Model-based RL

Recap: Planning algorithm for computing $\pi^\star$

We assumed that $P(s' \mid s, a), r(s, a)\, \forall s, a, s'$ are **known**

Value Iteration:

$$Q^{t+1}(s,a) \Leftarrow r(s,a) + \max_a \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} Q^t(s', a'), \forall s, a$$

Policy Iteration:

$$\pi^{t+1}(s) = \arg\max_a Q^{\pi^t}(s,a), \text{ for all } s$$

Recap: Value-based Learning

When $P(s' \mid s, a)$ is unknown, Q-learning aims to learn $Q^\star$ directly

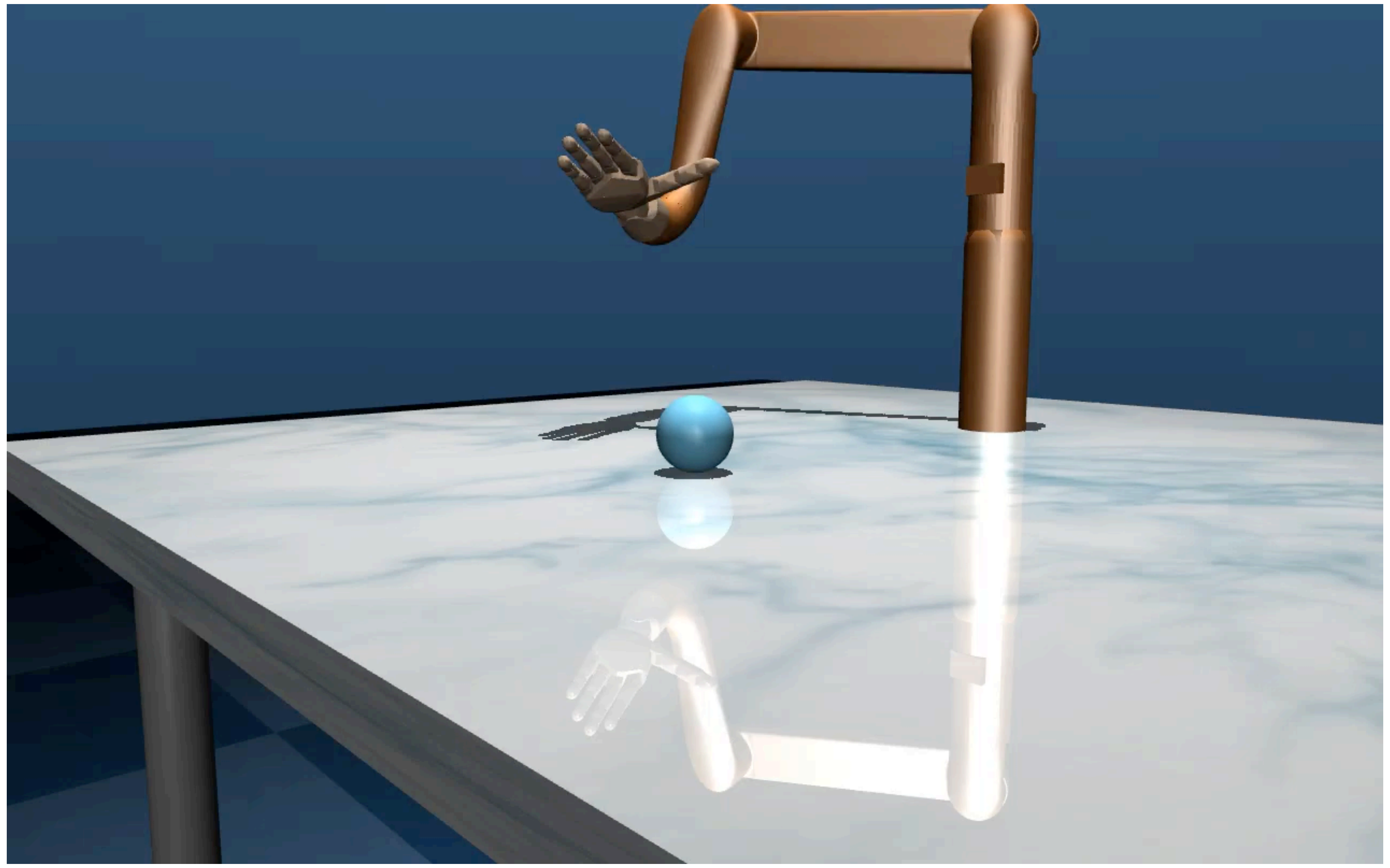
$$\hat{Q}(s,a) \Leftarrow \hat{Q}(s,a) + \left(r(s,a) + \max_{a'} \hat{Q}(s',a') - \hat{Q}(s,a) \right)$$

where $a \sim \epsilon$-greedy$(\hat{Q})$, and $s' \sim P(\cdot \mid s, a)$, $r = r(s,a)$

Questions for Today:

Can we **learn the transition** from data and then compute its optimal policy; and what performance guarantee we can get?

Motivation for Model-based Approach



While we cannot model the exact analytical dynamics,
we can learn it from data $\{s, a, s'\}$

Then we do planning: e.g.,
$\widehat{\pi}^\star = \text{VI}(\widehat{P}, r)$

(Often in practice we iterate the above process)

Motivation for Model-based Approach

Potential benefits of learning model over Q-learning

There are cases where model is much easier to learn than value function

Once model is learned, we can optimize different rewards (i.e., multi-task)

Outline:

1. Simulation lemma:
What is the performance of π under $(\widehat{P}, r)$

2. Algorithm: estimate $\widehat{P}$ from data
and compute $\widehat{\pi}^\star$—the optimal policy of $\widehat{P}$

3. Analyzing the performance $\widehat{\pi}^\star$ under (P, r)

State-action distribution

Given a policy π and s_0, we denote $P_h^\pi(s, a \mid s_0)$ as the **prob of reaching (s, a) at time h, given we start at s_0**

Denote $d_{s_0}^\pi$ as the average state-action distribution:

$$d_{s_0}^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h P_h^\pi(s, a \mid s_0)$$

A key fundamental question in Model-based RL:

Given two transitions $\hat{P}$ and P, how would π behave differently in $\hat{P}$ and P?

Denote:

$$\widehat{V}^{\pi}(s_0) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,|\, \pi, \widehat{P}\right]; \quad V^{\pi}(s_0) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,|\, \pi, P\right];$$

What is the difference between $\widehat{V}^{\pi}(s_0)$ & $V^{\pi}(s_0)$?

In other words, how does the model error propagate to values

Simulation Lemma

Simulation Lemma:

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

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

Simulation Lemma:

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

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

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

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

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

Summary so far:

Simulation Lemma:

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

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

Total model disagreement over the real traces

Outline:

✓ **1. Simulation lemma:**

What is the performance of π under any estimator $\widehat{P}$

2. Algorithm: estimate $(\widehat{P}, \widehat{r})$ from data and compute $\widehat{\pi}^\star$—the optimal policy of $(\widehat{P}, \widehat{r})$

3. Analyzing the performance $\widehat{\pi}^\star$ under (P, r)

A Model-based Algorithm

Assume reward r is known (just for analysis simplicity):

1. Model fitting:

$\forall s, a$: collect N next states, $s_i' \sim P(\cdot \mid s, a), i \in [N]$; set

$$\widehat{P}(s' \mid s, a) = \frac{\sum_{i=1}^{N} \mathbf{1}\{s_i' = s'\}}{N};$$

2. Planning w/ the learned model:

$$\widehat{\pi}^\star = \textbf{PI}\left(\widehat{P}, r\right)$$

Detour: estimating mean of Bernoulli distribution

Given: we have a biased coin:

With probability p, it gives +1, and w/ prob 1-p, it gives -1;

To estimate p: We flip the coin N times independently, get N outcomes, $\{x_i\}_{i=1}^N$,

$$x_i \in \{-1, +1\}$$

$$\hat{p} = \frac{\sum_{i=1}^N \mathbf{1}\{x_i = +1\}}{N}$$

(Informal) we can show that $|\hat{p} - p| \lesssim \sqrt{\frac{1}{N}}$

(proof out of scope)

Model-based RL

1. Model fitting:

$\forall s, a$: collect N next states,

$s_i' \sim P(\cdot \mid s, a), i \in [N]$; set

$$\widehat{P}(s' \mid s, a) = \frac{\sum_{i=1}^{N} \mathbf{1}\{s_i' = s'\}}{N};$$

Estimating the prob of observing s' at (s, a)

We can show (informally):

$$\|\hat{P}(\cdot \mid s, a) - P(\cdot \mid s, a)\|_1 \lesssim \sqrt{1/N}, \forall s, a$$

2. Planning w/ the learned model:

$$\widehat{\pi}^\star = \textbf{PI}\left(\widehat{P}, r\right)$$

Summary so far:

By collecting data (next state) at very (s, a), we build an estimator $\hat{P}$ that is close to P

(e.g., possible to show error shrinks $\approx 1/\sqrt{N}$)

Outline:

1. Simulation lemma:
What is the performance of π under any estimator $\widehat{P}$

2. Algorithm: estimate $(\widehat{P}, \hat{r})$ from data
and compute $\widehat{\pi}^\star$—the optimal policy of $(\widehat{P}, \hat{r})$

3. Analyzing the performance $\widehat{\pi}^\star$ under (P, r)

Performance of the learned policy

Now assume that $\hat{P}$ is nearly accurate, i.e., $\|\hat{P}(\cdot \mid s, a) - P(\cdot \mid s, a)\|_1 \leq \epsilon, \forall s, a$

How good is $\hat{\pi}^\star = \text{PI}(\hat{P}, r)$?

$$V^\star(s_0) - V^{\hat{\pi}^\star}(s_0)$$

Q: why this is true?

$$\leq V^\star(s_0) - \widehat{V}^{\pi^\star}(s_0) + \widehat{V}^{\hat{\pi}^\star}(s_0) - V^{\hat{\pi}^\star}(s_0)$$

(Simulation lemma)

$$\leq \frac{1}{(1-\gamma)^2}\left[\mathbb{E}_{s,a\sim d_{s_0}^{\pi^\star}}\|\widehat{P}(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 + \mathbb{E}_{s,a\sim d_{s_0}^{\hat{\pi}^\star}}\|\widehat{P}(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1\right]$$

$$\leq \frac{2}{(1-\gamma)^2} \cdot \epsilon,$$

Summary for Today:

1. A model-based RL **Algorithm for small-size MDP**

$$\widehat{P}(s' \mid s, a) = \frac{\sum_{i=1}^{N} \mathbf{1}\{s'_i = s'\}}{N}, \forall s, a; \quad \widehat{\pi}^\star = \textbf{PI}\left(\widehat{P}, r\right)$$

2. **Simulation lemma** allows us to link model error to policy's performance

3. Good model leads to a good policy, up to some error amplification based on effective horizon