Confident Off-Policy Evaluation and Selection through Self-Normalized Importance Weighting

Ilja Kuzborskij **Claire Vernade** **András György** **Csaba Szepesvári**
DeepMind

Abstract

We consider off-policy evaluation in the contextual bandit setting for the purpose of obtaining a robust off-policy *selection* strategy, where the selection strategy is evaluated based on the value of the chosen policy in a set of proposal (target) policies. We propose a new method to compute a lower bound on the value of an arbitrary target policy given some logged data in contextual bandits for a desired coverage. The lower bound is built around the so-called Self-normalized Importance Weighting (SN) estimator. It combines the use of a semi-empirical Efron-Stein tail inequality to control the concentration and Harris' inequality to control the bias. The new approach is evaluated on a number of synthetic and real datasets and is found to be superior to its main competitors, both in terms of tightness of the confidence intervals and the quality of the policies chosen.

1 Introduction

Consider the following offline stochastic decision making problem where an agent observes a collection of contexts, actions, and associated rewards collected by some *behavior policy* and has to choose a new policy from a finite set of *target policies*. The agent's goal is to select the policy that has the highest *value*, defined as its expected reward. We call this variant of the contextual bandit problem *best-policy selection* (in the off-policy setting).

This problem is encountered for instance in personalized recommendation and allocation problems, such as in medical applications, online advertising, and operations research: a *static behavior policy* (e.g. a randomized classifier) is run online and for each chosen action, only partial (bandit) feedback is received. The collected data must then be used to evaluate other *static target policies* (Agarwal et al., 2017) with the goal of choosing a policy that will perform better on average.

We emphasize that this work focuses on such *static* policies. These are an important class of policies that are preferred in practice in many cases. For example, recommender systems based on a batch-learnt classifier with predictable behavior or expert-designed rules as in medical applications.

At its core, this selection problem relies on off-policy evaluation (Bottou et al., 2013; Dudík et al., 2011; Swaminathan and Joachims, 2015a), which is concerned with accurately estimating the value of a target policy, using a logged dataset, and aiming for a good bias-variance trade-off. To guarantee that such trade-off holds in practice, one would ideally rely on high-probability confidence bounds. However, only few works on off-policy evaluation have provided practically computable, tight confidence bounds. It is recognized, though, that such bounds should depend on the empirical variance of the estimator (Bottou et al., 2013; Thomas et al., 2015a,b; Swaminathan and Joachims, 2015b; Metelli et al., 2018). In general, this is a nontrivial task and the standard tools such as sub-Gaussian tail concentration inequalities (e.g. Bernstein's inequality) are ill-suited for this job. Indeed, most estimators derive from *Importance Weighting (IW)*, a standard technique for estimating a property of a distribution while having access to a sample generated by another distribution. At the same time, arguably one of the most interesting scenarios is when the target and the behavior policies are misaligned, which corresponds to situations when the weights of IW exhibit a heavy-tailed behavior. In such cases, the control of the moments of the IW estimator, and therefore its concentration, is in general futile.

In this paper we revisit *Self-normalized Importance Weighting (SN)*, a self-normalized version of IW. This estimator is asymptotically unbiased, and is known

Proceedings of the 24$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

for its small variance in practice (Hesterberg, 1995). Moreover, unlike IW, all the moments of SN are simultaneously bounded. These favorable properties allow us to prove finite-sample concentration inequalities at the price of a (controllable) bias.

Contributions. Our main result is a new high-probability lower bound on the value of the SN estimator, stated in Section 3 and proved in Section 4. Moreover, we formulate the off-policy selection problem and propose a systematic, appropriate approach using off-policy evaluation tools. In this context, we demonstrate empirically (Section 6) that our bound achieves the best performance compared to all existing and proposed baselines.

2 Notation and preliminaries

Off-policy evaluation for contextual bandits. For the stochastic contextual bandit model, an off-policy evaluation problem is characterized by a triplet $(P_X, P_{R|X,A}, \pi_b)$, where P_X is a probability measure over contexts (we assume that the context space is any probability space $(\mathcal{X}, \Sigma_{\mathcal{X}}, P_X)$), $P_{R|X,A}$ is a probability kernel producing the reward distribution given the context $X \in \mathcal{X}$ and action $A \in [K] = \{1, \dots, K\}$, and $\pi_b : \mathcal{X} \to [K]$ is a behaviour policy, that is, a conditional distribution over actions given the context.

The decision maker observes (S, π_b), where $S = ((X_1, A_1, R_1), \dots, (X_n, A_n, R_n))$ is a tuple of independent context-action-reward triplets, obtained by following the behaviour policy π_b: for all $i \in [n]$, $A_i \sim \pi_b(\cdot|X_i)$, where $X_i \sim P_X$. The reward $R_i \sim P_{R|X,A}$ is a *bandit feedback* as it only reveals the value of the taken action A_i. For example, in a multi-class classification task, the reward may be a (noisy) binary random variable that indicates whether the chosen label is right or not, but the true label is not revealed. We assume that the rewards are bounded in $[0, 1]$ and that π_b is known and can be evaluated at any context-action pair. In many applications the behavior policy represents the policy running in the system, which is usually known by the practitioner and can be queried.

A policy π is any conditional distribution over the actions and its value $v(\pi)$ is defined by

$$v(\pi) = \int_{\mathcal{X}} \sum_{a \in [K]} \pi(a|x) r(x, a) \, \mathrm{d}P_X(x) \qquad (1)$$

where $r(x, a) = \int u \, \mathrm{d}P_{R|X,A}(u|x, a)$ is the mean reward for a given context-action pair (x, a). Similarly to π_b, we assume that any known policy π can be evaluated for any context-action pair. In general, the goal of off-policy evaluation is to return an estimate $\hat{v}^{\text{est}}(\pi)$ of the value $v(\pi)$ of some *target policy* with controlled bias and variance. In contrast, we are concerned with obtaining a data-dependent scoring function that allows the decision maker to choose the highest performing target policy in a set of candidate policies. We call this statistical problem *best-policy selection*.

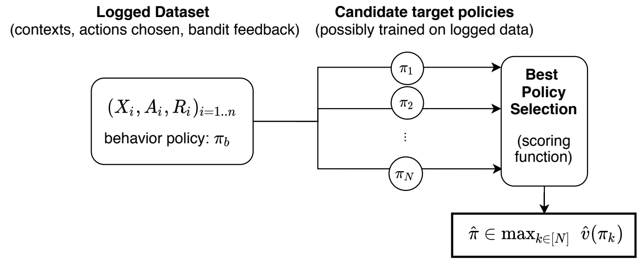

Figure 1: The best-policy selection problem

Best-policy selection. Given a finite set $\{\pi_1, \dots, \pi_N\}$ of policies, called the *target* policies, our goal is to design a decision algorithm that returns a policy $\hat{\pi}$ with the highest value. We denote $\pi^* \in \max_k v(\pi_k)$ a policy with maximum value in the target set, and the objective of the decision maker is to identify π^*. The decision maker uses a scoring function $\hat{v}^{\text{est}}$ as input and chooses the policy that has the highest score: $\hat{\pi} \in \arg\max_k \hat{v}^{\text{est}}(\pi_k)$ (see Figure 1 for an illustration of the problem). The quality of the selected $\hat{\pi}$ is measured by how close its value is to that of the optimal policy π^*. The choice of the estimation method used to make the decision is crucial. A naive approach would be to use directly a value estimator (see below for a review of classical methods). However, we demonstrate in this paper that this may lead to dramatic losses. In turn, we propose a scoring function based on a high-probability lower bound on the value.

Remark 1. *Best-policy selection is related to the problem of off-policy learning, which aims at designing data-based policies that have a high value. In our case, policies are created arbitrarily and possibly trained on the logged dataset, and our decision rule only guarantees that the decision maker returns the best performing one in a given set. We pose the problem of learning over a discrete class as a decision problem.*

Classical estimation approaches. Importance Weighting (IW) is the most widely known approach to obtain an unbiased value estimator: $\hat{v}^{\text{IW}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i|X_i)}{\pi_b(A_i|X_i)} R_i = \frac{1}{n} \sum_{i=1}^{n} W_i R_i$, which is unbiased since $\mathbb{E}_{\pi_b}[WR] = \mathbb{E}_{\pi}[R]$. Each data point is reweighted by its *importance weight*, which is related to the likelihood ratio of the event of selecting the given action under both policies. In this paper we focus on a self-normalized version of IW, called *Self-normalized Importance Weighting (SN)* (Hesterberg, 1995), defined as

$$\hat{v}^{\text{SN}}(\pi) = \frac{\sum_{k=1}^n W_k R_k}{\sum_{i=1}^n W_i} ,$$

where the sum of the weights (a random variable) is used instead of n for normalization. While the SN estimator is not unbiased, it is generally regarded as a good estimator. As a start, we may for example note that it gives values in the $[0,1]$ interval. To get a sense of the concentration of the SN estimator, it is instrumental to consider the bound that one may get from the Efron-Stein inequality on its variance (see, e.g. (Kuzborskij and Szepesvári, 2019)): A quick calculation gives $\mathrm{Var}(\hat{v}^{\text{SN}}(\pi)) \le n\,\mathbb{E}[\tilde{W}_1^2]$ where $\tilde{W}_1 = W_1/(W_1 + \cdots + W_n)$. Note that in the ideal case when the behavior and target policies coincide, $W_i = 1$ and thus $\tilde{W}_1 = 1/n$. In the Monte-Carlo simulation literature, $n_{\text{eff}} = (\tilde{W}_1^2 + \cdots + \tilde{W}_n^2)^{-1} \approx 1/(n\,\mathbb{E}[\tilde{W}_1^2])$ is known as the effective sample-size (Kong, 1992; Elvira et al., 2018), and it provides a quantitative complexity measure of the estimation problem for a target π: the quality of $\hat{v}^{\text{SN}}(\pi)$ is as good as if we had used n_{eff} samples from π instead of n samples from π_b. However, since n_{eff} is random, this is not entirely satisfactory as it does not lead to an easy finite-sample concentration bound using standard tools.

3 Confidence bound for SN Estimator

We now state our main result, a high-probability lower bound on the value of a policy π based on the SN estimator. Even though we use it here as an efficient scoring function for the off-policy selection problem, we believe this result is of independent interest.

Theorem 1. *As defined above, let $W_i = \pi(A_i|X_i)/\pi_b(A_i|X_i)$ for all i, and assume that $(W_i, R_i)_{i=1}^n$ are independent from each other and $R_i \in [0,1]$ a.s. (almost surely). Let $Z = W_1 + \cdots + W_n$ and $Z^{(k)} = Z + (W_k' - W_k)$. Then, for any $x \ge 2$, with probability at least $1 - 2e^{-x}$,*

$$v(\pi) \ge \left(B\left(\hat{v}^{\text{SN}}(\pi) - \epsilon\right)_+ - \sqrt{\frac{x}{2n}}\right)_+ \text{ where}$$

$$\epsilon = \sqrt{2\left(V^{\text{SN}} + U^{\text{SN}}\right)\left(x + \frac{1}{2}\ln\left(1 + \frac{V^{\text{SN}}}{U^{\text{SN}}}\right)\right)}$$

$$V^{\text{SN}} = \sum_{k=1}^n \mathbb{E}\left[\left(\frac{W_k}{Z} + \frac{W_k'}{Z^{(k)}}\right)^2 \,\middle|\, W_1^k, X_1^n\right] ,$$

$$U^{\text{SN}} = \mathbb{E}[V^{\text{SN}}|X_1^n] \quad B = \min\left(\mathbb{E}\left[\frac{n}{Z} \,\middle|\, X_1^n\right]^{-1}, 1\right)$$

where $(a)_+ = \max(a, 0)$ and $a_1^l = (a_1, \ldots, a_l)$.

Comments. This bound essentially depends on an *Efron-Stein estimate* of the variance of SN, V^{SN}, while always conditioning on the contexts. It can be qualified of *semi-empirical* as it relies on taking expectations over the weights. Thus, its computability relies on our ability to compute those expectations, which requires knowing π_b and being able to evaluate it on any context-action pair. Remarkably, V^{SN} can be computed without *any* knowledge of the context distribution P_X or the reward probability kernel $P_{R|X,A}$. Notably, the bias B quantifying policy mismatch appears as a multiplicative term in the bound. Indeed, when π coincides with π_b exactly, $B = 1$ and the estimator suffers no bias. Conversely, the more the mismatch, the larger the bias is. Similarly to V^{SN}, the bias can be computed since the distribution of the importance weights (conditioned on the contexts) is known. Finally, as discussed earlier, it is worth noting that the behavior of V^{SN} is closely related to that of the effective sample size n_{eff}: when the target and the behavior policies coincide, we have $V^{\text{SN}} = \mathcal{O}(1/n)$, while when policies are in full mismatch (e.g. Dirac deltas on two different actions) we have $V^{\text{SN}} = \Theta(1)$. In the intermediate regime of a partial mismatch, for instance when the distribution of importance weights is heavy-tailed, $V^{\text{SN}} = o(1)$ and thus $\epsilon \to 0$ as $n \to \infty$.

Theorem 1 generalizes and strengthens a similar result on the SN estimator developed in (Kuzborskij and Szepesvári, 2019, Theorem 7). Compared to the latter work, there are two key differences: First, our bound is designed for the contextual bandit setting as opposed to the context-free setting considered in (Kuzborskij and Szepesvári, 2019). Second, our bound is significantly tighter, due to a novel application of Harris' inequality to control the bias of the estimator.

Computation of the bound. Although the bound can be computed exactly, computing expectations can be expensive for large action spaces (exponential in K). In this paper we compute an approximation to all expectations in V^{SN}, U^{SN}, and B using Monte-Carlo simulations, as presented in Algorithm 1, and its Python implementation is presented in Appendix D. The algorithm simply updates averages, $\tilde{V}_t^{\text{SN}}$, $\tilde{U}_t^{\text{SN}}$, and $\tilde{B}_t$, over rounds $t = 1, 2, \ldots$, where in each round, a fresh tuple of weights $(W_i')_i$ is sampled from $\pi_b(\cdot|X_1) \times \cdots \times \pi_b(\cdot|X_n)$. The simulation needs to be run until we obtain good enough estimates for the quantities V^{SN},U^{SN}, and B. This can be checked via standard empirical concentration bounds on the estimation errors, such as the empirical Bernstein's inequality (Mnih et al., 2008; Maurer and Pontil, 2009). For example, denoting the sample variance of $\tilde{V}_t^{\text{SN}}$ by $\widehat{\mathrm{Var}}(\tilde{V}_t^{\text{SN}})$, stopping the simulation when

$$\sqrt{2\widehat{\mathrm{Var}}(\tilde{V}_t^{\text{SN}})/t} + 14x/(3t-3) \le \varepsilon \qquad (2)$$

Algorithm 1 Computation of estimates for a variance proxy $V^{\text{SN}}, U^{\text{SN}}$. Scalar operations are understood pointwise when applied to vectors.

Input: observed context-action pairs $S = (X_i, A_i))_{i=1}^n$, behavior / target policy π_b / π
Output: Estimate of a variance proxy $\tilde{V}^{\text{SN}}$ to be used in Theorem 1

$\boldsymbol{W} \leftarrow \left[\frac{\pi(A_1|X_1)}{\pi_b(A_1|X_1)}, \dots, \frac{\pi(A_n|X_n)}{\pi_b(A_n|X_n)}\right]$
$\bar{\boldsymbol{W}} \leftarrow [W_1, W_1 + W_2, \dots, W_1 + \dots + W_n]$
$\boldsymbol{V}, \boldsymbol{U} \leftarrow [0, \dots, 0] \in \mathbb{R}^n$, $\tilde{Z}^{\text{inv}} \leftarrow 0$
$t \leftarrow 1$
repeat
  $\boldsymbol{W}' \leftarrow \left[\frac{\pi(A_1'|X_1)}{\pi_b(A_1'|X_1)}, \dots, \frac{\pi(A_n'|X_n)}{\pi_b(A_n'|X_n)}\right]$ where $\boldsymbol{A}' \sim \pi_b(\cdot|X_1) \times \dots \times \pi_b(\cdot|X_n)$
  Sample $\boldsymbol{W}'', \boldsymbol{W}'''$ as independent copies of $\boldsymbol{W}'$
  $\bar{\boldsymbol{W}}'^{\text{rev}} \leftarrow [\sum_{i=1}^n W_i', \sum_{i=2}^n W_i', \dots, W_n']$
  $\boldsymbol{Z} \leftarrow \bar{\boldsymbol{W}}_{1:n-1} + \bar{\boldsymbol{W}}'^{\text{rev}}_{2:n}$ ▷ Partially simulated sums of weights
  $\boldsymbol{V} \leftarrow (1 - \frac{1}{t})\boldsymbol{V} + \frac{1}{t}\left(\frac{\boldsymbol{W}}{\boldsymbol{Z}}\right)^2$
  $\boldsymbol{U} \leftarrow (1 - \frac{1}{t})\boldsymbol{U} + \frac{1}{t}\left(\frac{\boldsymbol{W}''}{W_1'' + \dots + W_n''}\right)^2$
  $\tilde{Z}_t^{\text{inv}} \leftarrow (1 - \frac{1}{t})\tilde{Z}_t^{\text{inv}} + \frac{1}{t} \cdot \frac{1}{W_1''' + \dots + W_n'''}$
  $\tilde{V}_t^{\text{SN}} \leftarrow \boldsymbol{V}\boldsymbol{1}$, $\tilde{U}_t^{\text{SN}} \leftarrow \boldsymbol{U}\boldsymbol{1}$
  $t \leftarrow t + 1$
until Convergence of $\tilde{V}_t^{\text{SN}}$, $\tilde{U}_t^{\text{SN}}$, and $\tilde{Z}_t$ (see main text and Eq. (2))
$\tilde{B}_t \leftarrow \min\left\{1, \frac{1}{n\tilde{Z}_t^{\text{inv}}}\right\}$
return $\tilde{V}_t^{\text{SN}}$, $\tilde{U}_t^{\text{SN}}$, $\tilde{B}_t$

holds, guarantees that the simulation error $|V^{\text{SN}} - \tilde{V}_t^{\text{SN}}|$ is bounded by ε w.p. at least $1 - e^{-x}$ (the stopping conditions for the other quantities are deferred to Appendix A).[1] Then, denoting by T_ε the number of iterations until stopping, we can use Theorem 1 with V^{SN} replaced by $\tilde{V}_{T_\varepsilon}^{\text{SN}} + \varepsilon$. Of course, each application of the convergence tests needs to be combined with Theorem 1 through a union bound: for example, verifying the convergence for all three variables every 2^k steps for $k = 1, 2, \dots$ means that the final bound on the value holds w.p. at least $1 - (2 + 3\log_2(T_\varepsilon))e^{-x}$.

4 Proof of Theorem 1

We start with the decomposition of $v(\pi) - \hat{v}^{\text{SN}}(\pi)$ as

$$\underbrace{v(\pi) - \mathbb{E}\left[v(\pi) \mid X_1^n\right]}_{\text{Concentration of contexts}} + \underbrace{\mathbb{E}\left[v(\pi) \mid X_1^n\right] - \mathbb{E}\left[\hat{v}^{\text{SN}}(\pi) \mid X_1^n\right]}_{\text{Bias}} + \underbrace{\mathbb{E}\left[\hat{v}^{\text{SN}}(\pi) \mid X_1^n\right] - \hat{v}^{\text{SN}}(\pi)}_{\text{Concentration}} .$$

Each paragraph below focuses respectively on the *Concentration*, *Bias* and *Concentration of contexts* term.

Concentration. We use a conditioned form of the concentration inequality of Kuzborskij and Szepesvári (2019, Theorem 1) that we restate below without a proof.[2] The form of the result is slightly different from its original version to better suit our needs: the version stated here uses a filtration and eventually we use this with $\mathcal{F}_0$ defined as the σ-algebra generated by the contexts X_1^n.

Theorem 2. *Let $(\mathcal{F}_i)_{i=0}^n$ be a filtration and let $S = (Y_1, \dots, Y_n)$ be a sequence of random variables such that the components of S are independent given $\mathcal{F}_0$ and $(Y_k)_k$ is $(\mathcal{F}_k)_k$-adapted. Then, for any $x \geq 2$ and $y > 0$ we have with probability at least $1 - e^{-x}$,*

$$|f(S) - \mathbb{E}[f(S) \mid \mathcal{F}_0]| \leq \sqrt{2(V + y)\left(x + \ln\left(\sqrt{1 + V/y}\right)\right)}$$

where $V = \mathbb{E}\left[\sum_{k=1}^n (f(S) - f(S^{(k)}))^2 \;\middle|\; Y_1, \dots, Y_k\right]$ with $S^{(k)}$ being S with its kth element replaced with an independent copy of Y_k.

We apply the inequality with $f = \hat{v}^{\text{SN}}$, $S = ((W_1, R_1), \dots, (W_n, R_n))$ and $\mathcal{F}_k$ being the σ-algebra generated by X_1^k; then $((W_k, R_k))_k$ is $(\mathcal{F}_k)_k$-adapted, and taking $y = \mathbb{E}[V^{\text{SN}}|X_1^n]$, we get that for any $x \geq 2$, w.p. at least $1 - e^{-x}$,

$$\mathbb{E}[\hat{v}^{\text{SN}}(\pi) \mid X_1^n] - \hat{v}^{\text{SN}}(\pi) \geq -\epsilon \tag{3}$$

where ϵ is defined in Theorem 1, and we also used that $V \leq V^{\text{SN}}$ (see Proposition 5 in Appendix B.1).

Bias. Now we turn our attention to the *bias* term. Let $v(\pi|x)$ denote the value of a policy given a fixed context $x \in \mathcal{X}$: $v(\pi|x) = \sum_{a \in [K]} \pi(a|x) r(x, a)$. Then, since $A_k \sim \pi_b(\cdot|X_k)$,

$$\mathbb{E}\left[\sum_{k=1}^n W_k R_k \;\middle|\; X_1^n\right] = \sum_{k=1}^n \mathbb{E}\left[\frac{\pi(A_k|X_k)}{\pi_b(A_k|X_k)} R_k \;\middle|\; X_k\right] \tag{4}$$

$$= \sum_{k=1}^n \sum_{a \in [K]} \pi(a|X_k) r(X_k, a) = \sum_{k=1}^n v(\pi|X_k) . \tag{5}$$

To relate the above to the expectation of an SN estimator we use Harris' inequality, stated below for completeness. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *non-decreasing* (non-increasing) if it is non-decreasing (non-increasing) in each variable while keeping all other variables fixed at any value.

[1] The parameter ε needs to be specified by the user. A typical choice is $\varepsilon = 1/n$ since $V^{\text{SN}} \geq 1/n$ a.s.

[2] The proof of Kuzborskij and Szepesvári (2019) can be applied almost exactly with minimal, trivial changes.

Theorem 3 (Harris' inequality (Boucheron et al., 2013, Theorem 2.15)). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a non-increasing and $g : \mathbb{R}^n \to \mathbb{R}$ be a non-decreasing function. Then for real-valued random variables $(X_1, \dots, X_n)$ independent from each other, we have $\mathbb{E}[f(X_1,\dots,X_n)g(X_1,\dots,X_n)] \le \mathbb{E}[f(X_1,\dots,X_n)]\,\mathbb{E}[g(X_1,\dots,X_n)]$.*

By noting that $(w_1,\dots,w_n) \to \sum_i w_i R_i$ is non-decreasing a.s. and $(w_1,\dots w_n) \to (\sum_i w_i)^{-1}$ is non-increasing (since the R_i are non-negative), we apply Harris' inequality combined with Eq. (5) to get the desired upper bound

$$\begin{aligned}\mathbb{E}\left[\hat{v}^{\mathrm{SN}}(\pi) \mid X_1^n\right] &= \mathbb{E}\left[\frac{\sum_{k=1}^n W_k R_k}{\sum_{k=1}^n W_k} \,\middle|\, X_1^n\right] \qquad (6)\\ &\le \mathbb{E}\left[\frac{1}{\sum_{k=1}^n W_k} \,\middle|\, X_1^n\right]\left(\sum_{k=1}^n v(\pi|X_k)\right) .\end{aligned}$$

Concentration of contexts. All that is left to do is to account for the randomness of contexts. Since $(v(\pi|X_k))_{k\in[n]}$ are independent and they take values in the range $[0,1]$, by Hoeffding's inequality we have for $x \ge 0$, w.p. at least $1-e^{-x}$, that $\sum_{k=1}^n v(\pi|X_k) - nv(\pi) \le \sqrt{nx/2}$. Hence we bound the bias term as:

$$\begin{aligned}&v(\pi) - \mathbb{E}\left[\hat{v}^{\mathrm{SN}}(\pi) \mid X_1^n\right]\\ &\overset{(6)}{\ge} v(\pi) - \mathbb{E}\left[\frac{1}{\sum_{k=1}^n W_k} \,\middle|\, X_1^n\right]\sum_{k=1}^n v(\pi|X_k)\\ &\ge v(\pi)\left(1 - \mathbb{E}\left[\frac{n}{\sum_{k=1}^n W_k} \,\middle|\, X_1^n\right]\right) - \mathbb{E}\left[\frac{\sqrt{nx/2}}{\sum_{k=1}^n W_k} \,\middle|\, X_1^n\right]\end{aligned}$$

Combining the bias bound above with the concentration term Eq. (3) through the union bound we get, w.p. at least $1-2e^{-x}$, that

$$\begin{aligned}v(\pi) - \hat{v}^{\mathrm{SN}}(\pi) \ge v(\pi)\left(1 - \mathbb{E}\left[\frac{n}{\sum_{k=1}^n W_k} \,\middle|\, X_1^n\right]\right)\\ - \mathbb{E}\left[\frac{\sqrt{nx/2}}{\sum_{k=1}^n W_k} \,\middle|\, X_1^n\right] - \epsilon\ .\end{aligned}$$

Noticing that $v(\pi) \ge 0$ and rearranging gives

$$\begin{aligned}v(\pi) &\ge \left(\underbrace{\mathbb{E}\left[\frac{n}{\sum_{k=1}^n W_k} \,\middle|\, X_1^n\right]^{-1}}_{B'}\left(\hat{v}^{\mathrm{SN}}(\pi) - \epsilon\right) - \sqrt{\frac{x}{2n}}\right)_+\\ &\ge \left(\min\{1, B'\}\left(\hat{v}^{\mathrm{SN}}(\pi) - \epsilon\right)_+ - \sqrt{\frac{x}{2n}}\right)_+\end{aligned}$$

because $(ab-c)_+ = (a(b)_+ - c)_+ \ge (a'(b)_+ - c)_+$ for $a \ge a' \ge 0, b \in \mathbb{R}, c \ge 0$.

5 Related work and baseline confidence intervals

The benefits of using confidence bounds in off-policy evaluation and learning has been recognized in a number of works (Bottou et al., 2013; Thomas et al., 2015a; Swaminathan and Joachims, 2015a,b). Arguably, the standard tool in off-policy evaluation is the IW estimator that originates from the sampling literature (Owen, 2013). However, it has a high variance when the weights have a heavy-tailed distribution. There has been many attempts to stabilize this estimator, including truncation (Ionides, 2008; Thomas et al., 2015b; Bottou et al., 2013) or smoothing (Vehtari et al., 2015). These more stable estimators admit confidence intervals manifesting a bias-variance trade-off, but tuning the truncation or smoothing process is a hard problem on its own, lacking good practical solutions, as discussed in (Gilotte et al., 2018). One approach is to tune the level of truncation depending on the data (e.g. by looking at importance weight quantiles) (Bottou et al., 2013), however, this does not guarantee that the resulting estimator is unbiased. Another popular technique is tuning the truncation level on a hold-out sample (Thomas et al., 2015a; Swaminathan and Joachims, 2015c). A closely related approach to truncation is smoothing, and in this paper we compare against an asymptotically unbiased, smoothed version of IW (tuned in a data-agnostic way), described in Section 5.1.

In contrast, SN, has a low variance in practice and good concentration properties even when the distribution of the weights is (moderately) heavy-tailed. Asymptotic concentration results were already mentioned by Hesterberg (1995) and polynomial (low-probability) finite-time bounds were explored by Metelli et al. (2018). An alternative source of variance in IW is due to the randomness of the rewards; a popular method mitigating its effect is the so-called Doubly-Robust (DR) estimator (Dudík et al., 2014), further improved by Farajtabar et al. (2018), and stabilized by truncations in (Wang et al., 2017; Su et al., 2019a,b). DR and IW can be more generally and optimally mixed as in (Kallus, 2018) who prove asymptotic MSE error bounds. Unfortunately, to the best of our knowledge, many of those works present "sanity-check" bounds (e.g., verifying asymptotic lack of bias), which are practically uncomputable for problems like ours. A notable exception is the family of bounds with the aforementioned truncation: in this paper, for completeness, we present finite-sample confidence bounds for such stabilized estimators (see Section 5.1), which we use as additional baselines.

Finally, a somewhat different approach compared to all of the above was recently explored in (Karampatziakis et al., 2019) based on the *empirical likelihood* approach. Their estimator and the corresponding lower bound on the value relies on solving a convex optimization problem. However, in contrast to the above works, their bound only holds asymptotically, that is, in probability as $n \to \infty$. In the following section we discuss it in detail and compare in the forthcoming experiments.

5.1 Baseline confidence intervals

We derive high-probability lower bounds for the stabilized IW-λ and DR-λ estimators using the same technique as above. Proofs for all statements in this section are given in Appendix B.

IW-λ. Truncation of importance weights is a standard stabilization technique used to bound moments of the IW estimator (Bottou et al., 2013; Swaminathan and Joachims, 2015b). Here we focus on a closely related (albeit theoretically more appealing due to its smoothness) λ-*corrected* version of IW, $\hat{v}^{\text{IW-}\lambda}$, where instead of truncation we add a corrective parameter to the denominator of the importance weight, that is, $W_i^\lambda = \pi(A_k|X_k)/\left(\pi_b(A_k|X_k) + \lambda\right)$ for some $\lambda > 0$ (note that $W_i^\lambda \le \min(W_i, 1/\lambda)$). This ensures that weights are bounded, and setting $\lambda = 1/\sqrt{n}$ ensures that the estimator is asymptotically unbiased. Exploiting this fact, we prove the following confidence bound based on the empirical Bernstein's inequality (Maurer and Pontil, 2009) (also presented in Appendix B.3 for completeness).

Proposition 1. *For the IW-λ estimator we have with probability at least $1 - 3e^{-x}$, for $x > 0$,*

$$v(\pi) \ge \hat{v}^{\text{IW-}\lambda}(\pi) - \sqrt{\frac{2x}{n}\operatorname{Var}^{\text{IW-}\lambda}(X_1^n)} - \frac{7x}{3\lambda(n-1)} - \operatorname{Bias}^{\text{IW-}\lambda}(X_1^n) - \sqrt{\frac{x}{2n}}$$

where $\operatorname{Var}^{\text{IW-}\lambda}$ and $\operatorname{Bias}^{\text{IW-}\lambda}$ are, respectively, the empirical variance and bias of the estimator, defined in the full statement of the proposition in Appendix B.3.

DR-λ (Dudík et al., 2011; Farajtabar et al., 2018; Su et al., 2019b) combines a direct model estimator and IW, finding a compromise that should behave like IW with a reduced variance. As in the case of IW, we introduce a λ-corrected version of DR, $\hat{v}^{\text{DR-}\lambda}$ (given formally in Appendix B.4), where importance weights are replaced with W_i^λ. This allows to prove the following bound:

Proposition 2. *For the DR-λ estimator defined w.r.t. a fixed $\eta : \mathcal{X} \times [K] \to [0,1]$ we have with probability at least $1 - 3e^{-x}$, for $x > 0$,*

$$v(\pi) \ge \hat{v}^{\text{DR-}\lambda}(\pi) - \sqrt{\frac{2x}{n}\operatorname{Var}^{\text{DR-}\lambda}(X_1^n)} - \frac{7}{3}\left(1 + \frac{1}{\lambda}\right)\frac{x}{n-1} - \operatorname{Bias}^{\text{DR-}\lambda}(X_1^n) - \sqrt{\frac{x}{2n}}\,.$$

where $\operatorname{Var}^{\text{DR-}\lambda}$ and $\operatorname{Bias}^{\text{DR-}\lambda}$ are, respectively, the variance and bias estimates defined in Appendix B.4.

As before, setting $\lambda = 1/\sqrt{n}$ ensures that DR-λ is asymptotically unbiased.

Chebyshev-SN. A Chebyshev-type confidence bound for SN can also be proved relying on the fact that the moments of SN are bounded. We present this result as a (naive) alternative approach to the more involved one proposed in this work. This idea was explored in the context of Markov decision processes by Metelli et al. (2018).

Proposition 3. *With probability at least $1 - 3e^{-x}$ for $x > 0$,*

$$v(\pi) \ge \frac{N_x}{n}\left(\hat{v}^{\text{SN}}(\pi) - \sqrt{\frac{\sum_{k=1}^n \mathbb{E}[W_k^2|X_k]}{N_x^2}\, e^x}\right) - \sqrt{\frac{x}{2n}}\,,$$

where $N_x = n - \sqrt{2x\sum_{k=1}^n \mathbb{E}\left[W_k^2 \mid X_1^n\right]}\,.$

Empirical Likelihood (EL) esimator. The EL estimator for off-policy evaluation introduced by Karampatziakis et al. (2019) comes with *asymptotic* confidence intervals: for any error probability $\delta > 0$, the coverage probability of the confidence interval tends to $1 - \delta$ as the sample size $n \to \infty$. In particular, the EL estimator is based on a Maximum Likelihood Estimator (MLE) $\hat{v}^{\text{MLE}}(\pi) = \boldsymbol{W}^\top \boldsymbol{Q}^{\text{MLE}} \boldsymbol{R}$ where $\boldsymbol{W} = [W_1, \dots, W_n]^\top$, $\boldsymbol{R} = [R_1, \dots, R_n]^\top$, and the matrix $\boldsymbol{Q}^{\text{MLE}}$ is a solution of some (convex) empirical maximum likelihood optimization problem which can be solved efficiently. The empirical likelihood theory of Owen (2013) provides a way to get a slightly different estimator that comes with asymptotic confidence intervals. More precisely, assuming that $\boldsymbol{Q}^{\text{POP}} \succeq 0$ is some matrix satisfying $v(\pi) = \boldsymbol{W}^\top \boldsymbol{Q}^{\text{POP}} \boldsymbol{R}$, the asymptotic theorem of empirical likelihood (Owen, 2013) gives an asymptotic identity

$$\lim_{n\to\infty} \mathbb{P}\left(\operatorname{tr}(\ln \boldsymbol{Q}^{\text{MLE}}) - \operatorname{tr}(\ln \boldsymbol{Q}^{\text{POP}}) \le \tfrac{1}{2}\chi^2_{q:1-\delta}\right) = 1 - \delta$$

for any error probability δ where $\chi^2_{q:1-\delta}$ is the $1-\delta$ quantile of a χ^2 distribution with one degree-of-freedom. Based on this (Karampatziakis et al., 2019) estimated the value $v(\pi)$ by solving the optimization problem

$$\min_{\boldsymbol{Q}\succeq 0} \boldsymbol{W}^\top \boldsymbol{Q} \boldsymbol{R} \quad s.t.\ \boldsymbol{W}^\top \boldsymbol{Q}\boldsymbol{1} = 1, \boldsymbol{1}^\top \boldsymbol{Q}\boldsymbol{1} = 1,$$
$$x_n + \operatorname{tr}(\ln \boldsymbol{Q}) \ge \operatorname{tr}(\ln \boldsymbol{Q}^{\text{MLE}})$$

where x_n is a specific sequence satisfying $x_n \to \chi^2_{q:1-\delta}$ as $n \to \infty$. In our experiments we use the implementation of the authors of Karampatziakis et al. (2019) available in (Mineiro and Karampatziakis, 2019). More precisely, we use their function `asymptoticconfidenceinterval`, which returns the lower and upper bounds of the confidence interval to run our experiments (we use only the lower bound as returned value).

6 Experiments

Our experiments aim to verify two hypotheses: (i) the Efron-Stein Lower Bound (ESLB) of Theorem 1 is empirically tighter than its main competitors (discussed in Section 5), which is assessed through the value-gap and experiments for the best policy selection problem; (ii) best policy selection based on confidence bounds is superior to selection using just the bare estimators. Henceforth, we will be concerned mainly with comparisons between confidence bounds. Therefore, most estimators which do not come with practically computable confidence bounds are outside of the scope of the following experiments. Our experimental process is inspired by previous work on off-policy evaluation (Dudík et al., 2011, 2014; Farajtabar et al., 2018).

6.1 Policies and datasets

We summarize our experimental setup here; all details can be found in Appendix C. We consider a contextual bandit problem such that for every context $\boldsymbol{x}$ there is a single action with reward 1, denoted by $\rho(\boldsymbol{x}) \in [K]$, and the reward of all other actions is 0. This setup is closely related to multiclass classification problems: treating feature vectors as contexts (arriving sequentially) and the predicted label as the action, and defining the reward to be 1 for a correct prediction and 0 otherwise, we arrive at the above bandit problem; this construction has been used in off-policy evaluation (see, e.g., (Bietti et al., 2018)).

Throughout we consider Gibbs policies: we define an ideal Gibbs policy as $\pi^{\text{ideal}}(y \mid \boldsymbol{x}) \propto e^{\frac{1}{\tau}\mathbb{I}\{y=\rho(\boldsymbol{x})\}}$, where $\mathbb{I}$ denotes the indicator function[3] and $\tau > 0$ is a temperature parameter. The smaller τ is, the more peaky is the distribution on the predicted label. To create mismatching policies, we consider a *faulty* policy type for which the peak is shifted to another, wrong action for a set of faulty actions $F \subset [K]$ (i.e., if $\rho(\boldsymbol{x}) \in F$, the peak is shifted by 1 cyclically). In the following we consider faulty behavior policies, while one among the target policies is *ideal*.

Motivated by the large body of literature on off-policy learning (Swaminathan and Joachims, 2015a,c; Joachims et al., 2018), which considers the problem of directly learning a policy from logged bandit feedback, we also consider trained target policies: the policies have a parametric form $\pi^{\hat{\boldsymbol{\Theta}}_{\text{IW}}}(k|\boldsymbol{x}) \propto e^{\frac{1}{\tau}(\hat{\boldsymbol{\theta}}_{\text{IW}})_k^\top \boldsymbol{x}}$ and $\pi^{\hat{\boldsymbol{\Theta}}_{\text{SN}}}(k|\boldsymbol{x}) \propto e^{\frac{1}{\tau}(\hat{\boldsymbol{\theta}}_{\text{SN}})_k^\top \boldsymbol{x}}$, and their parameters are learned by respectively maximizing the empirical values $\hat{v}^{\text{IW}}$ and $\hat{v}^{\text{SN}}$ (through gradient descent), to imitate parameter fitting w.r.t. these estimators (see Appendix C.1 for details).

Some of our experiments require a precise control of the distribution of the contexts, as well as of the sample size. To accomplish this, we generate *synthetic datasets* from an underlying multiclass classification problem through the scikit-learn function `make_classification()`.

Finally, 8 *real* multiclass classification datasets are chosen from OpenML (Dua and Graff, 2017) (see Table 3 in Appendix C) with classification tasks of various sizes, dimensions and class imbalance.

6.2 Empirical tightness analysis: comparison of existing bounds

ESLB of Theorem 1, Chebyshev, λ-DR and λ-IW take as input a parameter $\delta \in (0, 1)$ that controls the theoretical error probability of the obtained lower bound (the coverage probability is $1 - \delta$). However, there is usually a gap between this theoretical value and the actual empirical coverage obtained in practice. We fix a synthetic problem with size $n = 10^4$, $\tau_b = 0.3$ (Gibbs behavior policy, with two faulty actions) and $\tau_t = 0.3$ (Gibbs target policy, with one different faulty action). As an indication of the difficulty of the problem, the effective sample size (see Sec. 2) here is $n_{\text{eff}} = 655$, which is an order of magnitude smaller than n (a moderate policy mismatch), but should allow a reasonable estimation.

For each value of δ, we repeated the same experiments 100 times, regenerating a new but identically distributed logged dataset and computing the estimates. The empirical distribution of the lower bound $\hat{v}_n(\pi)$ and the width $v(\pi) - \hat{v}_n(\pi)$ at $\delta = 0.05$ (the δ-value used in the experiments) are shown in Figure 2. We can observe that all lower bounds have a positive distance to the true value *for any* $\delta \in (0, 1)$. This means that none of the lower bounds is tight at this sample size. Nonetheless, ESLB is considerably tighter for low error probabilities ($\delta \le 0.1$). It is expected that this tightness is a key ingredient to make more accurate decisions.

Nonetheless, in all 100 repetitions, the lower bound is *always* below the true value, meaning that the empirical coverage of the estimators is $1 > 1 - \delta$ for all δ. This

[3] For an event E, $\mathbb{I}\{E\} = 1$ if E holds, and 0 otherwise.

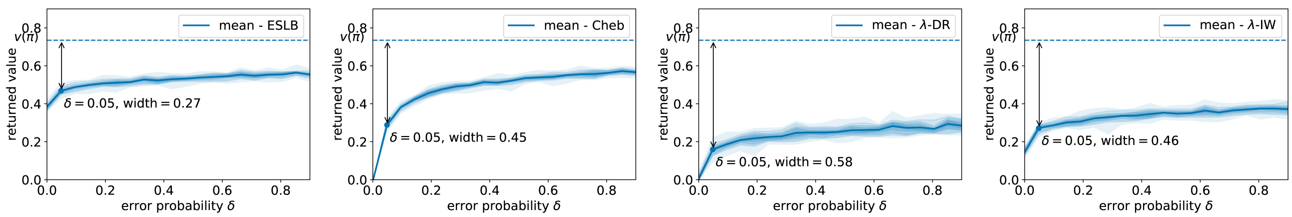

Figure 2: Empirical tightness: analysis on synthetic data. From left to right: ESLB, Chebyshev, λ-DR and λ-IW. 100 runs for each value of δ.

Table 1: Average test rewards for a 5-action problem on a synthetic benchmark. Symbol $-\infty$ indicates that no policy can be selected, since the confidence bound is always vacuous. Here the behaviour policy is faulty on two actions and the target policies are: ideal, fitted on $\hat{v}^{\text{IW}}$, and fitted on $\hat{v}^{\text{SN}}$.

Sample size	5000	10000	20000
ESLB	**1.000 ± 0.004**	**1.000 ± 0.004**	**1.000 ± 0.004**
λ-IW	0.710 ± 0.443	0.837 ± 0.356	0.900 ± 0.238
λ-DR	$-\infty$	0.837 ± 0.356	0.941 ± 0.140
Cheb-SN	$-\infty$	$-\infty$	$-\infty$
DR	0.896 ± 0.187	0.871 ± 0.279	0.951 ± 0.122
Emp.Lik. (Karampatziakis et al., 2019)	0.844 ± 0.312	0.819 ± 0.354	0.883 ± 0.293
Best policy on the test set	1.000 ± 0.004	1.000 ± 0.004	1.000 ± 0.004

Table 2: Average test rewards on a real benchmark. Here the behaviour policy is faulty on two actions and the target policies are: ideal, fitted on $\hat{v}^{\text{IW}}$, and fitted on $\hat{v}^{\text{SN}}$.

Name Size	Yeast 1484	PageBlok 5473	OptDigits 5620	SatImage 6435	isolet 7797	PenDigits 10992	Letter 20000	kropt 28056
ESLB	**0.90 ± 0.27**	**0.91 ± 0.27**	**0.91 ± 0.26**	**0.91 ± 0.26**	**0.90 ± 0.27**	**0.91 ± 0.27**	**0.91 ± 0.27**	**0.91 ± 0.27**
λ-IW	**0.91 ± 0.26**	**0.91 ± 0.27**	0.72 ± 0.40	0.70 ± 0.39	0.75 ± 0.40	**0.9 ± 0.27**	**0.90 ± 0.27**	**0.90 ± 0.27**
λ-DR	$-\infty$	**0.91 ± 0.27**	$-\infty$	$-\infty$	**0.90 ± 0.27**	**0.91 ± 0.26**	**0.91 ± 0.27**	**0.91 ± 0.27**
Cheb-SN	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	**0.90 ± 0.27**	$-\infty$
DR	0.52 ± 0.31	0.75 ± 0.36	0.68 ± 0.32	0.62 ± 0.39	0.21 ± 0.29	0.79 ± 0.31	0.63 ± 0.28	**0.91 ± 0.27**
Emp.Lik. (Karampatziakis et al., 2019)	0.31 ± 0.32	0.66 ± 0.40	0.28 ± 0.35	0.63 ± 0.40	0.21 ± 0.29	0.54 ± 0.42	0.24 ± 0.33	0.71 ± 0.29

is not the case for EL, as shown in Figure 3. This simulation highlights two interesting facts about EL. We can see that the returned lower bound is always very close to the true value, which should be a perfect property for our selection problem. But unfortunately, the estimated lower bound also suffers from a quite large variance, resulting an empirical coverage below $1-\delta$ when δ is small. This is likely the reason for the inferior performance of EL on our real data experiments, presented in the next section.

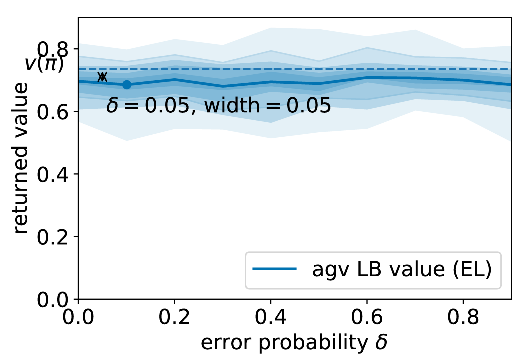

Figure 3: Empirical tightness of the EL lower bound estimator. The returned value is on average very close to the true value, but the empirical coverage is below $1-\delta$ for small values of δ.

6.3 Best-policy selection

We evaluate all estimators on the best-policy selection problem (see Figure 1). For all experiments, we use a behavior policy with two faulty actions and temperature $\tau = 0.2$. The set of candidate target policies is $\pi^{\text{ideal}}, \pi^{\hat{\Theta}_{\text{IW}}}, \pi^{\hat{\Theta}_{\text{SN}}}$ with temperature $\tau = 0.2$ for synthetic and (almost deterministic) $\tau = 0.1$ for real datasets. The performance of a selected policy is its average reward collected on a separate test sample over 10 independent trials ($5 \cdot 10^4$ examples in synthetic case).

In all cases, we set the error probability $\delta = 0.01$ except for OptDigits and SatImage where it is $\delta = 0.05$.[4] Since the best-policy selection problem relies on the comparison of $N = 3$ confidence bounds, by application of the union bound, the final result holds with a lower probability, i.e. δ is replaced by δ/N.

Table 1 presents the results on a synthetic benchmark. ESLB perfectly returns the best policy on all trials while other estimators fail at least once.

Results on real data, summarized in Table 2, show that ESLB also achieves the best average performance in all cases, but other confidence-based methods turn out to be reasonable alternatives — especially for large samples. This is in contrast with the pure estimators DR and EL (Karampatziakis et al., 2019) that consistently make selection mistakes and show significantly lower performance on average on all datasets. This confirms our hypothesis that selection must be performed by a confidence-based scoring function.

The detailed decomposition of bounds in both Tables 1 and 2 into constituent terms (concentration, bias, etc.) can be found in Appendix E.

6.4 Discussion

Confidence bound for the off-policy selection. In this paper we demonstrated that the high probability lower bound on the value can be used effectively to select the best policy. We considered only a finite set of N target policies, however bounds on the value can be combined using a *union bound* to provide a bound on the value of the selected policy. In such case Theorem 1 would hold with a slight modification, where we would pay a $x + \ln(N)$ term in place of x. In other words, for $\pi^\star \in \max_{\pi \in \{\pi_1, \dots, \pi_N\}} \mathrm{ESLB}_{x+\ln(N)}(\pi)$ where $\mathrm{ESLB}_x(\pi)$ denotes a right hand side of the bound in Theorem 1, with probability at least $1 - 2e^{-x}$ we have $v(\pi^\star) \geq \mathrm{ESLB}_{x+\ln(N)}(\pi^\star)$.

Naturally, one might wonder how to extend the above to the uncountable class of policies where one could maximize the lower bound over policies: Such optimization is known as the *off-policy policy optimization.* Several works explored such possibility for SN estimator by showing bounds in the Probably Approximately Correct (PAC) framework (Swaminathan and Joachims, 2015b; Athey and Wager, 2021). PAC bounds are known to be conservative in general: In our case they would hold with respect to the *worst* policy in a given policy class. The concentration inequality we build upon in this work (Theorem 2) is also available in a so-called PAC-Bayes formulation (Kuzborskij and Szepesvári, 2019), where the inequality holds for all probability measures over a given policy space simultaneously. Such PAC-Bayes formulation offers an alternative to the PAC formulation, while it was shown to be less conservative numerically (Dziugaite and Roy, 2017).

Upper bound on the value. Our concentration inequality is symmetrical and should in principle allow to obtain an upper-bound on the value as well. Note however that for our use case we aim at finding the policy with highest value and so for that task the lower bound suffices. Nonetheless, upper-bounds could be computed to get a sense of the *width* of the confidence interval and thus of its tightness.

7 Conclusions and future work

We derived the first high-probability truncation-free finite-sample confidence bound on the value of a contextual bandit policy that employs SN estimator, which turned into a practical and principled off-policy selection method. The sharpness of our bound is due to a careful handling of the empirical variance of SN estimator, but this desirable property comes at the cost of an increased computation complexity. Indeed, Monte-Carlo simulations are needed to obtain the key terms in the bound. These efforts allow us to achieve state-of-the-art performance on a variety of contextual bandit tasks. Nevertheless, we see several future directions for our work.

Off-policy learning. We have demonstrated that ESLB can be used to improve the behavior policy by choosing a better target policy provided there is one in the finite set of candidates. In principle, we can obtain even better improvement by selecting the policy from an uncountable class by maximizing the ESLB –or a similar– e.g. as in Swaminathan and Joachims (2015c).

ESLB for MDPs. Extending our analysis to stateful Markov Decision Processes is a challenging but promising direction. Indeed, off-policy evaluation and learning is a major topic of research in reinforcement learning. In that more general setting, for each learning episode, a policy generates not only a reward but an entire sequence of states, actions and rewards, which eventually characterize its value. We believe that similar techniques as those applied here may allow to obtain high-probability bounds on the policy value, at least in finite-horizon MDPs.

Acknowledgements

We are grateful to Tor Lattimore for many insightful comments.

[4] For those datasets, all confidence intervals were vacuous with $\delta = 0.01$ so we adjusted it to obtain exploitable results.

References

A. Agarwal, S. Basu, T. Schnabel, and T. Joachims. Effective evaluation using logged bandit feedback from multiple loggers. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2017.

L. Bottou, J. Peters, J. Quiñonero Candela, D. X. Charles, M. Chickering, E. Portugaly, D. Ray, P. Y. Simard, and E. Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learing (ICML)*, 2011.

A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(1):1731–1755, 2015a.

P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *Conference on Artificial Intelligence (AAAI)*, 2015a.

P. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learing (ICML)*, pages 2380–2388, 2015b.

A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Conference on Neural Information Processing Systems (NIPS)*, pages 3231–3239, 2015b.

A. M. Metelli, M. Papini, F. Faccio, and M. Restelli. Policy optimization via importance sampling. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 5442–5454, 2018.

T. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

I. Kuzborskij and C. Szepesvári. Efron-Stein PAC-Bayesian Inequalities. arXiv:1909.01931, 2019. URL https://arxiv.org/abs/1909.01931.

A. Kong. A note on importance sampling using standardized weights. University of Chicago, Dept. of Statistics, Tech. Rep, 1992.

V. Elvira, L. Martino, and C. P. Robert. Rethinking the effective sample size. arXiv:1809.04129, 2018.

V. Mnih, C. Szepesvári, and J.-Y. Audibert. Empirical bernstein stopping. In *International Conference on Machine Learing (ICML)*, 2008.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. In *Conference on Computational Learning Theory (COLT)*, 2009.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

A. B. Owen. *Monte Carlo theory, methods and examples*. 2013. URL https://statweb.stanford.edu/~owen/mc/.

E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

A. Gilotte, C. Calauzènes, T. Nedelec, A. Abraham, and S. Dollé. Offline a/b testing for recommender systems. In *International Conference on Web Search and Data Mining*, 2018.

A. Swaminathan and T. Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learing (ICML)*, pages 814–823, 2015c.

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

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

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

Y. Su, L. Wang, M. Santacatterina, and T. Joachims. Cab: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learing (ICML)*, pages 6005–6014, 2019a.

Y. Su, M. Dimakopoulou, A. Krishnamurthy, and M. Dudík. Doubly robust off-policy evaluation with shrinkage. *arXiv preprint arXiv:1907.09623*, 2019b.

N. Kallus. Balanced policy evaluation and learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 8895–8906, 2018.

N. Karampatziakis, J. Langford, and P. Mineiro. Empirical likelihood for contextual bandits. In *NeurIPS 2019 Optimization Foundations for Reinforcement Learning Workshop*, 2019.

P. Mineiro and N. Karampatziakis. Code to reproduce the results in the paper 'empirical likelihood for contextual bandits'. https://github.com/pmineiro/elfcb/blob/master/MLE/MLE/asymptoticconfidenceinterval.py, 2019. Accessed: 2020-09-30.

A. Bietti, A. Agarwal, and J. Langford. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064*, 2018.

T. Joachims, A. Swaminathan, and M. de Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations (ICLR)*, 2018.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Uncertainty in Artificial Intelligence (UAI)*, 2017.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.