Information-Theoretic Considerations in Batch RL

Jinglin Chen, Nan Jiang
University of Illinois at Urbana Champaign
What we study: theory of batch RL (ADP)—backbone for “deep RL”
What we study: theory of batch RL (ADP)—backbone for “deep RL”

Setting: learn good policy from batch data \{ (s, a, r, s') \} + value-function approximator $F$ (model $Q^*$)
What we study: theory of batch RL (ADP)—backbone for “deep RL”
Setting: learn good policy from batch data \{ (s, a, r, s') \} + value-function approximator $F$ (model $Q^*$)

Central question: When is sample-efficient ($poly(\log|F|, H)$) learning guaranteed?
What we study: theory of **batch RL (ADP)**—backbone for “deep RL”

Setting: learn **good policy** from batch data \{(s, a, r, s')\} + value-function approximator \(F\) (model \(Q^*\))

Central question: When is **sample-efficient** \(poly(\log|F|, H)\) learning guaranteed?

---

**Assumption on data**

- Data distribution
- Distribution induced by any policy \(\pi\)

---

**Assumption on \(F\)**
What we study: theory of batch RL (ADP)—backbone for “deep RL”
Setting: learn good policy from batch data \( \{(s, a, r, s')\} \) + value-function approximator \( F \) (model \( Q^* \))

Central question: When is sample-efficient \((poly(\log|F|, H))\) learning guaranteed?

**Assumption on data**
- Data distribution
- Distribution induced by any policy \( \pi \)

**Assumption on \( F \)**
- \( f \)
- \( \Pi_F \)
- \( \mathcal{F} \)
- \( \mathcal{T}_f \)
- small

[Remond’s03]

[Munos & Szepesvari ’05]
What we study: theory of batch RL (ADP)—backbone for “deep RL”

Setting: learn good policy from batch data \{(s, a, r, s')\} + value-function approximator \(F\) (model \(Q^*\))

Central question: When is sample-efficient \(\text{poly}(\log|F|, H)\) learning guaranteed?

**Assumption on data**
- Distribution shift
  - Divergence (ratio) between two distributions \(d \equiv \frac{\mu(s,a)}{\mu'(s,a)}\) is upper bounded by a constant \(C\) ("concentratability")

**Assumption on \(F\)**
- Are they necessary? (hardness results)
- Do they hold in interesting scenarios?

---

**Setting:**
- Learn a good policy from batch data \{\((s, a, r, s')\)\} with value-function approximator \(F\) (model \(Q^*\)).

**Central Question:**
- When is sample-efficient \(\text{poly}(\log|F|, H)\) learning guaranteed?

**Assumption on Data**
- Distribution shift
  - Divergence (ratio) between two distributions \(d \equiv \frac{\mu(s,a)}{\mu'(s,a)}\) is upper bounded by a constant \(C\) ("concentratability")

**Assumption on \(F\)**
- Necessary? (hardness results)
- Hold in interesting scenarios?

---

**Figure 2:** Distribution Shift

**Jinglin Chen**
- Artificial Intelligence Ph.D. Qualifying Exam
What we study: theory of batch RL (ADP)—backbone for “deep RL"

Setting: learn good policy from batch data \{(s, a, r, s')\} + value-function approximator \(F\) (model \(Q^*\))

Central question: When is sample-efficient \((\text{poly}(\log|F|, H))\) learning guaranteed?

Assumption on data

- Intuition: data should be exploratory

Assumption on \(F\)

Are they necessary? (hardness results)

Do they hold in interesting scenarios?

\[\text{Munos}'03\]

\[\text{Munos & Szepesvari '05}\]
What we study: theory of batch RL (ADP)—backbone for “deep RL”
Setting: learn good policy from batch data \{(s, a, r, s')\} + value-function approximator \(F\) (model \(Q^*\))

Central question: When is sample-efficient \((poly(\log|F|, H))\) learning guaranteed?

<table>
<thead>
<tr>
<th>Assumption on data</th>
<th>Are they necessary? (hardness results)</th>
<th>Do they hold in interesting scenarios?</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Intuition: data should be exploratory</td>
<td>• We show: also about MDP dynamics!</td>
<td></td>
</tr>
<tr>
<td>• Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Assumption on data:
- Distribution shift between data distribution and distribution induced by any policy \(\pi\).

Assumption on \(F\):
- \(F\) is a set of functions that are small in some sense.

\[\mathcal{F} = \{f(\cdot; \theta) : \theta \in \Theta\}\]

\([\text{Munos'}03]\) \([\text{Munos & Szepesvari '05}]\)
What we study: theory of batch RL (ADP)—backbone for “deep RL”
Setting: learn good policy from batch data \{ (s, a, r, s’) \} + value-function approximator $F$ (model $Q^*$)

Central question: When is sample-efficient ($\text{poly}(\log |F|, H)$) learning guaranteed?

**Assumption on data**

- Intuition: data should be exploratory
- We show: also about MDP dynamics!
- Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution

**Assumption on $F$**

- $F = \{ f(\cdot; \theta) : \theta \in \mathbb{R}^d \}$

**Do they hold in interesting scenarios?**

Similar to Jiang et al [2017]

[Munos'03]

[Munos & Szepesvari '05]

[Self Introduction]

Research Topic: Information-Theoretic Considerations in Batch RL

Selected Paper: Liu, Qiang, et al. Breaking the Curse of Horizon

Overview

Problem Setting

Assumptions

Results

Mild Distribution Shift

Divergence (ratio) between two distributions $d \propto (s, a) / \mu(s, a)$ is upper bounded by a constant $C$ (“concentratability”)

Data distribution: $\mathbb{P}_{(s, a)}$ Distribution induced by arbitrary policy in the class: $*_{s, a}$

Figure 2: Distribution Shift

Jinglin Chen
Artificial Intelligence Ph.D. Qualifying Exam

What we study: theory of batch RL (ADP)—backbone for “deep RL”
Setting: learn good policy from batch data \{ (s, a, r, s’) \} + value-function approximator $F$ (model $Q^*$)

Central question: When is sample-efficient ($\text{poly}(\log |F|, H)$) learning guaranteed?

**Assumption on data**

- Intuition: data should be exploratory
- We show: also about MDP dynamics!
- Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution

**Assumption on $F$**

- $F = \{ f(\cdot; \theta) : \theta \in \mathbb{R}^d \}$

**Do they hold in interesting scenarios?**

Similar to Jiang et al [2017]

[Munos'03]

[Munos & Szepesvari '05]
What we study: theory of batch RL (ADP)—backbone for “deep RL”

Setting: learn good policy from batch data \{(s, a, r, s')\} + value-function approximator \(F\) (model \(Q^*\))

Central question: When is sample-efficient \((\text{poly}(\log|F|, H))\) learning guaranteed?

**Assumption on data**
- Intuition: data should be exploratory
- We show: also about MDP dynamics!
- Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution

**Assumption on \(F\)**
- Conjecture: realizability alone is insufficient

**Are they necessary? (hardness results)**

**Do they hold in interesting scenarios?**

Similar to Jiang et al [2017]

---

Figure 2: Distribution Shift

Jinglin Chen
Artificial Intelligence Ph.D. Qualifying Exam

---

What we study: theory of batch RL (ADP)—backbone for “deep RL”

Setting: learn good policy from batch data \{(s, a, r, s')\} + value-function approximator \(F\) (model \(Q^*\))

Central question: When is sample-efficient \((\text{poly}(\log|F|, H))\) learning guaranteed?

**Assumption on data**
- Intuition: data should be exploratory
- We show: also about MDP dynamics!
- Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution

**Assumption on \(F\)**
- Conjecture: realizability alone is insufficient

**Are they necessary? (hardness results)**

**Do they hold in interesting scenarios?**

Similar to Jiang et al [2017]

---

Figure 2: Distribution Shift

Jinglin Chen
Artificial Intelligence Ph.D. Qualifying Exam
What we study: theory of batch RL (ADP)—backbone for “deep RL”

Setting: learn good policy from batch data \{(s, a, r, s')\} + value-function approximator \(F\) (model \(Q^*\))

Central question: When is sample-efficient \(\text{poly}(\log|\mathcal{F}|, H)\) learning guaranteed?

---

Assumption on data

- Intuition: data should be exploratory
- We show: also about MDP dynamics!
- Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution

Are they necessary? (hardness results)

- Conjecture: realizability alone is insufficient
- Alg-specific lower bound exists for decades
- Info-theoretic?

Assumption on \(F\)

- Similar to Jiang et al [2017]

Do they hold in interesting scenarios?

---

[Fig. 2: Distribution Shift]

Jinglin Chen
Artificial Intelligence Ph.D. Qualifying Exam
What we study: theory of batch RL (ADP)—backbone for “deep RL”
Setting: learn good policy from batch data \{(s, a, r, s')\} + value-function approximator $F$ (model $Q^*$)

Central question: When is sample-efficient ($poly(\log |F|, H)$) learning guaranteed?

**Are they necessary? (hardness results)**
- Intuition: data should be exploratory
- We show: also about MDP dynamics!
- Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution

**Assumption on data**
- Data distribution:
  - Distribution induced by any policy $\pi$

**Assumption on $F$**
- Conjecture: realizability alone is insufficient
- Alg-specific lower bound exists for decades
- Info-theoretic?
  - Negative results: two general proof styles excluded
  - e.g., construct an exponentially large MDP family $\Rightarrow$ fail!

Similar to Jiang et al [2017]
What we study: theory of batch RL (ADP)—backbone for “deep RL”
Setting: learn good policy from batch data \{(s, a, r, s')\} + value-function approximator \(F\) (model \(Q^*\))

\[
\text{Central question: When is sample-efficient } (\text{poly}(\log|F|, H)) \text{ learning guaranteed?}
\]

**Assumption on data**
- Distribution: \(\mathcal{F}\) piece-wise constant + \(\mathcal{F}\) closed under Bellman update =\(\Leftrightarrow\) bisimulation [Givan et al'03]

**Assumption on \(F\)**
- \(\mathcal{F}\) closed under bisimulation [Givan et al'03]
- \(\Pi_F \mathcal{F}\) small
- \(\mathcal{F}\) piece-wise constant

**Are they necessary? (hardness results)**
- Intuition: data should be exploratory
- We show: also about MDP dynamics!
- Unrestricted dynamics cause exponential lower bound even with the most exploratory distribution

**Do they hold in interesting scenarios?**
- Conjecture: realizability alone is insufficient
- Alg-specific lower bound exists for decades
- Info-theoretic?
  - Negative results: two general proof styles excluded
  - e.g., construct an exponentially large MDP family \(\Rightarrow\) fail!

**Similar to Jiang et al [2017]**
Implications and the Bigger Picture

Tabular RL

Online (exploration)

Batch

RL with function approximation tractable

Nice dynamics & exploratory data + realizability + ??

Nice dynamics & exploratory data + realizability

Gap confirmed

Gap?

Nice dynamics (low Bellman rank; Jiang et al’17) + realizability

Nice dynamics (low witness rank; Sun et al’18) + realizability

value-based

model-based

Poster: Tue Evening
Pacific Ballroom #209

Poster: Tue Evening
Pacific Ballroom #209