Optimal and Adaptive Off-Policy Evaluation in Contextual Bandits

Yu-Xiang Wang
Joint work with Alekh Agarwal, Miro Dudik
Off-Policy Evaluation: Answering the “what-if” question

- Targeted advertisement
  - A “policy” decides which ad to show based on “context”
  - Then the user may click or not click
  - The click-through rate measures how good the policy is

- What if I ran a different policy instead?
  - a.k.a., Counterfactual reasoning
Many applications

- For safe policy deployment
- For policy optimization
Contextual bandits

- Contexts:
  - $x_1, \ldots, x_n \sim \lambda$ drawn iid, possibly infinite domain
- Actions:
  - $a_i \sim \mu(a|x_i)$ Taken by a randomized “Logging” policy
- Reward:
  - $r_i \sim D(r|x_i, a_i)$ Revealed only for the action taken
- Value:
  - $\mu = \mathbb{E}_{x \sim \lambda} \mathbb{E}_{a \sim \mu(\cdot|x)} \mathbb{E}_D[r|x, a]$
  - We collect data $(x_i, a_i, r_i)_{i=1}^n$ by the above processes.

- What if we use a different policy $\pi$ (the “Target” policy)?
  - How do we estimate its value?
Importance sampling/Inverse propensity scoring

(Horvitz & Thompson, 1952)

\[
\hat{\nu}_{\text{IPS}}^\pi = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\mu(a_i|x_i)} r_i =: \rho_i
\]

Pros:

• No assumption on rewards
• Unbiased
• Computationally efficient

Cons:

• High variance when the weight is large
Model-based approach

• Fit a regression model of the reward

\[ \hat{r}(x, a) \approx \mathbb{E}(r|x, a) \]

using the data

• Then for any target policy

\[ \hat{u}_{\pi}^{\text{DM}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in A} \hat{r}(x_i, a) \pi(a|x_i) \]

Pros:

• Low-variance.
• Can evaluate on unseen contexts

Cons:

• Often high bias
• The model can be wrong/hard to learn
Variants and combinations

• Modifying importance weights:
  • Trimmed IPS  (Bottou et. al. 2013)
  • Truncated/Reweighted IPS  (Bembom and van der Laan, 2008)

• Doubly Robust estimators:
  • A systematic way of incorporating DM into IPS
  • Originated in statistics (see e.g., Robins and Rotnitzky, 1995; Bang and Robins, 2005)
  • Used for off-policy evaluation  (Dudík et al., 2014)
Many estimators are proposed. Are they optimal? How good is good enough?

In this work, we formally address these problems.

1. Minimax lower bound: IPS is optimal in the general case.

2. A new estimator --- SWITCH --- that can be even better than IPS in some cases.
What do we mean by “optimal”?

- Minimax theory
  - Find an estimator that works well for ALL problem within a class of problems.
  - An estimator $\hat{v}: (\mathcal{X} \times \mathcal{A} \times \mathbb{R})^n \to \mathbb{R}$
  - Minimax risk / rate:
    $$\inf_{\hat{v}} \sup_{\text{a class of problems}} \mathbb{E}(\hat{v}(\text{Data}) - v^\pi)^2$$
    Taken over data $\sim \mu$
  - Fix context distribution and policies $\left(\lambda, \mu, \pi\right)$
  - A class of problems = a class of reward distributions.
What do we mean by “optimal”? 

• The class of problems: (generalizing Li et. al. 2015)

\[ \mathcal{R}(\sigma, R_{\text{max}}) := \left\{ D(r|x,a) : 0 \leq \mathbb{E}_D[r|x,a] \leq R_{\text{max}}(x,a) \text{ and } \varnothing \right\} \]

\[ \text{Var}_D[r|x,a] \leq \sigma^2(x,a) \text{ for all } x,a \}

• The minimax risk

\[ \inf_{\hat{v}} \sup_{D(r|a,x) \in \mathcal{R}(\sigma^2,R_{\text{max}})} \mathbb{E}(\hat{v} - v^\pi)^2 \]
Lower bounding the minimax risk

- **Our main theorem:** under mild conditions

\[
\inf_{\hat{v}} \sup_{D(r|a,x)\in\mathcal{R}(\sigma^2, R_{\text{max}})} \mathbb{E}(\hat{v} - v^\pi)^2 = \Omega \left[ \frac{1}{n} \left( \mathbb{E}_\mu [\rho^2 \sigma^2] + \mathbb{E}_\mu [\rho^2 R_{\text{max}}^2] (1 - \tilde{O}(n^{\lambda_0})) \right) \right]
\]

- Randomness in reward
- Randomness due to context distribution
- Max prob. of a single \( x \)

- Subsumes lower bound for multi-arm bandit.

This implies that **IPS is optimal!**

- The high variance is required.
  - In contextual bandits with *large context spaces and non-degenerate context distribution*.
  - Model-free approach is fundamentally limited.

- Different from multi-arm bandit
  - Li et. al. (2015) showed that in k-arm bandit, IPS is strictly suboptimal.
The pursuit of **adaptive estimators**

- **Minimaxity**: perform optimally on **hard problems**.
- **Adaptivity**: perform better on **easier problems**.
Suppose we are given an oracle

\[ \hat{r}(x, a) \]

• Could be very good, or completely off.
• How to **make the best use** of the predictions?
Why not just use **doubly robust**?

- Originated in statistics (see e.g.: Robins and Rotnitzky, 1995; Bang and Robins, 2005)
- Proposed for off-policy evaluation previously:
  

- We show that: DR can be as bad as IPS
- Does not adapt even with **perfect oracle**:

\[
\hat{r}(x, a) = \mathbb{E}(r \mid x, a)
\]

\[
\text{MSE}(\hat{v}_{DR}) \leq \frac{1}{n} \left( \mathbb{E}_\mu(\rho^2 \sigma^2) + \mathbb{E}_\pi(R_{\text{max}}^2) \right)
\]

DR can suffer from high variance just like IPS!
SWITCH estimator

• Recall that IPS is bad because:

\[
\hat{v}_\text{IPS}^\pi = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\mu(a_i|x_i)} r_i
\]

• SWITCH estimator:

For each \( i = 1, \ldots, n \), for each action \( a \in A \):

if \( \frac{\pi(a|x_i)}{\mu(a|x_i)} \leq \tau \):
    Use IPS (or DR).
else:
    Use the oracle estimator.

The approach is related to MAGIC estimator (Thomas & Brunskill, 2016), but with important difference.
Error bounds for SWITCH

\[
\text{MSE}(\hat{\nu}_{\text{SWITCH}}) \leq \\
\frac{2}{n} \mathbb{E}_\mu \left[ (\sigma^2 + R_{\text{max}}^2) \rho^2 \mathbf{1}(\rho \leq \tau) \right] \\
+ \frac{2}{n} \mathbb{E}_\pi [R_{\text{max}}^2 \mathbf{1}(\rho > \tau)] \\
+ \frac{2}{n} \mathbb{E}_\pi [\epsilon | \rho > \tau] \mathbb{P}_\pi (\rho > \tau)^2
\]

1) Variance from IPS (reduced truncation)

2) Variance due to sampling x. Required even with perfect oracle

3) Bias from the oracle.
Error bounds for SWITCH

- For appropriately tuned "threshold" parameter, SWITCH is
  - Independent to $p$ when oracle is perfect.
  - Minimax when oracle is horrible.
  - Robust to large importance weight.

- Data dependent tuning of parameter? Check out our paper!
- Different from MAGIC (Thomas and Brunskill, 2016)
Experiment setup

• 10 UCI Classification data sets converted to bandits.
  • Action is to predict labels.
  • Reward is \{0,1\}, depending on whether the action is correct.

• Follow standard setup in
  • (Beygelzimer & Langford, 2009)
  • (Gretton et. al. 2008)
  • (Dudik et. al. 2011)
CDF of relative MSE over 10 UCI multiclass classification data sets.
With additional label noise

![Graph showing relative error w.r.t. IPS for different methods: IPS, DM, DR, SWITCH-DR, oracle-SWITCH-DR, SWITCH-DR-magic.](image)
Conclusion

• IPS is optimal.
  • Need to go beyond the model-free approach.

• DR is unsatisfactory.

• We propose an new estimator: SWITCH
  • that has good theoretical properties.
  • performs quite well in practice.
Thank you! Any questions?
Connections and future work

• Extension to reinforcement learning
  • Lower bound directly applies in some sense.
  • SWITCH-DR for reinforcement learning?

• Lower bound directly applies to “mean effect” estimation.
  • Basically it corresponds to a different “target policy”.
The conditions for the main Theorem

\[ \mathbb{E}_\mu[(\rho \sigma)^{2+\varepsilon}] \leq \infty \]
\[ \mathbb{E}_\mu[(\rho R_{\text{max}})^{2+\varepsilon}] \leq \infty \]
\[ \mathbb{E}_\mu[\sigma^2/R_{\text{max}}^2] < \infty \]

• Moment conditions:

• If \( n \) is sufficiently large

\[
\inf_{\hat{v}} \sup_{D(r|a,x) \in \mathcal{R}(\sigma^2, R_{\text{max}})} \mathbb{E}(\hat{v} - v^\pi)^2
\]
\[
= \Omega \left[ \frac{1}{n} \left( \mathbb{E}_\mu[\rho^2 \sigma^2] + \mathbb{E}_\mu[\rho^2 R_{\text{max}}^2] (1 - 110\lambda_0 \log(4/\lambda_0)) \right) \right]
\]
Automatic parameter tuning

- Conservative approximate MSE minimizing.

\[ \hat{\tau} = \arg\min_{\tau} \hat{\text{Var}}_{\tau} + \hat{\text{Bias}}_{\tau}^2. \]

- Details:

\[
Y_i(\tau) := r_i \cdot \rho_i \cdot 1(\rho_i \leq \tau) + \sum_{a \in A} \hat{r}(x_i, a) \cdot \pi(a|x_i) \cdot 1(\rho(x_i, a) > \tau) \quad \text{and} \quad \bar{Y}(\tau) = \frac{1}{n} \sum_{i=1}^{n} Y_i(\tau),
\]

\[
\text{Var}(\hat{\psi}_{\text{SWITCH} - \tau}) = \frac{1}{n} \text{Var}(\hat{\psi}_{\text{SWITCH} - \tau}(x_1)) \approx \frac{1}{n^2} \sum_{i=1}^{n} (Y_i(\tau) - \bar{Y}(\tau))^2 =: \hat{\text{Var}}_{\tau},
\]

\[
\text{Bias}^2(\hat{\psi}_{\text{SWITCH}}) \leq \mathbb{E}_\mu[\rho^2 | \rho > \tau] \pi(\rho > \tau)^2 \leq \mathbb{E}_\mu[\rho R_{\text{max}}^2 | \rho > \tau] \pi(\rho > \tau)^2
\]

\[
\approx \left[ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\pi}(R_{\text{max}}^2 | \rho > \tau, x_i) \right] \left[ \frac{1}{n} \sum_{i=1}^{n} \pi(\rho > \tau | x_i) \right]^2 =: \hat{\text{Bias}}_{\tau}^2.
\]
Experiment setup

• 10 UCI Classification data sets converted to bandits.
  • Action is to predict labels.
  • Reward is \{0,1\}, depending on whether the action is correct.
  • Target policy is prediction of logistic regression.
  • Logging policy obtained by the label probability of a logistic regression learned from covariate shifted data.

• We sample data of size \( n = [100, 200, 500, 1000, \ldots] \), from discrete distribution of of length \( N \).