Information-Theoretic Considerations in Batch Reinforcement Learning
Jinglin Chen (UIUC), Nan Jiang (UIUC)

Introduction
- Batch value-func approx (=ADP): backbone of many deep RL alg
  - e.g., FQI → DQN
- Prior works prove that they work under certain assumptions [1]
- Are they necessary? Do they hold in interesting scenarios?
  - We seek info-theoretic (alg-independent) hardness to justify necessity

Setting and Algorithms
Setting: learn near-optimal policy from data \([\{(s, a, r, s')\}] + \) function class \(F\)
- \((s, a)\) is drawn i.i.d. from “data distribution”

Fitted Q-Iteration [2]: Initialize \(f \in F\)
- \(f = \text{solution to regression problem } \{(s, a) \rightarrow r + \gamma \max f, (s', a')\} \) over \(F\)

Modified Bellman Residual Minimization [1]
- \(\text{argmin, } \sup_{f \in F} \| \Pi \Delta f \|_F \rightarrow 0\)
  - where \(\Delta f = \Pi \Delta (s, a, r, s')\)

Implication
- \(C\) measures how exploratory the data is
- More than \(\phi\): if MDP dynamics are unregulated, no distribution works!
- What kind of problems have “smooth dynamics”?

Example of “smooth dynamics”
- High-dimensional observations generated from finite & small hidden state space
- Same as environments that allow sample-efficient exploration [3]
- Can construct small \(C\) by taking mixture of distributions of several policies

Assumptions and Upper bounds

Data Assumptions
- Data distribution well covers states (and actions) visited by any policy \(\pi\)
- Measured by \(C\): worst-case (over state & policy) density ratio

“Concentratability Coefficient”

Representation Assumptions
- Realizability: \(Q^* \in F\)
- Need more! \(\sup_{f \in F} \| \Pi \Delta f \|_F \rightarrow 0\)
  - 
- “Inherent Bellman error”

Upper bounds
- Under above assumptions, \(\text{poly}(\log F, C, H)\) sample complexity [1]
- We provide simplified analyses under minimal setup
  - Error rate for modified BRM [1] improved \(n^{1/2} \rightarrow n^{1/4}\)

On Concentratability

Exponential lower bound when \(C\) is unbounded
- Known & dtrm dynamics, unknown reward
- \(F\) realizes \(Q^*\) for every possible MDP
- Similarly \(G = \phi\) for inherent Bellman error
- No efficient exploration algorithm exists
- Any data distribution + any batch alg
  - special case of exploration algorithm

On Inherent Bellman Error

Conjecture
- There exists a family of MDPs \(\mathcal{M}\) such that: any algorithm with realizable \(F\) as input cannot have \(\text{poly}(\log F, C, H)\) complexity.

Why should be true:
- No poly alg known under general func approx with realizability alone
- Divergence of ADP known for decades

Obvious? Info-theoretic lower bound?
- Construct an exponential-sized model family \(\Rightarrow \) fail!

Reason: Batch model-based RL only needs realizability
- Create “small” \((F, G)\) from \(\mathcal{M}\): realizable & no inherent Bellman error

Lesson: Need to rule out \textit{model-based} algorithms.

“Value-profile” idea doesn’t work in tabular constructions
- Hide info of \(s\) and only reveal \(f(s, a)\) \(f \in F, a \in A\) [4, 5]
- Issue with construction in [5]: not realizable
- When realizable: efficient learning exists using \(Q^*\)-irrelevant abstraction

Why care?
- If true, construction is seriously stochastic and “non-bandit”
- All known RL lower bound are nearly deterministic and bandit-structured — no reflection of the long-horizon challenge of RL

Connection to State Abstractions
- \(\phi\) is bisimulation \(\Leftrightarrow \Pi\) (piece-wise constant) has 0 inherent Bellman error
- \(\phi\) is trivial
- \(\equiv\)
  - Use \(f = 0\) to witness reward errors.
  - Use \(f = \arg \max_{f} 1/2<f(s, a) + \Pi(s', a' f)>\) for any aggregated \(s'\) to witness transition errors.

References