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 $f_0 \in F$
f_t = solution to regression problem $\{(s, a) \to r + \gamma \max_{a'} f_{t-1}(s', a')\}$ over F

Modified Bellman Residual Minimization [1]

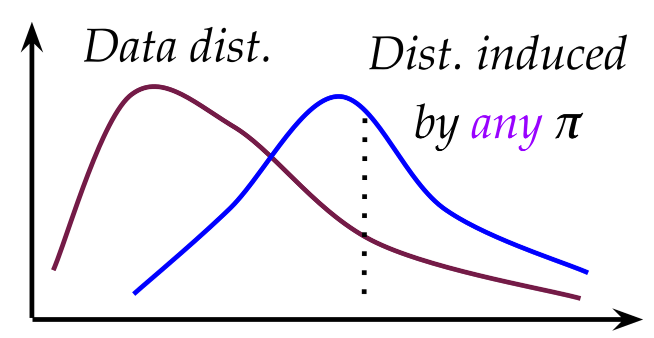
$\mathrm{argmin}_f \sup_g L_D(f; f) - L_D(g; f)$,
where $L_D(f; f') := \sum_{(s,a,r,s')} [f(s, a) - r - \gamma \max_{a'} f'(s', a')]^2$

Notations *Bellman update* : $(\mathcal{T}f)(s, a) = R(s, a) + \gamma\, \mathbf{E}_{s'|s,a}[\max_{a'} f(s', a')]$
Effective horizon : $H = 1/(1-\gamma)$

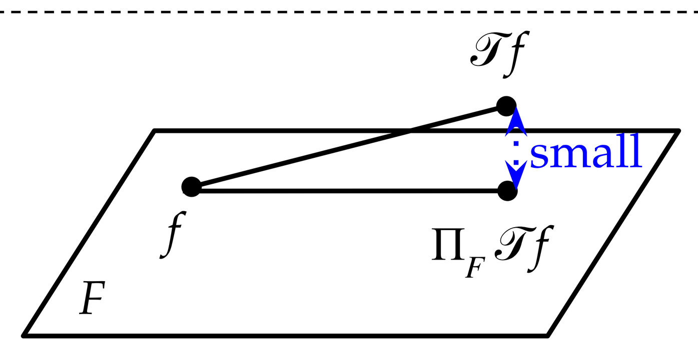
Assumptions and Upper bounds

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

"*Concentratability Coefficient*"



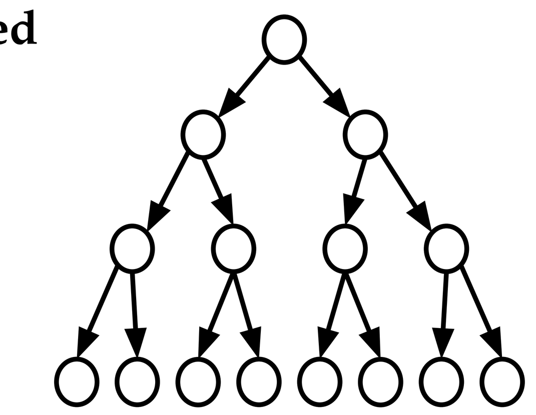
Representation Assumptions
- Realizability: $Q^\star \in F$
- Need more! $\sup_f \| \Pi_F \mathcal{T}f - \mathcal{T}f \| \approx 0$
  (or: $G \Rightarrow \sup_f \inf_g \| g - \mathcal{T}f \| \approx 0$)
  "*Inherent Bellman error*"

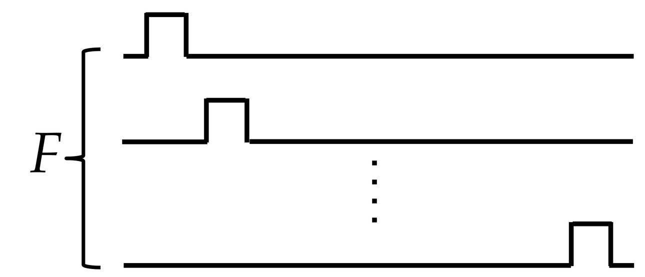


Upper bounds
- Under above assumptions, $poly(\log|F|, C, H)$ sample complexity [1]
- We provide simplified analyses under minimal setup
  - Error rate for modified BRM [1] improved $n^{-1/4} \to n^{-1/2}$

On Concentratability

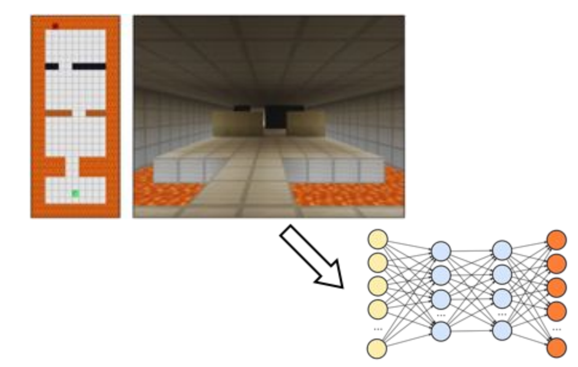
Exponential lower bound when C is unbounded
- Known & dtmn dynamics, unknown reward
- F realizes $Q^\star$ for every possible MDP
- Similarly G => no inherent Bellman error
- No efficient exploration algorithm exists
- Any data distribution + any batch alg = special case of exploration algorithm





Implication
- C measures how exploratory the data is
- More than that! 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



High-level implications

	Batch	Online (exploration)
Value-based	Smooth dynamics + exploratory data + Realizability + ???	Smooth dynamics (low Bellman rank [2]) + Realizability
Model-based	Smooth dynamics + exploratory data + Realizability	Smooth dynamics (low witness rank [3]) + Realizability

Gap? (between Value-based Batch and Value-based Online)
Gap? (between Value-based Batch and Model-based Batch)
Gap exists (between Value-based Online and Model-based Online)

On Inherent Bellman Error

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

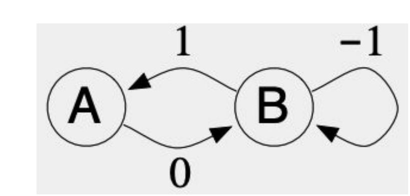
Why should be true:
- No poly alg known under general func approx with realizability alone
- Divergence of ADP known for decades

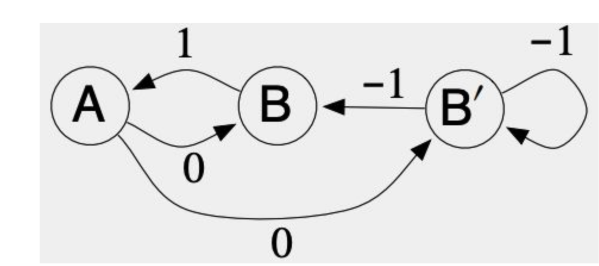
Obvious? Info-theoretic lower bound?

Construct an exponential-sized model family => fail!
Reason: Batch model-based RL only needs realizability
- Create "small" (F, G) from $\mathcal{M}$: realizable & no inherent Bellman error

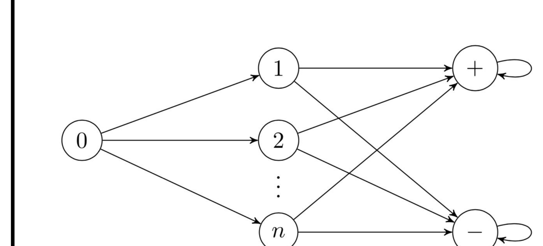
Lesson: Need to rule out model-based algorithms.

"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

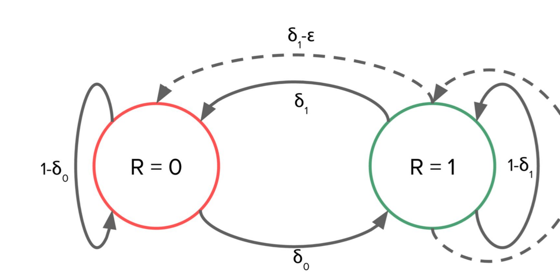




Why care?
- If true, construction is seriously stochastic and "non-bandit"
- All known RL lower bound are nearly deterministic and bandit-structured --- no reflection of the long-horizon challenge of RL



Left: Dann & Brunskill'15
Right: Osband & Van Roy'16



- May shed light on related questions
  - "True" horizon dependence in RL (JA18, COLT open problem)
  - Exploration with linear function approximation

Connection to State Abstractions

ϕ is bisimulation ⇔ F^ϕ (piece-wise constant) has 0 inherent Bellman error
- ⇒ is trivial
- ⇐:
  - Use $f = 0$ to witness reward errors.
  - Use f as the argmax of $<P(s^1, a) - P(s^2, a),\ f>$ for any aggregated s^1 and s^2 to witness transition errors.

References

[1] Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. Machine Learning, 71(1):89–129, 2008.
[2] Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. Journal of Machine Learning Research, 6:503–556, 2005.
[3] Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual Decision Processes with low Bellman rank are PAC-learnable. In International Conference on Machine Learning, 2017.
[4] Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based RL in Contextual Decision Processes: PAC bounds and Exponential Improvements over Model-free Approaches. In Conference on Learning Theory, 2019.
[5] Sutton, R. S. and Barto, A. G. Reinforcement learning: An introduction. MIT press, 2018.