Statistical Properties of Robust MDPs

Wenhao Yang

Peking Univeristy

What is (distributional) robustness?

- Models may be sensitive to estimation errors.

What is (distributional) robustness?

- Models may be sensitive to estimation errors.
- Example: suppose $X \sim P$ and θ is the parameter of interest.

What is (distributional) robustness?

- Models may be sensitive to estimation errors.
- Example: suppose $X \sim P$ and θ is the parameter of interest. The population risk minimization is:

$$\min_{\theta} \mathbb{E}_P f(X;\theta); \;\; \theta^* \in \arg\max_{\theta} \mathbb{E}_P f(X;\theta).$$

What is (distributional) robustness?

- Models may be sensitive to estimation errors.
- Example: suppose $X \sim P$ and θ is the parameter of interest. The population risk minimization is:

$$\min_{\theta} \mathbb{E}_P f(X;\theta); \;\; \theta^* \in \arg\max_{\theta} \mathbb{E}_P f(X;\theta).$$

The empirical risk minimization is:

$$\min_{\theta} \mathbb{E}_{\widehat{P}_n} f(X;\theta) := \frac{1}{n}\sum_i f(X_i;\theta); \;\; \widehat{\theta}_n^* \in \arg\max_{\theta} \mathbb{E}_{\widehat{P}_n} f(X;\theta)$$

What is (distributional) robustness?

- Models may be sensitive to estimation errors.
- Example: suppose $X \sim P$ and θ is the parameter of interest. The population risk minimization is:

$$\min_{\theta} \mathbb{E}_P f(X;\theta);\;\; \theta^* \in \arg\max_{\theta} \mathbb{E}_P f(X;\theta).$$

The empirical risk minimization is:

$$\min_{\theta} \mathbb{E}_{\widehat{P}_n} f(X;\theta) := \frac{1}{n}\sum_i f(X_i;\theta);\;\; \widehat{\theta}^*_n \in \arg\max_{\theta} \mathbb{E}_{\widehat{P}_n} f(X;\theta)$$

- $\widehat{\theta}^*_n$ may vary a lot with estimation errors of $\widehat{P}_n$.

What is (distributional) robustness?

- One solution: introduce (distributional) robustness.

What is (distributional) robustness?

- One solution: introduce (distributional) robustness.
- The population robust risk minimization:

$$\min_{\theta} \sup_{D(Q\|P)\leq\rho} \mathbb{E}_Q f(X;\theta); \text{ minimizer: } \theta_r^*.$$

What is (distributional) robustness?

- One solution: introduce (distributional) robustness.
- The population robust risk minimization:

$$\min_{\theta} \sup_{D(Q\|P)\leq\rho} \mathbb{E}_Q f(X;\theta); \text{ minimizer: } \theta_r^*.$$

- The empirical robust risk minimization:

$$\min_{\theta} \sup_{D(Q\|\widehat{P}_n)\leq\rho} \mathbb{E}_Q f(X;\theta); \text{ minimizer: } \widehat{\theta}_r^*.$$

What is (distributional) robustness?

- One solution: introduce (distributional) robustness.
- The population robust risk minimization:

$$\min_{\theta} \sup_{D(Q\|P)\le\rho} \mathbb{E}_Q f(X;\theta); \text{ minimizer: } \theta_r^*.$$

- The empirical robust risk minimization:

$$\min_{\theta} \sup_{D(Q\|\widehat{P}_n)\le\rho} \mathbb{E}_Q f(X;\theta); \text{ minimizer: } \widehat{\theta}_r^*.$$

- Why $\widehat{\theta}_r^*$ is less sensitive to randomness of $\widehat{P}_n$?

What is (distributional) robustness?

- One solution: introduce (distributional) robustness.
- The population robust risk minimization:

$$\min_{\theta} \sup_{D(Q\|P)\leq\rho} \mathbb{E}_Q f(X;\theta); \text{ minimizer: } \theta_r^*.$$

- The empirical robust risk minimization:

$$\min_{\theta} \sup_{D(Q\|\widehat{P}_n)\leq\rho} \mathbb{E}_Q f(X;\theta); \text{ minimizer: } \widehat{\theta}_r^*.$$

- Why $\widehat{\theta}_r^*$ is less sensitive to randomness of $\widehat{P}_n$?
- Image ρ is super large, like infinity.

Robust Markov Decision Processes

- Same parameters with MDPs: $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$.

Robust Markov Decision Processes

- Same parameters with MDPs: $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$.
- Additional parameters: uncertainty set $\mathcal{P}$.

Robust Markov Decision Processes

- Same parameters with MDPs: $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$.
- Additional parameters: uncertainty set $\mathcal{P}$.
- Robust value function:

$$V_r^\pi(s) := \inf_{P \in \mathcal{P}} V_P^\pi(s).$$

Robust Markov Decision Processes

- Same parameters with MDPs: $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$.
- Additional parameters: uncertainty set $\mathcal{P}$.
- Robust value function:

$$V_r^\pi(s) := \inf_{P \in \mathcal{P}} V_P^\pi(s).$$

- Robust Bellman operator $\mathcal{T}_r^\pi$:

$$\mathcal{T}_r^\pi V = R^\pi + \gamma \inf_{P \in \mathcal{P}} P^\pi V.$$

Robust Markov Decision Processes

- Same parameters with MDPs: $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$.
- Additional parameters: uncertainty set $\mathcal{P}$.
- Robust value function:

$$V_r^\pi(s) := \inf_{P \in \mathcal{P}} V_P^\pi(s).$$

- Robust Bellman operator $\mathcal{T}_r^\pi$:

$$\mathcal{T}_r^\pi V = R^\pi + \gamma \inf_{P \in \mathcal{P}} P^\pi V.$$

- Optimal robust Bellman operator $\mathcal{T}_r$:

$$\mathcal{T}_r V = \max_\pi R^\pi + \gamma \inf_{P \in \mathcal{P}} P^\pi V.$$

Robust Markov Decision Processes

- Same parameters with MDPs: $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$.
- Additional parameters: uncertainty set $\mathcal{P}$.
- Robust value function:

$$V_r^\pi(s) := \inf_{P \in \mathcal{P}} V_P^\pi(s).$$

- Robust Bellman operator $\mathcal{T}_r^\pi$:

$$\mathcal{T}_r^\pi V = R^\pi + \gamma \inf_{P \in \mathcal{P}} P^\pi V.$$

- Optimal robust Bellman operator $\mathcal{T}_r$:

$$\mathcal{T}_r V = \max_\pi R^\pi + \gamma \inf_{P \in \mathcal{P}} P^\pi V.$$

- Both are γ-contraction. Fixed points are V_r^π and $V_r^* = \max_\pi V_r^\pi$.

Uncertainty Set

Uncertainty Set

- Can we choose an arbitrary $\mathcal{P}$?

Uncertainty Set

- Can we choose an arbitrary $\mathcal{P}$?
  - No! It may be NP hard.[WKR13]

Uncertainty Set

- Can we choose an arbitrary $\mathcal{P}$?
  - No! It may be NP hard.[WKR13]
- [WKR13] Most common assumption on $\mathcal{P}$:

Uncertainty Set

- Can we choose an arbitrary $\mathcal{P}$?
  - No! It may be NP hard.[WKR13]
- [WKR13] Most common assumption on $\mathcal{P}$:
  - (s, a)-rectangular: $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_{s,a}$.

Uncertainty Set

- Can we choose an arbitrary $\mathcal{P}$?
  - No! It may be NP hard.[WKR13]
- [WKR13] Most common assumption on $\mathcal{P}$:
  - (s,a)-rectangular: $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_{s,a}$.
  - s-rectangular: $\mathcal{P} = \bigotimes_{s} \mathcal{P}_{s}$.

Uncertainty Set

- Can we choose an arbitrary $\mathcal{P}$?
  - No! It may be NP hard.[WKR13]
- [WKR13] Most common assumption on $\mathcal{P}$:
  - (s, a)-rectangular: $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_{s,a}$.
  - s-rectangular: $\mathcal{P} = \bigotimes_{s} \mathcal{P}_{s}$.
- Example: f-divergence set:

Uncertainty Set

- Can we choose an arbitrary $\mathcal{P}$?
  - No! It may be NP hard.[WKR13]
- [WKR13] Most common assumption on $\mathcal{P}$:
  - (s,a)-rectangular: $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_{s,a}$.
  - s-rectangular: $\mathcal{P} = \bigotimes_{s} \mathcal{P}_{s}$.
- Example: f-divergence set:
  - $\mathcal{P}_{s,a} = \{Q_{s,a} \in \Delta(\mathcal{S}) | \sum_{s' \in \mathcal{S}} f(\frac{Q_{s,a}(s')}{P_{s,a}(s')}) P_{s,a}(s') \leq \rho\}$.

Uncertainty Set

- Can we choose an arbitrary $\mathcal{P}$?
  - No! It may be NP hard.[WKR13]
- [WKR13] Most common assumption on $\mathcal{P}$:
  - (s, a)-rectangular: $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_{s,a}$.
  - s-rectangular: $\mathcal{P} = \bigotimes_{s} \mathcal{P}_{s}$.
- Example: f-divergence set:
  - $\mathcal{P}_{s,a} = \{Q_{s,a} \in \Delta(\mathcal{S}) | \sum_{s' \in \mathcal{S}} f(\frac{Q_{s,a}(s')}{P_{s,a}(s')}) P_{s,a}(s') \leq \rho\}$.
  - $\mathcal{P}_{s} = \{Q_{s,a} \in \Delta(\mathcal{S}) | \sum_{a \in \mathcal{A}, s' \in \mathcal{S}} f(\frac{Q_{s,a}(s')}{P_{s,a}(s')}) P_{s,a}(s') \leq |\mathcal{A}|\rho\}$.

Uncertainty Set

- Can we choose an arbitrary $\mathcal{P}$?
  - No! It may be NP hard.[WKR13]
- [WKR13] Most common assumption on $\mathcal{P}$:
  - (s, a)-rectangular: $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_{s,a}$.
  - s-rectangular: $\mathcal{P} = \bigotimes_{s} \mathcal{P}_{s}$.
- Example: f-divergence set:
  - $\mathcal{P}_{s,a} = \{Q_{s,a} \in \Delta(\mathcal{S}) | \sum_{s' \in \mathcal{S}} f(\frac{Q_{s,a}(s')}{P_{s,a}(s')}) P_{s,a}(s') \leq \rho\}$.
  - $\mathcal{P}_{s} = \{Q_{s,a} \in \Delta(\mathcal{S}) | \sum_{a \in \mathcal{A}, s' \in \mathcal{S}} f(\frac{Q_{s,a}(s')}{P_{s,a}(s')}) P_{s,a}(s') \leq |\mathcal{A}|\rho\}$.
- [WKR13] Optimal polices $\pi_r^* \in \arg\max_\pi V_r^\pi$:

Uncertainty Set

- Can we choose an arbitrary $\mathcal{P}$?
  - No! It may be NP hard.[WKR13]
- [WKR13] Most common assumption on $\mathcal{P}$:
  - (s,a)-rectangular: $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_{s,a}$.
  - s-rectangular: $\mathcal{P} = \bigotimes_{s} \mathcal{P}_{s}$.
- Example: f-divergence set:
  - $\mathcal{P}_{s,a} = \{Q_{s,a} \in \Delta(\mathcal{S}) | \sum_{s' \in \mathcal{S}} f(\frac{Q_{s,a}(s')}{P_{s,a}(s')}) P_{s,a}(s') \leq \rho\}$.
  - $\mathcal{P}_{s} = \{Q_{s,a} \in \Delta(\mathcal{S}) | \sum_{a \in \mathcal{A}, s' \in \mathcal{S}} f(\frac{Q_{s,a}(s')}{P_{s,a}(s')}) P_{s,a}(s') \leq |\mathcal{A}|\rho\}$.
- [WKR13] Optimal polices $\pi_r^* \in \arg\max_\pi V_r^\pi$:
  - Stationary, deterministic under (s,a)-rectangular assumption.

Uncertainty Set

- Can we choose an arbitrary $\mathcal{P}$?
  - No! It may be NP hard.[WKR13]
- [WKR13] Most common assumption on $\mathcal{P}$:
  - (s,a)-rectangular: $\mathcal{P} = \bigotimes_{s,a} \mathcal{P}_{s,a}$.
  - s-rectangular: $\mathcal{P} = \bigotimes_{s} \mathcal{P}_{s}$.
- Example: f-divergence set:
  - $\mathcal{P}_{s,a} = \{Q_{s,a} \in \Delta(\mathcal{S}) | \sum_{s' \in \mathcal{S}} f(\frac{Q_{s,a}(s')}{P_{s,a}(s')}) P_{s,a}(s') \leq \rho\}$.
  - $\mathcal{P}_{s} = \{Q_{s,a} \in \Delta(\mathcal{S}) | \sum_{a \in \mathcal{A}, s' \in \mathcal{S}} f(\frac{Q_{s,a}(s')}{P_{s,a}(s')}) P_{s,a}(s') \leq |\mathcal{A}|\rho\}$.
- [WKR13] Optimal polices $\pi_r^* \in \arg\max_\pi V_r^\pi$:
  - Stationary, deterministic under (s,a)-rectangular assumption.
  - Stationary, stochastic under s-rectangular assumption.

Data Generation Mechanism

Data Generation Mechanism

- P is always unknown!

Data Generation Mechanism

- P is always unknown!
- Generative model: for each (s, a), we obtain n samples $\{X_i^{(s,a)}\}_{i=1}^n \sim P_{s,a}(\cdot)$.

Data Generation Mechanism

- P is always unknown!
- Generative model: for each (s, a), we obtain n samples $\{X_i^{(s,a)}\}_{i=1}^n \sim P_{s,a}(\cdot)$.
- Estimation of P: $\widehat{P}_{s,a}(s') = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i^{(s,a)} = s')$.

Data Generation Mechanism

- P is always unknown!
- Generative model: for each (s, a), we obtain n samples $\{X_i^{(s,a)}\}_{i=1}^n \sim P_{s,a}(\cdot)$.
- Estimation of P: $\widehat{P}_{s,a}(s') = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i^{(s,a)} = s')$.
- $\widehat{\mathcal{P}} \to \mathcal{P}$, $\widehat{V}_r^\pi \to V_r^\pi$, $\widehat{V}_r^* \to V_r^*$.

Prior Results

Prior Results

- How many samples are sufficient to guarantee $\|V_r^* - \widehat{V}_r^*\|_\infty \leq \varepsilon$?

Prior Results

- How many samples are sufficient to guarantee $\|V_r^* - \widehat{V}_r^*\|_\infty \leq \varepsilon$?
- [ZBZ+21]: (s, a)-rectangular, $f(t) = t \log t$ (KL set), number of samples $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|\exp(\frac{1}{\beta(1-\gamma)})}{\varepsilon^2(1-\gamma)^2\rho^2}\right)$.

Prior Results

- How many samples are sufficient to guarantee $\|V_r^* - \widehat{V}_r^*\|_\infty \leq \varepsilon$?
- [ZBZ+21]: (s, a)-rectangular, $f(t) = t \log t$ (KL set), number of samples $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|\exp(\frac{1}{\beta(1-\gamma)})}{\varepsilon^2(1-\gamma)^2\rho^2}\right)$.
- It is counter-intuitive...

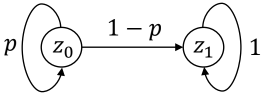
Lower Bound

Lower Bound

- A classic example with 2 states, 1 action:

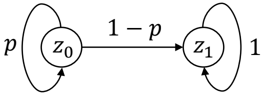
Lower Bound

- A classic example with 2 states, 1 action:



Lower Bound

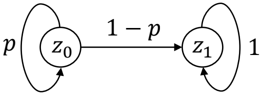
- A classic example with 2 states, 1 action:



- Robust value function: $V_r^*(z_0) = \frac{1}{1-\gamma g(p)}$,

Lower Bound

- A classic example with 2 states, 1 action:



- Robust value function: $V_r^*(z_0) = \frac{1}{1-\gamma g(p)}$,
  where $g(p) = \inf_{D_f(q\|p)\leq\rho} q$ and
  $D_f(q\|p) = pf(p/q) + (1-p)f(1-p/1-q)$.

Lower bound

Lower bound

- Consider a perturbation from p to $p + \delta$:

Lower bound

- Consider a perturbation from p to $p+\delta$:

$$\frac{1}{1-\gamma g(p+\delta)} - \frac{1}{1-\gamma g(p)} \geq \frac{\gamma \delta g'(p)}{(1-\gamma g(p))^2} = 2\varepsilon.$$

Lower bound

- Consider a perturbation from p to $p+\delta$:

$$\frac{1}{1-\gamma g(p+\delta)}-\frac{1}{1-\gamma g(p)}\geq\frac{\gamma\delta g'(p)}{(1-\gamma g(p))^2}=2\varepsilon.$$

- [AMK13] told us $n=\widetilde{\Omega}\left(\frac{p(1-p)g'(p)^2}{\varepsilon^2(1-\gamma g(p))^4}\right)$.

Lower bound

- Consider a perturbation from p to $p+\delta$:

$$\frac{1}{1-\gamma g(p+\delta)} - \frac{1}{1-\gamma g(p)} \geq \frac{\gamma \delta g'(p)}{(1-\gamma g(p))^2} = 2\varepsilon.$$

- [AMK13] told us $n = \widetilde{\Omega}\left(\frac{p(1-p)g'(p)^2}{\varepsilon^2(1-\gamma g(p))^4}\right)$.
- $f(t) = |t-1|$,

Lower bound

- Consider a perturbation from p to $p+\delta$:

$$\frac{1}{1-\gamma g(p+\delta)} - \frac{1}{1-\gamma g(p)} \geq \frac{\gamma\delta g'(p)}{(1-\gamma g(p))^2} = 2\varepsilon.$$

- [AMK13] told us $n = \widetilde{\Omega}\left(\frac{p(1-p)g'(p)^2}{\varepsilon^2(1-\gamma g(p))^4}\right)$.
- $f(t) = |t-1|$, $g(p) = p - \rho/2$,

Lower bound

- Consider a perturbation from p to $p+\delta$:

$$\frac{1}{1-\gamma g(p+\delta)} - \frac{1}{1-\gamma g(p)} \geq \frac{\gamma \delta g'(p)}{(1-\gamma g(p))^2} = 2\varepsilon.$$

- [AMK13] told us $n = \widetilde{\Omega}\left(\frac{p(1-p)g'(p)^2}{\varepsilon^2(1-\gamma g(p))^4}\right)$.
- $f(t) = |t-1|$, $g(p) = p - \rho/2$,$n = \widetilde{\Omega}\left(\frac{1-\gamma}{\varepsilon^2}\min\{\frac{1}{(1-\gamma)^4}, \frac{1}{\rho^4}\}\right)$.

Lower bound

- Consider a perturbation from p to $p+\delta$:

$$\frac{1}{1-\gamma g(p+\delta)} - \frac{1}{1-\gamma g(p)} \geq \frac{\gamma\delta g'(p)}{(1-\gamma g(p))^2} = 2\varepsilon.$$

- [AMK13] told us $n = \widetilde{\Omega}\left(\frac{p(1-p)g'(p)^2}{\varepsilon^2(1-\gamma g(p))^4}\right)$.
- $f(t) = |t-1|$, $g(p) = p - \rho/2, n = \widetilde{\Omega}\left(\frac{1-\gamma}{\varepsilon^2}\min\{\frac{1}{(1-\gamma)^4}, \frac{1}{\rho^4}\}\right)$.
- $f(t) = (t-1)^2$,

Lower bound

- Consider a perturbation from p to $p+\delta$:

$$\frac{1}{1-\gamma g(p+\delta)} - \frac{1}{1-\gamma g(p)} \geq \frac{\gamma\delta g'(p)}{(1-\gamma g(p))^2} = 2\varepsilon.$$

- [AMK13] told us $n = \widetilde{\Omega}\left(\frac{p(1-p)g'(p)^2}{\varepsilon^2(1-\gamma g(p))^4}\right)$.
- $f(t) = |t-1|$, $g(p) = p-\rho/2, n = \widetilde{\Omega}\left(\frac{1-\gamma}{\varepsilon^2}\min\{\frac{1}{(1-\gamma)^4}, \frac{1}{\rho^4}\}\right)$.
- $f(t) = (t-1)^2$,
  $g(p) = p - \sqrt{\rho p(1-p)}$,

Lower bound

- Consider a perturbation from p to $p+\delta$:

$$\frac{1}{1-\gamma g(p+\delta)} - \frac{1}{1-\gamma g(p)} \geq \frac{\gamma\delta g'(p)}{(1-\gamma g(p))^2} = 2\varepsilon.$$

- [AMK13] told us $n = \widetilde{\Omega}\left(\frac{p(1-p)g'(p)^2}{\varepsilon^2(1-\gamma g(p))^4}\right)$.
- $f(t) = |t-1|$, $g(p) = p - \rho/2, n = \widetilde{\Omega}\left(\frac{1-\gamma}{\varepsilon^2}\min\{\frac{1}{(1-\gamma)^4}, \frac{1}{\rho^4}\}\right)$.
- $f(t) = (t-1)^2$,
  $g(p) = p - \sqrt{\rho p(1-p)}, n = \widetilde{\Omega}\left(\frac{1}{\varepsilon^2(1-\gamma)^2}\min\{\frac{1}{1-\gamma}, \frac{1}{\rho}\}\right)$.

Upper bound

Upper bound

- Okay...How about upper bound?

Upper bound

- Okay...How about upper bound?
- No explicit expression of V_r^*...

Upper bound

- Okay...How about upper bound?
- No explicit expression of V_r^*...
- Let's take advantage of robust Bellman operator:

Upper bound

- Okay...How about upper bound?
- No explicit expression of V_r^*...
- Let's take advantage of robust Bellman operator:

$$\|V_r^* - \widehat{V}_r^*\|_\infty \leq \frac{1}{1-\gamma} \sup_{\pi\in\Pi, V\in[0,1/1-\gamma]^{|\mathcal{S}|}} \|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty.$$

Upper bound

- Okay...How about upper bound?
- No explicit expression of V_r^*...
- Let's take advantage of robust Bellman operator:

$$\|V_r^* - \widehat{V}_r^*\|_\infty \le \frac{1}{1-\gamma} \sup_{\pi\in\Pi, V\in[0,1/1-\gamma]^{|\mathcal{S}|}} \|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty.$$

- Uniform analysis on V can be unnecessary. But no harm!

Upper bound

- Okay...How about upper bound?
- No explicit expression of V_r^*...
- Let's take advantage of robust Bellman operator:

$$\|V_r^* - \widehat{V}_r^*\|_\infty \leq \frac{1}{1-\gamma} \sup_{\pi \in \Pi, V \in [0, 1/1-\gamma]^{|\mathcal{S}|}} \|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty.$$

- Uniform analysis on V can be unnecessary. But no harm! $\log \mathcal{N}(\Pi, \|\cdot\|_1) \approx \log \mathcal{N}([0, 1/1-\gamma]^{|\mathcal{S}|}, \|\cdot\|_\infty) \approx \Theta(|\mathcal{S}|)$.

Upper bound

Upper bound

- For any fixed π, V, we need concentration inequality to bound $\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty$.

Upper bound

- For any fixed π, V, we need concentration inequality to bound $\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty$.
- How...? Randomness is hidden in the constraints.

Upper bound

- For any fixed π, V, we need concentration inequality to bound $\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty$.
- How...? Randomness is hidden in the constraints. Try dual.

Upper bound

- For any fixed π, V, we need concentration inequality to bound $\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty$.
- How...? Randomness is hidden in the constraints. Try dual. By [Sha17]:

Upper bound

- For any fixed π, V, we need concentration inequality to bound $\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty$.
- How...? Randomness is hidden in the constraints. Try dual. By [Sha17]:

$$(P)\ \inf_{D_f(Q\|P)\le\rho} \sum_s Q(s)V(s).$$

$$(D)\ \sup_{\lambda\ge 0,\eta\in\mathbb{R}} -\lambda\sum_s P(s) f^*(\frac{\eta - V(s)}{\lambda}) - \lambda\rho + \eta.$$

Upper bound

- For any fixed π, V, we need concentration inequality to bound $\|\mathcal{T}_r^\pi V - \widehat{\mathcal{T}}_r^\pi V\|_\infty$.
- How...? Randomness is hidden in the constraints. Try dual. By [Sha17]:

$$(P) \inf_{D_f(Q\|P)\leq\rho} \sum_s Q(s)V(s).$$

$$(D) \sup_{\lambda\geq 0,\eta\in\mathbb{R}} -\lambda\sum_s P(s) f^*(\frac{\eta - V(s)}{\lambda}) - \lambda\rho + \eta.$$

- Next: calculations...

Upper bound

Upper bound

- Consider three f: $|t-1|$, $(t-1)^2$, $t\log t$, in (s,a)-rectangular assumption.

Upper bound

- Consider three f: $|t-1|$, $(t-1)^2$, $t\log t$, in (s,a)-rectangular assumption.
- Upper bound $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\varepsilon^2\rho^2(1-\gamma)^4}\right)$.

Upper bound

- Consider three f: $|t-1|$, $(t-1)^2$, $t\log t$, in (s,a)-rectangular assumption.
- Upper bound $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\varepsilon^2\rho^2(1-\gamma)^4}\right)$.
- For $f(t)=t\log t$, an additional parameter $(\min_{P^*(s'|s,a)>0} P^*(s'|s,a))^{-1}$.

Upper bound

- Consider three f: $|t-1|$, $(t-1)^2$, $t\log t$, in (s,a)-rectangular assumption.
- Upper bound $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\varepsilon^2\rho^2(1-\gamma)^4}\right)$.
- For $f(t) = t\log t$, an additional parameter $(\min_{P^*(s'|s,a)>0} P^*(s'|s,a))^{-1}$.
- Wait... Why infinity when $\rho \to 0$?

Upper bound

- Consider three f: $|t-1|$, $(t-1)^2$, $t\log t$, in (s,a)-rectangular assumption.
- Upper bound $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\varepsilon^2\rho^2(1-\gamma)^4}\right)$.
- For $f(t)=t\log t$, an additional parameter $(\min_{P^*(s'|s,a)>0} P^*(s'|s,a))^{-1}$.
- Wait... Why infinity when $\rho\to 0$?
- By fact $V_r^*\to V^*$ when $\rho\to 0$, alternative bound:

$$\|V_r^*-\widehat{V}_r^*\|_\infty \le \mathcal{O}\left(\frac{h(\rho)}{(1-\gamma)^2}\right)+\widetilde{\mathcal{O}}\left(\sqrt{\frac{|\mathcal{S}|}{(1-\gamma)^4 n}}\right).$$

Asymptotic Normality

Asymptotic Normality

- Confidence length of non-asymptotic results is $O_P(\sqrt{\log n / n})$.

Asymptotic Normality

- Confidence length of non-asymptotic results is $O_P(\sqrt{\log n/n})$.
- The non-asymptotic upper bound is not tight.

Asymptotic Normality

- Confidence length of non-asymptotic results is $O_P(\sqrt{\log n/n})$.
- The non-asymptotic upper bound is not tight.
- In large sample regime, rate of $\|V_r^* - \widehat{V}_r^*\|_\infty$ is $O_P(1/\sqrt{n})$.

Asymptotic Normality

- Confidence length of non-asymptotic results is $O_P(\sqrt{\log n/n})$.
- The non-asymptotic upper bound is not tight.
- In large sample regime, rate of $\|V_r^* - \widehat{V}_r^*\|_\infty$ is $O_P(1/\sqrt{n})$.
- Fix a π, by CLT and delta method:

$$\sqrt{n}(\widehat{\mathcal{T}}_r^\pi V_r^\pi - \mathcal{T}_r^\pi V_r^\pi) \xrightarrow{d} \mathcal{N}(0, \Lambda^\pi).$$

Asymptotic Normality

- Confidence length of non-asymptotic results is $O_P(\sqrt{\log n/n})$.
- The non-asymptotic upper bound is not tight.
- In large sample regime, rate of $\|V_r^* - \widehat{V}_r^*\|_\infty$ is $O_P(1/\sqrt{n})$.
- Fix a π, by CLT and delta method:

$$\sqrt{n}(\widehat{\mathcal{T}}_r^\pi V_r^\pi - \mathcal{T}_r^\pi V_r^\pi) \xrightarrow{d} \mathcal{N}(0, \Lambda^\pi).$$

- LHS $= -\widehat{M}^\pi \cdot \sqrt{n}(V_r^\pi - \widehat{V}_r^\pi) + o_P(\sqrt{n}\|\widehat{V}_r^\pi - V_r^\pi\|_\infty)$, where $\widehat{M}^\pi$ is the derivative of functional $I - \widehat{\mathcal{T}}_r^\pi$ at V_r^π.

Asymptotic Normality

- Confidence length of non-asymptotic results is $O_P(\sqrt{\log n/n})$.
- The non-asymptotic upper bound is not tight.
- In large sample regime, rate of $\|V_r^* - \widehat{V}_r^*\|_\infty$ is $O_P(1/\sqrt{n})$.
- Fix a π, by CLT and delta method:

$$\sqrt{n}(\widehat{\mathcal{T}}_r^\pi V_r^\pi - \mathcal{T}_r^\pi V_r^\pi) \xrightarrow{d} \mathcal{N}(0, \Lambda^\pi).$$

- LHS $= -\widehat{M}^\pi \cdot \sqrt{n}(V_r^\pi - \widehat{V}_r^\pi) + o_P(\sqrt{n}\|\widehat{V}_r^\pi - V_r^\pi\|_\infty)$, where $\widehat{M}^\pi$ is the derivative of functional $I - \widehat{\mathcal{T}}_r^\pi$ at V_r^π.
- Notice $\sqrt{n}(V_r^\pi - \widehat{V}_r^\pi) = O_P(1)$ and prove $\widehat{M}^\pi$ is consistent to M^π:

Asymptotic Normality

- Confidence length of non-asymptotic results is $O_P(\sqrt{\log n / n})$.
- The non-asymptotic upper bound is not tight.
- In large sample regime, rate of $\| V_r^* - \widehat{V}_r^* \|_\infty$ is $O_P(1/\sqrt{n})$.
- Fix a π, by CLT and delta method:

$$\sqrt{n}(\widehat{\mathcal{T}}_r^\pi V_r^\pi - \mathcal{T}_r^\pi V_r^\pi) \xrightarrow{d} \mathcal{N}(0, \Lambda^\pi).$$

- LHS $= -\widehat{M}^\pi \cdot \sqrt{n}(V_r^\pi - \widehat{V}_r^\pi) + o_P(\sqrt{n} \| \widehat{V}_r^\pi - V_r^\pi \|_\infty)$, where $\widehat{M}^\pi$ is the derivative of functional $I - \widehat{\mathcal{T}}_r^\pi$ at V_r^π.
- Notice $\sqrt{n}(V_r^\pi - \widehat{V}_r^\pi) = O_P(1)$ and prove $\widehat{M}^\pi$ is consistent to M^π:

$$\sqrt{n}(\widehat{V}_r^\pi - V_r^\pi) \xrightarrow{d} \mathcal{N}(0, (M^\pi)^{-1} \Lambda^\pi (M^\pi)^{-\top}).$$

Asymptotic Normality

Asymptotic Normality

- What about $\sqrt{n}(\widehat{V}_r^* - V_r^*)$?

Asymptotic Normality

- What about $\sqrt{n}(\widehat{V}_r^* - V_r^*)$?
- Need uniqueness assumption of $\pi^* \in \arg\max V_r^\pi$.

Asymptotic Normality

- What about $\sqrt{n}(\widehat{V}_r^* - V_r^*)$?
- Need uniqueness assumption of $\pi^* \in \arg\max V_r^\pi$. And replace π with π^*.

Asymptotic Normality

- What about $\sqrt{n}(\widehat{V}_r^* - V_r^*)$?
- Need uniqueness assumption of $\pi^* \in \arg\max V_r^\pi$. And replace π with π^*.
- If not. Still $\sqrt{n}$ rate, but not asymptotic normal.

Asymptotic Normality

- What about $\sqrt{n}(\widehat{V}_r^* - V_r^*)$?
- Need uniqueness assumption of $\pi^* \in \arg\max V_r^\pi$. And replace π with π^*.
- If not. Still $\sqrt{n}$ rate, but not asymptotic normal. The asymptotic distribution be like:

$$\bigvee_{\pi \in \Pi^*} \mathcal{N}(0, (M^\pi)^{-1} \Lambda^\pi (M^\pi)^{-\top}),$$

where $x \vee y = \max\{x, y\}$.

Asymptotic Normality

- What about $\sqrt{n}(\widehat{V}_r^* - V_r^*)$?
- Need uniqueness assumption of $\pi^* \in \arg\max V_r^\pi$. And replace π with π^*.
- If not. Still $\sqrt{n}$ rate, but not asymptotic normal. The asymptotic distribution be like:

$$\bigvee_{\pi \in \Pi^*} \mathcal{N}(0, (M^\pi)^{-1} \Lambda^\pi (M^\pi)^{-\top}),$$

where $x \vee y = \max\{x, y\}$.
- How to do inference?

1. Introduction
2. Robust MDPs
3. Statistical Results
4. **Further Discussion**
5. Reference

Discussion

Discussion

- How to construct an efficient robust estimator in linear MDPs?

Discussion

- How to construct an efficient robust estimator in linear MDPs?
- E.g. $P = \Phi\theta$, $\Phi \in \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|\times r}$ is known and $\theta \in \mathbb{R}_+^{r\times|\mathcal{S}|}$ is unknown. Offline dataset with coverage rate σ.

Discussion

- How to construct an efficient robust estimator in linear MDPs?
- E.g. $P = \Phi\theta$, $\Phi \in \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|\times r}$ is known and $\theta \in \mathbb{R}_+^{r\times|\mathcal{S}|}$ is unknown. Offline dataset with coverage rate σ.
  - Estimation of θ may be dependent on $|\mathcal{S}|$.

Discussion

- How to construct an efficient robust estimator in linear MDPs?
- E.g. $P = \Phi\theta$, $\Phi \in \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|\times r}$ is known and $\theta \in \mathbb{R}_+^{r\times|\mathcal{S}|}$ is unknown. Offline dataset with coverage rate σ.
    - Estimation of θ may be dependent on $|\mathcal{S}|$.
    - Least squares: $\mathbb{E}\|\widehat{\theta} - \theta\|_2^2 \leq \mathcal{O}(\sqrt{\frac{|\mathcal{S}|r^{5/2}}{n\sigma^2}})$. (Can we reduce it?)
- Currently the methods are model-based. ($\mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|)$ memory space).

Discussion

- How to construct an efficient robust estimator in linear MDPs?
- E.g. $P = \Phi\theta$, $\Phi \in \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|\times r}$ is known and $\theta \in \mathbb{R}_+^{r\times|\mathcal{S}|}$ is unknown. Offline dataset with coverage rate σ.
  - Estimation of θ may be dependent on $|\mathcal{S}|$.
  - Least squares: $\mathbb{E}\|\widehat{\theta} - \theta\|_2^2 \leq \mathcal{O}(\sqrt{\frac{|\mathcal{S}|r^{5/2}}{n\sigma^2}})$. (Can we reduce it?)
- Currently the methods are model-based. ($\mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|)$ memory space). Can we derive a model-free algorithm? (Tadashi and I are working on it.)

[AMK13] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen.
Minimax pac bounds on the sample complexity of reinforcement learning with a generative model.
Machine learning, 91(3):325–349, 2013.

[Sha17] Alexander Shapiro.
Distributionally robust stochastic programming.
SIAM Journal on Optimization, 27(4):2258–2275, 2017.

[WKR13] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem.
Robust markov decision processes.
Mathematics of Operations Research, 38(1):153–183, 2013.

[ZBZ+21] Zhengqing Zhou, Qinxun Bai, Zhengyuan Zhou, Linhai Qiu, Jose Blanchet, and Peter Glynn.
Finite-sample regret bound for distributionally robust offline tabular reinforcement learning.
In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 3331–3339, 2021.

Thanks!