Artificial Intelligence 348 (2025) 104419

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint

Minimax off-policy evaluation and learning with subgaussian and differentiable importance weighting

Alberto Maria Metelli *, Alessio Russo, Marcello Restelli

Politecnico di Milano, Piazza Leonardo da Vinci, 32, Milan, 20133, Italy

ARTICLE INFO

Keywords:
Off-policy estimation
Importance weighting
Power mean transformation
Subgaussian concentration
Differentiable importance weighting

ABSTRACT

In this work, we study the statistical properties of the *off-policy estimation* problem, i.e., estimating expectations under a target policy using samples collected from a different policy. We begin by presenting a novel minimax concentration lower bound that highlights the fundamental limits of off-policy estimation. We then analyze two well-known *importance weighting* (IW) techniques: vanilla IW and self-normalized importance weighting (SN). For both methods, we derive concentration and anti-concentration results, showing that their concentration rates are provably suboptimal compared to our lower bound. Observing that this undesired behavior arises from the *heavy-tailed* nature of the IW and SN estimators, we propose a new class of parametric estimators based on a transformation using the *power mean* (PM), which is no longer heavy-tailed. We study the theoretical properties of the PM estimator in terms of bias and variance. We show that, with suitable (possibly data-driven) tuning of its parameters, the PM estimator satisfies two key properties under certain conditions: (*i*) it achieves a *subgaussian* concentration rate that matches our lower bound and (*ii*) it maintains differentiability with respect to the target policy. Finally, we validate our approach through numerical simulations on both synthetic datasets and contextual bandits, comparing it against standard off-policy evaluation and learning baselines.[1]

1. Introduction

The technological and digital evolution of recent decades has increasingly emphasized the importance of data, which now appear in large volumes and in various forms, ranging from raw to more structured information. Extracting the knowledge contained in this data can help solve a wide range of real-world decision-making problems in domains such as health and care (e.g., [23,77]), industrial robot control (e.g., [33,32]), personalized advertising (e.g., [5,68]), and finance (e.g., [50]). Many of these problems can be addressed using *off-policy* techniques. We distinguish between (*i*) *off-policy evaluation*, where data collected with a behavioral policy is used to evaluate a different target policy and (*ii*) *off-policy learning*, where the available data is used to improve the performance of a different target policy. These techniques have been extensively applied in various frameworks, including the well-known *contextual multi-armed bandit* (CMAB, [36]).

Among the various techniques used to address these problems, a common approach is to adopt an *importance weighting* (IW, [53]) estimator, whose basic idea is to weight each sample proportionally to the probability of it being generated by the target policy. Vanilla IW has the advantage of being unbiased, but it may suffer from very high variance, particularly when the target and behavioral

* Corresponding author.
E-mail addresses: albertomaria.metelli@polimi.it (A.M. Metelli), alessio.russo@polimi.it (A. Russo), marcello.restelli@polimi.it (M. Restelli).

[1] A conference version of this work, including some of the results presented here, appeared in NeurIPS 2021 [48].

https://doi.org/10.1016/j.artint.2025.104419
Received 27 May 2025; Received in revised form 7 September 2025; Accepted 8 September 2025

Available online 15 September 2025

distributions are significantly different [14]. Indeed, IW tends to enlarge the range of the estimator according to the values taken by the ratio of the associated density functions. This range is finite for discrete distributions but is likely unbounded for continuous ones, potentially resulting in infinite variance. This behavior depends on its *heavy-tailed* properties [43], preventing the use of bounds stricter than polynomial ones (e.g., Chebyshev's inequality).

Over the years, many works have attempted to mitigate this issue, and several corrections to the importance weights have been proposed. Among the most popular are *weight truncation* (TR, [29,54]), which employs a clipping threshold M to prevent weight explosion, and the *self-normalization* (SN, [12]) technique, which bounds the weights by introducing interdependence among samples. While these transformations are general-purpose, other methods, such as the *doubly robust* (DR, [18]) estimator, are specifically designed for the CMAB setting. The DR estimator combines a *direct method* (DM), where the reward is estimated from historical data using regression, with an IW-based term used as a control variate.

Intuitively, IW corrections should be designed to effectively control the *variance*, while accepting the introduction of some *bias* compared to vanilla IW. However, evaluating the performance of such estimators using the *mean squared error* (MSE), i.e., the sum of the variance and the squared bias, may not be an appropriate performance index. As highlighted in several works [43,42], the MSE is a weak metric when heavy-tailed distributions are involved, especially when such estimators are intended for decision-making purposes [8]. Instead, *high-probability concentration bounds* are preferable, as they are stronger than MSE bounds and implicitly control all moments of the estimator. As noted in [43], an ideal estimator should satisfy *subgaussian* concentration, i.e., a particular instance of exponential concentration analogous to that of the sample mean of independent Gaussian random variables (e.g., Hoeffding's inequality). However, in striving for subgaussianity, an estimator may sacrifice other desirable properties such as *differentiability*, as is the case with the TR estimator [54], which can be particularly important when (gradient-based) learning methods are employed. Finally, existing estimators usually require setting parameters based on the desired confidence level and the properties of the behavioral and target distributions [8,54,47], which are typically characterized by some form of divergence (or dissimilarity index). Determining such parameters in a fully *data-driven* manner while maintaining theoretical guarantees remains largely an open question.

Original contributions. In this paper, we make a step towards understanding the statistical properties of off-policy estimation and, specifically, importance sampling and relative corrections. The contributions of this work are summarized as follows:

- We provide the first *minimax concentration lower bound* for off-policy estimation. This result establishes the minimum error that *any estimator* must suffer having fixed a confidence level, a number of samples, and a dissimilarity index between behavioral and target distributions. This shows how subgaussian concentration can be achieved at the price of a dependence on a dissimilarity index between the behavioral and target distribution (Section 3).
- We analyze the statistical properties of the vanilla importance weighting (IW) and self-normalized importance weighting (SN) estimators. Specifically, for both estimators, we provide *polynomial concentration bounds* and *polynomial anti-concentration bounds*, illustrating how these estimators are unable to match the minimax concentration rate (Section 4).
- We propose the *power-mean correction* (PM) importance weight estimator, a variant of standard IW. It uses two correction parameters (λ, s) that interpolate between the vanilla importance weights and 1, by performing a generalized mean with power s. This estimator is differentiable in the target policy and achieves exponential concentration. Under the knowledge of confidence level and of the dissimilarity index between the behavioral and target distributions, (λ, s) can be set to achieve subgaussian concentration, matching our minimax lower bound for a sufficiently large number of samples (Section 5).
- We devise a *data-driven approach* that allows setting (λ, s) based on the collected data. The approach requires solving a root-finding problem and, under the existence of a higher-order dissimilarity index, allows matching the minimax lower bound for a sufficiently large number of samples (Section 6).
- We critically compare the existing importance weighting corrections, highlighting their properties in terms of bias, variance, concentration, and differentiability (Section 7).
- We provide a numerical validation comparing our approach with traditional and modern off-policy baselines on synthetic domains and in the CMAB framework (Section 8).

Compared to the conference version of NeurIPS 2021 [48], the novel contributions of this paper are: the minimax concentration lower bound (Section 3); the anti-concentration bound for the SN estimator (Section 4); the analysis of the PM estimator for a general value of the parameter s (Section 5); and the data-driven approach extended to general order of the dissimilarity and value of the parameter s (Section 6). The proofs of all the results presented in the paper are reported in Appendix A.

2. Preliminaries

In this section, we provide the necessary background to understand the content of the paper. We focus on estimator concentration (Section 2.1), importance weighting (Section 2.2), the CMABs (Section 2.3), and off-policy evaluation and learning (Section 2.4).

2.1. Estimator concentration

Let $\mathcal{Y}$ be a set and $\mathfrak{F}_{\mathcal{Y}}$ be a σ-algebra over $\mathcal{Y}$, we denote with $\Delta^{\mathcal{Y}}$ the set of probability measures over the measurable space $(\mathcal{Y}, \mathfrak{F}_{\mathcal{Y}})$. Let $P \in \Delta^{\mathcal{Y}}$ be a probability measure and $f : \mathcal{Y} \to \mathbb{R}$ be a measurable function. Let $\overline{\mu}_n$ be an estimator for the expected

value of function f under measure P, i.e., $\mu = \mathbb{E}_{y\sim P}[f(y)]$, obtained with $n \in \mathbb{N}$ i.i.d. samples.[2] Suppose that, for every $\delta \in (0, \delta_{\max})$, with probability $1-\delta$ it holds that:

$$\left|\overline{\mu}_n - \mu\right| \leq g(n,\delta). \tag{1}$$

We allow $\overline{\mu}_n$ to explicitly depend on the number of samples n and on the confidence parameter δ. For $\beta_1, \beta_2 > 0$, we say that $\overline{\mu}_n$ admits: (i) *polynomial* concentration if $g(n,\delta) = \mathcal{O}\left(\frac{1}{n^{\beta_1}\delta^{\beta_2}}\right)$, this case corresponds to Chebyshev's inequality when $\beta_1 = \beta_2 = 1/2$; (ii) *exponential* concentration if $g(n,\delta) = \mathcal{O}\left(\frac{\left(\log\left(\frac{1}{\delta}\right)\right)^{\beta_2}}{n^{\beta_1}}\right)$; (iii) *subgaussian* concentration if (ii) holds with $\beta_1 = \beta_2 = 1/2$ [43], this case corresponds to Hoeffding's inequality.

2.2. Importance weighting

Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability measures, admitting p and q as density functions w.r.t. a reference measure. If $P \ll Q$, i.e., P is absolutely continuous w.r.t. Q, for any $\alpha \in (1,2]$, we introduce the integral:

$$I_\alpha(P\|Q) = \int_{\mathcal{Y}} p(y)^\alpha q(y)^{1-\alpha} \mathrm{d}y. \tag{2}$$

If $P = Q$ a.s. (almost surely) then $I_\alpha(P\|Q) = 1$. $I_\alpha(P\|Q)$ allows defining several commonly used divergences, like Rényi divergence [59]: $(\alpha-1)^{-1}\log I_\alpha(P\|Q)$ and Tsallis divergence [70]: $(\alpha-1)^{-1}\left(I_\alpha(P\|Q) - 1\right)$.

Let $f : \mathcal{Y} \to \mathbb{R}$ be a measurable function, *(vanilla) importance weighting* (IW, [53]) allows estimating the expected value 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. The estimator $\widehat{\mu}_n$ is obtained by reweighing each sample by the likelihood ratio $\omega(y) = \frac{p(y)}{q(y)}$ for all $y \in \mathcal{Y}$, also called *importance weight*:

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i\in[\![n]\!]} \omega(y_i) f(y_i). \tag{3}$$

It is well-known that $\widehat{\mu}_n$ is unbiased, i.e., $\mathbb{E}_{y_i\sim Q}[\widehat{\mu}_n] = \mu$ [53]. If f is bounded, the variance of the estimator can be upper-bounded as $\mathbb{V}\mathrm{ar}_{y_i\sim Q}[\widehat{\mu}_n] \leq \frac{1}{n}\|f\|_\infty I_2(P\|Q)$ [45]. More generally, the integral $I_\alpha(P\|Q)$ defines the α-moment of the importance weight $\omega(y)$ under the distribution Q.

A common approach to mitigate the variance of IW is to resort to *self-normalization* (SN, [53]). The estimator $\widetilde{\mu}_n$ is obtained by rescaling each weight $\omega(y)$ by the sum of the weights of all collected samples:

$$\widetilde{\mu}_n = \frac{\sum_{i\in[\![n]\!]} \omega(y_i) f(y_i)}{\sum_{i\in[\![n]\!]} \omega(y_i)}. \tag{4}$$

The SN estimator $\widetilde{\mu}_n$ has the desirable property of being bounded by $\|f\|_\infty$. However, since the division by the sum of the weights makes all the samples interdependent, it is no longer unbiased, while preserving consistency [53].

2.3. Contextual bandits

A *contextual multi-armed bandit* (CMAB, [36]) is represented by the tuple $\mathcal{C} = (\mathcal{X}, \mathcal{A}, \rho, p)$, where $\mathcal{X}$ is a measurable set of *contexts*, $\mathcal{A}$ is a measurable set of actions (or arms), $\rho \in \Delta^{\mathcal{X}}$ is the context distribution, and $p : \mathcal{X}\times\mathcal{A} \to \Delta^{\mathbb{R}}$ is the reward distribution. A *policy* $\pi : \mathcal{X} \to \Delta^{\mathcal{A}}$ characterizes the agent's behavior by mapping each context to a probability distribution over actions. At each round $t \in \mathbb{N}$, the agent observes a context $x_t \sim \rho$, chooses 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:

$$v(\pi) := \int_{\mathcal{X}} \rho(x) \int_{\mathcal{A}} \pi(a|x) \int_{\mathbb{R}} p(r|x,a) r \,\mathrm{d}r\mathrm{d}a\mathrm{d}x. \tag{5}$$

We denote with $r(x,a) := \int_{\mathbb{R}} p(r|x,a) r \mathrm{d}r$ the *reward function*. If a policy π^* maximizes the value function, i.e., $\pi^* \in \arg\max_{\pi\in\Pi} v(\pi)$, with $\Pi = \{\pi : \mathcal{X} \to \Delta^{\mathcal{A}}\}$ being the set of all policies, then the policy is optimal.

2.4. Off-policy evaluation and learning

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, [27]) problem consists in estimating the value function $v(\pi_e)$ of a target policy $\pi_e \in \Pi$ using the samples in $\mathcal{D}$. Differently, the *off-policy learning* (Off-PL, [18]) problem consists in estimating an optimal policy $\pi^* \in \Pi$ using the samples in $\mathcal{D}$. The simplest approach

[2] We denote with $[\![n]\!] := \{1, \dots, n\}$.

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 $\widehat{r}$. A different approach involves simply applying IW to reweight the samples of $\mathcal{D}$, leading to the *inverse propensity scoring* (IW, [23]) estimator. The *doubly-robust* (DR, [18]) estimator combines the two presented approaches. The DM estimate is corrected with an IW control variate to reduce the variance using the estimated reward $\widehat{r}$ (see also Table 12 in Appendix D).

3. Minimax concentration lower bound for off-policy estimation

In this section, we derive a minimax lower bound for the concentration rate that *any* off-policy estimator suffers. The following result highlights the main challenges of the off-policy estimation problem, showing, in particular, that a dependence on the dissimilarity index $I_\alpha(P\|Q)$ is unavoidable.

Theorem 3.1 *(Minimax Lower Bound). Let $\overline{\mu}_n$ be an estimator for $\mu = \mathbb{E}_{x\sim P}[f(x)]$ using samples $\{y_i\}_{i\in[\![n]\!]}$ i.i.d. collected from Q. There exist $P, Q \in \Delta^{\mathcal{Y}}$ two probability measures such that $P \ll Q$ and such that for some $\alpha \in (1,2]$, it holds that $I_\alpha(P\|Q) < +\infty$, and bounded measurable function $f : \mathcal{Y} \to \mathbb{R}$ such that, if $\delta \in (0,1/4)$, $I_\alpha(P\|Q) \geq 3$, and $n \geq \frac{\log\left(\frac{1}{\delta}\right)}{2\left(1-2^{\frac{1}{1-\alpha}}\right)}$, with probability at least δ it holds that:*

$$|\overline{\mu}_n - \mu| \geq \frac{\|f\|_\infty}{\sqrt{6}} \left(\frac{\log\left(\frac{1}{\delta}\right)}{n} \right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q)^{\frac{1}{\alpha}}. \tag{6}$$

Proof. We provide an explicit construction of P and Q. Let P and Q be two discrete distributions defined over the support $\mathcal{Y} = \{A, B\}$ such that, for $p, q \in [0,1]$, we have:

$$P(\{A\}) = p \quad \text{and} \quad P(\{B\}) = 1-p, \tag{P.1}$$

$$Q(\{A\}) = q \quad \text{and} \quad Q(\{B\}) = 1-q. \tag{P.2}$$

We consider two instances of the function f_s with $s \in \{-1, 1\}$, defined as follows:

$$f_s(y) = \begin{cases} \frac{s\epsilon}{p} & \text{if } y = A \\ 0 & \text{otherwise} \end{cases}, \tag{P.3}$$

where $\epsilon > 0$ will be specified later. Furthermore, we have that $\mu_s := \mathbb{E}_{y\sim P}[f_s(y)] = s\epsilon$ and $p = \frac{\epsilon}{\|f\|_\infty}$. Consider now the event that all samples $\{y_i\}_{i\in[\![n]\!]}$ sampled independently from Q are all equal to B, i.e., $\mathcal{E} := \{\forall i \in [\![n]\!] : y_i = B\}$, that holds with probability $\mathbb{P}_{y_i\sim Q}(\mathcal{E}) = (1-q)^n$ since all samples are independent. Thus, we have:

$$\begin{aligned}
2 \max_{s\in\{-1,1\}} \mathbb{P}_{y_i\sim Q}\left(|\overline{\mu}_n - \mu_s| \geq \epsilon\right) &\geq 2 \max_{s\in\{-1,1\}} \mathbb{P}_{y_i\sim Q}\left(|\overline{\mu}_n - \mu_s| \geq \epsilon \wedge \mathcal{E}\right) && \text{(P.4)} \\
&\geq \sum_{s\in\{-1,1\}} \mathbb{P}_{y_i\sim Q}\left(|\overline{\mu}_n - \mu_s| \geq \epsilon \wedge \mathcal{E}\right) && \text{(P.5)} \\
&\geq \mathbb{P}_{y_i\sim Q}\left(\max_{s\in\{-1,1\}} |\overline{\mu}_n - \mu_s| \geq \epsilon \wedge \mathcal{E}\right) && \text{(P.6)} \\
&\geq \mathbb{P}_{y_i\sim Q}\left(\min_{z\in\mathbb{R}} \max_{s\in\{-1,1\}} |z - s\epsilon| \geq \epsilon \wedge \mathcal{E}\right) && \text{(P.7)} \\
&\geq \mathbb{P}_{y_i\sim Q}(\mathcal{E}) = (1-q)^n, && \text{(P.8)}
\end{aligned}$$

where line (P.5) follows from bounding the maximum with the average, line (P.6) follows from a union bound, line (P.7) is obtained by considering the best choice of estimator $\overline{\mu}_n$, and line (P.8) is obtained from observing that $\min_{z\in\mathbb{R}} \max_{s\in\{-1,1\}} |z - s\epsilon| = \epsilon$ attained by $z = 0$. By setting the last expression equal to 2δ with $\delta \in (0,1/4)$, we obtain:

$$(1-q)^n = 2\delta \implies \log\left(\frac{1}{1-q}\right) = \frac{1}{n}\log\left(\frac{1}{2\delta}\right) \implies q \geq \frac{1}{n}\log\left(\frac{1}{2\delta}\right) \geq \frac{1}{2n}\log\left(\frac{1}{\delta}\right), \tag{P.9}$$

having exploited the bound $\log\frac{1}{1-q} \geq q$ for $q \in [0,1]$ and exploited $\delta \leq \frac{1}{4}$ to bound $\log\left(\frac{1}{2\delta}\right) \geq \frac{1}{2}\log\left(\frac{1}{\delta}\right)$. Let us now compute the dissimilarity index between P and Q, recalling that $p = \frac{\epsilon}{\|f\|_\infty}$:

$$I_\alpha(P\|Q) = p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha} \implies \epsilon = \|f\|_\infty q^{\frac{\alpha-1}{\alpha}} \left(I_\alpha(P\|Q) - (1-p)^\alpha(1-q)^{1-\alpha}\right)^{\frac{1}{\alpha}}. \tag{P.10}$$

By enforcing $(1-q)^{1-\alpha} \leq 2 \implies q \leq 1 - 2^{\frac{1}{1-\alpha}}$, exploiting Equation (P.9) and $I_\alpha(P\|Q) \geq 3 \implies I_\alpha(P\|Q) - 2 \geq I_\alpha(P\|Q)/3$, we have:

$$\epsilon = \|f\|_\infty q^{\frac{\alpha-1}{\alpha}} \left(I_\alpha(P\|Q) - (1-p)^\alpha (1-q)^{1-\alpha}\right)^{\frac{1}{\alpha}} \tag{P.11}$$

$$\geq \|f\|_\infty \left(\frac{1}{2n}\log\left(\frac{1}{\delta}\right)\right)^{\frac{\alpha-1}{\alpha}} \left(I_\alpha(P\|Q) - 2\right)^{\frac{1}{\alpha}} \tag{P.12}$$

$$\geq \frac{1}{2}\left(\frac{2}{3}\right)^{\frac{1}{\alpha}} \|f\|_\infty \left(\frac{1}{n}\log\left(\frac{1}{\delta}\right)\right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q)^{\frac{1}{\alpha}} \tag{P.13}$$

$$\geq \frac{\|f\|_\infty}{\sqrt{6}} \left(\frac{1}{n}\log\left(\frac{1}{\delta}\right)\right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q)^{\frac{1}{\alpha}}. \tag{P.14}$$

Finally, to ensure that $q \leq 1 - 2^{\frac{1}{1-\alpha}}$ and $q \geq \frac{1}{2n}\log\left(\frac{1}{\delta}\right)$ are compatible, we enforce:

$$\frac{1}{2n}\log\left(\frac{1}{\delta}\right) \leq 1 - 2^{\frac{1}{1-\alpha}} \implies n \geq \frac{\log\left(\frac{1}{\delta}\right)}{2\left(1 - 2^{\frac{1}{1-\alpha}}\right)}. \quad \square \tag{P.15}$$

The result of Theorem 3.1 provides a regime, for a sufficiently large number of samples n, and divergence term $I_\alpha(P\|Q)$, and small error probability δ, under which the concentration rate is of order $\Omega\left(\left(\frac{\log\left(\frac{1}{\delta}\right)}{n}\right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q)^{\frac{1}{\alpha}}\right)$. It is worth noting that if the variance of the importance weight exists finite, i.e., $I_2(P\|Q) < +\infty$, the rate reduces to the typical subgaussian one $\Omega\left(\sqrt{\frac{I_2(P\|Q)\log\left(\frac{1}{\delta}\right)}{n}}\right)$.[3]

To the best of our knowledge, Theorem 3.1 represents the first minimax concentration lower bound for off-policy estimation that clearly highlights the dependence on a dissimilarity index between the behavioral and the target distributions. As a consequence, we cannot aim to derive estimators whose concentration rate does not depend on the dissimilarity index $I_\alpha(P\|Q)$ and, consequently, assuming the finiteness of $I_\alpha(P\|Q)$ for some $\alpha \in (1,2]$ is necessary to obtain a meaningful concentration, at least in the considered regime. Existing lower bounds in the literature [39,74] differ from ours in several perspectives. First, they are limited to the CMAB setting, which, in particular, restricts to the case in which the number of actions is finite. Second, they restrict to the case in which the variance of the vanilla importance weight exists finite, i.e., analogous to our $I_2(P\|Q)$. Third, they focus on a different performance index, i.e., the *mean squared error* (MSE), rather than a concentration rate. In this sense, they represent bounds in expectation and not in probability as ours. It is worth noting that a high-probability bound can always be translated into an MSE one. We show this for the general case in Proposition A.3, that if applied to Theorem 3.1, leads to the MSE bound:

$$\mathop{\mathbb{E}}_{y_i \sim Q}\left[\left(\overline{\mu}_n - \mu\right)^2\right] \geq \sup_{\delta \in (0,1/4)} \delta \frac{\|f\|_\infty^2}{6}\left(\frac{\log\left(\frac{1}{\delta}\right)}{n}\right)^{\frac{2(\alpha-1)}{\alpha}} I_\alpha(P\|Q)^{\frac{2}{\alpha}} \geq \frac{\|f\|_\infty^2}{48}\left(\frac{\log 8}{n}\right)^{\frac{2(\alpha-1)}{\alpha}} I_\alpha(P\|Q)^{\frac{2}{\alpha}}, \tag{7}$$

which takes form $\mathcal{O}\left(\frac{I_2(P\|Q)}{n}\right)$ for $\alpha = 2$.

4. Polynomial concentration of common importance weighting estimators

In this section, we study the intrinsic limits of importance weighting, providing *concentration* and *anti-concentration* bounds for the vanilla importance weight (IW) estimator $\widehat{\mu}_n$ (Section 4.1) and the self-normalized importance weight (SN) estimator $\widetilde{\mu}_n$ (Section 4.2), showing that the polynomial concentration is tight.

4.1. Vanilla importance weighting estimator

Previous works have shown that the vanilla IW estimator $\widehat{\mu}_n$ admits *polynomial* concentration, under the assumption that the divergence $I_2(P\|Q)$, connected to the variance of the importance weight, is bounded [46]. The original result [46, Theorem 2] was based on Cantelli's inequality, which approaches Chebyshev's when the probability $\delta \to 0$. In the following, we derive a more general statement proving a Chebyshev-like inequality, under the assumption that for some $\alpha \in (1,2]$, the divergence $I_\alpha(P\|Q)$ is finite and f is bounded.[4]

[3] We use the asymptotic notations Ω and $\mathcal{O}$ to retain dependences on $\log\left(\frac{1}{\delta}\right)$, n, and $I_\alpha(P\|Q)$ only.

[4] Theorem 4.1 and Theorem 4.3 correct the constant in the bound compared to the one present in the original paper [48].

Theorem 4.1 *(Polynomial Concentration of IW Estimator). Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability measures such that $P \ll Q$ and such that for some $\alpha \in (1,2]$, it holds that $I_\alpha(P\|Q) < +\infty$, and let $f : \mathcal{Y} \to \mathbb{R}$ be a bounded measurable function. Let $\{y_i\}_{i\in[\![n]\!]}$ be sampled independently from Q, then for every $\delta \in (0,1)$, with probability at least $1-\delta$ it holds that:*

$$\left|\widehat{\mu}_n - \mu\right| \leq 4\|f\|_\infty \left(\frac{I_\alpha(P\|Q)}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}. \tag{8}$$

Proof. We first derive the following inequality concerning the α-absolute central moment of the estimator $\widehat{\mu}_n$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{y_i\sim Q}\left[\left|\widehat{\mu}_n - \mu\right|^\alpha\right] &\leq (3-\alpha) \mathop{\mathbb{E}}_{y_i\sim Q}\left[\left|\widehat{\mu}_n\right|^\alpha\right] && \text{(P.16)}\\
&= (3-\alpha)\frac{1}{n^\alpha} \mathop{\mathbb{E}}_{y_i\sim Q}\left[\left|\sum_{i\in[\![n]\!]} \omega(y_i) f(y_i)\right|^\alpha\right] && \text{(P.17)}\\
&\leq \frac{(3-\alpha)2^{2-\alpha}}{n^{\alpha-1}} \mathop{\mathbb{E}}_{y\sim Q}\left[|\omega(y)f(y)|^\alpha\right] && \text{(P.18)}\\
&\leq \frac{(3-\alpha)2^{2-\alpha}}{n^{\alpha-1}} \|f\|_\infty^\alpha I_\alpha(P\|Q), && \text{(P.19)}
\end{aligned}$$

where line (P.16) derives from applying Lemma A.9 to bound the α-absolute central moment with the α-absolute raw moment, line (P.18) follows from bounding the α-absolute moment of the sum with the α-absolute moment of each addendum as in Equation (1.11) of Pinelis et al. [57] using as constant $W_\alpha = 2^{2-\alpha}$ of Proposition 1.8 of Pinelis et al. [57], and line (P.19) derives from applying Hölder's inequality and the definition of I_α. Now, we can derive the concentration inequality:

$$\begin{aligned}
\mathop{\mathbb{P}}_{y_i\sim Q}\left(\left|\widehat{\mu}_n - \mu\right| \geq \epsilon\right) &= \mathop{\mathbb{P}}_{y_i\sim Q}\left(\left|\widehat{\mu}_n - \mu\right|^\alpha \geq \epsilon^\alpha\right) && \text{(P.20)}\\
&\leq \frac{\mathbb{E}_{y_i\sim Q}\left[\left|\widehat{\mu}_n - \mu\right|^\alpha\right]}{\epsilon^\alpha} && \text{(P.21)}\\
&\leq \frac{(3-\alpha)2^{2-\alpha}}{n^{\alpha-1}\epsilon^\alpha} \|f\|_\infty^\alpha I_\alpha(P\|Q), && \text{(P.22)}
\end{aligned}$$

where line (P.21) derives from Markov's inequality. By setting the right-hand side of the last equation equal to δ and bounding $(2^{2-\alpha}(3-\alpha))^{1/\alpha} \leq 4$, we get the result. □

Thus, according to the definition given in Section 2.1, Theorem 4.1 provides a polynomial concentration with $\beta_1 = 1 - 1/\alpha$ and $\beta_2 = 1/\alpha$. In particular, for $\alpha = 2$, we obtain the usual Chebyshev's inequality.

Remark 4.1 *(High-Probability vs Mean-Squared-Error Bounds).* As already mentioned, a different approach, often employed in the literature, consists in proving lower bounds expressed in MSE $\mathbb{E}_{y_i\sim Q}[(\widehat{\mu}_n - \mu)^2]$. As noted in [42], when the estimator is not well-concentrated around its mean (e.g., in the presence of heavy tails), the MSE is not adequate to capture the error, and high-probability bounds are more advisable. Indeed, by looking at Equation (P.19), setting $\alpha = 2$, we immediately observe that the IW estimator enjoys the following MSE bound:

$$\mathop{\mathbb{E}}_{y_i\sim Q}\left[\left(\widehat{\mu}_n - \mu\right)^2\right] \leq \|f\|_\infty^2 \frac{I_2(P\|Q)}{n}. \tag{9}$$

This matches the MSE minimax lower bound of Equation (7) up to multiplicative constant terms. Conversely, looking at the high-probability results, comparing the minimax lower bound of Theorem 3.1 and the upper bound of Theorem 4.1, we realize that they do not match in the dependence on δ, which is subgaussian in the former and polynomial in the latter. This further justifies the need for focusing on high-probability results rather than expectation ones when heavy-tailed distributions are concerned.

We now show that the concentration rate provided in Theorem 4.1 is tight, by deriving an anti-concentration bound for the difference $\left|\widehat{\mu}_n - \mu\right|$. Intuitively, we aim at proving the existence of two probability measures $P, Q \in \Delta^{\mathcal{Y}}$, having fixed a divergence $I_\alpha(P\|Q)$, for which the polynomial concentration rate of Theorem 4.1 is attained, up to constant factors.

Theorem 4.2 *(Anti-concentration of IW Estimator). There exist two distributions $P, Q \in \Delta^{\mathcal{Y}}$ 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 $n \geq \delta e \max\left\{1, \left(I_\alpha(P\|Q) - 1\right)^{\frac{1}{\alpha-1}}\right\}$, with probability at least δ it holds that:*

$$\left|\widehat{\mu}_n - \mu\right| \geq \|f\|_\infty \left(\frac{I_\alpha(P\|Q) - 1}{e\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}. \tag{10}$$

Proof Sketch. We construct a function f and two probability measures P and Q that fulfill the inequality. Let $\|f\|_\infty =: a > 0$, we consider $\mathcal{Y} = \{-a, 0, a\}$ and $f(y) = y$. We now define the probability distributions as follows:

$$P(\{-a\}) = P(\{a\}) = \frac{p}{2} \text{ and } P(\{0\}) = 1 - p, \tag{P.23}$$

$$Q(\{-a\}) = Q(\{a\}) = \frac{q}{2} \text{ and } Q(\{0\}) = 1 - q, \tag{P.24}$$

where $p = \left(\frac{a}{n\epsilon}\right)^{\alpha-1} (I_\alpha(P\|Q) - 1)$ and $q = \left(\frac{a}{n\epsilon}\right)^{\alpha} (I_\alpha(P\|Q) - 1)$, so that the divergence between P and Q is indeed $I_\alpha(P\|Q)$. Let us consider the vanilla IW estimator $\widehat{\mu}_n$, whose expectation is $\mu = 0$, and observe that $\mathbb{P}_{y_i\sim Q}\left(\left|\widehat{\mu}_n - \mu\right| \geq \epsilon\right) = 2\,\mathbb{P}_{y_i\sim Q}\left(\widehat{\mu}_n - \mu \geq \epsilon\right)$. We now consider the event $\mathcal{E}$ under which among the n samples, one is a and the remaining are 0:

$$\mathcal{E} := \left\{ \left|\left\{i \in [\![n]\!] : y_i = 0\right\}\right| = n - 1 \wedge \left|\left\{i \in [\![n]\!] : y_i = a\right\}\right| = 1 \right\}. \tag{P.25}$$

It is immediate to verify that, if event $\mathcal{E}$ occurs, we have that $\widehat{\mu}_n = \frac{pa}{qn} = \epsilon$ and, consequently, $\widehat{\mu}_n - \mu \geq \epsilon$. Thus, we now lower bound the probability:

$$\mathop{\mathbb{P}}_{y_i\sim Q}\left(\widehat{\mu}_n - \mu \geq \epsilon\right) \geq \mathop{\mathbb{P}}_{y_i\sim Q}(\mathcal{E}) = n\frac{q}{2}(1-q)^{n-1}. \tag{P.26}$$

At this point, we set the right-hand side to δ and solve for ϵ. Further mild conditions are to be enforced in order to ensure that all quantities are well-defined. The full proof is reported in Appendix A.1. □

Some observations are in order. First of all, we note that the anti-concentration bound in Theorem 4.2 displays the same order dependence on n and δ as the upper bound in Theorem 4.1. It is worth noting that the lower bound is vacuous when $I_\alpha(P\|Q) = 1$, i.e., when $P = Q$ a.s., since in an on-distribution setting and, being the function f bounded, subgaussian concentration bounds, like Hoeffding's inequality, hold. In particular, for $\alpha = 2$, n and $I_2(P\|Q)$ sufficiently large, the bound has order $\mathcal{O}\left(\sqrt{\frac{I_2(P\|Q)}{\delta n}}\right)$, matching Chebyshev's and the existing concentration inequalities for vanilla IW [45,46].

4.2. Self-normalized importance weighting estimator

We now move to the self-normalized importance weighting (SN) estimator. The literature has studied the concentration of this estimator by either providing polynomial concentration inequalities [45, Proposition D.3] or almost-exponential concentration inequalities [35] based on Efron-Stein arguments, still dependent on some unknown quantities that might compromise the concentration rate. We start by deriving a concentration bound for $\widetilde{\mu}_n$ under the assumption that for some $\alpha \in (1, 2]$ the divergence $I_\alpha(P\|Q)$ and f are bounded.

Theorem 4.3 *(Polynomial Concentration of SN Estimator). Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability measures such that $P \ll Q$ and such that for some $\alpha \in (1, 2]$, it holds that $I_\alpha(P\|Q) < +\infty$, and let $f : \mathcal{Y} \to \mathbb{R}$ be a bounded measurable function. Let $\{y_i\}_{i\in[\![n]\!]}$ be sampled independently from Q, then for every $\delta \in (0, 1)$, with probability at least $1 - \delta$ it holds that:*

$$\left|\widetilde{\mu}_n - \mu\right| \leq 2\|f\|_\infty \min\left\{1, 2\left(\frac{I_\alpha(P\|Q)}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\right\}. \tag{11}$$

Proof. We reduce the concentration of the SN estimator $\widetilde{\mu}_n$ to the concentration of the vanilla estimator $\widehat{\mu}_n$:

$$\widetilde{\mu}_n - \mu = \widetilde{\mu}_n - \mu \pm \widehat{\mu}_n = \widetilde{\mu}_n\left(1 - \frac{1}{n}\sum_{i\in[\![n]\!]}\omega(y_i)\right) + \left(\frac{1}{n}\sum_{i\in[\![n]\!]}\omega(y_i)f(y_i) - \mu\right). \tag{P.27}$$

We first derive the following inequality concerning the α-absolute central moment of the estimator $\widehat{\mu}_n$:

$$\mathop{\mathbb{E}}_{y_i\sim Q}\left[\left|\widetilde{\mu}_n - \mu\right|^\alpha\right] = \mathop{\mathbb{E}}_{y_i\sim Q}\left[\left|\widetilde{\mu}_n\left(1 - \frac{1}{n}\sum_{i\in[\![n]\!]}\omega(y_i)\right) + \left(\frac{1}{n}\sum_{i\in[\![n]\!]}\omega(y_i)f(y_i) - \mu\right)\right|^\alpha\right] \tag{P.28}$$

$$\leq 2^{\alpha-1}\mathop{\mathbb{E}}_{y_i\sim Q}\left[\left|\widetilde{\mu}_n\right|^\alpha\left|1 - \frac{1}{n}\sum_{i\in[\![n]\!]}\omega(y_i)\right|^\alpha + \left|\frac{1}{n}\sum_{i\in[\![n]\!]}\omega(y_i)f(y_i) - \mu\right|^\alpha\right] \tag{P.29}$$

$$\leq 2^{\alpha-1}\|f\|_\infty^\alpha \mathop{\mathbb{E}}_{y_i\sim Q}\left[\left|1 - \frac{1}{n}\sum_{i\in[\![n]\!]}\omega(y_i)\right|^\alpha\right] + 2^{\alpha-1}\mathop{\mathbb{E}}_{y_i\sim Q}\left[\left|\frac{1}{n}\sum_{i\in[\![n]\!]}\omega(y_i)f(y_i) - \mu\right|^\alpha\right] \tag{P.30}$$

$$\leq 2^{\alpha-1}(3-\alpha)\|f\|_\infty^\alpha \mathop{\mathbb{E}}_{y_i\sim Q}\left[\left|\frac{1}{n}\sum_{i\in[\![n]\!]}\omega(y_i)\right|^\alpha\right] + 2^{\alpha-1}(3-\alpha)\mathop{\mathbb{E}}_{y_i\sim Q}\left[\left|\frac{1}{n}\sum_{i\in[\![n]\!]}\omega(y_i)f(y_i)\right|^\alpha\right] \tag{P.31}$$

$$\leq \frac{2(3-\alpha)}{n^{\alpha-1}}\|f\|_\infty^\alpha I_\alpha(P\|Q), \tag{P.32}$$

where line (P.29) derives from observing that $|a+b|^\alpha \leq 2^{\alpha-1}(|a|^\alpha + |b|^\alpha)$ (Lemma A.7), line (P.30) follows from observing that $|\widetilde{\mu}_n| \leq \|f\|_\infty$, line (P.31) is obtained from bounding the α-absolute central moments with Lemma A.9, and line (P.32) is an application of Equation (1.11) of Pinelis et al. [57] using as constant $W_\alpha = 2^{2-\alpha}$ of Proposition 1.8 of Pinelis et al. [57] and based on the definition of I_α. Now we can derive the concentration inequality:

$$\mathop{\mathbb{P}}_{y_i\sim Q}\left(\left|\widetilde{\mu}_n - \mu\right| \geq \epsilon\right) = \mathop{\mathbb{P}}_{y_i\sim Q}\left(\left|\widetilde{\mu}_n - \mu\right|^\alpha \geq \epsilon^\alpha\right) \tag{P.33}$$

$$\leq \frac{\mathbb{E}_{y_i\sim Q}\left[\left|\widetilde{\mu}_n - \mu\right|^\alpha\right]}{\epsilon^\alpha} \tag{P.34}$$

$$\leq \frac{2(3-\alpha)}{\epsilon^\alpha n^{\alpha-1}}\|f\|_\infty^\alpha I_\alpha(P\|Q), \tag{P.35}$$

where line (P.34) derives from Markov's inequality. By setting the right-hand side of the last equation equal to δ, recalling hat $(2(3-\alpha))^{1/\alpha} \leq 4$, and observing that $\left|\widetilde{\mu}_n - \mu\right| \leq 2\|f\|_\infty$, we get the result. □

Compared to the upper bound of Theorem 4.1, we have the same dependence on δ, $I_\alpha(P\|Q)$ and n. However, the bound of Theorem 4.3 is tighter for large values of $I_\alpha(P\|Q)$ or for small values of δ and n thanks to the presence of the minimum, which accounts for the boundedness of the SN estimator $\widetilde{\mu}_n$.

We now show that, apart from multiplicative constants, the polynomial concentration in Theorem 4.3 is tight for the SN estimator by deriving the following anti-concentration bound.

Theorem 4.4 *(Anti-concentration of SN Estimator). There exist two distributions $P, Q \in \Delta^{\mathcal{Y}}$ 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 $n \geq \max\left\{\frac{\delta e}{I_\alpha(P\|Q)-1}, \left(\frac{I_\alpha(P\|Q)-1}{\delta}\right)^{\frac{1}{\alpha-1}}\right\}$, with probability at least δ it holds that:*

$$\left|\widetilde{\mu}_n - \mu\right| \geq \frac{\|f\|_\infty}{2}\left(\frac{I_\alpha(P\|Q)-1}{e\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}. \tag{12}$$

Thus, Theorem 4.4 illustrates that the SN estimator is characterized by the same anti-concentration rate as the vanilla IW estimator, at least for a specific regime of number of samples n. The proof relies on a similar construction of the distributions as in Theorem 4.2. The next section will show how a general transformation of the importance weights allows overcoming this undesirable concentration rate.

5. Power-mean correction of importance weighting

In this section, motivated by the negative results of Theorem 4.2 and Theorem 4.4, we devise a weight correction able to achieve subgaussian concentration. We start by introducing a class of corrections based on the notion of *power mean* [9] (Section 5.1) and study its properties in terms of concentration (Section 5.2) and differentiability (Section 5.3).

5.1. Power mean importance weight

Let us start with the following definition that introduces the notion of *power mean importance weight* (PM), which will be employed in the following analyses.

Definition 5.1 *(Power Mean Importance Weight).* Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability distributions such that $P \ll Q$, for every $s \in [-\infty, \infty]$ and $\lambda \in [0,1]$, let $\omega(y) = \frac{p(y)}{q(y)}$, then the (λ, s)-*corrected* power mean importance weight is defined as:

$$\omega_{\lambda,s}(y) := \Big((1-\lambda)\omega(y)^s + \lambda\Big)^{\frac{1}{s}}, \quad \forall y \in \mathcal{Y}. \tag{13}$$

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.[5] This transformation can be seen as a generalization of IW, as, regardless of the value of s, for

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

Table 1
Some notable choices of s for the (λ, s)-corrected importance weight of Definition 5.1.

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

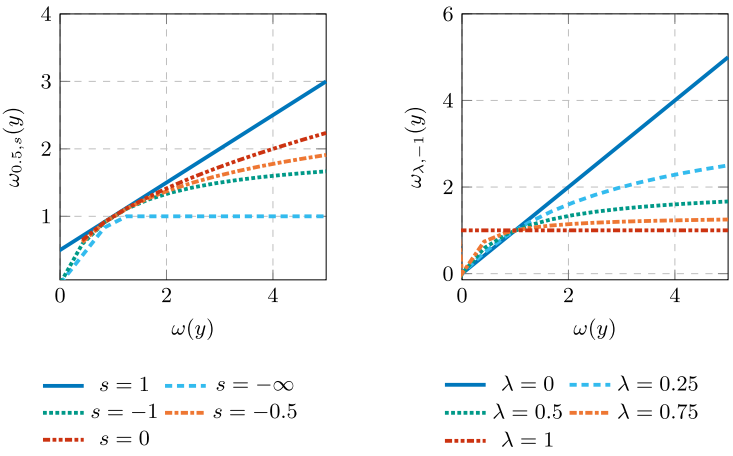

Fig. 1. Examples of importance weight corrections of Definition 5.1 for fixed λ (left) and fixed s (right).

$\lambda = 0$, we reduce to the vanilla importance weight $\omega_{0,s}(y) = \omega(y)$. By setting, instead, $\lambda = 1$, we have identically $\omega_{1,s}(y) = 1$. Another key property is the unbiasedness of the correction 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. The parameters s and λ govern the "intensity" of the correction in a continuous way. We note that the intensity of the correction increases as the value of λ moves towards 1 and the value of s moves away from 1. The smoothness of the transformation leads to a differentiable weight, unlike the truncation [29]. Some specific choices of s and λ are reported in Table 1 and in Fig. 1. The following result provides a preliminary characterization of the correction, independent of the properties of the two distributions.

Lemma 5.1. *Let $P, Q \in \Delta^{\mathcal{Y}}$ 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 $s \le s'$ then $\omega_{\lambda,s}(y) \le \omega_{\lambda,s'}(y)$;

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

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

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 IW derives from the heavy-tailed properties [45]. 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. Therefore, if we are looking for subgaussianity, we should restrict our attention to $s < 0$, which leads to lighter tails or even bounded weights. Specifically, in the rest of the paper, we focus on the sub-case $s \le -1$, leading to a well-behaved and analytically convenient form of the corrected weight.

5.2. Concentration properties

In this section, we study the class of PM importance weight estimators defined for $\lambda \in [0,1]$ and $s \le 0$, that we abbreviate as PM estimator:

$$\widehat{\mu}_{n,\lambda,s} := \frac{1}{n} \sum_{i \in [\![n]\!]} \omega_{\lambda,s}(y_i) f(y_i), \tag{14}$$

where $\{y_i\}_{i \in [\![n]\!]}$ are collected independently from Q. Then, we provide an exponential concentration inequality that, under certain circumstances, leads to subgaussian concentration. We start by providing an analysis of the bias and variance of the PM estimators (Section 5.2.1) and, then, we apply these results to obtain the desired concentration inequalities (Section 5.2.2).

5.2.1. Bias and variance analysis

Let us start by presenting the bias bound, holding under the assumption that the importance weight admits a finite moment of order $\alpha \in (1,2]$.

Lemma 5.2. *Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability distributions with $P \ll Q$. For every $\lambda \in [0,1]$ and $s \in [-\infty, -1]$, the (λ, s)-corrected importance weight induces a bias that is bounded for every $\alpha \in (1,2]$ as:*

$$\left| \mathop{\mathbb{E}}_{y_i \sim Q} \left[\widehat{\mu}_{n,\lambda,s} \right] - \mu \right| \leq \|f\|_\infty \lambda^{\frac{1-\alpha}{s}} (3-\alpha)^{\frac{1}{\alpha}} I_\alpha(P\|Q). \tag{15}$$

Proof. Let us consider the following derivation:

$$\begin{aligned}
\left| \mathop{\mathbb{E}}_{y_i \sim Q} \left[\widehat{\mu}_{n,\lambda,s} \right] - \mu \right| &= \left| \mathop{\mathbb{E}}_{y_i \sim Q} \left[\widehat{\mu}_{n,\lambda,s} - \widehat{\mu}_n \right] \right| && \text{(P.36)} \\
&\leq \mathop{\mathbb{E}}_{y_i \sim Q} \left[\left| \widehat{\mu}_{n,\lambda,s} - \widehat{\mu}_n \right| \right] && \text{(P.37)} \\
&\leq \|f\|_\infty \mathop{\mathbb{E}}_{y \sim Q} \left[\left| \omega_{\lambda,s}(y) - \omega(y) \right| \right], && \text{(P.38)}
\end{aligned}$$

where we applied Jensen's and Hölder's inequalities. Exploiting the definition of the (λ, s)-corrected importance weight and using the symbol $z = -s > 0$ for clarity of derivation, we have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{y \sim Q} \left[\left| \omega_{\lambda,s}(y) - \omega(y) \right| \right] &= \mathop{\mathbb{E}}_{y \sim Q} \left[\left| ((1-\lambda)\omega(y)^s + \lambda)^{\frac{1}{s}} - \omega(y) \right| \right] && \text{(P.39)} \\
&= \mathop{\mathbb{E}}_{y \sim Q} \left[\left| \frac{1}{\left(\frac{1-\lambda}{\omega(y)^z} + \lambda \right)^{\frac{1}{z}}} - \omega(y) \right| \right] && \text{(P.40)} \\
&= \mathop{\mathbb{E}}_{y \sim Q} \left[\left| \frac{1 - (1 + \lambda(\omega(y)^z - 1))^{\frac{1}{z}}}{\left(\frac{1-\lambda}{\omega(y)^z} + \lambda \right)^{\frac{1}{z}}} \right| \right] && \text{(P.41)} \\
&\leq \sup_{v \geq 0} \left(\frac{1}{\frac{1-\lambda}{v^z} + \lambda} \right)^{\frac{2-\alpha}{z}} \mathop{\mathbb{E}}_{y \sim Q} \left[\left| 1 - (1 + \lambda(\omega(y)^z - 1))^{\frac{1}{z}} \right| \left(\frac{1}{\frac{1-\lambda}{\omega(y)^z} + \lambda} \right)^{\frac{\alpha-1}{z}} \right] && \text{(P.42)} \\
&\leq \frac{1}{\lambda^{\frac{2-\alpha}{z}}} \underbrace{\mathop{\mathbb{E}}_{y \sim Q} \left[\left| 1 - (1 + \lambda(\omega(y)^z - 1))^{\frac{1}{z}} \right|^\alpha \right]^{\frac{1}{\alpha}}}_{\text{(a)}} \underbrace{\mathop{\mathbb{E}}_{y \sim Q} \left[\left(\frac{1}{\frac{1-\lambda}{\omega(y)^z} + \lambda} \right)^{\frac{\alpha}{z}} \right]^{\frac{\alpha-1}{\alpha}}}_{\text{(b)}}, && \text{(P.43)}
\end{aligned}$$

where line (P.42) follows from factorizing the denominator expression as $(\cdot) = (\cdot)^{2-\alpha}(\cdot)^{\alpha-1}$ and maximizing the first factor over $v = \omega(y)$, line (P.43) is obtained from Hölder's inequality with $p = \alpha$ and $q = \frac{\alpha}{\alpha-1}$ and by observing that since $z > 0$, function $\frac{1}{\frac{1-\lambda}{v^z} + \lambda}$ is monotonically increasing in v and, consequently:

$$\sup_{v \geq 0} \left(\frac{1}{\frac{1-\lambda}{v^z} + \lambda} \right)^{\frac{2-\alpha}{z}} = \lim_{v \to \infty} \left(\frac{1}{\frac{1-\lambda}{v^z} + \lambda} \right)^{\frac{2-\alpha}{z}} = \frac{1}{\lambda^{\frac{2-\alpha}{z}}}. \tag{P.44}$$

We now proceed with terms (a) and (b). Let us consider the term inside the expectation of (a) that can be bounded thanks to Lemma A.10:

$$\left| 1 - (1 + \lambda(\omega(y)^z - 1))^{\frac{1}{z}} \right| \leq \lambda^{\frac{1}{z}} \left| \omega(y) - 1 \right|. \tag{P.45}$$

Thus, term (a) becomes:

$$\begin{aligned}
\mathop{\mathbb{E}}_{y \sim Q} \left[\left| 1 - (1 + \lambda(\omega(y)^z - 1))^{\frac{1}{z}} \right|^\alpha \right] &\leq \lambda^{\frac{\alpha}{z}} \mathop{\mathbb{E}}_{y \sim Q} \left[|\omega(y) - 1|^\alpha \right] && \text{(P.46)} \\
&\leq \lambda^{\frac{\alpha}{z}} (3-\alpha) \mathop{\mathbb{E}}_{y \sim Q} \left[\omega(y)^\alpha \right] - \lambda^{\frac{\alpha}{z}} (\alpha - 1) && \text{(P.47)} \\
&\leq \lambda^{\frac{\alpha}{z}} (3-\alpha) \mathop{\mathbb{E}}_{y \sim Q} \left[\omega(y)^\alpha \right], && \text{(P.48)}
\end{aligned}$$

where we exploited Lemma A.9 in line (P.47). Let us consider term (b), we proceed as follows:

$$\mathop{\mathbb{E}}_{y\sim Q}\left[\left(\frac{1}{\frac{1-\lambda}{\omega(y)^z}+\lambda}\right)^{\frac{\alpha}{z}}\right] \le \mathop{\mathbb{E}}_{y\sim Q}\left[((1-\lambda)\omega(y)+\lambda)^{\alpha}\right] \tag{P.49}$$

$$\le \mathop{\mathbb{E}}_{y\sim Q}\left[(1-\lambda)\omega(y)^{\alpha}+\lambda\right] \tag{P.50}$$

$$= (1-\lambda)I_{\alpha}(P\|Q)+\lambda \tag{P.51}$$

$$\le I_{\alpha}(P\|Q), \tag{P.52}$$

where line (P.49) is obtained from the power mean inequality [9] by bounding the mean of order $s<0$ with the arithmetic mean and line (P.50) is obtained from Jensen's inequality recalling that $\alpha>1$. By combining everything, we obtain the result. □

As expected, the estimator presents zero bias for $\lambda=0$ and it increases with λ (note that the exponent $\frac{1-\alpha}{s}$ is non-negative) and with the divergence term $I_{\alpha}(P\|Q)$. For the specific case of $P=Q$ a.s., we already observed that the bias is null as the weight is always unitary. This phenomenon is not captured by our bound, which does not vanish when $P=Q$ because of a bound performed in the derivation (see Equation (P.47)). However, as we shall see, this feature does not have a significant effect on the overall concentration properties of the estimator. In particular, for $\alpha=2$, the bound becomes $\|f\|_{\infty}\lambda^{-\frac{1}{s}}I_2(P\|Q)$. Let us turn to the bound on the variance.

Lemma 5.3. *Let $P,Q\in\Delta^{\mathcal{Y}}$ be two probability distributions with $P\ll Q$. For every $\lambda\in[0,1]$ and $s\in[-\infty,0)$, the (λ,s)-corrected importance weight induces a variance that is bounded for every $\alpha\in(1,2]$ as:*

$$\mathop{\mathbb{V}\mathrm{ar}}_{y_i\sim Q}\left[\widehat{\mu}_{n,\lambda,s}\right] \le \frac{\|f\|_{\infty}^2}{\lambda^{\frac{\alpha-2}{s}}n}I_{\alpha}(P\|Q). \tag{16}$$

Proof. Let us consider the following derivation:

$$\mathop{\mathbb{V}\mathrm{ar}}_{y_i\sim Q}\left[\widehat{\mu}_{n,\lambda,s}\right] = \frac{1}{n}\mathop{\mathbb{V}\mathrm{ar}}_{y\sim Q}\left[\omega_{\lambda,s}(y)f(y)\right] \le \frac{1}{n}\mathop{\mathbb{E}}_{y\sim Q}\left[\omega_{\lambda,s}(y)^2f(y)^2\right] \le \frac{1}{n}\|f\|_{\infty}^2\mathop{\mathbb{E}}_{y\sim Q}\left[\omega_{\lambda,s}(y)^2\right]. \tag{P.53}$$

Exploiting the definition of the (λ,s)-corrected importance weight and using the symbol $z=-s>0$ for clarity of derivation, we have:

$$\mathop{\mathbb{E}}_{y\sim Q}\left[\omega_{\lambda,s}(y)^2\right] = \mathop{\mathbb{E}}_{y\sim Q}\left[((1-\lambda)\omega(y)^s+\lambda)^{\frac{2}{s}}\right] \tag{P.54}$$

$$= \mathop{\mathbb{E}}_{y\sim Q}\left[\left(\frac{1}{\frac{1-\lambda}{\omega(y)^z}+\lambda}\right)^{\frac{2}{z}}\right] \tag{P.55}$$

$$= \mathop{\mathbb{E}}_{y\sim Q}\left[\left(\frac{1}{\frac{1-\lambda}{\omega(y)^z}+\lambda}\right)^{\frac{\alpha}{z}}\left(\frac{1}{\frac{1-\lambda}{\omega(y)^z}+\lambda}\right)^{\frac{2-\alpha}{z}}\right] \tag{P.56}$$

$$\le \sup_{v\ge 0}\left(\frac{1}{\frac{1-\lambda}{v^z}+\lambda}\right)^{\frac{2-\alpha}{z}}\mathop{\mathbb{E}}_{y\sim Q}\left[\left(\frac{1}{\frac{1-\lambda}{\omega(y)^z}+\lambda}\right)^{\frac{\alpha}{z}}\right] \tag{P.57}$$

$$\le \frac{1}{\lambda^{\frac{2-\alpha}{z}}}\mathop{\mathbb{E}}_{y\sim Q}\left[((1-\lambda)+\lambda\omega(y))^{\alpha}\right] \tag{P.58}$$

$$\le \frac{1}{\lambda^{\frac{2-\alpha}{z}}}\mathop{\mathbb{E}}_{y\sim Q}\left[(1-\lambda)+\lambda\omega(y)^{\alpha}\right] \tag{P.59}$$

$$= \frac{1}{\lambda^{\frac{2-\alpha}{z}}}\left[(1-\lambda)+\lambda I_{\alpha}(P\|Q)\right] \tag{P.60}$$

$$\le \frac{1}{\lambda^{\frac{2-\alpha}{z}}}I_{\alpha}(P\|Q), \tag{P.61}$$

where line (P.58) is obtained from solving the maximization having observed that $z>0$ and by bounding the generalized mean of order s with the arithmetic mean, line (P.59) is obtained from Jensen's inequality since $\alpha\ge 1$, line (P.60) follows from the definition of $I_{\alpha}(P\|Q)$, and for line (P.61) we simply observe that $I_{\alpha}(P\|Q)\ge 1$. □

Different from the bias, the variance bound decreases in λ (note that the exponent $\frac{\alpha-2}{s}$ is non-negative) and increases with the divergence $I_{\alpha}(P\|Q)$. For $\alpha=2$, we obtain the bound $\frac{1}{n}\|f\|_{\infty}^2I_2(P\|Q)$. Note that when $P=Q$ a.s., we recover $\frac{1}{n}\|f\|_{\infty}^2$, which is

the Popoviciu's inequality for the variance [58]. From the derived results, we can see that our weight correction allows controlling bias and variance even for $I_2(P\|Q) = +\infty$, i.e., when the vanilla IW estimator might have infinite variance. Indeed, our transformed estimator has finite variance provided that there exists $\alpha \in (1,2)$ so that $I_\alpha(P\|Q) < +\infty$.

5.2.2. Concentration inequality

We are now ready to derive the core theoretical result. We prove an exponential concentration inequality for the (λ, s)-corrected PM importance weighting estimator, and we show that, if $I_2(P\|Q)$ is finite, we are able to achieve subgaussian concentration.[6] We start by providing a concentration inequality that holds for every $s \in [-\infty, -1]$ and $\lambda \in [0,1]$.

Lemma 5.4. *Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability distributions such that $P \ll Q$. Let $\{y_i\}_{i\in[\![n]\!]}$ be sampled independently from Q. For every $\alpha \in (1,2]$, $s \in [-\infty, -1]$, $\lambda \in [0,1]$, and $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that:*

$$\widehat{\mu}_{n,\lambda,s} - \mu \leq \|f\|_\infty \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n\lambda^{\frac{\alpha-2}{s}}} I_\alpha(P\|Q)} + \frac{2\|f\|_\infty \lambda^{\frac{1}{s}} \log\left(\frac{1}{\delta}\right)}{3n} + \|f\|_\infty (3-\alpha)^{\frac{1}{\alpha}} \lambda^{\frac{1-\alpha}{s}} I_\alpha(P\|Q). \tag{17}$$

Proof. The proof is a straightforward application of Bernstein's inequality [6] together with Lemma 5.2 and Lemma 5.3. First of all, we highlight the bias in the following decomposition:

$$\widehat{\mu}_{n,\lambda,s} - \mu = \underbrace{\widehat{\mu}_{n,\lambda,s} - \mathop{\mathbb{E}}_{y_i\sim Q}[\widehat{\mu}_{n,\lambda,s}]}_{\text{concentration}} + \underbrace{\mathop{\mathbb{E}}_{y_i\sim Q}[\widehat{\mu}_{n,\lambda,s}] - \mu}_{\text{bias}}. \tag{P.62}$$

The bias term is bounded by using Lemma 5.2, while, for the concentration term, we apply Bernstein's inequality. Let $\delta \in (0,1)$, with probability at least $1-\delta$ it holds that:

$$\widehat{\mu}_{n,\lambda,s} - \mathop{\mathbb{E}}_{y_i\sim Q}[\widehat{\mu}_{n,\lambda,s}] \leq \sqrt{2\mathop{\mathbb{V}\mathrm{ar}}_{y_i\sim Q}[\widehat{\mu}_{n,\lambda,s}]\log\left(\frac{1}{\delta}\right)} + \frac{2\sup_{\{y_i\}_{y\in[\![n]\!]}} |\mu_{n,\lambda,s}| \log\left(\frac{1}{\delta}\right)}{3n} \tag{P.63}$$

$$\leq \|f\|_\infty \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n\lambda^{\frac{\alpha-2}{s}}} I_\alpha(P\|Q)} + \frac{2\|f\|_\infty \lambda^{\frac{1}{s}} \log\left(\frac{1}{\delta}\right)}{3n}, \tag{P.64}$$

where the last line is obtained by bounding the variance with Lemma 5.3 and recalling that $\omega_{\lambda,s}(y_i) f(y_i) \leq \|f\|_\infty \lambda^{\frac{1}{s}}$ for every $i \in [\![n]\!]$. □

Given the general result provided by Lemma 5.4, we can now optimize over the pair (λ, s) in order to achieve a tighter bound.

Theorem 5.1. *Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability distributions such that $P \ll Q$. Let $\{y_i\}_{i\in[\![n]\!]}$ be sampled independently from Q. For every $\alpha \in (1,2]$ and $\delta \in (0,1)$ if:*

$$n \geq \frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 I_\alpha(P\|Q)}, \tag{18}$$

then, with probability at least $1-\delta$, it holds that:

$$\widehat{\mu}_{n,\lambda^*_\alpha,s^*_\alpha} - \mu \leq \frac{7\|f\|_\infty}{3} \left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2}\right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q)^{\frac{1}{\alpha}}, \tag{19}$$

having selected $(\lambda^*_\alpha, s^*_\alpha)$ such that:

$$\lambda^*_\alpha = \left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2 I_\alpha(P\|Q)}\right)^{-\frac{s^*_\alpha}{\alpha}}. \tag{20}$$

[6] We introduce our concentration inequalities as one-sided bounds just for simplicity, but they actually hold in both directions. Indeed, by replacing function f with function $-f$, we obtain the reversed one-sided bound.

Proof Sketch. The proof is obtained by minimizing the bound provided in Lemma 5.4 over the variable $\lambda^{\frac{1}{s}}$. Since the function is convex in the variable $\lambda^{\frac{1}{s}}$, this can be done by vanishing the derivative. The complete proof is reported in Appendix A.2. □

Some observations are in order. First, we notice that the dependence on the confidence level δ is the one typical of exponential concentration for every $\alpha \in (1, 2]$ and matches the minimax lower bound of Theorem 3.1 up to multiplicative constants. Second, we note that the exponential concentration bound provided in Theorem 5.1 holds for a sufficiently large number of samples n only (Equation 18). Indeed, for a too small number of samples, the optimal value of $(\lambda_\alpha^*, s_\alpha^*)$ requires setting $\lambda_\alpha^* = 1$, leading to the following bound:

$$\widehat{\mu}_{n,1,s} - \mu \leq \|f\|_\infty \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n} I_\alpha(P\|Q) + \frac{2\|f\|_\infty \log\left(\frac{1}{\delta}\right)}{3n}} + \|f\|_\infty (3-\alpha)^{\frac{1}{\alpha}} I_\alpha(P\|Q). \tag{21}$$

Third, the choices of λ_α^* and s_α^* are related. Indeed, one can fix s_α^* and derive the corresponding λ_α^*, or vice versa.[7] In other words, it is possible to obtain exponential concentration for every choice of $s_\alpha^* \in (-\infty, -1]$. Fourth, 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 [16]. Specifically, for $\alpha = 2$, with an optimal choice of (λ_2^*, s_2^*), we obtain the bound:

$$\lambda_2^* = \left(\frac{2\log\left(\frac{1}{\delta}\right)}{I_2(P\|Q)n}\right)^{-\frac{s_2^*}{2}} \implies \widehat{\mu}_{n,\lambda_2^*,s_2^*} - \mu \leq \frac{7\|f\|_\infty}{3}\sqrt{\frac{2I_2(P\|Q)\log\left(\frac{1}{\delta}\right)}{n}}. \tag{22}$$

Finally, from our high-probability bound, we can obtain an MSE upper bound by applying the general results of Proposition A.4 to Theorem 5.1, leading to the bound:

$$\mathop{\mathbb{E}}_{y_i \sim Q}\left[\left(\widehat{\mu}_{n,\lambda_\alpha^*,s_\alpha^*} - \mu\right)^2\right] \leq \frac{49}{9}\|f\|_\infty^2 \left(\frac{2I_\alpha(P\|Q)}{n^{\alpha-1}}\right)^{\frac{2}{\alpha}} \int_{\delta=0}^{1} \left(\log\left(\frac{1}{\delta}\right)\right)^{\frac{2(\alpha-1)}{\alpha}} \mathrm{d}\delta \leq \frac{98}{9}\|f\|_\infty^2 \left(\frac{I_\alpha(P\|Q)}{n^{\alpha-1}}\right)^{\frac{2}{\alpha}}. \tag{23}$$

In particular, for $\alpha = 2$, we obtain an upper bound of order $\mathcal{O}\left(\frac{I_2(P\|Q)}{n}\right)$, which matches Equation (7), as expected.

5.3. Differentiability

As we have already observed, our weight correction, differently from truncation, is smooth and, thus, differentiable in the target policy. This property is particularly important for learning approaches that focus on differentiable parametric policies and update the parameters through gradient ascent or descent [66]. Furthermore, having a bounded gradient allows for more convenient theoretical analysis (e.g., convergence). Thus, we focus on the properties of the *gradient* of the (λ, s)-corrected estimator and, to this purpose, we constrain the target distribution to belong to a parametric space of differentiable distributions $\mathcal{P}_\Theta = \{P_\theta \in \Delta^{\mathcal{Y}} : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$. The gradient of the corrected weight $\omega_{\lambda,s}$ w.r.t. the target policy parameters θ is given by:

$$\nabla_\theta \omega_{\lambda,s}(y) = \frac{(1-\lambda)\omega(y)^s}{((1-\lambda)\omega(y)^s + \lambda)^{1-\frac{1}{s}}} \nabla_\theta \log p_\theta(y), \quad \forall y \in \mathcal{Y}. \tag{24}$$

In particular, the following inequality involving the L_∞-norm of the gradient, for every $s < 0$ and $\lambda \in [0, 1]$ holds (Proposition A.1):

$$\left\|\nabla_\theta \omega_{\lambda,s}(y)\right\|_\infty \leq \frac{-s}{\lambda^{-\frac{1}{s}}(1-s)^{1+\frac{1}{s}}} \left\|\nabla_\theta \log p_\theta(y)\right\|_\infty, \quad \forall y \in \mathcal{Y}. \tag{25}$$

Thus, if the score $\nabla_\theta \log p_\theta$ 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 IW ($\lambda = 0$). Thus, we can also interpret λ as a regularization parameter for the gradient magnitude.

Furthermore, by using the optimal value of (λ, s), as prescribed by Theorem 5.1, we obtain the following bound on the L_∞-norm of the gradient of the corrected importance weight:

$$\left\|\nabla_\theta \omega_{\lambda,s}(y)\right\|_\infty \leq \frac{-s}{(1-s)^{1+\frac{1}{s}}} \left(\frac{n(\alpha-1)^2 I_\alpha(P\|Q)}{2\log\left(\frac{1}{\delta}\right)}\right)^{\frac{1}{\alpha}} \left\|\nabla_\theta \log p_\theta(y)\right\|_\infty, \quad \forall y \in \mathcal{Y}. \tag{26}$$

[7] A proposal for limiting this degree of freedom is provided in Appendix C.

As expected, the bound enlarges as the number of samples n increases, since, in such a scenario, the regularization enforced by λ gets reduced.

6. Data-driven tuning of (λ, s)

As shown in Section 5.2, the computation of the optimal correction pair $(\lambda_\alpha^*, s_\alpha^*)$ requires the knowledge of the divergence $I_\alpha(P\|Q)$, which, in turn, requires access to the form of P and Q. In this section, we first briefly present some basic approaches to overcome this limitation (Section 6.1) and, then, we focus on a novel adaptive approach that we discuss in detail (Section 6.2).

6.1. Basic approaches

In principle, we may perform a choice of $(\lambda_\alpha, s_\alpha)$ independent of $I_\alpha(P\|Q)$ that leads to a sub-optimal dependence on $I_\alpha(P\|Q)$ in the concentration bound, as shown in Proposition A.2:

$$\lambda_\alpha^\ddagger = \left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}\right)^{-\frac{s_\alpha^\ddagger}{\alpha}} \implies \widehat{\mu}_{n,\lambda_\alpha^\ddagger,s_\alpha^\ddagger} - \mu \leq \frac{7\|f\|_\infty}{3}\left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2}\right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q), \tag{27}$$

Even when P and Q are known, computing the $I_\alpha(P\|Q)$ can be challenging, especially for continuous distributions, since it involves the evaluation of a complex integral, as in Equation (2).[8] In principle, we could estimate the divergence $I_\alpha(P\|Q)$ from samples as the empirical moment of order α of the vanilla importance weights $\frac{1}{n}\sum_{i\in[\![n]\!]}\omega(y_i)^\alpha$, as done, for the case $\alpha = 2$, in previous works [45]. However, although possibly well-performing in practice [45], this approach would prevent any subgaussian concentration, as the behavior of the non-corrected $\omega(y)^\alpha$ will be surely heavy-tailed whenever $\omega(y)$ is. A general-purpose approach to avoid the divergence estimation is the *Lepski's adaptation method* [38], which requires knowing a lower bound and an upper bound (that might not be available) on $I_\alpha(P\|Q)$. Furthermore, this method is known to be computationally intensive.

6.2. Adaptive approach

In this section, we follow a different path inspired by the recent work of [76]. If a choice of the pair (λ, s) corrects the weight $\omega_{\lambda,s}$ leading to an ideal estimator $\widehat{\mu}_{n,\lambda,s}$, for the mean μ, we may expect that the empirical moment of order α of the corrected weights $\omega_{\lambda,s}$ will provide a reasonable estimate of $I_\alpha(P\|Q)$. Since, as observed in Section 5.2, for every choice of $s \in [-\infty, -1]$, there exists an optimal value of λ, we restrict the search to the problem of learning the optimal value of λ. Specifically, we propose to choose λ by solving the following equation, parametric in $\beta \in [0, 1/\alpha]$ whose value will be discussed later, in the variable $\lambda \in (0, n^{-\beta})$:

$$\widehat{h}_{\alpha,s}(\lambda) := \lambda^{-\frac{\alpha}{s}} \underbrace{\frac{1}{n}\sum_{i\in[\![n]\!]}\omega_{\lambda n^\beta,s}(y_i)^\alpha}_{\substack{\text{empirical moment}\\ \text{of order } \alpha}} = \frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 n}. \tag{28}$$

The intuition behind this approach can be stated as follows. If the empirical α-moment is close to the divergence, i.e., $\frac{1}{n}\sum_{i\in[\![n]\!]}\omega_{\lambda n^\beta,s}(y_i)^\alpha \simeq I_\alpha(P\|Q)$, having fixed $s \in [-\infty, -1]$, the solution $\widehat{\lambda}_\alpha$ of Equation (28) approaches the optimal parameters, i.e., $\widehat{\lambda}_\alpha \approx \lambda_\alpha^*$. The role of $\beta \in [0, 1/\alpha]$ allows, as we shall see later, to ensure that the solution of Equation (28) approaches λ_α^* with a fast-enough rate. The following result provides a sufficient condition under which, in high probability, Equation (28) admits a unique solution.

Lemma 6.1. *Let $\alpha \in (1, 2]$, $s \in [-\infty, -1]$, and $\delta \in (0, 1)$. If*

$$n \geq \left(\frac{6 I_\alpha(P\|Q)^{\frac{1}{\alpha-1}}\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}\right)^{\frac{1}{1+\frac{\alpha\beta}{s}}}, \tag{29}$$

then, with probability at least $1-\delta$, Equation (28) admits exactly one solution.

If Equation (28) admits a unique solution $\widehat{\lambda}_\alpha$, it is possible to relate it with the optimal solution λ_α^* prescribed by Theorem 5.1.

Lemma 6.2. *Let $\alpha \in (1, 2]$, $s \in [-\infty, -1]$, and $\delta \in (0, 1)$. Suppose that Equation (28) admits a unique solution, denoted by $\widehat{\lambda}_\alpha$. Then, if $I_{\alpha-s}(P\|Q)$ is finite, for every $\epsilon_1, \epsilon_2 \in (0, 1)$, with probability at least $1 - 2\delta$ it holds that:*

[8] For some specific classes of distributions, including Gaussians, the integral can be computed in closed form [22].

Algorithm 1 Root finding for Equation (28).

Input: function $\widehat{h}_{\alpha,s}(\cdot)$, threshold $\epsilon_3 > 0$

1: $k \leftarrow 1$, $\lambda_-^{(1)} \leftarrow 0$, $\lambda_+^{(1)} \leftarrow n^{-\beta}$ ▷ Initialization
2: **while** $\frac{\lambda_+^{(k)}}{\lambda_-^{(k)}} > 1 + \epsilon_3$ **do**
3: $\quad\lambda^{(k)} \leftarrow \frac{\lambda_+^{(k)} + \lambda_-^{(k)}}{2}$ ▷ Compute candidate root
4: $\quad$**if** $\widehat{h}_{\alpha,s}(\lambda^{(k)}) = 0$ **then**
5: $\quad\quad$**return** $\lambda^{(k)}$
6: $\quad$**end if**
7: $\quad$**if** $\operatorname{sign}(\widehat{h}_{\alpha,s}(\lambda^{(k)})) = \operatorname{sign}(\widehat{h}_{\alpha,s}(\lambda_-^{(k)}))$ **then**
8: $\quad\quad\lambda_-^{(k+1)} \leftarrow \lambda^{(k)}$
9: $\quad$**else**
10: $\quad\quad\lambda_+^{(k+1)} \leftarrow \lambda^{(k)}$
11: $\quad$**end if**
12: $\quad k \leftarrow k + 1$
13: **end while**
14: **return** $\lambda^{(k-1)}$

$$1 - \epsilon_2 \leq \frac{\widehat{\lambda}_\alpha}{\lambda_\alpha^*} \leq (1 + \epsilon_1)(1 + \epsilon_2), \tag{30}$$

under the condition that:

$$n \geq \max\left\{\left(\frac{16}{3\epsilon_1}\left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}\right)^{-\frac{s}{\alpha}} \frac{I_{\alpha-s}(P\|Q) - I_\alpha(P\|Q)}{I_\alpha(P\|Q)^{1-\frac{s}{\alpha}}}\right)^{\frac{1}{-\frac{s}{\alpha}-\beta}}, \left(\frac{(\alpha-1)^2}{\epsilon_2^2} \cdot 2^{\frac{1-2s+\alpha}{\alpha-1}} I_\alpha(P\|Q)^{\frac{2(\alpha-s)}{\alpha-1}}\right)^{-\frac{s}{\alpha\beta}}\right\}. \tag{31}$$

Proof. The result follows by combining Lemma A.6 and A.3. □

Some observations are in order. First of all, Lemma 6.2 ensures that, with high probability, the solution of the empirical Equation (28) $\widehat{\lambda}_\alpha$ is close to the optimal parameter λ_α^* up to multiplicative factors. This requires, nevertheless, the finiteness of a higher-order dissimilarity index $I_{\alpha-s}(P\|Q)$.[9] The discrepancy between $\widehat{\lambda}_\alpha$ and λ_α^* depends on two variables $\epsilon_1, \epsilon_2 \in (0,1)$ that can be arbitrarily tuned and affect the minimum number of samples n to ensure that inequality in Equation (30) holds. In particular, Lemma 6.2 accounts for two sources of errors. First, *approximation error*, due to the fact that, even by replacing the sample mean with the expectation in Equation (28), the root of the equation is different from λ_α^*. Indeed, such a solution is always an overestimation of λ_α^*, although it approaches it as $n \to +\infty$, leading to the first argument of the maximum in Equation (31) (Lemma A.6). Second, *estimation error*, due to the fact that Equation (28) is constructed as a sample mean that is close to the expectation for a sufficiently large number of samples. It is worth mentioning that the analysis of the concentration properties of function $\widehat{h}_{\alpha,s}$ leverages arguments based on the *self-bounding* functions [6] and leads to the second argument of the maximum in Equation (31) (Lemma A.3). The condition on the minimum number of samples of Equation (31) allows appreciating the role of the hyperparameter β. Indeed, it appears at the exponent of both arguments of the maximum with opposite roles.

Equation (28) does not admit a closed-form, in general. In Algorithm 1, we provide a bisection-based method to numerically find the root of Equation (28). After initializing the iteration counter k and the extremes of the search interval $[\lambda_-^{(1)}, \lambda_+^{(1)}] = [0, n^{-\beta}]$ (line 1), the algorithm checks whether the extremes satisfy the condition $\frac{\lambda_+^{(k)}}{\lambda_-^{(k)}} > 1 + \epsilon_3$, where $\epsilon_3 \in (0,1)$ is a threshold parameter. As we shall see, this condition ensures that the proposed solution is sufficiently close to the true one (line 2). Then, a candidate solution $\lambda^{(k)}$ is computed and evaluated (lines 3-6). If the candidate solution is not a root, based on the sign of the function $\widehat{h}_{\alpha,s}(\lambda^{(k)})$, the extremes of the interval are updated (lines 7-11). The following result provides a minimum number of iterations for the Algorithm 1 to stop.

Lemma 6.3. *Let $\alpha \in (1,2]$, $s \in [-\infty, -1]$, and $\delta \in (0,1)$. Suppose that Equation (28) admits a unique solution $\widehat{\lambda}_\alpha$. Then, for every $\epsilon_3 \in (0,1)$, it holds that:*

$$1 - \epsilon_3 \leq \frac{\lambda^{(k)}}{\widehat{\lambda}_\alpha} \leq 1 + \epsilon_3, \tag{32}$$

for a number of iterations $k \geq \left\lceil \log_2\left(\frac{n^{-\beta}}{\widehat{\lambda}_\alpha \epsilon_3}\right)\right\rceil$.

[9] We are currently unsure whether this additional requirement is an artifact of the analysis or an intrinsic phenomenon due to the choice of the PM correction.

It is worth noting that in Lemma 6.3 the number of iterations depends inversely on $\widehat{\lambda}_\alpha$ that, from Lemma 6.2, can be bounded by a fraction of λ^*_α. This allows obtaining the following final concentration bound.

Theorem 6.1. *Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability distributions such that $P \ll Q$. Let $\widehat{\lambda}_\alpha$ be the solution of Equation (28), then, if $I_{\alpha-s}(P\|Q)$ is finite, for every $\delta \in (0,1)$, with probability at least $1-4\delta$, it holds that:*

$$\widehat{\mu}_{n,\lambda^{(k)},s} - \mu \leq \frac{14\|f\|_\infty}{3}\left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2}\right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q)^{\frac{1}{\alpha}}, \tag{33}$$

under the conditions that:

$$n \geq \max\left\{\left(19\left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}\right)^{-\frac{s}{\alpha}} \frac{I_{\alpha-s}(P\|Q) - I_\alpha(P\|Q)}{I_\alpha(P\|Q)^{1-\frac{s}{\alpha}}}\right)^{\frac{1}{-\frac{s}{\alpha}-\beta}}, \left(26(\alpha-1)^2 2^{\frac{1-2s+\alpha}{\alpha-1}} I_\alpha(P\|Q)^{\frac{2(\alpha-s)}{\alpha-1}}\right)^{-\frac{s}{\alpha\beta}}, \tag{34}$$

$$\left(\frac{6I_\alpha(P\|Q)^{\frac{1}{\alpha-1}}\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}\right)^{\frac{1}{1+\frac{\alpha\beta}{s}}}\right\}, \tag{35}$$

$$k \geq 4 + \left(-\beta - \frac{s}{\alpha}\right)\log_2 n - \frac{s}{\alpha}\log_2 \frac{(1-\alpha)^2 I_\alpha(P\|Q)}{2\log\left(\frac{1}{\delta}\right)}. \tag{36}$$

Compared to Theorem 5.1, this result is weakened in three aspects. First, the inequality holds with a smaller probability $1-4\delta$ since multiple estimation processes with the same samples are needed, i.e., the computation of $\widehat{\lambda}_\alpha$ and the corrected estimator $\widehat{\mu}_{n,\widehat{\lambda}_\alpha,s}$. Second, and most important, the result holds for a number of samples n, which is larger than that of Theorem 5.1. Third, the multiplicative constant in the error bound is twice that of Theorem 5.1.

To visualize a more explicit result, let us instantiate the minimum number of samples n of Theorem 5.1 with $\alpha = 2$:

$$n \geq \Omega\left(\max\left\{\left(\frac{(I_3(P\|Q) - I_2(P\|Q))^2}{I_2(P\|Q)^3}\log\left(\frac{1}{\delta}\right)\right)^{\frac{1}{1-2\beta}}, I_2(P\|Q)^{\frac{3}{\beta}}, \left(I_2(P\|Q)\log\left(\frac{1}{\delta}\right)\right)^{\frac{1}{1-2\beta}}\right\}\right). \tag{37}$$

Now, we clearly see the effect of choosing $\beta \in [0, 1/2]$. Indeed, by increasing β, we enlarge the exponent of the first and last argument of the maximum, while decreasing the exponent of the middle one. Thus, a reasonable choice is $\beta = 3/7$, making all the exponents equal.

7. Related works

Importance weighting (IW) is a widely used statistical tool with a long-standing role in Monte Carlo simulation, where it serves as a powerful method for variance reduction in rare events scenarios and for conducting what-if analyses [31,63,12,26,60]. In the *machine learning* (ML) field, this method has been predominantly employed for off-policy evaluation and learning tasks (e.g., [14,44,69]), besides few exceptions (e.g., [11,24,49,55]).

7.1. Importance weighting and heavy tails

In the ML context, it is well-established that IW can have inconvenient behavior depending on the dissimilarity between the behavioral Q and the target P distributions [75,45,46]. In particular, the range of values taken by the IW estimators can scale up to $\operatorname{ess\,sup}_{y\sim Q} p(y)/q(y)$. For discrete distributions (if $P \ll Q$), this term is finite. However, when considering the continuous case, it can easily become unbounded [14]. Furthermore, in the continuous case, the standard IW estimator may have infinite variance and exhibit a heavy-tailed behavior [45,46]. In order to address this problem, some approaches have been proposed based on *robust* statistics, typically employed for mean estimation under heavy-tailed distributions [43]. Methods in this class include the *trimmed mean* [71,28], the *median of means* [51,30], and the *Catoni's* estimator [10]. For all of them, subgaussian guarantees were provided [43]. These techniques have also been successfully employed for regret minimization in finite [8,21] and continuous arm spaces [42]. In principle, these methods could be employed *as-is* in conjunction with IW, but, being general-purpose, they might overlook the peculiarities of the setting.

7.2. Importance weighting transformations

Different approaches have been developed to address this problematic IW behavior. An example is represented by the *self-normalization* (SN, [53]): $\widetilde{\omega}(y_i) = \omega(y_i) / \sum_{j \in [\![n]\!]} \omega(y_j)$ technique. The advantage of this approach lies in bounding the range of the produced estimator. However, this comes at the cost of generating interdependence among samples, thus introducing bias in the estimator. Although consistency is guaranteed [25,67], its finite-sample analysis is more challenging. In [45], a polynomial concentration inequality was provided and, more recently, exponential bounds based on Efron-Stein inequalities have been proposed [34,35]. 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 [35]. Another widely used strategy is the weight *truncation* (or *clipping*) (TR, [29,5]): $\omega_M^{\mathrm{TR}}(y) = \min\{\omega(y), M\}$, with $M > 0$ being the clipping threshold. The selection of the proper threshold has either been done using some empirical approaches [13,37], or by using theoretically principled techniques [4,74,54,40,19]. In particular, [54] derive a subgaussian deviation bound by suitably adapting the truncation threshold as a function of the number of samples n and the confidence parameter δ. However, because of the truncation itself, which involves the computation of a minimum, the resulting estimator is not differentiable. Another non-differentiable clipping transformation is the one proposed in [41], where the weight is transformed as $\omega_\tau^{\mathrm{TR2}}(y) = p(y) / \max\{\tau, q(y)\}$ in order to avoid the weight exploding.

Another interesting approach, designed for CMABs, is the *switch* estimator (DR-SW, [74]) that selects between DM and IW (or DR), based on the importance weight value, by also providing guarantees in MSE. Finally, a restricted body of works explores less crude transformations than truncation, called *smoothing* [73]. They typically take into account the tail behavior of the estimator [56], also providing asymptotic guarantees. Some transformations based on weight shrinkage were proposed. They are based on the minimization of different bounds on the MSE, in the CMABs [65] setting. In particular, the *optimistic shrinkage* (OS, [65]) leads to a transformation similar to ours $\omega_\tau^{\mathrm{OS}}(y) = \tau\omega(y)/(\omega(y)^2 + \tau)$. In Appendix E, by providing a more careful analysis than that of [48], we show that, when P and Q are known and τ is set adaptively, OS achieves subgaussian concentration but remains biased when $P = Q$ almost surely. More recently, *exponential smoothing* (ES) was proposed in [1] based on transformations of the form $\widehat{\omega}_\alpha^{\mathrm{ES}}(y) = p(y)/q(y)^\alpha$ or $\widetilde{\omega}_\beta^{\mathrm{ES}}(y) = (p(y)/q(y))^\beta$ for $\alpha, \beta > 0$. Bias, variance, and PAC Bayes generalization bounds are proposed, but no optimal values of α and β are computed, restricting to an empirical approach to set them.

An approach similar to our PM estimator is that of [52,20,2], where the *implicit exploration* (IX) estimator is employed: $\omega_\gamma^{\mathrm{IX}}(y) = p(y)/(q(y) + \gamma)$ for some $\gamma > 0$. Although these works do not provide an explicit bias-variance analysis, this estimator is not unbiased when $P = Q$ almost surely and is affected by a bias that depends on the volume of the decision space (Appendix F for details). Furthermore, an analysis of a general class of estimators is proposed in [62], suggesting, then, a *logarithmically smoothed* (LS) estimator of the form $\omega_\lambda^{\mathrm{LS}}(y) f(y) = -\log(1 - \lambda\omega(y) f(y))$. This estimator considers function f inside the transformation and shares similar properties with our PM one, when λ is selected optimally, but remains biased when $P = Q$ almost surely and works only when $f(y) \leq 0$. A similar idea of applying *log-sum-exp* to the product $\omega(y) f(y)$ is proposed in [3].

For a summary comparison of the most important estimators, we refer to Table 2.

8. Numerical simulations

In this section, we present numerical simulations for off-policy evaluation (Section 8.1) and learning (Section 8.2). Our objective is to demonstrate that the proposed estimators are competitive with traditional (e.g., vanilla IW and self-normalization) and modern baselines (e.g., truncation, optimistic shrinkage), while enjoying desirable theoretical properties. In the following experiments, we restricted our analysis to the harmonic correction obtained by setting $s = -1$ (see Table 1) and $\alpha = 2$, apart from the experiment in Table 4. The complete results of the experiments are reported in Appendix B.[10]

8.1. Off-policy evaluation

We present two off-policy evaluation experiments. The first one is a synthetic example involving Gaussian distributions, while the second set of experiments considers the CMAB setting.

8.1.1. Synthetic experiment

In this numerical simulation, we compare our corrected estimators with importance weighting 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.

Estimators comparison. Table 3 reports 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 the optimal value of λ according to Theorem 5.1 with $\alpha = 2$ and $s = -1$ (PM-λ^*) and the value estimated from samples as in Section 6 with $\beta = 1/2$ (PM-$\widehat{\lambda}$). We compare these estimators with vanilla importance weighting (IW), self-normalized importance weighting (SN), weight truncation (TR) with optimal threshold

[10] The code is provided at https://github.com/albertometelli/subgaussian-is.

Table 2
Comparison between several IW estimators assuming $\|f\|_\infty = 1$, $s = -1$, and for $\alpha = 2$ w.r.t. several indexes. For the SN 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 [35]. $\text{vol}(\mathcal{Y})$ represents the volume of the decision space.

Estimator	Maximum	Variance	Bias	Correction (order $\mathcal{O}$)	Concentration (order $\mathcal{O}$)	Is subgaussian?	Is unbiased when $P = Q$?	Is differentiable?
IW [53,45]	$\operatorname{ess\,sup} \frac{p}{q}$	$\frac{I_2(P\|Q)}{n}$	0	-	$\sqrt{\frac{I_2(P\|Q)}{\delta n}}$	✗ (poly)	✓	✓
SN [53,35]	1	V^{SN}	B^{SN}	-	$B^{\text{SN}} + \sqrt{V^{\text{ES}} \log \frac{1}{\delta}}$	✗ (exp)	✓	✓
TR(M) [29,54]	M	$\frac{I_2(P\|Q)}{n}$	$\frac{I_2(P\|Q)}{M}$	$\sqrt{\frac{nI_2(P\|Q)}{\log \frac{1}{\delta}}}$	$\sqrt{\frac{I_2(P\|Q) \log \frac{1}{\delta}}{n}}$	✓	✗	✗
OS(τ) [65]	$\frac{\sqrt{\tau}}{2}$	$\frac{I_2(P\|Q)}{n}$	$\frac{I_2(P\|Q)}{\sqrt{\tau}}$	$\frac{nI_2(P\|Q)}{\log \frac{1}{\delta}}$	$\sqrt{\frac{I_2(P\|Q) \log \frac{1}{\delta}}{n}}$	✓	✗	✓
IX(γ) [52]	$\frac{\operatorname{ess\,sup}_{y \sim P} p(y)}{\gamma}$	$\frac{I_2(P\|Q)}{n}$	$\sqrt{\gamma I_2(P\|Q) \text{vol}(\mathcal{Y})}$	$\left(\frac{\left(\operatorname{ess\,sup}_{y \sim P} p(y)\right)^2 \left(\log \frac{1}{\delta}\right)^2}{I_2(P\|Q)\text{vol}(\mathcal{Y}) n^2} \right)^{\frac{1}{3}}$	$\max_{\beta \in \{2,3\}} \sqrt[\beta]{\frac{I_2(P\|Q)(\operatorname{ess\,sup}_{y \sim P} p(y) \text{vol}(\mathcal{Y}))^{\beta-2} \log \frac{1}{\delta}}{n}}$	✗ (exp)	✗	✓
LS(λ)[(a)] [62]	$+\infty$	$\frac{I_2(P\|Q)}{n}$	$\lambda I_2(P\|Q)$	$\sqrt{\frac{\log \frac{1}{\delta}}{I_2(P\|Q) n}}$	$\sqrt{\frac{I_2(P\|Q) \log \frac{1}{\delta}}{n}}$	✓	✗	✓
PM(λ)[(b)] (**ours**)	$\frac{1}{\lambda}$	$\frac{I_2(P\|Q)}{n}$	$\lambda I_2(P\|Q)$	$\sqrt{\frac{\log \frac{1}{\delta}}{I_2(P\|Q) n}}$	$\sqrt{\frac{I_2(P\|Q) \log \frac{1}{\delta}}{n}}$	✓	✓	✓

[(a)] only when $f(y) \leq 0$. [(b)] for sufficiently large n.

Table 3
Absolute error in the illustrative example varying the number of samples n for the different estimators (mean ± std, 60 runs). For each column, the estimator with the 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	10	20	50	100	200	500	1000
IW	**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	**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
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
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
PM-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
PM-$\widehat{\lambda}$	**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 4
Absolute error in the illustrative example varying the parameter s of the corrected weight when $n = 500$ (mean ± std, 60 runs). The estimator with the smallest absolute error and the ones not statistically significantly different from that one (Welch's t-test with $p < 0.02$) are in bold.

s / λ	0	0.1	0.2	0.5
$-\infty$	3.12 ± 0.29	3.12 ± 0.29	3.12 ± 0.29	3.12 ± 0.29
−5	6.73 ± 1.21	**2.70 ± 0.30**	2.77 ± 0.30	2.57 ± 0.31
−2	6.73 ± 1.21	**2.45 ± 0.34**	**2.42 ± 0.32**	**2.28 ± 0.32**
−1	6.73 ± 1.21	**2.72 ± 0.47**	**2.47 ± 0.37**	**2.18 ± 0.32**
−0.5	6.73 ± 1.21	3.44 ± 0.64	**2.71 ± 0.47**	**2.20 ± 0.34**
0	6.73 ± 1.21	4.83 ± 0.89	3.66 ± 0.68	**2.38 ± 0.38**
0.5	6.73 ± 1.21	5.69 ± 1.05	4.85 ± 0.89	3.03 ± 0.52

selected as in [54], and IW with optimistic shrinkage (OS), where τ is computed by minimizing an MSE bound as in [65]. We notice that our estimators consistently outperform the traditional ones (IW and SN) and overall suffer smaller errors than TR and OS. Interestingly, the minimum error is often obtained by PM-$\widehat{\lambda}$, which uses an estimated value $\widehat{\lambda}$ that tends to get a higher value than λ^*. In this way, the correction is more intense, which, in this specific example, turns out to be beneficial.

Comparison of different values of s. We empirically test different values of the parameter s employed in Definition 5.1, in the same setting of Table 3 with $n = 500$ for the estimator PM-λ^* for different values of λ. The results are reported in Table 4. We can see that the best results are obtained with $s \in \{-1, -2\}$.

8.1.2. Contextual bandits

In this section, we report the experiments about off-policy evaluation in CMABs.

Setting. We follow the well-established setting of [18,74,64,65]. We consider 11 UCI [17] multi-class classification datasets (see Table 9 in 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 with a cross-fitting procedure on the full $\mathcal{D}$.

Estimators comparison. 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 [74], we plot in Fig. 2 the cumulative distribution function (CDF) of the absolute error normalized by the error of IW. 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 $\widehat{r}$ (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 PM-λ^* is very close to that of SN. This is likely because we are dealing with discrete distributions (actions are finite), which implicitly mitigates the degeneracy of the importance weight. Differently, the advantage w.r.t. the optimistic shrinkage (OS) is quite significant. Instead, for the model-based estimators, we observe 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 similar to ours. Overall, we observe that our estimators, although not radically outperforming the baselines, succeed in consistently displaying competitive performance.

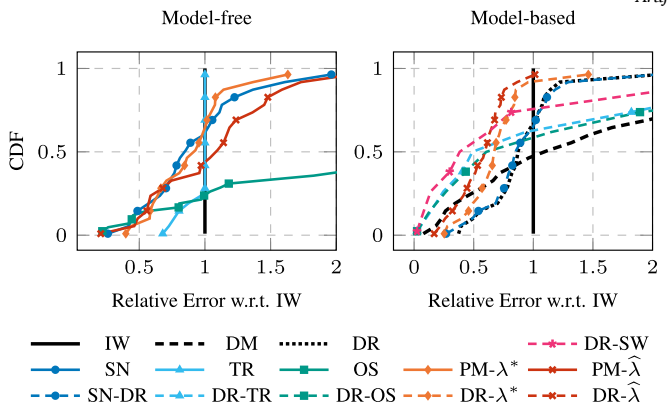

Fig. 2. CDF of the absolute error normalized by IW error for stochastic rewards with noise 0.25, across 110 conditions.

Table 5
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 the smallest absolute error and the ones not statistically significantly different from that one (Welch's t-test with $p < 0.05$) are in bold.

Estimator / n	100	200	500	1000	2000	5000	10000	20000
IW	**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	**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
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
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
PM-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
PM-$\widehat{\lambda}$	**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-$\widehat{\lambda}$	**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 6
Complementary cumulative distribution of the absolute error (multiplied by 100) $\mathbb{P}(E > \xi)$ in the *glass* dataset varying the number of samples n for the different estimators, when $\alpha_b = 0.9$ and $\alpha_e = 0.9999$ (5000 runs).

Estimator / ξ	10	20	50	100	200	500	1000
IW	0.6742	0.5414	0.1754	0.0326	0.0326	0.0198	0.0014
PM-λ^*	0.686	0.5416	0.176	0.0228	0.056	0	0
DR	0.6522	0.4094	0.117	0.0484	0.0378	0.0218	0.0022
DR-λ^*	0.65	0.4046	0.1088	0.03262	0.009	0	0

For the specific case of the *letter* dataset, we report in Table 5 the results obtained by setting $\alpha_b = 0.5$ and $\alpha_e = 0.9$ for different numbers 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 become more visible. In particular, the DR-SW and our corrected estimators (DR-λ^* and DR-$\widehat{\lambda}$) overall dominate the other baselines.

Tail behavior experiment. We run 5000 estimation processes using the *glass* dataset, $n = 30$, $\alpha_e = 0.9999$, and $\alpha_b = 0.9$. To compare the tail behavior between vanilla weights and our correction (for both model-free and model-based estimators), we consider the absolute error random variable E multiplied by 100 (as in Table 5) and we estimate the *complementary cumulative distribution* $\mathbb{P}(E > \xi)$. Thus, for large values of ξ, the larger $\mathbb{P}(E > \xi)$, the heavier the tail, since a larger amount of probability mass accumulates on the right of ξ. Table 6 reports the results for both model-based and model-free estimators. We observe that our corrected estimators consistently display a significantly lighter tail compared to the vanilla ones.

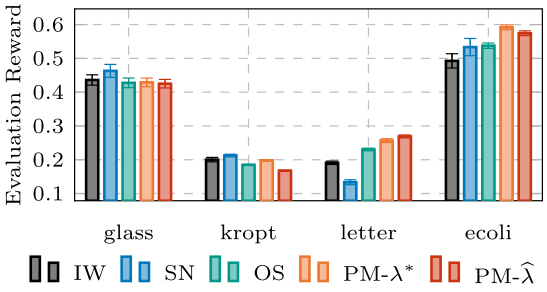

Fig. 3. Evaluation reward for four different datasets after 1000 iterations (4000 iterations for *letter*) of gradient ascent with a Boltzmann policy for the model-free estimators (mean ± std, 10 runs).

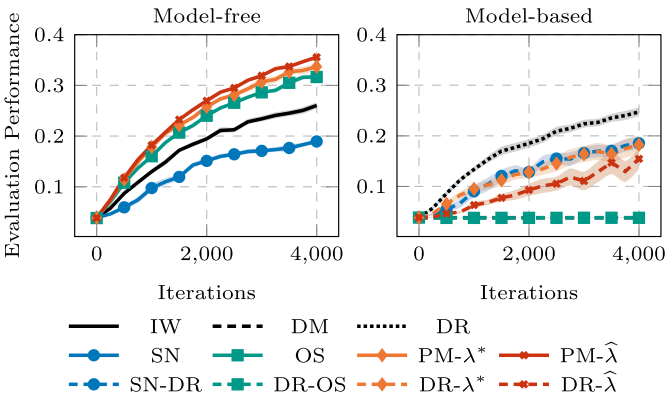

Fig. 4. Evaluation reward for the *letter* dataset comparing the learning curve of different estimators (mean ± std, 10 runs).

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 Fig. 3. We observe that our weight corrections (PM-λ^* and PM-$\widehat{\lambda}$) outperform the considered baselines (IW, SN, and OS) on *ecoli* and *letter* datasets, whereas SN emerges in the *glass* and *kropt* datasets. For the *letter* dataset, we report in Fig. 4 the learning curve, distinguishing between model-free (left) and model-based (right) estimators.[11] For the model-free ones, we observe the dominance of our estimators over the SN estimator, while the optimistic shrinkage estimator (OS) behaves 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 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 conceived a novel minimax lower bound for off-policy estimation, showing that exponential concentration is possible. Then, we derived an anti-concentration bound for the vanilla IW and SN estimators, 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, for a sufficiently large number of samples. Moreover, we introduced a data-driven approach that, under slightly more demanding assumptions, allows obtaining a similar concentration. The experimental evaluation has shown that our theoretically grounded transformation is competitive with the traditional and modern importance weighting 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.

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

CRediT authorship contribution statement

Alberto Maria Metelli: Writing – original draft, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Alessio Russo:** Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Marcello Restelli:** Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Funded by the European Union – Next Generation EU within the project NRPP M4C2, Investment 1.3 DD. 341 – 15 March 2022 – FAIR – Future Artificial Intelligence Research – Spoke 4 – PE00000013 – D53C22002380006.

Appendix A. Proofs and derivations

In this section, we report the proofs of the results that are reported in the main paper.

A.1. Proofs of Section 4

Theorem 4.2 (Anti-concentration of IW Estimator)**.** *There exist two distributions $P, Q \in \Delta^{\mathcal{Y}}$ 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 $n \geq \delta e \max\left\{1, \left(I_\alpha(P\|Q) - 1\right)^{\frac{1}{\alpha-1}}\right\}$, with probability at least δ it holds that:*

$$\left|\widehat{\mu}_n - \mu\right| \geq \|f\|_\infty \left(\frac{I_\alpha(P\|Q) - 1}{e\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}. \tag{10}$$

Proof. The proof is inspired by that of Proposition 6.2 of [10]. We construct a function f and two probability measures P and Q that fulfill the inequality. Let $a > 0$, we consider $\mathcal{Y} = \{-a, 0, a\}$ and $f(y) = y$. First of all, we observe that $a = \|f\|_\infty$. We now define the probability distributions as follows, for $p, q \in [0,1]$:

$$P(\{-a\}) = P(\{a\}) = \frac{p}{2} \text{ and } P(\{0\}) = 1 - p, \tag{P.65}$$

$$Q(\{-a\}) = Q(\{a\}) = \frac{q}{2} \text{ and } Q(\{0\}) = 1 - q. \tag{P.66}$$

We immediately observe that $\mathbb{E}_{y\sim P}[f(y)] = \mathbb{E}_{y\sim Q}[f(y)] = 0$. We select the values p and q as follows, for any $\alpha \in (1,2]$:

$$q = \left(\frac{a}{n\epsilon}\right)^\alpha \xi, \qquad p = \left(\frac{a}{n\epsilon}\right)^{\alpha-1} \xi, \tag{P.67}$$

where $\xi > 0$ will be specified later. First of all, we note that to make these probabilities valid, we need to enforce:

$$p \leq 1 \implies n \geq \frac{a}{\epsilon}\xi^{\frac{1}{\alpha}}, \qquad q \leq 1 \implies n \geq \frac{a}{\epsilon}\xi^{\frac{1}{\alpha-1}}. \tag{P.68}$$

This choice of p and q ensures that $a\frac{p}{q} = n\epsilon$. Let us now compute the divergence:

$$I_\alpha(P\|Q) = 2\left(\frac{p}{2}\right)^\alpha \left(\frac{q}{2}\right)^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha} \tag{P.69}$$

$$= p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha} \tag{P.70}$$

$$= \xi + \left(1 - \xi\left(\frac{a}{n\epsilon}\right)^{\alpha-1}\right)^\alpha \left(1 - \xi\left(\frac{a}{n\epsilon}\right)^\alpha\right)^{1-\alpha} \leq \xi + 1, \tag{P.71}$$

where the last inequality is obtained by upper bounding the second addendum under the assumption that $n \geq \frac{a}{\epsilon}\xi^{\frac{1}{\alpha-1}}$:

$$\left(1 - \xi\left(\frac{a}{n\epsilon}\right)^{\alpha-1}\right)^\alpha \left(1 - \xi\left(\frac{a}{n\epsilon}\right)^\alpha\right)^{1-\alpha} \leq \left(1 - \xi\left(\frac{a}{n\epsilon}\right)^{\alpha-1}\right)^\alpha \left(1 - \xi\left(\frac{a}{n\epsilon}\right)^{\alpha-1}\right)^{1-\alpha} \tag{P.72}$$

$$= 1 - \xi\left(\frac{a}{n\epsilon}\right)^{\alpha-1} \leq 1. \tag{P.73}$$

Thus, we select $\xi = I_\alpha(P\|Q) - 1$. Let us now consider the vanilla IW estimator $\widehat{\mu}_n$, whose expectation is $\mu = 0$, and the following derivation:

$$\mathop{\mathbb{P}}_{y_i\sim Q}\left(|\widehat{\mu}_n - \mu| \geq \epsilon\right) = \mathop{\mathbb{P}}_{y_i\sim Q}\left(\{\widehat{\mu}_n - \mu \leq -\epsilon\} \cup \{\widehat{\mu}_n - \mu \geq \epsilon\}\right) \tag{P.74}$$

$$= \mathop{\mathbb{P}}_{y_i\sim Q}\left(\widehat{\mu}_n - \mu \leq -\epsilon\right) + \mathop{\mathbb{P}}_{y_i\sim Q}\left(\widehat{\mu}_n - \mu \geq \epsilon\right) \tag{P.75}$$

$$= 2\mathop{\mathbb{P}}_{y_i\sim Q}\left(\widehat{\mu}_n - \mu \geq \epsilon\right), \tag{P.76}$$

where line (P.75) is obtained by observing that the two events are disjoint and line (P.76) comes from the symmetry of the events, given the construction of function f. We now consider the event $\mathcal{E}$ under which among the n samples, one is a and the remaining are 0:

$$\mathcal{E} := \left\{\left|\left\{i \in [\![n]\!] : y_i = 0\right\}\right| = n - 1 \wedge \left|\left\{i \in [\![n]\!] : y_i = a\right\}\right| = 1\right\}. \tag{P.77}$$

It is immediate to verify that if event $\mathcal{E}$ occurs we have that $\widehat{\mu}_n = \frac{pa}{qn} = \epsilon$, and, consequently, $\widehat{\mu}_n - \mu \geq \epsilon$. Thus, we now lower bound the probability:

$$\mathop{\mathbb{P}}_{y_i\sim Q}\left(\widehat{\mu}_n - \mu \geq \epsilon\right) \geq \mathop{\mathbb{P}}_{y_i\sim Q}(\mathcal{E}) \tag{P.78}$$

$$= n\frac{q}{2}(1-q)^{n-1} \tag{P.79}$$

$$= \frac{1}{2}\left(\frac{a}{\epsilon}\right)^\alpha n^{1-\alpha}\xi\left(1 - \left(\frac{a}{n\epsilon}\right)^\alpha \xi\right)^{n-1}. \tag{P.80}$$

Now, we derive a value of $\epsilon > 0$ such that the inequality holds with probability at least δ. We enforce the condition:

$$\frac{1}{2}\left(\frac{a}{\epsilon}\right)^\alpha n^{1-\alpha}\xi\left(1 - \left(\frac{a}{n\epsilon}\right)^\alpha \xi\right)^{n-1} \leq \delta \quad \Longrightarrow \quad \epsilon \geq a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1 - \left(\frac{a}{n\epsilon}\right)^\alpha \xi\right)^{\frac{n-1}{\alpha}}. \tag{P.81}$$

We claim that, for $\delta \in (0, e^{-1})$, any value of ϵ fulfilling condition (P.81) must be $\epsilon \leq \epsilon^\star$:

$$\epsilon^\star = a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1 - \frac{e\delta}{n}\right)^{\frac{n-1}{\alpha}}. \tag{P.82}$$

Indeed, we have:

$$a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1 - \left(\frac{a}{n\epsilon^\star}\right)^\alpha \xi\right)^{\frac{n-1}{\alpha}} \tag{P.83}$$

$$= a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1 - \left(\frac{a}{n}\right)^\alpha\left(a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1 - \frac{e\delta}{n}\right)^{\frac{n-1}{\alpha}}\right)^{-\alpha}\xi\right)^{\frac{n-1}{\alpha}} \tag{P.84}$$

$$= a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1 - \frac{\delta}{n}\left(1 - \frac{e\delta}{n}\right)^{-(n-1)}\right)^{\frac{n-1}{\alpha}} \tag{P.85}$$

$$\geq a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1 - \frac{\delta e}{n}\right)^{\frac{n-1}{\alpha}} = \epsilon^\star, \tag{P.86}$$

where the last inequality derives from observing that $\left(1 - \frac{e\delta}{n}\right)^{-(n-1)} \leq e$ if $\delta \in (0, e^{-1})$. Finally, we rephrase conditions (P.68):

$$n \geq \frac{a}{\epsilon^\star}\xi^{\frac{1}{\alpha}} \quad \Longrightarrow \quad n \geq n^{1-\frac{1}{\alpha}}\delta^{\frac{1}{\alpha}}\left(1 - \frac{e\delta}{n}\right)^{-\frac{n-1}{\alpha}} \quad \Longrightarrow \quad n \geq \delta e, \tag{P.87}$$

$$n \geq \frac{a}{\epsilon^\star}\xi^{\frac{1}{\alpha-1}} \quad \Longrightarrow \quad n \geq n^{1-\frac{1}{\alpha}}\delta^{\frac{1}{\alpha}}\xi^{\frac{1}{\alpha(\alpha-1)}}\left(1 - \frac{e\delta}{n}\right)^{-\frac{n-1}{\alpha}} \quad \Longrightarrow \quad n \geq \delta e\xi^{\frac{1}{\alpha-1}}, \tag{P.88}$$

having observed, again, that $\left(1 - \frac{e\delta}{n}\right)^{-\frac{n-1}{\alpha}} \leq e^{\frac{1}{\alpha}}$. Thus, we should enforce the condition $n \geq \delta e \max\left\{1, \xi^{\frac{1}{\alpha-1}}\right\}$. To make the statement more readable, we bound $\left(1 - \frac{e\delta}{n}\right)^{\frac{n-1}{\alpha}} \geq e^{-\frac{1}{\alpha}}$. □

Theorem 4.4 (Anti-concentration of SN Estimator)**.** *There exist two distributions $P, Q \in \Delta^{\mathcal{Y}}$ 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 $n \geq \max\left\{ \frac{\delta e}{I_\alpha(P\|Q)-1}, \left(\frac{I_\alpha(P\|Q)-1}{\delta}\right)^{\frac{1}{\alpha-1}} \right\}$, with probability at least δ it holds that:*

$$|\widetilde{\mu}_n - \mu| \geq \frac{\|f\|_\infty}{2} \left(\frac{I_\alpha(P\|Q) - 1}{e\delta n^{\alpha-1}} \right)^{\frac{1}{\alpha}}. \tag{12}$$

Proof. The proof follows the same steps of that of Theorem 4.2, with the same construction of function f and probability measures P and Q. We proceed under the conditions $p > q$ and $\epsilon \leq a$ that we enforce later in the proof. Thus, under event $\mathcal{E}$, then $\widetilde{\mu}_n - \mu \geq \frac{\epsilon}{2}$, as proven in the following:

$$\begin{aligned}
\widetilde{\mu}_n - \mu &= \widetilde{\mu}_n && \text{(P.89)}\\
&= \frac{\frac{p}{q}a}{(n-1)\frac{1-p}{1-q} + \frac{p}{q}} \geq \frac{\frac{p}{q}a}{(n-1) + \frac{p}{q}} && \text{(P.90)}\\
&= \frac{n\epsilon}{(n-1) + n\frac{\epsilon}{a}} && \text{(P.91)}\\
&\geq \frac{\epsilon}{1 + \frac{\epsilon}{a}} \geq \frac{\epsilon}{2}, && \text{(P.92)}
\end{aligned}$$

where line (P.89) is obtained by observing that $\mu = 0$, line (P.90) follows from the definition of the SN estimator, line (P.91) derives from $p \geq q$, consequently, $\frac{1-p}{1-q} \leq 1$ and from $\frac{p}{q} = \frac{n\epsilon}{a}$, and line (P.92) is obtained from simple manipulation and enforcing $\frac{\epsilon}{a} \leq 1$. By exploiting the ϵ expression of Theorem 4.2, we have:

$$|\widetilde{\mu}_n - \mu| \geq \frac{a}{2} \left(\frac{\xi}{\delta n^{\alpha-1}} \right)^{\frac{1}{\alpha}} \left(1 - \frac{\delta e}{n}\right)^{\frac{n-1}{\alpha}}. \tag{P.93}$$

In addition to the conditions enforced by Theorem 4.2, we need to require $p \geq q$ and $\epsilon \leq a$. To this end, we recall the conditions $1 - \frac{\delta e}{n} \leq 1$ and $\left(1 - \frac{e\delta}{n}\right)^{-(n-1)} \leq e$ if $\delta \in (0, e^{-1})$:

$$\epsilon \leq a \implies \left(\frac{\xi}{\delta n^{\alpha-1}} \right)^{\frac{1}{\alpha}} \left(1 - \frac{\delta e}{n}\right)^{\frac{n-1}{\alpha}} \leq 1 \quad \text{satisfied for} \quad n \geq \left(\frac{\xi}{\delta}\right)^{\frac{1}{\alpha-1}}, \tag{P.94}$$

$$p > q \implies n > \frac{a}{\epsilon} \implies n \left(\frac{\xi}{\delta n^{\alpha-1}} \right)^{\frac{1}{\alpha}} \left(1 - \frac{\delta e}{n}\right)^{\frac{n-1}{\alpha}} \geq 1 \quad \text{satisfied for} \quad n \geq \frac{e\delta}{\xi}. \tag{P.95}$$

Thus, collecting also the conditions of Theorem 4.2, we enforce $n \geq \delta e \max\left\{ 1, \frac{1}{\xi}, \xi^{\frac{1}{\alpha-1}}, \xi^{\frac{1}{\alpha-1}} \left(\frac{1}{\delta}\right)^{\frac{\alpha}{\alpha-1}} \frac{1}{e} \right\}$, that can be simplified, although by loosing some tightness, into $n \geq \max\left\{ \frac{\delta e}{\xi}, \left(\frac{\xi}{\delta}\right)^{\frac{1}{\alpha-1}} \right\}$. □

A.2. Proofs of Section 5

Lemma 5.1. *Let $P, Q \in \Delta^{\mathcal{Y}}$ 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 $s \leq s'$ then $\omega_{\lambda,s}(y) \leq \omega_{\lambda,s'}(y)$;*

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

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

Proof. Recall that $\omega_{s,\lambda}(y)$ is the power mean of exponent s between $\omega(y)$ and 1 and weights $(1-\lambda, \lambda)$. Consequently, (i) follows from the generalized mean inequality [9]. Let us move to (ii), if $s < 0$, we have:

$$\omega_{\lambda,s}(y) = \left((1-\lambda)\omega(y)^s + \lambda\right)^{\frac{1}{s}} = \frac{1}{\left(\frac{1-\lambda}{\omega(y)^{-s}} + \lambda\right)^{\frac{1}{-s}}} \leq \lambda^{\frac{1}{s}}. \tag{P.96}$$

Instead for $s > 0$, we have:

$$\omega_{\lambda,s}(y) = \left((1-\lambda)\omega(y)^s + \lambda\right)^{\frac{1}{s}} \geq \lambda^{\frac{1}{s}}. \tag{P.97}$$

Concerning (iii), let us first observe that for every $\lambda \in [0,1]$ and $s = 1$, it holds that $\mathbb{E}_{y\sim Q}[\omega_{\lambda,1}(y)] = 1$. Following from (i) and from the monotonicity of the expectation, we have that for $s < 1$:

$$\omega_{\lambda,s}(y) \leq \omega_{\lambda,1}(y) \quad \Longrightarrow \quad \mathop{\mathbb{E}}_{y\sim Q}[\omega_{\lambda,s}(y)] \leq \mathop{\mathbb{E}}_{y\sim Q}[\omega_{\lambda,1}(y)] = 1. \tag{P.98}$$

Symmetrically, for $s > 1$ we have:

$$\omega_{\lambda,s}(y) \geq \omega_{\lambda,1}(y) \quad \Longrightarrow \quad \mathop{\mathbb{E}}_{y\sim Q}[\omega_{\lambda,s}(y)] \geq \mathop{\mathbb{E}}_{y\sim Q}[\omega_{\lambda,1}(y)] = 1. \quad \square \tag{P.99}$$

Theorem 5.1. *Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability distributions such that $P \ll Q$. Let $\{y_i\}_{i\in[\![n]\!]}$ be sampled independently from Q. For every $\alpha \in (1,2]$ and $\delta \in (0,1)$ if:*

$$n \geq \frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 I_\alpha(P\|Q)}, \tag{18}$$

then, with probability at least $1-\delta$, it holds that:

$$\widehat{\mu}_{n,\lambda^*_\alpha,s^*_\alpha} - \mu \leq \frac{7\|f\|_\infty}{3}\left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2}\right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q)^{\frac{1}{\alpha}}, \tag{19}$$

having selected $(\lambda^*_\alpha, s^*_\alpha)$ such that:

$$\lambda^*_\alpha = \left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2 I_\alpha(P\|Q)}\right)^{-\frac{s^*_\alpha}{\alpha}}. \tag{20}$$

Proof. We start from the expression of the bound in Lemma 5.4 and highlight the dependence on λ and s:

$$\underbrace{\|f\|_\infty\sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n} I_\alpha(P\|Q)}}_{=:a}\lambda^{-\frac{\alpha-2}{2s}} + \underbrace{\frac{2\|f\|_\infty\log\left(\frac{1}{\delta}\right)}{3n}}_{=:b}\lambda^{\frac{1}{s}} + \underbrace{\|f\|_\infty(3-\alpha)^{\frac{1}{\alpha}} I_\alpha(P\|Q)}_{=:c}\lambda^{\frac{1-\alpha}{s}}. \tag{P.100}$$

For notational convenience, we rename $\eta := \lambda^{\frac{1}{s}}$. Thus, we need to minimize over $\eta \in [0,1]$ the function:

$$h(\eta) := a\eta^{-\frac{\alpha-2}{2}} + b\eta + c\eta^{1-\alpha}. \tag{P.101}$$

Since h is a continuously differentiable function in $(0,1]$, we vanish the derivative:

$$\frac{\partial h}{\partial \eta}(\eta) = \frac{1}{2}(2-\alpha)a\eta^{-\frac{\alpha}{2}} + b + (1-\alpha)c\eta^{-\alpha} = 0. \tag{P.102}$$

By setting $\xi^2 = \eta^{-\alpha}$, we obtain the second degree equation $\frac{1}{2}(2-\alpha)a\xi + b + (1-\alpha)c\xi^2 = 0$, leading to the only non-negative solution:

$$\lambda^{-\frac{\alpha}{2s}} = \frac{(2-\alpha)a + 2\sqrt{4bc(\alpha-1) + a^2(1-\alpha/2)^2}}{4c(\alpha-1)}. \tag{P.103}$$

By substituting the values of a, b, and c, we obtain:

$$\lambda^{-\frac{\alpha}{2s}} = \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2 I_\alpha(P\|Q)}} \cdot \underbrace{\frac{\left((2-\alpha) + 2\sqrt{\frac{8}{3}(3-\alpha)^{\frac{1}{\alpha}}(\alpha-1) + 2(1-\alpha/2)^2}\right)}{4(3-\alpha)^{\frac{1}{\alpha}}}}_{\leq\sqrt{2}}. \tag{P.104}$$

For analytical convenience and at the price of paying just a constant term, we make use of the following approximation of the optimal λ:

$$\lambda_\alpha^* = \left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2 I_\alpha(P\|Q)} \right)^{-\frac{s_\alpha^*}{\alpha}}. \tag{P.105}$$

By substituting this term into the bound, we obtain:

$$\widehat{\mu}_{n,\lambda_\alpha^*,s_\alpha^*} - \mu \leq \|f\|_\infty \left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2} \right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q)^{\frac{1}{\alpha}} \underbrace{\left((\alpha-1) + \frac{(\alpha-1)^2}{3} + (3-\alpha)^{\frac{1}{\alpha}} \right)}_{\leq 7/3} \tag{P.106}$$

$$\leq \frac{7\|f\|_\infty}{3} \left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2} \right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q)^{\frac{1}{\alpha}}, \tag{P.107}$$

where the last inequality is obtained by bounding the last factor over α. The condition on n is enforced so that $\lambda_\alpha^* \leq 1$. □

Proposition A.1. *Let $\lambda \in [0,1]$. For every $y \in \mathcal{Y}$, let $\omega(y) = \frac{p_\theta(y)}{q(y)}$, for a target distribution p_θ differentiable in θ. Then, it holds that:*

$$\left\|\nabla_\theta \omega_\lambda(y)\right\|_\infty \leq \frac{-s}{\lambda^{-\frac{1}{s}}(1-s)^{1+\frac{1}{s}}} \left\|\nabla_\theta \log p_\theta(y)\right\|_\infty. \tag{38}$$

Proof. Let us first compute the gradient explicitly, having set $z = -s > 0$:

$$\nabla_\theta \omega_{\lambda,s}(y) = \frac{\partial \omega_{\lambda,s}}{\partial \omega}(y) \nabla_\theta \omega(y) = \frac{(1-\lambda)\omega(y)^s}{((1-\lambda)\omega(y)^s + \lambda)^{1-\frac{1}{s}}} \nabla_\theta \log p_\theta(y) = \frac{(1-\lambda)\omega(y)}{(1-\lambda+\lambda\omega(y)^z)^{1+\frac{1}{z}}} \nabla_\theta \log p_\theta(y). \tag{P.108}$$

To get the result, we maximize the value of the following function:

$$g(v) = \frac{(1-\lambda)v}{(1-\lambda+\lambda v^z)^{1+\frac{1}{z}}}. \tag{P.109}$$

First of all, we observe that for $v = 0$ and $v \to +\infty$, the function has value 0. Thus, the maximum must lie in between. We vanish the derivative to find it:

$$\frac{\partial g(v)}{\partial v} = \frac{(1-\lambda)(1-\lambda-\lambda z v^z)}{(1-\lambda+\lambda v^z)^{2+\frac{1}{z}}} = 0 \quad \Longrightarrow \quad v^* = \left(\frac{1-\lambda}{\lambda z}\right)^{\frac{1}{z}}. \tag{P.110}$$

By substituting the found value, we obtain:

$$g(v^*) = \frac{z}{\lambda^{\frac{1}{z}}(1+z)^{1-\frac{1}{z}}} = \frac{-s}{\lambda^{-\frac{1}{s}}(1-s)^{1+\frac{1}{s}}}. \tag{P.111}$$

The result is obtained by applying the L_∞-norm. □

A.3. Proofs of Section 6

For the sake of simplicity, we will denote with $\eta := \lambda n^\beta$ for some $\beta \in \left[0, \frac{1}{\alpha}\right]$. Since $\lambda \in [0, n^{-\beta}]$, we have that $\eta \in [0,1]$. We introduce the following equation, representing an equivalent version of the expectation of Equation (28):

$$h_{\alpha,s}(\eta) = \eta^{-\frac{\alpha}{s}} \mathop{\mathbb{E}}_{y\sim Q}\left[\omega_{\eta,s}(y)^\alpha\right] = \frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 n^{1+\frac{\alpha\beta}{s}}}, \tag{39}$$

We introduce the corresponding empirical version, which is equivalent to Equation (28):

$$\widehat{h}_{\alpha,s}(\eta) = \eta^{-\frac{\alpha}{s}} \frac{1}{n} \sum_{i\in[\![n]\!]} \omega_{\eta,s}(y_i)^\alpha = \frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 n^{1+\frac{\alpha\beta}{s}}}, \tag{40}$$

Clearly, we have $\mathbb{E}_{y_i\sim Q}\left[\widehat{h}_{\alpha,s}(\eta)\right] = h_{\alpha,s}(\eta)$. The following result characterizes the boundedness of functions $\widehat{h}_{\alpha,s}(\eta)$ and $h_{\alpha,s}(\eta)$ and the existence of a unique solution of Equations (40) and (39).

Lemma A.1. *Let $\alpha \in (1,2]$ and $s \in [-\infty,-1]$. For every $\eta \in [0,1]$, it holds that $\widehat{h}_{\alpha,s}(\eta), h_{\alpha,s}(\eta) \in [0,1]$. Furthermore, let $Q(p>0) := Q(\{y \in \mathcal{Y} : p(y)>0\})$:*

(i) if

$$n \geq \left(\frac{2\log\left(\frac{1}{\delta}\right)}{Q(p>0)(\alpha-1)^2} \right)^{\frac{s}{s+\alpha\beta}},$$

then, Equation (39) admits exactly one solution;

(ii) if

$$\sum_{i \in [\![n]\!]} \mathbb{1}\{\omega(y_i)>0\} \geq \frac{2n^{-\frac{\alpha\beta}{s}}\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2},$$

then, Equation (40) admits exactly one solution.

Proof. For the boundedness, we immediately observe that $\widehat{h}_{\alpha,s}(\eta), h_{\alpha,s}(\eta) \geq 0$. Moreover, we have $\omega_{\eta,s}(y) \leq \eta^{-\frac{1}{s}}$, from which the result follows. We compute the derivatives of $\widehat{h}_{\alpha,s}(\eta)$ and $h_{\alpha,s}(\eta)$ in η. Let us start with $\widehat{h}_{\alpha,s}(\eta)$:

$$\frac{\partial}{\partial\lambda}\widehat{h}_{\alpha,s}(\eta) = -\frac{\alpha\eta^{-\frac{\alpha}{s}-1}}{sn}\sum_{i \in [\![n]\!]} \omega(y_i)^s((1-\eta)\omega(y_i)^s+\eta)^{-1+\frac{\alpha}{s}} > 0, \tag{P.112}$$

under the assumption that $\omega(y_i)>0$ for some $i \in [\![n]\!]$, concluding that $\widehat{h}_{\alpha,s}$ is increasing in $\eta \in (0,1)$. Let us move to $h_{\alpha,s}(\eta)$:

$$\frac{\partial}{\partial\eta}h_{\alpha,s}(\eta) = -\frac{\alpha\eta^{-\frac{\alpha}{s}-1}}{s}\mathop{\mathbb{E}}_{y\sim Q}\left[\omega(y)^s((1-\eta)\omega(y)^s+\eta)^{-1+\frac{\alpha}{s}}\right] > 0, \tag{P.113}$$

concluding that $h_{\alpha,s}$ is increasing in $\eta \in (0,1)$. By setting $\eta=0$, we have that $\widehat{h}_{\alpha,s}(0) = h_{\alpha,s}(0) = 0$. By setting $\eta=1$, we have that:

$$\widehat{h}_{\alpha,s}(1) = \sum_{i \in [\![n]\!]} \mathbb{1}\{\omega(y_i)>0\}, \qquad h_{\alpha,s}(1) = \mathop{\mathbb{E}}_{y\sim Q}[\mathbb{1}\{\omega(y)>0\}], \tag{P.114}$$

discarding the terms for which $\omega(y)=0$. Thus, we are guaranteed that the solution is unique if the right-hand side of Equations (40) and (39) lie in the interval $[\widehat{h}_{\alpha,s}(0), \widehat{h}_{\alpha,s}(1)]$ and $[h_{\alpha,s}(0), h_{\alpha,s}(1)]$, respectively. To conclude, consider now:

$$\mathop{\mathbb{E}}_{y\sim Q}[\mathbb{1}\{\omega(y)>0\}] = \mathop{\mathbb{E}}_{y\sim Q}[\mathbb{1}\{p(y)>0\}] = Q(\{y \in \mathcal{Y} : p(y)>0\}) = Q(p>0). \quad \square \tag{P.115}$$

In order to obtain a minimum number of samples for which, with high probability, the empirical equation admits a unique solution, we proceed as in the following lemma.

Lemma A.2. *Let $\alpha \in (1,2]$, $s \in [-\infty,-1]$, and $\delta \in (0,1)$. Let $Q(p>0) := Q(\{y \in \mathcal{Y} : p(y)>0\})$, if*

$$n \geq \left(\frac{6\log\left(\frac{1}{\delta}\right)}{Q(p>0)(\alpha-1)^2} \right)^{\frac{1}{1+\frac{\alpha\beta}{s}}},$$

then, with probability at least $1-\delta$, Equation (40) admits exactly one solution.

Proof. We use Lemma F.4 of [15] to bound the number of times $\omega(y_i)>0$ for $i \in [\![n]\!]$, since the random variables $\mathbb{1}\{\omega(y_i)>0\}$ are Bernoulli random variables with parameter $Q(p>0)$. With probability at least $1-\delta$, it holds that:

$$\sum_{i \in [\![n]\!]} \mathbb{1}\{\omega(y_i)>0\} \geq \frac{n}{2}Q(p>0) - \log\left(\frac{1}{\delta}\right). \tag{P.116}$$

Enforcing the following condition leads to the smallest number of samples requested:

$$\frac{n}{2}Q(p>0) - \log\left(\frac{1}{\delta}\right) \geq \frac{2n^{-\frac{\alpha\beta}{s}}\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}. \tag{P.117}$$

For analytical convenience, we enforce the more restrictive condition:

$$\frac{n}{2}Q(p>0) \geq \frac{3n^{-\frac{\alpha\beta}{s}}\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}. \quad \square \tag{P.118}$$

Lemma 6.1. *Let $\alpha \in (1,2]$, $s \in [-\infty,-1]$, and $\delta \in (0,1)$. If*

$$n \geq \left(\frac{6I_\alpha(P\|Q)^{\frac{1}{\alpha-1}}\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}\right)^{\frac{1}{1+\frac{\alpha\beta}{s}}}, \tag{29}$$

then, with probability at least $1-\delta$, Equation (28) admits exactly one solution.

Proof. We start from Lemma A.2 and proceed to lower bound $Q(p>0)$. We now argue that $Q(p>0)>0$ when there exists $\alpha \in (1,2]$ such that $I_\alpha(P\|Q) < +\infty$. To this end, we relate $Q(p>0)$ with the total variation divergence between P and Q. Let us define the measurable set $\mathcal{Y}_P := \{y \in \mathcal{Y} \,:\, p(y)>0\}$, we have:

$$\mathrm{TV}(P,Q) = \sup_{\mathcal{A}\subseteq\mathcal{Y} \text{ measurable}} |P(\mathcal{A})-Q(\mathcal{A})| \geq P(\mathcal{Y}_P)-Q(\mathcal{Y}_P) = 1-Q(\mathcal{Y}_P). \tag{P.119}$$

This entails that $Q(p>0)=Q(\mathcal{Y}_P)\geq 1-\mathrm{TV}(P,Q)$. To prove that $Q(p>0)>0$, we show that $\mathrm{TV}(P,Q)<1$ when there exists $\alpha\in(1,2]$ such that $I_\alpha(P\|Q)<+\infty$:

$$\mathrm{TV}(P,Q) \leq 1-\exp(\mathrm{KL}(P\|Q)) \leq 1-\exp\left(-\frac{1}{\alpha-1}\log I_\alpha(P\|Q)\right) = 1-I_\alpha(P\|Q)^{-\frac{1}{\alpha-1}} < 1, \tag{P.120}$$

where the first inequality follows from Equation 2.2 of [7] and the second from the bound between the KL-divergence and the Rényi divergence [72]. Thus, we have $Q(p>0)\geq I_\alpha(P\|Q)^{-\frac{1}{\alpha-1}}$. $\square$

Lemma A.3. *Let $\alpha\in(1,2]$, $s\in[-\infty,-1]$, and $\delta\in(0,1)$. Suppose that Equation (39) admits a unique solution, denoted by $\eta_\alpha^\dagger \in [0,1]$. Let $\lambda_\alpha^\dagger = \eta_\alpha^\dagger n^{-\beta}$. Then, for every $\epsilon_1>0$, it holds that:*

$$1 \leq \frac{\eta_\alpha^\dagger}{\eta_\alpha^*} \leq 1+\epsilon_1 \qquad \text{and} \qquad 1 \leq \frac{\lambda_\alpha^\dagger}{\lambda_\alpha^*} \leq 1+\epsilon_1, \tag{41}$$

where the second inequality holds if $n \geq \left(\frac{16}{3\epsilon_1}\left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}\right)^{-\frac{s}{\alpha}} \frac{I_{\alpha-s}(P\|Q)-I_\alpha(P\|Q)}{I_\alpha(P\|Q)^{1-\frac{s}{\alpha}}}\right)^{\frac{1}{-\frac{s}{\alpha}-\beta}}$, whenever $I_{\alpha-s}(P\|Q)$ is finite.

Proof. Let us first observe that:

$$\mathop{\mathbb{E}}_{y\sim Q}\left[\omega_{\eta,s}(y)^\alpha\right] = \mathop{\mathbb{E}}_{y\sim Q}\left[((1-\eta)\omega(y)^s+\eta)^{\frac{\alpha}{s}}\right] \tag{P.121}$$

$$\leq \mathop{\mathbb{E}}_{y\sim Q}\left[((1-\eta)\omega(y)+\eta)^\alpha\right] \tag{P.122}$$

$$\leq \mathop{\mathbb{E}}_{y\sim Q}\left[(1-\eta)\omega(y)^\alpha+\eta\right] \tag{P.123}$$

$$= (1-\eta)I_\alpha(P\|Q)+\eta \leq I_\alpha(P\|Q), \tag{P.124}$$

where line (P.122) derives from the inequality between the power mean with exponent $s\leq 1$ and the arithmetic mean and line (P.123) from Jensen's inequality. From the last inequality, we have:

$$h_{\alpha,s}(\eta) \leq \eta^{-\frac{\alpha}{s}} I_\alpha(P\|Q) \quad \Longrightarrow \quad \eta_\alpha^\dagger \geq \left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 I_\alpha(P\|Q) n^{1+\frac{\alpha\beta}{s}}}\right)^{-\frac{s}{\alpha}} \tag{P.125}$$

$$\Longrightarrow \quad \lambda_\alpha^\dagger \geq \left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 I_\alpha(P\|Q) n}\right)^{\frac{1}{\alpha}} = \lambda_\alpha^*. \tag{P.126}$$

Concerning the lower bound, we proceed with a second-order Taylor expansion centered in $\eta=0$:

$$((1-\eta)\omega(y)^s+\eta)^{\frac{\alpha}{s}} = \omega(y)^\alpha + \frac{\alpha\eta}{s}(\omega(y)^{-s}-1)\omega(y)^\alpha + \frac{\alpha\overline{\eta}^2}{2s^2}(\alpha-s)\omega(y)^{\alpha-2s}(\omega(y)^s-1)^2 \tag{P.127}$$

$$\geq \omega(y)^{\alpha} + \frac{\alpha\eta}{s}\left(\omega(y)^{-s} - 1\right)\omega(y)^{\alpha}, \tag{P.128}$$

for some $\overline{\eta} \in [0, \eta]$. From which, we obtain:

$$\mathop{\mathbb{E}}_{y\sim Q}\left[\left((1-\eta)\omega(y)^{s} + \eta\right)^{\frac{\alpha}{s}}\right] \geq \mathop{\mathbb{E}}_{y\sim Q}\left[\omega(y)^{\alpha} + \frac{\alpha\eta}{s}\left(\omega(y)^{-s} - 1\right)\omega(y)^{\alpha}\right] \tag{P.129}$$

$$= I_{\alpha}(P\|Q) + \frac{\alpha\eta}{s}(I_{\alpha-s}(P\|Q) - I_{\alpha}(P\|Q)). \tag{P.130}$$

By moving to function $h_{\alpha,s}(\eta)$, and recalling the Equation (39), we have:

$$h_{\alpha,s}(\eta) = \eta^{-\frac{\alpha}{s}} \mathop{\mathbb{E}}_{y\sim Q}\left[\omega_{\eta}(y)^{\alpha}\right] \geq \eta^{-\frac{\alpha}{s}} I_{\alpha}(P\|Q) + \frac{\alpha}{s}\eta^{1-\frac{\alpha}{s}}(I_{\alpha-s}(P\|Q) - I_{\alpha}(P\|Q)) \tag{P.131}$$

$$\implies \quad \eta^{-\frac{\alpha}{s}} I_{\alpha}(P\|Q) + \frac{\alpha}{s}\eta^{1-\frac{\alpha}{s}}(I_{\alpha-s}(P\|Q) - I_{\alpha}(P\|Q)) \leq \frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 n^{1+\frac{\alpha\beta}{s}}}. \tag{P.132}$$

We prove that for sufficiently large n, all solutions $\eta_{\alpha}^{\dagger}$ of the previous inequality satisfy $\eta \leq (1+\epsilon_1)\left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 I_{\alpha}(P\|Q) n^{1+\frac{\alpha\beta}{s}}}\right)^{-\frac{s}{\alpha}}$:

$$(1+\epsilon_1)^{-\frac{\alpha}{s}} \frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 I_{\alpha}(P\|Q) n^{1+\frac{\alpha\beta}{s}}} I_{\alpha}(P\|Q) + (1+\epsilon_1)^{1-\frac{\alpha}{s}} \frac{\alpha}{s}\left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 I_{\alpha}(P\|Q) n^{1+\frac{\alpha\beta}{s}}}\right)^{1-\frac{s}{\alpha}} \tag{P.133}$$

$$\times \left(I_{\alpha-s}(P\|Q) - I_{\alpha}(P\|Q)\right) \geq \frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 n^{1+\frac{\alpha\beta}{s}}} \tag{P.134}$$

$$\implies \quad n \geq \left(-\frac{(1+\epsilon_1)^{1-\frac{\alpha}{s}}}{\left((1+\epsilon_1)^{-\frac{\alpha}{s}} - 1\right)} \cdot \frac{\alpha}{s} \cdot \left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}\right)^{-\frac{s}{\alpha}} \cdot \frac{I_{\alpha-s}(P\|Q) - I_{\alpha}(P\|Q)}{I_{\alpha}(P\|Q)^{1-\frac{s}{\alpha}}}\right)^{\frac{1}{-\frac{s}{\alpha}-\beta}}. \tag{P.135}$$

Observing that when $\epsilon_1 \in (0,1)$, we have that:

$$\frac{(1+\epsilon_1)^{1-\frac{\alpha}{s}}}{\left((1+\epsilon_1)^{-\frac{\alpha}{s}} - 1\right)} \cdot \frac{\alpha}{s} \leq \frac{16}{3\epsilon_1}, \tag{P.136}$$

that we use to enforce the condition in the statement. This, implies that $\lambda_{\alpha}^{\dagger} \leq (1+\epsilon_1)\left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 I_{\alpha}(P\|Q) n}\right)^{\frac{1}{\alpha}} = (1+\epsilon_1)\lambda_{\alpha}^{*}$. □

Lemma A.4. *Let $\alpha \in (1,2]$, $s \in [-\infty, -1]$, and $\eta \in [0,1]$. It holds that*

$$\frac{\partial h_{\alpha,s}(\eta)}{\partial \eta^{-\frac{\alpha}{s}}} \geq \left(2 I_{\alpha}(P\|Q)\right)^{-\frac{1-s}{\alpha-1}}. \tag{42}$$

Proof. Let us first observe that:

$$\frac{\partial h_{\alpha,s}(\eta)}{\partial \eta^{-\frac{\alpha}{s}}} = \frac{\partial h_{\alpha,s}(\eta)}{\partial \eta}\frac{\partial \eta}{\partial \eta^{-\frac{\alpha}{s}}} = -\frac{\partial h_{\alpha,s}(\eta)}{\partial \eta}\left(-\frac{\alpha\eta^{-\frac{\alpha}{s}-1}}{s}\right)^{-1}. \tag{P.137}$$

The first factor was already computed in the proof of Lemma A.1. We now lower bound it. Let us first prove the following auxiliary inequality, using $p \geq 1$:

$$
\begin{aligned}
1 = \mathop{\mathbb{E}}_{y\sim Q}[\omega(y)] = \mathop{\mathbb{E}}_{y\sim Q}\Bigg[\Bigg(-\frac{\alpha\eta^{-\frac{\alpha}{s}-1}}{s}\omega(y)^s\left((1-\eta)\omega(y)^s+\eta\right)^{-1+\frac{\alpha}{s}}\Bigg)^{\frac{1}{p}} \\
\times\omega(y)\Bigg(-\frac{\alpha\eta^{-\frac{\alpha}{s}-1}}{s}\omega(y)^s\left((1-\eta)\omega(y)^s+\eta\right)^{-1+\frac{\alpha}{s}}\Bigg)^{-\frac{1}{p}}\Bigg] \\
\leq\Bigg(\underbrace{-\frac{\alpha\eta^{-\frac{\alpha}{s}-1}}{s}\mathop{\mathbb{E}}_{y\sim Q}\left[\omega(y)^s\left((1-\eta)\omega(y)^s+\eta\right)^{-1+\frac{\alpha}{s}}\right]}_{\frac{\partial h_{\alpha,s}(\eta)}{\partial\eta}}\Bigg)^{\frac{1}{p}} \\
\times\Bigg(-\frac{\alpha\eta^{-\frac{\alpha}{s}-1}}{s}\Bigg)^{-\frac{1}{p}}\mathop{\mathbb{E}}_{y\sim Q}\left[\omega(y)^{\frac{p-s}{p-1}}\left((1-\eta)\omega(y)^s+\eta\right)^{\frac{s-\alpha}{s(p-1)}}\right]^{\frac{p-1}{p}},
\end{aligned}
\tag{P.138}
$$

where the inequality follows from Hölder's inequality with exponents p and $\frac{p}{p-1}$. For the second factor, we proceed as follows:

$$
\begin{aligned}
\omega(y)^{\frac{p-s}{p-1}}\left((1-\eta)\omega(y)^s+\eta\right)^{\frac{s-\alpha}{s(p-1)}} &\leq \omega(y)^{\frac{p-s}{p-1}}\left((1-\eta)\omega(y)^{s(\alpha-s)}+\eta\right)^{-\frac{1}{s(p-1)}} && \text{(P.139)}\\
&\leq \omega(y)^{\frac{p-s}{p-1}}\left((1-\eta)^{-\frac{1}{s}}\omega(y)^{-(\alpha-s)}+\eta^{-\frac{1}{s}}\right)^{\frac{1}{p-1}} && \text{(P.140)}\\
&\leq \omega(y)^{\frac{p-\alpha}{p-1}}+\omega(y)^{\frac{p-s}{p-1}}. && \text{(P.141)}
\end{aligned}
$$

where line (P.139) follows from having observed that $\alpha - s \geq 1$ and applies Jensen's inequality, and line (P.140) is obtained from the subadditivity of $(\cdot)^{-\frac{1}{s}}$ with $s \leq -1$. We now take $p = \frac{\alpha-s}{\alpha-1} \geq 1$. Thus, we have:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{y\sim Q}\left[\omega(y)^{\frac{p-s}{p-1}}\left((1-\eta)\omega(y)^s+\eta\right)^{\frac{s-\alpha}{s(p-1)}}\right] &\leq \mathop{\mathbb{E}}_{y\sim Q}\left[\omega(y)^{\frac{p-\alpha}{p-1}}+\omega(y)^{\frac{p-s}{p-1}}\right] && \text{(P.142)}\\
&= \mathop{\mathbb{E}}_{y\sim Q}\left[\omega(y)^{\frac{-s-\alpha^2+2\alpha}{-s+1}}+\omega(y)^{\alpha}\right] && \text{(P.143)}\\
&= I_{\frac{-s-\alpha^2+2\alpha}{-s+1}}(P\|Q)+I_\alpha(P\|Q) \leq 2I_\alpha(P\|Q), && \text{(P.144)}
\end{aligned}
$$

having observed that $\frac{-s-\alpha^2+2\alpha}{-s+1} \in [1/2, 1]$. Now, using inequality (P.138), raised to the power p, we obtain:

$$
1 \leq \frac{\partial h_{\alpha,s}(\eta)}{\partial\eta}\left(-\frac{\alpha\eta^{-\frac{\alpha}{s}-1}}{s}\right)^{-1}\left(2I_\alpha(P\|Q)\right)^{\frac{1-s}{\alpha-1}} = \frac{\partial h_{\alpha,s}(\eta)}{\partial\eta^{-\frac{\alpha}{s}}}\left(2I_\alpha(P\|Q)\right)^{\frac{1-s}{\alpha-1}}. \quad \square \tag{P.145}
$$

Lemma A.5. *Let $\alpha \in (1,2]$, $s \in [-\infty, -1]$. Then, $n\widehat{h}_{\alpha,s}(\eta)$ is a self-bounding function. Therefore, for every $\eta \in [0,1]$ and $\epsilon > 0$, it holds that:*

$$
\mathop{\mathbb{P}}_{y_i\sim Q}\left(\widehat{h}_{\alpha,s}(\eta) - h_{\alpha,s}(\eta) \geq \epsilon\right) \leq \exp\left(\frac{-\epsilon^2 n}{2\left(h_{\alpha,s}(\eta)+\epsilon/3\right)}\right), \tag{43}
$$

$$
\mathop{\mathbb{P}}_{y_i\sim Q}\left(h_{\alpha,s}(\eta) - \widehat{h}_{\alpha,s}(\eta) \geq \epsilon\right) \leq \exp\left(\frac{-\epsilon^2 n}{2(h_{\alpha,s}(\eta)+\epsilon)}\right). \tag{44}
$$

Proof. We consider the definition of self-bounding function provided in [6, Definition 2]. We denote with $n\widehat{h}^{k,z}_{\alpha}(\eta)$ the function obtained from $n\widehat{h}_{\alpha,s}(\eta)$ by replacing $\omega(y_k)$ with $z \geq 0$. We show that $n\widehat{h}_{\alpha,s}(\eta)$ satisfies both conditions with $a = 1$ and $b = 0$:

$$
\begin{aligned}
n\widehat{h}_{\alpha,s}(\eta) - n\inf_{z\geq 0}\widehat{h}^{k,z}_{\alpha,s}(\eta) = \eta^{-\frac{\alpha}{s}}\left(\omega_{\eta,s}(y_k)^\alpha - \inf_{z\geq 0} z^\alpha\right) \leq \eta^{-\frac{\alpha}{s}}\omega_{\eta,s}(y_k)^\alpha \leq 1, && \text{(P.146)}\\
\sum_{k\in[\![n]\!]}\left(n\widehat{h}_{\alpha,s}(\eta) - n\inf_{z\geq 0}\widehat{h}^{k,z}_{\alpha,s}(\eta)\right)^2 = \sum_{k\in[\![n]\!]}\eta^{-\frac{\alpha}{s}}\left(\omega_{\eta,s}(y_k)^\alpha - \inf_{z\geq 0} z^\alpha\right) && \text{(P.147)}\\
= \sum_{k\in[\![n]\!]}\eta^{-\frac{\alpha}{s}}\omega_{\eta,s}(y_k)^\alpha && \text{(P.148)}\\
= n\widehat{h}_{\alpha,s}(\eta), && \text{(P.149)}
\end{aligned}
$$

having observed that $\eta^{-\frac{1}{s}}\omega_{\eta,s}(y_k) \leq 1$. □

Lemma A.6. *Let $\alpha \in (1,2]$, $s \in [-\infty,-1]$, and $\delta \in (0,1)$. Suppose that Equation (39) and (40) admit a unique solution, denoted by $\eta_\alpha^\dagger \in [0,1]$ and $\widehat{\eta}_\alpha \in [0,1]$, respectively. Let $\lambda_\alpha^\dagger = \eta_\alpha^\dagger n^{-\beta}$ and $\widehat{\lambda}_\alpha = \widehat{\eta}_\alpha n^{-\beta}$. Then, for every $\epsilon_2 \in (0,1)$, with probability at least $1-2\delta$, it holds that:*

$$1-\epsilon_2 \leq \frac{\widehat{\eta}_\alpha}{\eta_\alpha^\dagger} \leq 1+\epsilon_2 \qquad \text{and} \qquad 1-\epsilon_2 \leq \frac{\widehat{\lambda}_\alpha}{\lambda_\alpha^\dagger} \leq 1+\epsilon_2, \tag{45}$$

for $n \geq \left(\epsilon_2^{-2} \cdot 2^{\frac{1-2s+\alpha}{\alpha-1}} (\alpha-1)^2 I_\alpha(P\|Q)^{\frac{2(\alpha-s)}{\alpha-1}}\right)^{-\frac{s}{\alpha\beta}}$.

Proof. For the sake of this derivation, we will omit the subscripts α, s, simply denoting $h_{\alpha,s}$ with h and $\widehat{h}_{\alpha,s}$ with $\widehat{h}$. Let $\epsilon \in [0,1]$, consider the event $\left\{\left|\frac{\widehat{\eta}}{\eta^\dagger}-1\right| > \epsilon\right\}$. Under the sub-event $\{\widehat{\eta} > (1+\epsilon)\eta^\dagger\}$ recalling that function h and $\widehat{h}$ are increasing in η we have:

$$\begin{aligned}
\widehat{h}(\widehat{\eta}) - \widehat{h}(\eta^\dagger) &\geq \widehat{h}((1+\epsilon)\eta^\dagger) - \widehat{h}(\eta^\dagger) && \text{(P.150)}\\
&= \widehat{h}((1+\epsilon)\eta^\dagger) - \widehat{h}(\eta^\dagger) \pm h(\eta^\dagger) \pm h((1+\epsilon)\eta^\dagger) && \text{(P.151)}\\
&= \widehat{h}((1+\epsilon)\eta^\dagger) - h((1+\epsilon)\eta^\dagger) + h(\eta^\dagger) - \widehat{h}(\eta^\dagger) + h((1+\epsilon)\eta^\dagger) - h(\eta^\dagger) && \text{(P.152)}\\
&\geq \widehat{h}((1+\epsilon)\eta^\dagger) - h((1+\epsilon)\eta^\dagger) + h(\eta^\dagger) - \widehat{h}(\eta^\dagger) + \alpha(2I_2(P\|Q))^{-\frac{1-s}{\alpha-1}} \epsilon(\eta^\dagger)^{-\frac{\alpha}{s}}, && \text{(P.153)}
\end{aligned}$$

where the last inequality follows from Lemma A.4 having applied:

$$\begin{aligned}
h((1+\epsilon)\eta^\dagger) - h(\eta^\dagger) &\geq I_\alpha(P\|Q)^{-\frac{2}{\alpha-1}} \left((1+\epsilon)^\alpha - 1\right)(\eta^\dagger)^\alpha && \text{(P.154)}\\
&\geq \alpha(2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}} \epsilon(\eta^\dagger)^{-\frac{\alpha}{s}}, && \text{(P.155)}
\end{aligned}$$

having exploited the inequality $(1+\epsilon)^\alpha - 1 \geq \alpha\epsilon$ for $\alpha \in (1,2]$. Recalling that $\widehat{h}(\widehat{\eta}) = h(\eta^\dagger)$, the condition can be further simplified into $h((1+\epsilon)\eta^\dagger) - \widehat{h}((1+\epsilon)\eta^\dagger) \geq \alpha(2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}} \epsilon(\eta^\dagger)^{-\frac{\alpha}{s}}$. Symmetrically, under the sub-event $\{\widehat{\eta} < (1-\epsilon)\eta^\dagger\}$ we have:

$$\begin{aligned}
\widehat{h}(\widehat{\eta}) - \widehat{h}(\eta^\dagger) &\leq \widehat{h}((1-\epsilon)\eta^\dagger) - \widehat{h}(\eta^\dagger) && \text{(P.156)}\\
&= \widehat{h}((1-\epsilon)\eta^\dagger) - \widehat{h}(\eta^\dagger) \pm h(\eta^\dagger) \pm h((1-\epsilon)\eta^\dagger) && \text{(P.157)}\\
&= \widehat{h}((1-\epsilon)\eta^\dagger) - h((1-\epsilon)\eta^\dagger) + h(\eta^\dagger) - \widehat{h}(\eta^\dagger) + h((1-\epsilon)\eta^\dagger) - h(\eta^\dagger) && \text{(P.158)}\\
&\leq \widehat{h}((1-\epsilon)\eta^\dagger) - h((1-\epsilon)\eta^\dagger) + h(\eta^\dagger) - \widehat{h}(\eta^\dagger) - (2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}} \left(1-(1-\epsilon)^\alpha\right)(\eta^\dagger)^{-\frac{\alpha}{s}}, && \text{(P.159)}
\end{aligned}$$

that can be simplified, as before, into the condition $\widehat{h}((1-\epsilon)\eta^\dagger) - h((1-\epsilon)\eta^\dagger) \geq (2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}} \epsilon(\eta^\dagger)^{-\frac{\alpha}{s}}$ since $1-(1-\epsilon)^\alpha \geq \epsilon$ being $\epsilon < 1$. Thus, we have:

$$\begin{aligned}
\mathop{\mathbb{P}}_{y_i\sim Q}\left(\left|\frac{\widehat{\eta}}{\eta^\dagger}-1\right| > \epsilon\right) &= \mathop{\mathbb{P}}_{y_i\sim Q}\left(\widehat{\eta} > (1+\epsilon)\eta^\dagger\right) + \mathop{\mathbb{P}}_{y_i\sim Q}\left(\widehat{\eta} < (1-\epsilon)\eta^\dagger\right) && \text{(P.160)}\\
&\leq \mathop{\mathbb{P}}_{y_i\sim Q}\left(h((1+\epsilon)\eta^\dagger) - \widehat{h}((1+\epsilon)\eta^\dagger) \geq \alpha(2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}} \epsilon(\eta^\dagger)^{-\frac{\alpha}{s}}\right) && \text{(P.161)}\\
&\quad + \mathop{\mathbb{P}}_{y_i\sim Q}\left(\widehat{h}((1-\epsilon)\eta^\dagger) - h((1-\epsilon)\eta^\dagger) \geq (2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}} \epsilon(\eta^\dagger)^{-\frac{\alpha}{s}}\right). && \text{(P.162)}
\end{aligned}$$

First of all, we observe that $h((1+\epsilon)\eta^\dagger) = (1+\epsilon)^{-\frac{\alpha}{s}}(\eta^\dagger)^{-\frac{\alpha}{s}} \mathbb{E}_{y\sim Q}[\omega_{(1+\epsilon)\eta^\dagger,s}(y)^\alpha] \leq 2^{-\frac{\alpha}{s}}(\eta^\dagger)^{-\frac{\alpha}{s}} I_\alpha(P\|Q)$ and that $h((1-\epsilon)\eta^\dagger) \leq h(\eta^\dagger) \leq (\eta^\dagger)^{-\frac{\alpha}{s}} I_\alpha(P\|Q)$. Now, recalling that function h is self-bounding as proved in Lemma A.5, we have by Equation (44):

$$\begin{aligned}
&\mathop{\mathbb{P}}_{y_i\sim Q}\left(h((1+\epsilon)\eta^\dagger) - \widehat{h}((1+\epsilon)\eta^\dagger) \geq \alpha(2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}} \epsilon(\eta^\dagger)^{-\frac{\alpha}{s}}\right) && \text{(P.163)}\\
&\leq \exp\left(\frac{-\alpha^2(2I_\alpha(P\|Q))^{-\frac{2(1-s)}{\alpha-1}} \epsilon^2(\eta^\dagger)^{-\frac{2\alpha}{s}} n}{2\left(h((1+\epsilon)\eta^\dagger) + \alpha(2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}} \epsilon(\eta^\dagger)^{-\frac{\alpha}{s}}\right)}\right) && \text{(P.164)}\\
&\leq \exp\left(\frac{-\alpha^2(2I_\alpha(P\|Q))^{-\frac{2(1-s)}{\alpha-1}} \epsilon^2(\eta^\dagger)^{-\frac{2\alpha}{s}} n}{2\left(2^{-\frac{\alpha}{s}}(\eta^\dagger)^{-\frac{\alpha}{s}} I_\alpha(P\|Q) + \alpha(2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}} \epsilon(\eta^\dagger)^{-\frac{\alpha}{s}}\right)}\right) && \text{(P.165)}
\end{aligned}$$

$$\leq \exp\left(\frac{-\alpha^2\epsilon^2(\eta^\dagger)^{-\frac{\alpha}{s}}n}{2^{1+\frac{2(1-s)}{\alpha-1}}\left(2^{-\frac{\alpha}{s}}+\frac{\alpha}{4}\right)I_\alpha(P\|Q)^{\frac{1-2s+\alpha}{\alpha-1}}}\right) \tag{P.166}$$

$$\leq \exp\left(\frac{-4\epsilon^2(\eta^\dagger)^{-\frac{\alpha}{s}}n}{9(2I_\alpha(P\|Q))^{\frac{1-2s+\alpha}{\alpha-1}}}\right), \tag{P.167}$$

having crudely bounded $(2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}}\epsilon \leq \frac{1}{4}I_\alpha(P\|Q)$ and observed that $\frac{\alpha^2}{2^{-\alpha/s+\alpha/4}} \geq 4/9$. Similarly, by Equation (43), we have:

$$\mathbb{P}_{y_i\sim Q}\left(\widehat{h}((1-\epsilon)\eta^\dagger) - h((1-\epsilon)\eta^\dagger) \geq (2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}}\epsilon^2(\eta^\dagger)^{-\frac{\alpha}{s}}\right) \tag{P.168}$$

$$\leq \exp\left(\frac{-(2I_\alpha(P\|Q))^{-\frac{2(1-s)}{\alpha-1}}\epsilon^2(\eta^\dagger)^{-\frac{2\alpha}{s}}n}{2\left(h((1-\epsilon)\eta^\dagger)+\frac{1}{3}(2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}}\epsilon(\eta^\dagger)^{-\frac{\alpha}{s}}\right)}\right) \tag{P.169}$$

$$\leq \exp\left(\frac{-(2I_\alpha(P\|Q))^{-\frac{2(1-s)}{\alpha-1}}\epsilon(\eta^\dagger)^{-\frac{2\alpha}{s}}n}{2\left((\eta^\dagger)^{-\frac{\alpha}{s}}I_\alpha(P\|Q)+\frac{1}{3}(2I_\alpha(P\|Q))^{-\frac{1-s}{\alpha-1}}\epsilon(\eta^\dagger)^{-\frac{\alpha}{s}}\right)}\right) \tag{P.170}$$

$$\leq \exp\left(\frac{-3\epsilon^2(\eta^\dagger)^{-\frac{\alpha}{s}}n}{4(2I_\alpha(P\|Q))^{\frac{1-2s+\alpha}{\alpha-1}}}\right). \tag{P.171}$$

Putting these inequalities together, we obtain:

$$\mathbb{P}_{y_i\sim Q}\left(\left|\frac{\widehat{\eta}}{\eta^\dagger}-1\right|>\epsilon\right) \leq \exp\left(\frac{-4\epsilon^2(\eta^\dagger)^{\alpha}n}{9(2I_\alpha(P\|Q))^{\frac{1-2s+\alpha}{\alpha-1}}}\right)+\exp\left(\frac{-3\epsilon^2(\eta^\dagger)^{-\frac{\alpha}{s}}n}{4(2I_\alpha(P\|Q))^{\frac{1-2s+\alpha}{\alpha-1}}}\right) \tag{P.172}$$

$$\leq 2\exp\left(\frac{-4\epsilon^2(\eta^\dagger)^{\alpha}n}{9(2I_\alpha(P\|Q))^{\frac{1-2s+\alpha}{\alpha-1}}}\right), \tag{P.173}$$

leading to the inequality holding with probability at least $1-2\delta$:

$$\left|\frac{\widehat{\eta}}{\eta^\dagger}-1\right| \leq \sqrt{\frac{9(2I_\alpha(P\|Q))^{\frac{1-2s+\alpha}{\alpha-1}}\log\left(\frac{1}{\delta}\right)}{4n(\eta^\dagger)^{-\frac{\alpha}{s}}}}. \tag{P.174}$$

Thanks to Lemma A.3, we know that $\eta^\dagger \geq \left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2 I_\alpha(P\|Q)n^{1+\frac{\alpha\beta}{s}}}\right)^{-\frac{s}{\alpha}}$. From which we have:

$$\left|\frac{\widehat{\eta}}{\eta^\dagger}-1\right| \leq \sqrt{\frac{9\cdot 2^{\frac{1-2s+\alpha}{\alpha-1}}(\alpha-1)^2 I_\alpha(P\|Q)^{\frac{2(\alpha-s)}{\alpha-1}}}{8n^{-\frac{\alpha\beta}{s}}}}. \tag{P.175}$$

By setting the right-hand side equal to $\epsilon_2 \in (0,1)$, we obtain the minimum number of samples needed to guarantee that with probability at least $1-\delta$, it holds that $1-\epsilon_2 \leq \frac{\widehat{\eta}}{\eta^\dagger} \leq 1+\epsilon_2$:

$$n \geq \left(\epsilon_2^{-2}\cdot\frac{9}{8}\cdot 2^{\frac{1-2s+\alpha}{\alpha-1}}(\alpha-1)^2 I_\alpha(P\|Q)^{\frac{2(\alpha-s)}{\alpha-1}}\right)^{-\frac{s}{\alpha\beta}}. \quad \square \tag{P.176}$$

Lemma 6.3. *Let $\alpha \in (1,2]$, $s \in [-\infty,-1]$, and $\delta \in (0,1)$. Suppose that Equation (28) admitsa a unique solution $\widehat{\lambda}_\alpha$. Then, for every $\epsilon_3 \in (0,1)$, it holds that:*

$$1-\epsilon_3 \leq \frac{\lambda^{(k)}}{\widehat{\lambda}_\alpha} \leq 1+\epsilon_3, \tag{32}$$

for a number of iterations $k \geq \left\lceil \log_2\left(\frac{n^{-\beta}}{\widehat{\lambda}_\alpha\epsilon_3}\right)\right\rceil$.

Proof. From the properties of the bisection algorithm, we have that for every iteration $k \geq 1$, we have:

$$\left|\lambda^{(k)} - \widehat{\lambda}_\alpha\right| \leq \lambda_+^{(k)} - \lambda_-^{(k)} \leq \frac{\lambda_+^{(0)} - \lambda_-^{(0)}}{2^k}. \tag{P.177}$$

Recalling that $\lambda_+^{(0)} - \lambda_-^{(0)} = n^{-\beta}$, we have:

$$\left|\frac{\lambda^{(k)}}{\widehat{\lambda}_\alpha} - 1\right| \leq \frac{n^{-\beta}}{\widehat{\lambda}_\alpha 2^k}. \tag{P.178}$$

By setting the right-hand side equal to ϵ_3, we obtain the number of iterations. □

Theorem 6.1. *Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability distributions such that $P \ll Q$. Let $\widehat{\lambda}_\alpha$ be the solution of Equation (28), then, if $I_{\alpha-s}(P\|Q)$ is finite, for every $\delta \in (0,1)$, with probability at least $1-4\delta$, it holds that:*

$$\widehat{\mu}_{n,\lambda^{(k)},s} - \mu \leq \frac{14\|f\|_\infty}{3}\left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2}\right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q)^{\frac{1}{\alpha}}, \tag{33}$$

under the conditions that:

$$n \geq \max\left\{\left(19\left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}\right)^{-\frac{s}{\alpha}} \frac{I_{\alpha-s}(P\|Q) - I_\alpha(P\|Q)}{I_\alpha(P\|Q)^{1-\frac{s}{\alpha}}}\right)^{\frac{1}{-\frac{s}{\alpha}-\beta}}, \left(26(\alpha-1)^2 2^{\frac{1-2s+\alpha}{\alpha-1}} I_\alpha(P\|Q)^{\frac{2(\alpha-s)}{\alpha-1}}\right)^{-\frac{s}{\alpha\beta}}, \tag{34}$$

$$\left(\frac{6 I_\alpha(P\|Q)^{\frac{1}{\alpha-1}}\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}\right)^{\frac{1}{1+\frac{\alpha\beta}{s}}}\right\}, \tag{35}$$

$$k \geq 4 + \left(-\beta - \frac{s}{\alpha}\right)\log_2 n - \frac{s}{\alpha}\log_2 \frac{(1-\alpha)^2 I_\alpha(P\|Q)}{2\log\left(\frac{1}{\delta}\right)}. \tag{36}$$

Proof. By combining Lemma 6.2 and Lemma 6.3, we have that, with probability at least $1-2\delta$, it holds that:

$$(1-\epsilon_3)(1-\epsilon_2) \leq \frac{\lambda^{(k)}}{\lambda_\alpha^*} \leq (1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3). \tag{P.179}$$

Let $\xi := \frac{\lambda^{(k)}}{\lambda_\alpha^*}$, $\xi_+ := (1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3) > 1$, and $\xi_- = (1-\epsilon_2)(1-\epsilon_3) < 1$, starting from the proof of Theorem 5.1, we have that with probability $1-\delta$:

$$\widehat{\mu}_{n,\lambda^{(k)},s} - \mu \leq \|f\|_\infty\left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2}\right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q)^{\frac{1}{\alpha}}\left((\alpha-1)\xi^{\frac{2-\alpha}{2s}} + \frac{(\alpha-1)^2}{3}\xi^{\frac{1}{s}} + (3-\alpha)^{\frac{1}{\alpha}}\xi^{\frac{1-\alpha}{s}}\right). \tag{P.180}$$

By hypothesis of the theorem, noting the signs of the exponents of ξ, we can write:

$$(\alpha-1)\xi^{\frac{2-\alpha}{2s}} + \frac{(\alpha-1)^2}{3}\xi^{\frac{1}{s}} + (3-\alpha)^{\frac{1}{\alpha}}\xi^{\frac{1-\alpha}{s}} \leq (\alpha-1)\xi_-^{\frac{2-\alpha}{2s}} + \frac{(\alpha-1)^2}{3}\xi_-^{\frac{1}{s}} + (3-\alpha)^{\frac{1}{\alpha}}\xi_+^{\frac{1-\alpha}{s}} \tag{P.181}$$

$$\leq \underbrace{\left((\alpha-1) + \frac{(\alpha-1)^2}{3} + (3-\alpha)^{\frac{1}{\alpha}}\right)}_{\leq 7/3} \max\left\{\xi_-^{\frac{2-\alpha}{2s}}, \xi_-^{\frac{1}{s}}, \xi_+^{\frac{1-\alpha}{s}}\right\} \tag{P.182}$$

$$\leq \frac{7}{3}\max\left\{\xi_-^{-1}, \xi_+\right\}^{-\frac{1}{s}}. \tag{P.183}$$

We set ϵ_1 and $\epsilon_2 = \epsilon_3$ such that $(1-\epsilon_2)(1-\epsilon_3) = 1/2$ and $(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_3) = 2$ that leads to:

$$\epsilon_1 = 1 - \frac{\sqrt{2}}{2} \approx 0.293, \quad \epsilon_2 = \epsilon_3 = \frac{3 - 4(2 - \sqrt{2})}{1 + 4(2 - \sqrt{2})} \approx 0.197. \tag{P.184}$$

Thus, $\xi_- = 1/2$ and $\xi_+ = 2$ leading to:

$$\widehat{\mu}_{n,\lambda^{(k)},s} - \mu \leq \frac{14\|f\|_\infty}{3} \left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2} \right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q)^{\frac{1}{\alpha}}. \tag{P.185}$$

A union bound allows us to express the concentration inequality. Furthermore, we know that $\frac{\widehat{\lambda}_\alpha}{\lambda^*_\alpha} \geq 1 - \epsilon_2 \geq \frac{4}{5}$, from which, we have that:

$$\widehat{\lambda}_\alpha \geq \frac{4}{5}\lambda^*_\alpha = \frac{4}{5} \left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2 I_\alpha(P\|Q)} \right)^{-\frac{s}{\alpha}}. \tag{P.186}$$

From this, we can bound the number of iterations of Algorithm 1:

$$k \leq \log_2 \left\lceil \frac{n^{-\beta}}{\epsilon_3} \left(\frac{n(1-\alpha)^2 I_\alpha(P\|Q)}{2\log\left(\frac{1}{\delta}\right)} \right)^{-\frac{s}{\alpha}} \right\rceil \leq 4 + \left(-\beta - \frac{s}{\alpha}\right)\log_2 n - \frac{s}{\alpha}\log_2 \frac{(1-\alpha)^2 I_\alpha(P\|Q)}{2\log\left(\frac{1}{\delta}\right)}. \quad \square \tag{P.187}$$

Proposition A.2. *Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability distributions such that $P \ll Q$. Let $\{y_i\}_{i\in[\![n]\!]}$ sampled independently from Q. For every $\alpha \in (1,2]$ and $\delta \in (0,1)$, if:*

$$n \geq \frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2}, \tag{46}$$

then, with probability at least $1-\delta$, it holds that:

$$\widehat{\mu}_{n,\lambda^\ddagger_\alpha,s^\ddagger_\alpha} - \mu \leq \frac{7\|f\|_\infty}{3} \left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2} \right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q), \tag{47}$$

having selected $(\lambda^\ddagger_\alpha, s^\ddagger_\alpha)$ such that:

$$\lambda^\ddagger_\alpha = \left(\frac{2\log\left(\frac{1}{\delta}\right)}{(\alpha-1)^2} \right)^{-\frac{s^\ddagger_\alpha}{\alpha}}. \tag{48}$$

Proof. The result is simply obtained by substituting $\lambda^\ddagger$ into Equation (5.4):

$$\widehat{\mu}_{n,\lambda^\ddagger_\alpha,s^\ddagger_\alpha} - \mu \leq \|f\|_\infty \left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2} \right)^{\frac{\alpha-1}{\alpha}} \left((\alpha-1)\sqrt{I_\alpha(P\|Q)} + \frac{(\alpha-1)^2}{3} + (3-\alpha)^{\frac{1}{\alpha}} I_\alpha(P\|Q) \right) \tag{P.188}$$

$$\leq \frac{7\|f\|_\infty}{3} \left(\frac{2\log\left(\frac{1}{\delta}\right)}{n(\alpha-1)^2} \right)^{\frac{\alpha-1}{\alpha}} I_\alpha(P\|Q), \tag{P.189}$$

having recalled that $I_\alpha(P\|Q) \geq 1$. $\square$

A.4. Auxiliary results

Proposition A.3. *Let $\overline{\mu}_n$ be an estimator of μ for a given class of problems (e.g., off-policy estimation). If there exists a problem instance (e.g., $P, Q \in \Delta^{\mathcal{Y}}$) such that for every $\delta \in (\delta_{\min}, \delta_{\max})$, with probability at least δ, it holds that:*

$$\left|\overline{\mu}_n - \mu\right| \geq L(\delta), \tag{49}$$

for some $L : (\delta_{\min}, \delta_{\max}) \to \mathbb{R}_{\geq 0}$*, then, for any* $\beta \geq 0$*, it holds that:*

$$\mathbb{E}\left[\left|\overline{\mu}_n - \mu\right|^{\beta}\right] \geq \sup_{\delta \in (\delta_{\min}, \delta_{\max})} \delta L(\delta)^{\beta}. \tag{50}$$

Proof. Let us denote the bad event $\mathcal{F} = \left\{\left|\overline{\mu}_n - \mu\right| \geq L(\delta)\right\}$. We have:

$$\mathbb{E}\left[\left|\overline{\mu}_n - \mu\right|^{\beta}\right] \geq \mathbb{E}\left[\left|\overline{\mu}_n - \mu\right|^{\beta} \mathbb{1}\{\mathcal{F}\}\right] \geq L(\delta)^{\beta}\, \mathbb{P}(\mathcal{F}) \geq \delta L(\delta)^{\beta}. \tag{P.190}$$

By taking the supremum over the interval $(\delta_{\min}, \delta_{\max})$, we get the result. □

Proposition A.4. *Let* $\overline{\mu}_n$ *be an estimator of* μ *such that for every* $\delta \in (0, \delta_{\max})$*, with probability at most* δ*, it holds that:*

$$\left|\overline{\mu}_n - \mu\right| \leq U(\delta), \tag{51}$$

for some decreasing differentiable function $U : (0, \delta_{\max}) \to \mathbb{R}_{\geq 0}$ *such that* $\lim_{\delta \to 0^+} U(\delta) = +\infty$*, then, for any* $\beta > 0$*, it holds that:*

$$\mathbb{E}\left[\left|\overline{\mu}_n - \mu\right|^{\beta}\right] \leq (1 - \delta_{\max}) U(\delta_{\max})^{\beta} + \lim_{\delta \to 0^+} \delta U(\delta)^{\beta} + \int_{\delta=0}^{\delta_{\max}} U(\delta)^{\beta} \mathrm{d}\delta. \tag{52}$$

Proof. Relating the expected value to the cumulative distribution function, we have:

$$\mathbb{E}\left[\left|\overline{\mu}_n - \mu\right|^{\beta}\right] = \int_{t=0}^{+\infty} \mathbb{P}\left(\left|\overline{\mu}_n - \mu\right|^{\beta} \geq t\right) \mathrm{d}t \tag{P.191}$$

$$\leq U(\delta_{\max})^{\beta} + \int_{t=U(\delta_{\max})^{\beta}}^{U(0)^{\beta}} \mathbb{P}\left(\left|\overline{\mu}_n - \mu\right|^{\beta} \geq t\right) \mathrm{d}t \tag{P.192}$$

$$= U(\delta_{\max})^{\beta} + \int_{t=U(\delta_{\max})^{\beta}}^{U(0)^{\beta}} \mathbb{P}\left(\left|\overline{\mu}_n - \mu\right| \geq t^{1/\beta}\right) \mathrm{d}t. \tag{P.193}$$

By performing the change of variable $U(\delta) = t^{1/\beta}$ ($\mathrm{d}t = U'(\delta) U(\delta)^{\beta-1} \mathrm{d}\delta$), we have:

$$\mathbb{E}\left[\left|\overline{\mu}_n - \mu\right|^{\beta}\right] \leq U(\delta_{\max})^{\beta} + \int_{\delta=\delta_{\max}}^{0} \mathbb{P}\left(\left|\overline{\mu}_n - \mu\right| \geq U(\delta)\right) U'(\delta) U(\delta)^{\beta-1} \mathrm{d}\delta \tag{P.194}$$

$$\leq U(\delta_{\max})^{\beta} - \int_{\delta=0}^{\delta_{\max}} \delta U'(\delta) U^{\beta-1}(\delta) \mathrm{d}\delta, \tag{P.195}$$

$$= (1 - \delta_{\max}) U(\delta_{\max})^{\beta} + \lim_{\delta \to 0^+} \delta U(\delta)^{\beta} + \int_{\delta=0}^{\delta_{\max}} U(\delta)^{\beta} \mathrm{d}\delta, \tag{P.196}$$

having exploited that $\mathbb{P}\left(\left|\overline{\mu}_n - \mu\right| \geq U(\delta)\right) \leq \delta$ and having performed integration by parts. □

A.5. Technical lemmas

Lemma A.7. *For every* $a, b \in \mathbb{R}$ *and* $\alpha \geq 1$*, it holds that:*

$$|a + b|^{\alpha} \leq 2^{\alpha - 1} (|a|^{\alpha} + |b|^{\alpha}). \tag{53}$$

Proof. We will prove a more general statement. Let $\{a_i\}_{i\in[\![m]\!]}$, we have:

$$\left|\sum_{i\in[\![m]\!]} a_i\right| = \left|\sum_{i\in[\![m]\!]} 1\cdot a_i\right| \leq \left(\underbrace{\sum_{i\in[\![m]\!]} 1}_{m}\right)^{\frac{\alpha-1}{\alpha}} \left(\sum_{i\in[\![m]\!]} |a_i|^\alpha\right)^{\frac{1}{\alpha}}, \tag{P.197}$$

where we have applied Hoeffding's inequality with exponents α and $\frac{\alpha}{\alpha-1}$. Thus, we have:

$$\left|\sum_{i\in[\![m]\!]} a_i\right|^\alpha \leq m^{\alpha-1} \sum_{i\in[\![m]\!]} |a_i|^\alpha. \tag{P.198}$$

The result is obtained for $m=2$. □

Lemma A.8. *Consider the function, defined in terms of $a\in(1,2]$, $b\in[1,2)$, and $y\in\mathbb{R}$:*

$$h(y) := \frac{|y-1|^a}{b|y|^a - ay + 1}. \tag{54}$$

The following statements hold:

(i) $h(y)\geq 0$;
(ii) $h(0)=1$ *and* $\lim_{y\to\pm\infty} h(y) = 1/b$;
(iii) *if* $a\in(1,2)$ *then* $h(1)=0$ *and if* $a=2$ $h(1)=1$;
(iv) $\max_{y\in\mathbb{R}} h(y) = \max\{1,\eta(a,b)\}$, *where:*

$$\eta(a,b) := \frac{\left(\left(\frac{b}{a-1}\right)^{\frac{1}{2-a}} - 1\right)^a}{-a\left(\frac{b}{a-1}\right)^{\frac{1}{2-a}} + b\left(\frac{b}{a-1}\right)^{\frac{a}{2-a}} + 1}. \tag{55}$$

Proof. For (i), it suffices to show that the denominator of function h is non-negative. For $y\leq 0$, the result is obvious. Consider $y>0$, we have:

$$b|y|^a - ay + 1 \geq y^a - ay + 1 \geq 2 - a \geq 0, \tag{P.199}$$

where the second inequality follows from minimizing $y^a - ay + 1$ over $y>0$. (ii) and (iii) are trivial. For (iv), we consider the three cases. First, $y>1$, we compute and vanish the derivative:

$$\frac{\partial h}{\partial y}(y) = \frac{a(y-1)^{a-1}\left(by^a - (a-1)y^2\right)}{y(by^a - ay + 1)^2} = 0 \implies y^* = \left(\frac{b}{a-1}\right)^{\frac{1}{2-a}}. \tag{P.200}$$

We observe that, under the conditions on a and b, this point is always $y^*\geq 1$. Since $\lim_{y\to 1^+}\frac{\partial h}{\partial y}(y) > 0$, that $\lim_{y\to+\infty}\frac{\partial h}{\partial y}(y) < 0$, and that h is continuously differentiable for $y>1$, we have that y^* is a point of local maximum. By substituting into the expression of h, we get $\eta(a,b)$. Second, $y\in(0,1)$, by vanishing the derivative, we obtain that there are no stationary points. Since $\lim_{y\to 0^+}\frac{\partial h}{\partial y}(y) < 0$, that $\lim_{y\to 1^-}\frac{\partial h}{\partial y}(y) < 0$, and that h is continuously differentiable for $y\in(0,1)$, we conclude that h is non-increasing in $(0,1)$. Third, $y<0$, we compute and vanish the derivative:

$$\frac{\partial h}{\partial y}(y) = \frac{a(1-y)^{a-1}\left(b(-y)^a - (a-1)(-y)^2\right)}{y(b(-y)^a - a(-y) + 1)^2} = 0 \implies y^\dagger = -\left(\frac{b}{a-1}\right)^{\frac{1}{2-a}}. \tag{P.201}$$

Since $\lim_{y\to 0^-}\frac{\partial h}{\partial y}(y) > 0$, that $\lim_{y\to-\infty}\frac{\partial h}{\partial y}(y) < 0$, and that h is continuously differentiable for $y<0$, we have that $y^\dagger$ is a point of local minimum. It follows that:

$$\max_{y\in\mathbb{R}} h(y) = \max\{1,\eta(a,b)\}. \quad \square \tag{P.202}$$

Lemma A.9. *Let $P\in\Delta^{\mathcal{Y}}$ be a probability measure and $f:\mathcal{Y}\to\mathbb{R}$ be a measurable function such that $\mu = \mathbb{E}_{y\sim P}[f(y)]$. Then, for every $\alpha\in[1,2]$ it holds that:*

$$\mathop{\mathbb{E}}_{y\sim P}\left[|f(y)-\mu|^\alpha\right] \leq (3-\alpha)\mathop{\mathbb{E}}_{y\sim P}\left[|f(y)|^\alpha\right] - |\mu|^\alpha(\alpha-1) \leq (3-\alpha)\mathop{\mathbb{E}}_{y\sim P}\left[|f(y)|^\alpha\right]. \tag{56}$$

Proof. If $\mu = 0$, the inequality trivially holds. Consider $\mu \neq 0$ and the function:

$$g(y) = \frac{f(y)}{\mu}. \tag{P.203}$$

By construction, function g has mean under P equal to 1. We apply Lemma A.8, to obtain:

$$|g(y) - 1|^\alpha \leq \max\{1, \eta(\alpha, b)\} \left(b|g(y)|^\alpha - \alpha g(y) + 1\right). \tag{P.204}$$

By setting $b = 3 - \alpha$, we observe that $\eta(\alpha, 3-\alpha) \leq 1$. Thus, we have:

$$\mathop{\mathbb{E}}_{y\sim P}\left[|g(y) - 1|^\alpha\right] \leq (3-\alpha) \mathop{\mathbb{E}}_{y\sim P}\left[|g(y)|^\alpha\right] - \alpha + 1. \tag{P.205}$$

Passing to function f, we get the result. □

Lemma A.10. *Let $z \geq 1$, $\lambda \in [0,1]$, and $x \geq 0$. It holds that:*

$$\left|1 - ((1-\lambda) + \lambda x^z)^{\frac{1}{z}}\right| \leq \lambda^{\frac{1}{z}} |x - 1|. \tag{57}$$

Proof. Let us consider function:

$$f(x) := \left|1 - ((1-\lambda) + \lambda x^z)^{\frac{1}{z}}\right| - \lambda^{\frac{1}{z}} |x - 1|. \tag{P.206}$$

We observe that $f(1) = 0$. Thus, since f is continuously differentiable for $x \neq 1$, to show that $f(x) \leq 0$ for all $x \geq 0$, it suffices to show that $\frac{\partial f}{\partial x}(x)|_{x>1} \leq 0$ and $\frac{\partial f}{\partial x}(x)|_{x<1} \geq 0$. We start for $x > 1$, compute the derivative, and elaborate on it:

$$\frac{\partial f}{\partial x}(x)|_{x>1} = -\lambda^{\frac{1}{z}} + \frac{\lambda x^{z-1}}{((1-\lambda) + \lambda x^z)^{1-\frac{1}{z}}} \leq -\lambda^{\frac{1}{z}} + \frac{\lambda x^{z-1}}{(\lambda x^z)^{1-\frac{1}{z}}} \leq 0. \tag{P.207}$$

Consider now $x < 1$, we compute and elaborate on the derivative:

$$\frac{\partial f}{\partial x}(x)|_{x<1} = \lambda^{\frac{1}{z}} - \frac{\lambda x^{z-1}}{((1-\lambda) + \lambda x^z)^{1-\frac{1}{z}}} \geq \lambda^{\frac{1}{z}} - \frac{\lambda x^{z-1}}{(\lambda x^z)^{1-\frac{1}{z}}} \geq 0. \quad \square \tag{P.208}$$

Appendix B. Experiments

In this appendix, we report the experimental details and additional experimental results.

Infrastructure. The experiments have been run on a machine with two CPUs Intel(R) Xeon(R) CPU E7-8880 v4 @ 2.20GHz (22 cores, 44 threads, 55 MB cache) and 128 GB RAM.

Code. The code is built on top of the *Open Bandit Pipeline* [61, https://github.com/st-tech/zr-obp] which is licensed under the Apache 2.0 License. In the linked code, the source files that have been modified are marked with an appropriate comment at the beginning.

B.1. Off-policy evaluation

B.1.1. Synthetic example

Experimental details. To accurately estimate the expectation of function f under P, we generate at the beginning 10M from P and we estimate the expectation μ with the sample mean. For all estimators with optimal parameter (truncation threshold or λ), we employ the significance value $\delta = 0.1$.

For the optimistic shrinkage transformation (OS), we compute the correction parameter τ^*, by minimizing an upper bound on the MSE, derived from the one presented in the paper [65], accounting for the fact that we do not have a reward estimate (we are not considering here a DR estimator):

$$\tau^* \in \arg\min_{\tau \geq 0} \underbrace{\widehat{\mathbb{V}\mathrm{ar}}_{y_i \sim Q}\left[\omega_\tau^{\mathrm{OS}}(y_i) f(y_i)\right]}_{\text{estimated variance}} + \underbrace{\frac{\|f\|_\infty^2}{n} \sum_{i \in [\![n]\!]} \left(\omega_\tau^{\mathrm{OS}}(y_i) - \omega(y_i)\right)^2}_{\text{estimated bias}}, \tag{58}$$

where:

$$\widehat{\mathbb{V}\mathrm{ar}}_{y_i \sim Q}\left[\omega_\tau^{\mathrm{OS}}(y_i) r(y_i)\right] = \frac{1}{n} \sum_{i \in [\![n]\!]} \left(\omega_\tau^{\mathrm{OS}}(y_i) f(y_i) - \widehat{\mu}_\tau^{\mathrm{OS}}\right)^2, \qquad \widehat{\mu}_\tau^{\mathrm{OS}} = \frac{1}{n} \sum_{i \in [\![n]\!]} \omega_\tau^{\mathrm{OS}}(y_i) f(y_i). \tag{59}$$

Table 7
Variance values σ_Q^2 and σ_P^2 and divergence $I_2(P\|Q)$ for the different experiments.

σ_Q^2	σ_P^2	$I_2(P\|Q)$
1	1.5	1.904
1	1.9	27.949
1	1.99	$5.104e+11$
1	1.999	$8.379e+109$

Table 8
Absolute error in the illustrative examples varying the number of samples n for the different estimators and the different settings of Table 7 (mean ± std, 60 runs). The estimator with the smallest absolute error and the ones not statistically significantly different from that one (Welch's t-test with $p < 0.02$) are in bold.

$\sigma_Q^2 = 1$, $\sigma_P^2 = 1.5$

Estimator / n	10	20	50	100	200	500	1000
IS	23.52 ± 5.39	**15.39 ± 3.26**	**10.06 ± 1.93**	8.35 ± 0.73	6.29 ± 0.32	3.93 ± 0.12	**2.54 ± 0.06**
SN	23.09 ± 4.62	**14.37 ± 2.55**	**9.15 ± 1.32**	8.23 ± 0.63	6.32 ± 0.31	3.96 ± 0.12	**2.56 ± 0.06**
TR	**20.34 ± 4.66**	**13.48 ± 2.59**	**8.33 ± 1.08**	**7.38 ± 0.47**	**5.88 ± 0.27**	**3.60 ± 0.11**	**2.45 ± 0.06**
OS	**16.55 ± 4.13**	**11.87 ± 2.79**	**7.98 ± 1.21**	**6.53 ± 0.52**	**5.06 ± 0.26**	**3.21 ± 0.10**	**2.17 ± 0.05**
PM-λ^*	**18.86 ± 4.01**	**12.20 ± 2.30**	**7.44 ± 0.92**	**6.53 ± 0.43**	**5.14 ± 0.25**	**3.25 ± 0.10**	**2.20 ± 0.05**
PM-$\widehat{\lambda}$	**17.98 ± 3.83**	**11.30 ± 2.07**	**6.82 ± 0.77**	**5.89 ± 0.40**	**4.72 ± 0.23**	**3.03 ± 0.09**	**2.09 ± 0.05**

$\sigma_Q^2 = 1$, $\sigma_P^2 = 1.9$

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	**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
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
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
PM-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
PM-$\widehat{\lambda}$	**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**

$\sigma_Q^2 = 1$, $\sigma_P^2 = 1.99$

Estimator / n	10	20	50	100	200	500	1000
IS	24.42 ± 6.54	**25.03 ± 11.38**	15.72 ± 3.31	11.10 ± 1.89	8.96 ± 0.74	6.23 ± 0.32	4.77 ± 0.19
SN	25.50 ± 5.84	20.36 ± 3.36	13.99 ± 1.56	9.58 ± 1.08	8.73 ± 0.56	6.08 ± 0.27	4.64 ± 0.16
TR	24.42 ± 6.54	**25.03 ± 11.38**	15.72 ± 3.31	11.10 ± 1.89	8.96 ± 0.74	6.23 ± 0.32	4.77 ± 0.19
OS	**16.39 ± 4.48**	**16.89 ± 6.36**	**11.20 ± 1.96**	**7.66 ± 1.08**	6.80 ± 0.48	4.67 ± 0.21	3.62 ± 0.14
PM-λ^*	24.42 ± 6.54	**25.03 ± 11.38**	15.72 ± 3.31	11.10 ± 1.89	8.96 ± 0.74	6.23 ± 0.32	4.77 ± 0.19
PM-$\widehat{\lambda}$	**16.12 ± 4.19**	**12.50 ± 2.04**	**7.81 ± 0.77**	**5.19 ± 0.41**	**4.64 ± 0.24**	**2.92 ± 0.11**	**2.25 ± 0.05**

$\sigma_Q^2 = 1$, $\sigma_P^2 = 1.999$

Estimator / n	10	20	50	100	200	500	1000
IS	**32.44 ± 30.89**	**22.29 ± 11.21**	19.03 ± 5.26	19.39 ± 4.36	15.83 ± 2.03	9.21 ± 0.50	6.96 ± 0.26
SN	21.06 ± 5.75	18.00 ± 3.18	14.78 ± 2.10	11.81 ± 1.39	10.66 ± 0.89	7.94 ± 0.35	6.32 ± 0.20
TR	**32.44 ± 30.89**	**22.29 ± 11.21**	19.03 ± 5.26	19.39 ± 4.36	15.83 ± 2.03	9.21 ± 0.50	6.96 ± 0.26
OS	**21.32 ± 18.62**	**15.42 ± 6.75**	**12.18 ± 3.00**	13.50 ± 3.06	10.62 ± 1.25	6.15 ± 0.33	4.68 ± 0.16
PM-λ^*	**32.44 ± 30.89**	**22.29 ± 11.21**	19.03 ± 5.26	19.39 ± 4.36	15.83 ± 2.03	9.21 ± 0.50	6.96 ± 0.26
PM-$\widehat{\lambda}$	**13.36 ± 3.47**	**11.25 ± 1.67**	**7.52 ± 0.85**	**5.27 ± 0.43**	**3.68 ± 0.20**	**2.47 ± 0.10**	2.20 ± 0.05

Complete results. In all experiments, we employ $\mu_P = 0.5$ and $\mu_Q = 0$. The values of σ_P and σ_Q for the different experiments are reported in Table 7. In Table 8 and Fig. 5, we report the complete results for the different settings.

B.1.2. Contextual bandits

Experimental setting. The experimental evaluation is carried out over 11 UCI Machine Learning Repository datasets [17, https://archive.ics.uci.edu/ml/index.php] as reported in Table 9. For the estimators requiring the value of the significance, we select $\delta = 0.1$.

Complete results. In the comprehensive experiment, we consider 110 combinations obtained with a single run over the 11 datasets and 10 values of the pair (α_b, α_e) with $\alpha_b \in \{0.8, 0.9\}$ and $\alpha_e \in \{0.8, 0.85, 0.9, 0.95, 0.99\}$. The experiment with reward noise $\nu = 0.25$ is reported in the main paper (Fig. 2), whereas the noiseless $\nu = 0$ (deterministic rewards) is provided in Fig. 6. The results are in line with the stochastic case.

For the case of the *letter* dataset, we report the experiments with additional choices of α_e (Tables 10 and 11).

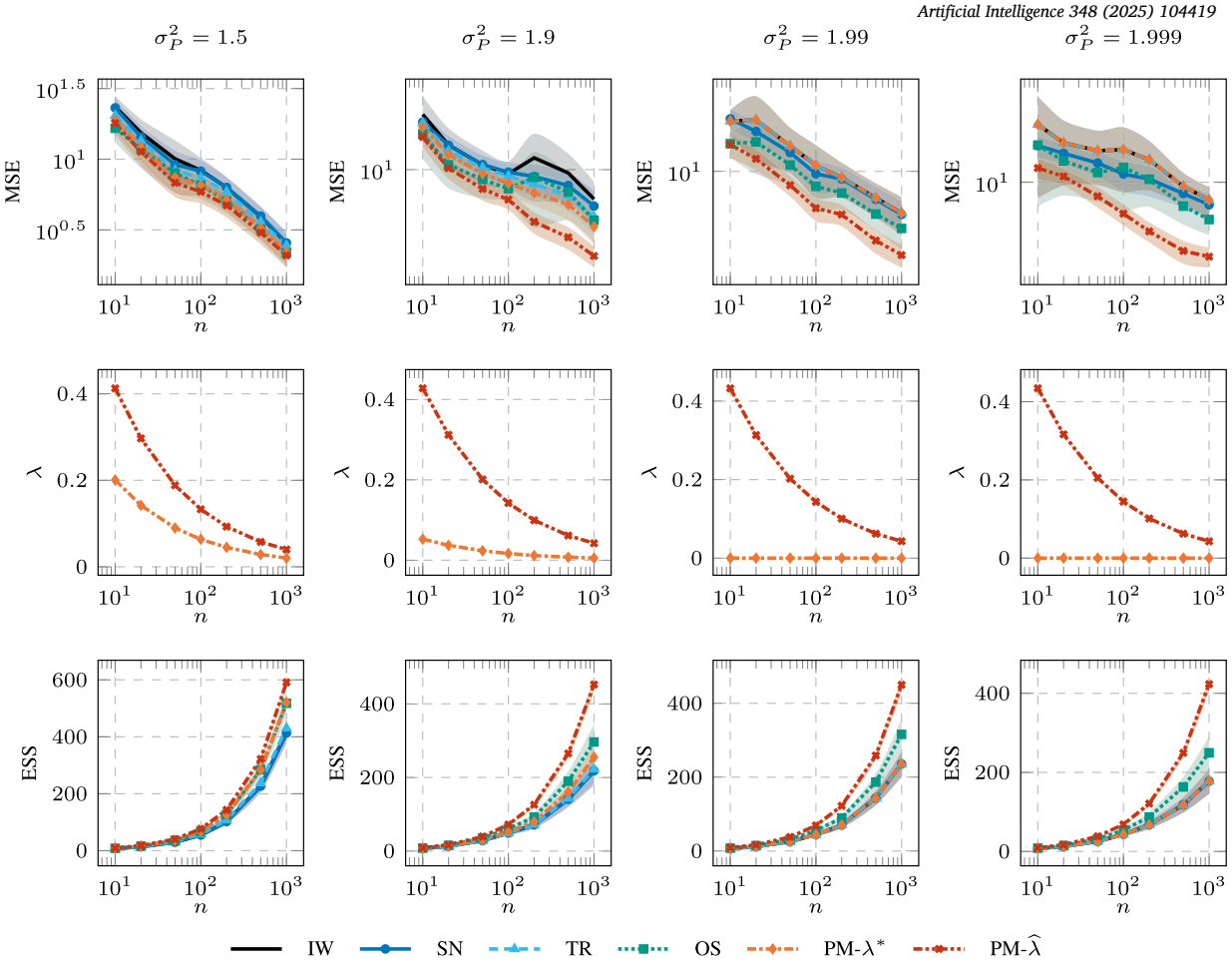

Fig. 5. Mean Squared Error (MSE), correction parameter λ, and Effective Sample Size (ESS), computed as $\frac{\left(\sum_{i\in[\![n]\!]}\omega(y_i)\right)^2}{\sum_{i\in[\![n]\!]}\omega(y_i)^2}$, as a function of the number of samples n for the different settings of Table 7 (mean $\pm$ 95% c.i., 60 runs).

Table 9
The 11 UCI dataset considered in the experiments. For each dataset, we report the number of examples n^*, dimensionality of the context, and number of classes K.

Dataset	ecoli	glass	isolet	kropt	letter	optdigits	page-blocks	pendigits	satimage	vehicle	yeast
Dataset size (n^*)	336	214	7797	28056	20000	5620	5473	10992	6435	846	1484
Context dimension	7	9	617	6	16	64	10	16	36	18	8
Classes (K)	8	6	26	18	26	10	5	10	6	4	10

B.2. Off-policy learning

Experimental setting. The optimization is performed by gradient ascent on the objective function:

$$\mathcal{L}(\boldsymbol{\theta}) = \widehat{v}(\pi_{\boldsymbol{\theta}}) - \frac{\zeta}{n}\sum_{i\in[\![n]\!]} I_2(\pi_{\boldsymbol{\theta}}(\cdot|x_i)\|\pi_b(\cdot|x_i)), \tag{60}$$

where $\widehat{v}(\pi_{\boldsymbol{\theta}})$ is the estimated value function using the different estimators, that is a function of the target policy $\pi_{\boldsymbol{\theta}}$. The second term is the empirical average of the divergence between the target $\pi_{\boldsymbol{\theta}}$ and the behavioral policy π_b. The regularizer is controlled by the regularization parameter $\zeta \geq 0$. The gradient optimization is performed in mini-batches made of 32 samples and the learning rate is selected with RMSprop, with 0.05 as base learning rate.

Complete results. In Fig. 7 and in Fig. 8 we report the complete results, in the setting presented in the main paper, for the non-regularized ($\zeta = 0$) and the regularized objective ($\zeta = 0.1$), respectively. The experiments with the regularized objective are limited

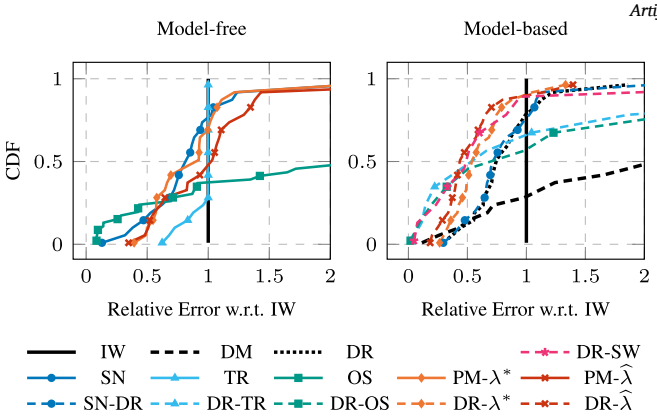

Fig. 6. CDF of the absolute error normalized by IW error for deterministic rewards, across 110 conditions for model-free estimators (left) and model-based ones (right).

Table 10
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 $\alpha_e = 0.99$ (mean ± std, 10 runs). For each column, the estimator with the smallest absolute error and the ones not statistically significantly different from that one (Welch's t-test with $p < 0.05$) are in bold.

Estimator / n	100	200	500	1000	2000	5000	10000	20000
IS	**20.04 ± 1.24**	**21.77 ± 2.46**	14.03 ± 0.57	8.40 ± 0.20	6.13 ± 0.09	2.77 ± 0.03	1.83 ± 0.01	1.10 ± 0.01
SN	27.34 ± 1.67	23.16 ± 1.40	16.86 ± 0.46	11.94 ± 0.25	7.37 ± 0.13	2.59 ± 0.03	1.74 ± 0.01	1.17 ± 0.01
TR	**20.04 ± 1.24**	**18.17 ± 1.60**	13.96 ± 0.57	8.40 ± 0.20	6.13 ± 0.09	2.77 ± 0.03	1.83 ± 0.01	1.10 ± 0.01
OS	24.47 ± 1.50	32.30 ± 1.17	15.37 ± 0.56	17.35 ± 0.46	16.46 ± 0.37	30.70 ± 0.15	34.03 ± 0.02	33.67 ± 0.01
PM-λ^*	**20.48 ± 1.33**	**16.77 ± 1.14**	10.06 ± 0.34	6.61 ± 0.16	5.30 ± 0.07	2.88 ± 0.03	2.08 ± 0.01	1.16 ± 0.01
PM-$\widehat{\lambda}$	**22.60 ± 1.52**	**17.06 ± 0.75**	10.22 ± 0.28	7.77 ± 0.16	5.61 ± 0.08	3.32 ± 0.03	2.50 ± 0.02	1.37 ± 0.01
DM	28.86 ± 1.92	27.56 ± 0.95	41.04 ± 0.26	41.94 ± 0.11	42.87 ± 0.05	47.06 ± 0.01	47.58 ± 0.01	47.51 ± 0.00
DR	**26.54 ± 4.51**	25.56 ± 2.43	16.69 ± 0.72	9.12 ± 0.20	5.62 ± 0.09	2.14 ± 0.02	**1.25 ± 0.01**	**0.83 ± 0.00**
SN-DR	**25.62 ± 3.21**	24.87 ± 1.79	18.94 ± 0.62	12.36 ± 0.23	7.19 ± 0.12	2.46 ± 0.02	**1.57 ± 0.01**	1.07 ± 0.01
DR-TR	**18.97 ± 1.12**	**16.54 ± 0.70**	20.95 ± 0.23	17.93 ± 0.09	17.90 ± 0.06	22.73 ± 0.01	23.45 ± 0.01	23.18 ± 0.00
DR-OS	**18.87 ± 1.18**	19.21 ± 0.55	17.15 ± 0.38	12.01 ± 0.23	8.67 ± 0.11	17.04 ± 0.06	17.88 ± 0.02	18.49 ± 0.02
DR-SW	23.97 ± 1.28	**16.66 ± 1.13**	**4.58 ± 0.18**	**4.64 ± 0.09**	4.76 ± 0.05	**0.75 ± 0.01**	**1.31 ± 0.01**	**0.77 ± 0.00**
DR-λ^*	**21.84 ± 2.30**	**18.16 ± 1.35**	11.26 ± 0.47	6.53 ± 0.14	**4.59 ± 0.07**	1.78 ± 0.02	**1.23 ± 0.01**	**0.72 ± 0.00**
DR-$\widehat{\lambda}$	**19.45 ± 1.62**	**14.35 ± 0.95**	7.89 ± 0.34	**4.88 ± 0.11**	**3.88 ± 0.06**	1.60 ± 0.02	**1.26 ± 0.01**	**0.68 ± 0.00**

Table 11
Absolute error (multiplied by 100) in the *letter* dataset varying the number of samples n for the different estimators, when $\alpha_b = 0.9$ and $\alpha_e = 0.99$ (mean ± std, 10 runs). For each column, the estimator with the smallest absolute error and the ones not statistically significantly different from that one (Welch's t-test with $p < 0.05$) are in bold.

Estimator / n	100	200	500	1000	2000	5000	10000	20000
IS	**10.08 ± 0.91**	**20.07 ± 5.66**	20.23 ± 1.60	13.52 ± 0.42	12.23 ± 0.24	6.49 ± 0.05	3.62 ± 0.03	2.74 ± 0.01
SN	15.85 ± 1.60	18.18 ± 1.71	26.34 ± 0.75	23.84 ± 0.35	12.91 ± 0.20	5.96 ± 0.05	4.15 ± 0.03	2.14 ± 0.01
TR	**10.08 ± 0.91**	**12.02 ± 1.66**	13.94 ± 0.65	11.34 ± 0.22	11.38 ± 0.20	6.40 ± 0.05	3.62 ± 0.03	2.74 ± 0.01
OS	26.61 ± 0.75	53.26 ± 5.63	38.57 ± 1.10	32.35 ± 0.09	30.73 ± 0.06	25.41 ± 0.02	24.48 ± 0.01	23.76 ± 0.00
PM-λ^*	**10.17 ± 0.91**	**12.03 ± 1.62**	13.11 ± 0.52	10.33 ± 0.13	8.89 ± 0.09	3.88 ± 0.03	2.74 ± 0.02	2.35 ± 0.01
PM-$\widehat{\lambda}$	**11.00 ± 0.93**	**9.73 ± 0.37**	9.72 ± 0.18	8.95 ± 0.09	7.82 ± 0.07	3.67 ± 0.03	2.98 ± 0.02	2.56 ± 0.01
DM	22.00 ± 1.92	**9.78 ± 0.47**	**7.27 ± 0.19**	**3.49 ± 0.08**	**2.83 ± 0.05**	9.16 ± 0.02	9.97 ± 0.01	9.89 ± 0.00
DR	**35.50 ± 15.08**	33.18 ± 7.27	30.82 ± 1.83	24.87 ± 0.82	14.16 ± 0.27	6.19 ± 0.06	3.46 ± 0.03	1.48 ± 0.01
SN-DR	19.47 ± 3.41	20.94 ± 2.62	24.69 ± 1.07	21.41 ± 0.43	13.66 ± 0.16	6.41 ± 0.06	3.57 ± 0.03	1.57 ± 0.01
DR-TR	12.26 ± 1.14	**10.90 ± 0.61**	**6.65 ± 0.23**	9.98 ± 0.09	10.29 ± 0.04	**2.34 ± 0.01**	**0.95 ± 0.00**	**0.60 ± 0.00**
DR-OS	12.57 ± 1.23	**8.59 ± 0.49**	**6.73 ± 0.27**	5.25 ± 0.13	9.05 ± 0.09	**2.60 ± 0.01**	1.97 ± 0.01	1.21 ± 0.00
DR-SW	12.46 ± 1.09	11.73 ± 0.64	**7.52 ± 0.25**	11.31 ± 0.10	11.59 ± 0.04	3.49 ± 0.01	2.09 ± 0.00	1.69 ± 0.00
DR-λ^*	18.78 ± 3.41	16.07 ± 2.09	15.26 ± 0.66	13.55 ± 0.31	8.96 ± 0.15	3.97 ± 0.04	2.44 ± 0.02	1.24 ± 0.01
DR-$\widehat{\lambda}$	14.58 ± 1.18	**10.32 ± 0.58**	8.42 ± 0.23	11.33 ± 0.19	11.02 ± 0.23	2.83 ± 0.02	2.00 ± 0.01	1.26 ± 0.00

to *glass* and *ecoli* datasets. We report the corresponding learning curves for the non-regularized (Fig. 9) and the regularized objectives (Fig. 10).

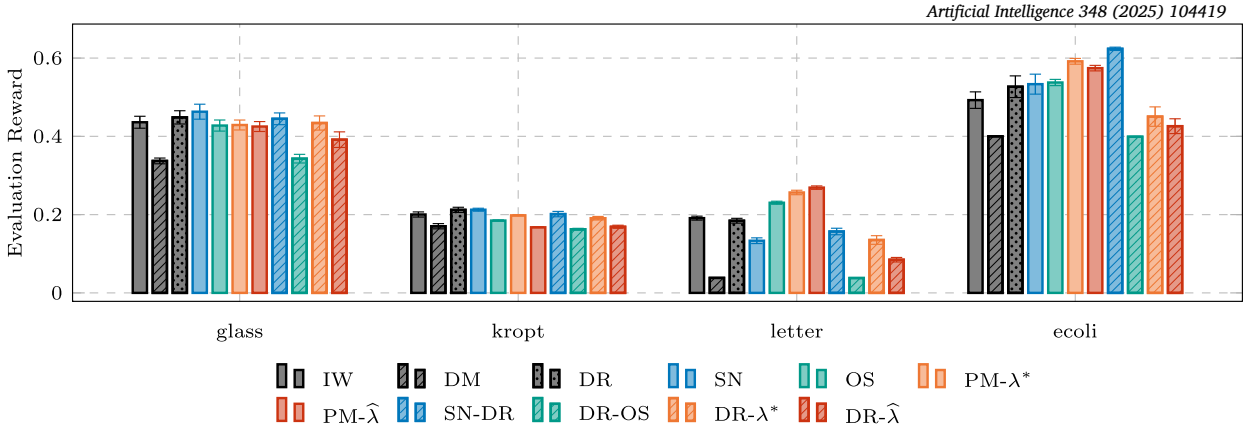

Fig. 7. Evaluation reward for four different datasets after 1000 iterations (4000 iterations for *letter*) of gradient ascent with a Boltzmann policy and the non-regularized objective ($\zeta = 0$) (mean $\pm$ std, 10 runs).

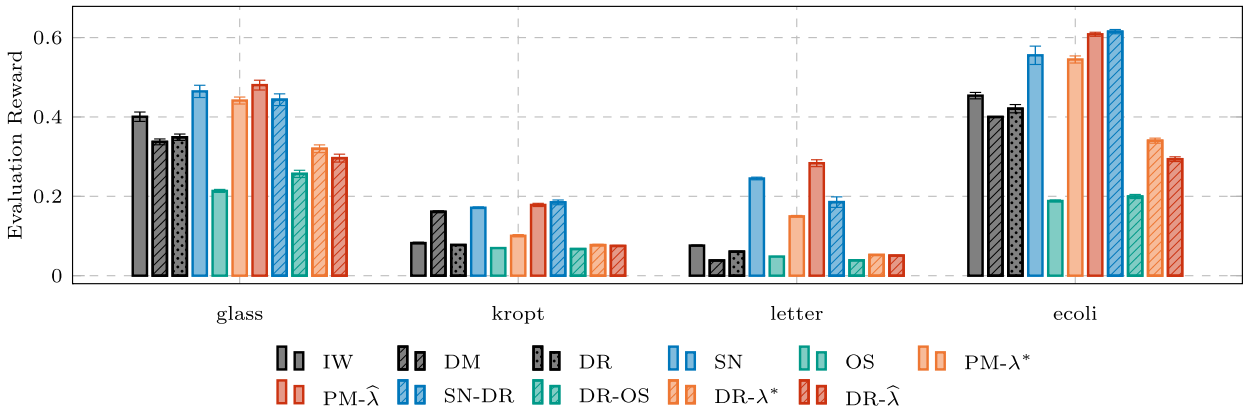

Fig. 8. Evaluation reward for four different datasets after 1000 iterations (4000 iterations for *letter*) of gradient ascent with a Boltzmann policy and the regularized objective ($\zeta = 0.1$) (mean $\pm$ std, 10 runs).

Appendix C. Choice of $(\lambda_\alpha, s_\alpha)$ values

From Theorem 5.1, we observed that the optimal $(\lambda_\alpha, s_\alpha)$ pair is defined by solving an optimization problem that is convex in $\lambda^{1/s}$. Hence, by fixing one value of $s \in [-\infty, -1]$, we can derive a corresponding optimal value of λ and vice-versa. Thus, we have a degree of freedom.

In this section, we show an approach that helps set a valid λ value and then find the corresponding optimal s parameter. This approach keeps the derived concentration guarantees unchanged. We remark that the bound provided in Lemma 5.4 is obtained from the results of Lemma 5.2 and 5.3. These results, in turn, are obtained through various relaxation steps (upper bounds), as shown in their proofs. Let us now define a tighter version of the bound provided in Lemma 5.4 holding under the same conditions. We get the following, holding with probability $1 - \delta$:

$$\widehat{\mu}_{n,\lambda,s} - \mu \leq \|f\|_\infty \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{n\lambda^{\frac{\alpha-2}{s}}}\Big((1-\lambda) + \lambda I_\alpha(P\|Q)\Big)} + \frac{2\|f\|_\infty \lambda^{\frac{1}{s}} \log\left(\frac{1}{\delta}\right)}{3n} +$$
$$+ \ \|f\|_\infty (3-\alpha)^{\frac{1}{\alpha}} \lambda^{\frac{1-\alpha}{s}} \Big((1-\lambda) I_\alpha(P\|Q) + \lambda\Big), \tag{61}$$

where this expression is simply obtained by avoiding the relaxation steps $(1-\lambda)I_\alpha(P\|Q) + \lambda \leq I_\alpha(P\|Q)$ and $(1-\lambda) + \lambda I_\alpha(P\|Q) \leq I_\alpha(P\|Q)$ employed in the proofs of Lemma 5.2 and Lemma 5.3, respectively. It turns out that this bound is no longer convex in $\lambda^{1/s}$ and hence cannot be analytically optimized. The approach we propose for the choice of the $(\lambda_\alpha, s_\alpha)$ pair is to plug the optimal value of $(\lambda_\alpha^*)^{\frac{1}{s_\alpha^*}}$ derived in Theorem 5.1 into the Equation (61) and then optimize with respect to λ. For simplicity, let us denote with $C_\alpha := (\lambda_\alpha^*)^{\frac{1}{s_\alpha^*}}$. We get the following:

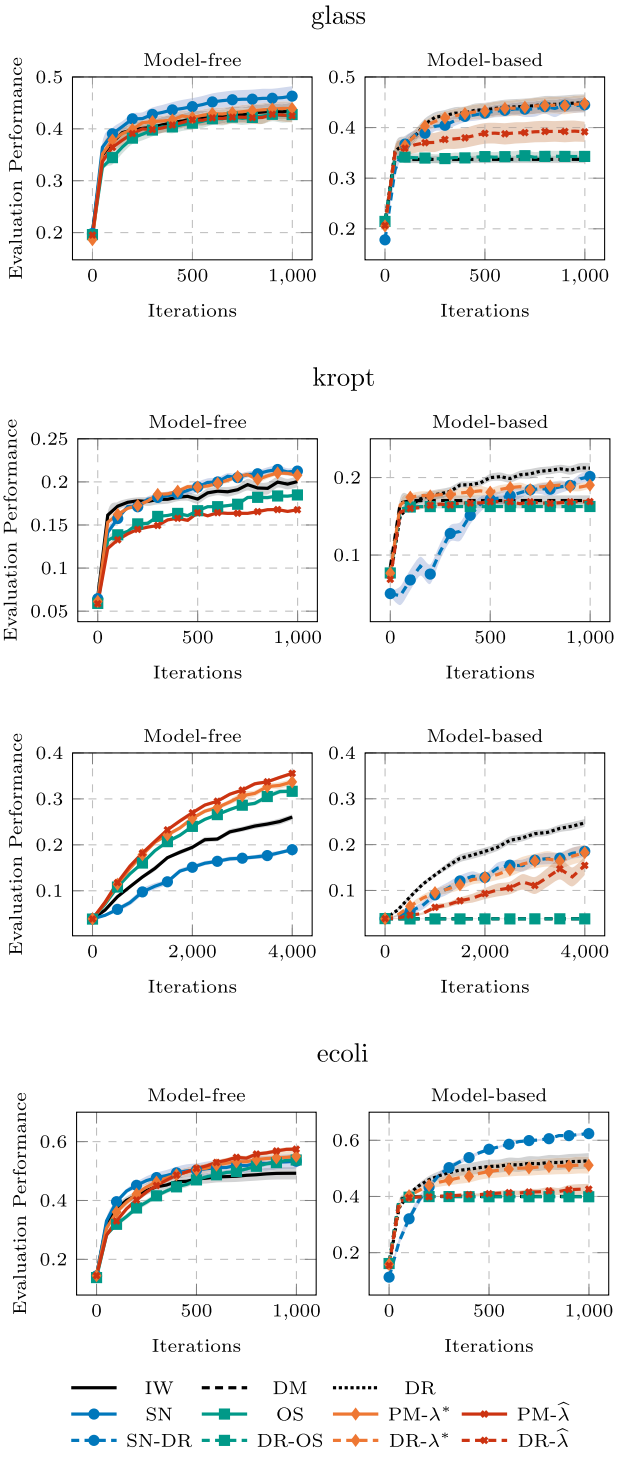

Fig. 9. Evaluation reward for the four datasets comparing the learning curve of different estimators with the non-regularized objective ($\zeta = 0$) (mean $\pm$ std, 10 runs).

$$\widehat{\mu}_{n,\lambda,s} - \mu \leq \|f\|_\infty \sqrt{\frac{2\log\left(\frac{1}{\delta}\right)}{nC_\alpha^{\alpha-2}}\Big((1-\lambda) + \lambda I_\alpha(P\|Q)\Big)} + \frac{2\|f\|_\infty C_\alpha \log\left(\frac{1}{\delta}\right)}{3n} +$$
$$+ \ \|f\|_\infty (3-\alpha)^{\frac{1}{\alpha}} C_\alpha^{1-\alpha}\Big((1-\lambda)I_\alpha(P\|Q) + \lambda\Big), \tag{62}$$

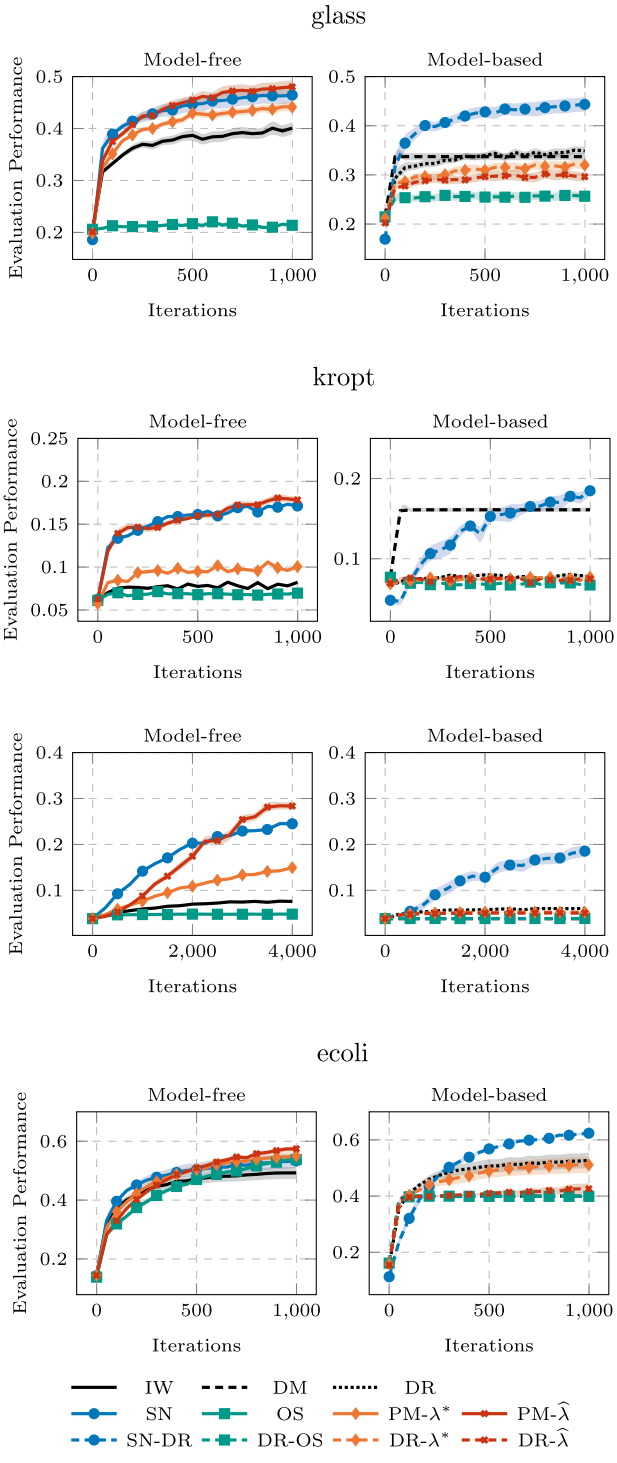

Fig. 10. Evaluation reward for the four datasets comparing the learning curve of different estimators with the regularized objective ($\zeta = 0.1$) (mean $\pm$ std, 10 runs).

Since we have freedom in the choice of $\lambda \in [0, 1]$, we can vanish the gradient of the relation above hence obtaining the optimal λ value, which can be expressed as:

$$\lambda_\alpha^* = \frac{C_\alpha^\alpha \log(\delta^{-1}) - 2n(3-\alpha)^{2/\alpha}}{2n(3-\alpha)^{2/\alpha}\left(I_\alpha(P\|Q) - 1\right)} = \frac{\frac{(\alpha-1)^2}{4(3-\alpha)^{2/\alpha}} I_\alpha(P\|Q) - 1}{I_\alpha(P\|Q) - 1}, \tag{63}$$

Table 12
Overview of the classical off-policy estimators for CMABs. π_b and π_e denote the behavioral and target policies respectively and $\widehat{r}$ the estimated reward function.

Estimator	Formula
Direct method (DM)	$\frac{1}{n}\sum_{i\in[\![n]\!]}\sum_{a\in\mathcal{A}}\pi_e(a\|x_i)\widehat{r}(x_i,a)$
Inverse propensity scoring (IPS)	$\frac{1}{n}\sum_{i\in[\![n]\!]}\frac{\pi_e(a_i\|x_i)}{\pi_b(a_i\|x_i)}r_i$
Doubly robust (DR)	$\frac{1}{n}\sum_{i\in[\![n]\!]}\sum_{a\in\mathcal{A}}\pi_e(a\|x_i)\widehat{r}(x_i,a)+\frac{1}{n}\sum_{i\in[\![n]\!]}\frac{\pi_e(a_i\|x_i)}{\pi_b(a_i\|x_i)}\left(r_i-\widehat{r}(x_i,a_i)\right)$

where the second equality is obtained by substituting the definition of C_α. We observe that the derived λ_α^* only depends on α and on the dissimilarity between the target and the behavioral distribution. Since $\frac{(\alpha-1)^2}{4(3-\alpha)^{2/\alpha}}<1$, we observe that the obtained value of λ_α^* defined in Equation (63) is valid whenever the numerator is positive, which holds when:

$$I_\alpha(P\|Q)\geq\frac{4(3-\alpha)^{2/\alpha}}{(\alpha-1)^2}.$$

Otherwise, we must take $\lambda_\alpha^*=0$. For the case with $\alpha=2$, we obtain:

$$\lambda_2^*=\frac{\frac{1}{4}I_\alpha(P\|Q)-1}{I_\alpha(P\|Q)-1},$$

which leads to valid values whenever $I_2(P\|Q)\geq 4$. From this, we can obtain a value for s_α^*:

$$s_\alpha^*=\frac{\log\lambda_\alpha^*}{\log C_\alpha}.$$

Appendix D. Comparison of estimators for CMABs

In Table 12, we report a comparison of the off-policy estimators for CMABs.

Appendix E. Analysis of the OS estimator

The OS (optimistic shrinkage) [65] is based on the weight transformation for $\tau\geq 0$:

$$\omega_\tau^{\mathrm{OS}}(y)=\frac{\tau\omega(y)}{\omega(y)^2+\tau}. \tag{64}$$

First of all, we notice that when $P=Q$ a.s. the weight becomes $\omega_\tau^{\mathrm{OS}}(y)=\frac{\tau}{\tau+1}$, so the estimator is biased. We start by observing that the corrected weight $\omega_\tau^{\mathrm{OS}}(y)$ converges to zero when the non-corrected weight is either zero or infinity. Thus, the maximum value of the weight must be in between. We compute it by vanishing the derivative:

$$\frac{\partial}{\partial\omega}\frac{\tau\omega}{\omega^2+\tau}=0\quad\Longrightarrow\quad\omega=\sqrt{\tau}. \tag{65}$$

From which, we obtain the maximum value of the weight equal to $\frac{\sqrt{\tau}}{2}$. We now focus on the following result concerning the bias and the variance of the OS estimator.[12]

Lemma E.1. *Let $P,Q\in\Delta^{\mathcal{Y}}$ be two probability distributions with $P\ll Q$. For every $\lambda\in[0,1]$, the bias and variance of the OS estimator are bounded as:*

$$\left|\mathop{\mathbb{E}}_{y_i\sim Q}\left[\widehat{\mu}_{n,\tau}^{OS}\right]-\mu\right|\leq\frac{\|f\|_\infty}{\sqrt{\tau}}I_2(P\|Q),\qquad\mathop{\mathbb{V}\mathrm{ar}}_{y_i\sim Q}\left[\widehat{\mu}_{n,\tau}^{OS}\right]\leq\frac{\|f\|_\infty^2}{n}I_2(P\|Q). \tag{66}$$

Proof. Let us start with the bias. We consider the following inequality:

$$\mathop{\mathbb{E}}_{y\sim Q}\left[\left|\omega_\tau^{\mathrm{OS}}(y)-\omega(y)\right|\right]=\mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)^3}{\omega(y)^2+\tau}\right]=\mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)^3}{\sqrt{\omega(y)^2+\tau}\sqrt{\omega(y)^2+\tau}}\right]\leq\mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)^3}{\omega(y)\sqrt{\tau}}\right]=\frac{I_2(P\|Q)}{\sqrt{\tau}}. \tag{P.209}$$

We consider now the variance term and derive a bound on the second moment of the OS weight:

[12] We make here a tighter analysis compared to the one presented in [48].

$$\mathop{\mathbb{E}}_{y\sim Q}\left[\omega_\tau^{\mathrm{OS}}(y)^2\right] = \mathop{\mathbb{E}}_{y\sim Q}\left[\left(\frac{\omega(y)\tau}{\omega(y)^2+\tau}\right)^2\right] \le \mathop{\mathbb{E}}_{y\sim Q}\left[\omega(y)^2\right] = I_2(P\|Q). \quad \square \tag{P.210}$$

We now move to the concentration result.

Theorem E.1. *Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability distributions with $P \ll Q$. For every $\delta \in (0,1)$, with an appropriate choice of τ, with probability at least $1-\delta$, it holds that:*

$$\widehat{\mu}_{n,\tau}^{OS} - \mu \le \|f\|_\infty \sqrt{\frac{2(5+2\sqrt{6})I_2(P\|Q)\log\frac{1}{\delta}}{3n}}. \tag{67}$$

Proof. We apply Bernstein's inequality to the estimator, starting for the bias and variance bounds of Lemma E.1:

$$\widehat{\mu}_{n,\tau}^{\mathrm{OS}} - \mu = \widehat{\mu}_{n,\tau}^{\mathrm{OS}} - \mathop{\mathbb{E}}_{y_i\sim Q}[\widehat{\mu}_{n,\tau}^{\mathrm{OS}}] + \mathop{\mathbb{E}}_{y_i\sim Q}[\widehat{\mu}_{n,\tau}^{\mathrm{OS}}] - \mu \tag{P.211}$$

$$\le \|f\|_\infty \sqrt{\frac{2I_2(P\|Q)\log\frac{1}{\delta}}{n}} + \frac{\|f\|_\infty\sqrt{\tau}\log\frac{1}{\delta}}{3n} + \frac{\|f\|_\infty}{\sqrt{\tau}} I_2(P\|Q). \tag{P.212}$$

We now minimize the bound as a function of $\sqrt{\tau}$ by vanishing the derivative to obtain:

$$\tau^* = \frac{3nI_2(P\|Q)}{\log\frac{1}{\delta}}. \tag{P.213}$$

By substituting τ^* we obtain the result. $\square$

Appendix F. Analysis of the IX estimator

The IX (implicit exploration) [52] is based on the weight transformation:

$$\omega_\gamma^{\mathrm{IX}}(y) = \frac{p(y)}{q(y)+\gamma}. \tag{68}$$

First of all, we notice that when $P = Q$ a.s. the weight becomes $\omega_\gamma^{\mathrm{IX}}(y) = \frac{1}{1+\gamma p(y)^{-1}}$, so the estimator is biased. We start by observing that the corrected weight $\omega_\gamma^{\mathrm{IX}}(y)$ takes maximum value depending on p, i.e., $\frac{1}{\gamma}\operatorname{ess\,sup}_{y\sim P} p(y)$. We now focus on the following result concerning the bias and the variance of the IX estimator.

Lemma F.1. *Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability distributions with $P \ll Q$. For every $\lambda \in [0,1]$, the bias and variance of the IX estimator are bounded as:*

$$\left|\mathop{\mathbb{E}}_{y\sim Q}\left[\widehat{\mu}_{n,\gamma}^{IX}\right] - \mu\right| \le \|f\|_\infty\sqrt{\gamma I_2(P\|Q)\mathrm{vol}(\mathcal{Y})}, \qquad \mathop{\mathbb{V}\mathrm{ar}}_{y_i\sim Q}\left[\widehat{\mu}_{n,\gamma}^{IX}\right] \le \frac{\|f\|_\infty^2}{n} I_2(P\|Q). \tag{69}$$

Proof. Let us start with the bias. We consider the following inequality:

$$\mathop{\mathbb{E}}_{y\sim Q}\left[\left|\omega_\gamma^{\mathrm{IX}}(y) - \omega(y)\right|\right] = \gamma \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{p(x)}{q(y)(q(y)+\gamma)}\right] \tag{P.214}$$

$$\le \gamma \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{p(y)^2}{q(y)^2}\right]^{1/2} \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{1}{(q(y)+\gamma)^2}\right]^{1/2} \tag{P.215}$$

$$\le \gamma \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{p(y)^2}{q(y)^2}\right]^{1/2} \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{1}{\gamma q(y)}\right]^{1/2} \tag{P.216}$$

$$\le \gamma\sqrt{I_2(P\|Q)}\frac{1}{\sqrt{\gamma}}\sqrt{\int_{\mathcal{Y}} \mathrm{d}y} \tag{P.217}$$

$$= \sqrt{\gamma I_2(P\|Q)\mathrm{vol}(\mathcal{Y})}, \tag{P.218}$$

having applied Cauchy-Schwarz's inequality. We consider now the variance term and derive a bound on the second moment of the IX weight:

$$\mathop{\mathbb{E}}_{y\sim Q}\left[\omega_\gamma^{\mathrm{IX}}(y)^2\right] = \mathop{\mathbb{E}}_{y\sim Q}\left[\left(\frac{p(y)}{q(y)+\gamma}\right)^2\right] \le \mathop{\mathbb{E}}_{y\sim Q}\left[\omega(y)^2\right] = I_2(P\|Q). \quad \square \tag{P.219}$$

We now move to the concentration result.

Theorem F.1. *Let $P, Q \in \Delta^{\mathcal{Y}}$ be two probability distributions with $P \ll Q$. For every $\delta \in (0,1)$, with an appropriate choice of γ, with probability at least $1-\delta$, it holds that:*

$$\widehat{\mu}^{\mathrm{IX}}_{n,\gamma} - \mu \leq \|f\|_\infty \sqrt{\frac{2 I_2(P\|Q) \log \frac{1}{\delta}}{n}} + \|f\|_\infty \sqrt[3]{\frac{9 I_2(P\|Q) \operatorname{ess\,sup}_{y\sim P} p(y) \mathrm{vol}(\mathcal{Y}) \log \frac{1}{\delta}}{4n}}. \tag{70}$$

Proof. We apply Bernstein's inequality to the estimator, starting with the bias and variance bounds of Lemma F.1:

$$\widehat{\mu}^{\mathrm{IX}}_{n,\gamma} - \mu = \widehat{\mu}^{\mathrm{IX}}_{n,\gamma} - \mathop{\mathbb{E}}_{y_i\sim Q}[\widehat{\mu}^{\mathrm{IX}}_{n,\gamma}] + \mathop{\mathbb{E}}_{y_i\sim Q}[\widehat{\mu}^{\mathrm{IX}}_{n,\gamma}] - \mu \tag{P.220}$$

$$\leq \|f\|_\infty \sqrt{\frac{2 I_2(P\|Q) \log \frac{1}{\delta}}{n}} + \frac{\|f\|_\infty \log \frac{1}{\delta}}{3\gamma n} \operatorname*{ess\,sup}_{y\sim P} p(y) + \|f\|_\infty \sqrt{\gamma I_2(P\|Q) \mathrm{vol}(\mathcal{Y})}. \tag{P.221}$$

We now minimize the bound as a function of γ by vanishing the derivative to obtain:

$$\gamma^* = \left(\frac{2 \left(\operatorname{ess\,sup}_{y\sim P} p(y)\right)^2 \left(\log \frac{1}{\delta}\right)^2}{3 I_2(P\|Q) \mathrm{vol}(\mathcal{Y}) n^2} \right)^{\frac{1}{3}}. \tag{P.222}$$

By substituting γ^* we obtain the result. □

Data availability

Data is publicly available.

References

[1] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, Anna Korba, Exponential smoothing for off-policy learning, in: International Conference on Machine Learning (ICML), vol. 202, PMLR, 2023, pp. 984–1017.
[2] Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Papini, Alberto Maria Metelli, Nicola Gatti, Online learning with off-policy feedback in adversarial mdps, in: International Joint Conference on Artificial Intelligence (IJCAI), 2024, pp. 3697–3705.
[3] Armin Behnamnia, Gholamali Aminian, Alireza Aghaei, Chengchun Shi, Vincent YF Tan, Hamid R. Rabiee, Batch learning via log-sum-exponential estimator from logged bandit feedback, in: ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists, 2024.
[4] Oliver Bembom, Mark J. van der Laan, Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators, Am. J. Epidemiol. (2008).
[5] Léon Bottou, Jonas Peters, Joaquin Quiñonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y. Simard, Ed Snelson, Counterfactual reasoning and learning systems: the example of computational advertising, J. Mach. Learn. Res. 14 (1) (2013) 3207–3260.
[6] Stéphane Boucheron, Gábor Lugosi, Pascal Massart, et al., On concentration of self-bounding functions, Electron. J. Probab. 14 (2009) 1884–1899.
[7] Jean Bretagnolle, Catherine Huber, Estimation des densités: risque minimax, Sémin. Probab. 12 (1978) 342–363.
[8] Sébastien Bubeck, Nicolò Cesa-Bianchi, Gábor Lugosi, Bandits with heavy tail, IEEE Trans. Inf. Theory 59 (11) (2013) 7711–7717.
[9] Peter S. Bullen, Handbook of Means and Their Inequalities, vol. 560, Springer Science & Business Media, 2013.
[10] Olivier Catoni, Challenging the empirical mean and empirical variance: a deviation study, Ann. Inst. Henri Poincaré Probab. Stat. 48 (2012) 1148–1185.
[11] Kamil Andrzej Ciosek, Shimon Whiteson, OFFER: off-environment reinforcement learning, in: AAAI Conference on Artificial Intelligence (AAAI), vol. 31, AAAI Press, 2017, pp. 1819–1825.
[12] William G. Cochran, Sampling Techniques, 3rd edition, John Wiley, ISBN 0-471-16240-X, 1977.
[13] Stephen R. Cole, Miguel A. Hernán, Constructing inverse probability weights for marginal structural models, Am. J. Epidemiol. 168 (6) (2008) 656–664.
[14] Corinna Cortes, Yishay Mansour, Mehryar Mohri, Learning bounds for importance weighting, Adv. Neural Inf. Process. Syst. 23 (2010) 442–450.
[15] Christoph Dann, Tor Lattimore, Emma Brunskill, Unifying pac and regret: uniform pac bounds for episodic reinforcement learning, Adv. Neural Inf. Process. Syst. 30 (2017) 5713–5723.
[16] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, Roberto I. Oliveira, et al., Sub-gaussian mean estimators, Ann. Stat. 44 (6) (2016) 2695–2725.
[17] Dheeru Dua, Casey Graff, UCI machine learning repository, http://archive.ics.uci.edu/ml, 2017.
[18] Miroslav Dudík, John Langford, Lihong Li, Doubly robust policy evaluation and learning, in: International Conference on Machine Learning (ICML), 2011, pp. 1097–1104.
[19] Filippo Fedeli, Alberto Maria Metelli, Francesco Trovò, Marcello Restelli, IWDA: importance weighting for drift adaptation in streaming supervised learning problems, IEEE Trans. Neural Netw. Learn. Syst. 34 (10) (2023) 6813–6823.
[20] Germano Gabbianelli, Gergely Neu, Matteo Papini, Importance-weighted offline learning done right, in: International Conference on Algorithmic Learning Theory (ALT), vol. 237, PMLR, 2024, pp. 614–634.
[21] Gianmarco Genalti, Lupo Marsigli, Nicola Gatti, Alberto Maria Metelli, (ϵ, u)-adaptive regret minimization in heavy-tailed bandits, in: Annual Conference on Learning Theory (COLT), vol. 247, PMLR, 2024, pp. 1882–1915.
[22] M. Gil, Fady Alajaji, Tamás Linder, Rényi divergence measures for commonly used univariate continuous distributions, Inf. Sci. 249 (2013) 124–131.
[23] Jinyong Hahn, On the role of the propensity score in efficient semiparametric estimation of average treatment effects, Econometrica (1998) 315–331.
[24] Josiah P. Hanna, Philip S. Thomas, Peter Stone, Scott Niekum, Data-efficient policy evaluation through behavior policy search, in: International Conference on Machine Learning (ICML), vol. 70, PMLR, 2017, pp. 1394–1403.
[25] Tim Hesterberg, Weighted average importance sampling and defensive mixture distributions, Technometrics 37 (2) (1995) 185–194.
[26] Timothy Classen Hesterberg, Advances in importance sampling, PhD thesis, Citeseer, 1988.

[27] Daniel G. Horvitz, Donovan J. Thompson, A generalization of sampling without replacement from a finite universe, J. Am. Stat. Assoc. 47 (260) (1952) 663–685.
[28] Peter J. Huber, Robust estimation of a location parameter, in: Breakthroughs in Statistics, Springer, 1992, pp. 492–518.
[29] Edward L. Ionides, Truncated importance sampling, J. Comput. Graph. Stat. 17 (2) (2008) 295–311.
[30] Mark Jerrum, Leslie G. Valiant, Vijay V. Vazirani, Random generation of combinatorial structures from a uniform distribution, Theor. Comput. Sci. 43 (1986) 169–188.
[31] Herman Kahn, Andy W. Marshall, Methods of reducing sample size in Monte Carlo computations, J. Oper. Res. Soc. Am. 1 (5) (1953) 263–278.
[32] Ozsel Kilinc, Giovanni Montana, Reinforcement learning for robotic manipulation using simulated locomotion demonstrations, Mach. Learn. 111 (2) (2022) 465–486.
[33] Jens Kober, Jan Peters, Learning Motor Skills - From Algorithms to Robot Experiments, Springer Tracts in Advanced Robotics, vol. 97, Springer, 2014.
[34] Ilja Kuzborskij, Csaba Szepesvári, Efron-Stein pac-bayesian inequalities, CoRR, arXiv:1909.01931, 2019.
[35] Ilja Kuzborskij, Claire Vernade, András György, Csaba Szepesvári, Confident off-policy evaluation and selection through self-normalized importance weighting, in: International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 130, PMLR, 2021, pp. 640–648.
[36] John Langford, Tong Zhang, The epoch-greedy algorithm for multi-armed bandits with side information, Adv. Neural Inf. Process. Syst. 20 (2007) 817–824.
[37] Brian K. Lee, Justin Lessler, Elizabeth A. Stuart, Weight trimming and propensity score weighting, PLoS ONE 6 (3) (2011) e18174.
[38] Oleg V. Lepski, Vladimir G. Spokoiny, Optimal pointwise adaptive methods in nonparametric estimation, Ann. Stat. (1997) 2512–2546.
[39] Lihong Li, Rémi Munos, Csaba Szepesvári, Toward minimax off-policy value estimation, in: International Conference on Artificial Intelligence and Statistics (AISTATS), vol. 38, 2015, pp. 608–616.
[40] Pierre Liotet, Francesco Vidaich, Alberto Maria Metelli, Marcello Restelli, Lifelong hyper-policy optimization with multiple importance sampling regularization, in: AAAI Conference on Artificial Intelligence (AAAI), vol. 36, AAAI Press, 2022, pp. 7525–7533.
[41] Ben London, Ted Sandler, Bayesian counterfactual risk minimization, in: International Conference on Machine Learning (ICML), vol. 97, PMLR, 2019, pp. 4125–4133.
[42] Shiyin Lu, Guanghui Wang, Yao Hu, Lijun Zhang, Optimal algorithms for Lipschitz bandits with heavy-tailed rewards, in: International Conference on Machine Learning (ICML), vol. 97, PMLR, 2019, pp. 4154–4163.
[43] Lugosi Gabor, Shahar Mendelson, Mean estimation and regression under heavy-tailed distributions: a survey, Found. Comput. Math. 19 (5) (2019) 1145–1190.
[44] Ashique Rupam Mahmood, Hado van Hasselt, Richard S. Sutton, Weighted importance sampling for off-policy learning with linear function approximation, Adv. Neural Inf. Process. Syst. 27 (2014) 3014–3022.
[45] Alberto Maria Metelli, Matteo Papini, Francesco Faccio, Marcello Restelli, Policy optimization via importance sampling, Adv. Neural Inf. Process. Syst. 31 (2018) 5447–5459.
[46] Alberto Maria Metelli, Matteo Papini, Nico Montali, Marcello Restelli, Importance sampling techniques for policy optimization, J. Mach. Learn. Res. 21 (2020) 141.
[47] Alberto Maria Metelli, Matteo Papini, Pierluca D'Oro, Marcello Restelli, Policy optimization as online learning with mediator feedback, in: AAAI Conference on Artificial Intelligence (AAAI), vol. 35, AAAI Press, 2021, pp. 8958–8966.
[48] Alberto Maria Metelli, Alessio Russo, Marcello Restelli, Subgaussian and differentiable importance sampling for off-policy evaluation and learning, Adv. Neural Inf. Process. Syst. 34 (2021) 8119–8132.
[49] Alberto Maria Metelli, Samuele Meta, Marcello Restelli, On the relation between policy improvement and off-policy minimum-variance policy evaluation, in: Uncertainty in Artificial Intelligence (UAI), vol. 216, PMLR, 2023, pp. 1423–1433.
[50] John E. Moody, Matthew Saffell, Learning to trade via direct reinforcement, IEEE Trans. Neural Netw. 12 (4) (2001) 875–889.
[51] Arkadij Semenovič Nemirovskij, David Borisovich Yudin, Problem Complexity and Method Efficiency in Optimization, 1983.
[52] Gergely Neu, Explore no more: improved high-probability regret bounds for non-stochastic bandits, Adv. Neural Inf. Process. Syst. 28 (2015) 3168–3176.
[53] Art B. Owen, Monte Carlo Theory, Methods and Examples, 2013.
[54] Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, Marcello Restelli, Optimistic policy optimization via multiple importance sampling, in: International Conference on Machine Learning (ICML), vol. 97, PMLR, 2019, pp. 4989–4999.
[55] Matteo Papini, Giorgio Manganini, Alberto Maria Metelli, Marcello Restelli, Policy gradient with active importance sampling, Reinf. Learn. J. 2 (2024) 645–675.
[56] James Pickands III, et al., Statistical inference using extreme order statistics, Ann. Stat. 3 (1) (1975) 119–131.
[57] Iosif Pinelis, et al., Best possible bounds of the von Bahr–Esseen type, Ann. Funct. Anal. 6 (4) (2015) 1–29.
[58] Tiberiu Popoviciu, Sur les équations algébriques ayant toutes leurs racines réelles, Mathematica 9 (1935) 129–145.
[59] Alfréd Rényi, On measures of entropy and information, Technical report, Hungarian Academy of Sciences, Budapest, Hungary, 1961.
[60] Brian D. Ripley, Stochastic Simulation, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, Wiley, 1987.
[61] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, Yusuke Narita, Open bandit dataset and pipeline: towards realistic and reproducible off-policy evaluation, in: NeurIPS Datasets and Benchmarks, 2021.
[62] Otmane Sakhi, Imad Aouali, Pierre Alquier, Nicolas Chopin, Logarithmic smoothing for pessimistic off-policy evaluation, selection and learning, Adv. Neural Inf. Process. Syst. 37 (2024) 80706–80755.
[63] David Siegmund, Importance sampling in the Monte Carlo study of sequential tests, Ann. Stat. (1976) 673–684.
[64] Yi Su, Lequn Wang, Michele Santacatterina, Thorsten Joachims, CAB: continuous adaptive blending for policy evaluation and learning, in: International Conference on Machine Learning (ICML), vol. 97, PMLR, 2019, pp. 6005–6014.
[65] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, Miroslav Dudík, Doubly robust off-policy evaluation with shrinkage, in: International Conference on Machine Learning (ICML), vol. 119, PMLR, 2020, pp. 9167–9176.
[66] Richard S. Sutton, David McAllester, Satinder Singh, Yishay Mansour, Policy gradient methods for reinforcement learning with function approximation, Adv. Neural Inf. Process. Syst. 12 (1999).
[67] Adith Swaminathan, Thorsten Joachims, The self-normalized estimator for counterfactual learning, Adv. Neural Inf. Process. Syst. 28 (2015) 3231–3239.
[68] Liang Tang, Rómer Rosales, Ajit Singh, Deepak Agarwal, Automatic ad format selection via contextual bandits, in: ACM International Conference on Information and Knowledge Management (CIKM'13), ACM, 2013, pp. 1587–1594.
[69] Philip S. Thomas, Georgios Theocharous, Mohammad Ghavamzadeh, High-confidence off-policy evaluation, in: AAAI Conference on Artificial Intelligence (AAAI), vol. 29, AAAI Press, 2015, pp. 3000–3006.
[70] Constantino Tsallis, Possible generalization of Boltzmann-Gibbs statistics, J. Stat. Phys. 52 (1–2) (1988) 479–487.
[71] John W. Tukey, Donald H. McLaughlin, Less vulnerable confidence and significance procedures for location based on a single sample: trimming/winsorization 1, Sankhya, Ser. A (1963) 331–352.
[72] Tim Van Erven, Peter Harremos, Rényi divergence and Kullback-Leibler divergence, IEEE Trans. Inf. Theory 60 (7) (2014) 3797–3820.
[73] Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, Jonah Gabry, Pareto smoothed importance sampling, J. Mach. Learn. Res. 25 (2024) 72.
[74] Yu-Xiang Wang, Alekh Agarwal, Miroslav Dudík, Optimal and adaptive off-policy evaluation in contextual bandits, in: International Conference on Machine Learning (ICML), vol. 70, PMLR, 2017, pp. 3589–3597.
[75] Changhe Yuan, Marek J. Druzdzel, How heavy should the tails be?, in: International Florida Artificial Intelligence Research Society Conference (FLAIRS), AAAI Press, 2005, pp. 799–805.
[76] Chao Zheng, A new principle for tuning-free Huber regression, Stat. Sin. (2020).
[77] Xin Zhou, Nicole Mayer-Hamblett, Umer Khan, Michael R. Kosorok, Residual weighted learning for estimating individualized treatment rules, J. Am. Stat. Assoc. 112 (517) (2017) 169–187.