High-Confidence Off-Policy Evaluation

Philip S. Thomas, Georgios Theocharous, Mohammad Ghavamzadeh

AAAI-15. 29th AAAI Conference on Artificial Intelligence, Austin, Texas, USA, January 2015.

Presenter: Robert L. Logan IV
Overview

This paper develops a method to compute the confidence that the expected return of a policy exceeds some lower bound, using only trajectories generated from other policies.

Terminology:

▶ *Evaluation Policy* - The policy which we wish to estimate the expected return of.
▶ *Behavior Policy* - The policies used to estimate the return of the evaluation policy.
"...execution of a new policy can be costly or dangerous if it performs worse than the policy that is currently being used..."

Examples:

- News recommendation systems
- Patient diagnosis systems
- Neuroprosthetic control
- Automatic drug administration
This work follows the standard Markov Decision Process (MDP) formalism:

- $\mathcal{S}$: State space.
- $\mathcal{A}$: Action space.
- $r_t \in [r_{max}, r_{min}]$: Reward at time $t$.
- $\gamma \in [0, 1]$: Discount factor.
- $\pi(a \mid s, \theta)$: Probability of taking action $a$ in state $s$ given policy parameters $\theta$.
- $\tau = \{s_1, a_1, r_1, \ldots, s_T, a_T, r_T\}$: A trajectory.
Problem Setup

Define the normalized and discounted return of a trajectory to be:

\[ R(\tau) = \frac{\left( \sum_t \gamma^{t-1} r_t \right) - R_-}{R_+ - R_-} \]

Where \( R_+ \) and \( R_- \) are upper and lower bounds for \( \sum_t \gamma^{t-1} r_t \).

Ideally, we want to know the expected return given the evaluation policy parameters \( \theta \):

\[ \rho(\theta) = \mathbb{E} [ R(\tau) | \theta ] \]
Generating Unbiased Estimates of $\rho(\theta)$

**Key Idea**
Given a dataset $D = \{(\tau_i, \theta_i) : \tau_i \text{ generated using } \theta_i\}$ estimate $\rho(\theta)$ using *importance sampling*.

**Importance Sampling**
Given a *target distribution* $p$ and *sampling distribution* $q$ and a function $f$:

$$\mathbb{E}_{x \sim p}(f(x)) = \frac{1}{n} \sum_{i=1}^{n} f(x_i) \frac{p(x_i)}{q(x_i)}$$

where $x_i \sim q$. 
Generating Unbiased Estimates of $\rho(\theta)$

In the context of our problem, the target distribution is $Pr(\tau_i \mid \theta)$ (e.g. the probability of a trajectory under the evaluation policy), and the sampling distribution is $Pr(\tau_i \mid \theta_i)$ (e.g. the probability of a trajectory under the behavior policy it was generated by).

The expected return of the evaluation policy can then be estimated by:

$$\rho(\theta) \approx \frac{1}{n} \sum_{i=1}^{n} R(\tau_i) \frac{Pr(\tau_i \mid \theta)}{Pr(\tau_i \mid \theta_i)}$$

$$= \frac{1}{n} \sum_{i=1}^{n} R(\tau_i) \prod_{t} \frac{\pi(a_t \mid s_t, \theta)}{\pi(a_t \mid s_t, \theta_i)}$$

$$\hat{\rho}(\theta, \tau_i, \theta_i)$$
Recall our motivation:
"..execution of a new policy can be costly or dangerous if it performs worse than the policy that is currently being used..."
”Classical” Results

In the following slides, we will review a few ”classical” results providing lower bounds for the expectation of a random variable that can be estimated using samples.

We introduce the following variables/assumptions:

- \( X_i \): real-valued, positive, bounded random variables (e.g. importance weighted returns).
- \( \mu \): \( E(X_i) \) for all \( X_i \).
- \( b \): A real-number satisfying \( Pr[X_i < b] = 1 \) for all \( X_i \).
"Classical" Results

Chernoff-Hoeffding (CH) inequality
With probability at least $1 - \delta$:

$$\mu \geq \frac{1}{n} \sum_{i=1}^{n} X_i - b \sqrt{\frac{\ln \left( \frac{1}{\delta} \right)}{2n}}$$

Anderson (AM) inequality
With probability at least $1 - \delta$:

$$\mu \geq z_n - \sum_{i=0}^{n-1} (z_{i+1} - z_i) \min \left\{ 1, \frac{i}{n} + \sqrt{\frac{\ln \left( \frac{2}{\delta} \right)}{2n}} \right\}$$

where $z_i$ are the samples $X_i$ in increasing order, and $z_0 = 0$. 
"Classical" Results

Maurer & Pontil’s empirical Bernstein (MPeB) inequality

With probability at least $1 - \delta$:

$$
\mu \geq \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{7b \ln \left(\frac{2}{\delta}\right)}{3(n - 1)} - \frac{1}{n} \sqrt{\frac{\ln \left(\frac{2}{\delta}\right)}{n - 1} \sum_{i,j=1}^{n} (X_i - X_j)^2}
$$
The author’s use the MPeB inequality to prove the following:

**Theorem 1**

Let $Y_i = \min\{X_i, c_i\}$ where $c_i > 0$, with probability at least $1 - \delta$:

$$
\mu \geq \left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \sum_{i=1}^{n} \frac{Y_i}{c_i} - \left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \frac{7n \ln (2/\delta)}{3(n - 1)} - \left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \sqrt{\frac{\ln (2/\delta)}{n - 1} \sum_{i,j=1}^{n} \left( \frac{Y_i}{c_i} - \frac{Y_j}{c_j} \right)^2}
$$
These inequalities can be inverted so that given a lower bound \( \mu_- \) we can determine the confidence (e.g. value of \( \delta \)) that \( \mu \geq \mu_- \).

Theorem 1 requires pre-specified thresholds \( c_i \). In the paper, the authors select a single value \( c^* \) and set all \( c_i = c^* \). They show how an optimal value of \( c^* \) can be determined by splitting the dataset and performing cross-validation.
Experiment 1: The Mountain Car Problem

Demo
https://www.youtube.com/watch?v=x_qDs2kA7H4&feature=youtu.be
Experiment 1: The Mountain Car Problem
Experiment 1: The Mountain Car Problem
Experiment 1: The Mountain Car Problem

<table>
<thead>
<tr>
<th>Method</th>
<th>$\rho_-$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thm. 1</td>
<td>0.154</td>
</tr>
<tr>
<td>CH</td>
<td>-5,831,000</td>
</tr>
<tr>
<td>MPeB</td>
<td>-129,703</td>
</tr>
<tr>
<td>AM</td>
<td>0.055</td>
</tr>
</tbody>
</table>

Table: 95% confidence lower bounds
Experiment 2: Digital Marketing using Real-World Data

Problem
Decide the optimal policy for user-specific targeting of advertisements. Rewards are measured by advertisement click rates (e.g. agent receives +1 if a user clicks the ad, and 0 otherwise).

Data
- From a website for a Fortune 50 company.
- > 100,000 visitors a day.
- Each user has 31 features.
- Agent must select from two clusters of advertisements.
Experiment 2: Digital Marketing using Real-World Data

![Graph showing the relationship between the number of trajectories and lower bounds for different methods: Thm 1, MPeB, AM, and Behavior. The graph indicates that as the number of trajectories increases, the lower bound decreases for Thm 1, increases for MPeB, and shows a mixed trend for AM and Behavior.]
Experiment 2: Digital Marketing using Real-World Data

Black - 2 million off-policy trajectories
Blue - 5 million off-policy trajectories
Red - 5 million off-policy trajectories + 1 million on-policy trajectories