Offline Reinforcement Learning with Realizability and Single-policy Concentrability

Jason D. Lee

February 24, 2022
Wenhao Zhan  
Princeton

Baihe Huang  
PKU

Audrey Huang  
UIUC

Nan Jiang  
UIUC

Figures borrowed from Yuxin Chen, Shicong Cen, and Simon Du.
Recent successes in RL
Markov decision process (MDP)

- A collection of MABs indexed by state $s \in S$.
- At time step $t$, an agent observes the state $s_t$, selects an action $a_t \sim \pi(\cdot|s_t)$, and then receives a reward $r(s_t, a_t)$.
- The environment transitions to a new state $s_{t+1} \sim P(\cdot|s_t, a_t)$. 

![Diagram showing state $s_t$, agent, and state transition](image-url)
Markov decision process (MDP)

- A collection of MABs indexed by state $s \in S$.
- At time step $t$, an agent observes the state $s_t$, selects an action $a_t \sim \pi(\cdot|s_t)$, and then receives a reward $r(s_t, a_t)$.
- The environment transitions to a new state $s_{t+1} \sim P(\cdot|s_t, a_t)$.
Markov decision process (MDP)

- A collection of MABs indexed by state $s \in S$.
- At time step $t$, an agent observes the state $s_t$, selects an action $a_t \sim \pi(\cdot|s_t)$, and then receives a reward $r(s_t, a_t)$.
- The environment transitions to a new state $s_{t+1} \sim P(\cdot|s_t, a_t)$.
Value function and state-action (Q) function of policy $\pi$:

$$
\forall s \in S : \quad V^\pi(s) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s \right]
$$

$$
\forall (s, a) \in S \times A : \quad Q^\pi(s, a) := \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a \right]
$$

- Long-term discounted reward: $\gamma \in [0, 1)$ is the discount factor
- Expectation is w.r.t. the sampled trajectory under $\pi$
Reinforcement learning (RL)

Reinforcement Learning: **online vs offline**

**online**
- Agent
- Environment
- Action
- Reward, state

**offline**
- Agent
- Dataset
- Environment
- Action
- Reward
- State

---

**Summary**
- **Online Learning:** The agent interacts directly with the environment to receive feedback.
- **Offline Learning:** The agent learns from a dataset that has already been collected, without direct interaction with the environment.
Reinforcement Learning: **online vs offline**

**online**
- Agent
- Environment
- Action
- Reward, State

**offline**
- Agent
- Dataset
- Environment
- Action
- Reward
- State

**offline:** no interaction with the environment!
Reinforcement learning (RL)

Challenges in RL: big $S$!

Go game: $\gtrsim 10^{700}$ states

Mario: $256^{256 \times 400}$
Reinforcement learning (RL)

Challenges in RL: big $\mathcal{S}$!

Go game: $\gtrsim 10^{700}$ states

Mario: $256^{256 \times 400}$

How to design provably efficient methods for RL?
Surely, RL has been solved?

**Best result** $B^*SH^3/\epsilon^2$ for Mario:

$10^{250000}$

$\frac{1}{12}$ of the output!

$XJWXB21$, $B^*$ is some measure of distribution shift.
Function Approximation

\[ f \in \mathcal{F} \]

\[
\begin{cases}
\text{Linear} \\
\text{Kernel} \\
\text{Neural Network}
\end{cases}
\]

With \( O\left(\frac{\log |\mathcal{F}|}{\epsilon^2}\right) \) samples we can learn \( \epsilon \)-optimal predictor by ERM.

\(|\mathcal{F}|: \) cardinality of \( \mathcal{F} \).
Let’s first look at Online RL + Function Approximation

Huge slew of negative results:

- Linear function approximation even with gap conditions is hard
- Simplest neural net function approximation is hard

Positive results:

- Bilinear classes is essentially the broadest class.
- Almost all positive results rely on **elliptic potential** lemma, so are linear in some way.

---

*WAJAYJS21, WWK21
†DYM21
‡DKLLMSW
Let’s first look at Online RL + Function Approximation

Huge slew of negative results:

- Linear function approximation even with gap conditions is hard*
- Simplest neural net function approximation is hard †

Positive results:

- Bilinear classes‡ is essentially the broadest class.
- Almost all positive results rely on elliptic potential lemma, so are linear in some way.

Basically only Linear Online RL is possible.

---

* WAJAYJS21, WWK21
† DYM21
‡ DKLLMSW
Is offline RL harder than online RL?

- After the bilinear paper, I became depressed about online/offline RL.
- My reasoning: offline RL is harder than online RL, and online is already impossible.
Is offline RL harder than online RL?

- After the bilinear paper, I became depressed about online/offline RL.
- My reasoning: offline RL is harder than online RL, and online is already impossible.

So, I went to work on the simulator setting where you can use Neural Nets*.

*HHKLLWa21,HHKLLWb21
Is offline RL harder than online RL?

• After the bilinear paper, I became depressed about online/offline RL.
• My reasoning: offline RL is harder than online RL, and online is already impossible.

Wait, you can aim lower in offline RL!

*HHKLLWa21,HHKLLWb21*
Easier Problem: Transfer Learning

Density Ratio $B^* := \max_x \frac{p_{tgt}(x)}{p_{src}(x)}$. For many function classes (e.g. kernel methods), the *transfer difficulty* is characterized by density ratio*:

$$\minimax \asymp \left(\frac{B^*/n}{c}\right)^c,$$

$c$ is the exponent without distribution shift.

**Analogous result for Offline RL**

The best you can hope for is $B^* \frac{\log|\mathcal{F}|\text{poly}(\frac{1}{1-\gamma})}{\epsilon^c}$, and all the hard part of online RL is hidden in $B^*$.

TLDR: Offline RL is easier, because we can aim lower!

*MPW2022*
Model and Notations

Model:

- infinite horizon MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu_0\}$.
- offline dataset $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$ where $(s_i, a_i) \sim d^D$, $r_i = r(s_i, a_i)$, $s'_i \sim P(\cdot|s_i, a_i)$.
- $d^D$ is unknown. Denote $d^D(a|s)$ by $\pi_D(a|s)$.
- $\mu_0$ is unknown: Assume access to i.i.d. samples $\mathcal{D}_0 = \{s_{0,j}\}_{j=1}^{n_0}$ from $\mu_0$.
Model and Notations

Model:
- infinite horizon MDP $\mathcal{M} = \{S, A, P, r, \gamma, \mu_0\}$.
- offline dataset $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^n$ where $(s_i, a_i) \sim d^D$, $r_i = r(s_i, a_i)$, $s'_i \sim P(\cdot|s_i, a_i)$.
- $d^D$ is unknown. Denote $d^D(a|s)$ by $\pi_D(a|s)$.
- $\mu_0$ is unknown: Assume access to i.i.d. samples $\mathcal{D}_0 = \{s_{0,j}\}_{j=1}^{n_0}$ from $\mu_0$.

Notations:
- $d^\pi$: discounted state visitation probability under policy $\pi$.
- $Q^\pi(s, a) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a, \pi \right]$.
- $V^\pi(s) = \mathbb{E} \left[ \sum_{t=0}^{\infty} r(s_t, a_t) | s_0 = s, \pi \right]$. 
Offline RL should be easy right?

What should $\mathcal{F}$ approximate?
Offline RL should be easy right?

What should $\mathcal{F}$ approximate?

**Value Function Approximation:** Approximate $Q^*$ via function class $\mathcal{F}$. 
Offline RL should be easy right?

What should $\mathcal{F}$ approximate?

**Value Function Approximation:** Approximate $Q^*$ via function class $\mathcal{F}$.

Can we attain $\text{poly}(B^*, \log |\mathcal{F}|, \frac{1}{\epsilon}, \frac{1}{1-\gamma})$ sample complexity to find optimal policy?
In concurrent work\(^*\), this has been shown to be impossible.

**Theorem (FKSI\text{X}21)**

There is a family of MDPs (with \( A = 2 \), \( B^* \leq 16 \), and realizable value function \( |F| = 2 \)) such that any algorithm needs \( n \geq S^{1/3} \) to attain

\[
J(\pi^*) - J(\hat{\pi}) \geq \frac{.01}{1 - \gamma}.
\]

Similar lower bound holds even under strong concentrability (all-policy concentrability).

First conjectured by Chen and Jiang in 2019.

\(^*\)FKSI\text{X}21
Should we give up?

The whole point is to break lower bounds!

Potential Assumptions:

• Completeness
• Super strong Concentrability
Completeness

Function class is closed under Bellman update:
For all $f \in \mathcal{F}$, $Tf \in \mathcal{F}$.

What is wrong with this?

- Non-monotone: increasing the approximation power of $\mathcal{F}$ may cause completeness to be more violated.
- Pretrained representation are realizable, yet do not work empirically under distribution shift in algorithms that require completeness*.

*WFK22
What if $\mathcal{F}$ is universal?

But my $\mathcal{F}$ is universal, so it has to be complete!
What if $\mathcal{F}$ is universal?

But my $\mathcal{F}$ is universal, so it has to be complete!
What if $\mathcal{F}$ is universal?

**But my $\mathcal{F}$ is universal, so it has to be complete!**

**NO!!!!!!**

- Have to use function classes of bounded complexity (e.g. RKHS norm ball, finite-capacity network)
- Bellman operator may not preserve the bounded complexity.
Algorithms that work with Completeness

• Approximate Dynamic Programming* (Fitted Q Iteration)
• Minimax FQI †
• Bellman-consistent Pessimism‡
• Many others...

*EGW05, CJ19
†CJ19
‡XCJMA21
Concentrability

Many types of distribution ratio/concentrability:

• Single-policy: \( \| \frac{d\pi^*}{d\mathcal{D}} \|_\infty \leq B^* \)

• All-policy: \( \| \frac{d\pi}{d\mathcal{D}} \|_\infty \leq B^\pi \) for all \( \pi \)

• Super-strong: \( \| \frac{p(.|s,a)}{d^\mathcal{P}(.)} \|_\infty \leq B^P \) for all \( s, a \)
Positive result under super-strong assuptions

Only positive result under realizability* is from Chen and Jiang:

\[ n \geq \text{poly}(B^P, \frac{1}{\epsilon}, \frac{1}{1 - \gamma}) \]

*Not comparing to model-based methods, since realizable implies completeness.
Only positive result under realizability* is from Chen and Jiang:

\[ n \geq \text{poly}(B^P, \frac{1}{\epsilon}, \frac{1}{1 - \gamma}) \]

When does this hold?

• Known example is when dynamics $P$ have low non-negative rank and $\mu$ is average of the rows of $P(s')$.

*Not comparing to model-based methods, since realizable implies completeness.
Transfer learning is possible under the weakest density ratio condition:

\[ \left\| \frac{p_{\text{tgt}}}{p_{\text{src}}} \right\|_\infty \leq B^* \text{ equiv to } \left\| \frac{d\pi^*}{dD} \right\|_\infty \leq B^* \]
Pessimism

Pessimism is a recently developed technique that allows us to use single-point density ratio:

- Pioneered in Linear MDP*
- Bellman-consistent Pessimism for general function class (under completeness) †
- All known algorithms that allow single-point or all-policy ratio require completeness.

---

*JYY20, earlier works also use it, but do not analyze.
†XJCJMA21
Challenges in offline RL

- Distribution shift → **Super strong concentrability**
- Function approximation → **Bellman-completeness**

Both assumptions are very **strong** and are **violated** in practice!
Challenges in offline RL

• Distribution shift $\rightarrow$ Super strong concentrability
• Function approximation $\rightarrow$ Bellman-completeness

Both assumptions are very strong and are violated in practice!

Is sample-efficiency possible with realizability and single-policy concentrability?
Back to the basics: LP

**Dual LP**

\[
\begin{align*}
\max_{d \geq 0} & \quad \mathbb{E}_{(s,a) \sim d}[r(s,a)] \\
\text{s.t.} & \quad d(s) = (1 - \gamma)\mu_0(s) + \gamma \sum_{s',a'} P(s'|s,a')d(s',a')
\end{align*}
\]

(1)

(2)

where \(d \in \mathbb{R}^{|S \times A|}, \quad d(s) = \sum_a d(s,a)\),

Bellman flow constraints \(\iff\) \(d\) is induced by a policy \(\pi\).
Primal-dual LP for MDPs

\[
\max_{d \geq 0} \min_v L_\alpha(v, w) := (1 - \gamma) \mathbb{E}_{s \sim \mu_0}[v(s) + \mathbb{E}_{(s,a) \sim d}[e_v(s, a)],
\]

where \( e_v(s, a) = r(s, a) + \gamma \sum_{P(s'|s,a)} v(s') - v(s) \).

• Inspired by bilinear \( \pi \)-learning* and OptiDice†

*W17, W19
†LJPLK21
Offline primal-dual

Change of variables: \( w(s, a) = \frac{d(s,a)}{d^D(s,a)} \)

Offline primal-dual LP for MDPs

\[
\max_{w \geq 0} \min_v L_\alpha(v, w) := (1 - \gamma) \mathbb{E}_{s \sim \mu_0}[v(s)] + \mathbb{E}_{(s, a) \sim d^D}[w(s, a)e_v(s, a)].
\]

Computable from samples!
Difficulties with primal-dual

\[
\max_{w \geq 0} \min_v L_\alpha(v, w) := (1 - \gamma) \mathbb{E}_{s \sim \mu_0}[v(s)] + \mathbb{E}_{(s,a) \sim d^D}[w(s, a)e_v(s, a)].
\]

- Not strongly concave in \( w \), so no uniqueness.
- Nature can randomize over instances, to force errors when there is zeroes in \( w \) (counterexample in the paper).
Problem: Regularized Maximin

\[
\max_{w \geq 0} \min_v L_\alpha(v, w) := (1 - \gamma) \mathbb{E}_{s \sim \mu_0}[v(s)] - \alpha \mathbb{E}_{(s, a) \sim \mathcal{D}}[f(w(s, a))] \\
+ \mathbb{E}_{(s, a) \sim \mathcal{D}}[w(s, a)e_v(s, a)],
\]

where \(e_v(s, a) = r(s, a) + \gamma \sum_{P(s'|s, a)} v(s') - v(s)\).

Denote the optimizer as \((v^*_\alpha, w^*_\alpha)\).
• Policy optimization:  \( \max_{\pi} J(\pi) = \mathbb{E}_{(s,a) \sim d_{\pi}}[r(s,a)] \).

• Density Regularization:

\[
\max_{\pi} J_{D,f}(\pi) = \mathbb{E}_{(s,a) \sim d_{\pi}}[r(s,a)] - \alpha D_f(d_{\pi} || d^D),
\]

where \( \alpha > 0, \ D_f(d_{\pi} || d^D) = \mathbb{E}_{(s,a) \sim d^D} \frac{d_{\pi}(s,a)}{d^D(s,a)} \) is an \( f \)-divergence.
Interpretation: Density Regularization

• Policy optimization: \( \max_{\pi} J(\pi) = \mathbb{E}_{(s,a) \sim d^\pi} [r(s,a)] \).

• Density Regularization:

\[
\max_{\pi} J_{D,f}(\pi) = \mathbb{E}_{(s,a) \sim d^\pi} [r(s,a)] - \alpha D_f (d^\pi \| d^D),
\]

where \( \alpha > 0 \), \( D_f (d^\pi \| d^D) = \mathbb{E}_{(s,a) \sim d^D} \left[ \frac{d^\pi(s,a)}{d^D(s,a)} \right] \) is an \( f \)-divergence.

Encourages \( d^\pi \) to stay close to \( d^D \).

• Suggested explanation from DICE family of algorithms and most offline algorithms.
Uniqueness: Density regularization leads to strong concavity in the primal-dual, and thus unique $w^*_\alpha$. Suppose $d^*_\alpha$ is the optimum of the regularized LP, then we can extract the regularized optimal policy $\pi^*_\alpha$ via:

$$\pi^*_\alpha(s|a) := \begin{cases} \frac{d^*_\alpha(s,a)}{\sum_a d^*_\alpha(s,a)}, & \text{for } \sum_a d^*_\alpha(s,a) > 0, \forall s \in S, a \in A. \\ \frac{1}{|A|}, & \text{else.} \end{cases}$$
Uniqueness: Density regularization leads to strong concavity in the primal-dual, and thus unique $w^*_\alpha$. Suppose $d^*_\alpha$ is the optimum of the regularized LP, then we can extract the regularized optimal policy $\pi^*_\alpha$ via:

$$
\pi^*_\alpha(s|a) := \begin{cases} 
\frac{d^*_\alpha(s,a)}{\sum_a d^*_\alpha(s,a)}, & \text{for } \sum_a d^*_\alpha(s,a) > 0, \\
\frac{1}{|A|}, & \text{else.}
\end{cases} \forall s \in \mathcal{S}, a \in \mathcal{A}.
$$

When $\alpha > 0$ and $f$ is strongly-convex, $d^*_\alpha$ and $\pi^*_\alpha$ are unique!
**Function classes:** \( \mathcal{V} \subseteq \mathbb{R}^{|S|} \) and \( \mathcal{W} \subseteq \mathbb{R}_+^{|S| \times |A|} \)

**Algorithm:** PRO-RL

\[
(\hat{w}, \hat{v}) = \arg \max_{w \in \mathcal{W}} \arg \min_{v \in \mathcal{V}} \hat{L}_\alpha(v, w),
\]

(4)

where

\[
\hat{L}_\alpha(v, w) := (1 - \gamma) \frac{1}{n_0} \sum_{j=1}^{n_0} [v(s_{0,j})] + \frac{1}{n} \sum_{i=1}^{n} [-\alpha f(w(s_i, a_i))]
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} [w(s_i, a_i)e_v(s_i, a_i, r_i, s_i')],
\]

(5)

and \( e_v(s, a, r, s') = r + \gamma v(s') - v(s) \).

Denote the optimizer as \((v^*_\alpha, w^*_\alpha)\).
Assume $\pi_D$ is known for now, $d^D(s, a) = d^D(s)\pi_D(a|s)$. Then the final learned policy is:

$$\hat{\pi}(a|s) = \begin{cases} 
\frac{\hat{w}(s,a)\pi_D(a|s)}{\sum_{a'} \hat{w}(s,a')\pi_D(a'|s)}, & \text{for } \sum_{a'} \hat{w}(s,a')\pi_D(a'|s) > 0, \\
\frac{1}{|A|}, & \text{else},
\end{cases}$$
Assume $\pi_D$ is known for now, $d^D(s, a) = d^D(s)\pi_D(a|s)$. Then the final learned policy is:

$$\hat{\pi}(a|s) = \begin{cases} \frac{\hat{w}(s,a)\pi_D(a|s)}{\sum_{a'} \hat{w}(s,a')\pi_D(a'|s)}, & \text{for } \sum_{a'} \hat{w}(s,a')\pi_D(a'|s) > 0, \\ \frac{1}{|A|}, & \text{else,} \end{cases}$$

When $\pi_D$ is unknown, use behavior cloning to extract the policy!
Assumptions

- **Concentrability**: \( \frac{d^*_\alpha(s,a)}{d\mathcal{D}(s,a)} \leq B^\alpha_w, \forall s \in \mathcal{S}, a \in \mathcal{A}. \)

- **Realizability**: \( v^*_\alpha \in \mathcal{V}, w^*_\alpha \in \mathcal{W}. \)
Assumptions

- **Concentrability**: \( \frac{d_{\alpha}(s,a)}{dD(s,a)} \leq B_w^{\alpha}, \forall s \in S, a \in A. \)

- **Realizability**: \( v^*_\alpha \in V, w^*_\alpha \in W. \)

- **Properties of \( f \):**
  - Strong Convexity: \( f \) is \( M_f \)-strongly-convex,
  - Boundedness: \( |f'(x)| \leq B_{f',\alpha}, |f(x)| \leq B_{f,\alpha}, \forall 0 \leq x \leq B^\alpha_w. \)
  - Non-negativity: \( f(x) \geq 0, \forall x \in \mathbb{R}. \)
Assumptions

• **Concentrability**: \( \frac{d^{\star}(s,a)}{dD(s,a)} \leq B^\alpha_w, \forall s \in S, a \in A. \)

• **Realizability**: \( v^\star_\alpha \in V, w^\star_\alpha \in W. \)

• **Properties of \( f \):**
  - Strong Convexity: \( f \) is \( M_f \)-strongly-convex,
  - Boundedness: \( |f'(x)| \leq B_{f',\alpha}, |f(x)| \leq B_{f,\alpha}, \forall 0 \leq x \leq B^\alpha_w. \)
  - Non-negativity: \( f(x) \geq 0, \forall x \in \mathbb{R}. \)

• **Boundedness of the function classes:**
  - \( 0 \leq w(s,a) \leq B^\alpha_w, \forall s \in S, a \in A, w \in W, \)
  - \( \|v\|_\infty \leq B_{v,\alpha} := \frac{\alpha B_{f',\alpha} + 1}{1-\gamma}, \forall v \in V. \)
Assumptions

• **Concentrability**: \( \frac{d^*_\alpha(s,a)}{dD(s,a)} \leq B^\alpha_w, \forall s \in S, a \in A. \)

• **Realizability**: \( v^*_\alpha \in V, w^*_\alpha \in W. \)

• **Properties of \( f \):**
  • Strong Convexity: \( f \) is \( M_f \)-strongly-convex,
  • Boundedness: \(|f'(x)| \leq B_{f',\alpha}, |f(x)| \leq B_{f,\alpha}, \forall 0 \leq x \leq B^\alpha_w. \)
  • Non-negativity: \( f(x) \geq 0, \forall x \in \mathbb{R}. \)

• **Boundedness of the function classes:**
  • \( 0 \leq w(s,a) \leq B^\alpha_w, \forall s \in S, a \in A, w \in W, \)
  • \( \|v\|_{\infty} \leq B_{v,\alpha} := \frac{\alpha B_{f',\alpha} + 1}{1-\gamma}, \forall v \in V. \)

*Single-policy concentrability and only realizability!*
Statistical error

Statistical error term that arises in analysis:

\[ \epsilon_{\text{stat}} := (1-\gamma)B_v \cdot \left( \frac{2 \log \frac{4|V|}{\delta}}{n} \right)^{\frac{1}{2}} + (\alpha B_f + B_w B_e) \cdot \left( \frac{2 \log \frac{4|V||W|}{\delta}}{n} \right)^{\frac{1}{2}}. \]
Statistical error term that arises in analysis:

\[ \epsilon_{\text{stat}} := (1 - \gamma) B_v \cdot \left( \frac{2 \log \frac{4|V|}{\delta}}{n} \right)^{\frac{1}{2}} + (\alpha B_f + B_w B_e) \cdot \left( \frac{2 \log \frac{4|W|}{\delta}}{n} \right)^{\frac{1}{2}}. \]

\( \epsilon_{\text{stat}} \) characterizes the statistical error \( \hat{L}_\alpha(v, w) - L_\alpha(v, w) \) based on elementary concentration (unbiased)!
Theorem (Sample complexity of learning $\pi^*_\alpha$)

Fix $\alpha > 0$. Suppose assumptions hold for the said $\alpha$. Then with at least probability $1 - \delta$, the output of PRO-RL satisfies:

$$J(\pi^*_\alpha) - J(\hat{\pi}) \leq \frac{4}{1 - \gamma} \sqrt{\frac{\epsilon_{\text{stat}}}{\alpha M_f}}.$$
Theorem (Sample complexity of learning $\pi^*_\alpha$)

Fix $\alpha > 0$. Suppose assumptions hold for the said $\alpha$. Then with at least probability $1 - \delta$, the output of $\text{PRO-RL}$ satisfies:

$$J(\pi^*_\alpha) - J(\hat{\pi}) \leq \frac{4}{1 - \gamma} \sqrt{\frac{\epsilon_{\text{stat}}}{\alpha M_f}}.$$

$$f(x) = \frac{M_f}{2} x^2 \rightarrow n = \tilde{O}\left(\frac{(B_{w,\alpha})^2}{(1-\gamma)^6(\alpha M_f)^2 \epsilon^4} + \frac{(B_{w,\alpha})^4}{(1-\gamma)^6 \epsilon^4}\right).$$
Sample complexity of competing with $\pi_0^*$

Corollary (Sample complexity of competing with $\pi_0^*$)

Suppose there exists $d_0^* \in D_0^*$ with concentrability (not unique). Assume the realizability holds for $\alpha = \alpha_\epsilon := \frac{\epsilon}{2B_{f,0}}$. For

$$n \gtrsim \frac{(\epsilon B_{f,\alpha_\epsilon} + 2B_{w,\alpha_\epsilon} B_{e,\alpha_\epsilon} B_{f,0})^2}{\epsilon^6 M_f^2 (1 - \gamma)^4} \log \frac{4|\mathcal{V}||\mathcal{W}|}{\delta},$$

the output of PRO-RL with input $\alpha = \alpha_\epsilon$ satisfies

$$J(\pi_0^*) - J(\hat{\pi}) \leq \epsilon,$$

with probability greater than $1 - \delta$. 
Corollary (Sample complexity of competing with $\pi_0^*$)

Suppose there exists $d_0^* \in D_0^*$ with concentrability (not unique). Assume the realizability holds for $\alpha = \alpha_\epsilon := \frac{\epsilon}{2B_{f,0}}$. For

$$n \gtrsim \frac{(\epsilon B_{f,\alpha_\epsilon} + 2B_{w,\alpha_\epsilon} B_{e,\alpha_\epsilon} B_{f,0})^2}{\epsilon^6 M_f^2 (1 - \gamma)^4} \log \frac{4|\mathcal{V}| |\mathcal{W}|}{\delta},$$

the output of PRO-RL with input $\alpha = \alpha_\epsilon$ satisfies

$$J(\pi_0^*) - J(\hat{\pi}) \leq \epsilon,$$

with probability greater than $1 - \delta$.

Efficient learning with single-policy concentrability and realizability!
Comparison with existing algorithms

| Algorithm          | Data                                                                 | Function Class                                                                 |
|--------------------|----------------------------------------------------------------------|--------------------------------------------------------------------------------|
| AVI                | $\|\frac{d\pi}{dD}\|_\infty \leq B_w, \forall \pi$                 | $\mathcal{T} f \in \mathcal{F}, \forall f \in \mathcal{F}$ (Munos and Szepesvári, 2008) |
| API                |                                                                      | $\mathcal{T}^\pi f \in \mathcal{F}, \forall f \in \mathcal{F}, \pi \in \Pi$ (Antos et al., 2008b) |
| BVFT              | **Stronger than above**                                             | $Q^* \in \mathcal{F}$ (Xie and Jiang, 2021b)                                  |
| Pessimism          | $\|\frac{d\pi_0}{dD}\|_\infty \leq B_w$                           | $\mathcal{T}^\pi f \in \mathcal{F}, \forall f \in \mathcal{F}, \pi \in \Pi$ (Xie et al., 2021) |
| PRO–RL (against $\pi_{\alpha}^*$) | $\|\frac{d\pi^*_\alpha}{dD}\|_\infty \leq B_w$                     | $w_0^* \in \mathcal{W}, Q^\pi \in \mathcal{F}, \forall \pi \in \Pi$ (Jiang and Huang, 2020) |
| PRO–RL             | $\|\frac{d\pi_0}{dD}\|_\infty \leq B_w$                           | $w_{\alpha}^* \in \mathcal{W}, v_{\alpha}^* \in \mathcal{V}$ (Theorem 1) |
| PRO–RL with $\alpha = 0$ | $\|\frac{d\pi^*_0}{dD}\|_\infty \leq B_w, \frac{d\pi^*_0(s)}{dD(s)} \geq B_{w,l}, \forall s$ | $w_{\alpha_{\epsilon}}^*, B_w \in \mathcal{W}, v_{\alpha_{\epsilon}}^*, B_w \in \mathcal{V}$ (Corollary 3) |
|                    | $\frac{d\pi^*(s)}{dD(s)} \leq B_{w,u}, \forall \pi, s$            | $w_0^* \in \mathcal{W}, v_0^* \in \mathcal{V}$ (Corollary 6) |
Comparison with existing algorithms

| Algorithm          | Data                                                                 | Function Class                                                                 |
|--------------------|----------------------------------------------------------------------|--------------------------------------------------------------------------------|
| AVI                | $\| \frac{d^{\pi}}{dD} \|_{\infty} \leq B_w, \forall \pi $          | $\mathcal{T}f \in \mathcal{F}, \forall f \in \mathcal{F}$ (Munos and Szepesvári, 2008) |
| API                |                                                                      | $\mathcal{T}^{\pi}f \in \mathcal{F}, \forall f \in \mathcal{F}, \pi \in \Pi$ (Antos et al., 2008b) |
| BVFT               | Stronger than above                                                  | $Q^* \in \mathcal{F}$ (Xie and Jiang, 2021b)                                     |
| Pessimism          | $\| \frac{d^{\pi}_{\alpha}}{dD} \|_{\infty} \leq B_w $             | $\mathcal{T}^{\pi}f \in \mathcal{F}, \forall f \in \mathcal{F}, \pi \in \Pi$ (Xie et al., 2021) |
| PRO–RL (against $\pi_{\alpha}^*$) |                                                                      | $w_0^* \in \mathcal{W}, Q^{\pi} \in \mathcal{F}, \forall \pi \in \Pi$ (Jiang and Huang, 2020) |
| PRO–RL             | $\| \frac{d^{\pi}_{\alpha}}{dD} \|_{\infty} \leq B_w $             | $w^*_\alpha \in \mathcal{W}, v^*_\alpha \in \mathcal{V}$ (Theorem 1)           |
| PRO–RL with $\alpha = 0$ | $\| \frac{d^{\pi}_{\alpha}}{dD} \|_{\infty} \leq B_w, \frac{d^{\pi}_{\alpha}(s)}{dD(s)} \geq B_{w,l}, \forall s$ | $w^*_{\alpha^*_s}, B_w \in \mathcal{W}, v^*_{\alpha^*_s}, B_w \in \mathcal{V}$ (Corollary 3) |

The first algorithm to achieve efficient learning with **single-policy concentrability** and **only realizability**!
Proof sketch for Theorem

Intuition: **invariance of saddle points**

**Lemma**

Suppose \((x^*, y^*)\) is a saddle point of \(f(x, y)\) over \(\mathcal{X} \times \mathcal{Y}\), then for any \(\mathcal{X}' \subseteq \mathcal{X}\) and \(\mathcal{Y}' \subseteq \mathcal{Y}\), if \((x^*, y^*) \in \mathcal{X}' \times \mathcal{Y}'\), we have:

\[
(x^*, y^*) \in \arg \min_{x \in \mathcal{X}'} \arg \max_{y \in \mathcal{Y}'} f(x, y),
\]

\[
(x^*, y^*) \in \arg \max_{y \in \mathcal{Y}'} \arg \min_{x \in \mathcal{X}'} f(x, y).
\]
Proof sketch for Theorem

Intuition: invariance of saddle points

Lemma

Suppose \((x^*, y^*)\) is a saddle point of \(f(x, y)\) over \(\mathcal{X} \times \mathcal{Y}\), then for any \(\mathcal{X}' \subseteq \mathcal{X}\) and \(\mathcal{Y}' \subseteq \mathcal{Y}\), if \((x^*, y^*) \in \mathcal{X}' \times \mathcal{Y}'\), we have:

\[
(x^*, y^*) \in \arg\min_{x \in \mathcal{X}'} \arg\max_{y \in \mathcal{Y}'} f(x, y),
\]

\[
(x^*, y^*) \in \arg\max_{y \in \mathcal{Y}'} \arg\min_{x \in \mathcal{X}'} f(x, y).
\]

Optimizing over \(V \times W\) instead of \(\mathbb{R}^{|S|} \times \mathbb{R}^{|S||A|}_+\) can still find \((v^*_\alpha, w^*_\alpha)\).
Concentration of $\hat{L}_\alpha(v, w)$

Step 1: bound $|\hat{L}_\alpha(v, w) - L_\alpha(v, w)|$ via Hoeffding’s inequality and union bound.

**Lemma**

With at least probability $1 - \delta$, for all $v \in \mathcal{V}$ and $w \in \mathcal{W}$ we have:

$$|\hat{L}_\alpha(v, w) - L_\alpha(v, w)| \leq \epsilon_{stat}.$$
Near-optimal \( \hat{w} \)

Step 2: bound \( \| \hat{w} - w^*_\alpha \|_{2,d^D} \) via strong concavity.

**Lemma**

*With at least probability* \( 1 - \delta \),

\[
L_\alpha(v^*_\alpha, w^*_\alpha) - L_\alpha(v^*_\alpha, \hat{w}) \leq 2\varepsilon_{stat}.
\]
Near-optimal $\hat{w}$

Step 2: bound $\|\hat{w} - w_\alpha^*\|_{2,d^D}$ via strong concavity.

**Lemma**

*With at least probability $1 - \delta$,*

$$L_\alpha(v_\alpha^*, w_\alpha^*) - L_\alpha(v_\alpha^*, \hat{w}) \leq 2\epsilon_{stat}.$$ 

**Lemma**

*With at least probability $1 - \delta$,*

$$\|\hat{w} - w_\alpha^*\|_{2,d^D} \leq \sqrt{\frac{4\epsilon_{stat}}{\alpha M_f}}.$$
Near-optimal $\hat{\pi}$

Step 3: bound $\mathbb{E}_{s \sim d^*_\alpha} [\| \pi^*_\alpha(s, \cdot) - \hat{\pi}(s, \cdot) \|_1]$ and $J(\pi^*_\alpha) - J(\hat{\pi})$ via performance difference lemma.

**Lemma**

$$\mathbb{E}_{s \sim d^*_\alpha} [\| \pi^*_\alpha(s, \cdot) - \hat{\pi}(s, \cdot) \|_1] \leq 2 \| \hat{\omega} - w^*_\alpha \|_{2,d^D}.$$
Near-optimal $\hat{\pi}$

Step 3: bound $\mathbb{E}_{s \sim d^*_\alpha} [||\pi^*_\alpha(s, \cdot) - \hat{\pi}(s, \cdot)||_1]$ and $J(\pi^*_\alpha) - J(\hat{\pi})$ via performance difference lemma.

**Lemma**

$\mathbb{E}_{s \sim d^*_\alpha} [||\pi^*_\alpha(s, \cdot) - \hat{\pi}(s, \cdot)||_1] \leq 2||\hat{w} - w^*_\alpha||_{2,d^D}.$

**Lemma**

$J(\pi^*_\alpha) - J(\hat{\pi}) \leq \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d^*_\alpha} [||\pi^*_\alpha(s, \cdot) - \hat{\pi}(s, \cdot)||_1].$
Other results (see paper)

• **Agnostic Learning I**: competes with the best in the function class.

• **Agnostic Learning II**: competes with the best policy that the dataset covers.

• Unknown behavior policy $\pi_D$: **behavior cloning**.

• Improved sample complexity: set $\alpha = 0$, requires stronger concentration assumptions or asymptotics.
Primal-dual formulation is the analog of ERM for offline RL.
Primal-dual formulation is the analog of ERM for offline RL.

Remaining Questions:

- Optimal sample complexity in $\epsilon$.
- Realizability wrt unregularized value function/density ratio in non-asymptotic setting.
- Markov games.