The Sample Complexity (with a Generative Model)

Sham Kakade and Wen Sun

CS 6789: Foundations of Reinforcement Learning

Announcements

- Reading assignments (see website)
  - sign up for a chapter (signup sheep will be up today)
  - start the assignment only after the we approve the chapter.
  - requirements:
    - one page report that summarizes the chapter
    - check all mathematical steps in the chapter
- Participation/effort Bonus
  - we will give extra credit for participation (class, ED, etc)
  - extra credit for reading assignments, finding bugs, project…
- The book will be updated often.
  - Feedback/questions/finding typos appreciated!

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 $\mathscr{M} = (S, A, P, r, \gamma)$ 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 sampling models: episodic setting and generative models.

Recap

Summary Table

	Value Iteration	Policy Iteration	LP-based Algorithms
Poly.	$S^2 A \frac{L(P,r,\gamma)\ \log \frac{1}{1-\gamma}}{1-\gamma}$	$(S^3 + S^2 A) \frac{L(P,r,\gamma) \log \frac{1}{1-\gamma}}{1-\gamma}$	$S^3 A\, L(P, r, \gamma)$
Strongly Poly.	X	$(S^3 + S^2 A) \cdot \min \left\{ \frac{A^S}{S}, \frac{S^2 A \log \frac{S^2}{1-\gamma}}{1-\gamma} \right\}$	$S^4 A^4 \log \frac{S}{1-\gamma}$

- VI: poly time for fixed γ, not strongly poly
- PI: poly and strongly-poly time for fixed γ
- LP approach: poly and strongly-poly time

(LP approach is only logarithmic in $1/(1-\gamma)$)

Today

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)

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:

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

    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)
- The total number of calls to our generative model is SAN.

Attempt 1:

the naive model based approach

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

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:
  $\max_{s,a} \|P(\cdot\,|\,s,a) - \widehat{P}(\cdot\,|\,s,a)\|_1 \leq (1-\gamma)^2\epsilon/2\,.$

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,
  $\|Q^\pi - \widehat{Q}^\pi\|_\infty \leq \epsilon / 2$

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 $\widehat{M}$.
  $\|Q^\star - Q^{\hat{\pi}^\star}\|_\infty \leq \epsilon$

Matrix Expressions

- Define P^π to be the transition matrix on state-action pairs (for deterministic π):

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

Matrix Expressions

- 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,

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

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

Matrix Expressions

- 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)

“Simulation” Lemma

“Simulation Lemma”: For all π,

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

"Simulation" Lemma

"Simulation Lemma": For all π,

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

Proof: Using our matrix equality for Q^π, we have:

$$\begin{aligned}
Q^\pi - \widehat{Q}^\pi &= Q^\pi - (I - \gamma \widehat{P}^\pi)^{-1} r \\
&= (I - \gamma \widehat{P}^\pi)^{-1}((I - \gamma \widehat{P}^\pi) - (I - \gamma P^\pi))Q^\pi \\
&= \gamma(I - \gamma \widehat{P}^\pi)^{-1}(P^\pi - \widehat{P}^\pi)Q^\pi \\
&= \gamma(I - \gamma \widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi
\end{aligned}$$

Proof of Claim 1

Proof of Claim 1

- Concentration of a distribution in the ℓ_1 norm:
  - For a fixed s, a. With pr greater than $1 - \delta$,

    $$\|P(\cdot \mid s, a) - \widehat{P}(\cdot \mid s, a)\|_1 \leq c\sqrt{\frac{S \log(1/\delta)}{N}}$$

    with N samples used to estimate $\widehat{P}(\cdot \mid s, a)$.

Proof of Claim 1

- Concentration of a distribution in the ℓ_1 norm:
  - For a fixed s, a. With pr greater than $1 - \delta$,

    $$\|P(\cdot \mid s, a) - \widehat{P}(\cdot \mid s, a)\|_1 \leq c\sqrt{\frac{S\log(1/\delta)}{N}}$$

    with N samples used to estimate $\widehat{P}(\cdot \mid s, a)$.
- The first claim now follows by the union bound.

Proof of Claim 2 (&3)

Proof of Claim 2 (&3)

For the second claim,

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

$$\leq \frac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\pi\|_\infty$$

$$\leq \frac{\gamma}{1-\gamma}\left(\max_{s,a}\|P(\cdot\,|\,s,a) - \widehat{P}(\cdot\,|\,s,a)\|_1\right)\|V^\pi\|_\infty$$

$$\leq \frac{\gamma}{(1-\gamma)^2}\max_{s,a}\|P(\cdot\,|\,s,a) - \widehat{P}(\cdot\,|\,s,a)\|_1$$

(why is the first inequality true?)

Proof of Claim 2 (&3)

For the second claim,

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

$$\leq \frac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\pi\|_\infty$$

$$\leq \frac{\gamma}{1-\gamma}\left(\max_{s,a}\|P(\cdot\,|\,s,a) - \widehat{P}(\cdot\,|\,s,a)\|_1\right)\|V^\pi\|_\infty$$

$$\leq \frac{\gamma}{(1-\gamma)^2}\max_{s,a}\|P(\cdot\,|\,s,a) - \widehat{P}(\cdot\,|\,s,a)\|_1$$

(why is the first inequality true?)

The proof for the Claim 3 immediately follows from the second claim.

Attempt 2:

obtaining sublinear sample complexity

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

Reference sheet (defs/notation)

Reference sheet (defs/notation)

- Remember: # samples from generative model $= SAN$

Reference sheet (defs/notation)

- Remember: # samples from generative model $= SAN$
- P^π is the transition matrix on state-action pairs for a deterministic policy π:

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

Reference sheet (defs/notation)

- Remember: # samples from generative model $= SAN$
- P^π is the transition matrix on state-action pairs for a deterministic policy π:

$$P^\pi_{(s,a),(s',a')} := P(s' \mid s, a) \qquad \text{if } a' = \pi(s')$$

$$0 \qquad \text{if } a' \neq \pi(s')$$

- With this notation,

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

Reference sheet (defs/notation)

- Remember: # samples from generative model $= SAN$
- P^π is the transition matrix on state-action pairs for a deterministic policy 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,

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

- $\dfrac{1}{1-\gamma}(I - \gamma P^\pi)^{-1}$ is a matrix whose rows are probability distributions (why?)

Reference sheet (defs/notation)

- Remember: # samples from generative model $= SAN$
- P^π is the transition matrix on state-action pairs for a deterministic policy 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,

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

- $\dfrac{1}{1-\gamma}(I - \gamma P^\pi)^{-1}$ is a matrix whose rows are probability distributions (why?)

- $\widehat{Q}^\star$: optimal value in estimated model $\widehat{M}$.
  $\widehat{\pi}^\star$: optimal policy in $\widehat{M}$.
  $Q^{\widehat{\pi}^\star}$: (true) value of estimated policy.

Attempt 2: Sublinear Sample Complexity

Attempt 2: Sublinear Sample Complexity

Proposition: (Crude Value Bound) With probability greater than $1 - \delta$,

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

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

Attempt 2: Sublinear Sample Complexity

Proposition: (Crude Value Bound) With probability greater than $1-\delta$,

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

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

What about the value of the policy?

$$\|Q^\star - Q^{\hat{\pi}^\star}\|_\infty \le \frac{\gamma}{(1-\gamma)^3}\sqrt{\frac{2\log(2SA/\delta)}{N}}$$

Component-wise Bounds Lemma

Lemma: we have that

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

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

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.

Proof: (& key idea for sublinearity)

Proof: (& key idea for sublinearity)

- Proof of the first claim:

Proof: (& key idea for sublinearity)

- Proof of the first claim:
  - By the previous lemma: $\|Q^\star - \widehat{Q}^\star\|_\infty \leq \frac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\star\|_\infty$

Proof: (& key idea for sublinearity)

- Proof of the first claim:
  - By the previous lemma: $\|Q^\star - \widehat{Q}^\star\|_\infty \leq \frac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\star\|_\infty$
  - Recall $\|V^\star\|_\infty \leq 1/(1-\gamma)$.

Proof: (& key idea for sublinearity)

- Proof of the first claim:
  - By the previous lemma: $\|Q^\star - \widehat{Q}^\star\|_\infty \leq \dfrac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\star\|_\infty$
  - 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: (& key idea for sublinearity)

- Proof of the first claim:
  - By the previous lemma: $\|Q^\star - \widehat{Q}^\star\|_\infty \le \frac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\star\|_\infty$
  - Recall $\|V^\star\|_\infty \le 1/(1-\gamma)$.
  - By Hoeffding's inequality and the union bound,
    $$\|(P - \widehat{P})V^\star\|_\infty = \max_{s,a}\left| E_{s'\sim P(\cdot|s,a)}[V^\star(s')] - E_{s'\sim \widehat{P}(\cdot|s,a)}[V^\star(s')]\right|$$
    $$\le \frac{1}{1-\gamma}\sqrt{\frac{2\log(2SA/\delta)}{N}}$$
    which holds with probability greater than $1-\delta$.
- Proof of second claim is similar (see the book)

Attempt 3:

minimax optimal sample complexity

idea: better variance control

("near") Minimax Optimal Sample Complexity

Theorem: (Azar et al. '13) With probability greater than $1-\delta$,

$$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \gamma\sqrt{\frac{c}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{N}} + \frac{c\gamma}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{N},$$

where c is an absolute constant.

("near") Minimax Optimal Sample Complexity

Theorem: (Azar et al. '13) With probability greater than $1-\delta$,

$$\|Q^\star - \widehat{Q}^\star\|_\infty \le \gamma\sqrt{\frac{c}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{N}} + \frac{c\gamma}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{N},$$

where c is an absolute constant.

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

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

("near") Minimax Optimal Sample Complexity

Theorem: (Azar et al. '13) With probability greater than $1-\delta$,

$$\|Q^\star - \widehat{Q}^\star\|_\infty \le \gamma \sqrt{\frac{c}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{N}} + \frac{c\gamma}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{N},$$

where c is an absolute constant.

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

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

Corollary: What about the policy? Naively, need $N/(1-\gamma)^2$ more samples.
We pay *another factor of* $1/(1-\gamma)^2$ *samples. Is this real?*

Minimax Optimal Sample Complexity
(on the policy)

Minimax Optimal Sample Complexity (on the policy)

Theorem: (Agarwal et al. '20) For $\epsilon < \sqrt{1/(1-\gamma)}$, provided

$N \geq \frac{c}{(1-\gamma)^3} \frac{\log(cSA/\delta)}{\epsilon^2}$ then with prob. greater than $1-\delta$),

$$\|Q^\star - Q^{\hat{\pi}^\star}\|_\infty \leq \epsilon$$

Minimax Optimal Sample Complexity (on the policy)

Theorem: (Agarwal et al. '20) For $\epsilon < \sqrt{1/(1-\gamma)}$, provided

$N \geq \frac{c}{(1-\gamma)^3} \frac{\log(cSA/\delta)}{\epsilon^2}$ then with prob. greater than $1-\delta$),

$$\|Q^\star - Q^{\widehat{\pi}^\star}\|_\infty \leq \epsilon$$

Lower Bound: We can't do better.

Proof sketch: part 1

- From "Component-wise Bounds" lemma, we want to bound:
$Q^\star - \widehat{Q}^\star \leq \gamma \| (I - \gamma \widehat{P}^{\pi^\star})^{-1} (P - \widehat{P}) V^\star \|_\infty \leq ??$

Proof sketch: part 1

- From "Component-wise Bounds" lemma, we want to bound:
$$Q^\star - \widehat{Q}^\star \leq \gamma \| (I - \gamma \widehat{P}^{\pi^\star})^{-1} (P - \widehat{P}) V^\star \|_\infty \leq ??$$

- From Bernstein's ineq, with pr. greater than $1-\delta$, we have (component-wise):
$$|(P - \widehat{P})V^\star| \leq \sqrt{\frac{2\log(2SA/\delta)}{N}} \sqrt{\mathrm{Var}_P(V^\star)} + \frac{1}{1-\gamma} \frac{2\log(2SA/\delta)}{3N} \vec{1}$$

Proof sketch: part 1

- From "Component-wise Bounds" lemma, we want to bound:
  $$Q^\star - \widehat{Q}^\star \leq \gamma \| (I - \gamma \widehat{P}^{\pi^\star})^{-1} (P - \widehat{P}) V^\star \|_\infty \leq ??$$

- From Bernstein's ineq, with pr. greater than $1 - \delta$, we have (component-wise):
  $$| (P - \widehat{P}) V^\star | \leq \sqrt{\frac{2\log(2SA/\delta)}{N}} \sqrt{\mathrm{Var}_P(V^\star)} + \frac{1}{1-\gamma} \frac{2\log(2SA/\delta)}{3N} \vec{1}$$

- Therefore
  $$Q^\star - \widehat{Q}^\star \leq \gamma \sqrt{\frac{2\log(2SA/\delta)}{N}} \| (I - \gamma \widehat{P}^{\pi^\star})^{-1} \sqrt{\mathrm{Var}_P(V^\star)} \|_\infty$$
  $+$"lower order term"

Bellman Equation for the Variance

- Variance: $\mathrm{Var}_P(V)(s,a) := \mathrm{Var}_{P(\cdot|s,a)}(V)$

  Component wise variance: $\mathrm{Var}_P(V) := P(V)^2 - (PV)^2$

Bellman Equation for the 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 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^π_M as the (total) variance of the discounted reward:

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

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

Key Lemma

Lemma: For any policy π and MDP M,

$$\left\| (I - \gamma P^{\pi})^{-1} \sqrt{\mathrm{Var}_P(V_M^{\pi})} \right\|_{\infty} \leq \sqrt{\frac{2}{(1-\gamma)^3}}$$

Proof idea: convexity + Bellman equations for the variance.

Putting it all together

Proof sketch: we have two MDPs M and $\widehat{M}$. need to bound:

Putting it all together

Proof sketch: we have two MDPs M and $\widehat{M}$. need to bound:

$$\|(I-\gamma \widehat{P}^{\pi^\star})^{-1}\sqrt{\mathrm{Var}_P(V^\star)}\|_\infty = \|(I-\gamma P^{\pi^\star}_{\widehat{M}})^{-1}\sqrt{\mathrm{Var}_P(V^{\pi^\star}_{M})}\|_\infty$$

$$\leq \|(I-\gamma P^{\pi^\star}_{\widehat{M}})^{-1}\sqrt{\mathrm{Var}_P(V^{\pi^\star}_{\widehat{M}})} + \text{"lower order"}$$

$$\leq \sqrt{\frac{2}{(1-\gamma)^3}} + \text{"lower order"}$$

Putting it all together

Proof sketch: we have two MDPs M and $\widehat{M}$. need to bound:

$$\|(I - \gamma \widehat{P}^{\pi^\star})^{-1}\sqrt{\mathrm{Var}_P(V^\star)}\|_\infty = \|(I - \gamma P^{\pi^\star}_{\widehat{M}})^{-1}\sqrt{\mathrm{Var}_P(V^{\pi^\star}_{M})}\|_\infty$$

$$\leq \|(I - \gamma P^{\pi^\star}_{\widehat{M}})^{-1}\sqrt{\mathrm{Var}_P(V^{\pi^\star}_{\widehat{M}})} + \text{"lower order"}$$

$$\leq \sqrt{\frac{2}{(1-\gamma)^3}} + \text{"lower order"}$$

First equality above: just notation

Putting it all together

Proof sketch: we have two MDPs M and $\widehat{M}$. need to bound:

$$\|(I-\gamma \widehat{P}^{\pi^\star})^{-1}\sqrt{\mathrm{Var}_P(V^\star)}\|_\infty = \|(I-\gamma P^{\pi^\star}_{\widehat{M}})^{-1}\sqrt{\mathrm{Var}_P(V^{\pi^\star}_{M})}\|_\infty$$

$$\leq \|(I-\gamma P^{\pi^\star}_{\widehat{M}})^{-1}\sqrt{\mathrm{Var}_P(V^{\pi^\star}_{\widehat{M}})} + \text{"lower order"}$$

$$\leq \sqrt{\frac{2}{(1-\gamma)^3}} + \text{"lower order"}$$

First equality above: just notation

Second step: concentration $\rightarrow$ we need to quantify:

$$\sqrt{\mathrm{Var}_P(V^{\pi^\star}_{M})} \approx \sqrt{\mathrm{Var}_P(V^{\pi^\star}_{\widehat{M}})}$$

Putting it all together

Proof sketch: we have two MDPs M and $\widehat{M}$. need to bound:

$$\|(I-\gamma \widehat{P}^{\pi^\star})^{-1}\sqrt{\mathrm{Var}_P(V^\star)}\|_\infty = \|(I-\gamma P^{\pi^\star}_{\widehat{M}})^{-1}\sqrt{\mathrm{Var}_P(V^{\pi^\star}_{M})}\|_\infty$$

$$\leq \|(I-\gamma P^{\pi^\star}_{\widehat{M}})^{-1}\sqrt{\mathrm{Var}_P(V^{\pi^\star}_{\widehat{M}})} + \text{"lower order"}$$

$$\leq \sqrt{\frac{2}{(1-\gamma)^3}} + \text{"lower order"}$$

First equality above: just notation

Second step: concentration → we need to quantify:

$$\sqrt{\mathrm{Var}_P(V^{\pi^\star}_{M})} \approx \sqrt{\mathrm{Var}_P(V^{\pi^\star}_{\widehat{M}})}$$

Last step: previous slide