The Sample Complexity with a Generative Model

Announcements

- Norms for class
  - Video: please feel free to turn your video!
    It's nice to see some of you once in a while.
  - Questions: by chat or just ask.

- Notes are posted in advance.
  Notes/Book on hand can help during lectures.

- HW1 posted later this week (due in two weeks)

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 an MDP $\mathscr{M} = (S, A, P, r, \gamma)$ how many observed transitions do we need to estimate $Q^\star$ (or find $\pi^\star$)?

  - We consider an abstract model (a generative model) to study the number of samples required for learning.

Recap

Summary Table

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

- VI Per iteration complexity: S^2A
- PI Per iteration complexity: S^3+S^2A
- The LP approach is only logarithmic in $1-\gamma$

Today

A Generative Model

- A generative model provides us with a sample $s' \sim P(\cdot \mid s, a)$ upon input of a state action pair (s, a).

A Generative Model

- A generative model provides us with a sample $s' \sim P(\cdot \mid s, a)$ upon input of a state action pair (s, a).
- most naive approach to learning: suppose we call our simulator N times at each state action pair. Let $\widehat{P}$ be our empirical model:

A Generative Model

- A generative model provides us with a sample $s' \sim P(\cdot \,|\, s, a)$ upon input of a state action pair (s, a).
- most naive approach to learning: suppose we call our simulator N times at each state action pair. Let $\widehat{P}$ be our empirical model:
- $\widehat{P}(s' \,|\, s, a) = \dfrac{\text{count}(s', s, a)}{N}$

  where count(s', s, a) is the #times the state-action (s, a) transitions to state s'.

A Generative Model

- A generative model provides us with a sample $s' \sim P(\cdot \,|\, s, a)$ upon input of a state action pair (s, a).
- most naive approach to learning: suppose we call our simulator N times at each state action pair. Let $\widehat{P}$ be our empirical model:
- $\widehat{P}(s' \,|\, s, a) = \dfrac{\text{count}(s', s, a)}{N}$

  where $\text{count}(s', s, a)$ is the #times the state-action (s, a) transitions to state s'.
- Each "sample" calls the generative model SA times (for each state action pair). The total number of calls to our generative model is SAN.

A Generative Model

- A generative model provides us with a sample $s' \sim P(\cdot \mid s, a)$ upon input of a state action pair (s, a).
- most naive approach to learning: suppose we call our simulator N times at each state action pair. Let $\widehat{P}$ be our empirical model:
- $\widehat{P}(s' \mid s, a) = \dfrac{\text{count}(s', s, a)}{N}$

  where $\text{count}(s', s, a)$ is the #times the state-action (s, a) transitions to state s'.
- Each "sample" calls the generative model SA times (for each state action pair). The total number of calls to our generative model is SAN.
- The generative model helps us disentangle the issue of fundamental statistical li from exploration.

How many samples do we need to learn?

- This is a reasonable abstraction to understand the statistical limit, without having to directly address exploration.

How many samples do we need to learn?

- This is a reasonable abstraction to understand the statistical limit, without having to directly address exploration.
- Note that since P has a S^2A parameters, a naive approach would be to estimate P accurately (using $O(S^2A)$ samples) and then use $\widehat{P}$ for planning.

How many samples do we need to learn?

- This is a reasonable abstraction to understand the statistical limit, without having to directly address exploration.
- Note that since P has a S^2A parameters, a naive approach would be to estimate P accurately (using $O(S^2A)$ samples) and then use $\widehat{P}$ for planning.

- *Do we require an accurate model of the world in order to find a near optimal policy?*

ε-optimal

model based RL

vs model free RL

Attempt 1:

the naive model based approach

let's build an accurate model with $O(S^2A)$ samples

$\hookrightarrow N = O(S)$ call.

call gen model

Model accuracy

SAN times

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 $1 - \delta$,

$O(S^2 A / \epsilon^2)$

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

(reward is known)

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

$O(S^2A)$ samples.

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: $\max_{s,a} \|P(\cdot\,|\,s,a) - \widehat{P}(\cdot\,|\,s,a)\|_1 \leq (1-\gamma)^2\epsilon/2$.
- Uniform value accuracy: For all policies π, $\|Q^\pi - \widehat{Q}^\pi\|_\infty \leq \epsilon/2$
- Near optimal planning: Suppose that $\widehat{\pi}$ is the optimal policy in $\widehat{M}$. ~~$\|\widehat{Q}^\pi - Q^\star\|_\infty \leq \epsilon$~~

value of the policy we find.

estimates of Q^*

$\|\widehat{Q}^* - Q^*\| \leq \varepsilon$, $\|Q^{\widehat{\pi}^*} - Q^*\| \leq \varepsilon$

Matrix Expressions

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

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

P^π $SA \times SA$

P matrix $SA \times S$

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,

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

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

$V^\pi \in \mathbb{R}^S$, $Q^\pi \in \mathbb{R}^{SA}$ vectors

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

$\Rightarrow (I - \gamma P^\pi) Q^\pi = r$

- And that:
  $Q^\pi = (I - \gamma P^\pi)^{-1} r$
  (where one can show the inverse exists)

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

“Simulation” Lemma

(Simulation Lemma) For all π, we have:

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

“Simulation” Lemma

(Simulation Lemma) For all π, we have:

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

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

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

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

when mdt. $\hat{Q}^\pi$

Proof: Claim 1

Concentration of a distribution in the ℓ_1 norm: for a fixed s, a

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

with pr greater than $1-\delta$, with m samples used to estimate $\widehat{P}(\cdot \mid s,a)$.

The first claim now follows by the union bound.

Proof: Claim 2

For the second claim, we have that:

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

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

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

(why is step 1 true?)

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

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,

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

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

- Estimated transition $\widehat{P}$, optimal value in estimated model $\widehat{Q}^\star$, optimal policy in optimal model $\widehat{\pi}^\star$, (true) value of estimated policy $Q^{\widehat{\pi}^\star}$

only need $O(SA)$ calls to estimate $\hat{Q}^\star$

Attempt 2:

sublinear sample complexity!

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

Attempt 2: Sublinear Sample Complexity

Attempt 2: Sublinear Sample Complexity

Proposition: (Crude Value Bound) Let $\delta \geq 0$. With probability greater than $1 - \delta$,

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

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

$\Rightarrow$ # samples $\geq$ $O\left(\frac{SA}{(1-\gamma)^4 \varepsilon^2}\right)$

$\Downarrow$

$\|Q^* - \widehat{Q}^{\pi^*}\| \leq \varepsilon$

Attempt 2: Sublinear Sample Complexity

Proposition: (Crude Value Bound) Let $\delta \geq 0$. With probability greater than $1 - \delta$,

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

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

amplification?

What about the value of the policy?

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

scaling of $Q^\star \leq O\left(\frac{1}{1-\gamma}\right)$

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:

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

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

Component-wise Bounds Lemma

Lemma: we have that:

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

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

proof:

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

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

using the simulation lemma in the final step.

See notes for the proof of second claim.

Proof: (& key idea for sublinearity!)

For the first claim,

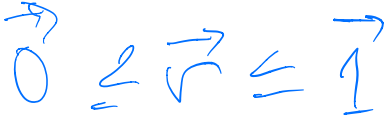
$$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \frac{\gamma}{1-\gamma}\|(P - \widehat{P})V^\star\|_\infty$$

Why?

Proof: (& key idea for sublinearity!)

For the first claim,

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



Why?

By applying Hoeffding's inequality $V^\star$ 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 $1 - \delta$.

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 $1-\delta$,

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

where c is an absolute constant.

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

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

lower bound says this is optimal

("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.

care about $Q^{\hat{\pi}^*}$

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

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

Corollary: What about the policy? Need $N/(1-\gamma)^2$ 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 $1-\delta$),

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

$Q^\star \approx O\left(\frac{1}{1-\gamma}\right)$.

Suppose want $\|Q^\star - Q^{\hat{\pi}\star}\|_\infty \leq \frac{\epsilon}{1-\gamma}$

sufficient $N \geq \frac{c}{(1-\gamma)} \frac{\log(SA)}{\epsilon^2}$

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 $1-\delta$),

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

Lower Bound: We can't do better.

Proof sketch: part 1

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

Proof sketch: part 1

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

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

- 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
  $$\|Q^\star - \widehat{Q}^\star\|_\infty \leq \gamma\sqrt{\frac{2\log(2SA/\delta)}{N}}\|(I - \gamma \widehat{P}^{\pi^\star})^{-1}\sqrt{\mathrm{Var}_P(V^\star)}\|_\infty$$
  $+$"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 (totoal) 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_M^π as the (totoal) 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:
  $\Sigma_M^\pi = \gamma^2 \mathrm{Var}_P(V_M^\pi) + \gamma^2 P^\pi \Sigma_M^\pi$

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

Putting it all together

Proof: 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: 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: 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 -> need to say

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

Putting it all together

Proof: 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 -> need to say

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

Last step: previous slide