cdot | s_0, a_0) \|_1 \right]
\]

disagreement \( \hat{P} \) vs. \( P \)

distribution under true \( P \)

**Proof:** First, we claim that

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

\[
+ \gamma \mathbb{E} \left[ \hat{V}^\pi(s_1) - V^\pi(s_1) \right]
\]

By iterating this expression \( K \) times,

\[
\hat{V}^\pi(s_0) - V^\pi(s_0) = \sum_{k=1}^{K} \gamma^k \mathbb{E} \left[ \mathbb{E} \left[ \hat{V}^\pi(s_k) \right] - \mathbb{E} \left[ V^\pi(s_k) \right] \right]
\]

\[
+ \gamma^K \mathbb{E} \left[ \hat{V}^\pi(s_k) - V^\pi(s_k) \right]
\]
letting $k \to \infty$, 
\[ \hat{V}^\pi(s_0) - V^\pi(s_0) = \frac{1}{1 - \gamma} \mathbb{E}_{s_0 \sim p(s, a)} \left[ \mathbb{E}_{s'} \left( \hat{V}^\pi(s') \right) - \mathbb{E}_{s' \sim p(s, a)} \left( \hat{V}^\pi(s') \right) \right] \]

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

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

\[ \leq \sum_{s' \in \mathcal{S}} \left| \hat{p}(s' | s, a) - p(s' | s, a) \right| \frac{1}{1 - \gamma} \]

\[ = \frac{1}{1 - \gamma} \sum_{s' \in \mathcal{S}} \left| \hat{p}(s' | s, a) - p(s' | s, a) \right| \]

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

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

(t=0 term is equal)

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

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

\[ = \gamma \mathbb{E} \left[ \hat{V}^\pi(s_1) - \hat{V}^\pi(s_1) + \hat{V}^\pi(s_1) - \hat{V}^\pi(s_1) \right] \]

\[ = 0 \]
Policy Iteration

Let $\hat{\pi^*} = \Pi (\hat{\mathcal{P}}, \mathcal{R})$ (ignore iteration approximation for now—assume $T > S + (HTW)$)

comparing to true optimal value:

$$V^*(s_0) - V^{\hat{\pi}^*}(s_0) \leq V^*(s_0) - V^{\hat{\pi}^*}(s_0) + \hat{\pi}^* \text{ is optimal on } \hat{\mathcal{P}}_0, s_0 \Rightarrow \hat{\pi}^*(s) \geq V^{\pi}(s_0) \forall \pi.$$

(simulation lemma 2x)

$$\leq \frac{\gamma}{(1-\gamma)^2} \left( \mathbb{E}_{s_0, a, \pi \sim \hat{\pi}^*} \| \mathbb{P}_0(s' | s, a) - \mathbb{P}_0(s' | s, a) \|_1 \right)$$

(model error bound)

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

Theorem: (Sample complexity)

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

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 iid. samples $s_i \sim N(0, \rho^2 I)$, $a_i \sim N(0, \rho^2 I)$.

2) estimate parameters by least squares

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

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,

1) parameter estimation

$$\| [\hat{A} - A, \hat{B} - B] \|_2 \lesssim \sqrt{\frac{(n_s + n_a) \log(1/\delta)}{N}}$$

matrix norm

II) Difference in value

$$\| P_t - \hat{P}_t \|_2 \lesssim \| [\hat{A} - A, \hat{B} - B] \|_2$$

III) Difference in performance

$$\tilde{V}_v^*(s_0) - V_v^*(s_0) \leq \| P_t - \hat{P}_t \|_2 \leq \sqrt{\frac{(n_s + n_a) \log(1/\delta)}{N}}$$

Sample complexity: $\epsilon$-optimal policy
after $N \geq \frac{(n_s + n_a)}{\epsilon^2}$ samples