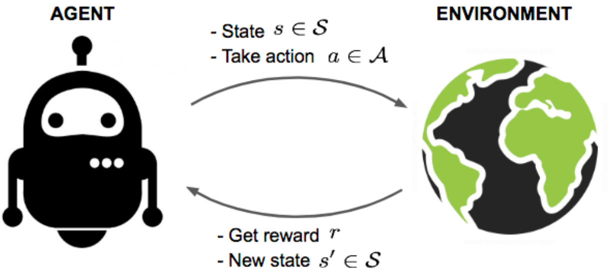
Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism

Cong Ma

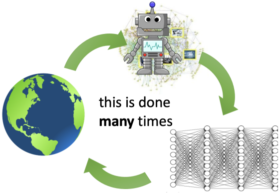
Department of Statistics, UChicago

Reinforcement learning (RL)



Goal: learn an optimal policy to maximize cumulative rewards

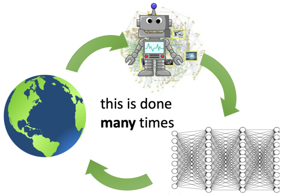
Two main paradigms of RL



Online RL

- interact with environment
- actively collect new data

Two main paradigms of RL

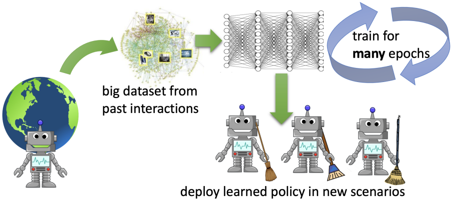


Online RL

- interact with environment
- actively collect new data

Offline/Batch RL

- no interaction
- data is given



Why offline RL?

—self-driving car

Why offline RL?

—self-driving car

online data collection

costly, dangerous, unethical

Why offline RL?

—self-driving car

online data collection

costly, dangerous, unethical

large-scale human driving data

$\Longrightarrow$ offline RL

An observation on offline RL

— and two motivating questions

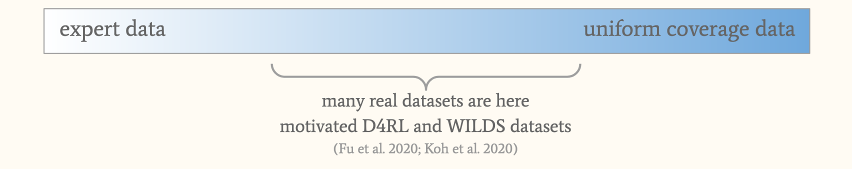
Two types of offline data

- Expert data: data from a good/optimal policy
- Uniform coverage data: data that cover state and action spaces

Disparate treatments in theory/practice

- Expert data:
  - imitation learning (imitate experts' behavior)
  - suboptimality decays at $1/N$ rate

- Uniform coverage data:
  - a different set of algorithms
  - suboptimality decays at $1/\sqrt{N}$ rate



Question 1: formulation

Question: Can we develop an offline RL framework that captures the entire data composition?

Question 1: formulation

Question: Can we develop an offline RL framework that captures the entire data composition?

Answer: Yes!

Single-policy concentrability coefficient $C^\star$:

$$C^\star \approx \text{distance}(\mu, \pi^\star)$$

—μ *corresponds to behavior data*
—$\pi^\star$ *corresponds to optimal policy*

Question 2: algorithm design

Question: Can we design an offline RL algorithm that works optimally for any data composition, without knowing $C^{\star}$?

Question 2: algorithm design

Question: Can we design an offline RL algorithm that works optimally for any data composition, without knowing $C^\star$?

Answer: Yes! This is where pessimism enters the picture

Pessimism via lower confidence bound:

$$\hat{\pi} = \arg\max_{\pi} \quad \widehat{\mathsf{LCB}}\left(J(\pi)\right)$$

— *compare to* $\pi^\star = \arg\max_{\pi} J(\pi)$

Outline

- Setup and notation
- Warm-up: multi-armed bandit
- Contextual bandit
- Markov decision process
- Conclusion and future directions

Setup and notation

Infinite-horizon Markov decision processes

MDP$(\mathcal{S}, \mathcal{A}, P, R, \rho, \gamma)$

- State space $\mathcal{S} = \{1, 2, \ldots, S\}$
- Action space $\mathcal{A} = \{1, 2, \ldots, A\}$
- Probability transition $P(s' \mid s, a)$
- Reward distributions $R(\cdot|s, a)$ on $[0, 1]$ with mean $r(s, a)$
- Initial state distribution $\rho(s)$
- Discount factor $\gamma \in [0, 1)$

Policy and value function

- Stationary deterministic policy $\pi : \mathcal{S} \mapsto \mathcal{A}$
- Value function: for all $s \in \mathcal{S}$, define

$$V^{\pi}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_t = \pi(s_t) \text{ for all } t \geq 0\right]$$

Policy and value function

- Stationary deterministic policy $\pi : \mathcal{S} \mapsto \mathcal{A}$
- Value function: for all $s \in \mathcal{S}$, define

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_t = \pi(s_t) \text{ for all } t \geq 0\right]$$

- Expected value of policy: $J(\pi) := \mathbb{E}_{s\sim\rho}[V^\pi(s)] = \sum_s \rho(s) V^\pi(s)$

Policy and value function

- Stationary deterministic policy $\pi : \mathcal{S} \mapsto \mathcal{A}$
- Value function: for all $s \in \mathcal{S}$, define

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_t = \pi(s_t) \text{ for all } t \geq 0\right]$$

- Expected value of policy: $J(\pi) := \mathbb{E}_{s\sim\rho}[V^\pi(s)] = \sum_s \rho(s) V^\pi(s)$
- There exists deterministic policy $\pi^\star$ that achieves $\max_\pi J(\pi)$

Offline learning in MDP

Given batch dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{1 \le i \le N}$, where $(s_i, a_i) \sim \mu$, $r_i \sim R(\cdot \mid s_i, a_i), s_i' \sim P(\cdot \mid s_i, a_i)$

Goal: minimize expected sub-optimality based on collected data

$$\mathbb{E}_{\mathcal{D}}\left[J(\pi^\star) - J(\hat{\pi})\right]$$

Question 1: formulation (revisited)

Question: Can we develop an offline RL framework that captures the entire data composition?

Answer: Yes!

Single-policy concentrability coefficient $C^\star$:

$$C^\star \approx \text{distance}(\mu, \pi^\star)$$

Question 1: formulation (revisited)

Question: Can we develop an offline RL framework that captures the entire data composition?

Answer: Yes!

Single-policy concentrability coefficient $C^\star$:

$$C^\star \approx \text{distance}(\mu, \pi^\star)$$

—need to translate $\pi^\star$ into distribution

Single-policy concentrability coefficient

Occupancy measure induced by $\pi^\star$

$$d^{\pi^\star}(s,a) \coloneqq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}_t(s_t = s, a_t = a; \pi^\star)$$

Single-policy concentrability coefficient

Occupancy measure induced by $\pi^\star$

$$d^{\pi^\star}(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}_t(s_t = s, a_t = a; \pi^\star)$$

Definition 1

We say $(\mu, \pi^\star)$ has $C^\star$ concentrability coefficient if

$$\max_{s,a} \quad \frac{d^{\pi^\star}(s,a)}{\mu(s,a)} \le C^\star$$

Single-policy concentrability coefficient

Occupancy measure induced by $\pi^\star$

$$d^{\pi^\star}(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}_t(s_t = s, a_t = a; \pi^\star)$$

Definition 1

We say $(\mu, \pi^\star)$ has $C^\star$ concentrability coefficient if

$$\max_{s,a} \quad \frac{d^{\pi^\star}(s,a)}{\mu(s,a)} \leq C^\star$$

- Possible values of $C^\star$: $C^\star \in [1, \infty)$
- $C^\star = 1$: expert data
- $C^\star > 1$: $\mathcal{D}$ may include "spurious" samples, i.e., state-action pairs not visited by $\pi^\star$

Offline learning in MDP (revisited)

Given batch dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{1 \le i \le N}$, where $(s_i, a_i) \sim \mu$, $r_i \sim R(\cdot \mid s_i, a_i), s_i' \sim P(\cdot \mid s_i, a_i)$

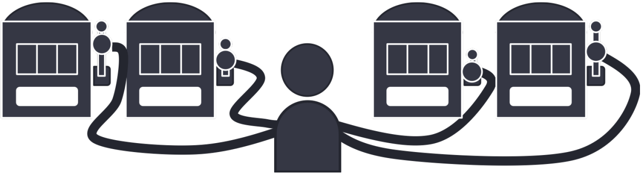
Goal: minimize expected sub-optimality based on collected data

$$\mathbb{E}_{\mathcal{D}}\left[J(\pi^\star) - J(\hat{\pi})\right]$$

Question: How does $\mathbb{E}_{\mathcal{D}}\left[J(\pi^\star) - J(\hat{\pi})\right]$ depend on $C^\star$?
Is the dependence optimal?

Warm-up: multi-armed bandit

Multi-armed bandit



- Action space: $\mathcal{A} = \{1, 2, \ldots, A\}$
- Reward distributions: $R(\cdot \mid a)$ with mean $r(a)$

—*correspond to MDP with single state and* $\gamma = 0$

Offline learning in multi-armed bandit

- Batch dataset $\mathcal{D} = \{(a_i, r_i)\}_{1 \le i \le N}$, where $a_i \sim \mu$, $r_i \sim R(\cdot \mid a_i)$
- Single-policy concentrability coefficient

$$\max_a \quad \frac{d^{\pi^\star}(a)}{\mu(a)} = \frac{1}{\mu(a^\star)} \le C^\star$$

Goal: minimize expected sub-optimality based on collected data

$$\mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\hat{a})]$$

Why empirical best arm fails?

A natural idea is to pick empirical best arm

$$\hat{a} := \arg\max_a \hat{r}(a)$$

— $\hat{r}(a)$ *empirical mean reward of arm* a

Why empirical best arm fails?

A natural idea is to pick empirical best arm

$$\hat{a} := \arg\max_{a} \hat{r}(a)$$

— $\hat{r}(a)$ *empirical mean reward of arm* a

Proposition 1

For any $\epsilon < 0.05$, $N \geq 500$, *there exists a bandit problem with two arms such that for* $\hat{a} = \text{argmax}_a \hat{r}(a)$, *one has*

$$\mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\hat{a})] \geq \epsilon.$$

Why empirical best arm fails?

A natural idea is to pick empirical best arm

$$\hat{a} := \arg\max_{a} \hat{r}(a)$$

— $\hat{r}(a)$ *empirical mean reward of arm* a

Proposition 1

For any $\epsilon < 0.05$, $N \geq 500$, *there exists a bandit problem with two arms such that for* $\hat{a} = \text{argmax}_a \hat{r}(a)$, *one has*

$$\mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\hat{a})] \geq \epsilon.$$

- Empirical best arm is sensitive to arms with few observations
- This happens even when $C^\star$ is small

Pessimism via lower confidence bound

Lessons learned from failure of empirical best arm

- Should not treat arms equally
- Need to be pessimistic about arms with few observations

Pessimism via lower confidence bound

Lessons learned from failure of empirical best arm

- Should not treat arms equally
- Need to be pessimistic about arms with few observations

Lower confidence bound for bandit: fix some $L > 0$, return

$$\hat{a} \coloneqq \arg\max_{a} \quad \hat{r}(a) - \frac{L}{\sqrt{N(a) \vee 1}}$$

—$N(a)$ *number of times arm* a *is seen*

A closer look at LCB

Lower confidence bound for bandit: fix some $L > 0$, return

$$\hat{a} := \arg\max_a \quad \hat{r}(a) - \frac{L}{\sqrt{N(a) \vee 1}}$$

—$N(a)$ *number of times arm a is seen*

- $\frac{L}{\sqrt{N(a) \vee 1}}$ is large when $N(a)$ is small
- View $\hat{r}(a) - \frac{L}{\sqrt{N(a) \vee 1}}$ as lower confidence bound of $r(a)$
- $\frac{L}{\sqrt{N(a) \vee 1}}$ arises from Hoeffding concentration inequality

Performance guarantees

Theorem 2

Set $L \asymp \sqrt{\log(AN)}$. Policy $\hat{a}$ returned by LCB algorithm obeys

$$\mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\hat{a})] \lesssim \sqrt{\frac{C^\star}{N}}$$

- LCB beats empirical best arm
- Performance of LCB degrades gracefully w.r.t. $C^\star$

Is LCB optimal for offline bandits?

— resort to minimax lower bounds in Statistics

Define problem class

$$\mathsf{MAB}(C^\star) = \{(\mu, R) \mid \frac{1}{\mu(a^\star)} \le C^\star\}$$

Is LCB optimal for offline bandits?

— *resort to minimax lower bounds in Statistics*

Define problem class

$$\mathsf{MAB}(C^\star) = \{(\mu, R) \mid \frac{1}{\mu(a^\star)} \le C^\star\}$$

Theorem 3

When $C^\star \ge 2$, one has

$$\inf_{\hat{a}} \sup_{\mathsf{MAB}(C^\star)} \mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\hat{a})] \gtrsim \sqrt{\frac{C^\star}{N}}$$

Is LCB optimal for offline bandits?

— resort to minimax lower bounds in Statistics

Define problem class

$$\mathsf{MAB}(C^\star) = \{(\mu, R) \mid \frac{1}{\mu(a^\star)} \leq C^\star\}$$

Theorem 3

When $C^\star \geq 2$, one has

$$\inf_{\hat{a}} \sup_{\mathsf{MAB}(C^\star)} \mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\hat{a})] \gtrsim \sqrt{\frac{C^\star}{N}}$$

When $C^\star \in (1, 2)$, one has

$$\inf_{\hat{a}} \sup_{\mathsf{MAB}(C^\star)} \mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\hat{a})] \gtrsim \exp\left(-N(2 - C^\star) \cdot \log\left(\frac{2}{C^\star - 1}\right)\right)$$

Imitation learning is better when $C^\star \in (1, 2)$

When $C^\star \in (1, 2)$, one has $\mu(a^\star) > 1/2$. Reasonable to pick most played arm

$$\hat{a} = \mathsf{argmax}_a \, N(a)$$

—$N(a)$ *number of times arm* a *is seen*

Imitation learning is better when $C^\star \in (1, 2)$

When $C^\star \in (1, 2)$, one has $\mu(a^\star) > 1/2$. Reasonable to pick most played arm

$$\hat{a} = \mathsf{argmax}_a \, N(a)$$

—$N(a)$ *number of times arm a is seen*

Proposition 2

Assume that $C^\star \in [1, 2)$. For $\hat{a} = \mathsf{argmax}_a \, N(a)$, we have

$$\mathbb{E}_{\mathcal{D}}[r(a^\star) - r(\hat{a})] \le \exp\left(-N \cdot \mathsf{KL}\left(\mathrm{Bern}\left(\tfrac{1}{2}\right) \,\middle\|\, \mathrm{Bern}\left(\tfrac{1}{C^\star}\right)\right)\right).$$

- Matches the exponential rate

Non-adaptivity of LCB for bandit

Recall LCB for bandit

$$\hat{a} := \arg\max_{a} \quad \hat{r}(a) - \frac{L}{\sqrt{N(a) \vee 1}}$$

We showed with $L \asymp \sqrt{\log N}$, LCB is optimal for $C^\star \geq 2$

Can LCB with $L \asymp \sqrt{\log N}$ be optimal for $C^\star \in (1, 2)$?

Non-adaptivity of LCB for bandit

Recall LCB for bandit

$$\hat{a} := \arg\max_a \quad \hat{r}(a) - \frac{L}{\sqrt{N(a) \vee 1}}$$

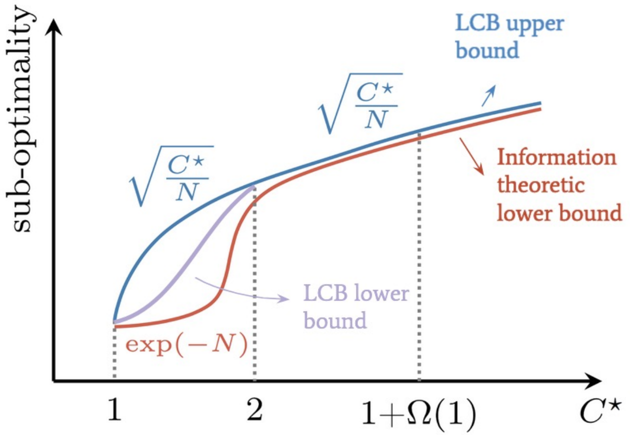
We showed with $L \asymp \sqrt{\log N}$, LCB is optimal for $C^\star \geq 2$

Can LCB with $L \asymp \sqrt{\log N}$ be optimal for $C^\star \in (1, 2)$?

—*No*

- LCB cannot achieve $\exp(-N)$ with $L \asymp \sqrt{\log n}$ when $C^\star \in (1, 2)$
- Need to set $L \asymp N$ to achieve $\exp(-N)$ rate; however this choice fails to yield $1/\sqrt{N}$ rate when $C^\star \geq 2$

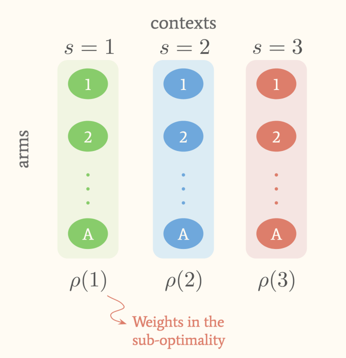
Summary of LCB for bandit



case when $L \asymp \sqrt{\log N}$

Contextual bandit

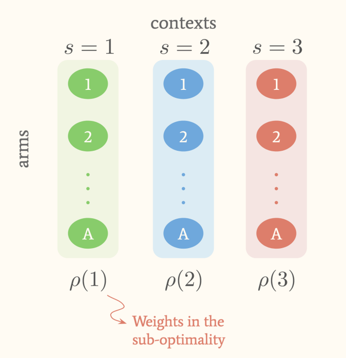
Contextual bandit



- State space $\mathcal{S} = \{1, 2, \ldots, S\}$
- Action space $\mathcal{A} = \{1, 2, \ldots, A\}$
- Reward distributions $R(\cdot \mid s, a)$ with mean $r(s, a)$
- Batch dataset $\mathcal{D} = \{(s_i, a_i, r_i)\}_{1 \le i \le N}$, where $(s_i, a_i) \sim \mu$, $r_i \sim R(\cdot \mid s_i, a_i)$

— *correspond to MDP with* $\gamma = 0$

Contextual bandit



- State space $\mathcal{S} = \{1, 2, \ldots, S\}$
- Action space $\mathcal{A} = \{1, 2, \ldots, A\}$
- Reward distributions $R(\cdot \mid s, a)$ with mean $r(s, a)$
- Batch dataset $\mathcal{D} = \{(s_i, a_i, r_i)\}_{1 \le i \le N}$, where $(s_i, a_i) \sim \mu$, $r_i \sim R(\cdot \mid s_i, a_i)$

— *correspond to MDP with* $\gamma = 0$

Goal: minimize expected sub-optimality based on collected data

$$\mathbb{E}_{\mathcal{D}}\left[J(\pi^\star) - J(\hat{\pi})\right]$$

Assumption and algorithm

- Single-policy concentrability coefficient

$$\max_{s} \; \frac{\rho(s)}{\mu(s, \pi^{\star}(s))} \leq C^{\star}$$

- LCB algorithm: fix some $L > 0$, return

$$\hat{\pi}(s) \coloneqq \arg\max_{a} \quad \hat{r}(s,a) - \frac{L}{\sqrt{N(s,a) \vee 1}}$$

Performance guarantees

Theorem 4

Consider $S \geq 2$. *Set* $L \asymp \sqrt{\log(SAN)}$. *Policy* $\hat{\pi}$ *returned by LCB algorithm obeys*

$$\mathbb{E}_{\mathcal{D}}\left[J(\pi^\star) - J(\hat{\pi})\right] \lesssim \sqrt{\frac{S(C^\star - 1)}{N}} + \frac{S}{N}$$

Remarks:

- When $C^\star$ is close to 1, $1/N$ rate, as in imitation learning
- When $C^\star$ is large, $1/\sqrt{N}$ rate, as for uniform coverage data
- Rate smoothly transitions from $1/N$ to $1/\sqrt{N}$ as $C^\star$ increases

Heuristic argument

Sub-optimality bound of LCB for contextual bandit

$$\mathbb{E}_{\mathcal{D}}\left[J(\pi^\star) - J(\hat{\pi})\right] \lesssim \sqrt{\frac{S(C^\star - 1)}{N}} + \frac{S}{N}$$

In particular, we would like to understand

- What are sources of error?
- Why not $\sqrt{\frac{SC^\star}{N}}$?

Source 1: missing mass

When $N(s, \pi^\star(s)) = 0$,

Source 1: missing mass

When $N(s, \pi^\star(s)) = 0$, one has error

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\sum_s \rho(s)\left[r(s,\pi^\star(s)) - r(s,\hat{\pi}(s))\right]\mathbb{1}\{N(s,\pi^\star(s))=0\}\right] \\
&\le \mathbb{E}_{\mathcal{D}}\left[\sum_s \rho(s)\mathbb{1}\{N(s,\pi^\star(s))=0\}\right] \\
&= \sum_s \rho(s)(1-\mu(s,\pi^\star(s)))^N \\
&\le \sum_s C^\star \mu(s,\pi^\star(s))(1-\mu(s,\pi^\star(s)))^N \lesssim \frac{SC^\star}{N}
\end{aligned}
$$

—$\max_{x\in[0,1]} x(1-x)^N \le 4/(9N)$

Source 1: missing mass

When $N(s, \pi^\star(s)) = 0$, one has error

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\sum_s \rho(s)\left[r(s,\pi^\star(s)) - r(s,\hat{\pi}(s))\right]\mathbb{1}\{N(s,\pi^\star(s)) = 0\}\right] \\
&\le \mathbb{E}_{\mathcal{D}}\left[\sum_s \rho(s)\mathbb{1}\{N(s,\pi^\star(s)) = 0\}\right] \\
&= \sum_s \rho(s)(1-\mu(s,\pi^\star(s)))^N \\
&\le \sum_s C^\star \mu(s,\pi^\star(s))(1-\mu(s,\pi^\star(s)))^N \lesssim \frac{SC^\star}{N}
\end{aligned}
$$

—$\max_{x\in[0,1]} x(1-x)^N \le 4/(9N)$

—*need* $S \ge 2$

Source 2: estimation error

When $N(s, \pi^\star(s)) \geq 1$, one has $|r(s,a) - \hat{r}(s,a)| \lesssim \frac{1}{\sqrt{N(s,a)}}$

Source 2: estimation error

When $N(s,\pi^\star(s)) \geq 1$, one has $|r(s,a) - \hat{r}(s,a)| \lesssim \frac{1}{\sqrt{N(s,a)}}$

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[J(\pi^\star) - J(\hat{\pi})\right] &= \mathbb{E}_{\mathcal{D},\rho}\left[r(s,\pi^\star(s)) - r(s,\hat{\pi}(s))\right] \\
&\lesssim \mathbb{E}_{\mathcal{D},\rho}\left[\frac{1}{\sqrt{N(s,\pi^\star(s))}}\right] \\
&\approx \mathbb{E}_{\rho}\left[\frac{1}{\sqrt{N\mu(s,\pi^\star(s))}}\right] \\
&= \sum_s \rho(s)\frac{1}{\sqrt{N\mu(s,\pi^\star(s))}} \lesssim \sqrt{\frac{SC^\star}{N}}
\end{aligned}$$

Source 2: estimation error

When $N(s, \pi^\star(s)) \geq 1$, one has $|r(s,a) - \hat{r}(s,a)| \lesssim \frac{1}{\sqrt{N(s,a)}}$

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[J(\pi^\star) - J(\hat{\pi})\right] &= \mathbb{E}_{\mathcal{D},\rho}\left[r(s, \pi^\star(s)) - r(s, \hat{\pi}(s))\right] \\
&\lesssim \mathbb{E}_{\mathcal{D},\rho}\left[\frac{1}{\sqrt{N(s, \pi^\star(s))}}\right] \\
&\approx \mathbb{E}_{\rho}\left[\frac{1}{\sqrt{N\mu(s, \pi^\star(s))}}\right] \\
&= \sum_s \rho(s) \frac{1}{\sqrt{N\mu(s, \pi^\star(s))}} \lesssim \sqrt{\frac{SC^\star}{N}}
\end{aligned}
$$

—hmm, where is $C^\star - 1$?

Where does $C^\star - 1$ come from?

Key observation: instead of

$$r(s, \pi^\star(s)) - r(s, \hat{\pi}(s)) \lesssim \frac{1}{\sqrt{N(s, \pi^\star(s))}}$$

One actually has

$$r(s, \pi^\star(s)) - r(s, \hat{\pi}(s)) \lesssim \frac{1}{\sqrt{N(s, \pi^\star(s))}} \mathbb{1}\{\hat{\pi}(s) \neq \pi^\star(s)\}$$

Where does $C^\star - 1$ come from?

Key observation: instead of

$$r(s,\pi^\star(s)) - r(s,\hat{\pi}(s)) \lesssim \frac{1}{\sqrt{N(s,\pi^\star(s))}}$$

One actually has

$$r(s,\pi^\star(s)) - r(s,\hat{\pi}(s)) \lesssim \frac{1}{\sqrt{N(s,\pi^\star(s))}} \mathbb{1}\{\hat{\pi}(s) \neq \pi^\star(s)\}$$

Identify clean set $\mathcal{S}_{\text{clean}}$ such that for $s \in \mathcal{S}_{\text{good}}$, $\hat{\pi}(s) = \pi^\star(s)$ with high prob.,

Where does $C^\star - 1$ come from?

Key observation: instead of

$$r(s,\pi^\star(s)) - r(s,\hat{\pi}(s)) \lesssim \frac{1}{\sqrt{N(s,\pi^\star(s))}}$$

One actually has

$$r(s,\pi^\star(s)) - r(s,\hat{\pi}(s)) \lesssim \frac{1}{\sqrt{N(s,\pi^\star(s))}} \mathbb{1}\{\hat{\pi}(s) \neq \pi^\star(s)\}$$

Identify clean set $\mathcal{S}_{\text{clean}}$ such that for $s \in \mathcal{S}_{\text{good}}$, $\hat{\pi}(s) = \pi^\star(s)$ with high prob., and

$$\sum_{s \notin \mathcal{S}_{\text{clean}}} \rho(s) \frac{1}{\sqrt{N\mu(s,\pi^\star(s))}} \lesssim \sqrt{\frac{S(C^\star - 1)}{N}}$$

Optimality of LCB in offline contextual bandit

As before, define problem class

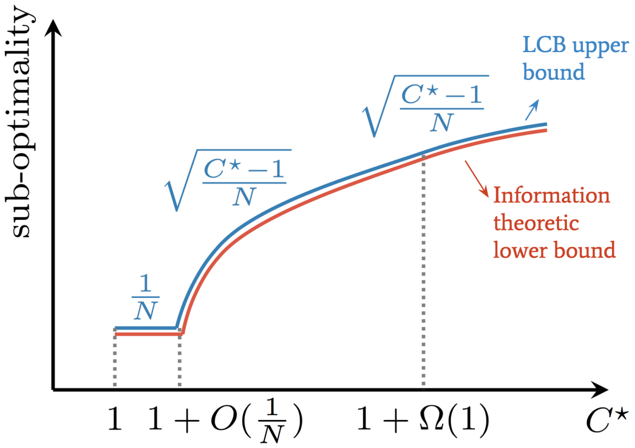
$$\mathsf{CB}(C^\star) := \{(\rho, \mu, R) \mid \max_s \frac{\rho(s)}{\mu(s, \pi^\star(s))} \leq C^\star\}$$

Theorem 5

Assume that $S \geq 2$. For any $C^\star \geq 1$, one has

$$\inf_{\hat{\pi}} \sup_{\mathsf{CB}(C^\star)} \mathbb{E}_{\mathcal{D}}[J(\pi^\star) - J(\hat{\pi})] \gtrsim \sqrt{\frac{S(C^\star - 1)}{N}} + \frac{S}{N}$$

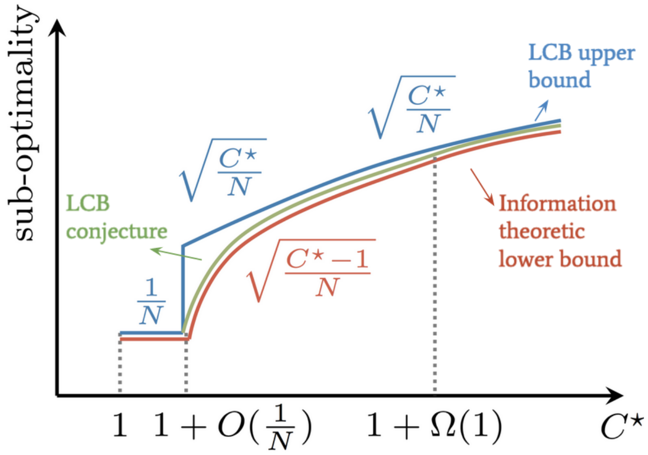
Summary of LCB in offline contextual bandits



LCB achieves optimality without knowing $C^\star$

Markov decision process

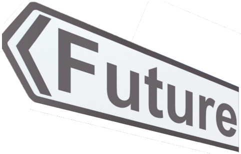
One-page result for MDP



- Combine value iteration with LCB
- Hoeffding confidence bounds yield sub-optimal dependence on $\frac{1}{1-\gamma}$

Future directions

- Close the gap in MDP
- Other measures of quality of behavior data
- Extensions to continuous state-action space and function approximation



Paper:
"Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism," to appear in Neurips 2021,

P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, S. Russell, arXiv:2103.12021