Recap++

Two Fundamental Questions in Sample Complexity

Sampling models: episodic, offline, gen models

Generative model setting: input: (s, a) output: a sample $s' \sim P(\cdot \mid s, a)$ and $r(s, a)$

1. The Model Size Question (Sublinear Learning)
   - Q1: Can we find an ϵ-optimal policy with **sublinear** sample complexity?

2. The Horizon Question (Error Amplification)
   - Q2: What is the "**horizon amplification**"? i.e. the dependence on $1/(1-\gamma)$

Attempt 1:

the naive model based approach

Matrix Expressions

- View P as a matrix of size $SA \times S$ (rows are probability distributions)
- Define P^π to be the transition matrix on state-action pairs (for deterministic π):

$$P^\pi_{(s,a),(s',a')} := \begin{cases} P(s' \mid s, a) & \text{if } a' = \pi(s') \\ 0 & \text{if } a' \neq \pi(s') \end{cases}$$

- With this notation,

$$Q^\pi = r + \gamma P V^\pi$$

$$Q^\pi = r + \gamma P^\pi Q^\pi$$

- Also,

$$Q^\pi = (I - \gamma P^\pi)^{-1} r$$

(where one can show the inverse exists)

Model accuracy

Proposition: c is an absolute constant. $\epsilon > 0$. For $N \geq \frac{c\gamma}{(1-\gamma)^4} \frac{S \log(cSA/\delta)}{\epsilon^2}$

and with probability greater than $1 - \delta$,

- Model accuracy: The transition model is ϵ has error bounded as:
  $\max_{s,a} \|P(\cdot \mid s, a) - \widehat{P}(\cdot \mid s, a)\|_1 \leq (1-\gamma)^2\epsilon/2$.
- Uniform value accuracy: For all policies π,
  $\|Q^\pi - \widehat{Q}^\pi\|_\infty \leq \epsilon/2$
- Near optimal planning: Suppose that $\hat{\pi}^\star$ is the optimal policy in $\widehat{M}$.
  $\|Q^\star - Q^{\hat{\pi}^\star}\|_\infty \leq \epsilon$

Attempt 2:

obtaining sublinear sample complexity

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

Sample Size Corollaries

Corollary: for $\epsilon < 1$, provided $N \geq \frac{c}{(1-\gamma)^4} \frac{\log(cSA/\delta)}{\epsilon^2}$ then

$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \epsilon$ (with prob. greater than $1-\delta$)

What about the policy?

Corollary: for $\epsilon < 1$, provided $N \geq \frac{c}{(1-\gamma)^6} \frac{\log(cSA/\delta)}{\epsilon^2}$ then

$\|Q^\star - Q^{\widehat{\pi}\star}\|_\infty \leq \epsilon$ (with prob. greater than $1-\delta$)

Component-wise Bounds Lemma

Lemma: we have that

$$Q^\star - \widehat{Q}^\star \leq \gamma(I - \gamma \widehat{P}^{\pi^\star})^{-1}(P - \widehat{P})V^\star$$

$$Q^\star - \widehat{Q}^\star \geq \gamma(I - \gamma \widehat{P}^{\widehat{\pi}^\star})^{-1}(P - \widehat{P})V^\star$$

Proof:

For the first claim, the optimality of $\pi^\star$ in M implies:

$$Q^\star - \widehat{Q}^\star = Q^{\pi^\star} - \widehat{Q}^{\widehat{\pi}^\star} \leq Q^{\pi^\star} - \widehat{Q}^{\pi^\star} = \gamma(I - \gamma \widehat{P}^{\pi^\star})^{-1}(P - \widehat{P})V^\star,$$

using the simulation lemma in the final step.

See notes for the proof of second claim.

Attempt 3:

minimax optimal sample complexity

idea: better variance control

Revisiting proof attempt 2: where is there slop?

- Proof of the first claim:
  - By comp. lemma: $\|Q^\star - \widehat{Q}^\star\|_\infty \leq \frac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\star\|_\infty$
  - Recall $\|V^\star\|_\infty \leq 1/(1-\gamma)$.
  - By Hoeffding's inequality and the union bound,

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

$$\leq \frac{1}{1-\gamma}\sqrt{\frac{2\log(2SA/\delta)}{N}}$$

  which holds with probability greater than $1-\delta$.
- Proof of second claim is similar (see the book)

Minimax Optimal Sample Complexity

Corollary: for $\epsilon < 1$, provided $N \geq \dfrac{c}{(1-\gamma)^3}\dfrac{\log(cSA/\delta)}{\epsilon^2}$ then

$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \epsilon$ (with prob. greater than $1-\delta$)

What about the policy?

Naively, we need $N/(1-\gamma)^2$ more samples.

A different thm gives: With the same N,

$\|Q^\star - Q^{\widehat{\pi}^\star}\|_\infty \leq \epsilon$ (with prob. greater than $1-\delta$)

Lower Bound: We can't do better.

Proof sketch: part 1

- From Bernstein's ineq, with pr. greater than $1-\delta$, we have (component-wise):

$$|(P - \widehat{P})V^\star| \le \sqrt{\frac{2\log(2SA/\delta)}{N}}\sqrt{\mathrm{Var}_P(V^\star)} + \frac{1}{1-\gamma}\frac{2\log(2SA/\delta)}{3N}\vec{1}$$

- How to use this: again from "Component-wise Bounds" lemma,

$$Q^\star - \widehat{Q}^\star \le \gamma \| (I - \gamma \widehat{P}^{\pi^\star})^{-1}(P - \widehat{P})V^\star \|_\infty \le ??$$

- Therefore

$$Q^\star - \widehat{Q}^\star \le \gamma \sqrt{\frac{2\log(2SA/\delta)}{N}} \| (I - \gamma \widehat{P}^{\pi^\star})^{-1}\sqrt{\mathrm{Var}_P(V^\star)} \|_\infty$$

$+$"lower order term"

Bellman Equation for the (total) Variance

- Variance: $\mathrm{Var}_P(V)(s,a) := \mathrm{Var}_{P(\cdot|s,a)}(V)$
  Component wise variance: $\mathrm{Var}_P(V) := P(V)^2 - (PV)^2$

- Let's keep around the MDP M subscripts.
  Define Σ_M^π as the (total) variance of the discounted reward:

$$\Sigma_M^\pi(s,a) := E\left[\left(\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) - Q_M^\pi(s,a)\right)^2 \middle| s_0 = s, a_0 = a\right]$$

- Bellman equation for the total variance:
  $\Sigma_M^\pi = \gamma^2 \mathrm{Var}_P(V_M^\pi) + \gamma^2 P^\pi \Sigma_M^\pi$