Model-based RL under the Generative Model Setting

Recap: Infinite Horizon MDP

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

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Policy $\pi : S \mapsto A$

Bellman Equation:

$$V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, \pi(s))} V^\pi(s')$$

Recap: Infinite Horizon MDP

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

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Policy $\pi : S \mapsto A$

Bellman Equation:

$$V^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} V^\pi(s')$$

Bellman Optimality:

$$Q^\star(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a' \in A} Q^\star(s',a') \right]$$

$$V^\star(s) = \max_a \left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right)$$

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

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

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

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

$Q^t \to Q^*, \pi^t$

$\forall s, a$

Value Iteration:

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

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: $\rightarrow$ Approximate

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

Policy Iteration: $\longrightarrow \pi^*$

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

$$Q^{\pi^{t+1}}(s,a) \geq Q^{\pi^t}(s,a), \ \forall s, a$$

Recap: State-action distribution

Given some s_0, and policy π, we denote $d^{\pi}_{s_0}(s, a)$ as:

$$d^{\pi}_{s_0}(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}^{\pi}_h(s, a; s_0)$$

Recap: State-action distribution

Given some s_0, and policy π, we denote $d^{\pi}_{s_0}(s, a)$ as:

$$d^{\pi}_{s_0}(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}^{\pi}_h(s, a; s_0)$$

Probability of π visiting (s, a) at step h starting from the fixed initial state s_0

A **new** setting: Generative Model

$P(s' | s, a), \forall s, a, s'$

In VI, PI, DP (for tabular MDP and LQR), we have **known** P, r

$Ax + Bu + w$, $w \sim N(0, \sigma^2 I)$

$P(x' | x, u) = N(Ax + Bu, \sigma^2 I)$

A **new** setting: Generative Model

In VI, PI, DP (for tabular MDP and LQR), we have **known** P, r

We will focus on generative model setting here:

We can reset to any (s, a), and get a sample $s' \sim P(\cdot \mid s, a)$

A **new** setting: Generative Model

In VI, PI, DP (for tabular MDP and LQR), we have **known** P, r

We will focus on generative model setting here:

We can reset to any (s, a), and get a sample $s' \sim P(\cdot \mid s, a)$

This is weaker than the known setting,
and valid for problems such as board games, control/planning in simulation etc

Questions for Today:

Under the generative model setting, how we learn to compute $\pi^\star$;
and what performance guarantee we can get?

Questions for Today:

Under the generative model setting, how we learn to compute $\pi^\star$; and what performance guarantee we can get?

(We will see the first sample complexity analysis..)

If I want an ε-optimal policy,

How many samples do I need ?

Outline:

1. Simulation lemma:

$\widehat{P} \approx P$

What is the performance of π under $(\widehat{P}, r)$

2. Algorithm: estimate $\widehat{P}$ from data

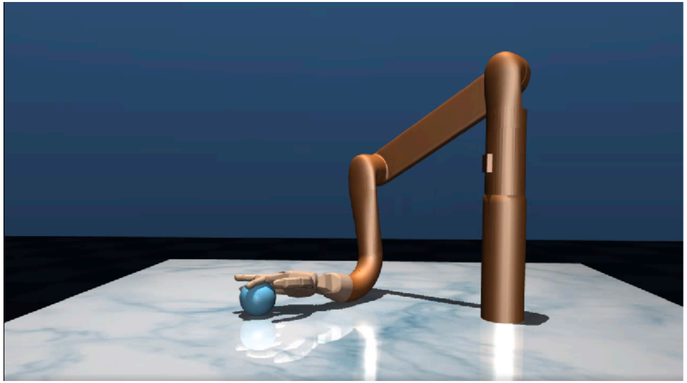
and compute $\widehat{\pi}^\star$—the optimal policy of $\widehat{P}$

$\leftarrow \mathrm{PI}(\widehat{P}, r)$

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

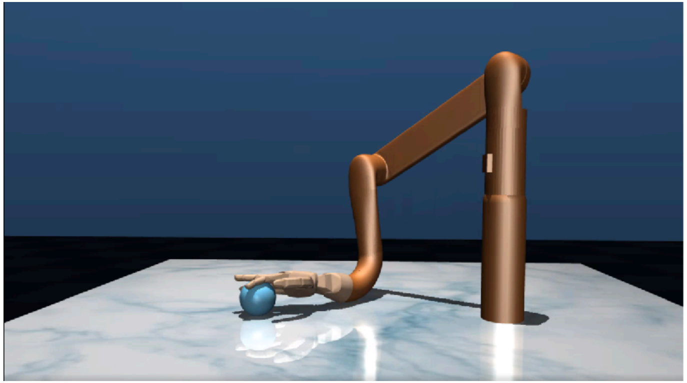
Motivation for Model-based Approach

It is a very common and default approach to try in practice



Motivation for Model-based Approach

It is a very common and default approach to try in practice



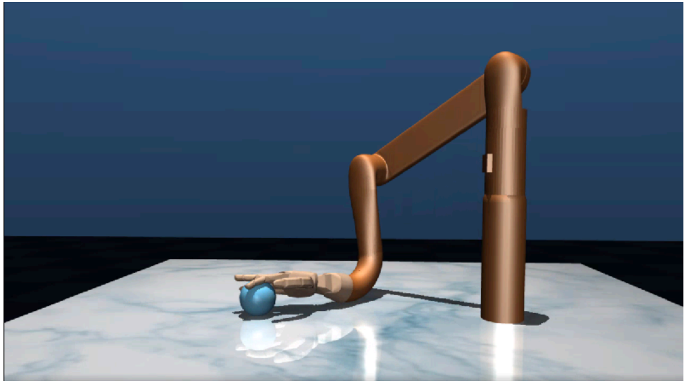
While we cannot write out the exact analytical dynamics,
we can learn it from data $\{s, a, s'\}$

$$\min_{f} \sum_{i=1}^{N} \| f(s_i, a_i) - s_i' \|_2^2, \quad s \in \mathbb{R}^d$$

$$\max_{\hat{P}} \sum_{i=1}^{N} \ln\left(\hat{P}(s_i' \mid s_i, a_i)\right)$$

Motivation for Model-based Approach

It is a very common and default approach to try in practice



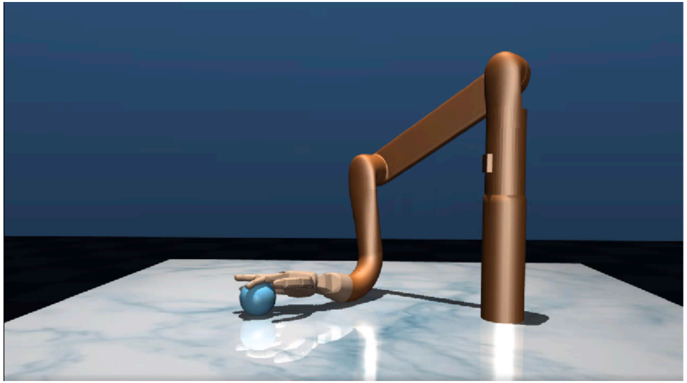
While we cannot write out the exact analytical dynamics,
we can learn it from data $\{s, a, s'\}$

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

$\mathrm{PI}(\widehat{P}, r)$

Motivation for Model-based Approach

It is a very common and default approach to try in practice



While we cannot write out the exact analytical dynamics,

we can learn it from data $\{s, a, s'\}$

Then we do planning: e.g.,

$\widehat{\pi}^\star = \mathrm{VI}(\widehat{P}, r)$

(Often in practice we iterate the above process)

A key fundamental question in Model-based RL:

Notations:

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

$$\left| \widehat{V}^{\pi}(s_0) - V^{\pi}(s_0) \right| \leq ?? \leftarrow f(\widehat{P}, P)$$

A key fundamental question in Model-based RL:

Notations:

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

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

A key fundamental question in Model-based RL:

Notations:

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

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:

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

A key fundamental question in Model-based RL:

Notations:

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

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

$P, \ Q$

$$\left| \mathbb{E}_{x\sim P} f(x) - \mathbb{E}_{x\sim Q} f(x) \right|, \quad \forall f$$

Simulation Lemma:

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

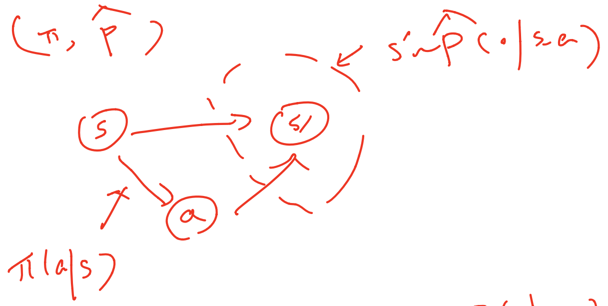
Distribution of π under the true model P

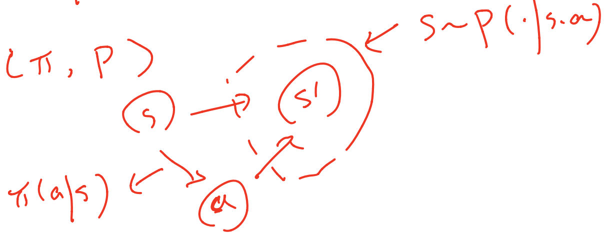
Simulation Lemma Explanation

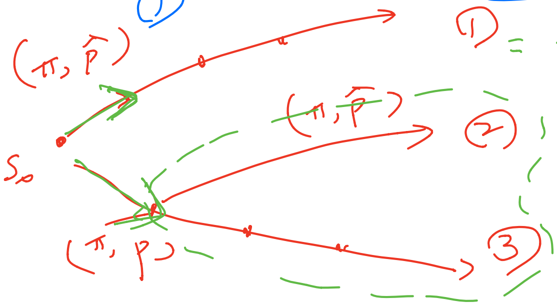
Simulation Lemma:

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

① ②







① $= \widehat{V}^{\pi}(s_0) =$

② $= r(s_0, \pi(s_0)) + \mathbb{E}_{s'\sim P(s_0,\pi(s_0))}\widehat{V}^{\pi}(s')$

③ $= V^{\pi}(s_0)$

$$① - ③ = \underbrace{① - ②}_{h=0} + \underbrace{② - ③}_{\text{Recursion}}$$

Simulation Lemma Proof

Simulation Lemma:

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

Simulation Lemma Proof

Simulation Lemma:

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

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

$$\underline{r(s_0,\pi(s_0))} + \gamma\,\mathbb{E}_{s'\sim\widehat{P}(\cdot|s_0,\pi(s_0))}\widehat{V}^{\pi}(s')$$

$$V^{\pi}(s_0) = \underline{r(s_0,\pi(s_0))} + \gamma\,\mathbb{E}_{s'\sim P(\cdot|s_0,\pi(s_0))}V^{\pi}(s')$$

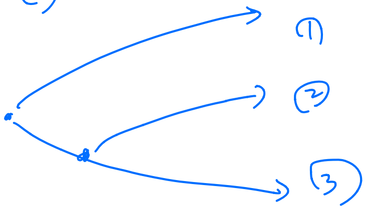
Simulation Lemma Proof

Simulation Lemma:

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

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

$$= \gamma\mathbb{E}_{a_0\sim\pi(\cdot|s_0)}\left[\mathbb{E}_{s_1\sim \widehat{P}(s_0,a_0)}\widehat{V}^\pi(s_1) \underbrace{- \mathbb{E}_{s_1\sim P(s_0,a_0)}\widehat{V}^\pi(s_1) + \mathbb{E}_{s_1\sim P(s_0,a_0)}\widehat{V}^\pi(s_1)}_{2} - \mathbb{E}_{s_1\sim P(s_0,a_0)}V^\pi(s_1)\right]$$



$\mathbb{P}^{\pi}_{h}(s,a;s_0)$

$\mathbb{P}^{\pi}_{h=1}(s,a;s_0)$

Simulation Lemma Proof

Simulation Lemma:

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

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

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

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

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

① - ②

② - ③ ← do one more step Recursion

$\gamma + \gamma^2 + \gamma^3 \cdots = \frac{\gamma}{1-\gamma}$

Summary so far:

$\{s, a, s'\}$

$$\hat{f} = \min_{f} \sum_{i=1}^{N} \| f(s,a) - s' \|_2^2$$

$$\hat{f}(s,a) \approx f^*(s,a)$$

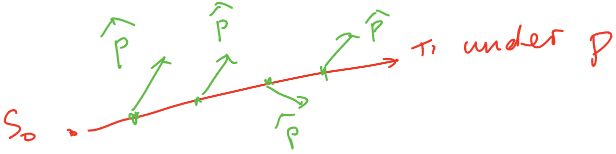
Simulation Lemma:

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

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

$$\widehat{V}^{\pi} \in \left[0, \frac{1}{1-\gamma}\right]$$

Total model disagreement over the real traces



P, Q

$$\left| \mathbb{E}_{x \sim P} f(x) - \mathbb{E}_{x \sim Q} f(x) \right| \leq \sup_x |f(x)| \, \| P - Q \|_1$$

$$= \frac{1}{2} \sup_x |f(x)| \, \| P - Q \|_{TV}$$

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):

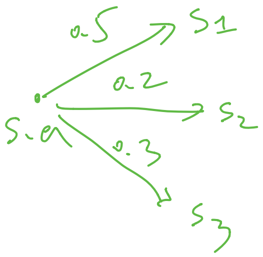
A Model-based Algorithm

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

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



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

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

$\widehat{P}(s' \mid s, a) \to P(s' \mid s, a)$
when $N \to \infty$

2. Planning w/ the learned model:

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

Steps of Analysis

1. Model fitting:

Generative model

$\forall s, a$: 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:

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

Steps of Analysis

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:

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

1. How good is our learned model? I.e., $\widehat{P}(\cdot \mid s, a) \approx P(\cdot \mid s, a)$??

$$\left\| \widehat{P}(\cdot \mid s, a) - P(\cdot \mid s, a) \right\|_1 \leq ??$$

Steps of Analysis

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:

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

optimal policy under $\widehat{P}$

1. How good is our learned model? I.e., $\widehat{P}(\cdot \mid s, a) \approx P(\cdot \mid s, a)$??

2. How model error propagates to the performance of $\widehat{\pi}^\star$ (simulation lemma)

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;

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 do experiments:

We flip the coin N times independently, get N outcomes, $\{x_i\}_{i=1}^N, x_i \in \{-1, +1\}$

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 do experiments:

We flip the coin N times independently, get N outcomes, $\{x_i\}_{i=1}^N, x_i \in \{-1, +1\}$

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

$\hat{p} \to p, \quad N \to \infty$

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 do experiments:

We flip the coin N times independently, get N outcomes, $\{x_i\}_{i=1}^N, x_i \in \{-1, +1\}$

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

$\hat{p} + \sqrt{\frac{\ln(1/\delta)}{N}}$

p

$\hat{p}$

$\hat{p} - \sqrt{\frac{\ln(1/\delta)}{N}}$

W/ probability at least $1-\delta$, we will have $|\hat{p} - p| \leq O\left(\sqrt{\frac{\ln(1/\delta)}{N}}\right)$

(concentration bound; Hoeffding's inequality; proof out of scope)

$x \in \mathbb{R}^d$

$\|x\|_1 = \sum_i |x_i|$

Steps of Analysis: model error

1. Model fitting:

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

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

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

1. How good is our learned model? I.e., $\widehat{P}(\cdot \mid s, a) \approx P(\cdot \mid s, a)$??

$\forall s, a$ wp $P(s' \mid s, a)$, we get s'

wp $1 - P(s' \mid s, a)$, we observe something else

$$\left| \widehat{P}(s' \mid s, a) - P(s' \mid s, a) \right| \le \sqrt{\frac{1}{N}} \Rightarrow$$

$$\| \widehat{P}(\cdot \mid s, a) - P(\cdot \mid s, a) \|_1 = \sum_{s' \in S} \left| \widehat{P}(s' \mid s, a) - P(s' \mid s, a) \right| \le S \sqrt{\frac{1}{N}}$$

Steps of Analysis: model error

1. Model fitting:

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

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

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

1. How good is our learned model? I.e., $\widehat{P}(\cdot \mid s, a) \approx P(\cdot \mid s, a)$??

Lemma (proof is out of scope): with probability $1 - \delta$, we have that for all s, a,

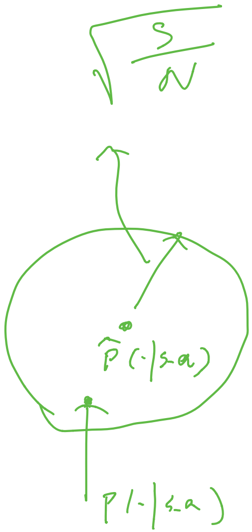
$$\left\| \widehat{P}(\cdot \mid s, a) - P(\cdot \mid s, a) \right\|_1 \leq \sqrt{\frac{S \ln(2SA/\delta)}{N}} \approx \sqrt{\frac{S}{N}}$$

Summary so far:

We get a confidence ball (under ℓ_1 norm) for P:

W/ probability at least $1-\delta$:

$$\forall s,a \quad \left\| \widehat{P}(\cdot \mid s,a) - P(\cdot \mid s,a) \right\|_1 \leq \sqrt{\frac{S \ln(2SA/\delta)}{N}}$$



Steps of Analysis: performance of the learned policy

Lemma (proof is out of scope): with probability $1-\delta$, we have that for all s, a,

$$\left\| \widehat{P}(\cdot \mid s,a) - P(\cdot \mid s,a) \right\|_1 \leq \sqrt{\frac{S \ln(2SA/\delta)}{N}}$$

2. Planning w/ the learned model: $\widehat{\pi}^\star = \textbf{PI}\left(\widehat{P}, r\right)$

optimal for $\widehat{P}$

Steps of Analysis: performance of the learned policy

Lemma (proof is out of scope): with probability $1-\delta$, we have that for all s, a,

$$\left\| \widehat{P}(\cdot \mid s,a) - P(\cdot \mid s,a) \right\|_1 \leq \sqrt{\frac{S \ln(2SA/\delta)}{N}}$$

2. Planning w/ the learned model: $\widehat{\pi}^\star = \textbf{PI}\left(\widehat{P}, r \right)$

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

Steps of Analysis: performance of the learned policy

Lemma (proof is out of scope): with probability $1-\delta,$ we have that for all $s, a,$

$$\left\| \widehat{P}(\cdot \mid s, a) - P(\cdot \mid s, a) \right\|_1 \leq \sqrt{\frac{S \ln(2SA/\delta)}{N}}$$

2. Planning w/ the learned model: $\widehat{\pi}^\star = \textbf{PI}\left(\widehat{P}, r \right)$

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

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

Δ

Steps of Analysis: performance of the learned policy

Lemma (proof is out of scope): with probability $1-\delta$, we have that for all s, a,

$$\left\| \widehat{P}(\cdot \mid s,a) - P(\cdot \mid s,a) \right\|_1 \le \sqrt{\frac{S \ln(2SA/\delta)}{N}}$$

$\widehat{V}^{\pi}$: performance of π under $\widehat{P}$

2. Planning w/ the learned model: $\widehat{\pi}^\star = \text{PI}\left(\widehat{P}, r\right)$

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

$$\le V^\star(s_0) \underbrace{- \widehat{V}^{\pi^\star}(s_0) + \widehat{V}^{\widehat{\pi}^\star}(s_0)}_{\geq 0} - V^{\widehat{\pi}^\star}(s_0)$$

Q: why this is true?

$\widehat{\pi}^*$ is optimal under $\widehat{P}$

$\widehat{V}^{\widehat{\pi}^*} \geq \widehat{V}^{\pi^*}$

Steps of Analysis: performance of the learned policy

Lemma (proof is out of scope): with probability $1-\delta,$ we have that for all s, a,

$$\left\| \widehat{P}(\cdot \mid s,a) - P(\cdot \mid s,a) \right\|_1 \leq \sqrt{\frac{S \ln(2SA/\delta)}{N}}$$

2. Planning w/ the learned model: $\widehat{\pi}^\star = \textbf{PI}\left(\widehat{P}, r \right)$

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

Q: why this is true?

$\leq V^\star(s_0) - \widehat{V}^{\pi^\star}(s_0) + \widehat{V}^{\widehat{\pi}^\star}(s_0) - V^{\widehat{\pi}^\star}(s_0)$ ← apply simulation-lemma

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

Steps of Analysis: performance of the learned policy

Lemma (proof is out of scope): with probability $1-\delta$, we have that for all s, a,

$$\left\| \widehat{P}(\cdot \mid s,a) - P(\cdot \mid s,a) \right\|_1 \leq \sqrt{\frac{S \ln(2SA/\delta)}{N}}$$

2. Planning w/ the learned model: $\widehat{\pi}^\star = \mathbf{PI}\left(\widehat{P}, r\right)$

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

Q: why this is true?

$\leq V^\star(s_0) - \widehat{V}^{\pi^\star}(s_0) + \widehat{V}^{\widehat{\pi}^\star}(s_0) - V^{\widehat{\pi}^\star}(s_0)$ ← optimality of $\widehat{\pi}^\star$ under $\widehat{P}$

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

← simulation lemma

$\leq \frac{2}{(1-\gamma)^2}\sqrt{\frac{S \ln(2SA/\delta)}{N}}$, wp $1-\delta$; ← Confidence Ball

Summary so far:

Theorem (Sample Complexity):

Fix $\delta \in (0,1)$, $\epsilon \in (0, 1/(1-\gamma))$, set $N = \dfrac{4S\ln(2SA/\delta)}{\epsilon^2(1-\gamma)^4}$;

with probability at least $1-\delta$, we have:

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

$$V^*(s_0) - V^{\hat{\pi}^*}(s_0) \le \sqrt{\frac{S}{N}} \cdot \frac{1}{(1-\gamma)^2} = \varepsilon$$

$$\Rightarrow \quad N = \frac{S}{\varepsilon^2(1-\gamma)^4}$$

$$SA \cdot N = S^2A / (\varepsilon^2(1-\gamma)^4)$$

Summary so far:

Theorem (Sample Complexity):

Fix $\delta \in (0,1), \epsilon \in (0,1/(1-\gamma))$, set $N = \dfrac{4S\ln(2SA/\delta)}{\epsilon^2(1-\gamma)^4}$;

with probability at least $1-\delta$, we have:

$$V^\star(s_0) - V^{\widehat{\pi}^\star}(s_0) \le \epsilon;$$

Key ingredients:

Confidence Ball construction + Simulation lemma

Summary for Today:

1. A model-based **Algorithm** under generative model:

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

Summary for Today:

1. A model-based **Algorithm** under generative model:

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

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

Summary for Today:

1. A model-based **Algorithm** under generative model:

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

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

3. **Analysis**: W/ simulation lemma, we achieve ϵ-near optimality w/ # of samples $\widetilde{O}\left(\frac{S^2 A}{\epsilon^2(1-\gamma)^4}\right)$ (improvement is possible, but out of scope)