Adaptive Estimator Selection for Off-Policy Evaluation

Yi Su [1] **Pavithra Srinath** [2] **Akshay Krishnamurthy** [2]

Abstract

We develop a generic data-driven method for estimator selection in off-policy policy evaluation settings. We establish a strong performance guarantee for the method, showing that it is competitive with the oracle estimator, up to a constant factor. Via in-depth case studies in contextual bandits and reinforcement learning, we demonstrate the generality and applicability of the method. We also perform comprehensive experiments, demonstrating the empirical efficacy of our approach and comparing with related approaches. In both case studies, our method compares favorably with existing methods.

1. Introduction

In practical scenarios where safety, reliability, or performance is a concern, it is typically infeasible to directly deploy a reinforcement learning (RL) algorithm, as it may compromise these desiderata. This motivates research on *off-policy evaluation*, where we use data collected by a (presumably safe) logging policy to estimate the performance of a given target policy, without ever deploying it. These methods help determine if a policy is suitable for deployment at minimal cost and, in addition, serve as the statistical foundations of sample-efficient policy optimization algorithms. In light of the fundamental role off-policy evaluation plays in RL, it has been the subject of intense research over the last several decades (Horvitz & Thompson, 1952; Dudík et al., 2014; Swaminathan et al., 2017; Kallus & Zhou, 2018; Sutton, 1988; Bradtke & Barto, 1996; Precup et al., 2000; Jiang & Li, 2016; Thomas & Brunskill, 2016; Farajtabar et al., 2018; Liu et al., 2018; Voloshin et al., 2019).

As many off-policy estimators have been developed, practitioners face a new challenge of choosing the best estimator for their application. This selection problem is critical to high quality estimation as has been demonstrated in recent empirical studies (Voloshin et al., 2019). However, data-driven estimator selection in these settings is fundamentally different from hyperparameter optimization or model selection for supervised learning. In particular, cross validation or bound minimization approaches fail because there is no unbiased and low variance approach to compare estimators. As such, the current best practice for estimator selection is to leverage domain expertise or follow guidelines from the literature (Voloshin et al., 2019).

Domain knowledge can suggest a particular form of estimator, but a second selection problem arises, as many estimators themselves have hyperparameters that must be tuned. In most cases, these hyperparameters modulate a bias-variance tradeoff, where at one extreme the estimator is unbiased but has high variance, and at the other extreme the estimator has low variance but potentially high bias. Hyperparameter selection is critical to performance, but high-level prescriptive guidelines are less informative for these low-level selection problems. We seek a data-driven approach.

In this paper, we study the estimator-selection question for off-policy evaluation. We provide a general technique, that we call SLOPE, that applies to a broad family of estimators, across several distinct problem settings. On the theoretical side, we prove that the selection procedure is competitive with oracle tuning, establishing an *oracle inequality*. To demonstrate the generality of our approach, we study two applications in detail: (1) bandwidth selection in contextual bandits with continuous actions, and (2) horizon selection for "partial importance weighting estimators" in RL. In both examples, we prove that our theoretical results apply, and we provide a comprehensive empirical evaluation. In the contextual bandits application, our selection procedure is competitive with the skyline oracle tuning (which is unimplementable in practice) and outperforms any fixed parameter in aggregate across experimental conditions. In the RL application, our approach substantially outperforms standard baselines including MAGIC (Thomas & Brunskill, 2016), the only comparable estimator selection method.

A representative experimental result for the RL setting is displayed in Figure 1. Here we consider two different domains from Voloshin et al. (2019) and compare our new estimator, SLOPE, with well-known baselines. Our method selects

[1]Cornell University, Ithaca, NY [2]Microsoft Research, New York, NY. Correspondence to: Akshay Krishnamurthy <akshaykr@microsoft.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

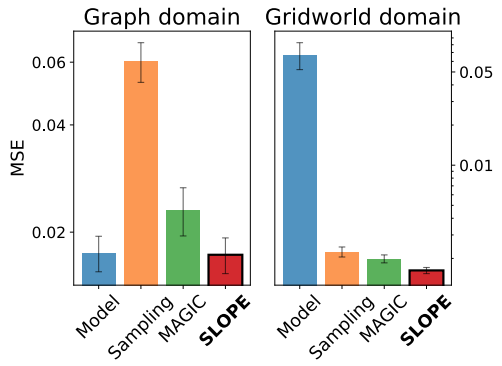

Figure 1. Representative experiments with SLOPE. SLOPE is consistently one of the best approaches, regardless of whether model-based or importance-sampling based estimators are better.

a false horizon η, uses an unbiased importance sampling approach up to horizon η, and then prematurely terminates the episode with a value estimate from a parametric estimator (in this case trained via Fitted Q iteration). Model selection focuses on choosing the false horizon η, which yields parametric and trajectory-wise importance sampling estimators as special cases ("Model" and "sampling" in the figure). Our experiments show that regardless of which of these approaches dominates, SLOPE is competitive with the best approach. Moreover, it outperforms MAGIC, the only other tuning procedure for this setting. Section 5 contains more details and experiments.

At a technical level, the fundamental challenge with estimator selection is that there is no unbiased and low-variance approach for comparing parameter choices. This precludes the use of cross validation and related approaches, as estimating the error of a method is itself an off-policy evaluation problem! Instead, adapting ideas from nonparametric statistics (Lepski & Spokoiny, 1997; Mathé, 2006), our selection procedure circumvents this error estimation problem by only using variance estimates, which are easy to obtain. At a high level, we use confidence bands for each estimator around their (biased) expectation to find one that approximately balances bias and variance. This balancing corresponds to the oracle choice, and so we obtain our performance guarantee.

Related work. As mentioned, off-policy evaluation is a vibrant research area with contributions from machine learning, econometrics, and statistics communities. Two settings of particular interest are contextual bandits and general RL. For the former, recent and classical references include (Horvitz & Thompson, 1952; Dudík et al., 2014; Hirano et al., 2003; Farajtabar et al., 2018; Su et al., 2020). For the latter, please refer to Voloshin et al. (2019).

Parameter tuning is quite important for many off-policy evaluation methods. Munos et al. (2016) observe that methods like RETRACE are fairly sensitive to the hyperparameter. Similarly conclusions can be drawn from the experiments of Su et al. (2019) in the contextual bandits context. Yet, when tuning is required, most works resort to heuristics. For example, in Kallus & Zhou (2018), a bandwidth hyperparameter is selected by performing an auxiliary nonparametric estimation task, while in Liu et al. (2018), it is selected as the median of the distances between points. In both cases, no theoretical guarantees are provided for such methods.

Indeed, despite the prevalence of hyperparameters in these methods, we are only aware of two methods for estimator selection: the MAGIC estimator (Thomas & Brunskill, 2016), and the bound minimization approach studied by Su et al. (2020) (see also Wang et al. (2017)). Both approaches use MSE surrogates for estimator selection, where MAGIC under-estimates the MSE and the latter uses an over-estimate. The guarantees for both methods (asymptotic consistency, competitive with unbiased approaches) are much weaker than our oracle inequality, and SLOPE substantially outperforms MAGIC in experiments.

Our approach is based on Lepski's principle for bandwidth selection in nonparametric statistics (Lepski, 1992; Lepskii, 1991; 1993; Lepski & Spokoiny, 1997). In this seminal work, Lepski studied nonparametric estimation problems and developed a data-dependent bandwidth selection procedure that achieves optimal adaptive guarantees, in settings where procedures like cross validation do not apply (e.g., estimating a regression function at a single given point). Since its introduction, Lepski's methodology has been applied to other statistics problems (Birgé, 2001; Mathé, 2006; Goldenshluger & Lepski, 2011; Kpotufe & Garg, 2013; Page & Grünewälder, 2018), but its use in machine learning has been limited. To our knowledge, Lepski's principle has not been used for off-policy evaluation, which is our focus.

2. Setup

We formulate the estimator selection problem generically, where there is an abstract space $\mathcal{Z}$ and a distribution $\mathcal{D}$ over $\mathcal{Z}$. We would like to estimate some parameter $\theta^\star := \theta(\mathcal{D}) \in \mathbb{R}$, where θ is some known real-valued functional, given access to $z_1, \dots, z_n \overset{iid}{\sim} \mathcal{D}$. Let $\hat{\mathcal{D}}$ denote the empirical measure, that is the uniform measure over the points $z_{1:n}$.

To estimate $\theta^\star$ we use a finite set of M estimators $\{\theta_i\}_{i=1}^M$, where each $\theta_i : \Delta(\mathcal{Z}) \to \mathbb{R}$. Given the dataset, we form the estimates $\hat{\theta}_i := \theta_i(\hat{\mathcal{D}})$. Ideally, we would choose the index that minimizes the absolute error with $\theta^\star$, that is $\operatorname{argmin}_{i\in[M]} \left|\hat{\theta}_i - \theta^\star\right|$. Of course this *oracle* index depends on the unknown parameter $\theta^\star$, so it cannot be computed from the data. Instead we seek a data-driven approach for selecting an index $\hat{i}$ that approximately minimizes the error.

To fix ideas, in the RL context, we may think of θ as the value of a *target policy* π_{T} and $z_{1:n}$ as n trajectories col-

lected by some *logging policy* π_{L}. The estimators $\{\theta_i\}_{i=1}^M$ may be partial importance weighting estimators (Thomas & Brunskill, 2016), that account for policy mismatch on trajectory prefixes of different length. These estimators modulate a bias variance tradeoff: importance weighting short prefixes will have high bias but low variance, while importance weighting the entire trajectory will be unbiased but have high variance. We will develop this example in detail in Section 5.

For performance guarantees, we decompose the absolute error into two terms: the bias and the deviation. For this decomposition, define $\bar{\theta}_i := \mathbb{E}[\theta_i(\hat{D})]$ where the expectation is over the random samples $z_{1:n}$. Then we have

$$\left|\hat{\theta}_i - \theta^\star\right| \le \left|\bar{\theta}_i - \theta^\star\right| + \left|\hat{\theta}_i - \bar{\theta}_i\right| =: \textsc{bias}(i) + \textsc{dev}(i).$$

As $\textsc{dev}(i)$ involves statistical fluctuations only, it is amenable to concentration arguments, so we will assume access to a high probability upper bound. Namely, our procedure uses a confidence function $\textsc{cnf}$ that satisfies $\textsc{dev}(i) \le \textsc{cnf}(i)$ for all $i \in [M]$ with high probability. On the other hand, estimating the bias is much more challenging, so we do not assume that the estimator has access to $\textsc{bias}(i)$ or any sharp upper bound. Our goal is to select an index $\hat{i}$ achieving an *oracle inequality* of the form

$$\left|\hat{\theta}_{\hat{i}} - \theta^\star\right| \le \textsc{const} \times \min_i \left\{\textsc{b}(i) + \textsc{cnf}(i)\right\}, \qquad (1)$$

that holds with high probability where $\textsc{const}$ is a universal constant and $\textsc{b}(i)$ is a sharp upper bound on $\textsc{bias}(i)$.[1] This guarantee certifies that the selected estimator is competitive with the error bound for the best estimator under consideration.

We remark that the above guarantee is qualitatively similar, but weaker than the ideal guarantee of competing with the actual error of the best estimator (as opposed to the error bound). In theory, this difference is negligible as the two guarantees typically yield the same statistical conclusions in terms of convergence rates. Empirically we will see that (1) does yield strong practical performance.

3. General Development

To obtain an oracle inequality of the form in (1), we require some benign assumptions. When we turn to the case studies, we will verify that these assumptions hold for our estimators.

Validity and Monotonicity. The first basic property on the bias and confidence functions is that they are valid in the sense that they actually upper bound the corresponding terms in the error decomposition.

[1] Some assumptions prevent us from setting $\textsc{b}(i) = \textsc{bias}(i)$.

Assumption 1 (Validity)**.** *We assume*

1. *(Bias Validity)* $\left|\bar{\theta}_i - \theta^\star\right| \le \textsc{b}(i)$ *for all* i.
2. *(Confidence Validity) With probability at least* $1-\delta$, $|\hat{\theta}_i - \bar{\theta}_i| \le \textsc{cnf}(i)$ *for all* i.

Typically $\textsc{cnf}$ can be constructed using straightforward concentration arguments such as Bernstein's inequality. Importantly, $\textsc{cnf}$ does not have to account for the bias, so the term $\textsc{dev}$ that we must control is centered. We also note that $\textsc{cnf}$ need not be deterministic, for example it can be derived from empirical Bernstein inequalities. We emphasize again that the estimator does not have access to $\textsc{b}(i)$.

We also require a monotonicity property on these functions.

Assumption 2 (Monotonicity)**.** *We assume that there exists a constant* $\kappa > 0$ *such that for all* $i \in [M-1]$

1. $\textsc{b}(i) \le \textsc{b}(i+1)$.
2. $\kappa \cdot \textsc{cnf}(i) \le \textsc{cnf}(i+1) \le \textsc{cnf}(i)$.

In words, the estimators should be ordered so that the bias is monotonically increasing and the confidence is decreasing but not too quickly, as parameterized by the constant κ. This structure is quite natural when estimators navigate a bias-variance tradeoff: when an estimator has low bias it typically also has high variance and vice versa. It is also straightforward to enforce a decay rate for $\textsc{cnf}$ by selecting the parameter set appropriately. We will see how to do this in our case studies.

The SLOPE procedure. SLOPE is an acronym for "Selection by Lepski's principle for Off-Policy Evaluation." As the name suggests, the approach is based on Lepski's principle (Lepski & Spokoiny, 1997) and is defined as follows. We first define intervals

$$I(i) := [\hat{\theta}_i - 2\textsc{cnf}(i), \hat{\theta}_i + 2\textsc{cnf}(i)],$$

and we then use the intersection of these intervals to select an index $\hat{i}$. Specifically, the index we select is

$$\hat{i} := \max\left\{i \in [M] : \cap_{j=1}^{i} I(j) \neq \emptyset\right\}.$$

In words, we select the largest index such that the intersection of all previous intervals is non-empty. See Figure 2 for an illustration.

The intuition is to adopt an optimistic view of the bias function $\textsc{b}(i)$. First observe that if $\textsc{b}(i) = 0$ then, by Assumption 1, we must have $\theta^\star \in I(i)$. Reasoning optimistically, it is possible that we have $\textsc{b}(i) = 0$ for all $i \le \hat{i}$, since by the definition of $\hat{i}$ there exists a choice of $\theta^\star$ that is consistent with all intervals. As $\textsc{cnf}(\hat{i})$ is the smallest among these, index $\hat{i}$ intuitively has lower error than all previous indices. On the other hand, it is not possible to have $\textsc{b}(\hat{i}+1) = 0$,

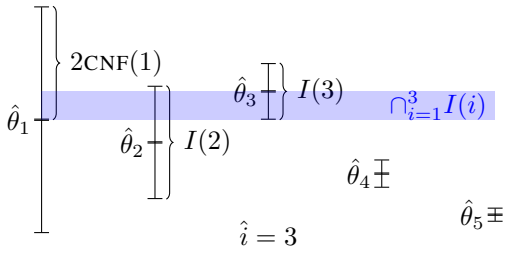

Figure 2. Illustration of SLOPE with $M = 5$. As $\cap_{i=1}^{3} I(i)$ is non-empty but $I(4)$ does not intersect with $I(3)$, we select $\hat{i} = 3$.

since there is no consistent choice for $\theta^\star$ and the bias is monotonically increasing. In fact, if $\theta^\star \in I_{\hat{i}}$, then we must actually have $\text{B}(\hat{i}+1) \geq \text{CNF}(\hat{i}+1)$, since the intervals have width $4\text{CNF}(\cdot)$. Finally, since $\text{CNF}(\cdot)$ does not shrink too quickly, all subsequent indices cannot be much better than $\hat{i}$, the index we select. Of course, we may not have $\theta^\star \in I_{\hat{i}}$, so this argument does not constitute a proof of correctness, which is deferred to Appendix A.

Theoretical analysis. We now state the main guarantee.

Theorem 3. *Under Assumption 1 and Assumption 2, we have that with probability at least* $1-\delta$*:*

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

The theorem verifies that the index $\hat{i}$ satisfies an oracle inequality as in (1), with $\text{CONST} = 6(1+\kappa^{-1})$. This is the best guarantee one could hope for, up to the constant factor and the caveat that we are competing with the error bound instead of the error, which we have already discussed. For off-policy evaluation, we are not aware of any other approaches that achieve any form of oracle inequality. The closest comparison is the bound minimization approach of Su et al. (2020), which is provably competitive only with unbiased indices (with $\text{B}(i) = 0$). However in finite sample, these indices could have high variance and consequently worse performance than some biased estimator. In this sense, the SLOPE guarantee is much stronger.

While our main result gives a high probability absolute error bound, it is common in the off-policy evaluation literature to instead provide bounds on the mean squared error. Via a simple translation from the high-probability guarantee, we can obtain a MSE bound here as well. For this result, we use the notation $\text{CNF}(i;\delta)$ to highlight the fact that the confidence bounds hold with probability $1-\delta$.

Corollary 4 (MSE bound)**.** *In addition to Assumption 1 and Assumption 2, assume that* $\theta^\star, \hat{\theta}_i \in [0, R]$ *a.s., $\forall i$, and that* CNF *is deterministic.*[2] *Then for any* $\delta > 0$*,*

$$\mathbb{E}(\hat{\theta}_{\hat{i}} - \theta^\star)^2 \leq C/\kappa^2 \min_i \left\{\text{B}(i)^2 + \text{CNF}(i;\delta)^2\right\} + R^2\delta,$$

where $C > 0$ *is a universal constant.*[3]

We state this bound for completeness but remark that it is typically loose in constant factors because it is proven through a high probability guarantee. In particular, we typically require $\text{CNF}(i) > \sqrt{\text{VAR}(i)}$ to satisfy Assumption 1, which is already loose in comparison with a more direct MSE bound. Unfortunately, Lepski's principle cannot provide direct MSE bounds without estimating the MSE itself, which is precisely the problem we would like to avoid. On the other hand, the high probability guarantee provided by Theorem 3 is typically more practically meaningful.

4. Application 1: continuous contextual bandits

For our first application, we consider a contextual bandit setting with continuous action space, following Kallus & Zhou (2018). Let $\mathcal{X}$ be a context space and $\mathcal{A} = [0,1]$ be a univariate real-valued action space. There is a distribution $\mathcal{P}$ over context-reward pairs, which is supported on $(\mathcal{X}, \mathcal{A} \to [0,1])$. We have a stochastic logging policy $\pi_\text{L} : \mathcal{X} \to \Delta(\mathcal{A})$ which induces the distribution $\mathcal{D}$ by generating tuples $\{(x_i, a_i, r_i(a_i), \pi_\text{L}(a_i))\}_{i=1}^n$, where $(x_i, r_i) \sim \mathcal{P}$, $a_i \sim \pi_\text{L}(x_i)$, only $r_i(a_i)$ is observed, and $\pi_\text{L}(a_i)$ denotes the density value. This is a bandit setting as the distribution $\mathcal{P}$ specifies rewards for all actions, but only the reward for the chosen action is available for estimation.

For off-policy evaluation, we would like to estimate the value of some target policy π_T, which is given by $V(\pi_\text{T}) := \mathbb{E}_{(x,r)\sim\mathcal{P}, a\sim\pi_\text{T}(x)}[r(a)]$. Of course, we do not have sample access to $\mathcal{P}$ and must resort to the logged tuples generated by π_L. A standard off-policy estimator in this setting is

$$\hat{V}_h(\pi_\text{T}) := \frac{1}{nh} \sum_{i=1}^n \frac{K(|\pi_\text{T}(x_i) - a_i|/h) r_i(a_i)}{\pi_\text{L}(a_i)},$$

where $K : \mathbb{R}_+ \to \mathbb{R}$ is a *kernel function* (e.g., the boxcar kernel $K(u) = \frac{1}{2}\mathbf{1}\{u \leq 1\}$). This estimator has appeared in recent work (Kallus & Zhou, 2018; Krishnamurthy et al., 2019). The key parameter is the bandwidth h, which modulates a bias-variance tradeoff, where smaller bandwidths have lower bias but higher variance.

4.1. Theory

For a simplified exposition, we instantiate our general framework when (1) π_L is the uniform logging policy, (2) K is the boxcar kernel, and (3) we assume that $\pi_\text{T}(x) \in [\gamma_0, 1-\gamma_0]$ for all x. These simplifying assumptions help clarify the results, but they are not fundamentally limiting.

Fix $\gamma \in (0,1)$ and let $\mathcal{H} := \{\gamma_0 \gamma^{M-i} : 1 \leq i \leq M\}$

[2] The restriction to deterministic confidence functions can easily be removed with another concentration argument.

[3] We have not attempted to optimize the constant, which can can be extracted from our proof in Appendix A.

denote a geometrically spaced grid of M bandwidth values. Let $\hat{\theta}_i := \hat{V}_{h_i}(\pi_{\mathrm{T}})$. For the confidence term, in Appendix A, we show that we can set

$$\mathrm{CNF}(i) := \sqrt{\frac{2\log(2M/\delta)}{nh_i}} + \frac{2\log(2M/\delta)}{3nh_i} \qquad (2)$$

and this satisfies Assumption 1. With this form, it is not hard to see that the second part of Assumption 2 is also satisfied, and so we obtain the following result.

Theorem 5. *Consider the setting above with uniform π_{L}, boxcar kernel, and $\mathcal{H}$ as defined above. Let* B *be any valid and monotone bias function, and define* CNF *as in (2). Then Assumption 1 and Assumption 2 are satisfied with $\kappa = \gamma$, so the guarantee in Theorem 3 applies.*

In particular, if γ, γ_0 are constants and rewards are L-Lipschitz, for $\omega(\sqrt{\log(\log(n))/n}) \le L \le O(n)$, then

$$\left|\hat{\theta}_{\hat{i}} - \theta^\star\right| \le O\left(\left(\frac{L\log(\log(n)/\delta)}{n}\right)^{1/3}\right)$$

with probability at least $1-\delta$, without knowledge of the Lipschitz constant L.

For the second statement, we remark that if the Lipschitz constant were known, the best error rate achievable is $O((L\log(1/\delta)/n)^{1/3})$. Thus, SLOPE incurs almost no price for adaptation. We also note that it is typically impossible to know this parameter in practice.

It is technical but not difficult to derive a more general result, relaxing many of the simplifying assumptions we have made. To this end, we provide a guarantee for non-uniform π_{L} in the appendix. We do not pursue other extensions here, as the necessary techniques are well-understood (c.f., Kallus & Zhou (2018); Krishnamurthy et al. (2019)).

4.2. Experiments

We empirically evaluate using SLOPE for bandwidth selection in a synthetic environment for continuous action contextual bandits. We summarize the experiments and findings here with detailed description in Appendix B.[4]

The environment. We use a highly configurable synthetic environment, which allows for action spaces of arbitrary dimension, varying reward function, reward smoothness, kernel, target, and logging policies. In our experiments, we focus on $\mathcal{A} = [0,1]$. We vary all other parameters, as summarized in Table 1.

The simulator prespecifies a mapping $x \mapsto a^\star(x)$ which is the global maxima for the reward function. We train deterministic policies by regressing from the context to this global maxima. For the logging policy, we use two "softening" approaches for randomization, following Farajtabar et al. (2018). We use two regression models (linear, decision tree), and two softenings in addition to uniform logging, for a total of 5 logging and 2 target policies.

Reward fn $\in$ {quadratic, absolute value}
Lipschitz const $\in \{0.1, 0.3, 1, 3, 10\}$
Kernel $\in$ {boxcar, Epanechnikov}
$\pi_{\mathrm{T}} \in$ {linear, tree}
$\pi_{\mathrm{L}} \in$ {linear, tree}
Randomization $\in$ {uniform, friendly, adversarial}
Samples $\in \{10^i : i \in \{1,2,3,4,5\}\}$

Table 1. Contextual Bandit experimental conditions.

Methods. We consider 7 different choices of geometrically spaced bandwidths $\mathcal{H} := \{2^{-i} : i \in [7]\}$. We evaluate the performance of these fixed bandwidths in comparison with SLOPE, which selects from $\mathcal{H}$. For SLOPE, we simplify the implementation by replacing the confidence function in (2), with twice the empirical standard deviation of the corresponding estimate. This approximation is a valid asymptotic confidence interval and is typically sharper than (2), so we expect it to yield better practical performance. We also manually enforce monotonicity of this confidence function.

We are not aware of other viable baselines for this setting. In particular, the heuristic method of Kallus & Zhou (2018) is too computationally intensive to use at our scale.

Experiment setup. We have 1000 conditions determined by: logging policy, target policy, reward functional form, reward smoothness, kernel, and number of samples n. For each condition, we first obtain a Monte Carlo estimate of the ground truth $V(\pi_{\mathrm{T}})$ by collecting 100k samples from π_{T}. Then we collect n trajectories from π_{L} and evaluate the squared error of each method $(\hat{V} - V(\pi_{\mathrm{T}}))^2$. We perform 30 replicates of each condition with different random seeds and calculate the correspond mean squared error (MSE) for each method: $\mathrm{MSE} := \frac{1}{30}\sum_{i=1}^{30}(\hat{V}_i - V(\pi_{\mathrm{T}}))^2$ where $\hat{V}_i$ is the estimate on the i^{th} replicate.

Results. The left panel of Figure 3 aggregates results via the empirical CDF of the normalized MSE, where we normalize by the worst MSE in each condition. The point (x, y) indicates that on y-fraction of conditions the method has normalized MSE at most x, so better methods lie in the top-left quadrant. We see that SLOPE is the top performer in comparison with the fixed bandwidths.

In the center panel, we record the results of pairwise comparisons between all methods. Entry (i, j) of this array is the fraction of conditions where method i is significantly better than method j (using a paired t-test with significance level

[4] Code for this section is publicly available at https://github.com/VowpalWabbit/slope-experiments.

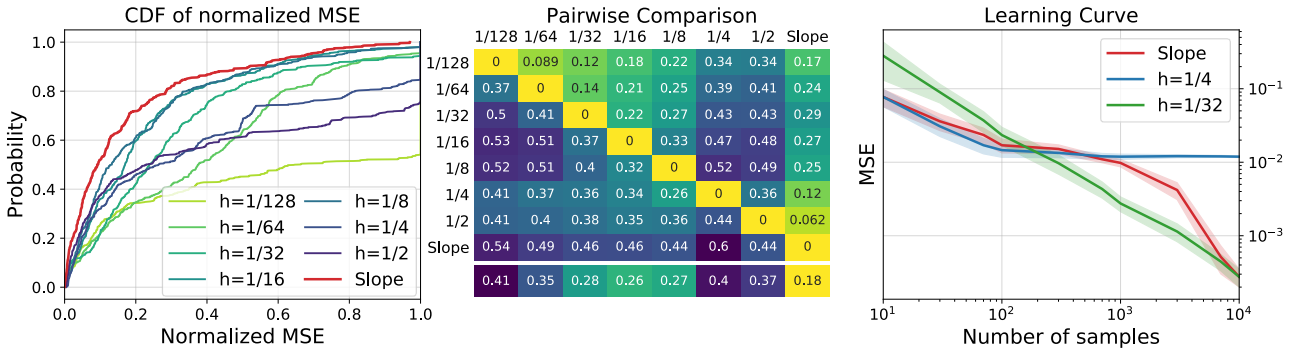

Figure 3. Experimental results for contextual bandits with continuous actions. Left: CDF of normalized MSE across all 480 conditions. Normalization is by the worst MSE for that condition. Middle: Pairwise comparison matrix, entry $P_{i,j}$ counts the fraction of conditions where method i is statistically significantly better than method j, so larger numbers in the rows (or smaller numbers in the columns) is better. Right: asymptotic behavior of SLOPE selecting between two bandwidths.

0.05). Better methods have smaller numbers in their column, which means they are typically not significantly worse than other methods. The final row summarizes the results by averaging each column. In this aggregation, SLOPE also outperforms each individual fixed bandwidth, demonstrating the advantage in data-dependent estimator selection.

Finally, in the right panel, we demonstrate the behavior of SLOPE in a single condition as n increases. Here SLOPE is only selecting between two bandwidths $\mathcal{H} := \{1/4, 1/32\}$. When n is small, the small bandwidth has high variance but as n increases, the bias of the larger bandwidth dominates. SLOPE effectively navigates this tradeoff, tracking $h = 1/4$ when n is small, and switching to $h = 1/32$ as n increases.

Summary. SLOPE is the top performer when compared with fixed bandwidths in our experiments. This is intuitive as we do not expect a single fixed bandwidth to perform well across all conditions. On the other hand, we are not aware of other approaches for bandwidth selection in this setting, and our experiments confirm that SLOPE is a viable and practically effective approach.

5. Application 2: reinforcement learning

Our second application is multi-step reinforcement learning (RL). We consider episodic RL where the agent interacts with the environment in episodes of length H. Let $\mathcal{X}$ be a state space and $\mathcal{A}$ a finite action space. In each episode, a trajectory $\tau := (x_1, a_1, r_1, x_2, a_2, r_2, \ldots, x_H, a_H, r_H)$ is generated where (1) $x_1 \in \mathcal{X}$ is drawn from a starting distribution P_0, (2) rewards $r_h \in \mathbb{R}$ and next state $x_{h+1} \in \mathcal{X}$ are drawn from a system descriptor $(r_h, x_{h+1}) \sim P_+(x_h, a_h)$ for each h (with the obvious definition for time H), and (3) actions $a_1, \ldots, a_H \in \mathcal{A}$ are chosen by the agent. A policy $\pi : \mathcal{X} \mapsto \Delta(\mathcal{A})$ chooses a (possibly stochastic) action in each state and has value $V(\pi) := \mathbb{E}\left[\sum_{h=1}^{H} \gamma^{h-1} r_h \mid a_{1:H} \sim \pi\right]$, where $\gamma \in (0, 1)$ is a discount factor. For normalization, we assume that rewards are in $[0, 1]$ almost surely.

Environment	GW	MC	Graph	PO-Graph
Horizon	25	250	16	16
MDP	Yes	Yes	Yes	No
Sto Env	Both	No	Yes	Yes
Sto Rew	No	No	Both	Both
Sparse Rew	No	No	Both	Both
Model class	Tabular	NN	Tabular	Tabular
Samples	$2^{7:9}$	$2^{8:10}$	$2^{7:10}$	$2^{7:10}$
# of policies	5	4	2	2

Table 2. RL Environment Details

For off-policy evaluation, we have a dataset of n trajectories $\{(x_{i,1}, a_{i,1}, r_{i,1}, \ldots, x_{i,H}, a_{i,H}, r_{i,H})\}_{i=1}^{n}$ generated by following some logging policy π_{L}, and we would like to estimate $V(\pi_{\mathrm{T}})$ for some other target policy. The importance weighting approach is also standard here, and perhaps the simplest estimator is

$$\hat{V}_{\mathrm{IPS}}(\pi_{\mathrm{T}}) := \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{H} \gamma^{h-1} \rho_{i,h} r_{i,h}, \tag{3}$$

where $\rho_{i,h} := \prod_{h'=1}^{h} \frac{\pi_{\mathrm{T}}(a_{i,h'}|x_{i,h'})}{\pi_{\mathrm{L}}(a_{i,h'}|x_{i,h'})}$ is the step-wise importance weight. This estimator is provably unbiased under very general conditions, but it suffers from high variance due to the H-step product of density ratios.[5] An alternative approach is to directly model the value function using supervised learning, as in a regression based dynamic programming algorithm like Fitted Q Evaluation (Riedmiller, 2005; Szepesvári & Munos, 2005). While these "direct modeling" approaches have very low variance, they are typically highly biased because they rely on supervised learning models that cannot capture the complexity of the environment. Thus

[5] We note that there are variants with improved variance. As our estimator selection question is somewhat orthogonal, we focus on the simplest estimator.

they lie on the other extreme of the bias-variance spectrum.

To navigate this tradeoff, Thomas & Brunskill (2016) propose a family of *partial importance weighting estimators*. To instantiate this family, we first train a direct model $\hat{V}_{\mathrm{DM}} : \mathcal{X} \times [H] \to \mathbb{R}$ to approximate $(x, h) \mapsto \mathbb{E}_\pi \left[\sum_{h'=h}^{H} \gamma^{h'-h} r_{h'} \mid x_h = x\right]$, for example via Fitted Q Evaluation. Then, the estimator is

$$\hat{V}_\eta(\pi_{\mathrm{T}}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{h=1}^{\eta} \gamma^{h-1}\rho_{i,h} r_{i,h} + \frac{1}{n}\sum_{i=1}^{n} \gamma^{\eta}\rho_{i,\eta}\hat{V}_{\mathrm{DM}}(x_{i,\eta+1}, \eta+1), \tag{4}$$

The estimator has a parameter η that governs a *false horizon* for the importance weighting component. Specifically, we only importance weight the rewards up until time step η and we complete the trajectory with the predictions from a direct modeling approach. The model selection question here centers around choosing the false horizon η at which point we truncate the unbiased importance weighted estimator.

5.1. Theory

We instantiate our general estimator selection framework in this setting. Let $\hat{\theta}_i := \hat{V}_{H-i+1}(\pi_{\mathrm{T}})$ for $i \in \{1, \ldots, H+1\}$. Intuitively, we expect that the variance of $\hat{\theta}_i$ is large for small i, since the estimator involves a product of many density ratios. Indeed, in the appendix, we derive a confidence bound and prove that it verifies our assumptions. The bound is quite complicated so we do not display it here, but we refer the interested reader to (7) in Appendix A. The bound is a Bernstein-type bound which incorporates both variance and range information. We bound these as

$$\mathrm{Variance}(\hat{V}_\eta(\pi_{\mathrm{T}})) \le 3V_{\max}^2(1 + \sum_{h=1}^{\eta}\gamma^{2(h-1)}p_{\max}^h)$$
$$\mathrm{Range}(\hat{V}_\eta(\pi_{\mathrm{T}})) \le 3V_{\max}(1 + \sum_{h=1}^{\eta}\gamma^{h-1}p_{\max}^h),$$

where $V_{\max} := (1-\gamma)^{-1}$ is the range of the value function and $p_{\max} := \sup_{x,a} \frac{\pi_{\mathrm{T}}(a|x)}{\pi_{\mathrm{L}}(a|x)}$ is the maximum importance weight, which should be finite. Equipped with these bounds, we can apply Bernstein's inequality to obtain a valid confidence interval.[6] Moreover, it is not hard to show that this confidence interval is monotonic with $\kappa := (1+\gamma p_{\max})^{-1}$. This yields the following theorem.

Theorem 6 (Informal)**.** *Consider the episodic RL setting with $\hat{\theta}_i := \hat{V}_{H-i+1}(\pi_{\mathrm{T}})$ defined in (4). Let* B *be any valid and monotone bias function. Then with* CNF(i) *as in (7) in the appendix, Assumption 1 and Assumption 2 with $\kappa := (1+\gamma p_{\max})^{-1}$ hold, so Theorem 3 applies.*

A more precise statement is provided in Appendix A, and we highlight some salient details here. First, our analysis actually applies to a doubly-robust variant of the estimator $\hat{V}_\eta$, in the spirit of (Jiang & Li, 2016). Second, $\mathrm{B}(i) := \frac{\gamma^{H-i+1}-\gamma^H}{1-\gamma}$ is valid and monotone, and can be used to obtain a concrete error bound. However, the oracle inequality yields a stronger conclusion, since it applies for *any* valid and monotone bias function. This universality is particularly important when using the doubly robust variant, since it is typically not possible to sharply bound the bias.

The closest comparison is MAGIC (Thomas & Brunskill, 2016), which is strongly consistent in our setting. However, it does not satisfy any oracle inequality and is dominated by SLOPE in experiments.

5.2. Experiments

We evaluate SLOPE in RL environments spanning 106 different experimental conditions. We also compare with previously proposed estimators and assess robustness under various conditions. Our experiments closely follow the setup of Voloshin et al. (2019). Here we provide an overview of the experimental setup and highlight the salient differences from theirs. All experimental details are in Appendix C.[7]

The environments. We use four RL environments: Mountain Car, Gridworld, Graph, and Graph-POMDP (abbreviated MC, GW, Graph, and PO-Graph). All four environments are from Voloshin et al. (2019), and they provide a broad array of environmental conditions, varying in terms of horizon length, partial observability, stochasticity in dynamics, stochasticity in reward, reward sparsity, whether function approximation is required, and overlap between logging and target policies. Logging and target policies are from Voloshin et al. (2019). A summary of the environments and their salient characteristics is displayed in Table 2.

Methods. We compare four estimators: the direct model (DM), a self-normalized doubly robust estimator (WDR), (c) MAGIC, and (d) SLOPE. All four methods use the same direct model, which we train either by Fitted Q Evaluation or by $Q^\pi(\lambda)$ (Munos et al., 2016), following the guidelines in Voloshin et al. (2019). The doubly robust estimator is the most competitive estimator in the family of full-trajectory importance weighting. It is similar to (3), except that the direct model is used as a control variate and the normalizing constant n is replaced with the sum of importance weights. MAGIC, as we have alluded to, is the only other estimator

[6] This yields a relatively concise deviation bound, but we note that it is not the sharpest possible.

[7] Code for this section is available at https://github.com/clvoloshin/OPE-tools.

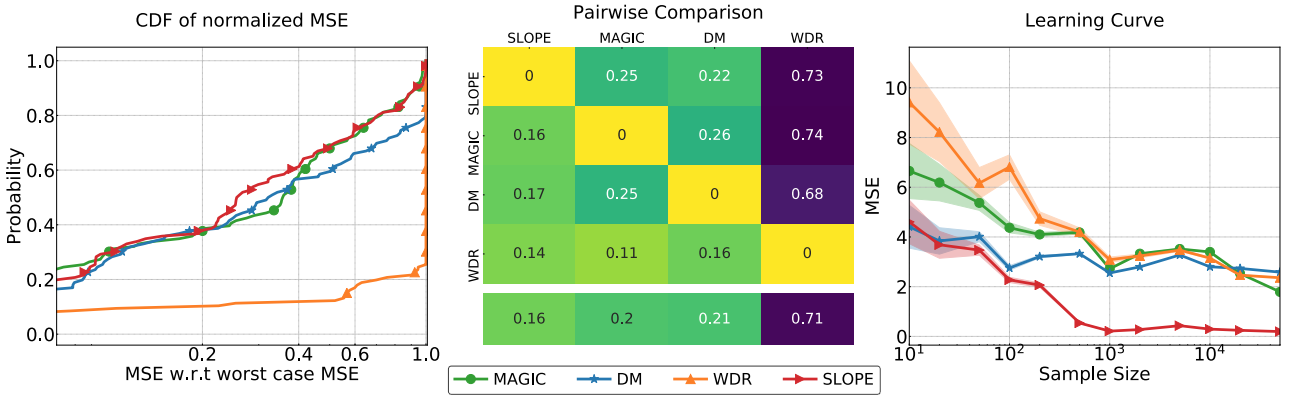

Figure 4. Left: Cumulative distribution function of the normalized MSE for all conditions, Middle: Pairwise comparison matrix P for the methods, over all conditions. Element P_{ij} denotes the percentage of times that method i outperforms method j. The last row shows the column average for each method, the lower the better. Right: Learning Curve for the Hybrid domain.

selection procedure we are aware of for this setting. It aggregates partial importance weighting estimators to optimize a surrogate for the MSE. For SLOPE, we use twice the empirical standard deviation as the confidence function, which is asymptotically valid and easier to compute.

We do not consider other baselines for two reasons. First, DM, WDR, and MAGIC span the broad estimator categories (importance weighted, direct, hybrid) within which essentially all estimators fall. Secondly, many other estimators have hyperparameters that must be tuned, and we believe SLOPE will also be beneficial when used in these contexts.

Experiment Setup. We have 106 experimental conditions determined by environment, stochasticity of dynamics and reward, reward sparsity, logging policy, target policy, and number of trajectories n. For each condition, we calculate the MSE for each method by averaging over 100 replicates.

Results. In the left panel of Figure 4, as in Section 4.2, we first visualize the aggregate results via the cumulative distribution function (CDF) of the normalized MSE in each condition (normalizing by the worst performing method in each condition). As a reminder, the figure reads as follows: for each x value, the corresponding y value is the fraction of conditions where the estimator has normalized MSE at most x. In this aggregation, we see that WDR has the worst performance, largely due to intolerably high variance. MAGIC and DM are competitive with each other with MAGIC having a slight edge. SLOPE appears to have the best aggregate performance; for the most part its CDF dominates the others.

In the central panel, we display an array of statistical comparisons between pairs of methods. As before, entry (i, j) of this array is computed by counting the fraction of conditions where method i beats j in a statistically significant manner (we use paired t-test on the MSE with significance level 0.05). The column-wise averages are also displayed.

In this aggregation, we see clearly that SLOPE dominates the three other methods. First, SLOPE has column average that is smaller than the other methods. More importantly, SLOPE is favorable when compared with each other method individually. For example, SLOPE is (statistically) significantly worse than MAGIC on 16% of the conditions, but it is significantly better on 25%. Thus, this visualization clearly demonstrates that SLOPE is the best performing method in aggregate across our experimental conditions.

Before turning to the final panel of Figure 4, we recall Figure 1, where we display results for two specific conditions. Here, we see that SLOPE outperforms or is statistically indistinguishable from the best baseline, regardless of whether direct modeling is better than importance weighting! We are not aware of any selection method that enjoys this property.

Learning curves. The final panel of Figure 4 visualizes the performance of the four methods as the sample size increases. Here we consider the Hybrid domain from Thomas & Brunskill (2016), which is designed specifically to study the performance of partial importance weighting estimators. The domain has horizon 22, with partial observability in the first two steps, but full observability afterwards. Thus a (tabular) direct model is biased since it is not expressive enough for the first two time steps, but $\hat{V}_2(\pi_T)$ is a great estimator since the direct model is near-perfect afterwards.

The right panel of Figure 4 displays the MSE for each method as we vary the number of trajectories, n (we perform 128 replicates and plot bars at ± 2 standard errors). We see that when n is small, DM dominates, but its performance does not improve as n increases due to bias. Both WDR and MAGIC catch up as n increases, but SLOPE is consistently competitive or better across all values of n, outperforming the baselines by a large margin. Indeed, this is because

SLOPE almost always chooses the optimal false horizon index of $\eta = 2$ (e.g., 90% of the replicates when $n = 500$).

Summary. Our experiments show that SLOPE is competitive, if not the best, off-policy evaluation procedure among SLOPE, MAGIC, DM, and WDR. We emphasize that SLOPE is not an estimator, but a selection procedure that in principle can select hyperparameters for many estimator families. Our experiments with the partial importance weighting family are quite convincing, and we believe this demonstrates the potential for SLOPE when used with other estimator families for off-policy evaluation in RL.[8]

6. Discussion

In summary, this paper presents a new approach for estimator selection in off-policy evaluation, called SLOPE. The approach applies quite broadly; in particular, by appropriately spacing hyperparameters, many common estimator families can be shown to satisfy the assumptions for SLOPE. To demonstrate this, we provide concrete instantiations in two important applications. Our theory yields, to our knowledge, the first oracle-inequalities for off-policy evaluation in RL. Our experiments demonstrate strong empirical performance, suggesting that SLOPE may be useful in many off-policy evaluation contexts.

Acknowledgements

We thank Mathias Lecuyer for comments on a preliminary version of this paper and Miro Dudik for formative discussions. YS is supported by the Bloomberg Data Science Fellowship.

References

Birgé, L. An alternative point of view on lepski's method. *Lecture Notes-Monograph Series*, 2001.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 1996.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv:1606.01540*, 2016.

Dudík, M., Erhan, D., Langford, J., and Li, L. Doubly robust policy evaluation and optimization. *Statistical Science*, 2014.

[8] In settings where straightforward empirical variance estimates are not available, the bootstrap may provide an alternative approach for constructing the CNF function. Experimenting with such estimators is a natural future direction.

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

Goldenshluger, A. and Lepski, O. V. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 2011.

Hirano, K., Imbens, G. W., and Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 2003.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 1952.

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

Kallus, N. and Zhou, A. Policy evaluation and optimization with continuous treatments. In *Artificial Intelligence and Statistics*, 2018.

Kpotufe, S. and Garg, V. Adaptivity to local smoothness and dimension in kernel regression. In *Advances in Neural Information Processing Systems*, 2013.

Krishnamurthy, A., Langford, J., Slivkins, A., and Zhang, C. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. In *Conference on Learning Theory*, 2019.

Lepski, O. V. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 1992.

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

Lepskii, O. V. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 1991.

Lepskii, O. V. Asymptotically minimax adaptive estimation. ii. schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications*, 1993.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, 2018.

Mathé, P. The lepskii principle revisited. *Inverse problems*, 2006.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, 2016.

Page, S. and Grünewälder, S. The goldenshluger-lepski method for constrained least-squares estimators over RKHSs. *arXiv:1811.01061*, 2018.

Precup, D., Sutton, R. S., and Singh, S. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning*, 2000.

Riedmiller, M. Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, 2005.

Su, Y., Wang, L., Santacatterina, M., and Joachims, T. Cab: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learning*, 2019.

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

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, 1988.

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudik, M., Langford, J., Jose, D., and Zitouni, I. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*, 2017.

Szepesvári, C. and Munos, R. Finite time bounds for sampling based fitted value iteration. In *International Conference on Machine Learning*, 2005.

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

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

Wang, Y.-X., Agarwal, A., and Dudik, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, 2017.