Minimax Methods for Off-policy Evaluation and Policy Optimization

Nan Jiang

Apr 4, 2020
@ iDS2 Seminar

Based on works with my students
Masatoshi Uehara, Jiawei Huang, Tengyang Xie



Value-based RL

- e.g., FQI: learn Q^* from a batch dataset $\{(s, a, r, s')\}$ with function approximator F

$$f_t = \arg\min_{f \in \mathcal{F}} \sum_{(s,a,r,s') \in D} \left(f(s,a) - \left(r + \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s', a') \right) \right)^2$$

- Aspects we've got used to
  - Squared loss: surrogate (expected return is a plain average)
  - Bootstrapped targets (iterative / changing optimization obj): error accumulation (sometimes exponentially in horizon)
- This talk: off-policy RL without squared loss, without bootstrapped targets, nice control of error accumulation, etc
  - 3-line derivations (assume you know Bellman eq)

Off-policy Evaluation (OPE)

- Evaluate policy π using "off-policy" data (e.g., sampled using a different policy π_b)
- Why care?
  - Key to the success of supervised learning: **training/validation**
  - OPE is the validation process for RL
- How difficult?
  - Without further assumptions, worst-case error is exponential in horizon (unless π and π_b are extremely close) for any method —"curse of horizon"

Off-policy Evaluation (OPE)

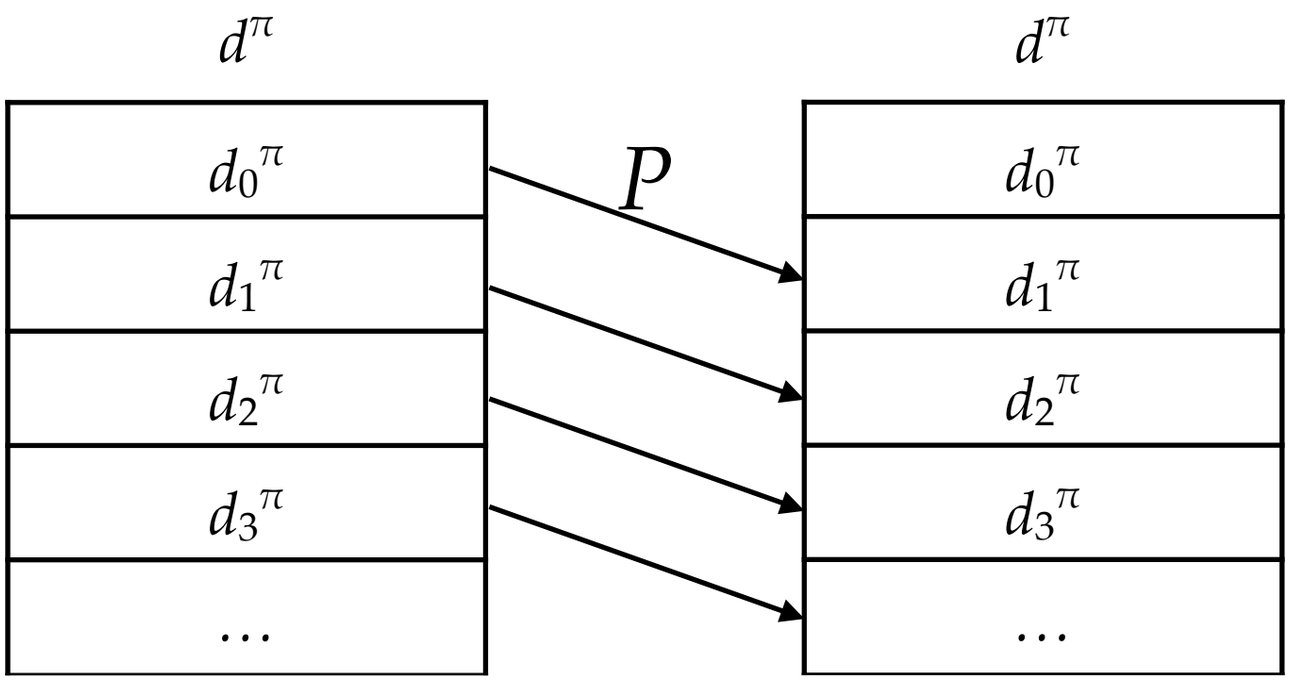
- π *will* induce random trajectory $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$
  - s_0 is a known deterministic start state (wlog)
  - $a_t \sim \pi(\cdot \mid s_t)$
  - $r_t = R(s_t, a_t)$, $s_{t+1} \sim P(\cdot \mid s_t, a_t)$
  - Want to estimate: $J(\pi) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi\right]$
    - $J(\pi) = Q^\pi(s_0, \pi)$, where $Q^\pi(s, a) = \mathbb{E}_{r,s'|s,a}[r + \gamma Q^\pi(s', \pi)]$
  - Data: i.i.d. (s, a, r, s') tuples. $(s, a) \sim \mu$

OPE Methods

Method	Func Approx	Stat Error	Comment
Importance Sampling / Doubly Robust	No	Exponential in horizon	Needs trajectory data & stochastic behavior policy
Model-based	Realizable model class	Complexity of function approximation	
Fitted-Q	$\mathcal{T}^\pi q \in \mathcal{Q}, \forall q \in \mathcal{Q}$		Linear: divergent under realizability
MWL	$w^\pi \in \mathcal{W}$, $Q^\pi \in \text{sp}(\mathcal{Q})$		Linear: LSTDQ only 1 realizability needed
MQL	$Q^\pi \in \mathcal{Q}$, $w^\pi \in \text{sp}(\mathcal{W})$		
MCI	Either $Q^\pi \in \mathcal{Q}$ or $w^\pi \in \mathcal{W}$ (for convex $\mathcal{Q}$ and $\mathcal{W}$)		Confidence Interval

More notations

- π will induce random trajectory $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$
  - s_0 is a known deterministic start state (wlog)
  - Want to estimate: $J(\pi) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi\right]$
- Let d_t^π denote the marginal distribution of (s_t, a_t)
- Occupancy measure: $d^\pi := \sum_{t=0}^{\infty} \gamma^t d_t^\pi$.
  - Note $J(\pi) = \mathbb{E}_{(s,a)\sim d^\pi,\ r=R(s,a)}[r]$
  - Bellman eq for occupancy



More notations

- π will induce random trajectory $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$
  - s_0 is a known deterministic start state (wlog)
  - Want to estimate: $J(\pi) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid \pi\right]$
- Let d_t^π denote the marginal distribution of (s_t, a_t)
- Occupancy measure: $d^\pi := \sum_{t=0}^{\infty} \gamma^t d_t^\pi$.
  - Note $J(\pi) = \mathbb{E}_{(s,a)\sim d^\pi,\ r=R(s,a)}[r]$
  - Bellman eq for occupancy
- Data: i.i.d. (s, a, r, s') tuples where $(s, a) \sim \mu$
- We assume data "well covers" π : $\max_{s,a} \frac{d^\pi(s,a)}{\mu(s,a)} \leq C$

Marginalized Importance Sampling

- Idea: if we have access to

$$w^\pi(s,a) := \frac{d^\pi(s,a)}{\mu(s,a)}$$

$d^\pi(s,a)$: occupancy of π

$\mu(s,a)$: data dist.

OPE is easy:

$$J(\pi) = \mathbb{E}_{(s,a)\sim\mu}[w^\pi(s,a)\cdot r] =: \mathbb{E}_{w^\pi}[r]$$

  - Note: μ omitted in shorthand for simplicity

  - Can we learn w^π ?

  - If we have function classes $\mathcal{W}$ and $\mathcal{Q}$ to well approximate w^π and Q^π, resp, then we are in business!

Derivation

- Easy fact 1: $J(\pi) - Q^{\pi}(s_0, \pi) = 0$

$$\mathbb{E}_w[\ldots] := \mathbb{E}_{(s,a)\sim\mu}[w(s,a) \cdot \ldots]$$

- Easy fact 2: $0 = \mathbb{E}_w[r + \gamma Q^{\pi}(s', \pi) - Q^{\pi}(s, a)]$ for any w
  - Q^{π} respects Bellman eq on every $(s,\ a)$
- Combined:

$$J(\pi) - Q^{\pi}(s_0, \pi) = \mathbb{E}_w[r + \gamma Q^{\pi}(s', \pi) - Q^{\pi}(s, a)]$$

- Move terms:

$$J(\pi) - \mathbb{E}_w[r] = Q^{\pi}(s_0, \pi) + \mathbb{E}_w[\gamma Q^{\pi}(s', \pi) - Q^{\pi}(s, a)]$$

- Our goal: find w s.t. $J(\pi) \approx \mathbb{E}_w[r]$
- Approach: find w that minimizes the (abs. value of) RHS
- Don't know Q^{π}, but if $Q^{\pi} \in \text{sp}(\mathcal{Q})$, can do:

$$\arg\min_{w\in\mathcal{W}} \max_{q\in\mathcal{Q}} \left| q(s_0, \pi) + \mathbb{E}_w[\gamma q(s', \pi) - q(s, a)] \right|$$

MWL (Minimax Weight Learning)

- Key lemma:
  $J(\pi) - \mathbb{E}_w[r] = Q^\pi(s_0, \pi) + \mathbb{E}_w[\gamma Q^\pi(s', \pi) - Q^\pi(s, a)]$
- Algorithm:
  $\arg\min_{w \in \mathcal{W}} \max_{q \in \mathcal{Q}} \left| q(s_0, \pi) + \mathbb{E}_w[\gamma q(s', \pi) - q(s, a)] \right|$
- If $Q^\pi \in \text{sp}(\mathcal{Q})$, $\left| J(\pi) - \mathbb{E}_w[r] \right| \leq \max_{q \in \mathcal{Q}} |\ldots|$
  - When $w = w^\pi$, $\max_{q \in \mathcal{Q}} |\ldots| = 0$
    - Proof: Bellman equation for occupancy measure
  - If $Q^\pi \in \text{sp}(\mathcal{Q})$ and $w^\pi \in \mathcal{W}$, no guarantee that w^π will be learned, but $\mathbb{E}_w[r]$ will be accurate!
    - Can learn w^π if $\mathcal{Q}$ is “rich enough”

Related work

- First algorithm of this kind: Liu et al'18 "breaking the curse of horizon".
- Followup work: DualDICE, GenDICE, …
- One-shot case ($\gamma = 0$) with RKHS $\mathcal{Q}$: Kernel mean matching
  - to see this, assume an initial state distribution d_0

$$\arg\min_{w \in \mathcal{W}} \max_{q \in \mathcal{Q}} \left| \mathbb{E}_{s_0 \sim d_0}[q(s_0, \pi)] - \mathbb{E}_w[q(s, a)] \right|$$

  - learning w to convert source distribution μ to target distribution $d_0 \times \pi$
  - Difference: w can be an independent weight for each data point in KMM, but needs to be parameterized in the multi-step case (unless with deterministic dynamics?)

OPE Methods

Method	Func Approx	Statistical Error	Comment
Importance Sampling / Doubly Robust	No	Exponential in horizon	Needs trajectory data & stochastic behavior policy
Model-based	Realizable model class	Complexity of function approximation	
Fitted-Q	$\mathcal{T}^\pi q \in \mathcal{Q}, \forall q \in \mathcal{Q}$		Linear: divergence
MWL	$w^\pi \in \mathcal{W}$, $Q^\pi \in \text{sp}(\mathcal{Q})$		Linear: LSTDQ only 1 realizability needed
MQL	$Q^\pi \in \mathcal{Q}$, $w^\pi \in \text{sp}(\mathcal{W})$		
MCI	Either $Q^\pi \in \mathcal{Q}$ or $w^\pi \in \mathcal{W}$ (for convex $\mathcal{Q}$ and $\mathcal{W}$)		Confidence Interval

MQL (Minimax Q-function Learning)

- Takeaway: using value function as discriminators can help learning a good weighting function w for OPE
  - w becomes first-class citizen, q secondary
- What if we want to learn q s.t. $q(s_0, \pi) \approx J(\pi)$?
- Lemma: $J(\pi) - q(s_0, \pi) = \mathbb{E}_{w^\pi}[r + \gamma q(s', \pi) - q(s, a)]$
- Algorithm: $\arg\min_{q \in \mathcal{Q}} \max_{w \in \mathcal{W}} |\mathbb{E}_w[r + \gamma q(s', \pi) - q(s, a)]|$
- A form of Bellman error (residual) minimization, but
  - Classical residual estimation considers state-wise Bellman error: $(q(s, a) - \mathbb{E}_{r,s'|s,a}[r + \gamma q(s', \pi)])^2$
  - unbiased estimation requires "double sampling" [Baird'95]
  - We estimate "average" Bellman errors w.r.t. different distributions, no double sampling needed

OPE Methods

Method	Func Approx	Statistical Error	Comment
Importance Sampling / Doubly Robust	No	Exponential in horizon	Needs trajectory data & stochastic behavior policy
Model-based	Realizable model class	Complexity of function approximation	
Fitted-Q	$\mathcal{T}^\pi q \in \mathcal{Q}, \forall q \in \mathcal{Q}$		Linear: divergence
MWL	$w^\pi \in \mathcal{W}$, $Q^\pi \in \text{sp}(\mathcal{Q})$		Linear: LSTDQ only 1 realizability needed
MQL	$Q^\pi \in \mathcal{Q}$, $w^\pi \in \text{sp}(\mathcal{W})$		
MCI	Either $Q^\pi \in \mathcal{Q}$ or $w^\pi \in \mathcal{W}$ (for convex $\mathcal{Q}$ and $\mathcal{W}$)		Confidence Interval

Compare MWL/MQL to Fitted-Q in linear setting

- If Q^π is linear in features ϕ, how do different algorithms perform?
- ADP (e.g., Fitted-Q): divergence [Gordon'95, Tsitsiklis & Van Roy'96]
- MQL?
  - Use the same class as $\mathscr{W}$, even if $\mathscr{W}$ doesn't capture w^π
  - Population ver of MQL (no stat. error) is accurate!
  - It's a familiar alg: LSTDQ [Lagoudakis & Parr'03]

$$\textbf{for each } (s,a,r,s') \in D$$
$$\widetilde{\mathbf{A}} \leftarrow \widetilde{\mathbf{A}} + \phi(s,a)\Big(\phi(s,a) - \gamma\phi\big(s',\pi(s')\big)\Big)^\intercal$$
$$\widetilde{b} \leftarrow \widetilde{b} + \phi(s,a)r$$

$$\widetilde{w}^\pi \leftarrow \widetilde{\mathbf{A}}^{-1}\widetilde{b}$$

Compare MWL/MQL to Fitted-Q in linear setting

- If Q^π is linear in features ϕ, how do different algorithms perform?
- ADP (e.g., Fitted-Q): divergence [Gordon'95, Tsitsiklis & Van Roy'96]
- MQL?
  - Use the same class as $\mathscr{W}$, even if $\mathscr{W}$ doesn't capture w^π
    - Population ver of MQL (no stat. error) is accurate!
    - It's a familiar alg: LSTDQ [Lagoudakis & Parr'03]
    - (Who thought LSTDQ is good for estimating $J(\pi)$?!)
    - LSTDQ "is not TD": it's not really using squared loss, not using bootstrapped targets, but doing moment matching!
  - MWL with the same linear classes: also LSTDQ :)
  - Just as ADP work with non-linear func approx, MWL/MQL is the natural generalization of LSTDQ

Policy optimization?

- MQL solves Bellman eq for policy evaluation with the help of importance weight discriminators
- Can we solve Bellman optimality eq (for Q*) in the same way?
- $\arg\min_{q \in \mathcal{Q}} \max_{w \in \mathcal{W}} |\mathbb{E}_w[r + \gamma \max_{a'} q(s', a') - q(s, a)]|$
  - Analyses and assumptions way cleaner than AVI methods
    - AVI incurs $O(H^2)$ error [Scherrer and Lesner'12]; we incur $O(H)$
- If $\mathcal{W}$ is an RKHS, this is (almost) kernel-loss [Feng et al'19]
  - overlapping authors with Liu et al'18 missed the connections between the two papers…

Beyond (double) realizability

- Back to OPE: **both** $\mathscr{W}$ and $\mathscr{Q}$ need to be realizable
- Recall that $w^\pi(s,a) := \frac{d^\pi(s,a)}{\mu(s,a)}$: this object may not even exist (if denominator = 0)!
- How to address model misspecification?
- Another issue:
  - MWL: $\arg\min_{w\in\mathscr{W}} \max_{q\in\mathscr{Q}} |q(s_0,\pi) + \mathbb{E}_w[\gamma q(s',\pi) - q(s,a)]|$
  - MQL: $\arg\min_{q\in\mathscr{Q}} \max_{w\in\mathscr{W}} |\mathbb{E}_w[r + \gamma \max_{a'} q(s',a') - q(s,a)]|$
  - MWL does not use rewards, MQL does not use s_0
- Yet another one: I don't want two algorithms…
- Let's kill 3 birds with 1 stone…!

Minimax Confidence Interval

- Back to our key lemma for MWL:

$$J(\pi) = Q^\pi(s_0, \pi) + \mathbb{E}_w[r + \gamma Q^\pi(s', \pi) - Q^\pi(s, a)] =: L(w, Q^\pi)$$

- If we have poor $\mathscr{W}$ but realizable $\mathscr{Q}$:

$$\min_q L(w, q) \le J(\pi) = L(w, Q^\pi) \le \max_q L(w, q),\ \forall w$$

$$\max_w \min_q L(w, q) \le J(\pi) \le \min_w \max_q L(w, q)$$

- What if we start with the key lemma for MQL?

$$J(\pi) = q(s_0, \pi) + \mathbb{E}_{w^\pi}[r + \gamma q(s', \pi) - q(s, a)] = L(w^\pi, q)$$

- If we have poor $\mathscr{Q}$ but realizable $\mathscr{W}$:

$$\max_q \min_w L(w, q) \le J(\pi) \le \min_q \max_w L(w, q)$$

- For convex $\mathscr{Q}$ and $\mathscr{W}$, a pair of **reversed** intervals
- All components of data (rewards, initial state) are used
- Blue is Frenchel AlgaeDICE [Nachum et al'19]

Minimax Confidence Interval

- These intervals are not only useful for policy evaluation
- Consider non-exploratory batch data ($\mathscr{W}$ is poor)
  - Exploitation/robust/pessimism: $\max_\pi \max_w \min_q L(w, q; \pi)$
  - Exploration/optimism: $\max_\pi \min_w \max_q L(w, q; \pi)$
    - Tabular: Rmax
    - Function approximation variant: OLIVE [Jiang et al'17] for low Bellman rank problems

OPE Methods

Method	Func Approx	Statistical Error	Comment
Importance Sampling / Doubly Robust	No	Exponential in horizon	Needs trajectory data & stochastic behavior policy
Model-based	Realizable model class	Complexity of function approximation	
Fitted-Q	$\mathcal{T}^\pi q \in \mathcal{Q}, \forall q \in \mathcal{Q}$		Linear: divergence
MWL	$w^\pi \in \mathcal{W}$, $Q^\pi \in \text{sp}(\mathcal{Q})$		Linear: LSTDQ only 1 realizability needed
MQL	$Q^\pi \in \mathcal{Q}$, $w^\pi \in \text{sp}(\mathcal{W})$		
MCI	Either $Q^\pi \in \mathcal{Q}$ or $w^\pi \in \mathcal{W}$ (for convex $\mathcal{Q}$ and $\mathcal{W}$)		Confidence Interval

Challenges

- Optimization
- Without realizability?
- Quantification of statistical errors
- Practical procedure without hyper-parameter tuning

References

- Minimax Weight and Q-Function Learning for Off-Policy Evaluation.

- Minimax Confidence Interval for Off-Policy Evaluation and Policy Optimization.

- *Q** Approximation Schemes for Batch Reinforcement Learning: A Theoretical Comparison.