Learning in Generative Model

Wen Sun

CS 6789: Foundations of Reinforcement Learning

Today:

Today:

- Recap: computational complexity
  - Question: Given an MDP $\mathscr{M} = (S, A, P, r, \gamma)$ can we exactly compute $Q^\star$ (or find $\pi^\star$) in polynomial time?

Today:

- Recap: computational complexity
  - Question: Given an MDP $\mathcal{M} = (S, A, P, r, \gamma)$ can we exactly compute $Q^\star$ (or find $\pi^\star$) in polynomial time?

- Today: statistical complexity
  - Question: Given only sampling access to an unknown MDP $\mathcal{M} = (S, A, P, r, \gamma)$ how many observed transitions do we need to estimate $Q^\star$ (or find $\pi^\star$)?

Today:

- Recap: computational complexity
  - Question: Given an MDP $\mathscr{M} = (S, A, P, r, \gamma)$ can we exactly compute $Q^\star$ (or find $\pi^\star$) in polynomial time?

- Today: statistical complexity
  - Question: Given only sampling access to an unknown MDP $\mathscr{M} = (S, A, P, r, \gamma)$ how many observed transitions do we need to estimate $Q^\star$ (or find $\pi^\star$)?

Two natural models for learning in an unknown MDP

- Episodic setting:
  - in every episode, $s_0 \sim \mu$.
  - the learner acts for some finite number of steps and observes the trajectory.
  - The state is then resets to $s_0 \sim \mu$.

Two natural models for learning in an unknown MDP

- Episodic setting:
  - in every episode, $s_0 \sim \mu$.
  - the learner acts for some finite number of steps and observes the trajectory.
  - The state is then resets to $s_0 \sim \mu$.
- Generative model setting:
  - input: (s, a)
  - output: a sample $s' \sim P(\cdot \mid s, a)$ and $r(s, a)$

Two natural models for learning in an unknown MDP

- Episodic setting:
  - in every episode, $s_0 \sim \mu$.
  - the learner acts for some finite number of steps and observes the trajectory.
  - The state is then resets to $s_0 \sim \mu$.
- Generative model setting:
  - input: (s, a)
  - output: a sample $s' \sim P(\cdot \mid s, a)$ and $r(s, a)$

- Sample complexity of RL:
  how many transitions do we need observe in order to find a near optimal policy?
  - Episodic setting: we must actively explore to gather information
  - Generative model setting: lets us disentangle the issue of fundamental statistical limits from exploration.

How many samples do we need to learn?

- What is the minmax optimal sample complexity, with generative modeling access? (using *any* algorithm)
  - Since P has S^2A parameters, we may hope that $O(S^2A)$ samples are sufficient for learning.

How many samples do we need to learn?

- What is the minmax optimal sample complexity, with generative modeling access? (using *any* algorithm)
  - Since P has S^2A parameters, we may hope that $O(S^2A)$ samples are sufficient for learning.
- Questions:
  - Is a naive model-based approach optimal?
    i.e. estimate P accurately (using $O(S^2A)$ samples) and then use $\widehat{P}$ for planning.
  - Is sublinear learning possible?
    (i.e. learn with fewer than $\Omega(S^2A)$ samples)

How many samples do we need to learn?

- What is the minmax optimal sample complexity, with generative modeling access? (using *any* algorithm)
  - Since P has S^2A parameters, we may hope that $O(S^2A)$ samples are sufficient for learning.
- Questions:
  - Is a naive model-based approach optimal?
    i.e. estimate P accurately (using $O(S^2A)$ samples) and then use $\widehat{P}$ for planning.
  - Is sublinear learning possible?
    (i.e. learn with fewer than $\Omega(S^2A)$ samples)
- If sublinear learning is possible, then we do not need an accurate model of the world in order to act near-optimally?

The most naive approach: model based

- Today: let us assume access to a generative model

The most naive approach: model based

- Today: let us assume access to a generative model
- most naive approach to learning:
  - Call our simulator N times at each state action pair.
  - Let $\widehat{P}$ be our empirical model:

$$\widehat{P}(s'|s,a) = \frac{\text{count}(s',s,a)}{N}$$

    where $\text{count}(s',s,a)$ is the #times (s,a) transitions to state s'.
  - we also know the rewards after one call.

    (for simplicity, we often assume $r(s,a)$ is determinstic)
  - Compute $\widehat{\pi}^\star = \text{PI}(\widehat{P}, r)$

The most naive approach: model based

- Today: let us assume access to a generative model
- most naive approach to learning:
  - Call our simulator N times at each state action pair.
  - Let $\widehat{P}$ be our empirical model:

    $$\widehat{P}(s'|s,a) = \frac{\text{count}(s', s, a)}{N}$$

    where $\text{count}(s', s, a)$ is the #times (s, a) transitions to state s'.
  - we also know the rewards after one call.

    (for simplicity, we often assume $r(s, a)$ is determinstic)
  - Compute $\widehat{\pi}^\star = \text{PI}(\widehat{P}, r)$
- The total number of calls to our generative model is SAN.

Attempt 1:

the naive model based approach

Model accuracy

Proposition: With probability greater than $1 - \delta$,

Model accuracy

Proposition: With probability greater than $1 - \delta$,

- Model accuracy: The transition model is ϵ has error bounded as:

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

$P(\cdot|s,a) = [\ ]_{|S|}$

Model accuracy

Proposition: With probability greater than $1 - \delta$,

- Model accuracy: The transition model is ϵ has error bounded as:

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

Model accuracy

Proposition: With probability greater than $1 - \delta$,

- Model accuracy: The transition model is ϵ has error bounded as:

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

(HW1 will walk you through the proof)

“Simulation” Lemma

Given policy π, does $P \approx \widehat{P}$ imply $V^\pi \approx V^\pi_{\widehat{P}}$?

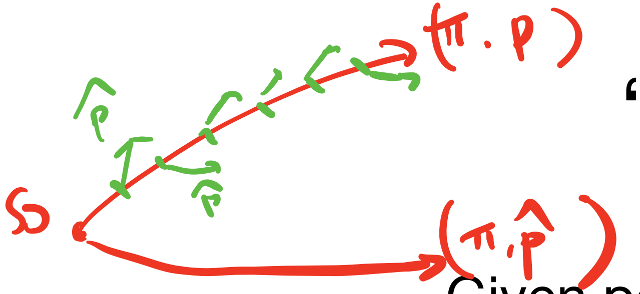
Proposition

"Simulation" Lemma

Given policy π, does $P \approx \widehat{P}$ imply $V^\pi \approx V^\pi_{\widehat{P}}$?

Proposition

- Given any two transitions P and $\widehat{P}$, and any policy π, we have:



"Simulation" Lemma

Given policy π, does $P \approx \widehat{P}$ imply $V^{\pi} \approx V^{\pi}_{\widehat{P}}$?

Proposition

- Given any two transitions P and $\widehat{P}$, and any policy π, we have:

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

"Simulation" Lemma

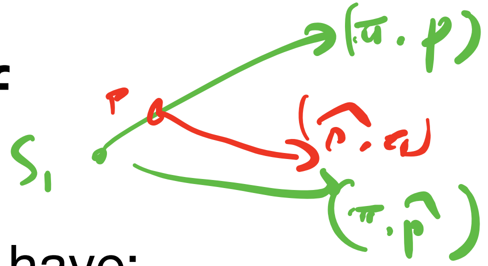
Given policy π, does $P \approx \widehat{P}$ imply $V^\pi \approx V^\pi_{\widehat{P}}$?

Proposition

- Given any two transitions P and $\widehat{P}$, and any policy π, we have:

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

“Simulation” Lemma: proof



Proposition

- Given any two transitions P and $\widehat{P}$, and any policy π, we have:

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

$$V^\pi(s_0) - V^\pi_{\widehat{P}}(s_0) = Q^\pi(s_0, \pi(s_0)) - Q^\pi_{\widehat{P}}(s_0, \pi(s_0))$$

$$= \gamma \mathbb{E}_{s_1\sim P(s_0,\pi(s_0))} V^\pi(s_1) - \gamma \mathbb{E}_{s_1\sim \widehat{P}(s_0,\pi(s_0))} V^\pi_{\widehat{P}}(s_1)$$

$$= \gamma \mathbb{E}_{s_1\sim P(s_0,\pi(s_0))} V^\pi(s_1) - \gamma \mathbb{E}_{s_1\sim P} V^\pi_{\widehat{P}}(s_1) + \gamma \mathbb{E}_{s_1\sim P} V^\pi_{\widehat{P}}(s_1) - \gamma \mathbb{E}_{s_1\sim \widehat{P}} V^\pi_{\widehat{P}}(s_1)$$

$$\Rightarrow \gamma \left(P(s_0, \pi(s_0)) - \widehat{P}(s_0, \pi(s_0))\right)^\top V^\pi_{\widehat{P}}$$

Combine Model Accuracy and Simulation

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

Combine Model Accuracy and Simulation

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

$$= V^\star - V^{\pi^\star}_{\widehat{P}} + V^{\pi^\star}_{\widehat{P}} - V^{\widehat{\pi}^\star} + V^{\widehat{\pi}^\star}_{\widehat{P}} - V^{\widehat{\pi}^\star}_{\widehat{P}}$$

Combine Model Accuracy and Simulation

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

$$= V^\star - V^{\pi^\star}_{\widehat{P}} + V^{\pi^\star}_{\widehat{P}} - V^{\widehat{\pi}^\star} + V^{\widehat{\pi}^\star}_{\widehat{P}} - V^{\widehat{\pi}^\star}_{\widehat{P}}$$

$$\leq V^\star - V^{\pi^\star}_{\widehat{P}} - V^{\widehat{\pi}^\star} + V^{\widehat{\pi}^\star}_{\widehat{P}}$$

Combine Model Accuracy and Simulation

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

$$= V^\star - V^{\pi^\star}_{\widehat{P}} + V^{\pi^\star}_{\widehat{P}} - V^{\widehat{\pi}^\star} + V^{\widehat{\pi}^\star}_{\widehat{P}} - V^{\widehat{\pi}^\star}_{\widehat{P}}$$

$$\leq V^\star - V^{\pi^\star}_{\widehat{P}} - V^{\widehat{\pi}^\star} + V^{\widehat{\pi}^\star}_{\widehat{P}} \quad \lesssim \frac{1}{(1-\gamma)^2}\sqrt{\frac{S\ln(SA/\delta)}{N}}$$

Combine Model Accuracy and Simulation

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

$$= V^\star - V^{\pi^\star}_{\widehat{P}} + V^{\pi^\star}_{\widehat{P}} - V^{\widehat{\pi}^\star} + V^{\widehat{\pi}^\star}_{\widehat{P}} - V^{\widehat{\pi}^\star}_{\widehat{P}}$$

$$\leq V^\star - V^{\pi^\star}_{\widehat{P}} - V^{\widehat{\pi}^\star} + V^{\widehat{\pi}^\star}_{\widehat{P}} \quad \lesssim \frac{1}{(1-\gamma)^2}\sqrt{\frac{S\ln(SA/\delta)}{N}}$$

Set it to ϵ, solve for N

Conclusion on the naive approach

Given δ. If we draw $\frac{S^2 A}{\epsilon^2(1-\gamma)^4} \cdot \ln \frac{SA}{\delta}$ many total samples, w/ prob at at least $1-\delta$, we have $V^{\pi^\star}(s) - V^{\widehat{\pi}^\star}(s) \leq \epsilon, \forall s$

Conclusion on the naive approach

Given δ. If we draw $\frac{S^2 A}{\epsilon^2(1-\gamma)^4} \cdot \ln \frac{SA}{\delta}$ many total samples, w/ prob at at least $1-\delta$, we have $V^{\pi^\star}(s) - V^{\widehat{\pi}^\star}(s) \leq \epsilon, \forall s$

Q: can we do better than a linear scaling in $S^2 A$?

Attempt 2:

obtaining sublinear sample complexity

idea: use concentration only on $V^{\star}$

Model error projected on $V^\star$

Model error projected on $V^\star$

- Recall $\|V^\star\|_\infty \leq 1/(1-\gamma)$.

Model error projected on $V^\star$

- Recall $\|V^\star\|_\infty \leq 1/(1-\gamma)$.
- By Hoeffding's inequality and the union bound,

$$\max_{s,a} \left| E_{s' \sim P(\cdot|s,a)}[V^\star(s')] - E_{s' \sim \widehat{P}(\cdot|s,a)}[V^\star(s')] \right|$$

$$\leq O\left(\frac{1}{1-\gamma} \sqrt{\frac{\log(SA/\delta)}{N}} \right)$$

which holds with probability greater than $1-\delta$.

Bounding error $\|Q^\star - \widehat{Q}^\star\|_\infty$

Recall $Q^\star$ the optimal Q function in P, $\widehat{Q}^\star$ the optimal Q in $\widehat{P}$, we have (proof next slide)

$$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \frac{1}{(1-\gamma)^2}\sqrt{\frac{2\log(2SA/\delta)}{N}}$$

Bounding error $\|Q^\star - \widehat{Q}^\star\|_\infty$

Recall $Q^\star$ the optimal Q function in P, $\widehat{Q}^\star$ the optimal Q in $\widehat{P}$, we have (proof next slide)

$$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \frac{1}{(1-\gamma)^2}\sqrt{\frac{2\log(2SA/\delta)}{N}}$$

Recall $\widehat{\pi}^\star(s) = \arg\max_a \widehat{Q}^\star(s,a)$, we immediately have (recall VI analysis):

$$\|V^\star - V^{\widehat{\pi}^\star}\|_\infty \leq \frac{1}{(1-\gamma)^3}\sqrt{\frac{\ln(SA/\delta)}{N}}$$

Bounding error $\|V^\star - \widehat{V}^\star\|_\infty$

$$V^\star(s_0) - \widehat{V}^\star(s_0) = Q^\star(s_0, \pi^\star(s_0)) - \widehat{Q}^\star(s_0, \hat{\pi}^\star(s_0))$$

Bounding error $\|V^\star - \widehat{V}^\star\|_\infty$

$$V^\star(s_0) - \widehat{V}^\star(s_0) = Q^\star(s_0, \pi^\star(s_0)) - \widehat{Q}^\star(s_0, \hat{\pi}^\star(s_0))$$

$$\leq Q^\star(s_0, \pi^\star(s_0)) - \widehat{Q}^\star(s_0, \pi^\star(s_0))$$

Bounding error $\|V^\star - \widehat{V}^\star\|_\infty$

$$V^\star(s_0) - \widehat{V}^\star(s_0) = Q^\star(s_0, \pi^\star(s_0)) - \widehat{Q}^\star(s_0, \hat{\pi}^\star(s_0))$$

$$\leq Q^\star(s_0, \pi^\star(s_0)) - \widehat{Q}^\star(s_0, \pi^\star(s_0)) \ = \gamma \mathbb{E}_{s_1 \sim P(s_0, \pi^\star(s_0))} V^\star(s_1) - \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} \widehat{V}^\star(s_1)$$

Bounding error $\|V^\star - \widehat{V}^\star\|_\infty$

$$V^\star(s_0) - \widehat{V}^\star(s_0) = Q^\star(s_0, \pi^\star(s_0)) - \widehat{Q}^\star(s_0, \hat{\pi}^\star(s_0))$$

$$\leq Q^\star(s_0, \pi^\star(s_0)) - \widehat{Q}^\star(s_0, \pi^\star(s_0)) \ = \gamma \mathbb{E}_{s_1 \sim P(s_0, \pi^\star(s_0))} V^\star(s_1) - \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} \widehat{V}^\star(s_1)$$

$$= \gamma \mathbb{E}_{s_1 \sim P(s_0, \pi^\star(s_0))} V^\star(s_1) - \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} V^\star(s_1) + \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} V^\star(s_1) - \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} \widehat{V}^\star(s_1)$$

Bounding error $\|V^\star - \widehat{V}^\star\|_\infty$

$$V^\star(s_0) - \widehat{V}^\star(s_0) = Q^\star(s_0, \pi^\star(s_0)) - \widehat{Q}^\star(s_0, \hat{\pi}^\star(s_0))$$

$$\leq Q^\star(s_0, \pi^\star(s_0)) - \widehat{Q}^\star(s_0, \pi^\star(s_0)) \ = \gamma \mathbb{E}_{s_1 \sim P(s_0, \pi^\star(s_0))} V^\star(s_1) - \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} \widehat{V}^\star(s_1)$$

$$= \gamma \mathbb{E}_{s_1 \sim P(s_0, \pi^\star(s_0))} V^\star(s_1) - \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} V^\star(s_1) + \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} V^\star(s_1) - \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} \widehat{V}^\star(s_1)$$

$$= \gamma \left(P(s_0, \pi^\star(s_0)) - \widehat{P}(s_0, \pi^\star(s_0)) \right)^\top V^\star + \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} \left(V^\star(s_1) - \widehat{V}^\star(s_1) \right)$$

Bounding error $\|V^\star - \widehat{V}^\star\|_\infty$

$$V^\star(s_0) - \widehat{V}^\star(s_0) = Q^\star(s_0, \pi^\star(s_0)) - \widehat{Q}^\star(s_0, \hat{\pi}^\star(s_0))$$

$$\leq Q^\star(s_0, \pi^\star(s_0)) - \widehat{Q}^\star(s_0, \pi^\star(s_0)) \ = \gamma \mathbb{E}_{s_1 \sim P(s_0, \pi^\star(s_0))} V^\star(s_1) - \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} \widehat{V}^\star(s_1)$$

$$= \gamma \mathbb{E}_{s_1 \sim P(s_0, \pi^\star(s_0))} V^\star(s_1) - \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} V^\star(s_1) + \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} V^\star(s_1) - \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} \widehat{V}^\star(s_1)$$

$$= \gamma \left(P(s_0, \pi^\star(s_0)) - \widehat{P}(s_0, \pi^\star(s_0)) \right)^\top V^\star + \gamma \mathbb{E}_{s_1 \sim \widehat{P}(s_0, \pi^\star(s_0))} \left(V^\star(s_1) - \widehat{V}^\star(s_1) \right)$$

$$\leq \frac{\gamma}{1-\gamma} \mathbb{E}_{s,a \sim d^{\pi^\star}_{\widehat{P}}} \left(P(s,a) - \widehat{P}(s,a) \right)^\top V^\star$$