Doubly Robust Off-policy Value Evaluation for Reinforcement Learning

Nan Jiang¹,², Lihong Li²
¹University of Michigan, ²Microsoft Research

Abstract

What is the problem
Evaluating a policy using data produced by a different 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? X ✓ ✓
Controlled bias? ✓ X ✓

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

Notations

- MDP \( M = (S, A, P, R) \), initial state distribution \( \mu \), horizon \( H \)
- Behavior policy \( \pi_0 \), target policy \( \pi_1 \)
- Dataset \( D = \{ (s_1, a_1, r_1, s_2, \ldots, s_{H+1}) \}_{i=1}^N \), \( a_t \sim \pi_t(\cdot|s_t) \)
- Objective: estimating the value of \( \pi_1 \)
  \[ V_{\pi_1} = \mathbb{E} \left[ \sum_{t=1}^{H} r_t | a_t \sim \pi_1(\cdot|s_t) \right], \]
  abbreviated as \( V_1 \)

Existing Solutions

- Importance Sampling\(^{(1)}\) (step-wise version) \[ V_{IS}^{step-wise} = \sum_{t=1}^{H} \pi_{1,t} r_t \]
  where \( \pi_{1,t} = \pi_1(\cdot|s_t) / \pi_0(\cdot|s_t) \) and \( \pi_{1,t} := \prod_{s=1}^{t} \pi_{1,s} \)
  - Unbiased, high variance (exp. in horizon)
- Regression-based estimator (a.k.a., “model-based”, “direct method”)
  e.g., in contextual bandits, regress \( \tilde{R}(s,a) \) from \( \{s,a \rightarrow r\} \)
  \[ V_{REG} := \tilde{V}(s) = \sum_{t=1}^{H} \pi_t(\cdot|s_t) \tilde{R}(s, a) \]
  (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:
\[ V_{IS}^{step-wise} = \rho_1(r_1) + V_{IS}^{step-wise} \]
\[ V_{IS}^{step-wise} := Q(s_1, \pi_1(s_1)) \]
Unbiased estimate of \( V_{\pi_1} \)

Apply DR trick at each horizon: (see bandit version in \([2]\])
\[ V_{DR}^{H+1} = \tilde{V}(s_{H+1}) + \rho_1(r_{H+1}) + V_{DR}^{H} - Q(s_{H+1}, a_{H+1}) \]

Properties of DR:
- Unbiased, regardless of how poor \( \tilde{Q} \) is (hat terms cancel in expectation).
- 0 variance if MDP is deterministic and \( \tilde{Q} = Q \) (hence \( \tilde{V} = V \)).
- step-wise IS = DR with \( \tilde{Q} = Q \)
  \( \Rightarrow \) DR can have lower variance if \( \tilde{Q} \) is better than a trivial function!

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 \( \pi_{\text{train}} \) from 1st half;
  2. Mix \( \pi_{\text{train}} \) and \( \pi_0 \) 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

Approximation of LCB, i.e., \( \text{LCB} = \text{mean} - C \times \text{standard error} \).

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
\[ \sum_{H=1}^{N} \rho_{H} (r_{H+1}) + V_{DR}^{H} - Q(s_{H+2}, a_{H+2}) \]

(simplification: only reward at step \( H+1 \))

Improves with a good \( \tilde{Q} \)

References