Subgaussian and Differentiable Importance Sampling for Off-Policy Evaluation and Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Importance Sampling (IS) is a widely used building block for a large variety of off-policy estimation and learning algorithms. However, empirical and theoretical studies have progressively shown that vanilla IS leads to poor estimations whenever the behavioral and target policies are too dissimilar. In this paper, we analyze the theoretical properties of the IS estimator by deriving a novel anticoncentration bound that formalizes the intuition behind its undesired behavior. Then, we propose a new class of IS transformations, based on the notion of power mean. To the best of our knowledge, the resulting estimator is the first to achieve, under certain conditions, two key properties: (i) it displays a subgaussian concentration rate; (ii) it preserves the differentiability in the target distribution. Finally, we provide numerical simulations on both synthetic examples and contextual bandits, in comparison with off-policy evaluation and learning baselines.

1 Introduction

The availability of historically collected data is a common scenario in many real-world decision-making problems, including medical treatments [e.g., 17, 65], recommendation systems [e.g., 33, 16], personalized advertising [e.g., 3, 58], finance [e.g., 41], and industrial robot control [e.g., 27, 26]. Historical data can be leveraged to address two classes of problems. First, given data collected with a *behavioral* policy, we want to estimate the performance of a different *target* policy. This problem is known as *off-policy evaluation* [Off-PE, 21]. Second, we want to employ the available data to improve the performance of a baseline policy. This latter problem is named *off-policy learning* [Off-PL 14]. Off-policy methods are studied by both the *reinforcement learning* [RL, 56] and *contextual multi-armed bandit* [CMAB, 30] communities. Given its intrinsic simplicity compared to RL, off-policy methods are nowadays well understood in the CMAB framework [e.g., 42, 1, 14, 62]. Among them, the *doubly robust* estimator [DR, 14] is one of the most promising off-policy methods for CMABs. DR combines a *direct method* (DM), in which the reward is estimated from historical data via regression, with an *importance sampling* [IS, 44] control variate.

IS plays a crucial role in the off-policy methods and counterfactual reasoning. However, IS tends to exhibit problematic behavior for general distributions. This is formalized by its *heavy-tailed* properties [39], which prevent the application of exponential concentration bounds [4]. To cope with this issue, typically, corrections are performed on the importance weight including *truncation* [23] and *self-normalization* [SN, 44], among the most popular. Significant results have recently been derived for both techniques [45, 29]. Nevertheless, we believe that the widespread use of IS calls for a better theoretical understanding of its properties and for the design of principled weight corrections.

Defining the desirable properties of an off-policy estimator is a non-trivial task. Some works employed the *mean squared error* (MSE) as an index of the estimator quality [34, 62]. However, controlling the MSE, while effectively capturing the bias-variance trade-off, does not provide any guarantee on the concentration properties of the estimator, which might still display a heavy-tailed behavior [37].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

For this reason, we believe that a more suitable approach is to require that the estimator deviations concentrate at a *subgaussian* rate [12]. Subgaussianity implicitly controls the tail behavior and leads to tight exponential concentration inequalities. Unlike MSE, the probabilistic requirements are non-asymptotic (finite-sample), from which guarantees on the MSE can be derived. While subgaussianity can be considered a satisfactory requirement for Off-PE, additional properties are advisable when switching to Off-PL. In particular, the *differentiability* w.r.t. the target policy parameters is desirable whenever Off-PL is carried out via gradient optimization. For instance, weight truncation, as presented in [45], allows achieving subgaussianity but leads to a non-differentiable objective. Consequently, the optimization phase requires additional care, which sometimes leads to computationally heavy discretizations [45]. On the contrary, the SN estimator is differentiable in the target policy, but fails to achieve subgaussian concentration for general distributions.

In this paper, we take a step towards a better understanding of IS. After having introduced the necessary background (Section 2), we derive an anticoncentration bound for the mean estimation with vanilla IS. We show that polynomial concentration (Chebychev's inequality) is tight in this setting (Section 3). This result formalizes the intuition behind the undesired behavior of these estimators for general distributions. Hence, we propose a class of importance weight corrections, based on the notion of power mean (Section 4). The rationale behind these corrections is to "smoothly shrink" the weights towards the mean, with different intensities. In this way, we mitigate the heavy-tailed behavior and, in the meantime, we exert control over the induced bias. Then, we derive bounds on the bias and variance that allow obtaining an exponential concentration inequality and, under certain conditions, subgaussian concentration (Section 5). Furthermore, the smooth transformation allows preserving the differentiability in the target policy, unlike some existing transformations, like weight truncation. To the best of our knowledge, this is the first IS correction that preserves the differentiablity and is proved to be subgaussian. This correction, however, requires knowledge of a distributional divergence between the target and behavioral policies, which may be unknown or difficult to compute. To this end, we introduce an approach to empirically estimate the correction parameter, preserving the desirable concentration properties (Section 6). After providing a comparative review of the literature (Section 7), we present an experimental study comparing our approach with traditional and modern off-policy baselines on synthetic domains and in the CMAB framework (Section 8). The proofs of the results presented in the main paper can be found in Appendix A.

2 Preliminaries

We start introducing the background about probability, importance sampling and contextual bandits.

Probability We denote with $\mathscr{P}(\mathcal{Y})$ the set of probability measures over a $(\mathcal{Y},\mathfrak{F}_{\mathcal{Y}})$. Let $P\in\mathscr{P}(\mathcal{Y})$, $f:\mathcal{Y}\to\mathbb{R}$ be a function, and $\overline{\mu}_n$ be an estimator for the mean $\mu=\mathbb{E}_{x\sim P}[f(x)]$ obtained with n i.i.d. samples. Suppose that with probability $1-\delta$ it holds that $|\overline{\mu}_n-\mu|\leqslant\sqrt{g(n,\delta)}$. For $\beta>0$, we say that $\overline{\mu}_n$ admits: (i) *polynomial* concentration if $g(n,\delta)=\mathcal{O}\big(1/(n\delta)^{\beta}\big)$; (ii) *exponential* concentration if $g(\delta)=\mathcal{O}\big(\big(\log(1/\delta)/n\big)^{\beta}\big)$; (iii) *subgaussian* concentration if (ii) holds with $\beta=1$ [37]. These cases correspond to Chebyshev's, Bernstein's, and Höeffding's inequalities respectively [4].

Importance Sampling Let $P,Q\in\mathscr{P}(\mathcal{Y})$ admitting p and q as density functions, if $P\ll Q$, i.e., P is absolutely continuous w.r.t. Q, for any $\alpha\in(1,2]$, we introduce the integral: $I_\alpha(P\|Q)=\int_{\mathcal{Y}}p(y)^{\alpha}q(y)^{1-\alpha}\mathrm{d}y$. If $P=Q$ a.s. (almost surely) then $I_\alpha(P\|Q)=1$. $I_\alpha(P\|Q)$ allows defining several divergences, like Rényi [49]: $(\alpha-1)^{-1}\log I_\alpha(P\|Q)$. Let $f:\mathcal{Y}\to\mathbb{R}$ be a function, *(vanilla) importance sampling* [IS, 44] allows estimating the expectation of f under the *target* distribution P, i.e., $\mu=\mathbb{E}_{y\sim P}[f(y)]$, using i.i.d. samples $\{y_i\}_{i\in[n]}$ collected with the *behavioral* distribution Q:

$$\widehat{\mu}_n=\frac{1}{n}\sum_{i\in[n]}\omega(y_i)f(y_i),\qquad\text{where}\qquad\omega(y)=\frac{p(y)}{q(y)},\quad\forall y\in\mathcal{Y}.$$

It is well-known that $\widehat{\mu}_n$ is unbiased, i.e., $\mathbb{E}_{y_i\sim Q}[\widehat{\mu}_n]=\mu$ [44]. If f is bounded, the variance of the estimator can be upper-bounded as $\mathbb{V}\mathrm{ar}_{y_i\sim Q}[\widehat{\mu}_n]\leqslant\frac{1}{n}\|f\|_\infty I_2(P\|Q)$ [39]. More in general, the integral $I_\alpha(P\|Q)$ represents the α-moment of the importance weight $\omega(y)$ under Q.

Contextual Bandits A *contextual multi-armed bandit* [CMAB, 30] is represented by the tuple $\mathcal{C}=(\mathcal{X},\mathcal{A},\rho,p)$, where $\mathcal{X}$ is the set of *contexts*, $\mathcal{A}$ is the finite set of actions (or arms), $\rho\in\mathscr{P}(\mathcal{X})$ is the context distribution, and $p:\mathcal{X}\times\mathcal{A}\to\mathscr{P}(\mathbb{R})$ is the reward distribution. The agent's behavior is

encoded by a *policy* $\pi:\mathcal{X}\to\mathscr{P}(\mathcal{A})$. At each round $t\in\mathbb{N}$, the agent observes a context $x_t\sim\rho$, plays an action $a_t\sim\pi(\cdot|x_t)$, gets the reward $r_t\sim p(\cdot|x_t,a_t)$ and the system moves on to the next round. The *value* of a policy π is given by $v(\pi)=\int_{\mathcal{X}}\rho(x)\sum_{a\in\mathcal{A}}\pi(a|x)\int_{\mathbb{R}}p(r|x,a)r\mathrm{d}r\mathrm{d}x$. We denote with $r(x,a)=\int_{\mathbb{R}}p(r|x,a)r\mathrm{d}r$ the *reward function*. A policy π^* is optimal if it maximizes the value function, i.e., $\pi^*\in\arg\max_{\pi\in\Pi}v(\pi)$, where $\Pi=\{\pi:\mathcal{X}\to\mathscr{P}(\mathcal{A})\}$ is the set of all policies.

Let $\mathcal{D}=\{(x_t,a_t,r_t)\}_{t\in[n]}$ be a dataset of samples collected in a CMAB with a behavioral policy $\pi_b\in\Pi$. The *off-policy evaluation* [Off-PE, 21] problem consists in estimating the value function $v(\pi_e)$ of a target policy $\pi_e\in\Pi$ using the samples in $\mathcal{D}$. The *off-policy learning* [Off-PL, 14] problem consists in estimating an optimal policy $\pi^*\in\Pi$ using the samples in $\mathcal{D}$. The simplest approach to address the Off-PE/Off-PL problem is to learn from $\mathcal{D}$ an estimate $\widehat{r}(x,a)$ of the reward function $r(x,a)$ via regression. This approach is known as *direct method* (DM) and its properties heavily depend on the quality of the estimate $\widehat{r}$. Another approach is to simply apply IS to reweight the samples of $\mathcal{D}$, leading to the *inverse propensity scoring* [IS, 17] estimator. The two approaches are combined in the *doubly-robust* [DR, 14] estimator, in which the DM estimate is corrected with an IS control variate to reduce the variance using the estimated reward $\widehat{r}$ (see also Table 10 in Appendix D).

3 Anticoncentration of Vanilla Importance Sampling

In this section, we analyze the intrinsic limitations of the vanilla IS. It is well-known that under the assumption that for some $\alpha\in(1,2]$ the divergence $I_\alpha(P\|Q)$ is finite and f is bounded, the vanilla IS estimator $\widehat{\mu}_n$ admits *polynomial* concentration, i.e., with probability at least $1-\delta$:[1]

$$|\widehat{\mu}_n-\mu|\leqslant\|f\|_\infty\left(\frac{2^{2-\alpha}I_\alpha(P\|Q)}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}. \tag{1}$$

We now show that the concentration in Equation (1) is tight, by deriving an anticoncentration bound for $|\widehat{\mu}_n-\mu|$; then, we discuss its implications and compare it with previous works.

Theorem 3.1. *There exist two distributions $P,Q\in\mathscr{P}(\mathcal{Y})$ with $P\ll Q$ and a bounded measurable function $f:\mathcal{Y}\to\mathbb{R}$ such that for every $\alpha\in(1,2]$ and $\delta\in(0,e^{-1})$ if $n\geqslant\delta e\max\left\{1,(I_\alpha(P\|Q)-1)^{\frac{1}{\alpha-1}}\right\}$, with probability at least δ it holds that:*

$$|\widehat{\mu}_n-\mu|\geqslant\|f\|_\infty\left(\frac{I_\alpha(P\|Q)-1}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1-\frac{e\delta}{n}\right)^{\frac{n-1}{\alpha}}.$$

First of all, we note the polynomial dependence on the confidence level δ. The bound is vacuous when $I_\alpha(P\|Q)=1$, i.e., when $P=Q$ a.s., since in an on-policy setting and, being the function f bounded, subgaussian concentration bounds (like Höeffding's inequality) hold. In particular, for $\alpha=2$, n and $I_2(P\|Q)$ sufficiently large, the bound has order $\mathcal{O}\big(\sqrt{I_2(P\|Q)/(\delta n)}\big)$, matching Chebyshev's and the existing concentration inequalities for vanilla importance sampling [39, 40].

Our result is of independent interest and applies for general distributions. Previous works considered the MAB [34] and CMAB [62] settings proving *minimax* lower bounds in *mean squared error* (MSE) $\mathbb{E}_{y\sim Q}[(\widehat{\mu}_n-\mu)^2]$. These results differ from ours in several respects. First, we focus on a *specific estimator*, the vanilla one, while those result are minimax. Second, they provide lower bounds to the MSE, while we focus on the deviations in probability.[2] From our probabilistic result, it is immediate to derive an MSE guarantee (Corollary A.1 of Appendix A.1). Finally, they assume that the second moment of the importance weight $I_2(P\|Q)$ is finite, whereas our result allows considering scenarios in which only moments of order $\alpha<2$ are finite.

4 Power-Mean Correction of Importance Sampling

In this section, motivated by the negative result of Theorem 3.1, we look for a weight correction able to achieve subgaussian concentration. Specifically, we introduce a class of corrections based on the notion of *power mean* [6] and we study its properties. Let us start with the following definition.

[1] The original result [40, Theorem 2] was limited to $\alpha=2$ and based on Cantelli's inequality which approaches Chebyshev's when $\delta\to0$. See Theorem A.1 in Appendix A.1, for a proof of Equation (1).

[2] As noted in [36], when the estimator is not well-concentrated around its mean (e.g., in presence of heavy tails), the MSE is not adequate to capture the error and high-probability bounds are more advisable.

s	$\omega_{s,\lambda}(y)$
$-\infty$ (minimum)	$\min\{\omega(y),1\}$
-1 (harmonic)	$\dfrac{\omega(y)}{1-\lambda+\lambda\omega(y)}$
0 (geometric)	$\omega(y)^{1-\lambda}$
1 (arithmetic)	$(1-\lambda)\omega(y)+\lambda$

Table 1: Choices of s for the (λ,s)-corrected importance weight of Definition 4.1.

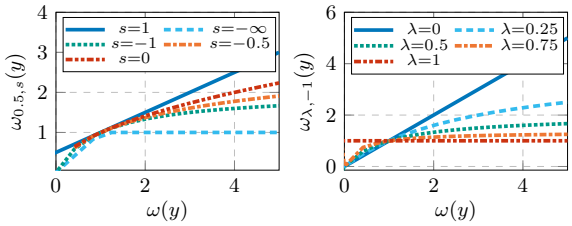

Figure 1: Examples of importance weight corrections of Definition 4.1 for fixed λ (left) and fixed s (right).

Definition 4.1. *Let $P,Q\in\mathscr{P}(\mathcal{Y})$ be two probability distributions such that $P\ll Q$, for every $s\in[-\infty,\infty]$ and $\lambda\in[0,1]$, let $\omega(y)=p(y)/q(y)$, the* (λ,s)-corrected *importance weight is defined as:*

$$\omega_{\lambda,s}(y)=\Big((1-\lambda)\omega(y)^s+\lambda\Big)^{\frac{1}{s}},\quad \forall y\in\mathcal{Y}.$$

The correction can be seen as the weighted *power mean* with exponent s between the vanilla importance weight $\omega(y)$ and 1 with weights $1-\lambda$ and λ respectively.[3] We immediately notice that, regardless of the value of s, for $\lambda=0$, we reduce to the vanilla importance weight $\omega_{0,s}(y)=\omega(y)$ and for $\lambda=1$, we have identically $\omega_{1,s}(y)=1$. Furthermore, the correction is unbiased when $P=Q$ a.s. regardless s and λ. Thus, the correction "smoothly interpolates" between the vanilla weight $\omega(y)$ and its mean under Q, i.e., 1. s and λ govern the "intensity" of the correction in a continuous way. Differently from the truncation [23], this transformation leads to a differentiable weight. Some specific choices of s and λ are reported in Table 1 and in Figure 1. The following result provides a preliminary characterization of the correction, independent of the properties of the two distributions.

Lemma 4.1. *Let $P,Q\in\mathscr{P}(\mathcal{Y})$ be two probability distributions with $P\ll Q$, then for every $\lambda\in[0,1]$ and $y\in\mathcal{Y}$ it holds that:*

(i) if $s\leqslant s'$ then $\omega_{\lambda,s}(y)\leqslant\omega_{\lambda,s'}(y)$;

(ii) if $s<0$ then $\omega_{\lambda,s}(y)\leqslant\lambda^{\frac{1}{s}}$, otherwise if $s>0$ then $\omega_{\lambda,s}(y)\geqslant\lambda^{\frac{1}{s}}$;

(iii) if $s<1$ then $\mathbb{E}_{y\sim Q}[\omega_{\lambda,s}(y)]\leqslant 1$, otherwise if $s>1$ then $\mathbb{E}_{y\sim Q}[\omega_{\lambda,s}(y)]\geqslant 1$.

From point *(ii)* we observe that the corrected weight is bounded from below when $s>0$ and bounded from above when $s<0$. It is well-known that the inconvenient behavior of IS derives from the heavy-tailed properties [39]. Thus, the *arithmetic* correction ($s=1$) performs just a convex combination between the vanilla weight and 1, having no effect on the tail properties. Any correction with $s>1$ increases the weight value, making the tail even heavier. So, if we are looking for subgaussianity, we should restrict our attention to $s<0$, which leads to lighter tails or even bounded weights.

5 Subgaussian and Differentiable Importance Sampling

In this section, we focus on the *harmonic* correction ($s=-1$), which leads to a weight of the form:[4] $\omega_{\lambda,-1}(y)=\frac{\omega(y)}{1-\lambda+\lambda\omega(y)}$. We start analyzing the bias and variance of this class of estimators. Then, we provide an exponential concentration inequality that, under certain circumstances, results to be subgaussian. Finally, we show that the resulting estimator is differentiable in the target distribution. To lighten the notation we neglect the -1 subscript, abbreviating $\widehat{\mu}_\lambda=\widehat{\mu}_{\lambda,-1}$.

Bias and Variance We now derive bounds for the bias and the variance induced by the $(\lambda,-1)$-corrected importance weight.

Lemma 5.1. *Let $P,Q\in\mathscr{P}(\mathcal{Y})$ be two probability distributions with $P\ll Q$. For every $\lambda\in[0,1]$, the bias and variance of the $(\lambda,-1)$-corrected importance weight can be bounded for every $\alpha\in(1,2]$ as:*

$$\Big|\mathop{\mathbb{E}}_{y\sim Q}[\widehat{\mu}_{n,\lambda}]-\mu\Big|\leqslant\|f\|_\infty\lambda^{\alpha-1}I_\alpha(P\|Q),\qquad \mathop{\mathbb{V}\mathrm{ar}}_{y_i\sim Q}[\widehat{\mu}_{n,\lambda}]\leqslant\frac{\|f\|_\infty^2}{n\lambda^{2-\alpha}}I_\alpha(P\|Q).$$

As expected, the bias is zero for $\lambda=0$; it increases with λ and with the divergence term $I_\alpha(P\|Q)$. Indeed, we already observed that the bias is null when $P=Q$ a.s.. In particular, for $\alpha=2$, the

[3] For $s\in\{-\infty,0,\infty\}$ the power mean is defined as a limit.

[4] The choice of $s=-1$ is mainly for analytical convenience and, as we shall see, it already allows enforcing the desired properties. We leave investigating the other values of s for future work.

bound becomes $\|f\|_\infty \lambda I_2(P\|Q)$. Instead, the variance bound decreases in λ and increases with the divergence $I_\alpha(P\|Q)$. For $\alpha=2$, we obtain the bound $\frac{1}{n}\|f\|_\infty^2 I_2(P\|Q)$. Note that when $P=Q$ a.s., we recover $\frac{1}{n}\|f\|_\infty^2$, which is the Popoviciu's inequality for the variance [48]. Thus, our weight correction allows controlling bias and variance even when $I_2(P\|Q)=\infty$, i.e., when the vanilla IS estimator might have infinite variance. Indeed, our transformed estimator has finite variance provided that there exists $\alpha\in(1,2)$ so that $I_\alpha(P\|Q)<\infty$. Tighter (but less intelligible) bounds on bias and variance are reported in Appendix A.3.

Concentration Inequality We are now ready to derive the core theoretical result. We prove an exponential concentration inequality for the $(\lambda,-1)$-corrected IS estimator and we show that, if $I_2(P\|Q)$ is finite, we are able to achieve subgaussian concentration.

Theorem 5.1. *Let $P,Q\in\mathscr{P}(\mathcal{Y})$ be two probability distributions such that $P\ll Q$. For every $\alpha\in(1,2]$ and $\delta\in(0,1)$, if we select $\lambda=\lambda_\alpha^*$ then, with probability at least $1-\delta$ it holds that:*

$$\widehat{\mu}_{n,\lambda_\alpha^*}-\mu\leqslant\|f\|_\infty(2+\sqrt{3})\left(\frac{2I_\alpha(P\|Q)^{\frac{1}{\alpha-1}}\log\frac{1}{\delta}}{3(\alpha-1)^2 n}\right)^{1-\frac{1}{\alpha}},\quad \textit{with}\quad \lambda_\alpha^*=\left(\frac{2\log\frac{1}{\delta}}{3(\alpha-1)^2 I_\alpha(P\|Q)n}\right)^{\frac{1}{\alpha}}.$$

We immediately notice that the dependente on the confidence level δ is the one typical of exponential concentration for every $\alpha\in(1,2]$. In particular, we observe that the bound is subgaussian when $\alpha=2$, requiring that $I_2(P\|Q)<\infty$. Recalling that $I_2(P\|Q)$ governs the variance of the estimator, this result is in line with the general theory of estimators for which the existence of the variance is an unavoidable requirement to achieve subgaussian concentration [12]. Specifically, for $\alpha=2$ the optimal value of the parameter is $\lambda_2^*=\sqrt{(2\log(1/\delta))/(3I_2(P\|Q)n)}$, leading to the bound:

$$\widehat{\mu}_{n,\lambda_2^*}-\mu\leqslant\|f\|_\infty(2+\sqrt{3})\sqrt{\frac{2I_2(P\|Q)\log\frac{1}{\delta}}{3n}}. \tag{2}$$

A tighter bound, based on a different choice of the correction parameter λ_α^{**} is derived in Appendix A.3 and it is omitted here for clarity of presentation.

Differentiability As we have already observed, our weight correction, differently from truncation, is smooth and, thus, differentiable in the target policy. We now focus on the properties of the *gradient* of the $(\lambda,-1)$-corrected estimator and, to this purpose, we constrain the target distribution to belong to a parametric space differentiable distributions $\mathcal{P}_\Theta=\{P_{\boldsymbol{\theta}}\in\mathscr{P}(\mathcal{X}):\boldsymbol{\theta}\in\Theta\}$, where $\Theta\subseteq\mathbb{R}^d$. The gradient of the corrected weight ω_λ w.r.t. the target policy parameters $\boldsymbol{\theta}$ is given by:

$$\nabla_{\boldsymbol{\theta}}\omega_\lambda(y)=\nabla_{\boldsymbol{\theta}}\frac{p_{\boldsymbol{\theta}}(y)}{q(y)}=\frac{(1-\lambda)\omega(y)}{(1-\lambda+\lambda\omega(y))^2}\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(y),\quad \forall y\in\mathcal{Y}.$$

In particular, it can be proved that $\|\nabla_{\boldsymbol{\theta}}\omega_\lambda(y)\|_\infty\leqslant\frac{1}{4\lambda}\|\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(y)\|_\infty$ (Proposition A.1 of Appendix A.3). Thus, if the score $\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}$ is bounded, the gradient will be bounded whenever $\lambda>0$. This property is advisable for gradient optimization and it is not guaranteed, for example, for vanilla IS ($\lambda=0$). Thus, we can also interpret λ as a regularization parameter for the gradient magnitude.

6 Data-driven Tuning of λ

The computation of the parameter λ_2^* requires the knowledge of the divergence $I_2(P\|Q)$. Even when P and Q are known, computing the $I_2(P\|Q)$ can be challenging, especially for continuous distributions, since it involves the evaluation of a complex integral.[5] In principle, we could estimate the divergence $I_2(P\|Q)$ from samples as the empirical second moment of the vanilla importance weights $\frac{1}{n}\sum_{i\in[n]}\omega(y_i)^2$. However, although possibly well-performing in practice [39], this approach would prevent any subgaussian concentration, as the behavior of the non-corrected $\omega(y)^2$ will be surely heavy-tailed whenever $\omega(y)$ is. A general-purpose approach to avoid the divergence estimation is the *Lepski's adaptation method* [32], which only requires knowing an upper and a lower bound on $I_2(P\|Q)$. Unfortunately, this method is known to be computationally intensive.

In this section, we follow a different path inspired by the recent work [64]. If a choice of the parameter λ corrects the weight ω_λ leading to an ideal estimator $\widehat{\mu}_{n,\lambda}$, for the mean μ, we may expect that the empirical second moment of the corrected weights ω_λ will provide a reasonable estimation of

[5] For some common distributions, including Gaussians, the integral can be computed in closed form [15].

$I_2(P\|Q)$. Based on this observation, we propose to choose λ by solving the following equation:

$$\lambda^2 \underbrace{\frac{1}{n}\sum_{i\in[n]} \omega_{\lambda n^{1/4}}(y_i)^2}_{\text{empirical second moment}} = \frac{2\log\frac{1}{\delta}}{3n}. \tag{3}$$

The intuition behind this approach can be stated as follows. If the empirical second moment is close to the divergence, i.e., $\frac{1}{n}\sum_{i\in[n]}\omega_{\lambda n^{1/4}}(y_i)^2 \simeq I_2(P\|Q)$, the solution $\widehat{\lambda}$ of Equation (3) approaches the optimal parameter, i.e., $\widehat{\lambda} \simeq \sqrt{(2\log(1/\delta))/(3I_2(P\|Q)n)} = \lambda_2^*$. We formalize this reasoning in Appendix A.4, proving that Equation (3) admits a unique root $\widehat{\lambda}\in[0,1]$ (Lemma A.4) and that when the number of samples n grows to infinity, $\widehat{\lambda}$ converges indeed to λ_2^* (Lemma A.8). The following result provides the concentration properties of the estimator $\widehat{\mu}_{n,\lambda}$ when using $\widehat{\lambda}$ instead of λ_2^*, under slightly more demanding requirements on the moments of the importance weights.

Theorem 6.1. *Let $P,Q\in\mathscr{P}(\mathcal{Y})$ be two probability distributions such that $P\ll Q$. Let $\widehat{\lambda}$ be the solution of Equation* (3)*, then, if $I_3(P\|Q)$ is finite, for sufficiently large n, for every $\delta\in(0,1)$, with probability at least $1-2\delta$ it holds that:*

$$\widehat{\mu}_{n,\widehat{\lambda}} - \mu \leqslant \|f\|_\infty \frac{5+2\sqrt{3}}{2}\sqrt{\frac{2I_2(P\|Q)\log\frac{1}{\delta}}{3n}}.$$

Compared to Theorem 5.1, this result is weakened in two aspects. First, the inequality holds with a smaller probability $1-2\delta$ since two estimation processes with the same samples are needed, i.e., the computation of $\widehat{\lambda}$ and the corrected estimator $\widehat{\mu}_{n,\widehat{\lambda}}$. Second, and most important, the result holds for large n, whose minimum value is reported in the proof and depends on $I_3(P\|Q)$, which must be finite. We think this is a not too strong requirement considering that even the variance of an empirical estimate of $I_2(P\|Q)$ would depend on the fourth moment of the importance weight, i.e., $I_4(P\|Q)$.[6]

7 Related Works

Importance Sampling has a long history in Monte Carlo simulation as an effective technique for variance reduction in presence of rare events and for what-if analysis [25, 53, 20, 9, 50]. Apart from sparse exceptions [e.g., 8, 18], in the machine learning community, IS is primarily employed for off-policy estimation and learning [e.g., 11, 38, 59]. In this setting, it is well-known that IS might display an inconvenient behavior, depending on the behavioral Q and target P distributions [63, 39]. In particular, IS tends to enlarge the range of the estimator up to $\operatorname{ess\,sup}_{y\sim Q} p(y)/q(y)$. Although this term is finite for discrete distributions (if $P\ll Q$), it is likely unbounded for continuous ones [11]. Furthermore, in the latter case, the vanilla IS estimator might have infinite variance and tends to exhibit a heavy-tailed behavior [39, 40]. These properties suggest that a way of addressing this phenomenon is to resort to *robust* statistics, typically employed for mean estimation under heavy-tailed distributions [37]. Methods in this class include the *trimmed mean* [60, 22], the *median of means* [43, 24], and the *Catoni's* estimator [7]. For all of them, subgaussian guarantees were provided [37]. These techniques have been also successfully employed for regret minimization in finite [5] and continuous arm spaces [36]. In principle, these methods could be employed *as-is* in combination with IS, but, being general-purpose, they might disregard the peculiarities of the setting.

Several *ad-hoc* methods to cope with the problematic IS behavior have been progressively developed. An example, devised by the statistical community, is *self-normalization* [SN-IS, 44]: $\widetilde{\omega}(y_i) = \omega(y_i)/\sum_{j\in[n]}\omega(y_j)$. This approach has the advantage of controlling the range of the estimator at the price of making all samples interdependent and generating a bias. Although the asymptotic consistency is guaranteed [19, 57], its finite-sample analysis is more challenging. In [39], a polynomial concentration inequality was provided and, more recently, exponential bounds based on Efron-Stein inequalities have been proposed [28, 29]. Nevertheless, the resulting inequality is not guaranteed to decrease with $\mathcal{O}(1/\sqrt{n})$ for general distributions and it is difficult to formally relate its concentration rate to the tail properties of the involved distributions [29]. Another popular technique is the weight *truncation* (or *clipping*) [IS-TR, 23, 3]: $\omega_M^{\text{TR}}(y) = \min\{\omega(y), M\}$, where $M>0$ is the

[6] It is possible to circumvent the computation of $I_2(P\|Q)$ by choosing λ independently from $I_2(P\|Q)$ at the price of downgrading the concentration from subgaussian to exponential (Corollary A.2 of Appendix A.4).

Estimator	Maximum	Variance	Bias	Correction (order $\mathcal{O}$)	Concentration (order $\mathcal{O}$)	Is subgaussian?	Is unbiased when $P=Q$?	Is differentiable?
IS [44, 39]	$\operatorname{ess\,sup}\frac{p}{q}$	$\frac{I_2(P\|Q)}{n}$	0	-	$\sqrt{\frac{I_2(P\|Q)}{\delta n}}$	✗ (poly)	✓	✓
SN-IS [44, 29]	1	V^{SN}	B^{SN}	-	$B^{\text{SN}}+\sqrt{V^{\text{ES}}\log\frac{1}{\delta}}$	✗ (exp)	✓	✓
IS-TR(M) [23, 45]	M	$\frac{I_2(P\|Q)}{n}$	$\frac{I_2(P\|Q)}{M}$	$\sqrt{\frac{nI_2(P\|Q)}{\log\frac{1}{\delta}}}$	$\sqrt{\frac{I_2(P\|Q)\log\frac{1}{\delta}}{n}}$	✓	✗	✗
IS-OS(τ) [54]	$\frac{\sqrt{\tau}}{2}$	$\frac{I_2(P\|Q)}{n}$	$\frac{I_3(P\|Q)}{\tau}$	$\sqrt[3]{\left(\frac{nI_3(P\|Q)}{\log\frac{1}{\delta}}\right)^2}$	$\max_{\beta\in\{2,3\}}\sqrt[\beta]{\frac{I_\beta(P\|Q)\left(\log\frac{1}{\delta}\right)^{\beta-1}}{n^{\beta-1}}}$	✗ (exp)	✗	✓
IS-λ (ours)	$\frac{1}{\lambda}$	$\frac{I_2(P\|Q)}{n}$	$\lambda I_2(P\|Q)$	$\sqrt{\frac{\log\frac{1}{\delta}}{I_2(P\|Q)n}}$	$\sqrt{\frac{I_2(P\|Q)\log\frac{1}{\delta}}{n}}$	✓	✓	✓

Table 2: Comparison between several IS estimators assuming $\|f\|_\infty=1$ and for $\alpha=2$ w.r.t. several indexes. For the SN-IS estimator V^{SN} is the Efron-Stein estimate of the variance and B^{SN} is the bias. V^{SN} and B^{SN} converge to 0 as $n\to\infty$, but no convergence rate is provided in [29].

threshold. Some works rely on empirical selections of the truncation threshold [31, 10], while others focus on more theoretically principled approaches [2, 62, 45]. In particular, in [45] a subgaussian deviation bound is derived by suitably adapting the truncation threshold as a function of the number of samples n and the confidence parameter δ. Another interesting approach, designed for CMABs is the *switch* estimator [DR-SW, 62] that selects between DM and IS (or DR), based on the importance weight value, with also guarantees in MSE. Finally, a not so large part of the literature focuses on less crude transformations than truncation, called *smoothing* [61]. They typically take into explicit consideration the estimator tails [46], also providing asymptotic guarantees. Very recently, shrinkage transformations of the weight were proposed, based on the minimization of different bounds on the MSE, in the CMABs [54] setting. Specifically, the *optimistic shrinkage* [IS-OS, 54] leads to a transformation similar to ours $\omega_\tau^{\text{OS}}(y)=\tau\omega(y)/(\omega(y)^2+\tau)$. Unfortunately, even when knowing P and Q and setting τ adaptively, IS-OS is unable to achieve subgaussian concentration and requires $I_3(P\|Q)$ to be finite (Appendix E for details). Refer to Table 2 for a comparison of the estimators.

8 Numerical Simulations

In this section, we provide numerical simulations for off-policy evaluation (Section 8.1) and learning (Section 8.2), with the goal of showing that our estimators, while enjoying desirable theoretical properties, are competitive with traditional (e.g., vanilla IS and self-normalization) and modern baselines (e.g., truncation, optimistic shrinkage). The complete results can be found in Appendix B.

8.1 Off-Policy Evaluation

We present two off-policy evaluation experiments. We start with a synthetic example with Gaussian distributions and, then, we move to the CMAB setting.

Synthetic Example In this experiment, we compare our corrected estimators with IS baselines in a *continuous-distribution* off-policy estimation problem. Specifically, we consider a Gaussian behavioral policy $Q=\mathcal{N}(\mu_Q,\sigma_Q^2)$ and a Gaussian target policy $P=\mathcal{N}(\mu_P,\sigma_P^2)$. We generate n i.i.d. samples from Q and we estimate the expectation of function $f(y)=100\cos(2\pi y)$ under P. We select $\mu_Q=0$, $\mu_P=0.5$, $\sigma_Q^2=1$ and $\sigma_P^2=1.9$, leading to a divergence $I_2(P\|Q)\simeq 27.9$. The results with different choices of the σ_P^2 are reported in Appendix B.1.1. In Table 3, we report the absolute error between the estimated and the true mean for the different importance sampling estimators. For our correction, we report the results obtained with optimal value of λ according to Theorem 5.1 with $\alpha=2$ (IS-λ^*), a value of λ that optimizes a tighter bound reported in Appendix A.3 (IS-λ^{**}), and the value estimated from samples as in Section 6 (IS-$\widehat{\lambda}$). We compare these estimators with vanilla IS (IS), self-normalized IS (SN-IS), weight truncation (IS-TR) with optimal threshold selected as in [45], and IS with optimistic shrinkage (IS-OS), where τ is computed by minimizing an MSE bound as in [54]. We notice that our estimators consistently outperform the traditional ones (IS and SN-IS) and overall suffer smaller errors than IS-TR and IS-OS. Interestingly, the minimum error is often obtained by IS-$\widehat{\lambda}$, which uses an estimated value $\widehat{\lambda}$ that tends to get a higher value than both λ^* and λ^{**}. In this way, the correction is more intense, which, in this specific example, turns out to be beneficial.

Contextual Bandits For the off-policy evaluation in CMABs, we follow the well-established setting of [14, 62, 55, 54]. We consider 11 UCI [13] multi-class classification datasets (see Table 7 in

Estimator / n	10	20	50	100	200	500	1000
IS	**27.43±13.33**	**15.70±4.83**	10.89±1.81	9.26±0.92	12.41±1.88	9.42±0.68	5.84±0.27
SN-IS	**23.89±5.77**	15.62±2.62	10.96±1.18	9.53±0.74	8.82±0.62	7.48±0.37	5.14±0.20
IS-TR	**23.47±7.52**	**14.03±2.75**	10.32±1.47	8.89±0.79	7.68±0.46	6.21±0.28	4.22±0.15
IS-OS	**19.25±8.68**	**10.93±3.29**	**8.37±1.35**	**7.06±0.61**	**8.69±1.44**	6.65±0.47	3.97±0.16
IS-λ^*	**21.75±6.36**	**13.17±2.45**	**9.26±1.19**	7.76±0.62	6.53±0.38	5.29±0.23	3.52±0.12
IS-λ^{**}	**20.66±4.08**	**12.62±2.19**	**8.86±1.08**	**7.39±0.57**	5.94±0.32	4.74±0.20	3.19±0.10
IS-$\hat{\lambda}$	**18.19±3.93**	**10.27±1.64**	**7.03±0.75**	**5.79±0.38**	**3.85±0.21**	**2.90±0.10**	**2.06±0.05**

Table 3: Absolute error in the illustrative example varying the number of samples n for the different estimators (mean $\pm$ std, 60 runs). For each column, the estimator with smallest absolute error and the ones not statistically significantly different from that one (Welch's t-test with $p<0.02$) are in bold.

Estimator / n	100	200	500	1000	2000	5000	10000	20000
IS	**17.38±1.27**	**22.26±2.00**	15.98±0.82	8.36±0.21	**4.67±0.07**	2.68±0.03	2.15±0.02	1.10±0.00
SN-IS	**23.95±1.68**	**19.39±1.30**	17.94±0.50	11.43±0.22	7.10±0.13	2.54±0.03	**1.61±0.01**	1.10±0.00
IS-TR	**17.38±1.27**	**18.92±1.36**	15.88±0.82	8.36±0.21	**4.67±0.07**	2.68±0.03	2.15±0.02	1.10±0.00
IS-OS	24.91±1.45	31.93±1.15	15.38±0.56	17.25±0.45	16.41±0.37	30.63±0.15	33.95±0.02	33.61±0.01
IS-λ^*	**17.22±1.35**	**17.10±1.03**	11.57±0.45	**5.66±0.17**	4.86±0.06	2.73±0.02	2.47±0.02	1.27±0.01
IS-λ^{**}	**17.21±1.39**	**16.20±0.93**	10.93±0.37	**5.55±0.15**	5.05±0.06	2.82±0.02	2.68±0.02	1.43±0.01
IS-$\hat{\lambda}$	**18.16±1.49**	**16.52±0.85**	11.23±0.29	**6.48±0.15**	5.85±0.07	3.01±0.03	2.89±0.02	1.50±0.01
DM	**20.52±1.18**	25.28±0.97	36.19±0.31	36.04±0.08	36.95±0.06	41.99±0.01	42.70±0.01	42.71±0.00
DR	**23.00±1.88**	**25.79±2.38**	20.02±0.92	8.30±0.17	**4.37±0.08**	2.16±0.02	**1.38±0.01**	**0.64±0.00**
SN-DR	**20.89±1.45**	**23.38±1.91**	20.79±0.74	10.99±0.17	6.48±0.11	2.54±0.02	**1.52±0.01**	0.99±0.00
DR-TR	**18.48±1.13**	**15.96±0.72**	18.58±0.23	15.52±0.09	15.45±0.07	20.33±0.01	21.05±0.01	20.78±0.00
DR-OS	**18.47±1.17**	**18.84±0.60**	17.10±0.39	12.19±0.22	8.86±0.11	17.52±0.06	18.40±0.02	19.04±0.02
DR-SW	**22.83±1.25**	**16.81±1.14**	**4.59±0.18**	**4.70±0.09**	4.86±0.06	**0.77±0.01**	**1.38±0.01**	**0.78±0.00**
DR-λ^*	**20.03±1.25**	**18.70±1.33**	13.04±0.61	**6.22±0.13**	**3.82±0.07**	1.79±0.02	**1.37±0.01**	**0.61±0.00**
DR-λ^{**}	**19.40±1.22**	**17.29±1.17**	11.39±0.53	**5.60±0.12**	**3.65±0.06**	1.67±0.02	**1.38±0.01**	**0.63±0.00**
DR-$\hat{\lambda}$	**18.53±1.21**	**14.92±0.98**	9.18±0.44	**4.91±0.10**	**3.39±0.06**	1.61±0.02	**1.40±0.01**	**0.65±0.00**

Table 4: Absolute error (multiplied by 100) in the *letter* dataset varying the number of samples n for the different estimators, when $\alpha_b=0.5$ and $\alpha_e=0.9$ (mean $\pm$ std, 10 runs). For each column, the estimator with smallest absolute error and the ones not statistically significantly different from that one (Welch's t-test with $p<0.05$) are in bold.

Appendix B.1.2). Each dataset $\mathcal{D}^*=\{(x_i,a_i^*)\}_{i\in[n^*]}$ is mapped to a CMAB problem with action set $\mathcal{A}=[K]$. Every sample (x_i,a) leads to a reward given by $\mathbb{1}\{a=a_i^*\}$. To model noise, the reward is switched with probability $\nu\in[0,1]$. Each dataset is split into a training set $\mathcal{D}_{\text{train}}$ and an evaluation $\mathcal{D}_{\text{eval}}$ with proportions 30% and 70%. A multi-class classifier $\mathtt{C}$ is trained on $\mathcal{D}_{\text{train}}$. The behavioral policy is obtained as: $\pi_b(a|x)=\alpha_b+\frac{1-\alpha_b}{K}$ if $a=\mathtt{C}(x)$ and $\pi_b(a|x)=\frac{1-\alpha_b}{K}$ otherwise, where $\alpha_b\in[0,1]$. The target policy π_e is obtained as the behavioral one by training another classifier on $\mathcal{D}_{\text{train}}$ and using $\alpha_e\in[0,1]$. We employ π_b to generate a dataset $\mathcal{D}=\{(x_i,a_i,r_i)\}_{i\in[n]}$ sampling x_i from $\mathcal{D}_{\text{eval}}$ where $a_i\sim\pi_b(\cdot|x_i)$ and r_i is computed as described before. The ground truth value function is computed as $v(\pi_e)=\frac{1}{n}\sum_{x\in\mathcal{D}_{\text{eval}}}\sum_{a\in\mathcal{A}}\pi_e(a|x)r(x,a)$. For DM and DR, we employ a regressor to learn the reward function with a cross-fitting procedure on the whole $\mathcal{D}$.

We use logistic regression as classifiers and linear regression to fit the reward function. We consider several settings that vary the values of α_b and α_e across all the 11 datasets, generating 110 combinations and a reward noise of $\nu=0.25$. Details and results for the noiseless case are reported in Appendix B.1.2. To summarize the results, following the approach of [62], we plot in Figure 2 the cumulative distribution function (CDF) of the absolute error normalized by the error of IS. A lower error corresponds to a CDF curve towards the upper-left corner. We distinguish between the approaches that do not make use of the reward estimate $\hat{r}$ (model-free, left) and the ones that do (model-based, right). As for the model-free ones, we note that the performance of our estimator IS-λ^* is very close to that of SN-IS. This is likely because we are dealing with discrete distributions (actions are finite), which implicitly mitigate the degeneracy of the importance weight. Differently, the advantage w.r.t. the optimistic shrinkage (IS-OS) is quite significant. Instead, for the model-based estimators, we observe

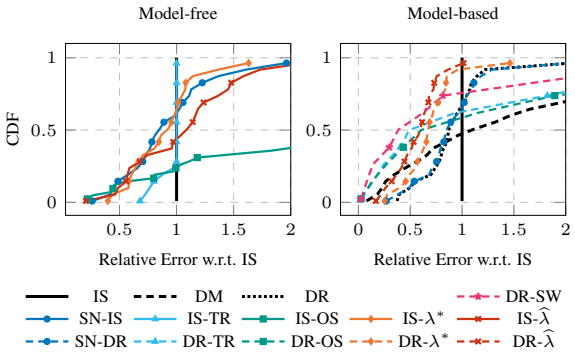

Figure 2: CDF of the absolute error normalized by IS error for stochastic rewards with noise 0.25, across 110 conditions.

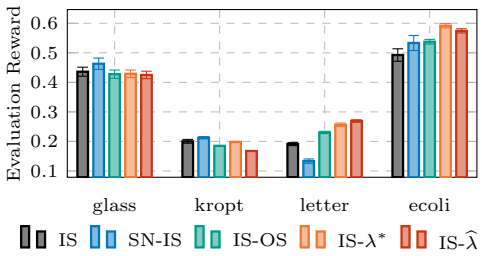

Figure 3: Evaluation reward for four different datasets after 1000 iterations (4000 iterations for *letter*) of gradient ascent with a Boltzmann policy for the model-free estimators (mean $\pm$ std, 10 runs).

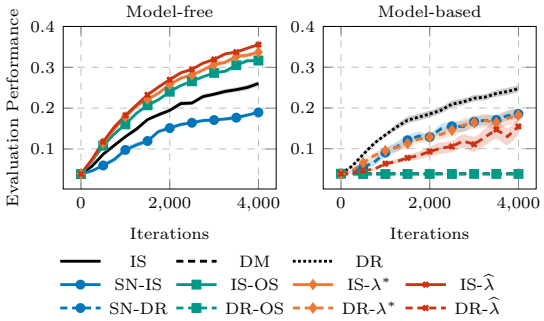

Figure 4: Evaluation reward for the *letter* dataset comparing the learning curve of different estimators (mean $\pm$ std, 10 runs).

that our weight correction combined with the DR estimator (DR-λ* and DR-$\widehat{\lambda}$) outperforms the standard DR and its combinations with SN (SN-DR), truncation (DR-TR), and optimistic shrinkage (DR-OS). Instead, the switch estimator (DR-SW) displays a performance comparable to ours.

For the specific case of the *letter* dataset, we report in Table 4 the results obtained by setting $\alpha_b = 0.5$ and $\alpha_e = 0.9$ for different number of samples n. We notice essentially two behaviors. When the number of samples is very low (e.g., 100, 200) all estimators perform similarly, with poor performance. As n increases, the benefits of the DR-like estimators becomes more visible. In particular, the DR-SW and our corrected estimators (DR-λ*, DR-λ**, and DR-$\widehat{\lambda}$) overall dominate the other baselines.

8.2 Off-Policy Learning

Finally, we provide an experiment in which we employ the off-policy methods to improve a baseline policy in the CMAB framework. We refer to the same setting of Section 8.1 with a uniform behavioral policy ($\alpha_b = 0$). For the target policy, we consider a Boltzmann policy in some featurization of the context $\pi_{\boldsymbol{\theta}}(a|x) \propto \exp\left(\boldsymbol{\theta}_a^T \boldsymbol{\phi}(x)\right)$. We optimize the estimated value function in the parameters $\boldsymbol{\theta}$ via gradient ascent. Further details and experiments with regularized objectives are reported in Appendix B.2. We perform Off-PL on four datasets and the results for the model-free estimators are reported in Figure 3. We observe that our weight corrections (DR-λ* and DR-$\widehat{\lambda}$) outperform the considered baselines (IS, SN-IS, and IS-OS) on *ecoli* and *letter* datasets, whereas SN-IS emerges in the *glass* and *kropt* datasets. For the *letter* dataset, we report in Figure 4 the learning curve, distinguishing between model-free (left) and model-based (right) estimators.[7] For the model-free ones, we observe the dominance of our estimators over SN estimator (SN-IS), while the optimistic shrinkage estimator (IS-OS) behaving similarly to ours. Interestingly, for the model-based estimators, plain DR beats the other estimators, including self-normalization that performs almost identically with our DR-λ*, and IS-OS that fails completely to learn the task.

9 Discussion and Conclusions

In this paper, we have deepened the study of the importance sampling technique for off-policy evaluation and learning. We derived an anticoncentration bound for the vanilla IS estimator, proving polynomial concentration is tight for this setting. Then, we introduced and analyzed a class of importance weight corrections based on the intuition of smoothly shrinking the weight towards one. Assuming that the second moment of the importance weight exists, we have introduced the first transformation that achieves subgaussian concentration and maintains the differentiability of the estimator in the target policy parameters. The experimental evaluation has shown that our theoretically-grounded transformation is competitive with the traditional and modern IS baselines (including self-normalization, truncation, and optimistic shrinkage) in the CMAB framework for both evaluation and learning. The advantages of our correction are more visible in the case of continuous distributions, where the degeneracy of importance sampling is amplified. Future works include the extension of these corrections to the more challenging RL setting with continuous actions.

[7] Clearly, the truncated (IS-TR) and the switch (DR-SW) estimators cannot be directly employed in this setting, being non-differentiable.

References

[1] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[2] Oliver Bembom and Mark J van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. 2008.

[3] Léon Bottou, Jonas Peters, Joaquin Quiñonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y. Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *J. Mach. Learn. Res.*, 14(1):3207–3260, 2013.

[4] Stéphane Boucheron, Gábor Lugosi, Pascal Massart, et al. On concentration of self-bounding functions. *Electronic Journal of Probability*, 14:1884–1899, 2009.

[5] Sébastien Bubeck, Nicolò Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Trans. Inf. Theory*, 59(11):7711–7717, 2013.

[6] Peter S Bullen. *Handbook of means and their inequalities*, volume 560. Springer Science & Business Media, 2013.

[7] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

[8] Kamil Andrzej Ciosek and Shimon Whiteson. OFFER: off-environment reinforcement learning. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 1819–1825. AAAI Press, 2017.

[9] William G Cochran. *Sampling techniques*. John Wiley & Sons, 2007.

[10] Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

[11] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 442–450. Curran Associates, Inc., 2010.

[12] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, Roberto I Oliveira, et al. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[13] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[14] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In Lise Getoor and Tobias Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 1097–1104. Omnipress, 2011.

[15] M. Gil, Fady Alajaji, and Tamás Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Inf. Sci.*, 249:124–131, 2013.

[16] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline A/B testing for recommender systems. In Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek, editors, *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 198–206. ACM, 2018.

[17] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

[18] Josiah P. Hanna, Philip S. Thomas, Peter Stone, and Scott Niekum. Data-efficient policy evaluation through behavior policy search. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1394–1403. PMLR, 2017.

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

[20] Timothy Classen Hesterberg. *Advances in importance sampling*. PhD thesis, Citeseer, 1988.

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

[22] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

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

[24] Mark Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theor. Comput. Sci.*, 43:169–188, 1986.

[25] Herman Kahn and Andy W Marshall. Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, 1953.

[26] Ozsel Kilinc, Yang Hu, and Giovanni Montana. Reinforcement learning for robotic manipulation using simulated locomotion demonstrations. *CoRR*, abs/1910.07294, 2019.

[27] Jens Kober and Jan Peters. *Learning Motor Skills - From Algorithms to Robot Experiments*, volume 97 of *Springer Tracts in Advanced Robotics*. Springer, 2014.

[28] Ilja Kuzborskij and Csaba Szepesvári. Efron-stein pac-bayesian inequalities. *arXiv preprint arXiv:1909.01931*, 2019.

[29] Ilja Kuzborskij, Claire Vernade, András György, and Csaba Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. 130:640–648, 2021.

[30] John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 817–824. Citeseer, 2007.

[31] Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174, 2011.

[32] Oleg V Lepski and Vladimir G Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, pages 2512–2546, 1997.

[33] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In Irwin King, Wolfgang Nejdl, and Hang Li, editors, *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 297–306. ACM, 2011.

[34] Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015.

[35] Friedrich Liese and Igor Vajda. *Convex statistical distances*, volume 95. Teubner, 1987.

[36] Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4154–4163. PMLR, 2019.

[37] Gabor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.*, 19(5):1145–1190, 2019.

[38] Ashique Rupam Mahmood, Hado van Hasselt, and Richard S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3014–3022, 2014.

[39] Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5447–5459, 2018.

[40] Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. Importance sampling techniques for policy optimization. *J. Mach. Learn. Res.*, 21:141:1–141:75, 2020.

[41] John E. Moody and Matthew Saffell. Learning to trade via direct reinforcement. *IEEE Trans. Neural Networks*, 12(4):875–889, 2001.

[42] Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

[43] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[44] Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.

[45] Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, and Marcello Restelli. Optimistic policy optimization via multiple importance sampling. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4989–4999. PMLR, 2019.

[46] James Pickands III et al. Statistical inference using extreme order statistics. *Annals of statistics*, 3(1):119–131, 1975.

[47] Iosif Pinelis et al. Best possible bounds of the von bahr–esseen type. *Annals of Functional Analysis*, 6(4):1–29, 2015.

[48] Tiberiu Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*, 9:129–145, 1935.

[49] Alfréd Rényi. On measures of entropy and information. Technical report, Hungarian Academy of Sciences Budapest Hungary, 1961.

[50] Brian D Ripley. *Stochastic simulation*, volume 316. John Wiley & Sons, 2009.

[51] Yuta Saito, Aihara Shunsuke, Matsutani Megumi, and Narita Yusuke. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. *arXiv preprint arXiv:2008.07146*, 2020.

[52] Igal Sason. On f-divergences: Integral representations, local behavior, and inequalities. *Entropy*, 20(5):383, 2018.

[53] David Siegmund. Importance sampling in the monte carlo study of sequential tests. *The Annals of Statistics*, pages 673–684, 1976.

[54] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9167–9176. PMLR, 2020.

[55] Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. CAB: continuous adaptive blending for policy evaluation and learning. 97:6005–6014, 2019.

[56] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[57] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3231–3239, 2015.

[58] Liang Tang, Rómer Rosales, Ajit Singh, and Deepak Agarwal. Automatic ad format selection via contextual bandits. In Qi He, Arun Iyengar, Wolfgang Nejdl, Jian Pei, and Rajeev Rastogi, editors, *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 1587–1594. ACM, 2013.

[59] Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 3000–3006. AAAI Press, 2015.

[60] John W Tukey and Donald H McLaughlin. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/winsorization 1. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 331–352, 1963.

[61] Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

[62] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3589–3597. PMLR, 2017.

[63] Changhe Yuan and Marek J. Druzdzel. How heavy should the tails be? In Ingrid Russell and Zdravko Markov, editors, *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference, Clearwater Beach, Florida, USA*, pages 799–805. AAAI Press, 2005.

[64] Chao Zheng. A new principle for tuning-free huber regression. *Statistica Sinica*, 2020.

[65] Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, and Michael R Kosorok. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 9.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] The paper is mostly a theoretical contribution and we think it will not have any potential negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sections 3, 4, and 5.
   (b) Did you include complete proofs of all theoretical results? [Yes] They are reported in Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code is included in the supplementary material, with a README reporting the commands to reproduce the results. The datasets employed in the CMAB experiments are taken from the UCI repository (`https://archive.ics.uci.edu/ml/index.php`).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 8 and Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Their meaning is reported in the corresponding figure captions.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix B
   (b) Did you mention the license of the assets?[Yes] See Appendix B
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The code is attached to the supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]