On the Design of Estimators for Bandit Off-Policy Evaluation

Nikos Vlassis* Aurelien Bibaut‡ Maria Dimakopoulou* Tony Jebara*

*NETFLIX ‡UC Berkeley

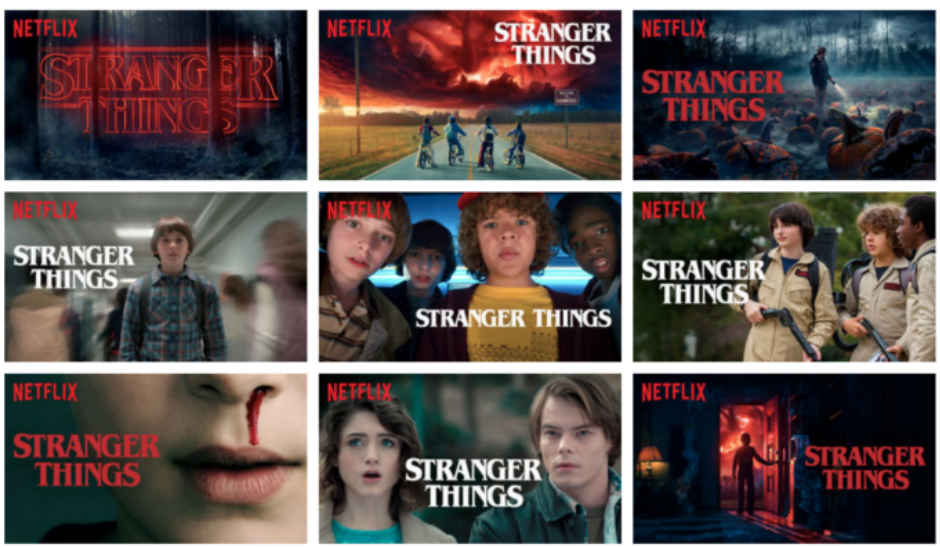
ICML 2019

Off-policy evaluation

The problem of estimating the value of a target policy, using data collected under a different logging policy.

[*Robins & Rotnitzky, 1995; Hahn, 1998; Hirano et al., 2003;Dudik et al., 2014; Li et al., 2015; Thomas & Brunskill, 2016; Swaminathan et al., 2017; Wang et al., 2017; Athey & Wager, 2017; Kallus & Zhou, 2018; Farajtabar et al., 2018; Joachims et al., 2018.*]

Very relevant for applications, e.g., artwork optimization at Netflix.



Bernoulli k-armed bandits

K Bernoulli arms, with (unknown) parameters P_a, for $a = 1, \ldots, K$

The logged data consist of n i.i.d. pairs (a_i, r_i) generated as follows:

1. K binary rewards are drawn as $r_a \sim P_a$, for $a = 1, \ldots, K$
2. Actions are drawn as $a_i \sim \mu$ where μ is the **logging** policy
3. We observe the rewards $r_i = r_{a_i}$

Using the logged data, we want to estimate the value of a **target** policy π:

$$v = \sum_a \pi_a P_a$$

Two popular estimators

The IPS estimator:

$$\hat{v}_{\text{IPS}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi_i}{\mu_i} r_i = \sum_a \pi_a \frac{n_a^+}{n\mu_a}$$

The REG estimator:

$$\hat{v}_{\text{REG}} = \sum_a \pi_a \frac{n_a^+}{n_a} \mathbb{I}[n_a > 0]$$

where $n_a = \sum \text{plays}(a)$, and $n_a^+ = \sum \text{rewards}(a)$.

This work: Design new instance-dependent estimators

Define parametrized estimator X via some function $f : (\mathbb{N}, \mathbb{N}, \mathbb{R}) \mapsto \mathbb{R}$

$$X = \sum_a f(n_a, n_a^+, \theta_a)$$

When f is **polynomial** in n_a, n_a^+, the population risk (MSE) of X is **closed-form**.

Proof: Expand the MSE as a sum of expectations of functions of n_a and n_a^+. Then use the following result for multinomial counts (Mosimann, 1962):

$$\mathbb{E}[(n_a)_m] = (n)_m \, \mu_a^m,$$

where $(x)_m = x(x-1)\cdots(x-m+1)$ is the m'th order **falling factorial** of x, and the fact that any polynomial can be written as a linear combination of falling factorials (e.g., Newton series).

Example: Parametrized control variates

Function f is linear in n_a^+ and **bilinear** in θ_a and n_a:

$$X = \hat{v}_{\mathsf{IPS}} + \frac{1}{n} \sum_a \theta_a n_a$$

Risk of X is quadratic in θ:

$$n\,\mathsf{MSE}[X] = n\Big(\sum_a \mu_a \theta_a\Big)^2 - \Big(v + \sum_a \mu_a \theta_a\Big)^2 + \sum_a \mu_a \theta_a^2 + 2\sum_a P_a \pi_a \theta_a + \sum_a \frac{\pi_a^2}{\mu_a} P_a$$

How to deal with terms involving P_a?

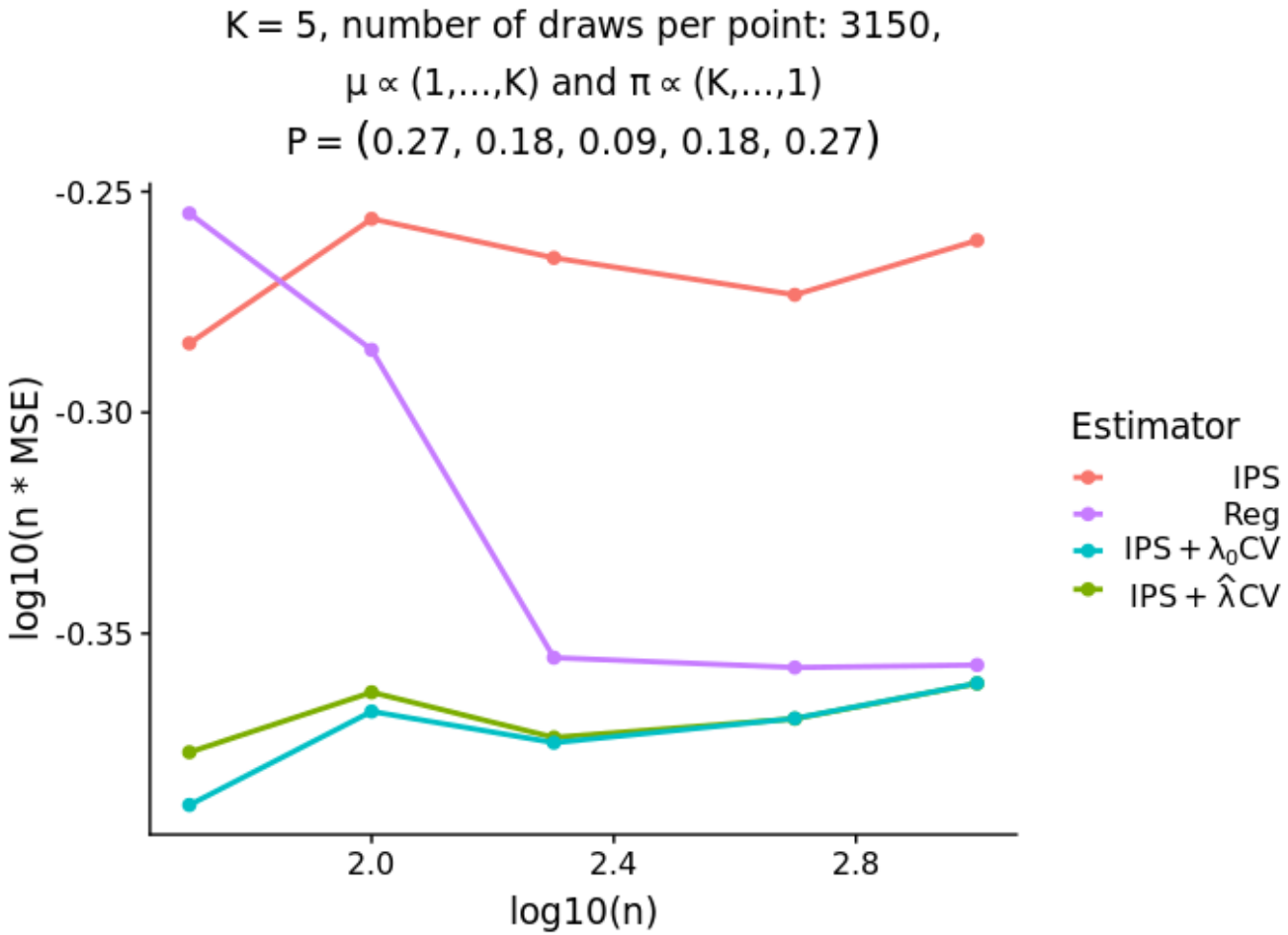
- Eliminate them (via known bounds, minimax solution, etc.). Several interesting research questions here.
- Approximate them using the logged data. Examples:

$$\sum_a P_a \pi_a \theta_a \approx \sum_a \frac{n_a^+}{n\mu_a} \pi_a \theta_a$$

$$\sum_a P_a \pi_a \theta_a \approx \sum_a \frac{n_a^+}{n_a} \mathbb{I}[n_a > 0] \pi_a \theta_a$$

[*Thomas & Brunskill, 2016; Swaminathan et al., 2017; Farajtabar et al., 2018.*]

Experiments: K-armed bandits



Experiments: Contextual bandits

Multiclass classification with bandit feedback (UCI datasets):

Dataset	ecoli	glass	satimage	vehicle	yeast
Classes (K)	8	6	6	4	10
Sample size (n)	336	214	6435	846	1484
DM	0.4837	0.3170	0.3259	0.4090	0.1914
IPS	0.3074	0.3092	0.0724	0.1901	0.1111
DR	0.2136	0.2497	0.0402	0.1298	0.0840
MRDR	0.1673	0.3185	**0.0302**	0.1194	0.0824
DR IPS CV	0.2099	**0.2271**	0.0400	0.1266	0.0827
MRDR IPS CV	**0.1665**	0.3103	**0.0302**	**0.1182**	**0.0818**

$$\text{DR IPS CV} = \text{DR} - \kappa_{\text{DR}} \cdot \text{IPS}\ , \qquad \hat{\kappa}_{\text{DR}} = \frac{\widehat{\text{Cov}}(\text{DR},\text{IPS})}{\widehat{\mathbb{E}}[\text{IPS}]^2 + \widehat{\text{Var}}(\text{IPS})}$$

Takehomes

There exist estimators that improve REG for k-armed bandits in the finite-sample regime.

We can improve DR for contextual bandits by operating outside the manifold of DR estimators.

Key open question: Are there instance-dependent control variates that guarantee risk improvement uniformly over the values of P_a?