Subgaussian Importance Sampling for Off-Policy Evaluation and Learning

Alberto Maria Metelli [1] **Alessio Russo** [1] **Marcello Restelli** [1]

Abstract

Importance Sampling (IS) is a widely used building block for a large variety of off-policy estimation and learning algorithms. However, empirical and theoretical studies have progressively shown that vanilla IS leads to poor estimations whenever the behavioral and target policies are too dissimilar. In this paper, we analyze the theoretical properties of the IS estimator by deriving a probabilistic deviation lower bound that formalizes the intuition behind its undesired behavior. Then, we propose a class of IS transformations, based on the notion of power mean, that are able, under certain circumstances, to achieve a subgaussian concentration rate. Differently from existing methods, like weight truncation, our estimator preserves the differentiability in the target distribution.

1. Introduction

The availability of historically collected data is a common scenario in many real-world decision-making problems, including medical treatments (e.g., Hahn, 1998; Zhou et al., 2017), recommendation systems (e.g., Li et al., 2011; Gilotte et al., 2018), personalized advertising (e.g., Bottou et al., 2013; Tang et al., 2013), finance (e.g., Moody & Saffell, 2001), and industrial robot control (e.g., Kober & Peters, 2014; Kilinc et al., 2019). Historical data can be leveraged to address two classes of problems. First, given data collected with a *behavioral* policy, we want to estimate the performance of a different *target* policy. This problem is known as *off-policy evaluation* (Off-PE, Horvitz & Thompson, 1952). Second, we want to employ the available data to improve the performance of a baseline policy. This second problem is named *off-policy learning* (Off-PL Dudík et al., 2011). Off-policy methods are studied by both the *reinforcement learning* (RL, Sutton & Barto, 2018) and *contextual multi-armed bandit* (CMAB, Langford & Zhang, 2007) communities. Given its intrinsic simplicity compared to RL, off-policy methods are nowadays well understood in the CMAB framework (e.g., Murphy et al., 2001; Bang & Robins, 2005; Dudík et al., 2011; Wang et al., 2017). Among them, the *doubly robust* estimator (DR, Dudík et al., 2011) is one of the most promising off-policy methods for CMABs. DR combines a *direct method* (DM), in which the reward is estimated from historical data via regression, with an *importance sampling* (IS, Owen, 2013) control variate.

More generally, IS plays a crucial role in the off-policy methods and counterfactual reasoning. It is established that IS tends to exhibit problematic behavior for general distributions. This is formalized by its *heavy-tailed* properties (Metelli et al., 2018), which prevents the application of exponential concentration bounds (Boucheron et al., 2003). To cope with this issue, typically, corrections are performed on the importance weight including *truncation* (Ionides, 2008) and *self-normalization* (Owen, 2013) among the most popular. Significant results have recently been derived for both techniques (Papini et al., 2019; Kuzborskij et al., 2020). Nevertheless, we believe that the widespread use of IS calls for a better theoretical understanding of its properties and for the design of general principled weight corrections.

Defining the desirable properties of an off-policy estimator is a non-trivial task. Some works employed the *mean squared error* (MSE) as an index of the estimator quality (Li et al., 2015; Wang et al., 2017). However, controlling the MSE, while effectively capturing the bias-variance trade-off, does not provide any guarantee on the concentration properties of the estimator, which might still display a heavy-tailed behavior (Lugosi & Mendelson, 2019). For this reason, we believe that a more suitable approach is to require that the estimator deviations concentrate at a *subgaussian* rate (Devroye et al., 2016). Subgaussianity implicitly controls the tail behavior and leads to exponential concentration inequalities. Unlike MSE, the probabilistic requirements are non-asymptotic (finite-sample), from which guarantees on the MSE can be derived. While subgaussianity can be considered a satisfactory requirement for Off-PE, additional properties are advisable when switching to Off-PL. In particular, the *differentiability* w.r.t. the target policy parameters is desirable whenever Off-PL is carried out via gradient optimization. For instance, weight truncation, as presented in (Papini et al., 2019; Metelli et al., 2021), allows achieving subgaussianity but leads to a non-differentiable objective.

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy. Correspondence to: Alberto Maria Metelli <albertomaria.metelli@polimi.it>.

Workshop on Reinforcement Learning Theory @ ICML 2021.

Consequently, the optimization phase requires additional care, which sometimes leads to computationally heavy discretizations (Papini et al., 2019). Thus, while truncation remains a powerful theoretical tool, it struggles when trying to scale to more realistic scenarios, including learning.

In this paper, we take a step towards a better understanding of IS. We start by deriving a probabilistic deviation lower bound for the mean estimation with vanilla IS. We show that polynomial concentration (Chebychev's inequality) is tight in this setting (Section 3). This result formalizes the intuition behind the undesired behavior of these estimators for general distributions. Hence, we propose a class of importance weight corrections, based on the notion of power mean (Section 4). The rationale behind these corrections is to "shrink" the weights towards the mean, with different intensities. In this way, we mitigate the heavy-tailed behavior and, in the meantime, we exert control over the induced bias. Then, we derive bounds on the bias and variance that allow obtaining an exponential concentration inequality (Section 5). To the best of our knowledge, this is the first IS correction that preserves the *differentiablity* in the target policy and is proved to achieve a *subgaussian* concentration rate. The proofs of all the results presented in the main paper can be found in Appendix C.

2. Preliminaries

We denote with $\mathscr{P}(\mathcal{Y})$ the set of probability measures over a measurable space $(\mathcal{Y}, \mathfrak{F}_{\mathcal{Y}})$. Let $P \in \mathscr{P}(\mathcal{Y})$, whenever needed, we assume that P admits a probability density function w.r.t. a reference measure, denoted with p. Let $P, Q \in \mathscr{P}(\mathcal{Y})$ if $P \ll Q$, i.e., P is absolutely continuous w.r.t. Q, for any $\alpha \in (1, 2]$, we introduce the integral:

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

Note that if $P = Q$ a.s. (almost surely) then $I_\alpha(P\|Q) = 1$. $I_\alpha(P\|Q)$ is the basic block of several distributional divergences. For instance, the Rényi divergence (Rényi, 1961) can be expressed as $(\alpha-1)^{-1} \log I_\alpha(P\|Q)$ and the Tsallis divergence (Tsallis, 1988) as $(\alpha-1)^{-1}(I_\alpha(P\|Q) - 1)$.

Let $P, Q \in \mathscr{P}(\mathcal{Y})$ and let $f : \mathcal{Y} \to \mathbb{R}$ be a measurable function. If $P \ll Q$, *(vanilla) importance sampling* (IS, Owen, 2013) allows estimating the expectation of f under the *target* distribution P, i.e., $\mu = \mathbb{E}_{y\sim P}[f(y)]$, using i.i.d. samples $\{y_i\}_{i\in[n]}$ collected with the *behavioral* distribution Q:

$$\widehat{\mu} = \frac{1}{n} \sum_{i\in[n]} \omega(y_i) f(y_i),$$

where $\omega(y) = p(y)/q(y)$ is the importance weight. It is well-known that $\widehat{\mu}$ is unbiased, i.e., $\mathbb{E}_{y_i \overset{\text{iid}}{\sim} Q}[\widehat{\mu}] = \mu$ (Owen, 2013). If f is bounded, the variance of the estimator can be upper-bounded as $\mathbb{V}\mathrm{ar}_{y_i \overset{\text{iid}}{\sim} Q}[\widehat{\mu}] \leqslant \frac{1}{n}\|f\|_\infty^2 I_2(P\|Q)$ (Metelli et al., 2018). Notice that the integral $I_\alpha(P\|Q)$ in Equation (1) represents the α-moment of the importance weight under Q. A common approach to mitigate the variance of IS is to resort to *self-normalization* (SN-IS, Owen, 2013):

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

The SN-IS estimator $\widetilde{\mu}$ has the desirable property of being bounded by $\|f\|_\infty$. However, it is no longer unbiased, while preserving consistency (Owen, 2013).

3. Probabilistic Limits of Vanilla Importance Sampling

In this section, we analyze the intrinsic limitations of the vanilla IS by deriving a probabilistic lower bound of the deviation of the estimator $\widehat{\mu}$ from the true mean μ. We start by introducing the result, then, we discuss its implications and compare it with previous work.

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

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

We note the polynomial dependence on the confidence level δ, typical of Chebyshev's inequalities (Boucheron et al., 2003). The bound is vacuous when $I_\alpha(P\|Q) = 1$, i.e., when $P = Q$ a.s.. Indeed, in this case, we are in an on-policy setting and, since the function f is bounded, exponential concentration bounds (like Höeffding's inequality) hold. In particular, for $\alpha = 2$, n and $I_2(P\|Q)$ sufficiently large, the bound has order $\mathcal{O}\left(\sqrt{\frac{I_2(P\|Q)}{\delta n}}\right)$. This form matches the deviation upper bound previously presented in (Metelli et al., 2018; 2020), proving that Chebyshev's inequality is actually tight for vanilla IS.[1]

Our result is of independent interest and applies for general distributions. Previous works considered the MAB (Li et al., 2015) and CMAB (Wang et al., 2017) settings proving deviation *minimax* lower bounds in *mean squared error* (MSE) $\mathbb{E}_{y \overset{\text{iid}}{\sim} Q}[(\widehat{\mu} - \mu)^2]$. These results differ from ours for three aspects. First, they focus on minimax optimality, while we derive an anticoncentration bound for vanilla IS. Second, they provide lower bounds to the MSE, while we focus on the deviations in probability. From our probabilistic result,

[1] In (Metelli et al., 2018) the provided bound was based on Cantelli's inequality which approaches Chebyshev's when $\delta \to 0$.

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

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

it is immediate to derive an MSE guarantee (Corollary C.1 of Appendix C.1). Third, they assume that the second moment of the importance weight $I_2(P\|Q)$ is finite, whereas our result allows us to consider scenarios in which only moments of order $\alpha<2$ are finite.

4. Power-Mean Correction of Importance Sampling

In this section, motivated by the negative result of Theorem 3.1, we look for a transformation of the importance weights able to achieve exponential concentration. Specifically, we introduce a class of corrections based on the notion of *power mean* (Bullen, 2013) and we study its properties. Let us start with the following definition.

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

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

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.[2] We immediately notice that, regardless of the value of s, for $\lambda=0$, we reduce to the vanilla importance weight $\omega_{0,s}(y)=\omega(y)$ and for $\lambda=1$, we have identically $\omega_{1,s}(y)=1$. Furthermore, the correction is unbiased when $P=Q$ a.s. regardless of the values of s and λ. Thus, the rationale behind the correction is to *interpolate* between the vanilla importance weight $\omega(y)$ and its mean under Q, i.e., 1. Some specific choices of s are reported in Table 1 and some examples are shown in Figure 1. We note that the intensity of the correction increases as λ moves towards 1 and s moves away from 1.

The following result provides a preliminary characterization of the correction, which is independent of the properties of the two distributions P and Q.

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

(i) if $s\leqslant s'$ then $\omega_{\lambda,s}(y)\leqslant\omega_{\lambda,s'}(y)$;
(ii) if $s<0$ then $\omega_{\lambda,s}(y)\leqslant\lambda^{\frac{1}{s}}$, otherwise if $s>0$ then $\omega_{\lambda,s}(y)\geqslant\lambda^{\frac{1}{s}}$;
(iii) if $s<1$ then $\mathbb{E}_{y\sim Q}[\omega_{\lambda,s}(y)]\leqslant 1$, otherwise if $s>1$ then $\mathbb{E}_{y\sim Q}[\omega_{\lambda,s}(y)]\geqslant 1$.

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

Thus, 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 importance sampling derives from the heavy-tailed properties (Metelli et al., 2018). 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 $s>1$ will increase the value of the weight, making the tail even heavier. So, if we are looking for subgaussianity, we should restrict our attention to $s<0$, which leads to lighter tails or even bounded weights.

5. Subgaussian Importance Sampling

In this section, we focus on the *harmonic* correction ($s=-1$), which leads to a weight of the form:[3]

$$\omega_{\lambda,-1}(y)=\frac{\omega(y)}{1-\lambda+\lambda\omega(y)}.$$

We analyze the bias and variance (Section 5.1) of this class of estimators and, finally, we provide an exponential and, under certain circumstances, subgaussian concentration inequality (Section 5.2). To lighten the notation we neglect the -1 subscript, abbreviating $\mu_\lambda=\mu_{\lambda,-1}$.

5.1. Bias and Variance

We derive bounds for the bias and the variance induced by the $(\lambda,-1)$-corrected importance weight. We start with the following result concerning the bias.

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

$$\left|\mathop{\mathbb{E}}_{y\overset{iid}{\sim}Q}[\widehat{\mu}_\lambda]-\mu\right|\leqslant\|f\|_\infty\lambda^{\alpha-1}I_\alpha(P\|Q).$$

As expected, the bias is zero for $\lambda=0$ and increases with λ. Furthermore, the bias increases with the divergence term $I_\alpha(P\|Q)$. Indeed, we already observed that the bias is null when $P=Q$ a.s.. In particular, for $\alpha=2$, the bound becomes $\|f\|_\infty\lambda I_2(P\|Q)$. Let us now turn to the variance bound.

Lemma 5.2. *Let $P,Q\in\mathscr{P}(\mathcal{Y})$ two probability distributions with $P\ll Q$. For every $\lambda\in[0,1]$, the $(\lambda,-1)$-corrected importance weight induces a variance that can be bounded for*

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

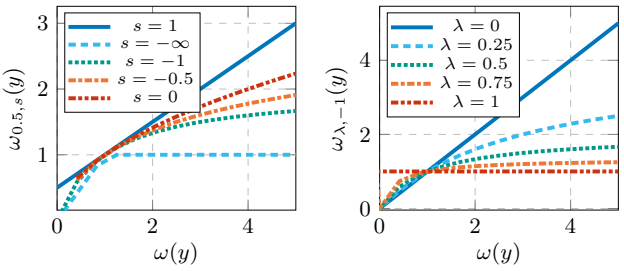

Figure 1. Examples of importance weight corrections of Definition 4.1 for fixed λ (left) and fixed s (right). Note that $s=1$ and $\lambda=0$ lead to no correction.

every $\alpha\in(1,2]$ *as:*

$$\underset{y_i\overset{iid}{\sim}Q}{\mathbb{V}\mathrm{ar}}\left[\widehat{\mu}_\lambda\right]\leqslant\frac{\|f\|_\infty^2}{n\lambda^{2-\alpha}}I_\alpha(P\|Q).$$

The variance bound decreases in λ and increases with the distributional divergence $I_\alpha(P\|Q)$. For $\alpha=2$, we obtain the bound $\frac{1}{n}\|f\|_\infty^2 I_2(P\|Q)$. Note that when $P=Q$ a.s., we recover $\frac{1}{n}\|f\|_\infty^2$, which is the Popoviciu's inequality for the variance (Popoviciu, 1935). These results show that the proposed weight correction allows controlling bias and variance even when $I_2(P\|Q)=\infty$, i.e., when the vanilla IS estimator might have infinite variance. Indeed, our transformed estimator has finite variance provided that there exists $\alpha\in(1,2)$ so that $I_\alpha(P\|Q)<\infty$. Tighter (but less intelligible) bounds on bias and variance are reported in Appendix C.3.

5.2. Concentration Inequality

We now use the results presented in the previous section to derive an exponential concentration inequality for the corrected IS estimator and to show that if $I_2(P\|Q)$ is finite we also achieve a subgaussian concentration rate.

Theorem 5.1. *Let* $P,Q\in\mathscr{P}(\mathcal{Y})$ *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* $\delta\in(0,1)$, *let*

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

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

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

Let us notice that the concentration inequality has an exponential dependence on the confidence parameter δ, for every $\alpha\in(1,2]$. However, we observe that the bound is subgaussian only 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 (Devroye et al., 2016). Specifically, for $\alpha=2$ the optimal value of the parameter is $\lambda_2^*=\sqrt{\frac{2\log\frac{1}{\delta}}{3I_2(P\|Q)n}}$ and we obtain the bound:

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

Note that the constant we obtain is $(2+\sqrt{3})\sqrt{2/3}\simeq 3.047$, while the optimal constant for subgaussian estimators is known to be $\sqrt{2}$ (Devroye et al., 2016). A tighter bound is derived in Lemma C.3 of Appendix C.3 and it is omitted here for clarity of presentation and space reasons.

The computation of the optimal parameter λ_2^* requires the knowledge of the divergence term $I_2(P\|Q)$, which, in turn, requires access to the form of P and Q. To this end, in Appendix B, we introduce an approach to empirically estimate the parameter preserving desirable concentration properties.

5.3. Differentability in the Target Distribution

In this section, we show that our estimator is differentiable in the target distribution and that the magnitude of the resulting gradient can be controlled via the hyperparameter λ. To this end, we assume that the target distribution P belongs to a parametric space of differentiable distributions $\mathcal{P}_\Theta=\{P_{\boldsymbol{\theta}}\in\mathscr{P}(\mathcal{X}):\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^d\}$, where Θ is the parameter space. Let us first focus on the importance weight gradient:

$$\nabla_{\boldsymbol{\theta}}\omega_\lambda(y)=\frac{(1-\lambda)\omega(y)}{(1-\lambda+\lambda\omega(y))^2}\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(y).$$

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

6. Discussion and Conclusions

In this paper, we have deepened the study of the importance sampling technique. We derived a lower bound of the deviation between the vanilla IS estimator and the true mean, proving that it allows for polynomial concentration only. Then, we introduced and analyzed a class of importance weight corrections based on the intuition of shrinking the weight towards 1. Assuming that the second moment of the importance weight exists, we have introduced the first transformation that both achieves subgaussian concentration rate and maintains the differentiability of the estimator in the target policy parameters. Future work includes studying the properties of other importance weight transformations, as well as applying these techniques to the contextual bandits and RL settings.

References

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, 2005.

Bembom, O. and van der Laan, M. J. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. 2008.

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

Boucheron, S., Lugosi, G., and Bousquet, O. Concentration inequalities. In *Summer School on Machine Learning*, pp. 208–240. Springer, 2003.

Boucheron, S., Lugosi, G., Massart, P., et al. On concentration of self-bounding functions. *Electronic Journal of Probability*, 14:1884–1899, 2009.

Bubeck, S., Cesa-Bianchi, N., and Lugosi, G. Bandits with heavy tail. *IEEE Trans. Inf. Theory*, 59(11):7711–7717, 2013. doi: 10.1109/TIT.2013.2277869.

Bullen, P. S. *Handbook of means and their inequalities*, volume 560. Springer Science & Business Media, 2013.

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

Ciosek, K. A. and Whiteson, S. OFFER: off-environment reinforcement learning. In Singh, S. P. and Markovitch, S. (eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 1819–1825. AAAI Press, 2017.

Cochran, W. G. *Sampling techniques*. John Wiley & Sons, 2007.

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

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pp. 442–450. Curran Associates, Inc., 2010.

Devroye, L., Lerasle, M., Lugosi, G., Oliveira, R. I., et al. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In Getoor, L. and Scheffer, T. (eds.), *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 1097–1104. Omnipress, 2011.

Gil, M., Alajaji, F., and Linder, T. Rényi divergence measures for commonly used univariate continuous distributions. *Inf. Sci.*, 249:124–131, 2013. doi: 10.1016/j.ins.2013.06.018.

Gilotte, A., Calauzènes, C., Nedelec, T., Abraham, A., and Dollé, S. Offline A/B testing for recommender systems. In Chang, Y., Zhai, C., Liu, Y., and Maarek, Y. (eds.), *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pp. 198–206. ACM, 2018. doi: 10.1145/3159652.3159687.

Hahn, J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pp. 315–331, 1998.

Hanna, J. P., Thomas, P. S., Stone, P., and Niekum, S. Data-efficient policy evaluation through behavior policy search. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1394–1403. PMLR, 2017.

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

Hesterberg, T. C. *Advances in importance sampling*. PhD thesis, Citeseer, 1988.

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

Huber, P. J. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pp. 492–518. Springer, 1992.

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

Jerrum, M., Valiant, L. G., and Vazirani, V. V. Random generation of combinatorial structures from a uniform distribution. *Theor. Comput. Sci.*, 43:169–188, 1986. doi: 10.1016/0304-3975(86)90174-X.

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

Kilinc, O., Hu, Y., and Montana, G. Reinforcement learning for robotic manipulation using simulated locomotion demonstrations. *CoRR*, abs/1910.07294, 2019.

Kober, J. and Peters, J. *Learning Motor Skills - From Algorithms to Robot Experiments*, volume 97 of *Springer Tracts in Advanced Robotics*. Springer, 2014. ISBN 978-3-319-03193-4. doi: 10.1007/978-3-319-03194-1.

Kuzborskij, I. and Szepesvári, C. Efron-stein pac-bayesian inequalities. *arXiv preprint arXiv:1909.01931*, 2019.

Kuzborskij, I., Vernade, C., György, A., and Szepesvári, C. Confident off-policy evaluation and selection through self-normalized importance weighting. *CoRR*, abs/2006.10460, 2020.

Langford, J. and Zhang, T. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 817–824. Citeseer, 2007.

Lee, B. K., Lessler, J., and Stuart, E. A. Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174, 2011.

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

Li, L., Chu, W., Langford, J., and Wang, X. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In King, I., Nejdl, W., and Li, H. (eds.), *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pp. 297–306. ACM, 2011. doi: 10.1145/1935826.1935878.

Li, L., Munos, R., and Szepesvári, C. Toward minimax off-policy value estimation. In Lebanon, G. and Vishwanathan, S. V. N. (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015.

Liese, F. and Vajda, I. *Convex statistical distances*, volume 95. Teubner, 1987.

Lu, S., Wang, G., Hu, Y., and Zhang, L. Optimal algorithms for lipschitz bandits with heavy-tailed rewards. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4154–4163. PMLR, 2019.

Lugosi, G. and Mendelson, S. Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.*, 19(5):1145–1190, 2019. doi: 10.1007/s10208-019-09427-x.

Mahmood, A. R., van Hasselt, H., and Sutton, R. S. Weighted importance sampling for off-policy learning with linear function approximation. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 3014–3022, 2014.

Metelli, A. M., Papini, M., Faccio, F., and Restelli, M. Policy optimization via importance sampling. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 5447–5459, 2018.

Metelli, A. M., Papini, M., Montali, N., and Restelli, M. Importance sampling techniques for policy optimization. *J. Mach. Learn. Res.*, 21:141:1–141:75, 2020.

Metelli, A. M., Papini, M., D'Oro, P., and Restelli, M. Policy optimization as online learning with mediator feedback. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 8958–8966. AAAI Press, 2021.

Moody, J. E. and Saffell, M. Learning to trade via direct reinforcement. *IEEE Trans. Neural Networks*, 12(4):875–889, 2001. doi: 10.1109/72.935097.

Murphy, S. A., van der Laan, M. J., Robins, J. M., and Group, C. P. P. R. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.

Owen, A. B. *Monte Carlo theory, methods and examples*. 2013.

Papini, M., Metelli, A. M., Lupo, L., and Restelli, M. Optimistic policy optimization via multiple importance sampling. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4989–4999. PMLR, 2019.

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

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

Rényi, A. On measures of entropy and information. Technical report, Hungarian Academy of Sciences Budapest Hungary, 1961.

Ripley, B. D. *Stochastic simulation*, volume 316. John Wiley & Sons, 2009.

Sason, I. On f-divergences: Integral representations, local behavior, and inequalities. *Entropy*, 20(5):383, 2018. doi: 10.3390/e20050383.

Siegmund, D. Importance sampling in the monte carlo study of sequential tests. *The Annals of Statistics*, pp. 673–684, 1976.

Su, Y., Dimakopoulou, M., Krishnamurthy, A., and Dudík, M. Doubly robust off-policy evaluation with shrinkage. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9167–9176. PMLR, 2020.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3231–3239, 2015.

Tang, L., Rosales, R., Singh, A., and Agarwal, D. Automatic ad format selection via contextual bandits. In He, Q., Iyengar, A., Nejdl, W., Pei, J., and Rastogi, R. (eds.), *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pp. 1587–1594. ACM, 2013. doi: 10.1145/2505515.2514700.

Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. High-confidence off-policy evaluation. In Bonet, B. and Koenig, S. (eds.), *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pp. 3000–3006. AAAI Press, 2015.

Tsallis, C. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1-2):479–487, 1988.

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

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

Wang, Y., Agarwal, A., and Dudík, M. Optimal and adaptive off-policy evaluation in contextual bandits. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3589–3597. PMLR, 2017.

Yuan, C. and Druzdzel, M. J. How heavy should the tails be? In Russell, I. and Markov, Z. (eds.), *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference, Clearwater Beach, Florida, USA*, pp. 799–805. AAAI Press, 2005.

Zheng, C. A new principle for tuning-free huber regression. *Statistica Sinica*, 2020.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187, 2017.

A. Related Works

Importance Sampling has a long history in Monte Carlo simulation as an effective technique for variance reduction in presence of rare events and for what-if analysis (Kahn & Marshall, 1953; Siegmund, 1976; Hesterberg, 1988; Cochran, 2007; Ripley, 2009). With rare exceptions (e.g., Ciosek & Whiteson, 2017; Hanna et al., 2017), in the machine learning community, IS is primarily employed for off-policy estimation and learning (e.g., Cortes et al., 2010; Mahmood et al., 2014; Thomas et al., 2015).

In this setting, it is well-known that IS might display an inconvenient behavior, depending on the behavioral Q and target P distributions (Yuan & Druzdzel, 2005; Metelli et al., 2018). In particular, IS has the effect of enlarging the range of the estimator up to $\operatorname{ess\,sup}_{y\sim Q} \frac{p(y)}{q(y)}$. Although this term is finite for discrete distributions (if $P \ll Q$), it is likely unbounded for continuous ones (Cortes et al., 2010). Furthermore, in the latter case, the vanilla IS estimator might have infinite variance under certain circumstances and tends to exhibit a heavy-tailed behavior (Metelli et al., 2018; 2020). These properties suggest that a way of addressing this phenomenon is to resort to *robust* statistics typically employed for mean estimation under heavy-tailed distributions (Lugosi & Mendelson, 2019). Methods in this class include the *trimmed mean* (Tukey & McLaughlin, 1963; Huber, 1992), the *median of means* (Nemirovskij & Yudin, 1983; Jerrum et al., 1986), and the *Catoni's* estimator (Catoni, 2012). For all of them, subgaussian guarantees were provided (Lugosi & Mendelson, 2019), but all of them, expect for the Catoni's estimator lead to non-differentiable estimators. These techniques have been also successfully employed for regret minimization algorithms for both finite (Bubeck et al., 2013) and continuous arm spaces (Lu et al., 2019). These methods could be employed *as-is* in combination with IS, but, being general-purpose, they might disregard the peculiarities of the setting.

Several *ad-hoc* methods to cope with the problematic IS behavior have been developed. An example, devised by the statistical community, is *self-normalization* (Owen, 2013). This approach has the advantage of controlling the range of the estimator at the price of making all samples interdependent and generating a bias. Although the asymptotic consistency is guaranteed (Hesterberg, 1995; Swaminathan & Joachims, 2015), its finite-sample analysis is more challenging. In (Metelli et al., 2018) a polynomial concentration inequality was provided and, more recently, exponential bounds based on Efron-Stein inequalities have been proposed (Kuzborskij & Szepesvári, 2019; Kuzborskij et al., 2020). Nevertheless, the resulting inequality contains terms that are not easy to estimate (Kuzborskij et al., 2020). Another popular technique is the weight *truncation* (or *clipping*) (Ionides, 2008; Bottou et al., 2013). Some works rely on empirical selections of the truncation threshold (Lee et al., 2011; Cole & Hernán, 2008), while others focus on more theoretically principled approaches (Bembom & van der Laan, 2008; Wang et al., 2017; Papini et al., 2019). In particular, in (Wang et al., 2017) an approach designed for CMABs combines truncation with DR estimator, deriving theoretical guarantees in MSE. Instead, in (Papini et al., 2019) a subgaussian deviation bound is derived by suitably adapting the truncation threshold as a function of the number of samples n and the confidence parameter δ. Finally, a not so large part of the literature focuses on less crude transformations than truncation, called *smoothing* (Vehtari et al., 2015). They typically take into explicit consideration the estimator tails (Pickands III et al., 1975), also providing asymptotic guarantees. Very recently, shrinkage transformations of the weight was proposed, based on the minimization of different bounds on the MSE, in the specific setting of CMABs (Su et al., 2020).

B. Data-driven Tuning of λ

The computation of the optimal parameter λ_2^* requires the knowledge of the divergence term $I_2(P\|Q)$, which, in turn, requires access to the form of P and Q. Even when P and Q are known, it may be complex to compute the divergence, in particular for continuous distributions, since it involves the evaluation of a complex integral.[4] In principle, we could estimate the divergence $I_2(P\|Q)$ from samples, by computing the empirical second moment of the vanilla importance weights $\frac{1}{n}\sum_{i\in[n]}\omega(y_i)^2$, as done in previous works (Metelli et al., 2018). However, this approach would prevent any subgaussian concentration property, as the behavior of the non-corrected $\omega(y)^2$ will be surely heavy-tailed whenever $\omega(y)$ is. A general-purpose approach to circumvent this issue and avoid the divergence estimation is the *Lepski's adaptation method* (Lepski & Spokoiny, 1997), which only requires knowing an upper and a lower bound on the quantity to adapt to, $I_2(P\|Q)$ in our case. However, Lepski's method is known to be typically computationally intensive.

In this section, we follow a different path inspired to the recent work by Zheng (2020). If a choice of the parameter λ corrects the weight ω_λ leading to an ideal estimator $\widehat{\mu}_\lambda$, for the mean μ, we may expect that the empirical second moment of the corrected weights will provide a reasonable estimation of $I_2(P\|Q)$. Based on this, we propose to choose λ by solving

[4] For some known distributions, including Gaussians, the integral can be computed in closed form (Gil et al., 2013).

the equation:

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

The intuition behind this approach can be stated as follows. If the empirical second moment is close to the divergence, i.e., $\frac{1}{n}\sum_{i\in[n]}\omega_{\lambda n^{1/4}}(y_i)^2 \simeq I_2(P\|Q)$, the solution $\widehat{\lambda}$ of Equation (2) approaches the optimal parameter, i.e., $\widehat{\lambda} \simeq \sqrt{\frac{2\log\frac{1}{\delta}}{3I_2(P\|Q)n}} = \lambda_2^*$. We formalize this reasoning in Appendix C.4, proving that Equation (2) admits a unique root $\widehat{\lambda}\in[0,1]$ (Lemma C.4) and that when the number of samples n grows to infinity, $\widehat{\lambda}$ actually converges to λ_2^* (Lemma C.8).

The following result provides the concentration properties of the estimator $\widehat{\mu}_\lambda$ when using $\widehat{\lambda}$ instead of λ_2^*, under slightly more demanding requirements on the moments of the importance weights.

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

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

Compared to Theorem 5.1, this result is weakened in three aspects. First, the constant $5(2+\sqrt{3})/\sqrt{6}\simeq 7.62$ is larger. Second, the inequality holds with a smaller probability $1-2\delta$. This is explained by the fact that two estimation processes with the same samples are needed, i.e., the computation of $\widehat{\lambda}$ and the corrected estimator $\widehat{\mu}_{\widehat{\lambda}}$. Third, and most important, the result holds for sufficiently large n, whose value is reported in the proof and depends on $I_3(P\|Q)$, which must be finite. We think this is not a too strong requirement considering that even the variance of a empirical estimate of $I_2(P\|Q)$ would depend on the fourth moment of the importance weight, i.e., $I_4(P\|Q)$. It is worth noting that we could, in principle, select a value of λ that is independent of $I_2(P\|Q)$. In such a case, we are able to get a bound of order $\mathcal{O}\left(I_2(P\|Q)\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$, with a higher exponent for $I_2(P\|Q)$ (Corollary C.2 of Appendix C.4).

C. Proofs and Derivations

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

C.1. Proofs of Section 3

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

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

Proof. The proof is inspired to that of Proposition 6.2 of (Catoni, 2012). 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 $p,q\in[0,1]$:

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

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 $\alpha\in(1,2]$:

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

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

$$p \leqslant 1 \implies n \geqslant \frac{a}{\epsilon}\xi^{\frac{1}{\alpha}}, \tag{P.1}$$

$$q \leqslant 1 \implies n \geqslant \frac{a}{\epsilon}\xi^{\frac{1}{\alpha-1}}. \tag{P.2}$$

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

$$\begin{aligned} I_\alpha(P\|Q) &= 2\left(\frac{p}{2}\right)^\alpha \left(\frac{q}{2}\right)^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha} \\ &= p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha} \\ &= \xi + \left(1-\xi\left(\frac{a}{n\epsilon}\right)^{\alpha-1}\right)^\alpha \left(1-\xi\left(\frac{a}{n\epsilon}\right)^\alpha\right)^{1-\alpha} \leqslant \xi + 1, \end{aligned}$$

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

$$\left(1-\xi\left(\frac{a}{n\epsilon}\right)^{\alpha-1}\right)^\alpha \left(1-\xi\left(\frac{a}{n\epsilon}\right)^\alpha\right)^{1-\alpha} \leqslant \left(1-\xi\left(\frac{a}{n\epsilon}\right)^{\alpha-1}\right)^\alpha \left(1-\xi\left(\frac{a}{n\epsilon}\right)^{\alpha-1}\right)^{1-\alpha} = 1-\xi\left(\frac{a}{n\epsilon}\right)^{\alpha-1} \leqslant 1.$$

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

$$\begin{aligned} \mathop{\mathbb{P}}_{y_i \overset{\text{iid}}{\sim} Q}(|\widehat{\mu} - \mu| > \epsilon) &= \mathop{\mathbb{P}}_{y_i \overset{\text{iid}}{\sim} Q}(\{\widehat{\mu} - \mu < -\epsilon\} \cup \{\widehat{\mu} - \mu > \epsilon\}) \\ &= \mathop{\mathbb{P}}_{y_i \overset{\text{iid}}{\sim} Q}(\widehat{\mu} - \mu < -\epsilon) + \mathop{\mathbb{P}}_{y_i \overset{\text{iid}}{\sim} Q}(\widehat{\mu} - \mu > \epsilon) \qquad \text{(P.3)} \\ &= 2\mathop{\mathbb{P}}_{y_i \overset{\text{iid}}{\sim} Q}(\widehat{\mu} - \mu > \epsilon), \qquad \text{(P.4)} \end{aligned}$$

where line (P.3) is obtained by observing that the two events are disjoint and line (P.4) comes from the symmetry of the events. We now lower bound the probability:

$$\begin{aligned} \mathop{\mathbb{P}}_{y_i \overset{\text{iid}}{\sim} Q}(\widehat{\mu} - \mu > \epsilon) &\geqslant \mathop{\mathbb{P}}_{y_i \overset{\text{iid}}{\sim} Q}(\text{among the } n \text{ samples, one is } a \text{ and the remaining are } 0) \\ &= n\frac{q}{2}(1-q)^{n-1} \\ &= \frac{1}{2}\left(\frac{a}{\epsilon}\right)^\alpha n^{1-\alpha}\xi\left(1-\left(\frac{a}{n\epsilon}\right)^\alpha \xi\right)^{n-1}. \end{aligned}$$

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

$$\frac{1}{2}\left(\frac{a}{\epsilon}\right)^\alpha n^{1-\alpha}\xi\left(1-\left(\frac{a}{n\epsilon}\right)^\alpha \xi\right)^{n-1} \leqslant \delta \quad \implies \quad \epsilon \geqslant a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1-\left(\frac{a}{n\epsilon}\right)^\alpha \xi\right)^{\frac{n-1}{\alpha}}. \tag{P.5}$$

We claim that, for $\delta \in (0, e^{-1})$, any value of ϵ fulfilling condition (P.5) must be $\epsilon \leqslant \epsilon^\star$:

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

Indeed, we have:

$$\begin{aligned} a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1-\left(\frac{a}{n\epsilon^\star}\right)^\alpha \xi\right)^{\frac{n-1}{\alpha}} &= a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1-\left(\frac{a}{n}\right)^\alpha\left(a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1-\frac{e\delta}{n}\right)^{\frac{n-1}{\alpha}}\right)^{-\alpha}\xi\right)^{\frac{n-1}{\alpha}} \\ &= a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1-\frac{\delta}{n}\left(1-\frac{e\delta}{n}\right)^{-(n-1)}\right)^{\frac{n-1}{\alpha}} \\ &\geqslant a\left(\frac{\xi}{\delta n^{\alpha-1}}\right)^{\frac{1}{\alpha}}\left(1-\frac{\delta e}{n}\right)^{\frac{n-1}{\alpha}} = \epsilon^\star, \end{aligned}$$

where the last inequality derives from observing that $\left(1-\frac{e\delta}{n}\right)^{-(n-1)} \leqslant e$ if $\delta \in (0, e^{-1})$. Finally, we rephrase conditions (P.1) and (P.2):

$$n \geqslant \frac{a}{\epsilon^\star}\xi^{\frac{1}{\alpha}} \quad \implies \quad n \geqslant n^{1-\frac{1}{\alpha}}\delta^{\frac{1}{\alpha}}\left(1-\frac{e\delta}{n}\right)^{-\frac{n-1}{\alpha}} \quad \implies \quad n \geqslant \delta e,$$

$$n \geqslant \frac{a}{\epsilon^\star}\xi^{\frac{1}{\alpha-1}} \quad \Longrightarrow \quad n \geqslant n^{1-\frac{1}{\alpha}}\delta^{\frac{1}{\alpha}}\xi^{\frac{1}{\alpha(\alpha-1)}}\left(1-\frac{e\delta}{n}\right)^{-\frac{n-1}{\alpha}} \quad \Longrightarrow \quad n \geqslant \delta e \xi^{\frac{1}{\alpha-1}},$$

having observed, again, that $\left(1-\frac{e\delta}{n}\right)^{-\frac{n-1}{\alpha}} \leqslant e^{\frac{1}{\alpha}}$. Thus, we enforce the condition $n \geqslant \delta e \max\left\{1, \xi^{\frac{1}{\alpha-1}}\right\}$. □

Corollary C.1. *There exist two distributions $P, Q \in \mathscr{P}(\mathcal{Y})$ with $P \ll Q$ and a bounded measurable function $f : \mathcal{Y} \to \mathbb{R}$ such that for every $\alpha \in (1,2]$ it holds that:*

$$\mathop{\mathbb{E}}_{y_i \overset{iid}{\sim} Q}\left[|\widehat{\mu} - \mu|^\alpha\right] \geqslant \|f\|_\infty^\alpha \frac{I_\alpha(P\|Q) - 1}{n^{\alpha-1}}.$$

Proof. Let us denote the bad event:

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

From Theorem 3.1, we know that $\mathbb{P}_{y_i \overset{iid}{\sim} Q}(\mathcal{E}) \geqslant \delta$. Let us consider the expected absolute error with exponent $\alpha \in (1,2]$ and apply the law of total expectation:

$$\begin{aligned}\mathop{\mathbb{E}}_{y_i \overset{iid}{\sim} Q}\left[|\widehat{\mu} - \mu|^\alpha\right] &= \mathop{\mathbb{E}}_{y_i \overset{iid}{\sim} Q}\left[(\widehat{\mu} - \mu)^\alpha \,|\mathcal{E}\right] \mathop{\mathbb{P}}_{y_i \overset{iid}{\sim} Q}(\mathcal{E}) + \mathop{\mathbb{E}}_{y_i \overset{iid}{\sim} Q}\left[(\widehat{\mu} - \mu)^\alpha \,|\mathcal{E}^c\right] \mathop{\mathbb{P}}_{y_i \overset{iid}{\sim} Q}(\mathcal{E}^c) \\ &\geqslant \|f\|_\infty^\alpha \frac{I_\alpha(P\|Q)-1}{\delta n^{\alpha-1}} \left(1-\frac{e\delta}{n}\right)^{n-1} \delta + 0.\end{aligned}$$

The result is obtained by setting $\delta \to 0$. □

C.2. Proofs of Section 4

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

(i) if $s \leqslant s'$ then $\omega_{\lambda,s}(y) \leqslant \omega_{\lambda,s'}(y)$;
(ii) if $s < 0$ then $\omega_{\lambda,s}(y) \leqslant \lambda^{\frac{1}{s}}$, otherwise if $s > 0$ then $\omega_{\lambda,s}(y) \geqslant \lambda^{\frac{1}{s}}$;
(iii) if $s < 1$ then $\mathbb{E}_{y\sim Q}[\omega_{\lambda,s}(y)] \leqslant 1$, otherwise if $s > 1$ then $\mathbb{E}_{y\sim Q}[\omega_{\lambda,s}(y)] \geqslant 1$.

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 (Bullen, 2013). Let us move to (ii), if $s < 0$, we have:

$$\omega_{\lambda,s}(y) = \Big((1-\lambda)\omega(y)^s + \lambda\Big)^{\frac{1}{s}} = \frac{1}{\left(\frac{1-\lambda}{\omega(y)^{-s}} + \lambda\right)^{\frac{1}{-s}}} \leqslant \lambda^{\frac{1}{s}}.$$

Instead for $s > 0$, we have:

$$\omega_{\lambda,s}(y) = \Big((1-\lambda)\omega(y)^s + \lambda\Big)^{\frac{1}{s}} \geqslant \lambda^{\frac{1}{s}}.$$

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

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

Symmetrically, for $s > 1$ we have:

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

□

C.3. Proofs of Section 5

Before going to the proofs, we introduce the following integral:

$$J_\alpha(P\|Q) = \int_{\mathcal{Y}} q(y)\left|p(y)^\alpha q(y)^{-\alpha} - 1\right| \mathrm{d}y.$$

For $\alpha=1$, $J_1(P\|Q)$ reduces to the total variation divergence. For general values of α, $J_\alpha(P\|Q)$ represents the χ^α-divergence (Liese & Vajda, 1987; Sason, 2018). $J_\alpha(P\|Q)$ can be also seen as the α-absolute central moment of the importance weight $\omega(y)=p(y)/q(y)$. Consequently, we immediacy conclude that $J_\alpha(P\|Q)\leqslant I_\alpha(P\|Q)$. In particular, for $\alpha=2$, we have $J_2(P\|Q)=I_2(P\|Q)-1$.

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

$$\left|\mathop{\mathbb{E}}_{y_i\overset{iid}{\sim}Q}[\widehat{\mu}_\lambda]-\mu\right|\leqslant\|f\|_\infty\lambda^{\alpha-1}J_\alpha(P\|Q)^{\frac{1}{\alpha}}\left[(1-\lambda)I_\alpha(P\|Q)+\lambda\right]^{1-\frac{1}{\alpha}}.$$

Proof. Let us consider the following derivation:

$$\left|\mathop{\mathbb{E}}_{y_i\overset{iid}{\sim}Q}[\widehat{\mu}_\lambda]-\mu\right|=\left|\mathop{\mathbb{E}}_{y_i\overset{iid}{\sim}Q}[\widehat{\mu}_\lambda-\widehat{\mu}]\right|\leqslant=\mathop{\mathbb{E}}_{y_i\overset{iid}{\sim}Q}[|\widehat{\mu}_\lambda-\widehat{\mu}|]\leqslant\|f\|_\infty\mathop{\mathbb{E}}_{y\sim Q}[|\omega_\lambda(y)-\omega(y)|].$$

Thus, we have for $\alpha\in(1,2]$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{y\sim Q}[|\omega_\lambda(y)-\omega(y)|]&=\mathop{\mathbb{E}}_{y\sim Q}\left[\left|\frac{\omega(y)}{1-\lambda+\lambda\omega(y)}-\omega(y)\right|\right]\\
&=\lambda\mathop{\mathbb{E}}_{y\sim Q}\left[\frac{|\omega(y)-1|}{\frac{1-\lambda}{\omega(y)}+\lambda}\right]\\
&=\lambda\mathop{\mathbb{E}}_{y\sim Q}\left[|\omega(y)-1|\left(\frac{1}{\frac{1-\lambda}{\omega(y)}+\lambda}\right)^{\alpha-1}\left(\frac{1}{\frac{1-\lambda}{\omega(y)}+\lambda}\right)^{2-\alpha}\right]\\
&\leqslant\lambda\sup_{v\geqslant0}\left(\frac{1}{\frac{1-\lambda}{v}+\lambda}\right)^{2-\alpha}\mathop{\mathbb{E}}_{y\sim Q}\left[|\omega(y)-1|\left(\frac{1}{\frac{1-\lambda}{\omega(y)}+\lambda}\right)^{\alpha-1}\right].
\end{aligned}$$

Concerning the first term, we observe that the function $\frac{1}{\frac{1-\lambda}{v}+\lambda}$ is monotonically increasing in v and, consequently:

$$\sup_{v\geqslant0}\left(\frac{1}{\frac{1-\lambda}{v}+\lambda}\right)^{2-\alpha}=\lim_{v\to\infty}\left(\frac{1}{\frac{1-\lambda}{v}+\lambda}\right)^{2-\alpha}=\frac{1}{\lambda^{2-\alpha}}.$$

Concerning the second term, we proceed as follows:

$$\begin{aligned}
\mathop{\mathbb{E}}_{y\sim Q}\left[|\omega(y)-1|\left(\frac{1}{\frac{1-\lambda}{\omega(y)}+\lambda}\right)^{\alpha-1}\right]&\leqslant\mathop{\mathbb{E}}_{y\sim Q}[|\omega(y)-1|^\alpha]^{\frac{1}{\alpha}}\mathop{\mathbb{E}}_{y\sim Q}\left[\left(\frac{1}{\frac{1-\lambda}{\omega(y)}+\lambda}\right)^{\alpha}\right]^{1-\frac{1}{\alpha}} &\text{(P.6)}\\
&\leqslant\mathop{\mathbb{E}}_{y\sim Q}[|\omega(y)-1|^\alpha]^{\frac{1}{\alpha}}\mathop{\mathbb{E}}_{y\sim Q}[((1-\lambda)\omega(y)+\lambda)^\alpha]^{1-\frac{1}{\alpha}} &\text{(P.7)}\\
&\leqslant\mathop{\mathbb{E}}_{y\sim Q}[|\omega(y)-1|^\alpha]^{\frac{1}{\alpha}}\mathop{\mathbb{E}}_{y\sim Q}[(1-\lambda)\omega(y)^\alpha+\lambda]^{1-\frac{1}{\alpha}} &\text{(P.8)}\\
&=J_\alpha(P\|Q)^{\frac{1}{\alpha}}\left[(1-\lambda)I_\alpha(P\|Q)+\lambda\right]^{1-\frac{1}{\alpha}},
\end{aligned}$$

where line (P.6) derives from Hölder's inequality with exponents α and $\frac{\alpha}{\alpha-1}$, line (P.7) is obtained from the power mean inequality (Bullen, 2013) by bounding the harmonic mean with the arithmetic mean, line (P.8) follows from Jensen's inequality having observed that the function $\cdot^\alpha$ is a convex function. □

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

$$\left|\mathop{\mathbb{E}}_{y\overset{iid}{\sim}Q}[\widehat{\mu}_\lambda]-\mu\right|\leqslant\|f\|_\infty\lambda^{\alpha-1}I_\alpha(P\|Q).$$

Proof. The result follows immediately from Lemma C.1 by recalling that $J_\alpha(P\|Q)\leqslant I_\alpha(P\|Q)$ and observing that $(1-\lambda)I_\alpha(P\|Q)+\lambda\leqslant I_\alpha(P\|Q)$ as $I_\alpha(P\|Q)\geqslant1$. □

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

$$\underset{y_i\overset{iid}{\sim}Q}{\mathbb{V}\mathrm{ar}}\left[\widehat{\mu}_\lambda\right]\leqslant\frac{\|f\|_\infty^2}{n\lambda^{2-\alpha}}\left[(1-\lambda)I_\alpha(P\|Q)+\lambda\right].$$

Proof. Let us consider the following derivation:

$$\underset{y_i\overset{\text{iid}}{\sim}Q}{\mathbb{V}\mathrm{ar}}\left[\widehat{\mu}_\lambda\right]=\frac{1}{n}\underset{y\sim Q}{\mathbb{V}\mathrm{ar}}[\omega_\lambda(y)f(y)]\leqslant\frac{1}{n}\underset{y\sim Q}{\mathbb{E}}\left[\omega_\lambda(y)^2f(y)^2\right]\leqslant\frac{1}{n}\|f\|_\infty^2\underset{y\sim Q}{\mathbb{E}}\left[\omega_\lambda(y)^2\right].$$

Thus, we have for $\alpha\in(1,2]$:

$$\begin{aligned}
\underset{y\sim Q}{\mathbb{E}}\left[\omega_\lambda(y)^2\right]&=\underset{y\sim Q}{\mathbb{E}}\left[\left(\frac{1}{\frac{1-\lambda}{\omega(y)}+\lambda}\right)^2\right]\\
&=\underset{y\sim Q}{\mathbb{E}}\left[\left(\frac{1}{\frac{1-\lambda}{\omega(y)}+\lambda}\right)^\alpha\left(\frac{1}{\frac{1-\lambda}{\omega(y)}+\lambda}\right)^{2-\alpha}\right]\\
&\leqslant\sup_{v\geqslant0}\left(\frac{1}{\frac{1-\lambda}{v}+\lambda}\right)^{2-\alpha}\underset{y\sim Q}{\mathbb{E}}\left[\left(\frac{1}{\frac{1-\lambda}{\omega(y)}+\lambda}\right)^\alpha\right]\\
&\leqslant\frac{1}{\lambda^{2-\alpha}}\left[(1-\lambda)I_\alpha(P\|Q)+\lambda\right],
\end{aligned}$$

where the last line is obtained by employing analogous derivations as in Lemma C.1. □

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

$$\underset{y_i\overset{iid}{\sim}Q}{\mathbb{V}\mathrm{ar}}\left[\widehat{\mu}_\lambda\right]\leqslant\frac{\|f\|_\infty^2}{n\lambda^{2-\alpha}}I_\alpha(P\|Q).$$

Proof. The result is obtained from Lemma C.2 by observing that $(1-\lambda)I_\alpha(P\|Q)+\lambda\leqslant I_\alpha(P\|Q)$ as $I_\alpha(P\|Q)\geqslant1$. □

Lemma C.3. *Let $P,Q\in\mathscr{P}(\mathcal{Y})$ 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 $\delta\in(0,1)$ then, for every $\lambda\in[0,1]$, with probability at least $1-\delta$ it holds that:*

$$\widehat{\mu}_\lambda-\mu\leqslant\|f\|_\infty\sqrt{\frac{2\log\frac{1}{\delta}}{n\lambda^{2-\alpha}}\left[(1-\lambda)I_\alpha(P\|Q)+\lambda\right]}+\frac{2\|f\|_\infty\log\frac{1}{\delta}}{3\lambda n}+\|f\|_\infty\lambda^{\alpha-1}J_\alpha(P\|Q)^{\frac{1}{\alpha}}\left[(1-\lambda)I_\alpha(P\|Q)+\lambda\right]^{1-\frac{1}{\alpha}}.$$

Proof. The proof is a straightforward application of Bernstein's inequality together with Lemma C.1 and Lemma C.2. First of all, we highlight the bias in the following decomposition:

$$\widehat{\mu}_\lambda-\mu=\underbrace{\widehat{\mu}_\lambda-\underset{y_i\overset{\text{iid}}{\sim}Q}{\mathbb{E}}\left[\widehat{\mu}_\lambda\right]}_{\text{concentration}}+\underbrace{\underset{y_i\overset{\text{iid}}{\sim}Q}{\mathbb{E}}\left[\widehat{\mu}_\lambda\right]-\mu}_{\text{bias}}.$$

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

$$\begin{aligned}
\widehat{\mu}_\lambda-\underset{y_i\overset{\text{iid}}{\sim}Q}{\mathbb{E}}\left[\widehat{\mu}_\lambda\right]&\leqslant\sqrt{2\underset{y_i\overset{\text{iid}}{\sim}Q}{\mathbb{V}\mathrm{ar}}\left[\widehat{\mu}_\lambda\right]\log\frac{1}{\delta}}+\frac{2\|\mu_\lambda\|_\infty\log\frac{1}{\delta}}{3n}\\
&\leqslant\|f\|_\infty\sqrt{\frac{2\log\frac{1}{\delta}}{n\lambda^{2-\alpha}}\left[(1-\lambda)I_\alpha(P\|Q)+\lambda\right]}+\frac{2\|f\|_\infty\log\frac{1}{\delta}}{3\lambda n},
\end{aligned}$$

where the last line is obtained by bounding the variance with Lemma C.2 and recalling that $\|\mu_\lambda\|_\infty\leqslant\frac{\|f\|_\infty}{\lambda}$. □

We discuss how to optimize this bound in λ in Appendix D. We now move to a simplified version of the bound.

Theorem 5.1. *Let $P, Q \in \mathscr{P}(\mathcal{Y})$ 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 $\delta \in (0,1)$, let*

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

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

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

Proof. The derivation is analogous to that of Lemma C.3 using Bernstein's inequality, Lemma 5.1, and Lemma 5.2, leading to the inequality:

$$\widehat{\mu}_\lambda - \mu \leqslant \|f\|_\infty \sqrt{\frac{2\log\frac{1}{\delta}}{n\lambda^{2-\alpha}} I_\alpha(P\|Q)} + \frac{2\log\frac{1}{\delta}}{3\lambda n}\|f\|_\infty + \|f\|_\infty \lambda^{\alpha-1} I_\alpha(P\|Q) \tag{P.9}$$

This is a convex function of λ that can be minimized by vanishing the derivative. The derivative is actually a quadratic function in $\lambda^{\frac{\alpha}{2}}$ and its positive solution has a quite complex expression:

$$\lambda_\alpha^\# := \left(\frac{-3\alpha+\sqrt{3}\sqrt{(\alpha+2)(3\alpha-2)}+6}{6\sqrt{2}(\alpha-1)}\right)^{\frac{2}{\alpha}} \left(\frac{\log\frac{1}{\delta}}{I_\alpha(P\|Q)n}\right)^{\frac{2}{\alpha}} \leqslant \left(\frac{2}{3(\alpha-1)^2}\right)^{\frac{1}{\alpha}} \left(\frac{\log\frac{1}{\delta}}{I_\alpha(P\|Q)n}\right)^{\frac{1}{\alpha}} =: \lambda_\alpha^*,$$

where the inequality holds with equality when $\alpha = 2$. By substituting this value of λ_α^* we obtain the bound:

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

having observed that $\left(2-\sqrt{3}+\alpha(-2+\sqrt{3}+\alpha)\right)$ is a monotonically increasing function of α. □

Remark C.1. *In the proof of Theorem 5.1, we did not consider the possibility that $\lambda_\alpha^* > 1$, that would lead to a non-valid correction parameter. We claim that this circumstance occurs for very small values of n and δ only. Indeed:*

$$\lambda_\alpha^* \leqslant 1 \quad \Longrightarrow \quad n \geqslant \frac{2\log\frac{1}{\delta}}{3(\alpha-1)^2 I_\alpha(P\|Q)}.$$

In any case, if it occurs that $\lambda_\alpha^* > 1$, we conventionally clip it to 1.

Proposition C.1. *Let $\lambda \in [0,1]$. For every $(x,a) \in \mathcal{X}\times\mathcal{A}$, let $\omega(a|x) = \frac{\pi_{\boldsymbol{\theta}}(a|x)}{\pi_b(a|x)}$, for a target policy $\pi_{\boldsymbol{\theta}}$ differentiable in $\boldsymbol{\theta}$. Then, it holds that:*

$$\|\nabla_{\boldsymbol{\theta}} \omega_\lambda(a|x)\|_\infty \leqslant \frac{1}{4\lambda}\|\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|x)\|_\infty.$$

Proof. Let us first compute the gradient explicitly:

$$\nabla_{\boldsymbol{\theta}} \omega_\lambda(a|x) = \frac{\partial \omega_\lambda}{\partial \omega}(a|x) \nabla_{\boldsymbol{\theta}} \omega(a|x) = \frac{1-\lambda}{(1-\lambda+\lambda\omega(a|x))^2} \omega(a|x) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|x)$$

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

$$g(v) = \frac{(1-\lambda)v}{(1-\lambda+\lambda v)^2}.$$

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:

$$\frac{\partial g(v)}{\partial v} = \frac{(1-\lambda)(1-\lambda-\lambda v)}{(1-\lambda+\lambda v)^3} = 0 \quad \Longrightarrow \quad v^* = \frac{1}{\lambda} - 1.$$

By substituting the found value, we obtain:

$$g(v^*) = \frac{1}{4\lambda}.$$

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

C.4. Proofs of Section B

For the sake of simplicity, we will denote with $\eta = \lambda n^{1/4}$. We introduce the following equation:

$$h(\eta) = \eta^2 \mathop{\mathbb{E}}_{y\sim Q}\left[\omega_\eta(y)^2\right] = \frac{2\log\frac{1}{\delta}}{3\sqrt{n}},$$

and we denote with $\eta^\dagger$ a solution of this equation. We introduce the corresponding empirical version, that is equivalent to Equation (2):

$$\widehat{h}(\eta) = \frac{\eta^2}{n}\sum_{i\in[n]}\omega_\eta(y_i)^2 = \frac{2\log\frac{1}{\delta}}{3\sqrt{n}},$$

having $\widehat{\eta}$ as solution. Clearly, we have $\mathbb{E}_{y_i \overset{\text{iid}}{\sim} Q}\left[\widehat{h}(\eta)\right] = h(\eta)$.

Lemma C.4. *Let $h(\eta) = \eta^2\,\mathbb{E}_{y\sim Q}\left[\omega_\eta(y)^2\right]$. The following properties hold:*

(i) for every $\eta\in[0,1]$ we have $h(\eta)\in[0,1]$;
(ii) for every $c\in(0,1]$, the equation $h(\eta) = c$ admits at most one solution.

Proof. For (i) we immediately observe that $h(\eta)\geqslant 0$. Moreover, we have $\omega_\eta(y)\leqslant\eta^{-1}$, from which the result follows. For (ii) we show that $h(\eta)$ is monotonically increasing in η:

$$\frac{\partial h}{\partial\eta}(\eta) = 2\eta \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)^2}{(1-\eta+\eta\omega(y))^3}\right] > 0.$$

□

Remark C.2. *It might be the case that the equation $\widehat{h}(\eta) = \frac{2\log\frac{1}{\delta}}{3\sqrt{n}}$ admits no solution, for instance when $\frac{2\log\frac{1}{\delta}}{3\sqrt{n}} > 1$ or when $\sup_{\eta\in[0,1]}\widehat{h}(\eta) < 1$. In these cases, we conventionally set the solution $\eta^\dagger = 1$. We stress that this circumstance occurs only for small values of n, as in Remark C.1. Indeed, the right hand side $\frac{2\log\frac{1}{\delta}}{3\sqrt{n}} \to 0$ when $n\to\infty$.*

Lemma C.5. *Let $h(\eta) = \eta^2\,\mathbb{E}_{y\sim Q}\left[\omega_\eta(y)^2\right]$. Let $\eta^\dagger\in[0,1]$ such that:*

$$h(\eta^\dagger) = \frac{2\log\frac{1}{\delta}}{3\sqrt{n}} \quad \textit{and} \quad \lambda^\dagger = \eta^\dagger n^{-1/4}$$

then it holds that:

$$\lambda_2^* \leqslant \lambda^\dagger \leqslant \sqrt{2}\lambda_2^*,$$

where the second inequality holds if $n \geqslant \frac{4096(I_3(P\|Q)-I_2(P\|Q))^4\left(\log\frac{1}{\delta}\right)^2}{9I_2(P\|Q)^6}$, whenever $I_3(P\|Q)$ is finite.

Proof. Let us first observe that:

$$\mathop{\mathbb{E}}_{y\sim Q}\left[\omega_\eta(y)^2\right] = \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{1}{\left(\frac{1-\eta}{\omega(y)}+\eta\right)^2}\right] \leqslant \mathop{\mathbb{E}}_{y\sim Q}\left[((1-\eta)\omega(y)+\eta)^2\right] = (I_2(P\|Q)-1)\eta^2 + 1 \leqslant I_2(P\|Q),$$

where the first inequality derives from the inequality between the harmonic and arithmetic mean. From the last inequality, we have:

$$h(\eta)\leqslant\eta^2 I_2(P\|Q) \quad\Longrightarrow\quad \eta^\dagger \geqslant \sqrt{\frac{2\log\frac{1}{\delta}}{3I_2(P\|Q)\sqrt{n}}} \quad\Longrightarrow\quad \lambda^\dagger = \sqrt{\frac{2\log\frac{1}{\delta}}{3I_2(P\|Q)n}} = \lambda_2^*.$$

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

$$\frac{1}{\left(\frac{1-\eta}{\omega(y)}+\eta\right)^2} = \omega(y)^2 - 2\omega(y)^2(\omega(y)-1)\eta + 3(\omega(y)-1)^2\omega(y)^2\overline{\eta}^2 \geqslant \omega(y)^2 - 2\omega(y)^2(\omega(y)-1)\eta,$$

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

$$\mathop{\mathbb{E}}_{y\sim Q}\left[\frac{1}{\left(\frac{1-\eta}{\omega(y)}+\eta\right)^2}\right] \geqslant \mathop{\mathbb{E}}_{y\sim Q}\left[\omega(y)^2 - 2\omega(y)^2(\omega(y)-1)\eta\right] = I_2(P\|Q) - 2\eta(I_3(P\|Q) - I_2(P\|Q)).$$

By moving to function $h(\eta)$, and recalling the equation $h(\eta) = \frac{2\log\frac{1}{\delta}}{3\sqrt{n}}$, we have:

$$h(\eta) = \eta^2 \mathop{\mathbb{E}}_{y\sim Q}\left[\omega_\eta(y)^2\right] \geqslant \eta^2 I_2(P\|Q) - 2\eta^3(I_3(P\|Q) - I_2(P\|Q))$$
$$\implies \quad \eta^2 I_2(P\|Q) - 2\eta^3(I_3(P\|Q) - I_2(P\|Q)) \leqslant \frac{2\log\frac{1}{\delta}}{3\sqrt{n}}.$$

We prove that for sufficiently large n, all solutions $\eta^\dagger$ of the previous inequality satisfy $\eta \leqslant \sqrt{\frac{4\log\frac{1}{\delta}}{3I_2(P\|Q)\sqrt{n}}}$:

$$\frac{4\log\frac{1}{\delta}}{3I_2(P\|Q)\sqrt{n}} I_2(P\|Q) - 2\left(\frac{4\log\frac{1}{\delta}}{3I_2(P\|Q)\sqrt{n}}\right)^{\frac{3}{2}}(I_3(P\|Q) - I_2(P\|Q)) > \frac{2\log\frac{1}{\delta}}{3I_2(P\|Q)\sqrt{n}}$$
$$\implies \quad n \geqslant \frac{4096\,(I_3(P\|Q) - I_2(P\|Q))^4\left(\log\frac{1}{\delta}\right)^2}{9I_2(P\|Q)^6}.$$

This, implies that $\lambda^\dagger \leqslant \sqrt{\frac{4\log\frac{1}{\delta}}{3I_2(P\|Q)n}} = \sqrt{2}\lambda_2^*$. □

Lemma C.6. *Let $h(\eta) = \eta^2\,\mathbb{E}_{y\sim Q}\left[\omega_\eta(y)^2\right]$, then it holds that:*

$$\frac{\partial h(\eta)}{\partial \eta^2} \geqslant I_2(P\|Q)^{-2}.$$

Proof. Let us first observe that:

$$\frac{\partial h(\eta)}{\partial \eta^2} = \frac{\partial h(\eta)}{\partial \eta}\frac{\partial \eta}{\partial \eta^2} = \frac{\partial h(\eta)}{\partial \eta}\frac{1}{2\eta}.$$

The first factor was already computed in the proof of Lemma C.4. We now lower bound it. Let us first prove the following auxiliary inequality:

$$\begin{aligned} 1 = \mathop{\mathbb{E}}_{y\sim Q}[\omega(y)]^2 = \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)}{1-\lambda+\lambda\omega(y)}(1-\lambda+\lambda\omega(y))\right]^2 &\leqslant \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)^2}{(1-\lambda+\lambda\omega(y))^2}\right]\mathop{\mathbb{E}}_{y\sim Q}\left[(1-\lambda+\lambda\omega(y))^2\right] \\ &\leqslant \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)^2}{(1-\lambda+\lambda\omega(y))^2}\right] I_2(P\|Q), \end{aligned} \tag{P.10}$$

where the first inequality follows from Cauchy-Schwarz's and the second one by recalling that $\mathbb{E}_{y\sim Q}\left[(1-\lambda+\lambda\omega(y))^2\right] \leqslant I_2(P\|Q)$. Now, we proceed with Hölder's inequality with $p = \frac{3}{2}$ and $q = 3$:

$$\begin{aligned} \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)^2}{(1-\lambda+\lambda\omega(y))^2}\right] \leqslant \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)^{\frac{4}{3}}}{(1-\lambda+\lambda\omega(y))^2}\omega(y)^{\frac{2}{3}}\right] &\leqslant \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)^2}{(1-\lambda+\lambda\omega(y))^3}\right]^{\frac{2}{3}}\mathop{\mathbb{E}}_{y\sim Q}\left[\omega(y)^2\right]^{\frac{1}{3}} \\ &= \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)^2}{(1-\lambda+\lambda\omega(y))^3}\right]^{\frac{2}{3}} I_2(P\|Q)^{\frac{1}{3}}. \end{aligned} \tag{P.11}$$

Putting together Equation (P.10) and Equation (P.11), we have:

$$\mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)^2}{(1-\lambda+\lambda\omega(y))^3}\right] \geqslant \mathop{\mathbb{E}}_{y\sim Q}\left[\frac{\omega(y)^2}{(1-\lambda+\lambda\omega(y))^2}\right]^{\frac{3}{2}} I_2(P\|Q)^{-\frac{1}{2}} \geqslant I_2(P\|Q)^{-2}.$$

□

Lemma C.7. *Let $h(\eta) = \eta\,\mathbb{E}_{y\sim Q}[\omega_\eta(y)^2]$ and $\widehat{h}(\eta) = \frac{\eta^2}{n}\sum_{i\in[n]}\omega_\eta(y_i)^2$. Then, $n\widehat{h}(\eta)$ is a self-bounding function. Therefore,*

for every $\eta\in[0,1]$ *it holds that:*

$$\Pr_{y_i\overset{iid}{\sim}Q}\left(\widehat{h}(\eta)-h(\eta)\geqslant\epsilon\right)\leqslant\exp\left(\frac{-\epsilon^2 n}{2\left(h(\eta)+\frac{\epsilon}{3}\right)}\right)\quad \textit{with}\quad \epsilon>0, \tag{3}$$

$$\Pr_{y_i\overset{iid}{\sim}Q}\left(h(\eta)-\widehat{h}(\eta)\geqslant\epsilon\right)\leqslant\exp\left(\frac{-\epsilon^2 n}{2h(\eta)}\right)\quad \textit{with}\quad 0<\epsilon<h(\eta). \tag{4}$$

Proof. We consider the definition of self-bounding function provided in (Boucheron et al., 2009, Definition 1). We denote with $n\widehat{h}^{k,z}(\eta)$ the function obtained from $n\widehat{h}(\eta)$ by replacing $\omega(y_k)$ with $z\geqslant 0$. We show that $n\widehat{h}(\eta)$ satisfies both conditions:

$$n\widehat{h}(\eta)-n\widehat{h}^{k,z}(\eta)=\eta^2\left(\omega_\eta(y_k)^2-z^2\right)\leqslant\eta^2\omega_\eta(y_k)^2\leqslant 1,$$
$$\sum_{k\in[n]}\left(n\widehat{h}(\eta)-n\widehat{h}^{k,z}(\eta)\right)^2=\sum_{k\in[n]}\left(\omega_\eta(y_k)^2-z^2\right)^2\leqslant\sum_{k\in[n]}\left(\eta^2\omega_\eta(y_k)^2\right)^2\leqslant\sum_{k\in[n]}\eta^2\omega_\eta(y_k)^2=n\widehat{h}(\eta).$$

having observed that $\eta\omega_\eta(y_k)\leqslant 1$. By applying the concentration inequalities for the self-bounding functions (Boucheron et al., 2009), we obtain that for every $\eta\in[0,1]$ and $\epsilon>0$ it holds that:

$$\Pr_{y_i\overset{iid}{\sim}Q}\left(\widehat{h}(\eta)-h(\eta)\geqslant\epsilon\right)\leqslant\exp\left(\frac{-\epsilon^2 n}{2\left(h(\eta)+\frac{\epsilon}{3}\right)}\right).$$

Similarly, for every $\eta\in[0,1]$ and $0<\epsilon<h(\eta)$ it holds that:

$$\Pr_{y_i\overset{iid}{\sim}Q}\left(h(\eta)-\widehat{h}(\eta)\geqslant\epsilon\right)\leqslant\exp\left(\frac{-\epsilon^2 n}{2h(\eta)}\right).$$

□

Lemma C.8. *Let* $\eta^\dagger$ *be the solution of* $h(\eta^\dagger)=\frac{2\log\frac{1}{\delta}}{3\sqrt{n}}$ *and* $\widehat{\eta}$ *be the solution of* $\widehat{h}(\widehat{\eta})=\frac{2\log\frac{1}{\delta}}{3\sqrt{n}}$*. Then, for any* $\delta\in(0,1)$*, with probability at least* $1-\delta$ *it holds that:*

$$\frac{1}{2}\leqslant\frac{\widehat{\eta}}{\eta^\dagger}\leqslant\sqrt{2}\quad\textit{and}\quad\frac{1}{2}\leqslant\frac{\widehat{\lambda}}{\lambda^\dagger}\leqslant\sqrt{2},$$

for $n\geqslant\max\left\{544I_2(P\|Q)^{12}\left(\frac{\log\frac{2}{\delta}}{\log\frac{1}{\delta}}\right)^2,\frac{4096(I_3(P\|Q)-I_2(P\|Q))^4\left(\log\frac{1}{\delta}\right)^2}{9I_2(P\|Q)^6}\right\}$.

Proof. 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:

$$\begin{aligned}\widehat{h}(\widehat{\eta})-\widehat{h}(\eta^\dagger)&\geqslant\widehat{h}((1+\epsilon)\eta^\dagger)-\widehat{h}(\eta^\dagger)\\&=\widehat{h}((1+\epsilon)\eta^\dagger)-\widehat{h}(\eta^\dagger)\pm h(\eta^\dagger)\pm h((1+\epsilon)\eta^\dagger)\\&=\widehat{h}((1+\epsilon)\eta^\dagger)-h((1+\epsilon)\eta^\dagger)+h(\eta^\dagger)-\widehat{h}(\eta^\dagger)+h((1+\epsilon)\eta^\dagger)-h(\eta^\dagger)\\&\geqslant\widehat{h}((1+\epsilon)\eta^\dagger)-h((1+\epsilon)\eta^\dagger)+h(\eta^\dagger)-\widehat{h}(\eta^\dagger)+2I_2(P\|Q)^{-2}\epsilon(\eta^\dagger)^2,\end{aligned}$$

where the last inequality follows from Lemma C.6 having applied:

$$h((1+\epsilon)\eta^\dagger)-h(\eta^\dagger)\geqslant I_2(P\|Q)^{-2}\left((1+\epsilon)^2-1\right)(\eta^\dagger)^2=I_2(P\|Q)^{-2}(2+\epsilon)\epsilon(\eta^\dagger)^2\geqslant 2I_2(P\|Q)^{-2}\epsilon(\eta^\dagger)^2.$$

Recalling that $\widehat{h}(\widehat{\eta})=h(\eta^\dagger)$, the condition can be further simplified into $h((1+\epsilon)\eta^\dagger)-\widehat{h}((1+\epsilon)\eta^\dagger)\geqslant 2I_2(P\|Q)^{-2}\epsilon(\eta^\dagger)^2$. Symmetrically, under the sub-event $\{\widehat{\eta}<(1-\epsilon)\eta^\dagger\}$ we have:

$$\begin{aligned}\widehat{h}(\widehat{\eta})-\widehat{h}(\eta^\dagger)&\leqslant\widehat{h}((1-\epsilon)\eta^\dagger)-\widehat{h}(\eta^\dagger)\\&=\widehat{h}((1-\epsilon)\eta^\dagger)-\widehat{h}(\eta^\dagger)\pm h(\eta^\dagger)\pm h((1-\epsilon)\eta^\dagger)\\&=\widehat{h}((1-\epsilon)\eta^\dagger)-h((1-\epsilon)\eta^\dagger)+h(\eta^\dagger)-\widehat{h}(\eta^\dagger)+h((1-\epsilon)\eta^\dagger)-h(\eta^\dagger)\\&\leqslant\widehat{h}((1-\epsilon)\eta^\dagger)-h((1-\epsilon)\eta^\dagger)+h(\eta^\dagger)-\widehat{h}(\eta^\dagger)-I_2(P\|Q)^{-2}\left(1-(1-\epsilon)^2\right)(\eta^\dagger)^2,\end{aligned}$$

that can be simplified, as before, into the condition $\widehat{h}((1-\epsilon)\eta^\dagger)-h((1-\epsilon)\eta^\dagger)\geqslant I_2(P\|Q)^{-2}\epsilon(\eta^\dagger)^2$ since $1-(1-\epsilon)^2=\epsilon(2-\epsilon)\geqslant\epsilon$

being $\epsilon < 1$. Thus, we have:

$$\begin{aligned}\Pr_{y\overset{\text{iid}}{\sim}Q}\left(\left|\frac{\widehat{\eta}}{\eta^\dagger}-1\right|>\epsilon\right) &= \Pr_{y\overset{\text{iid}}{\sim}Q}\left(\widehat{\eta}>(1+\epsilon)\eta^\dagger\right)+\Pr_{y\overset{\text{iid}}{\sim}Q}\left(\widehat{\eta}<(1-\epsilon)\eta^\dagger\right)\\ &\leqslant \Pr_{y\overset{\text{iid}}{\sim}Q}\left(h((1+\epsilon)\eta^\dagger)-\widehat{h}((1+\epsilon)\eta^\dagger)\geqslant 2I_2(P\|Q)^{-2}\epsilon(\eta^\dagger)^2\right)\\ &\quad+\Pr_{y\overset{\text{iid}}{\sim}Q}\left(\widehat{h}((1-\epsilon)\eta^\dagger)-h((1-\epsilon)\eta^\dagger)\geqslant I_2(P\|Q)^{-2}\epsilon(\eta^\dagger)^2\right).\end{aligned}$$

First of all, we observe that $h((1+\epsilon)\eta^\dagger)=(1+\epsilon)^2(\eta^\dagger)^2\,\mathbb{E}_{y\sim Q}[\omega_{(1+\epsilon)\eta^\dagger}(y)^2]\leqslant 4(\eta^\dagger)^2 I_2(P\|Q)$. Now, recalling that function h is self-bounding as proved in Lemma C.7, we have by Equation (4):

$$\begin{aligned}\Pr\left(h((1+\epsilon)\eta^\dagger)-\widehat{h}((1+\epsilon)\eta^\dagger)\geqslant 2I_2(P\|Q)^{-2}\epsilon(\eta^\dagger)^2\right) &\leqslant \exp\left(\frac{-4I_2(P\|Q)^{-4}\epsilon^2(\eta^\dagger)^4 n}{2h((1+\epsilon)\eta^\dagger)}\right)\\ &\leqslant \exp\left(\frac{-4I_2(P\|Q)^{-4}\epsilon^2(\eta^\dagger)^4 n}{8(\eta^\dagger)^2 I_2(P\|Q)}\right)\\ &= \exp\left(\frac{-\epsilon^2(\eta^\dagger)^2 n}{2I_2(P\|Q)^5}\right),\end{aligned}$$

provided that $2I_2(P\|Q)^{-2}\epsilon(\eta^\dagger)^2\leqslant h((1+\epsilon)\eta^\dagger)$, that is fulfilled for every $\epsilon\in[0,1]$. Indeed, recalling that $h((1+\epsilon)\eta^\dagger)=(1+\epsilon)^2(\eta^\dagger)^2\,\mathbb{E}_{y\sim Q}[\omega_{(1+\epsilon)\eta^\dagger}(y)^2]\geqslant(1+\epsilon)^2(\eta^\dagger)^2 I_2(P\|Q)^{-2}$ (from Equation (P.10)), we have that $2I_2(P\|Q)^{-2}\epsilon(\eta^\dagger)^2\leqslant(1+\epsilon)^2(\eta^\dagger)^2 I_2(P\|Q)^{-2}$ is fulfilled for every $\epsilon\in[0,1]$. Similarly, by Equation (3) and recalling that $h((1-\epsilon)\eta^\dagger)\leqslant h(\eta^\dagger)\leqslant(\eta^\dagger)^2 I_2(P\|Q)$, we have:

$$\begin{aligned}\Pr\left(\widehat{h}((1-\epsilon)\eta^\dagger)-h((1-\epsilon)\eta^\dagger)\geqslant I_2(P\|Q)^{-2}\epsilon(\eta^\dagger)^2\right) &\leqslant \exp\left(\frac{-I_2(P\|Q)^{-4}\epsilon^2(\eta^\dagger)^4 n}{2\left(h((1-\epsilon)\eta^\dagger)+\frac{1}{3}I_2(P\|Q)^{-2}\epsilon(\eta^\dagger)^2\right)}\right)\\ &\leqslant \exp\left(\frac{-I_2(P\|Q)^{-4}\epsilon^2(\eta^\dagger)^4 n}{2(\eta^\dagger)^2 I_2(P\|Q)+\frac{2}{3}I_2(P\|Q)^{-2}\epsilon(\eta^\dagger)^2}\right)\\ &\quad \exp\left(\frac{-3\epsilon^2(\eta^\dagger)^2 n}{8I_2(P\|Q)^5}\right),\end{aligned}$$

having crudely bounded $I_2(P\|Q)^{-2}\epsilon\leqslant I_2(P\|Q)$. Putting these inequalities together, we obtain:

$$\Pr\left(\left|\frac{\widehat{\eta}}{\eta^\dagger}-1\right|>\epsilon\right)\leqslant\exp\left(\frac{-\epsilon^2(\eta^\dagger)^2 n}{2I_2(P\|Q)^5}\right)+\exp\left(\frac{-3\epsilon^2(\eta^\dagger)^2 n}{2I_2(P\|Q)^5}\right)\leqslant 2\exp\left(\frac{-3\epsilon^2(\eta^\dagger)^2 n}{8I_2(P\|Q)^5}\right),$$

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

$$\left|\frac{\widehat{\eta}}{\eta^\dagger}-1\right|\leqslant\sqrt{\frac{8I_2(P\|Q)^5\log\frac{2}{\delta}}{3n(\eta^\dagger)^2}}.$$

Under Lemma C.5, we know that $\eta^\dagger\geqslant\sqrt{\frac{2\log\frac{1}{\delta}}{3I_2(P\|Q)\sqrt{n}}}$. From which we have:

$$\left|\frac{\widehat{\eta}}{\eta^\dagger}-1\right|\leqslant\sqrt{\frac{4I_2(P\|Q)^6\log\frac{2}{\delta}}{\sqrt{n}\log\frac{1}{\delta}}}.$$

Simple calculations allow to conclude that $\frac{1}{2}\leqslant\frac{\widehat{\eta}}{\eta^\dagger}\leqslant\sqrt{2}$ for $n\geqslant 544 I_2(P\|Q)^{12}\left(\frac{\log\frac{2}{\delta}}{\log\frac{1}{\delta}}\right)^2$. □

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

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

Proof. Let us start observing that if we substitute a value of λ that is proportional to λ_2^* into Equation (P.9), we are able to provide the

following bound for $\beta > 0$:

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

Now, we provide sufficient conditions so that $\frac{1}{2}\lambda_2^* \leqslant \widehat{\lambda} \leqslant 2\lambda_2^*$. First of all, we know from Lemma C.5 that for sufficiently large n we have $1 \leqslant \frac{\lambda^\dagger}{\lambda_2^*} \leqslant \sqrt{2}$. Second, from Lemma C.7, we know that for sufficiently large n and with probability at least $1-\delta$, we have $\frac{1}{2} \leqslant \frac{\widehat{\lambda}}{\lambda^\dagger} \leqslant \sqrt{2}$. Thus, putting together these results we enforce $\frac{1}{2}\lambda_2^* \leqslant \widehat{\lambda} \leqslant 2\lambda_2^*$. Therefore, it holds with probability at least $1-2\delta$ and sufficiently large n that:

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

□

Corollary C.2. *Let $P, Q \in \mathscr{P}(\mathcal{Y})$ two probability distributions such that $P \ll Q$. Let $\{y_i\}_{i\in[n]}$ sampled independently from Q. For every $\delta \in (0,1)$, let*

$$\lambda^\ddagger = \sqrt{\frac{\log\frac{1}{\delta}}{n}}$$

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

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

Proof. The result is simply obtained by substituting $\lambda^\ddagger$ into Equation (P.9). □

D. Bound Comparison and Optimization

In this appendix, we provide a comparison between the bounds of Lemma C.3 and Theorem 5.1 and show how to numerically optimize the former. For the sake of simplicity, we restrict our attention to $\alpha = 2$ and we denote with $B^{**}(\lambda)$ the bound of Lemma C.3, with λ^{**} its global minimum, with $B^*(\lambda)$ the bound of Theorem 5.1, and with λ^* its global mimimum.

$B^{**}(\lambda)$ displays a pretty intricate dependence on λ that is not easy to optimize. As we can notice from Figure 2, the bound based on the values of its terms admits either one or two local minima. In any case $\lambda = 1$ is a value of interest, leading to a bound of the form:

$$\widehat{\mu}_1 - \mu \leqslant \|f\|_\infty\sqrt{\frac{2\log\frac{1}{\delta}}{n}} + \frac{2\|f\|_\infty\log\frac{1}{\delta}}{3n} + \|f\|_\infty\sqrt{J_2(P\|Q)}.$$

In such a case, we are replacing the importance weight with the value of 1 and we are estimating the mean under the target distribution with the mean of the behavioral distribution, paying the whole bias $\sqrt{J_2(P\|Q)} = \sqrt{I_2(P\|Q)-1}$. Clearly, this circumstance is convenient only when n is sufficiently small.

The bound of Theorem 5.1 B^* is looser compared with that of Lemma C.3 B^{**}. We can see in Figure 3 that bound of B^* is convex and yeilds an optimal value of λ^* that is smaller compared to the optimal value λ^{**} of B^{**}.

D.1. Numerical Optimization of the Bound of Lemma C.3

We now discuss how to find the global minimum of the bound presented in Lemma C.3 $B^{**}(\lambda)$. First of all, we observe that $B^{**}(\lambda)$ is continuously differentiable in λ:

$$\frac{\partial B^{**}(\lambda)}{\partial\lambda} = \sqrt{(I_2(P\|Q)-1)((1-\lambda)I_2(P\|Q)+\lambda)} - \frac{2\log\frac{1}{\delta}}{3n\lambda^2} - \frac{(I_2(P\|Q)-1)\left(\sqrt{2\log\frac{1}{\delta}}+\lambda\sqrt{(I_2(P\|Q)-1)n}\right)}{2\sqrt{n((1-\lambda)I_2(P\|Q)+\lambda)}}.$$

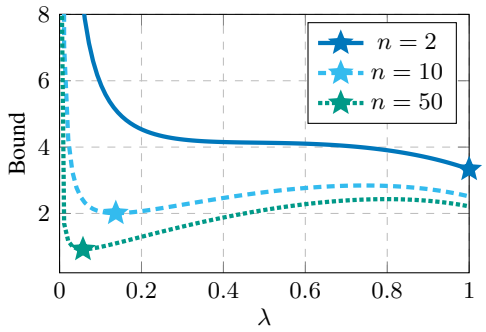

Figure 2. The bound of Lemma C.3 for $\alpha=2$, $I_2(P\|Q)=5$, $\delta=e^{-1}$, and $n\in\{2,10,50\}$. The minima are highlighted with the star.

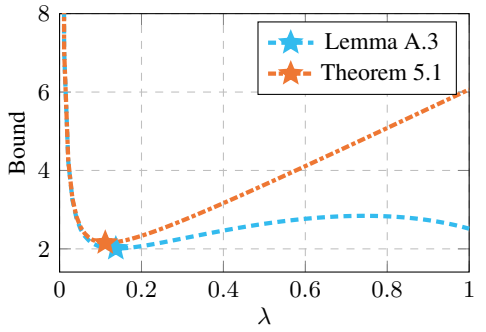

Figure 3. Comparison between the bounds of Lemma C.3 and Theorem 5.1 for $\alpha=2$, $I_2(P\|Q)=5$, $\delta=e^{-1}$, and $n=10$. The minima are highlighted with the star.

We start proving that $\frac{\partial B^{**}(\lambda)}{\partial\lambda}$ is a strictly concave function of λ:

$$\frac{\partial^2}{\partial\lambda^2}\left(\frac{\partial B^{**}(\lambda)}{\partial\lambda}\right)=\frac{\partial^3 B^{**}(\lambda)}{\partial\lambda^3}=-\frac{4\log\frac{1}{\delta}}{n\lambda^4}-\frac{3(I_2(P\|Q-1)^{7/2}\lambda}{8((1-\lambda)I_2(P\|Q)+\lambda)^{5/2}}$$
$$-\frac{3(I_2(P\|Q)-1)^{5/2}}{4((1-\lambda)I_2(P\|Q)+\lambda)^{3/2}}-\frac{3(I_2(P\|Q)-1)^3\sqrt{\log\frac{1}{\delta}}}{4\sqrt{2n}((1-\lambda)I_2(P\|Q)+\lambda)^{5/2}}<0.$$

We now prove that $\frac{\partial B^{**}(\lambda)}{\partial\lambda}$ admits at most two roots. By contradiction, suppose $\frac{\partial B^{**}(\lambda)}{\partial\lambda}$ admits three roots $\lambda_1<\lambda_2<\lambda_3$. By Rolle's theorem, there must exist $\lambda_1<\lambda_{12}<\lambda_2$ and $\lambda_2<\lambda_{23}<\lambda_3$ such that $\frac{\partial^2 B^{**}(\lambda)}{\partial\lambda^2}(\lambda_{12})=\frac{\partial^2 B^{**}(\lambda)}{\partial\lambda^2}(\lambda_{23})=0$. Again, by Rolle's theorem, there must exist $\lambda_{12}<\lambda_{1223}<\lambda_{23}$ such that $\frac{\partial^3 B^{**}(\lambda)}{\partial\lambda^3}(\lambda_{1223})=0$, which is a contradiction being $\frac{\partial B^{**}(\lambda)}{\partial\lambda}$ concave. Thus we consider three cases:

- $\frac{\partial B^{**}(\lambda)}{\partial\lambda}$ admits no roots. It follows that the global minimum of B^{**} is on the border $\{0,1\}$. Since $\lim_{\lambda\to0^+}B^{**}(\lambda)=\infty$, the minimum is in $\lambda^{**}=1$.
- $\frac{\partial B^{**}(\lambda)}{\partial\lambda}$ admits one root. It is simple to prove that for sufficiently large λ (possibly larger than 1, but this does not matter of the sake for the function study) we have $\frac{\partial B^{**}(\lambda)}{\partial\lambda}<0$. Being also $\lim_{\lambda\to0^+}\frac{\partial B^{**}(\lambda)}{\partial\lambda}=-\infty$, we conclude that the root must be a saddle point and, consequently, $\lambda^{**}=1$.
- $\frac{\partial B^{**}(\lambda)}{\partial\lambda}$ admits two roots $\lambda_1<\lambda_2$. Thus, there must exist $\lambda_1<\lambda_{12}<\lambda_2$ such that $\frac{\partial^2 B^{**}(\lambda)}{\partial\lambda^2}(\lambda_{12})=0$. Since $\frac{\partial^2 B^{**}(\lambda)}{\partial\lambda^2}$ is non-increasing, being $\frac{\partial B^{**}(\lambda)}{\partial\lambda}$ concave, it must be that $\frac{\partial^2 B^{**}(\lambda)}{\partial\lambda^2}(\lambda_1)>0$ and $\frac{\partial^2 B^{**}(\lambda)}{\partial\lambda^2}(\lambda_2)<0$. Thus, λ_1 is a local minimum and λ_2 a local maximum. It follows that $\lambda^{**}\in\arg\min_{\lambda\in\{\lambda_1,1\}}B^{**}(\lambda)$.

Thus, based on the function study, it suffices to find numerically the smallest root λ_1 (whenever it exists) of $\frac{\partial B^{**}(\lambda)}{\partial\lambda}$ and

Algorithm 1 Root finding for bound B^{**} of Lemma C.3

Compute the bound derivative $\frac{\partial B^{**}(\lambda)}{\partial \lambda}$
Apply Newton's method with λ^* as initial guess obtaining λ_1 as numerical root (if exists)
if Newton's method failed to converge **or** $B^{**}(\lambda_1) < B(1)$ **then**
  return 1
else
  return λ_1
end if

compare its bound value $B^{**}(\lambda_1)$ with $B^{**}(1)$. This task can be carried out using numerical root finding, e.g., Newton's method, using as initial guess 0 or λ^*, having observed that in the optimal correction parameter λ^* of the simplified bound B^* the derivative $\frac{\partial B^{**}(\lambda)}{\partial \lambda}$ is negative. The procedure is summarized in Algorithm 1

E. Bias^2 + Variance Minimization

In this appendix, we discuss the effect of employing the bounds on bias and variance we have derived when our goal consists in minimizing the MSE (Bias^2 + Variance) instead of a high-probability deviation inequality. By using Lemma 5.1 and Lemma 5.2, we have:

$$\text{MSE} \leqslant \|f\|_\infty^2 \lambda^{2(\alpha-1)} I_\alpha(P\|Q)^2 + \frac{\|f\|_\infty^2}{n\lambda^{2-\alpha}} I_\alpha(P\|Q). \tag{5}$$

We find the unique stationary point in λ, by vanishing the derivative:

$$\lambda_\alpha^{\S} = \left(\frac{2-\alpha}{2(\alpha-1) I_\alpha(P\|Q) n} \right)^{\frac{1}{\alpha}}.$$

Since, in general, Equation (5) is non-convex in λ, it might be the case that the minimum lies in the extremes $\{0, 1\}$. In particular, for the relevant case $\alpha = 2$, only the bias term depend on λ, suggesting $\lambda_2^{\S} = 0$. This is explained by the fact that the bound on the variance is independent from λ as it was meant to be employed in a high-probability concentration inequality rather than in an MSE bound.