Doubly Robust Off-policy Value Evaluation for Reinforcement Learning

Nan Jiang[1,2], Lihong Li[2]

[1]University of Michigan, [2]Microsoft Research

Abstract

What is the problem

Evaluating a policy using data produced by a **different** policy.

target policy *behavior policy*

When do we encounter the problem

Verify the safety of a new policy before deploying it in the real system
-- a critical step of RL in real-world applications, e.g.

- Adaptive medical treatment
- Dialog systems
- Customer relationship management

	Importance Sampling	Regression-based methods	**Our Doubly Robust estimator**
Low variance?	✗	✓	✓
Controlled bias?	✓	✗	✓

We also proved statistical lower bound of the problem, and the DR estimator matches the bound in certain scenarios.

Notations

- MDP $M = \langle S, A, P, R \rangle$, initial state distribution μ, horizon H
- Behavior policy π_0, target policy π_1
- Dataset $D = \left\{ (s_1, a_1, r_1, s_2, \ldots, s_{H+1}),\ a_t \sim \pi_0(\cdot|s_t) \right\}$
- Objective: estimating the value of π_1

  $V^{\pi_1} = \mathbb{E}\left[\sum_{t=1}^{H} r_t \mid a_t \sim \pi_1(\cdot|s_t)\right]$, abbreviated as V

Existing Solutions

- ***I**mportance **S**ampling*[1] (step-wise version) $V_{\text{step-IS}} := \sum_{t=1}^{H} \rho_{1:t} r_t$

  where $\rho_t = \pi_1(a_t|s_t)/\pi_0(a_t|s_t)$ and $\rho_{1:t} := \prod_{t'=1}^{t} \rho_{t'}$
  - Unbiased, high variance (exp. in horizon)
- ***Reg**ression-based estimator* (a.k.a., "model-based", "direct method")

  e.g., in contextual bandits, regress $\widehat{R}$ from $\{(s,a) \mapsto r\}$

  $V_{\text{REG}} := \widehat{V}(s) = \sum_{a'} \pi_1(a') \, \widehat{R}(s,a)$ (also need to regress P in the MDP case)
  - Typically low variance with function approximation (FA).
  - FA introduces uncontrolled bias.

Doubly Robust Estimator for RL

Re-expression of step-wise IS in recursive form:

$$V^{H+t-1}_{\text{step-IS}} = \rho_t \left(r_t + V^{H+t}_{\text{step-IS}}\right)$$

$V(s_t) = Q(s_t, \pi_1(s_t))$ $\Leftarrow$ $Q(s_t, a_t)$ Unbiased estimate of $V(s_{t+1})$

Apply DR trick at each horizon: (see bandit version in [2])

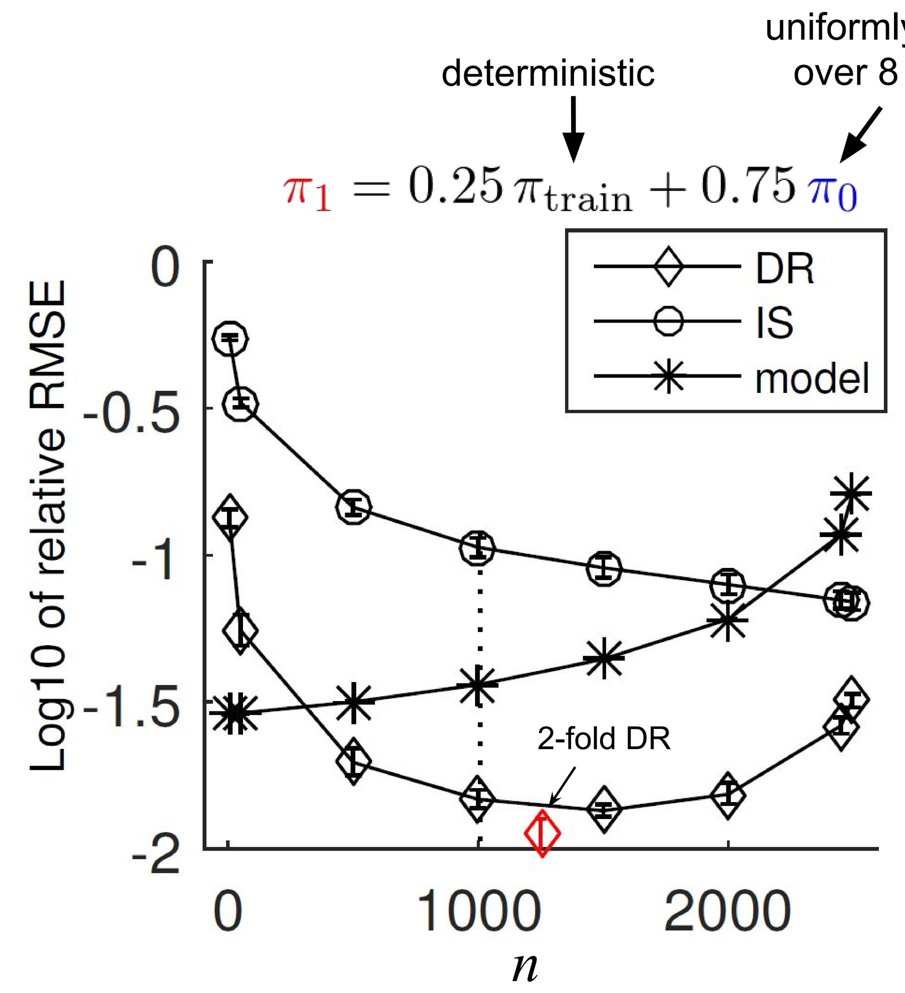
$$V^{H-t+1}_{\text{DR}} = \widehat{V}(s_t) + \rho_t\left(r_t + V^{H-t}_{\text{DR}} - \widehat{Q}(s_t, a_t)\right)$$

Properties of DR:

- Unbiased, regardless of how poor $\widehat{Q}$ is (hat terms cancel in expectation).
- 0 variance if MDP is deterministic and $\widehat{Q} \equiv Q$ (hence $\widehat{V} \equiv V$).
- step-wise IS = DR with $\widehat{Q} \equiv 0$.

⇒ DR can have lower variance if $\widehat{Q}$ is better than a trivial function!

Experiment: Comparing Point Estimates

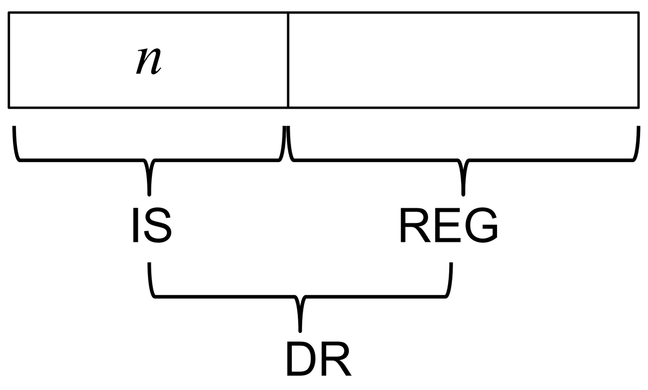


Domain: *Sailing*

Horizon: $H = 20$

$(\max \rho_{1:H} : 6 \times 10^8)$

Data: 2500 trajectories



Experiment: Safe Policy Improvement

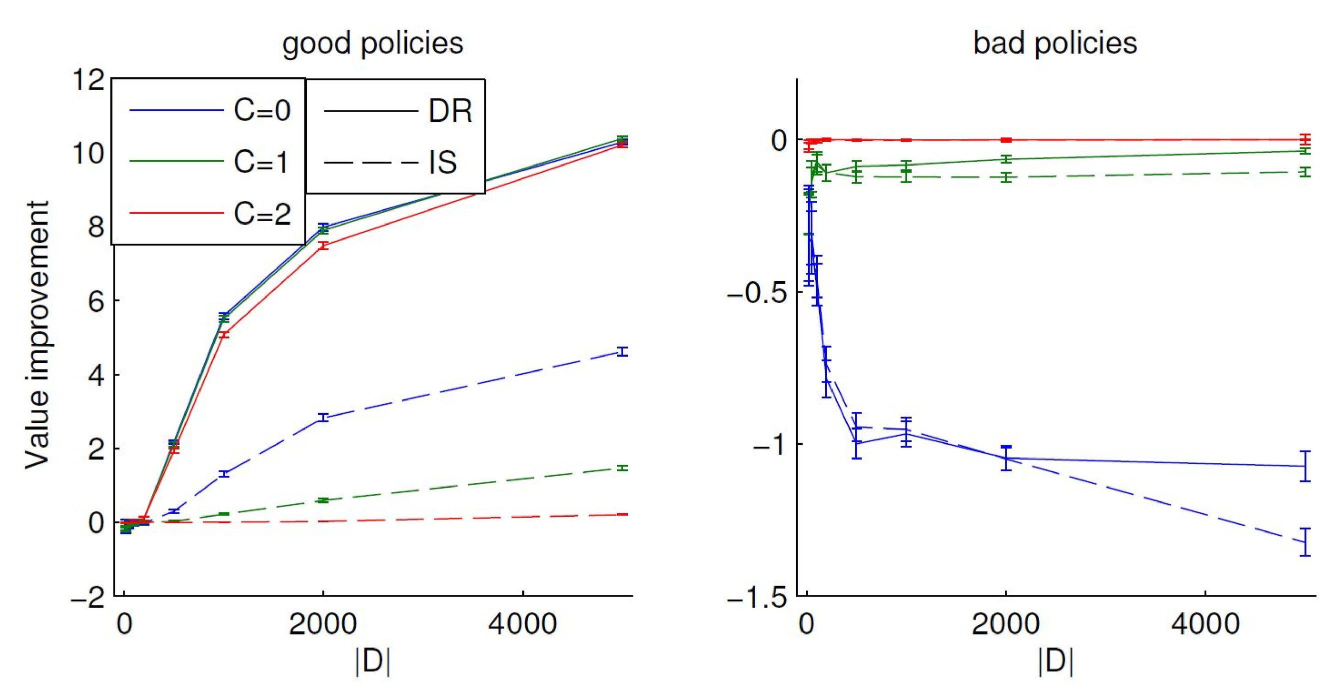
Setting: given batch data, recommend better policies (and reject bad ones)

Detailed Experiment Setup: (domain: Mountain Car)

1. Split data into two halves, compute π_{train} from 1st half;
2. Mix π_{train} and π_0 with various ratios;
3. Evaluate the mixed policies on the 2nd half of the data;
4. Recommend policy with the highest lower confidence bound (LCB).

Compared methods: DR and step-wise IS

+Gaussian approximation of LCB, i.e., LCB = mean - *C* * standard error.



DR recommends good policies more aggressively without sacrificing safety against bad policies.

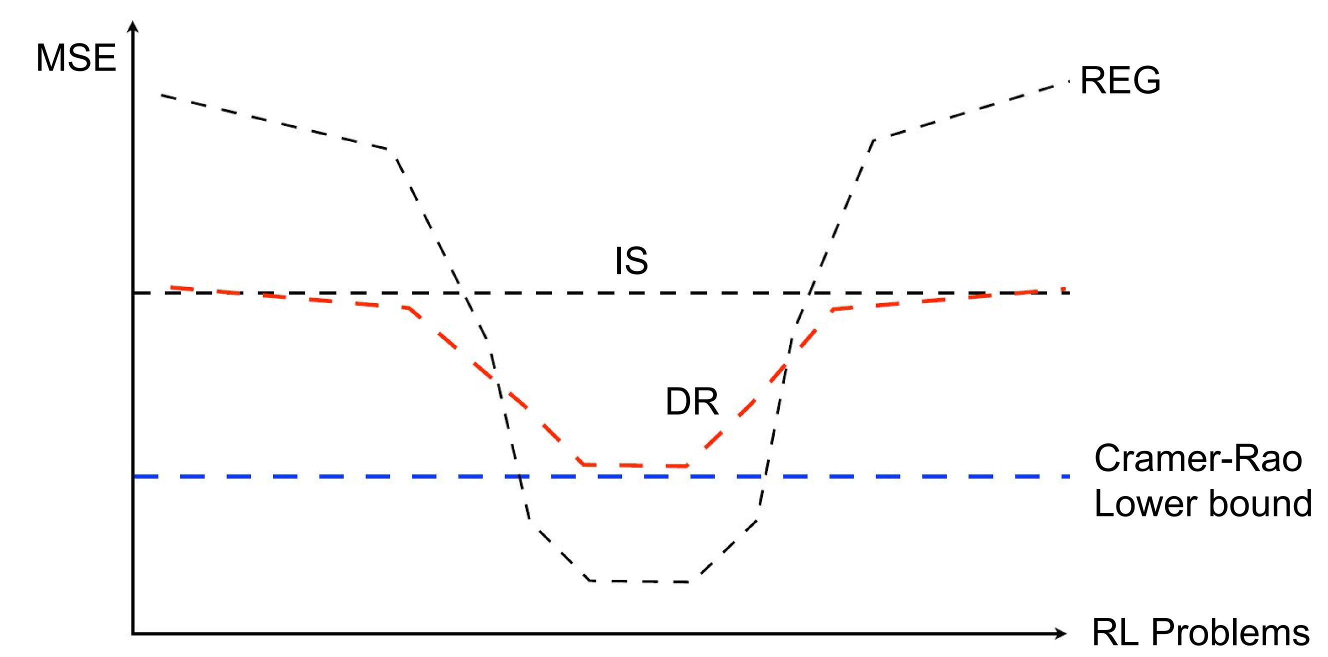
On the Hardness of the Problem

A most difficult situation

- Partially Observable MDP.
- Want the most credible evaluation: no assumption in evaluation phase.

Variance of DR in this case (simplification: only reward at step H+1)

$$\sum_{t=1}^{H+1} \mathbb{E}_{s_1:a_{t-1}}\left[\rho_{1:(t-1)}^2\left(\underbrace{\mathbb{V}_{s_t|s_1:a_{t-1}}[V(s_t)]}_{\text{lower bound (intrinsic variance)}} + \underbrace{\mathbb{V}_{a_t|s_t}\left[\rho_t\left(Q(s_t,a_t) - \widehat{Q}(s_t,a_t)\right)\right]}_{\text{improves with a good } \widehat{Q}}\right)\right]$$



References

[1] Precup, Sutton, and Singh. Eligibility traces for off-policy policy evaluation. In Proc. of the 17th Int. Conf. on Machine Learning, pages 759–766, 2000.

[2] Dudık, Langford, and Li. Doubly robust policy evaluation and learning. In Proc. of the 28th Int. Conf. on Machine Learning, pages 1097–1104, 2011.