Batch Value-Function Approximation with Only Realizability

Nan Jiang
University of Illinois at Urbana-Champaign

Value-function approximation

- Use a restricted class of functions to approximate the optimal value function Q^*
- Batch mode: passively given data & no access to environment
  - Important for real-life RL: medical, customer relationship management, experience personalization, etc.
- When can we guarantee sample-efficient learning?

A "Batch RL 101" Result?

- Supervised learning
  - Data: $(x, y) \sim P_{X,Y}$
  - A class of predictors F (assume finite), one of which is good
  - Can find a good predictor w/ $O(\log|F|)$ samples (info-theoretic)
- Reinforcement learning (batch-mode, VFA)
  - Data: (s, a, r, s') from MDP (to be defined)
    - Needs to be exploratory (to be formalized)
  - F (assume finite) s.t. $Q^* \in F$ (*realizability*)

    seems too weak
  - Can we find a near-optimal policy using $O(\log|F|)$ samples?

- Long-standing open problem
- Believed to be info-theoretically hard
- This talk: Break the barrier!

Markov Decision Process (MDP)

- For $t = 0, 1, 2, \ldots,$ the agent
  - observes state $s_t \in S$ (very large)
  - chooses action $a_t \in A$ (finite & small)
  - receives reward $r_t = R(s_t, a_t)$

transition dynamics
$P: S \times A \to \Delta(S)$

reward function
$R: S \times A \to [0,1]$

- Policy $\pi: S \to A$
- Expected return $J(\pi) := \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 \sim d_0; \pi]$
- Key solution concepts
  - Bellman eq: $Q^\star = \mathcal{T} Q^\star$, where for any f,
    $$(\mathcal{T} f)(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[\max_{a'} f(s',a')]$$
  - Optimal policy π^* is greedy w.r.t. Q^*
  - Occupancy: $d^\pi(s,a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}[s_t = s, a_t = a \mid \pi]$

Batch learning in large MDPs

standard-ish def:
$C = \max_\pi \|d^\pi / \mu\|_\infty$

- Dataset $D = \{(s, a, r, s')\}$
  - $(s, a) \sim \mu$ ("data distribution"), $r = R(s, a)$, $s' \sim P(\cdot \mid s, a)$
  - Measure exploratoriness: *concentrability coefficient* C [Munos'03'07]
- Function class F (finite) s.t. $Q^* \in F$ (*realizability*)
  - see approximate ver. in paper (not considered in talk)
- Goal: find $f \approx Q^*$ s.t. its greedy policy is ε-optimal

Back to the earlier question:

Can we achieve sample complexity
poly($\log|F|$, $1/(1-\gamma)$, $1/\varepsilon$, $1/\delta$, C)?

Prior work—no, unless w/ stronger func-approx assumptions

- e.g., $\forall f \in \mathcal{F}, \mathcal{T} f \in \mathcal{F}$, *no "inherent Bellman error"* [Antos'08]

Why realizability seems insufficient?

Intuition 1: Fitted Q-Iteration (FQI)

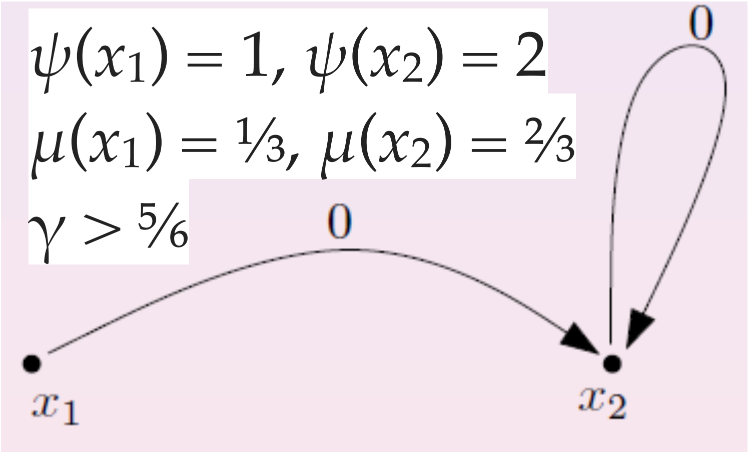
- Initialize $f_0 \in F$ arbitrarily
- In iteration k, convert D to least-square regression dataset

$$\{((s,a), r + \gamma \max_{a'} f_{k-1}(s',a'))\}$$

  and let f_k be the ERM

  bootstrapped target

- Can diverge even w/ realizable linear class & infinite data
  - Problem: the regression may NOT be realizable for $f_{k\text{-}1} \neq Q^*$
  - Resolved by $\forall f \in \mathcal{F}, \mathcal{T} f \in \mathcal{F}$ ($\mathcal{T} f_{k-1}$ is Bayes optimal)



[Tsitsiklis & Van Roy'96]

Why realizability seems insufficient?

Intuition 2: minimize $\|f - \mathcal{T}f\|$ (BRM)

- Naive: $\frac{1}{|D|}\sum_{(s,a,r,s')\in D}\left(f(s,a) - (r + \gamma \max_{a'} f(s',a'))\right)^2$
- Issue: expected = $\|f - \mathcal{T}f\|_{2,\mu}^2 + \boxed{\gamma^2 \mathbb{E}_{(s,a)\sim\mu} \mathrm{Var}_{s'\sim P(s,a)}[\max_{a'} f(s',a')]}$
- Sol 1, "double sampling" [Baird'95]: produce 2 iid s' from each (s, a)
- Sol 2, modified BRM [Antos et al'08]

$$\arg\min_{f\in\mathcal{F}} \max_{g\in\mathcal{G}} \sum_{(s,a,r,s')} \left(f(s,a) - \left(r + \gamma \max_{a'\in\mathcal{A}} f(s',a')\right)\right)^2 - \left(g(s,a) - \left(r + \gamma \max_{a'\in\mathcal{A}} f(s',a')\right)\right)^2$$

  - requires: $\mathcal{T}f \in \mathcal{G}\ \forall f \in \mathcal{F}$ ($|F|$ realizability assumptions)
  - special case of $G = F$ => no inherent Bellman error

Why realizability seems insufficient?

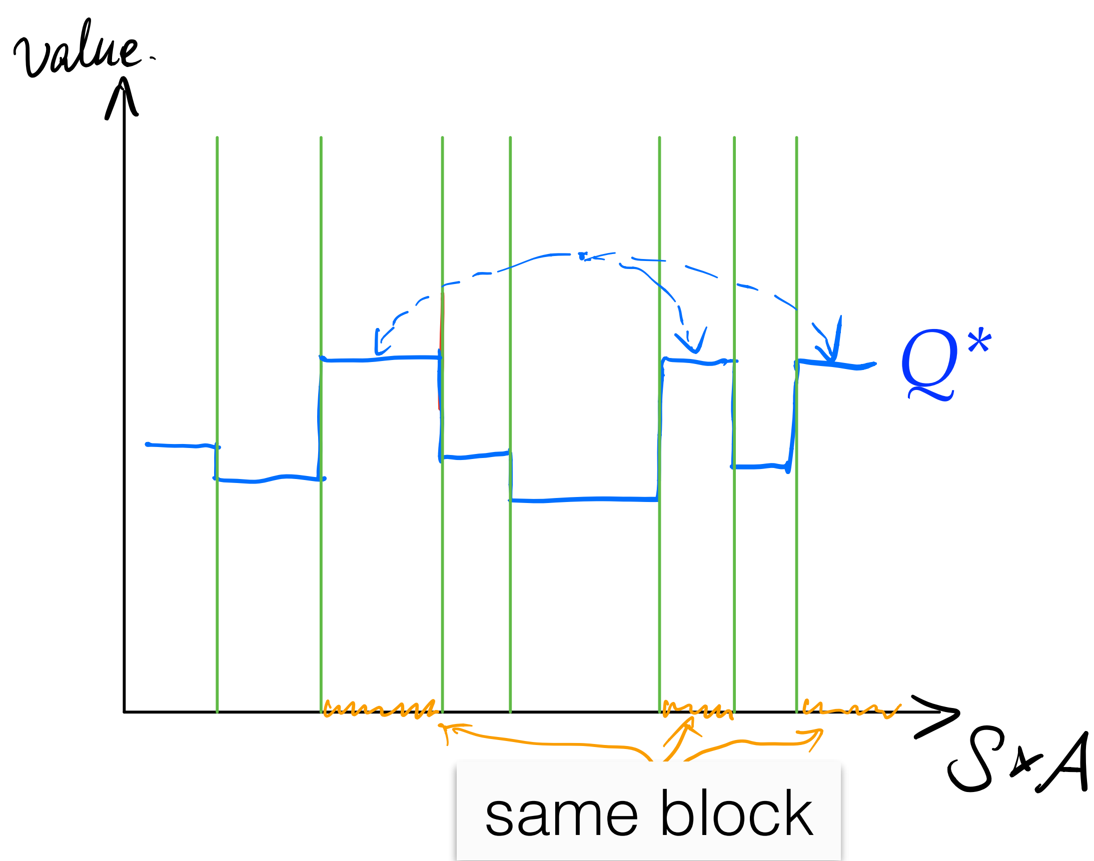
- All known algorithms fail under realizability, e.g.,
  - ADP diverges
  - BRM over-estimates
  - “ALP-style” methods need to model $d^\pi/\mu, \forall\pi$ [Xie & Jiang’20a]
  - Importance sampling has exponential variance
  - etc, etc
- Algorithmic ideas seems exhausted
  - … really?

Hint from State Abstractions

- Learning w/ "Q^*-irrelevant abstraction" is consistent [Gordon'95, Li et al'06]
- Essentially: piecewise constant function class + realizability
  - aggregate (s, a) pairs if Q^* values are the same
  - Solve the problem as if it were tabular (or FQI)
  - Sample complexity (vaguely) depends on #blocks
- More formal: If μ is supported on SxA (can relax), Q^* is the unique fixed point of $\mathcal{T}_\phi^\mu$ Bellman op + projection
  - $\mathcal{T}_\phi^\mu$ is always γ-contraction
  - Empirical ver $\widehat{\mathcal{T}}_\phi^\mu$: let $\mathcal{G}_\phi$ be the piecewise-constant class $\widehat{\mathcal{T}}_\phi^\mu f := \arg\min_{g \in \mathcal{G}_\phi} \frac{1}{|D|} \sum_{(s,a,r,s')} [(g(s,a) - r - \gamma \max_{a'} f(s',a'))^2]$

Hint from State Abstractions

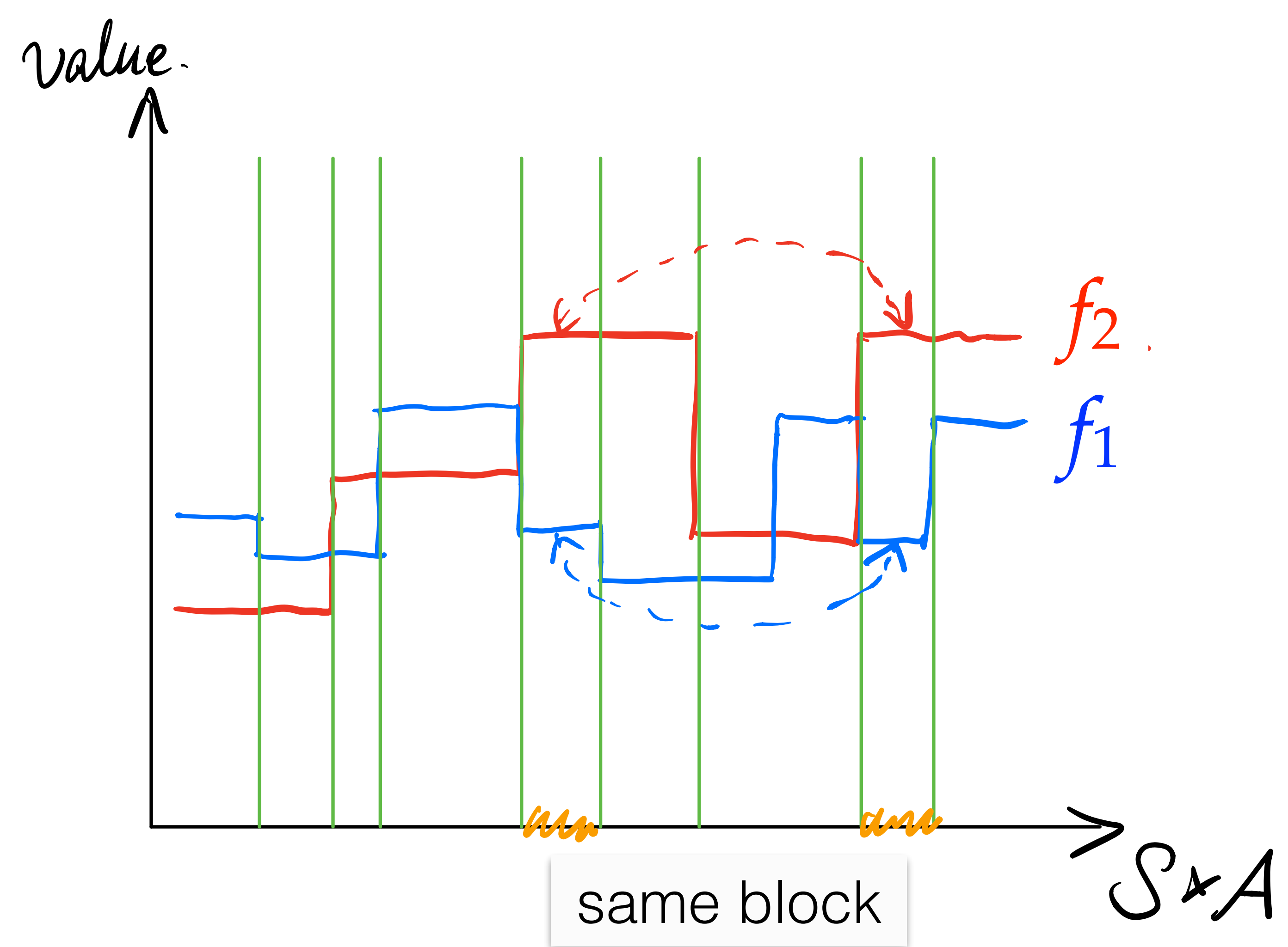
- Does a low-complexity ϕ always exist?
- YES! Just partition SxA according to Q^*
  - Size of ϕ: $O(1/\varepsilon)$ (ε is discretization error)



- Chicken-and-egg: only if I knew Q^*…

Pairwise Comparison

- Ultimately want to handle exponentially large F
- But problem is still nontrivial even when $|F|$=2!
  - One f_1, f_2 of is Q^*: how to find out from data?
- Partition SxA according to both functions in F simultaneously!
  - size of ϕ: $O(1/\varepsilon^2)$ — affordable!!!
- Fixed point of $\widehat{\mathcal{T}}_\phi^\mu$ will be close to Q^* => choose the one w/ lower $\|f - \widehat{\mathcal{T}}_\phi^\mu f\|$
- Extend to large F?
  - Naive: generate partition of size $O(1/\varepsilon^{|F|})$ X



Batch Value-Function Tournament [Xie & Jiang'20b]

- Algorithm: $\arg\min_{f\in\mathcal{F}} \max_{f'\in\mathcal{F}} \|f - \widehat{\mathcal{T}}_{\phi_{f,f'}} f\|_{2,D}$

  partition created out of f and f'

  - Inspired by Scheffé tournament & tournament algorithms for model selection in RL [Hallak et al'13, Jiang et al'15]
- Concern: not every ϕ is "good" (i.e., Q^*-irrelevant)
  - For $f = Q^*$: always tested on good ϕ => small error for all f'
  - For bad f : tested on a good ϕ when $f' = Q^*$ => large max error

Finite-sample analysis

- Previous reasoning builds on consistency of Q*-irrelevant abstractions
- Finite-sample guarantee additionally requires:

1. Concentration bounds: $\|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,D} \approx \|f - \mathcal{T}_\phi^\mu f\|_{2,\mu}$
   - Part of it is to show $\widehat{\mathcal{T}}_\phi^\mu f \approx \mathcal{T}_\phi^\mu f$, i.e., ERM close to population minimizer for non-realizable least-square!
   - Proof idea: all regression problems are *effectively* realizable in the eyes of histogram regressor
   - The other part: $\|\cdot\|_{2,D} \approx \|\cdot\|_{2,\mu}$ with $1/\sqrt{n}$ rate
2. Error-propagation: how $\|f - \mathcal{T}_\phi^\mu f\|_{2,\mu}$ controls $\|f - Q^\star\|_{2,\mu}$
   - In BRM: $f - Q^\star = \boxed{(f - \mathcal{T} f)} + \boxed{(\mathcal{T} f - \mathcal{T} Q^\star)}$
   - In BVFT: $f - Q^\star = \boxed{(f - \mathcal{T}_\phi^\mu f)} + \boxed{(\mathcal{T}_\phi^\mu f - \mathcal{T}_\phi^\mu Q^\star)}$

controlled by alg

determines error prop

Error propagation

How $\|f - \mathcal{T}_\phi^\mu f\|_{2,\mu}$ controls $\|f - Q^\star\|_{2,\mu}$

- Standard assumption: μ puts enough prob in each "block" of ϕ
- Corresponds to well-conditioned design matrix for linear class
- Problem: our ϕ is quite arbitrary
- Any assumption that is independent of ϕ?

Assumption 1. We assume that $\mu(s,a) > 0 \; \forall s, a$. We further assume that
(1) There exists constant $1 \le C_\mathcal{A} < \infty$ such that for any $s \in \mathcal{S}, a \in \mathcal{A}$, $\mu(a|s) \ge 1/C_\mathcal{A}$.
(2) There exists constant $1 \le C_\mathcal{S} < \infty$ such that for any $s \in \mathcal{S}, a \in \mathcal{A}, s' \in \mathcal{S}$, $P(s'|s,a)/\mu(s') \le C_\mathcal{S}$. Also $d_0(s)/\mu(s) \le C_\mathcal{S}$.
It will be convenient to define $C = C_\mathcal{S} C_\mathcal{A}$.

sample complexity:

$$\tilde{O}\left(\frac{C^2 \ln \frac{|\mathcal{F}|}{\delta}}{\epsilon^4 (1-\gamma)^8}\right)$$

- Key part: $P(s'|s,a)/\mu(s') \le C_\mathcal{S}$ [Munos'03]
- Satisfiable in MDPs whose transition matrix admits low-rank stochastic factorization

Limitations & Possibilities

Computationally intractable for training

Tractable for validation / model selection ✔

(choose among Q-functions produced by different training algs)

- Stronger than existing results (e.g., [Jiang et al'15])
- Potentially practical—ongoing empirical evaluation

Data assumption is very strong

- Open: standard concentrability (more next slide)?
- More challenging: data w/ insufficient coverage?

Finite-sample analyses of batch VFA

Example:	low-rank stoch. fac.	low-rank MDP	linear F & $\mathbb{E}_\mu[\varphi\varphi^\top] \succ 0$
	$\max_{s,a,s'} P(s'\|s,a)/\mu(s')$	$\max_{\pi} \\|d^\pi/\mu\\|_\infty$	$\max_{\pi,f,f'} \frac{\\|f-f'\\|_{d^\pi}}{\\|f-f'\\|_\mu}$
$\forall f \in \mathcal{F}, \mathcal{T}f \in \mathcal{F}$	✔	✔	✔
$Q^\star \in \mathcal{F}$	**X?**	**X?** (conj. [Chen&Jiang'19])	**X?**

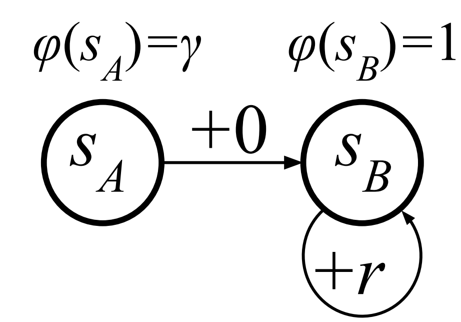
speculation prior to 2020

- Variations in data assumptions are minor
- Linear F may be easy?

Both Wrong!

Finite-sample analyses of batch VFA

Example:	low-rank stoch. fac.	low-rank MDP	linear F & $\mathbb{E}_\mu[\varphi\varphi^\top] \succ 0$
	$\max_{s,a,s'} P(s'\|s,a)/\mu(s')$	$\max_{\pi} \\|d^\pi/\mu\\|_\infty$	$\max_{\pi,f,f'} \frac{\\|f-f'\\|_{d^\pi}}{\\|f-f'\\|_\mu}$
$\forall f \in \mathcal{F}, \mathcal{T} f \in \mathcal{F}$	✔	✔	✔
$Q^\star \in \mathcal{F}$	✔ (BVFT, general F)	? (conj. [Chen&Jiang'19])	X (even w/ linear F)



Amortila et al'20, inspired by Wang et al'20

speculation prior to 2020

- Variations in data assumptions are minor
- Linear F may be easy?

Both Wrong!

Batch Value-function Approximation with Only Realizability.
Tengyang Xie, Nan Jiang. arXiv-20.

Additional References

- A Variant of the Wang-Foster-Kakade Lower Bound for the Discounted Setting. Philip Amortila, Nan Jiang, Tengyang Xie. arXiv-20.
- Q* Approximation Schemes for Batch Reinforcement Learning: A Theoretical Comparison. Tengyang Xie, Nan Jiang. UAI-20.
- Information-Theoretic Considerations in Batch Reinforcement Learning. Jinglin Chen, Nan Jiang. ICML-19.
- Nan Jiang, Alex Kulesza, Satinder Singh. Abstraction Selection in Model-based Reinforcement Learning. ICML-15.

Thank you!
Questions?