Improved Estimator Selection for Off-Policy Evaluation

George Tucker [1] **Jonathan Lee** [2 *]

Abstract

Off-policy policy evaluation is a fundamental problem in reinforcement learning. As a result, many estimators with different tradeoffs have been developed; however, selecting the best estimator is challenging with limited data and without additional interactive data collection. Recently, Su et al. (2020b) developed a data-dependent selection procedure that competes with the oracle selection up to a constant and demonstrate its practicality. We refine the analysis to remove an extraneous assumption and improve the procedure. The improved procedure results in a tighter oracle bound and stronger empirical results on a contextual bandit task.

1. Introduction

Off-policy policy evaluation is a fundamental problem in reinforcement learning. Moreover, in applications of reinforcement learning, the ability to evaluate potential policies without deploying them is paramount to reducing risk. As a result, developing off-policy estimators has been a subject of intense research activity (e.g., (Horvitz & Thompson, 1952; Precup, 2000; Dudík et al., 2014; Jiang & Li, 2016; Thomas & Brunskill, 2016; Wang et al., 2017; Farajtabar et al., 2018; Liu et al., 2018; Kallus & Zhou, 2018; Voloshin et al., 2019)).

However, selecting the best estimator for a particular application is challenging. There is no straightforward analogue to cross-validation in supervised learning. Recently, Su et al. (2020b) introduced SLOPE, a generic, data-driven selection procedure, which is provable competitive with the oracle selection (up to a constant) and they demonstrate that it empirically outperforms other selection approaches based on MSE surrogates (Thomas & Brunskill, 2016; Wang et al., 2017; Su et al., 2020a).

*Work done while at Google. [1]Google Research, Brain Team [2]Stanford University. Correspondence to: George Tucker <gjt@google.com>.

Workshop on Reinforcement Learning Theory at the 38^{th} International Conference on Machine Learning, 2021.

By refining the analysis underlying their approach, we can remove an extraneous assumption in their procedure and tighten the oracle bound on performance. Furthermore, we show that the improved procedure outperforms SLOPE on the task introduced in Su et al. (2020b).

2. Background

We work in the same abstract setup as Su et al. (2020b). Briefly, we would like to estimate a real-valued statistic $\theta^* := \theta(D)$ defined over an unknown data distribution D. Given a finite set of estimators $\{\theta_i(\cdot)\}$ defined over empirical data distributions and a (uniformly-weighted) empirical data distribution $\hat{D}$ formed from an i.i.d. sample of data $\{x_i\}$ from D, our goal is to choose the estimator that minimizes the absolute error to the statistic of interest θ^*. In other words, we would like a procedure that that selects an estimator index $\hat{i}$ such that

$$|\hat{\theta}_{\hat{i}} - \theta^*| \leq \text{CONST} \times \min_i |\hat{\theta}_i - \theta^*|,$$

where $\hat{\theta}_i := \theta_i(\hat{D})$ and CONST is a universal constant. To start, we can bound the distance as

$$|\hat{\theta}_i - \theta^*| \leq |\bar{\theta}_i - \theta^*| + |\hat{\theta}_i - \bar{\theta}_i| := \text{BIAS}(i) + \text{DEV}(i),$$

where $\bar{\theta}_i = \mathbb{E}_{\hat{D}}\left[\hat{\theta}_i\right]$. $\text{DEV}(i)$ characterizes statistical fluctuations, and we can generally bound it with high confidence, whereas $\text{BIAS}(i)$ is typically unknown.

2.1. SLOPE

Su et al. (2020b) introduce the SLOPE procedure based on Lepski's principle (Lepskii, 1991; 1992; 1993; Lepski & Spokoiny, 1997; Mathé, 2006). The procedure assumes a bound on bias $B(i) \geq \text{BIAS}(i)$[1] and access to a high confidence bound on $\text{DEV}(i)$ (i.e., with probability at least $1-\delta$, $\text{DEV}(i) \leq \text{CNF}(i)$ for all i). The procedure also requires monotonicity assumptions:

1. $B(i) \leq B(i+1)$
2. $\exists \kappa > 0$ such that $\kappa \,\text{CNF}(i) \leq \text{CNF}(i+1) \leq \text{CNF}(i)$

Then defining $I(i) = [\hat{\theta}_i - 2\,\text{CNF}(i), \hat{\theta}_i + 2\,\text{CNF}(i)]$ and

$$\hat{i} = \max\{i : \cap_{j=1}^{i} I(j) \neq \emptyset\},$$

[1]$\text{B}(i)$ is not used in the procedure but appears in the bound.

with probability at least $1-\delta$,

$$|\hat{\theta}_{\hat{i}} - \theta^*| \leq 6\left(1+\kappa^{-1}\right) \min_i \left(\mathrm{B}(i) + \mathrm{CNF}(i)\right).$$

So the SLOPE procedure provides an oracle inequality on the competitiveness of the procedure. It is natural for many families of estimators to exhibit a similar trade-off of complexity, captured by the confidence bounds, with bias.

3. Methods

We improve the procedure by refining the analysis.

Theorem 1. *Given $\delta > 0$, high confidence bounds $\mathrm{CNF}(i)$ on the deviations (i.e., with probability at least $1-\delta$, $\mathrm{DEV}(i) \leq \mathrm{CNF}(i)$ for all i), and that we have ordered*[2] *the estimators such that $\mathrm{CNF}(i) \geq \mathrm{CNF}(i+1)$. Selecting our estimator as*

$$\hat{i} = \max\{i : |\hat{\theta}_i - \hat{\theta}_j| \leq \mathrm{CNF}(i) + (\sqrt{6}-1)\,\mathrm{CNF}(j), j < i\},$$

ensures that with probability at least $1-\delta$,

$$|\hat{\theta}_{\hat{i}} - \theta^*| \leq \left(\sqrt{6}+3\right) \min_i \left(\max_{j \leq i} \mathrm{BIAS}(j) + \mathrm{CNF}(i)\right).$$

In the common case that the estimator family has increasing bias and tighter bounds on deviation with increasing i (a family of estimators with a bias/variance trade-off e.g., importance weight clipping), then the bound simplifies to

$$|\hat{\theta}_{\hat{i}} - \theta^*| \leq \left(\sqrt{6}+3\right) \min_i \left(\mathrm{BIAS}(i) + \mathrm{CNF}(i)\right).$$

In other words, we are able to compete with the oracle estimator (up to a problem-independent constant). Compared to Su et al. (2020b), we do not assume $\mathrm{CNF}(i)$ decreases slowly and thus our bound does not depend on the problem-dependent parameter κ that controls this decrease. As $\frac{1}{\kappa} \geq 1$, we have $\sqrt{6}+3 < 6 < 6(1+1) \leq 6(1+\kappa^{-1})$ for the constant on the oracle bound.

Proof. We will prove a generalized bound. Let $\mathrm{B}(i) \geq \mathrm{BIAS}(i)$ be a bound on bias satisfying $\mathrm{B}(i) \leq \mathrm{B}(i+1)$. Then, define $\hat{i} = \max\{i : |\hat{\theta}_i - \hat{\theta}_j| \leq a\,\mathrm{CNF}(i) + b\,\mathrm{CNF}(j), j < i\}$, where a and b are constants that we will optimize.

Lepski's principle leverages the fact that comparing pairs of estimators simplifies the problem. In our case, the difficulty arises from the unknown bias of our estimators, however, we do know that for a pair of estimators

$$\begin{aligned} B(j) - B(i) &\leq |\bar{\theta}_i - \bar{\theta}_j| \\ &\leq |\hat{\theta}_i - \hat{\theta}_j| + |\hat{\theta}_i - \hat{\theta}_j - \left(\bar{\theta}_i - \bar{\theta}_j\right)| \\ &\leq |\hat{\theta}_i - \hat{\theta}_j| + \mathrm{CNF}(i) + \mathrm{CNF}(j), \quad (1) \end{aligned}$$

which allows us to control the increase in bias in terms of *known* quantities.

Let $i^* = \min_i \mathrm{B}(i) + \mathrm{CNF}(i)$. When $i^* = \hat{i}$, we are done, so we consider the other two cases: 1) $\hat{i} < i^*$ and 2) $\hat{i} > i^*$.

1. We know that $B(\hat{i}) \leq B(i^*)$, so we need to ensure that CNF has not decreased too much. By definition of $\hat{i}$, there exists $j \leq \hat{i}$ such that

$$\begin{aligned} &b\,\mathrm{CNF}(j) + a\,\mathrm{CNF}(\hat{i}+1) < |\hat{\theta}_j - \hat{\theta}_{\hat{i}+1}| \\ &\quad \leq B(j) + B(\hat{i}+1) + \mathrm{CNF}(j) + \mathrm{CNF}(\hat{i}+1). \end{aligned}$$

   Implying that

$$\begin{aligned} (b-1)\,\mathrm{CNF}(\hat{i}) &\leq (b-1)\,\mathrm{CNF}(j) \\ &< B(j) + B(\hat{i}+1) + (1-a)\,\mathrm{CNF}(\hat{i}+1)) \\ &\leq 2B(i^*) + (1-a)\,\mathrm{CNF}(\hat{i}+1)) \\ &\leq 2B(i^*) \end{aligned}$$

   as long as $a \geq 1$. So, with $b > 1$, we conclude that

$$\begin{aligned} B(\hat{i}) + \mathrm{CNF}(\hat{i}) &< \left(\frac{2}{b-1}+1\right) B(i^*) \\ &\leq \left(\frac{2}{b-1}+1\right)\left(B(i^*) + \mathrm{CNF}(i^*)\right) \end{aligned}$$

2. In the second case, $\mathrm{CNF}(\hat{i}) \leq \mathrm{CNF}(i^*)$, so we need to show that the bias is not too much larger. By definition of $\hat{i}$ and Eq. 1,

$$\begin{aligned} B(\hat{i}) - B(i^*) &\leq |\hat{\theta}_{\hat{i}} - \hat{\theta}_{i^*}| + \mathrm{CNF}(\hat{i}) + \mathrm{CNF}(i^*) \\ &\leq (a+1)\,\mathrm{CNF}(\hat{i}) + (b+1)\,\mathrm{CNF}(i^*) \\ &\leq (a+b+2)\,\mathrm{CNF}(i^*) \end{aligned}$$

   implying that

$$\begin{aligned} B(\hat{i}) + \mathrm{CNF}(\hat{i}) &\leq B(i^*) + (a+b+3)\,\mathrm{CNF}(i^*) \\ &\leq (a+b+3)(B(i^*) + \mathrm{CNF}(i^*)). \end{aligned}$$

Together, we conclude that

$$\begin{aligned} &B(\hat{i}) + \mathrm{CNF}(\hat{i}) \leq \\ &\quad \max\left(\left(\frac{2}{b-1}+1\right), (a+b+3)\right)\left(B(i^*) + \mathrm{CNF}(i^*)\right) \end{aligned}$$

Minimizing over a and b results in $a=1$ and $b=\sqrt{6}-1$. Then choosing $B(i) = \max_{j \leq i} \mathrm{BIAS}(j)$ finishes the proof. □

Previous works have focused on MSE as the evaluation metric, so as a corollary, we have

[2] We can increase $\mathrm{CNF}(i)$ to accommodate any ordering at the expense of a worse bound.

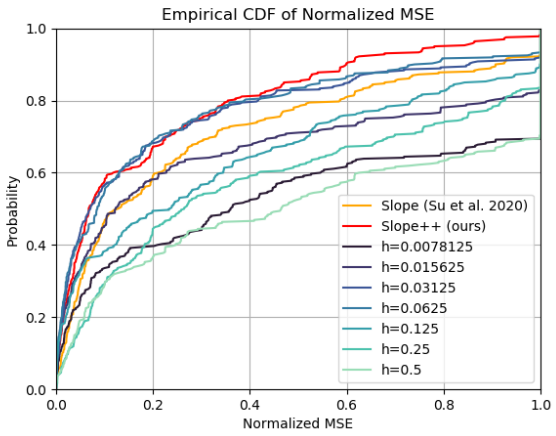

Figure 1. Experimental results for bandwidth selection of OPE estimators for contextual bandits with continuous actions. The plot shows the empirical CDF of normalized MSE (normalized by the worst MSE for that condition) across all conditions.

Corollary 1.1. *With the same assumptions as Thm 1 and additionally that* $\theta^*, \hat{\theta}_i \in [0, R]$ *a.s. for all* i *and that* CNF *is deterministic. Then, for any* $\delta > 0$*,*

$$\mathbb{E}\left[\left(\hat{\theta}_{\hat{i}} - \theta^*\right)^2\right] \leq C \min_i \left(\max_{j \leq i} \mathrm{BIAS}(j)^2 + \mathrm{CNF}(i;\delta)^2\right) + R^2\delta,$$

where C *is a universal constant.*

The proof follows immediately from Thm 1 and the proof of (Corollary 4; Su et al., 2020b).

4. Experiments

Following Su et al. (2020b), we empirically evaluate the procedure on bandwidth selection in a synthetic environment for continuous action contextual bandits. We follow the same experimental protocol and modify the code provided with Su et al. (2020b) at `https://github.com/VowpalWabbit/slope-experiments` to implement our procedure.

Briefly, the experiment evaluates the procedure on environments with varying target and logging policies (NN, tree), "softening" approaches for randomization following Farajtabar et al. (2018) (friendly, adversarial), Lipshitz constant (0.1, 1, 10), and samples (10, 100, 1000)[3].

We use 7 different choices of geometrically spaced bandwidths $\{2^{-i} : i \in [7]\}$. We evaluate the performance of each of these fixed choices, SLOPE, and the improved procedure (labeled SLOPE++ in the plot). As in Su et al. (2020b), we use the empirical standard deviation to construct the high confidence bound, which is valid asymptotically and usually yields better practical performance than a concentration inequality.

To measure performance, we estimate the ground truth statistic with $100k$ samples from the target policy. To aggregate results across conditions, we normalize by the worst MSE in each condition, then compute the empirical CDF of the normalized MSE (Figure 1). Increasingly performant methods lie in the top-left quadrant. We see that SLOPE++ is the top performer compared with the fixed bandwidths and SLOPE.

[3]From the provided configuration file `https://github.com/VowpalWabbit/slope-experiments/blob/master/scripts/ak_commands.py`

References

Dudík, M., Erhan, D., Langford, J., Li, L., et al. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456. PMLR, 2018.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Kallus, N. and Zhou, A. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pp. 1243–1251. PMLR, 2018.

Lepski, O. V. and Spokoiny, V. G. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, pp. 2512–2546, 1997.

Lepskii, O. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35 (3):454–466, 1991.

Lepskii, O. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

Lepskii, O. Asymptotically minimax adaptive estimation. ii. schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications*, 37(3):433–448, 1993.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *arXiv preprint arXiv:1810.12429*, 2018.

Mathé, P. The lepskii principle revisited. *Inverse problems*, 22(3):L11, 2006.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Su, Y., Dimakopoulou, M., Krishnamurthy, A., and Dudík, M. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pp. 9167–9176. PMLR, 2020a.

Su, Y., Srinath, P., and Krishnamurthy, A. Adaptive estimator selection for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9196–9205. PMLR, 2020b.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR, 2016.

Voloshin, C., Le, H. M., Jiang, N., and Yue, Y. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

Wang, Y.-X., Agarwal, A., and Dudık, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pp. 3589–3597. PMLR, 2017.