Entropy minimizing distributions are worst-case optimal importance proposals

Frédéric Cérou, Patrick Héas, Mathias Rousset

Entropy minimizing distributions are worst-case optimal importance proposals

Frédéric Cérou[1], Patrick Héas[1], and Mathias Rousset[1]

[1]IRMAR and Inria, University of Rennes, France.

Abstract

Importance sampling of target probability distributions belonging to a given convex class is considered. Motivated by previous results, the cost of importance sampling is quantified using the relative entropy of the target with respect to proposal distributions. Using a reference measure as a reference for cost, we prove under some general conditions that the worst-case optimal proposal is precisely given by the distribution minimizing entropy with respect to the reference within the considered convex class of distributions. The latter conditions are in particular satisfied when the convex class is defined using a push-forward map defining atomless conditional measures. Applications in which the optimal proposal is Gibbsian and can be practically sampled using Monte Carlo methods are discussed.

Contents

1 Introduction

Importance Sampling (IS) is a generic Monte Carlo methodology that aims at computing averages with respect to a given probability distribution $\eta \in \mathcal{P}(E)$ in a state space E, usually called *the target distribution*, by using weighted samples distributed according to a different distribution $\mu \in \mathcal{P}(E)$, usually called the *proposal distribution*.

A a general concept, IS is at the basis of most Monte Carlo strategies ever since its introduction in computational statistical physics in the early fifties. The non-expert reader may consult the second chapter of the monograph [11] or the review paper [15] as an introduction. One well-known problem is the lack of robustness in the choice of the proposal distribution which leads to the degeneracy of the importance weights, especially in high dimension (see *e.g.* [1]). Motivated in part by this issue, a considerable amount of sophisticated strategies incorporating IS have been developed. Broadly speaking, two type of ideas have emerged in all trends. First, one can smooth the sampling task by considering a pre-defined family or flow of targets and then perform sampling sequentially by starting with the easiest ones. Second, on may try to optimize the choice of the proposal, using some information on the model, or previous sampling attempts. These ideas are developed both in Sequential Monte Carlo (a.k.a. Particle Filters) methods in which the proposal itself is known up to a normalization and sampled iteratively (classical papers include [13, 6]), or in methods relying on a proposals in a parametric family with some adaptive features (mainly by optimization of the proposal based on empirical estimation of the 'cross-entropy' of the target [14, 8]).

All the above mentioned methods do suggest the following question: what is the optimal proposal associated to a given target distribution, or to a given subset of target distributions ?

In many applications, the target η is given through a non-normalized density with respect to a reference distribution $\pi \in \mathcal{P}(E)$; π being very easy to sample. The present work is motivated by the following additional context: the evaluation of the density defining η is computationally very expensive, but the user knows *a priori* that η belongs to a known *convex* class of 'admissible' distributions denoted $\mathcal{C} \subset \mathcal{P}(E)$, which is described in a much more simple way. For instance, $\mathcal{C}$ may be rigorously given by bounds on the considered model. Another scenario may occur when preliminary attempts of sampling the target η have been performed, yielding a confidence set $\mathcal{C}$ to which the true target belongs with a very high probability.

We will address and give a quite generic answer to the following problem: which importance proposal μ is (worst-case) optimal when the set of admissible target distributions is $\mathcal{C}$? The results will be stated for a subset of admissible target distributions $\mathcal{A} \subset \mathcal{C}$ with appropriate properties, but we will restrict – without significant loss of generality – in the present introduction to the case $\mathcal{A} = \mathcal{C}$ for expository purpose .

The first choice we have to make is the choice of the *cost* of performing importance sampling. A way to do that is through the *required sample size* N^* of i.i.d. μ-distributed variables. It can be defined as the sample size required to yield at least a $\delta > 0$ accuracy with probability $1 - p_\alpha$ when estimating properly normalized test functions; δ, p_α being given. Instead of using the classical Chebyshev lower bound to estimate N^* using variance, we will rather use the relative entropy (Kullback-Leibler divergence):

$$\ln N^* \simeq \mathrm{Ent}(\eta \mid \mu). \tag{1.1}$$

As will be discussed in Section 5, some *lower and upper* bounds estimates are given in [3] which show that, under some uniform tail conditions on the density η/μ, (1.1) is a rigorous equivalent when $\mathrm{Ent}(\eta \mid \mu)$ becomes large. This is not the case of variance, as will be demonstrated in Section 5.2. Some care is needed however with this argument: the estimates are not sharp and quite inaccurate for non-asymptotic practical purpose. In this paper, we will reformulate and comment the results of [3] in Section 5, and then simply proceed

with our analysis using relative entropy as a *definition* of the logarithmic cost of importance sampling. Note that relative entropy is also extensively used as a cost function in the so-called *cross-entropy or adaptive importance sampling methods* ([14, 8]) in order to optimize importance proposal distributions, see Section 4.3. Recent works have also questioned the use of variance in practical empirical estimations (the so-called Effective Sample Size) of the divergence between the target and the proposal, proposing to consider other Rényi entropies (see [9, 12, 10], and Section A for comments on Rényi entropies).

Now that we have a cost functional to compare importance sampling between various proposals, one must deal with the problem that worst-case costs are usually infinite. We simply solve this issue by performing a comparison of the log-cost of importance sampling between: i) a proposal μ, and ii) the reference proposal π. If $\mathcal{C} \subset \mathcal{P}(E)$ is a given convex set of admissible target distributions, one is led to define the *worst-case logarithmic cost* – of importance sampling with μ as compared to importance sampling with π – by the quantity:

$$\mathrm{WLC}_h(\mu \mid \pi) \stackrel{\mathrm{def}}{=} \sup_{\eta \in \mathcal{C},\, \mathrm{Ent}(\eta|\pi) \leq h} \mathrm{Ent}(\eta \mid \mu) - h. \tag{1.2}$$

The quantity (1.2) is the logarithm of the (normalized) worst-case cost for importance sampling with μ, for targets with a given maximal reference log-cost.

We will first prove in this paper (Theorem 3.1), that, under some specific assumptions, the worst-case optimal proposal distribution is unique and is given by the classical entropy minimizing distribution associated with $\mathcal{C}$:

$$\mu_* = \arg\min_{\mu \in \mathcal{P}(E)} \mathrm{WLC}_h(\mu \mid \pi) = \arg\min_{\mu \in \mathcal{C}} \mathrm{Ent}(\mu \mid \pi). \tag{1.3}$$

We will show more precisely that the optimal worst-case log-cost is given by

$$\mathrm{WLC}_h(\mu_* \mid \pi) + h = h - \mathrm{Ent}(\mu_* \mid \pi) \quad \geq 0,$$

and that the difference with an other proposal is given by

$$\mathrm{WLC}_h(\mu \mid \pi) - \mathrm{WLC}_h(\mu_* \mid \pi) \geq \mathrm{Ent}(\mu \mid \mu_*).$$

The main sufficient condition ensuring these results is the following: any strict half-space (defined by measurable functions) containing the

optimizer $\mu_* \in \mathcal{C}$ must also contain a distribution $\eta \in \mathcal{C}$ with prescribed relative entropy $\mathrm{Ent}(\eta \mid \pi) = h$ (at least up to an arbitrary precision).

This abstract condition will be made more concrete in Theorem 3.3, where a sufficient setting is proposed in the case where the set admissible targets is defined through a push-forward map T:

$$\mathcal{C} = \{\eta : T\sharp\eta \in \mathcal{C}_T \subset \mathcal{P}(F)\},$$

$T\sharp\eta$ denoting the push-forward by any measurable map $T : E \to F$ and $\mathcal{C}_T$ denoting an arbitrary convex set. A sufficient condition (H_T) ensuring the abstract condition (H) and the main results stated above is then: i) the conditional distribution $\pi(\ . \mid T = t)$ has an atomless distribution $T\sharp\pi(\mathrm{d}t)$-almost everywhere, and ii) $\mathcal{C}$ contains distributions defined as indicator densities with respect to μ_*. A counterexample on a two atom discrete space is provided showing the necessity of the atomless assumption.

An important more practical example is for T vector valued and $\mathcal{C} = \{\eta : \eta(T) \in C\}$ for C convex. It will be discussed in Section 4. In that case the optimal proposal belongs to the Gibbs (canonical) exponential family:

$$\mu_* \propto \exp(\langle \beta_*, T \rangle)\mathrm{d}\pi;$$

and can be simulated using some Monte Carlo procedure (*e.g.* Sequential Monte Carlo, or Direct), albeit in a cheapest way as compared to the target η.

Finally, it is interesting to remark that our main min-max characterization theorem is similar, albeit different, from the classical Pythagorean theorem for relative entropy in information geometry. We will recall in Section 3 that the latter is equivalent to the min-max property:

$$\mu_* = \operatorname*{arg\,min}_{\mu \in \mathcal{P}(E)} \sup_{\eta \in \mathcal{C}} \mathrm{Ent}(\eta \mid \mu) - \mathrm{Ent}(\eta \mid \pi),$$

which holds in general, without specific assumption. It will be discussed in Section 2 why the latter is not very satisfactory for a practical interpretation in terms of importance sampling. The main problem is that the relative log-cost $\mathrm{Ent}(\eta \mid \mu) - \mathrm{Ent}(\eta \mid \pi)$ quantifies the *relative improvement* of importance sampling by μ as compared π, and it turns out that for the optimal proposal $\mu = \mu_*$, the worst-case improvement

is always attained for the 'cheapest' trivial target $\eta = \mu_*$; yet, in practice, one is not interested in improving the cheapest targets of the admissible set $\mathcal{C}$. Our main results might be interpreted as a variant of this min-max formulation of the Pythagorean theorem in which the supremum has to be reached by target distributions with large relative entropy $\mathrm{Ent}(\eta \mid \pi) \gg 1$, which are relevant in importance sampling.

The paper is structured as follows. We will recall in Section 2 the definition and the main properties (*e.g.* the Pythagorean theorem) of the distribution μ_* that minimizes entropy relative to a reference π over a convex subset. In Section 3, we will then state and prove the two main theorems of this work, Theorem 3.1 and Theorem 3.3, with a counter-example for discrete state spaces. Some comments on application to Gibbs exponential families will be presented in Section 4. Finally, a reformulation of, and some comments on, the results of the reference [3] on the sample size required for importance sampling will be done in Section 5.

Notation

$T\sharp\pi$ denotes the push-forward (measure image) of π by the map T. $\pi(\varphi) = \int \varphi \mathrm{d}\pi$ denotes integration of test functions. η/π denotes the Radon-Nikodym derivative between two non-negative measures with domination relation $\eta \ll \pi$.

2 Entropy minimizing distributions

In this section, we recall some basic facts about the entropy minimizing distribution μ_* associated with a given convex subset $\mathcal{C}$ of probability distributions and a reference probability π.

2.1 Definition and Pythagorean theorem

Let (E, π) denotes the pair given by a (standard Borel) state space E endowed with reference probability distribution π. Let $\mathcal{C} \subset \mathcal{P}(E)$ be a convex subset. For simplicity, we assume there exists an *entropy minimizing distribution* (with finite entropy) associated with the pair

$(\pi, \mathcal{C})$. It will be denoted

$$\mu_* \stackrel{\text{def}}{=} \arg\min_{\mu \in \mathcal{C}} \mathrm{Ent}(\mu \mid \pi). \tag{2.1}$$

By strict convexity of entropy, the latter is uniquely defined. A general notion of entropy minimizing distribution exists in information geometry, where it is called *information projection* [4].

The 'Euler-Lagrange' equation or 'first-order condition' characterizing (2.1) are known in information geometry as the *Pythagorean theorem* for relative entropy.

Theorem 2.1 (Theorem 1, [4])**.** *Let (E, π) be a probability space. Let $\mathcal{C} \subset \mathcal{P}(E)$ be convex and contain the entropy minimizing distribution denoted μ_*. The following condition on $\mu \in \mathcal{P}(E)$:*

$$\forall \eta \in \mathcal{C}, \qquad \mathrm{Ent}(\eta \mid \pi) \geq \mathrm{Ent}(\eta \mid \mu) + \mathrm{Ent}(\mu_* \mid \pi), \tag{2.2}$$

has a unique solution given by μ_*.

Remark 2.2. *μ_* is also the unique distribution satisfying the more constraining condition:*

$$\forall \eta \in \mathcal{C}, \qquad \mathrm{Ent}(\eta \mid \pi) \geq \mathrm{Ent}(\eta \mid \mu) + \mathrm{Ent}(\mu \mid \pi).$$

Remark 2.3. *The above theorem is a kind of first-order optimality condition. To see this, let us denote by D_μ the formal (Fréchet) derivative on $\mathcal{P}(E)$ for the usual addition of measures. At least in a formal sense, $D_\mu \mathrm{Ent}(\mu \mid \pi)$ is a test function, and if $\eta \in \mathcal{C}$, the difference of probability measures $\eta - \mu$ is a tangent vector pointing inside $\mathcal{C}$. Formal first-order optimality condition precisely requires that:*

$$\begin{aligned} 0 \leq (\eta - \mu)\,(D_\mu \mathrm{Ent}(\mu \mid \pi)) &= (\eta - \mu)(\ln \mu/\pi) \\ &= \mathrm{Ent}(\eta \mid \pi) - \mathrm{Ent}(\eta \mid \mu) - \mathrm{Ent}(\mu \mid \pi), \end{aligned}$$

hence the above condition.

The following direct corollary will be useful in our proofs.

Corollary 2.4. *Under the assumptions of Theorem 2.1, if $\eta \in \mathcal{C}$ satisfies $\eta\left(\left|\ln \frac{\mu_*}{\pi}\right|\right) < +\infty$, then it holds*

$$\eta(\ln \frac{\mu_*}{\pi}) \geq \mu_*(\ln \frac{\mu_*}{\pi}).$$

2.2 Min-max formulation

It is especially of interest to the present work to reformulate the above theorem as a min-max problem as follows:

Corollary 2.5. *Under the assumptions of Theorem 2.1, one has*

$$\mu_* = \arg\min_{\mu\in\mathcal{P}(E)} \sup_{\eta\in\mathcal{C}} \mathrm{Ent}(\eta \mid \mu) - \mathrm{Ent}(\eta \mid \pi).$$

Moreover, for $\mu = \mu_*$, the supremum is attained by $\eta = \mu_*$:

$$\sup_{\eta\in\mathcal{C}} \mathrm{Ent}(\eta \mid \mu_*) - \mathrm{Ent}(\eta \mid \pi) = -\mathrm{Ent}(\mu_* \mid \pi).$$

Proof. The condition (2.2) is equivalent to the condition

$$\sup_{\eta\in\mathcal{C}} \mathrm{Ent}(\eta \mid \mu) - \mathrm{Ent}(\eta \mid \pi) \leq -\mathrm{Ent}(\mu_* \mid \pi),$$

so that the statement of the Pythagorean theorem is equivalent to the two conditions: i) $\sup_{\eta\in\mathcal{C}} \mathrm{Ent}(\eta \mid \mu) - \mathrm{Ent}(\eta \mid \pi) > -\mathrm{Ent}(\mu_* \mid \pi)$ if $\mu \neq \mu_*$, and ii) $\sup_{\eta\in\mathcal{C}} \mathrm{Ent}(\eta \mid \mu_*) - \mathrm{Ent}(\eta \mid \pi) \leq -\mathrm{Ent}(\mu_* \mid \pi)$ in which the upperbound is attained for $\eta = \mu_*$. □

Admitting that we quantify the logarithm of the cost of importance sampling of a target $\eta \in \mathcal{C}$ with a proposal μ using relative entropy, Corollary 2.5 is already of interest to our sampling interpretation. Indeed let us define the *worst relative logarithmic cost* by

$$\mathrm{WRLC}_{\mathcal{C}}(\mu \mid \pi) = \sup_{\eta\in\mathcal{C}} \mathrm{Ent}(\eta \mid \mu) - \mathrm{Ent}(\eta \mid \pi)$$

, that is the logarithm of worst-case *improvement ratio* between performing importance sampling with proposal μ as compared to with reference proposal π. The Pythagorean theorem is equivalent to say that the entropy minimizing distribution μ_* is the unique optimal proposal in terms of $\mathrm{WRLC}_{\mathcal{C}}(\mu \mid \pi)$. The latter is nonetheless not fully satisfactory from a practical perspective because for $\mu = \mu_*$ the worst case target is attained by the proposal itself $\eta = \mu_*$. Unfortunately, it might happen in principle that $\eta = \mu_*$ is the unique worst-case target in the sense of the above WRLC criteria. The Pythagorean theorem thus says nothing *a priori* about the improvement of cost for those targets η with higher relative entropy which are of practical significance.

In a way, it is the purpose of the present paper to modify the min-max formulation of Corollary 2.5 in order to consider *absolute rather than relative* worst-case costs. This justifies the definition of the worst-case log-cost (1.2).

2.3 Gibbs case

We end this section with the most standard example of entropy minimizing distributions.

Example 2.6 (Gibbs exponential family)**.** *Let $T : E \to \mathbb{R}^d$ be given with finite exponential moments $\pi(\mathrm{e}^{\langle \beta, T\rangle}) < +\infty$ for all $\beta \in \mathbb{R}^d$. Assume that*

$$\mathcal{C} = \{\eta \mid \eta(T) \in C\}$$

where C is closed convex with $C \cap \mathrm{supp}(T\sharp\pi) \neq \emptyset$. Then,

$$\mu_* = \mu_{\beta^*} = \frac{1}{Z_{\beta_*}} \mathrm{e}^{\langle \beta_*, T\rangle} \mathrm{d}\pi,$$

for a unique $\beta_* \in \mathbb{R}^d$. Morever, β_* is the unique $\beta \in \mathbb{R}^d$ satisfying the first order optimality condition

$$\forall t \in C, \qquad \langle \beta, t\rangle \geq \langle \beta, \mu_\beta(T)\rangle.$$

In the case where $C = \{t_0\}$, $\mathcal{C}$ is called a linear family*, the equality case is satisfied in the Pythagorean theorem, and β_* is the unique β satisfying*

$$\mu_\beta(T) = t_0.$$

3 Main results

We now assume in this section that one wants to compare the cost of importance sampling between: i) a proposal μ, and ii) the reference proposal π.

We will use the exponential of relative entropy (rather than variance) to quantify the sample size N^* required by importance sampling of η by μ, that is $\mathrm{Ent}(\eta \mid \mu) \simeq \ln N^*$. This will be thoroughly discussed and justified in Section 5. From now on, $\mathrm{Ent}(\eta \mid \mu)$ will be called the *log-cost* of performing importance sampling of target η with proposal μ.

Denoting by $\mathcal{A} \subset \mathcal{P}(E)$ a given set of admissible target distributions, it is then natural to define the *worst-case log-cost* – of importance sampling with μ as compared to importance sampling with π – by the quantity (1.2), that we recall here:

$$\mathrm{WLC}_h(\mu \mid \pi) \stackrel{\mathrm{def}}{=} \sup_{\eta \in \mathcal{A},\, \mathrm{Ent}(\eta|\pi) \leq h} \mathrm{Ent}(\eta \mid \mu) - h. \tag{3.1}$$

The quantity (3.1) is the logarithm of the worst-case cost of importance sampling with proposal μ for targets with a given maximal reference log-cost of importance sampling with π.

Note that since μ_* is entropy minimizing, the definition of WLC_h is well-defined only for $h \geq \mathrm{Ent}(\mu_* \mid \pi)$. For $h_* = \mathrm{Ent}(\mu_* \mid \pi)$, the set defining the supremum is given by the singleton μ_*, and one trivially obtains $\mathrm{WLC}_h(\mu \mid \pi) = \mathrm{Ent}(\mu_* \mid \mu) - \mathrm{Ent}(\mu_* \mid \pi)$ which has minimum 0 for $\mu = \mu_*$.

In (3.1), the 'worst-case scenario' is defined using the subset $\{\mathrm{Ent}(\ . \mid \pi) \leq h\} \cap \mathcal{A}$ defined by the admissible target distributions with a given maximal reference log-cost of importance sampling (with the reference proposal π). As detailed in the previous section, the definition of log-cost (3.1) has to be compared with the variant of the Pythagorean theorem in Corollary (2.5). The latter involves the worst-case *relative* log-cost, where the relative log-cost is defined by the difference $\mathrm{Ent}(\eta \mid \mu) - \mathrm{Ent}(\eta \mid \pi)$, whereas in 3.1 the comparison is made with a reference worst-case value h. This ensures that the minimization of the worst-case log-cost WLC_h really focuses on the worse target distributions.

3.1 The general min-max theorem

We can then state the first of the two main theorems of this paper. The main issue is to give a general condition under which the optimal proposal distribution minimizing the worst-case log-cost WLC_h is indeed the entropy minimizing over a convex set $\mathcal{C}$ of distributions containing $\mathcal{A}$:

$$\mathcal{A} \subset \mathcal{C}.$$

In short, this condition asks that in any half-space (as defined by bounded measurable functions) containing strictly μ_*, one can find a target $\eta \in \mathcal{A}$ with log-cost $\mathrm{Ent}(\eta \mid \pi)$ arbitrarily close to the reference log-cost h; or, in other words, that μ_* and those target distributions

η with $\mathrm{Ent}(\eta \mid \pi)$ close to h cannot be strictly separated by an hyperplane.

We will in fact show a much more precise results. First we will show that the optimal worst-case log-cost obtained for $\mu = \mu_*$ is given in fact by the opposite of the minimal entropy on $\mathcal{C}$: $-\mathrm{Ent}(\mu_* \mid \pi)$. This is similar to what happens in the Pythagorean theorem of Theorem 2.1. We will also prove the inequality (3.3) below, that states that the difference between i) a worst-case log-cost for any proposal μ, and ii) the optimal worst-case log-cost for proposal μ_*, is in fact greater than the log-cost of μ itself, $\mathrm{Ent}(\mu \mid \mu_*)$. This will immediately yield the desired characterization of μ_* as optimal worst-case proposal, with a quantification of optimality given by $\mathrm{Ent}(\mu \mid \mu_*)$.

Theorem 3.1. *Let (E, π) be a standard probability state space, with $\mathcal{C} \subset \mathcal{P}(E)$ a given convex subset of probability distributions containing the unique entropy minimizing distribution $\mu_* = \arg\min_{\mathcal{C}} \mathrm{Ent}(\,.\mid \pi)$.*

Let $h \geq \mathrm{Ent}(\mu_* \mid \pi)$ be given, and assume that for all bounded measurable real-valued function $f : E \to \mathbb{R}$, and all $\varepsilon > 0$, there exists an admissible distribution $\eta_{f,h,\varepsilon} \in \mathcal{A} \subset \mathcal{C}$ such that

$$|\mathrm{Ent}(\eta_{f,h,\varepsilon} \mid \pi) - h| \leq \varepsilon, \quad \& \quad \eta_{f,h,\varepsilon}(f) \geq \mu_*(f) - \varepsilon. \tag{3.2}$$

Then one first has that:

$$\mathrm{WLC}_h(\mu_* \mid \pi) = -\mathrm{Ent}(\mu_* \mid \pi).$$

has:

$$\forall \mu \in \mathcal{P}(E), \qquad \mathrm{WLC}_h(\mu \mid \pi) \geq \mathrm{WLC}_h(\mu_* \mid \pi) + \mathrm{Ent}(\mu \mid \mu_*). \tag{3.3}$$

where in the above the log-cost is defined by (3.1).

The above result immediately implies our main finding that

$$\mu_* = \operatorname*{argmin}_{\mu \in \mathcal{P}(E)} \mathrm{WLC}_h(\mu \mid \pi),$$

and when the condition (3.3) is satisfied for any h, we also obtain that:

$$\mu_* = \operatorname*{argmin}_{\mu \in \mathcal{P}(E)} \limsup_{h \to +\infty} \mathrm{WLC}_h(\mu \mid \pi),$$

the quantity $\limsup_{h \to +\infty} \mathrm{WLC}_h$ being a natural definition of the worst-case log-cost when no constraint is applied to admissible targets.

Remark 3.2. *The main assumption* (3.2) *in Theorem 3.1 above have a nice geometric interpretation.* μ_* *is the unique distribution obtained as the intersection of two convex sets* $\mathcal{C}$ *and* $\{\mathrm{Ent}(|\,\pi) \leq h_*\}$*, where* $h_* = \mathrm{Ent}(\mu_* \mid \pi)$ *is such that the latter two are 'tangent' with each other. The condition* (3.2) *exactly asks that in* any *open*[1] *half-space of distributions that contains* μ_**, one can find an element of* $\mathcal{A}$ *with prescribed entropy.*

Proof of Theorem 3.1. First recall that $\mathrm{Ent}(\mu_* \mid \pi) \leq \inf_{\eta\in\mathcal{A}} \mathrm{Ent}(\eta \mid \pi)$

Step 1. We claim that

$$\sup_{\eta\in\mathcal{A}\bigcap\{\mathrm{Ent}(\,.\,|\pi)\leq h\}} \mathrm{Ent}(\eta \mid \mu_*) = h - \mathrm{Ent}(\mu_* \mid \pi).$$

Indeed, let $\eta \in \mathcal{A}$ be a probability such that $\mathrm{Ent}(\eta \mid \pi) < +\infty$. By the Pythagorean inequality (2.2) one has $\mathrm{Ent}(\eta \mid \mu_*) < +\infty$ and $\eta\left(|\ln \pi/\mu_*|\right) < +\infty$. One can then consider the following decomposition:

$$\mathrm{Ent}(\eta \mid \mu_*) = \mathrm{Ent}(\eta \mid \pi) + \eta(\ln \frac{\pi}{\mu_*}), \tag{3.4}$$

and (2.2) becomes $\eta(\ln \frac{\pi}{\mu_*}) \leq -\inf_{\eta\in\mathcal{A}} \mathrm{Ent}(\eta \mid \pi) \leq -\mathrm{Ent}(\mu_* \mid \pi)$. By (3.2), there exists a sequence in $\mathcal{A}$ such that $\lim_n \mathrm{Ent}(\eta_n \mid \pi) = h$ and $\lim_n \eta_n(\ln \frac{\pi}{\mu_*}) \geq -\mathrm{Ent}(\mu_* \mid \pi)$. Hence the claim.

Step 2. We claim that if $\mu \neq \mu_*$, then

$$\sup_{\eta\in\mathcal{A}\bigcap\{\mathrm{Ent}(\,.\,|\pi)\leq h\}} \mathrm{Ent}(\eta \mid \mu) > h - \mathrm{Ent}(\mu_* \mid \pi).$$

Indeed, let η with $\mathrm{Ent}(\eta \mid \pi) < +\infty$, and $\mu \neq \mu_*$ be given, and assume without lost of generality that $\mathrm{Ent}(\eta \mid \mu) < +\infty$, which implies $\eta\left(\left|\ln \frac{\mu}{\pi}\right|\right) < +\infty$. One has the decomposition

$$\mathrm{Ent}(\eta \mid \mu) = \mathrm{Ent}(\eta \mid \pi) + \eta(\ln \frac{\pi}{\mu}) \tag{3.5}$$

By (3.2), there exists a sequence such that $\mathrm{Ent}(\eta_n \mid \pi) = h$ and $\lim_n \eta_n(\ln \pi/\mu) \geq \mu_*(\ln \pi/\mu) = -\mathrm{Ent}(\mu_* \mid \pi) + \mathrm{Ent}(\mu_* \mid \mu)$. Hence the claim. □

[1] for the locally convex weak topology on finite measures making evaluation of measurable bounded function continuous. This topology is sometimes called the weak τ-topology as in [7]

3.2 A sufficient condition in the atomless case

In this section, we will show that the main assumption in Theorem 3.1 is satisfied in a quite generic context related to applications. This is Theorem 3.3 below, which is the second main theorem of this paper.

Let $T : E \to F$ be a measurable map onto a secondary (standard Borel) measurable space F. Let $\mathcal{C}_T$ denotes any convex subset of $\mathcal{P}(F)$. Let us assume that the convex set containing admissible target distributions is defined as the pull-back by T of $\mathcal{C}_T$:

$$\mathcal{C} = \{T\sharp\eta \in \mathcal{C}_T \subset \mathcal{P}(F)\}\,. \tag{3.6}$$

This is a completely generic context; any convex set $\mathcal{C}$ can be written in this way, for instance trivially is a measurable isomorphism. This corresponds to the practical situation in which the prior information on the possible targets is given by a condition on the push-forward by T.

Theorem 3.3. *Let (E, π) be a standard probability space. Let $\mathcal{C}$ be given by* (3.6) *and contains the unique associated entropy minimizing distribution. Assume that:*

1. *The conditional distributions $\pi(\ .\ |\ T = t)$ are atomless $T\sharp(\mathrm{d}t)$-almost everywhere.*
2. *The set of admissible target distributions $\mathcal{A}$ contains at least all distributions having an indicator density with respect to μ_*.*

Then for any $h \geq \mathrm{Ent}(\mu_* \mid \pi) = h_*$, any $f : E \to \mathbb{R}$ bounded measurable, and any $\varepsilon > 0$, $\mathcal{A}$ contains a distribution $\eta_{f,h,\varepsilon}$ such that (3.2) *is verified. In particular, the statement of Theorem 3.1 holds true for any $h \geq h_*$.*

Proof. We first claim that: i) the conditional distribution of entropy minimizer μ_* is the same as the conditional distribution of π:

$$\mu_*(\mathrm{d}x \mid T(x) = t) = \pi(\mathrm{d}x \mid T(x) = t), \qquad T\sharp\pi(\mathrm{d}t) - \text{a.e.},$$

and ii) the entropy minimizer on $\mathcal{C}_T$ relative to $T\sharp\pi$ is $T\sharp\mu_*$ is. Indeed, chain rule of conditional entropy reads:

$$\mathrm{Ent}(\mu \mid \pi) = \int_F \mathrm{Ent}(\mu(\ .\ |\ T = t) \mid \pi(\ .\ |\ T = t))T\sharp\mu(\mathrm{d}t) + \mathrm{Ent}(T\sharp\mu \mid T\sharp\pi);$$

for μ ranging in $\mathcal{C}$, the first term of the right hand side uniquely ($T\sharp\pi$ almost everywhere) vanish for $\mu(\ .\mid T=t)=\pi(\ .\mid T=t)$, while second term is minimized only if $T\sharp\mu$ minimizes entropy on $\mathcal{C}_T$ relative to $T\sharp\pi$.

Next, let f be an arbitrary bounded measurable and $h\geq h_*$ be given. The atomless assumption ensures that there exists a set $A_{f,h}\subset E$ satisfying for $T\sharp\mu_*$ almost all t

$$\pi(A_{f,h}\mid T=t)=\mathrm{e}^{h_*-h}\quad\text{and}\quad\pi(f\mid A_{f,h}\,\&\,T=t)\geq\pi(f\mid T=t).$$

and Lemma B.2 (based on the existence of an independent complement to T) ensures that $A_{f,h}$ can be chosen among Borel subsets.

One can then define a target distribution candidate to satisfy (3.2).

$$\eta_{f,h}(\mathrm{d}x)\stackrel{\text{def}}{=}\mu_*(\mathrm{d}x\mid A_{f,h})=\mathrm{e}^{-h_*+h}\,\mathbb{1}_{A_{f,h}}(x)\mu_*(\mathrm{d}x).$$

We can now conclude by showing that $\eta_{f,h}$ indeed satisfies (3.2) for $\varepsilon=0$:

Step 1.Since $\mu_*(\ \mid T=t)=\pi(\ \mid T=t)$, one can remark that by construction of $A_{f,h}$, one also have $\mu_*(f\mid A_{f,h}\,\&\,T=t)\geq\mu_*(f\mid T=t)$, so that it holds $\eta_{f,h}(f)\geq\mu_*(f)$.

Step 2. We remark that by construction $\pi(A_{f,h}\mid T=t)$ is independent of t so that the push-forward distribution is unchanged $T\sharp\eta_{f,h}=T\sharp\mu_*$. Moreover, since $\eta_{f,h}(\ .\mid t=T)$ has an indicator density with respect to $\mu_*(\ \mid T=t)=\pi(\ \mid T=t)$, a routine calculation yields that

$$\mathrm{Ent}(\eta_{f,h}(\ .\mid T=t)\mid\pi(\ .\mid T=t))=-\ln\eta_{f,h}(A_{f,h}\mid T=t).$$

Using again the chain rule of conditional entropy we obtain

$$\begin{aligned}\mathrm{Ent}(\eta_{f,h}\mid\pi)&=\mathrm{Ent}(T\sharp\mu_*\mid T\sharp\pi)+\int-\ln\pi(A_{f,h}\mid T=t)T\sharp\eta(\mathrm{d}t)\\&=\mathrm{Ent}(\mu_*\mid\pi)+h-h_*=h.\end{aligned}$$

□

3.3 A simple two atoms counter-example

We next provide a very simple counter-example showing that the atomless asumption is critical in Theorem 3.3. For simplicity we consider

$$\mathcal{A} = \mathcal{C} = \mathcal{P}(E),$$

so that entropy minimizing distribution is simply the reference π:

$$\mu_* = \pi, \qquad h_* = 0.$$

Similar counter-examples can be constructed in more complex cases.

The problem arises when the discrete structure is not uniform: there are atoms with very different masses. In that discrete case, the (absolute) worst-case target distribution for the proposal π is exactly a Dirac distribution on the atom with smallest probability. The worst-case optimal proposal is then the uniform discrete measure: proposing with π is sub-optimal because of the smallest atom. However, for medium values of h, the discrete uniform distribution competes with the reference π; while for smaller values of h, targets concentrated on the smallest atom are not allowed and π is optimal.

Proposition 3.4. *Let $E = \{1, 2\}$ and $\pi \in \mathcal{P}(E)$ non-degenerate with $\pi(1) \geq \pi(2) > 0$. Let us denote $\mathrm{WLC}_h(\mu \mid \pi) \stackrel{\mathrm{def}}{=} \sup_{\eta:\, \mathrm{Ent}(\eta|\pi) \leq h} \mathrm{Ent}(\eta \mid \mu)$. Remark that $\mathrm{Ent}(\delta_i \mid \pi) = -\ln \pi(i)$. If $h \in [-\ln \pi(1), -\ln \pi(2)]$, let us denote by π_h the unique distribution such that*

$$\mathrm{Ent}(\pi_h \mid \pi) = h.$$

If $h \leq -\ln \pi(1)$ then

$$\pi = \arg\min_{\mu \in \mathcal{P}(E)} \mathrm{WLC}_h(\mu \mid \pi),$$

If $h \in [-\ln \pi(1), -\ln \pi(2)]$, then

$$\pi_h = \arg\min_{\mu \in \mathcal{P}(E)} \mathrm{WLC}_h(\mu \mid \pi),$$

If $h \geq -\ln \pi(2)$ then

$$\mathrm{Unif}(\{1, 2\}) = \arg\min_{\mu \in \mathcal{P}(E)} \mathrm{WLC}_h(\mu \mid \pi)).$$

Proof. We start with the case $h \leq -\ln \pi(1) \leq -\ln \pi(2)$. We claim that the assumptions of Theorem 3.1 are satisfied. Let $f : E \to \mathbb{R}$ be given, and let $i \in \{1, 2\}$ be such that $f_i = \max f$. The map $\theta \mapsto \mathrm{Ent}(\theta\pi + (1-\theta)\delta_i \mid \pi)$ is continuous on $[0, 1]$ taking values in $[0, -\ln \pi(i)]$ so that the value $h \leq -\ln \pi(i)$ is attained; and we set $\eta_{f,h} = \theta_h \pi + (1-\theta_h)\delta_i$ where θ_h is such that $\eta_{f,h} = \pi_h$. One also has $\eta_h(f) \geq \pi(f)$ by construction and the claim follows.

The case $h \geq -\ln \pi(2)$ is quite simple. The map $\eta \mapsto \mathrm{Ent}(\eta \mid \mu)$ is strictly convex with two admissible extrema: $\eta = \delta_1$ and $\eta = \delta_2$ suprema with respective entropies $-\ln \mu(1)$ and $-\ln \mu(2)$. This implies $\mathrm{WLC}_h(\mu \mid \pi) = -\ln \min(\mu(1), \mu(2)) \geq -\ln 1/2$, this last inequality being an equality if and ony if μ is the uniform distribution. This yields the result.

In the case $h \in [-\ln \pi(1), -\ln \pi(2)]$, one also consider the continuous strictly convex map $\eta \mapsto \mathrm{Ent}(\eta \mid \mu)$ which has now two admissible extrema: i) the Dirac distribution $\eta = \delta_1$ on the one hand, and ii) π_h. The associated admissible extrema defining WLC_h are then respectively: $-\ln \mu(1)$ and h. h is the greatest so that $\mathrm{WLC}_h(\mu \mid \pi) = \mathrm{Ent}(\pi_h \mid \mu)$. Hence the result.

□

4 Gibbs exponential families and applications

In this section, we consider a slightly more practical viewpoint. We assume that the main objective is to set up an importance sampling method aiming at numerically estimate averages with respect to a given target distribution η_{true}, defined up to a normalizing constant,

$$\eta_{\mathrm{true}}(\mathrm{d}x) \propto f(x)\pi(\mathrm{d}x).$$

We first discuss Gibbs exponential families, which are the entropy minimizing proposals that arise when considering targets in a convex set defined using the push-forward by a vector-valued function T. Then, we will present an applicative scenario and discuss the relationship with the well-known *cross-entropy* method.

4.1 Gibbs exponential families

Let us now consider a vector-valued bounded measurable function defined on the considered state-space:

$$T : E \to \mathbb{R}^d.$$

For instance, T may be given by various statistics, or physical observables of special interest. We also assume that we know that averages of T with respect to the target distribution belong to a given closed convex set C of $\mathbb{R}^d$:

$$\eta_{\text{true}}(T) \in C \subset \mathbb{R}^d.$$

C is assumed to have a non void intersection with the support of $T\sharp\pi$. This defines the admissible convex set of distributions $\mathcal{C} = \{\eta : \eta(T) \in C\}$. If one assumes that the conditional distributions $\pi(\ .\ |\ T = t)$ are atomless distribution $T\sharp\pi$ almost everywhere, one can directly apply Theorem 3.3 and obtain that the entropy minimizing distribution μ_* is the worst-case optimal proposal in the sense of the log-cost (3.1).

μ_* is well-known and is given by

$$\mu_* = \frac{1}{Z_{\beta_*}} \mathrm{e}^{\langle \beta_*, T \rangle} \mathrm{d}\pi.$$

It is the unique distribution in $\mathcal{C} = \{\eta : \eta(T) \in C\}$ and in the Gibbs (a.k.a canonical) exponential family $\mu_\beta \propto \mathrm{e}^{\langle \beta, T \rangle} \mathrm{d}\pi$, $\beta \in \mathbb{R}^d$, which minimizes the entropy relative to π.

It is also interesting to remark that it is not necessary to be able to sample exactly according to μ_* in order to implement the method. Indeed, μ_* is described by T up to a normalization, so that a Sequential (or other) Monte Carlo routine (*e.g.* [6]) can be used to sample according to the latter and estimate the associated normalization. The latter can be coupled with an iterative convex minimization routine calculating β_*

4.2 An applicative context

In some practical cases, the preliminary information on the target distribution η_{true} may be given by preliminary importance sampling in the form of a weighted empirical probability distribution

$$\eta^N = \sum_{n=1}^{N} W^n \delta_{X_n},$$

defined by a sequence of random states $X_1, \ldots, X_N \in E^N$, on which the density $f(X_1), \ldots, f(X_N)$ have been previously evaluated.

Let us again consider a vector-valued bounded measurable function defined on the considered state-space: $T : E \to \mathbb{R}^d$. For instance, T may be given by various statistics, or physical observables of special interest. It may happen then that the user is able to define a confidence convex set $C_{\eta^N} \subset \mathbb{R}^d$ – which heavily depends on the precise construction of η^N – which asserts that with very high probability

$$\mathbb{P}[\eta_{\text{true}} \in \mathcal{C}_{\eta^N}] \simeq 1 \tag{4.1}$$

where one has defined the confidence set of target distributions

$$\mathcal{C}_{\eta^N} \stackrel{\text{def}}{=} \left\{ \eta \in \mathcal{P}(E) : \eta(T) \in C_{\eta^N} \subset \mathbb{R}^d \right\}.$$

One can then choose as a worst-case optimal proposal the Gibbs distribution $\mu_* \propto \mathrm{e}^{\langle \beta_*, T \rangle} \mathrm{d}\pi$ minimizing entropy with the constraint that averages of T belong to C_{η^N}. This optimal proposal can then be computed using various combination of Monte Carlo sampling and convex optimization routines. Although each ingredient (sampling and convex optimization) are well-known, such combinations and the proposed application are not standard up to our knowledge, and a detailed study is left for future work.

An interesting point consists in remarking that the proposed optimization of the proposal can be done iteratively, leading to a kind of adaptive importance sampling algorithm. This typically also happens in the so-called *cross-entropy method* that we are to briefly discuss below.

4.3 Relation to the cross-entropy method

Broadly speaking, the cross-entropy method is an adaptive importance sampling method performed iteratively (with main step hereafter indexed k). On this topic, the interested reader can refer to the book [14], or to [5] for a shorter introduction. See also yhe related adaptive importance sampling in [8].

Let $\eta^N_{(k)}$ denotes a weighted empirical distribution approximating η_{true} and constructed with a previously obtained samples, as discussed in the previous section. In cross-entropy methods, importance proposals are chosen in a parametric family $(\mu_\theta)_{\theta \in \Theta}$, and in the most basic

versions of the algorithm, at step k, an i.i.d. sequence $X_1^{(k)}, \ldots, X_N^{(k)}$ is sampled with distribution $\mu_{\theta^N_{(k-1)}}$ for an iteratively chosen parameter $\theta^N_{(k-1)}$. The approximating weighted empirical distribution discussed in the previous section is then explicitly given by

$$\eta^N_{(k)} \propto \sum_{n=1}^{N} \frac{f}{\mathrm{d}\mu_{\theta^N_{(k-1)}}} \left(X_n^{(k)}\right) \delta_{X_n^{(k)}}.$$

Finally, the specific parameter $\theta^N_{(k)}$ at which importance sampling is performed at step $k+1$ is chosen by solving the minimization problem

$$\theta^N_{(k)} = \arg\min_{\theta} \mathrm{Ent}\left(\eta^N_{(k)} \mid \mu^N_{(k),\theta}\right), \tag{4.2}$$

where in the above we denote the empirical version of a proposal by

$$\mu^N_{(k),\theta} \propto \sum_{n=1}^{N} \frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu_{\theta^N_{(k-1)}}} \left(X_n^{(k)}\right) \delta_{X_n^{(k)}}.$$

In many cases, (with the obvious exception of mixtures), μ_θ is chosen in an exponential family that can be set in a canonical Gibbs form:

$$\mu_\theta \equiv \mu_\beta \propto \mathrm{e}^{\langle \beta, T\rangle} \mathrm{d}\pi.$$

At each step of the method, the minimizer is then precisely given by $\mu_* = \mu_{\beta_*}$ where β_* is the unique vector such that

$$\mu_{\beta_*}(T) = \eta^N(T).$$

The novelty of the present paper consist in showing that the distribution μ_{β_*}, according to Theorem 3.3, is also the worst-case optimal proposal *among all possible proposals*, as soon as one considers the following convex set of admissible targets:

$$\mathcal{C}_{\eta^N} = \left\{\eta \in \mathcal{P}(E) : \eta(T) = \eta^N(T)\right\}.$$

In other words, we have thus shown that the minimization of relative entropy between an empirical target η^N and proposals in an exponential family is *equivalent* to the minimization of *worst-case* relative entropy for targets η with fixed average $\eta(T) = \eta^N(T)$; the minimization being obtained for proposals spanning *all* distributions. This seems to be an original interpretation of the cross-entropy method.

Our result also suggests a possible way to improve or control cross-entropy methods. In practice, confidence intervals in the form of (4.1) could be used instead of the singleton $C = \{\eta^N(T)\} \subset \mathbb{R}^d$. This might improve the robustness of cross-entropy algorithms, especially in high dimension.

5 The sample size required for importance sampling

In this section, $\eta \in \mathcal{P}(E)$ will denote a generic target probability distributions and $\mu \in \mathcal{P}(E)$ a generic proposal distribution.

In the so-called *importance sampling* method, the density $\mathrm{d}\eta/\mathrm{d}\mu$ is assumed to be computable (perhaps only up to a normalizing constant), and averages of the form $\eta(\varphi)$ are estimated using the empirical distribution of n i.i.d. variables $(X_i)_{i=1\ldots n}$ distributed according to μ. The estimator is given by:

$$\eta^N(\varphi) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} \varphi(X_i)\,\frac{\mathrm{d}\eta}{\mathrm{d}\mu}(X_i) \xrightarrow[n\to+\infty]{\text{a.s.}} \eta(\varphi). \tag{5.1}$$

In the above, only the product $\varphi\,\frac{\mathrm{d}\eta}{\mathrm{d}\mu}$ has to be evaluated numerically; if the normalizing constant Z is unknown, the test function φ must be defined as a product of this unknown normalization Z with another computable function. From now on, we will denote

$$Y \stackrel{\text{def}}{=} \frac{\mathrm{d}\eta}{\mathrm{d}\mu}(X) \geq 0, \qquad X \sim \mu.$$

In the special case when the test function $\varphi = Z$ is the normalizing constant, the relative variance of the estimation of Z can be immediately computed:

$$\mathbb{V}\mathrm{ar}(\eta^N(\mathbf{1})) = \frac{\mathbb{V}\mathrm{ar}(Y)}{N} = \frac{\mathrm{e}^{\mathrm{Ent}_2(\eta|\mu)} - 1}{N} \tag{5.2}$$

where $\mathrm{Ent}_2(\eta \mid \mu) = \ln \eta\left(\frac{\mathrm{d}\eta}{\mathrm{d}\mu}\right)$ is the order 2 Rényi entropy (see below for a definition). Using (5.2) and various standard concentration inequalities, one can then obtain an upper bound on (some appropriate notion of) the *required* sample size N^*. This upper bound

is usually comparable to the variance (using Chebyshev inequality) $N^* \leq c\,\mathbb{V}\mathrm{ar}(Y)$, or to the square of an upper bound on the support of Y $N^* \leq c\ell^2$ if $Y \leq \ell$, or a combination of both; in the above c is a numerical constant.

5.1 The Chatterjee-Diaconis bounds

We shall use here an alternative approach motivated by [3]. In the latter reference, some theorems (Theorem 1.1, 1.2 and 1.3) are proved showing that the sample size $N^* = N^*_{\delta,p_\alpha}(\eta \mid \mu)$ required to obtain an importance sampling estimation at a given precision δ and with a given probability p_α is, quite generically, and at logarithmic scales, given by the relative entropy of the target η with respect to the proposal μ:

$$\ln N^* \simeq \mathrm{Ent}(\eta \mid \mu) = \mathbb{E}(Y \ln Y). \tag{5.3}$$

The failure (or equivalently success) probability of importance sampling at a given precision threshold $\delta > 0$ is rigorously defined as follows. The deviation probability of importance sampling is defined using $\mathbb{L}^2(\nu)$-normalized test functions, and is given by:

$$p_{\mathrm{dev},\delta}(N) \stackrel{\mathrm{def}}{=} \sup_{\varphi:\, \eta(\varphi^2)=1} \mathbb{P}\left(\left|\eta^N(\varphi) - \eta(\varphi)\right| \geq \delta\right), \tag{5.4}$$

where one considers estimators of the form (5.1). A sample size denoted $N^* = N^*_{p_\alpha,\delta}(\eta \mid \mu)$ and satisfying

$$p_{\mathrm{dev},\delta}(N^*) = p_\alpha.$$

is called a critical sample size *required* for importance sampling. It is a sample size achieving δ-accuracy with success probability exactly $1 - p_\alpha$.

For simplicity, we now state a slightly weaker version of Theorem 1.1 [3] (based on lower order moment conditions) for the special case where the failure probability is defined with the estimator of the normalizing constant rather than (5.4):

$$p_{\mathrm{dev},\delta}(N) \stackrel{\mathrm{def}}{=} \mathbb{P}\left(\left(\eta^N(\mathbb{1}) - 1\right) \geq \delta\right).$$

This case enables to present the result as a general result on sum i.i.d. real valued random variables. The case (5.4) can be treated in the same way (since the lower bound in Theorem 1.1 [3] is obtained for constant test functions $\varphi = \mathbb{1}$). Such results and their proofs can be found in Section A.

Theorem 5.1 (Corollary of Theorem 1.1 [3]). *Let $Y_i \geq 0$, $i = 1 \dots N$ i.i.d. random variables with unit mean $\mathbb{E}Y = 1$. Denote by N^* a sample size verifying*

$$\mathbb{P}\left(\left|\frac{1}{N^*}\sum_{i=1}^{N^*} Y_i - 1\right| \geq \delta\right) = p_\alpha,$$

for some $\delta, p_\alpha \in (0,1)$. Then the following estimate holds

$$|\ln N^* - \mathbb{E}(Y \ln Y)| \leq \inf_{\theta\in[0,1]}\left[\frac{1}{\theta}\ln\left(\mathbb{E}(Y^{1+\theta})\mathbb{E}(Y^{1-\theta))}\right) + \frac{\ln c(\delta, p_\alpha)}{\theta}\right], \tag{5.5}$$

where c_{δ,p_α} is a numerical constant depending only on p_α and δ.

The estimate (5.5) result can be interpreted a concentration/anti-concentration inequality with lower order moments condition. The proof, discussed in Section A, is based on: i) an upper bound on the deviation probability $p_{\mathrm{dev},\delta}(N)$ that decreases with the sample size according to a (slow) power-law $N^{-\theta/4}$, ii) an upper bound on the success probability $1 - p_{\mathrm{dev},\delta}(N)$ that increases with the sample size according to a (slow) power-law $N^{\theta/2}$.

In the context of importance sampling, one has $\frac{1}{N^*}\sum_{i=1}^{N^*} Y_i = \eta^N(\mathbb{1})$, $\mathbb{E}(Y \ln Y) = \mathrm{Ent}(\eta \mid \mu)$, and the logarithmic moments

$$\frac{1}{\theta}\ln\left(\mathbb{E}(Y^{1+\theta}\right) = \frac{1}{\theta}\ln\int\left(\frac{\mathrm{d}\eta}{\mathrm{d}\mu}\right)^\theta \mathrm{d}\eta \stackrel{\mathrm{def}}{=} \mathrm{Ent}_{1+\theta}(\eta \mid \mu), \qquad \theta \in [-1, +\infty],$$

are the order $1+\theta$ Rényi entropy of η relative to μ. The latter is increasing with θ, and it is continuous when finite. $\mathrm{Ent}_1 = \mathrm{Ent}$ is the usual relative entropy. We refer to [16] for a review of properties of those entropies.

In our context, the estimate in (5.5) is meaningful only in cases where the relative entropy $\mathbb{E}(Y \ln Y)$ is large as compared to the difference of Rényi entropies $\mathrm{Ent}_{1+\theta}(\eta \mid \mu) - \mathrm{Ent}_{1-\theta}(\eta \mid \mu)$. It is argued in [3] that this situation happens quite generically, see the discussion in Section A. Some care however is required in practice because the constants in the right hand side of (5.5) can be unsatisfactory; they are however certainly not sharp, and it may turn out that improved estimates can be obtained.

A conservative user of importance sampling might prefer to minimize the usual variance $\mathbb{V}\mathrm{ar}(Y) = \eta(\frac{\mathrm{d}\eta}{\mathrm{d}\mu}) - 1$, rather that $\mathrm{Ent}(\eta \mid \mu)$.

However, minimizing upper bounds might lead to inefficient results depending on context. The estimate (5.5) shows that $\mathrm{Ent}(\eta \mid \mu)$ is, loosely speaking, a compromise between an upper bound and a lower bound. This makes relative entropy an interesting practical criteria.

Moreover, denoting the right hand side in (5.5) by

$$R(Y) \stackrel{\text{def}}{=} \inf_{\theta \in [0,1]} \left[\frac{1}{\theta} \ln \left(\mathbb{E}(Y^{1+\theta}) \mathbb{E}(Y^{1-\theta}) \right) + \frac{\ln c}{\theta} \right],$$

we provide in the next section an important class of examples for which, in a quite generic asymptotics, the following holds:

$$\ln \mathbb{V}\mathrm{ar}(Y) - \mathbb{E}(Y \ln Y) \gg R(Y). \tag{5.6}$$

This implies, according to (5.5), that the sample size estimate with $\mathrm{e}^{\mathbb{E}(Y \ln Y)}$ is arbitrarily more accurate than $\mathbb{V}\mathrm{ar}(Y)$, in the sense that:

$$|\ln N^* - \ln \mathbb{V}\mathrm{ar}(Y)| \gg |\ln N^* - \mathbb{E}(Y \ln Y)|,$$

that is, in other words, variance is unwarrantedly too large.

5.2 A motivating example

Consider the class distribution given by

$$\begin{cases} \mathbb{P}(Y = 0) = 1 - p_1 - p_2, \\ \mathbb{P}(Y = l_1) = p_1, \\ \mathbb{P}(Y = l_2) = p_2, \end{cases}$$

with the condition

$$\alpha \stackrel{\text{def}}{=} p_1 l_1 = 1 - p_2 l_2,$$

which is equivalent to $\mathbb{E}Y = 1$. We assume that

$$r \stackrel{\text{def}}{=} \frac{l_2}{l_1} \to 0, \qquad l_1 \to +\infty,$$

together with

$$\alpha \to 0,$$

which means that a large value l_1 has a small contribution α in the average $\mathbb{E}(Y)$.

Straightforward calculations yields:

$$\ln \mathbb{E}(Y^2) = \ln(\alpha l_1 + (1-\alpha) l_2) = \ln l_1 + \ln(\alpha + (1-\alpha) r),$$

and we now assume that

$$\alpha \gg r^{1-\alpha} \geq r$$

which implies that the large value l_1 has a dominating contribution in the variance (contrary to the mean).

One can now compute

$$\mathbb{E}(Y \ln Y) = \alpha \ln l_1 + (1-\alpha) \ln l_2 = \ln l_1 + (1-\alpha) \ln r,$$

so that the difference with the log-variance diverge:

$$\ln \mathbb{E}(Y^2) - \mathbb{E}(Y \ln Y) \sim \ln(\alpha) - (1-\alpha) \ln r = \ln \frac{\alpha}{r^{1-\alpha}} \to +\infty.$$

On the other hand

$$\frac{1}{\theta} \ln \mathbb{E}(Y^{1+\theta}) = \frac{1}{\theta} \ln \left(\alpha l_1^\theta + (1-\alpha) l_2^\theta \right) = \ln l_1 + \frac{1}{\theta} \ln \left(\alpha + (1-\alpha) r^\theta \right),$$

so that

$$\frac{1}{\theta} \ln \left(\mathbb{E}(Y^{1+\theta}) \mathbb{E}(Y^{1-\theta}) \right) = \frac{1}{\theta} \ln \left(1 + \alpha(1-\alpha)(r^\theta + r^{-\theta} - 2) \right).$$

Now, taking

$$\theta = \frac{-\ln \alpha}{-\ln r} \to 0$$

implies that $\alpha r^{-\theta}$ remains bounded so that

$$\inf_{\theta \in [0,1]} \left[\frac{1}{\theta} \ln \left(\mathbb{E}(Y^{1+\theta}) \mathbb{E}(Y^{1-\theta}) \right) + \frac{\ln c}{\theta} \right] \leq \frac{-\ln r}{-\ln \theta} \ln(2c),$$

and since

$$\frac{-\ln r}{-\ln \theta} \ll -\ln \frac{r^{1-\alpha}}{\alpha},$$

which implies that (5.6) holds true.

A More on the Chatterjee-Diaconis bounds

In this section, $\eta \in \mathcal{P}(E)$ will denote generic target probability distributions, and $\mu \in \mathcal{P}(E)$ a generic proposal. The importance sampling estimator is given by for any test function φ.

$$\eta^N(\varphi) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \varphi(X_i) \frac{\mathrm{d}\eta}{\mathrm{d}\mu}(X_i) \xrightarrow[n \to +\infty]{\text{a.s.}} \eta(\varphi).$$

The failure probability of importance sampling is defined using a supremum over $\mathbb{L}^2(\nu)$-normalized test functions, and is given, for a precision threshold δ by:

$$p_{\text{dev},\delta}(N) \stackrel{\text{def}}{=} \sup_{\varphi:\, \eta(\varphi^2)=1} \mathbb{P}\left(\left|\eta^N(\varphi) - \eta(\varphi)\right| \geq \delta\right). \tag{A.1}$$

One can alternatively consider the failure probability of estimating a normalizing constant

$$p_{\text{dev},\delta}(N) \stackrel{\text{def}}{=} \mathbb{P}\left(\left|\eta^N(\mathbb{1}) - \eta(\mathbb{1})\right| \geq \delta\right). \tag{A.2}$$

If p_α denotes a failure probability, a critical sample size $N^* = N^*_{p_\alpha,\delta}(\eta \mid \mu)$ *required* for importance sampling is then defined by

$$p_{\text{dev},\delta}(N^*) = p_\alpha. \tag{A.3}$$

which is precisely the sample size achieving δ-accuracy with success probability $1 - p_\alpha$.

The main theorems of [3] and the subsequent examples are not so easy to state and interpret. We propose in this section a slightly weaker reformulation of the main result Theorem 1.1 of [3] using Rényi entropies.

Rényi entropies are the power-law generalization of relative entropy and are defined by

$$\mathrm{Ent}_\alpha(\eta \mid \mu) \stackrel{\text{def}}{=} \frac{1}{\alpha} \ln \int \left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)^\alpha \mathrm{d}\nu, \qquad \alpha \in [0, +\infty]$$

if $\nu \ll \mu$, while $\mathrm{Ent}_\alpha(\eta \mid \mu) = +\infty$ otherwise. $\alpha \mapsto \mathrm{Ent}_\alpha(\nu \mid \mu)$ is increasing and continuous when finite. $\mathrm{Ent}_1 = \mathrm{Ent}$ is the usual relative entropy. We refer to [16] for a review of properties of those entropies.

A corollary of the main result (Theorem 1.1) in [3] is then the following:

Corollary A.1 (Theorem 1.1, [3]). *Let $\eta, \mu \in \mathcal{P}(E)$ be two given probability distribution, and let N^* be a sample size verifying* (A.1)-(A.3) *or* (A.1)-(A.3) *with $\delta, p_\alpha \in (0,1)$. There is a numerical constant $c(\delta, p_\alpha) \leq \max\left(\left(\frac{3}{p_\alpha \delta}\right)^4, \left(\frac{2}{(1-p_\alpha)(1-\delta)}\right)^2\right)$ such that*

$$|\ln N^* - \mathrm{Ent}(\eta \mid \mu)| \leq \inf_{\theta \in [0,1]} \left[2\left(\mathrm{Ent}_{1+\theta}(\eta \mid \mu) - \mathrm{Ent}_{1-\theta}(\eta \mid \mu)\right) + \frac{\ln c}{\theta}\right]$$

Corollary A.1 implies that

$$\ln N_{p_\alpha, \delta}(\eta \mid \mu) \sim \mathrm{Ent}(\eta \mid \mu)$$

for asymptotics such that

$$\mathrm{Ent}(\eta \mid \mu) \gg \inf_{\theta \in (0,1)} \ \mathrm{Ent}_{1+\theta}(\eta \mid \mu) - \mathrm{Ent}_{1-\theta}(\eta \mid \mu) + \frac{1}{\theta} \ln c(\delta, p_\alpha), \tag{A.4}$$

where we stress that $c(\delta, p_\alpha)$ is numerical (independent of η, μ).

In Section 3 of [3], S. Chatterjee and P. Diaconis argued that many (high dimensional) toy models inspired by statistical mechanics indeed satisfy behaviors similar to (A.4) when $\mathrm{Ent}(\eta \mid \mu)$ is large. More specifically, one can first remark that Rényi entropies satisfy (if finite for $|\theta|$ small):

$$\mathrm{Ent}_{1+\theta}(\eta \mid \mu) = \mathrm{Ent}(\eta \mid \mu) + \frac{\theta}{2} \mathrm{var}_\eta(\ln \frac{\eta}{\mu}) + \mathcal{O}(\theta^3).$$

The authors then showed that in various cases of interest it holds $\mathrm{var}_\eta(\ln \frac{\eta}{\mu}) = \mathcal{O}\left(\mathrm{Ent}(\eta \mid \mu) = \eta(\ln \frac{\eta}{\mu})\right)$, which in other words means tha the mean and variance of the log-likelihood $\ln \frac{\eta}{\mu}$ are of the same order. In examples of Section 3 of [3], these two quantities scale like the dimension of the considered system. Choosing θ of order

$$\theta = \mathcal{O}(\mathrm{Ent}(\eta \mid \mu)^{-1/2})$$

then yields the asymptotic condition (A.4). We provided in Section 5.2 a simpler, practically generic, example for which this condition is also satisfied.

Proof of Corollary A.1. Let us recall the notation of Theorem 1.1 in [3]. The importance sampling estimator $\eta^N(\varphi)$ is denoted $I^n(f)$.

The authors also use the notation $L = \mathrm{Ent}(\eta \mid \mu)$ and $t = \pm(\ln N - L)$.We will also denote $\mathrm{Ent}_{1+\theta}(\nu \mid \mu) = L_{1+\theta}$.

Denote $t = \ln N - L \geq 0$. Using Markov inequality, the first inequality in Theorem 1.1 can be rewritten:

$$p_{\mathrm{dev},\delta}(N) \leq \frac{1}{\delta}\mathrm{e}^{-t/4} + \frac{2}{\delta}\sqrt{\eta(\mathbb{1}_{\ln\rho > D+t/2})}.$$

A routine upper bound of Chernoff-type yields, for any $\theta \geq 0$,

$$\eta(\mathbb{1}_{\ln\rho > L+t/2}) \leq \mathrm{e}^{-\theta L-\theta t/2} \underbrace{\eta(\rho^{\theta})}_{\exp\theta\mathrm{Ent}_{1+\theta}(\mu|\eta)} = \mathrm{e}^{\theta(L_{1+\theta}-L)-\theta t/2},$$

so that for $\theta < 1$:

$$p_{\mathrm{dev},\delta}(N) \leq \frac{3}{\delta}\mathrm{e}^{-\theta t/4+\theta(L_{1+\theta}-L)/2}.$$

Denoting $c = \left(\frac{3}{p_\alpha \delta}\right)^4$ for $p_{\mathrm{dev},\delta}(N^*) = p_\alpha$, it yields

$$\ln N - L = t \leq 2\,(L_{1+\theta} - L) + \frac{\ln c}{\theta},$$

the claimed upper bound on $\ln N^* - L$ follows.

Similarly, let us now denote $t = -\ln N + L \geq 0$. The second inequality in Theorem 1.1 in [3] can be rewritten:

$$\begin{aligned} 1 - p_{\mathrm{dev},\delta}(N) &\leq P(\eta^N(\mathbb{1}) - 1 \geq -\delta) \\ &\leq \frac{1}{1-\delta}\mathrm{e}^{-t/2} + \frac{1}{1-\delta}\eta(\mathbb{1}_{\ln\rho \leq D-t/2}). \end{aligned}$$

A similar routine upper bound of Chernoff-type yields

$$\eta(\mathbb{1}_{\ln\rho \leq L-t/2}) \leq \mathrm{e}^{\theta L-\theta t/2} \underbrace{\eta(\rho^{-\theta})}_{\exp\theta\mathrm{Ent}_{1-\theta}(\mu|\eta)} = \mathrm{e}^{\theta(L-L_{1-\theta})-\theta t/2},$$

so that

$$1 - p_{\mathrm{dev},\delta}(N) \leq \frac{2}{1-\delta}\mathrm{e}^{-\theta t/2+\theta(L-L_{1-\theta})}.$$

Denoting $c = \left(\frac{2}{(1-p_\alpha)(1-\delta)}\right)^2$ for $p_{\mathrm{dev},\delta}(N^*) = p_\alpha$, it yields

$$L - \ln N = t \leq 2\,(L - L_{1-\theta} - L) + \frac{\ln c}{\theta},$$

the claimed lower bound on $\ln N^* - L$ follows. □

B A measurable selection lemma

We first start by a general result of measure theory (see Theorem 10.8.3in Bogachev Vol. 2 [2]), which states the existence of an independent "complement" of atomless conditional measures.

Theorem B.1 (Independent Complement)**.** *Let $T : E \to F$ denotes a measurable map between two standard Borel spaces. Let $\pi \in \mathcal{P}(E)$ denotes a probability such that for $T\sharp\pi$-almost all $t \in F$, $\pi(\ . \mid T = t)$ is atomless. Let λ denotes the usual Lebesgue measure on the interval $[0, 1]$. Then there exists a measurable function $S : E \to [0, 1]$ such that, i) $(T, S) : E \to F \times [0, 1]$ is one-to-one and, ii)*

$$(T, S)\sharp\pi = T\sharp\pi \otimes \lambda;$$

that is, if X has distribution π, then $T(X)$ and $S(X)$ are independent with distribution $(T\sharp)\,\pi$ and λ respectively.

We next use the above theorem to prove the existence of distributions in a given half-space with prescribed conditional entropy.

Lemma B.2. *Let $T : E \to F$ denotes a measurable map between two standard Borel spaces. Let $\pi \in \mathcal{P}(E)$ denotes a probability such that for $T\sharp\pi$-almost all t, $\pi(\ . \mid T = t)$ is atomless. For each $\varepsilon \in (0, 1)$ and all bounded measurable function $F : E \to \mathbb{R}$, there exists a measurable set $A_{\varepsilon,F} \subset E$ verifying for $T\sharp\pi$-almost all t*

$$\pi(A_{f,h} \mid T = t) = \varepsilon$$

as well as

$$\pi(F \mid A_{f,h}, T = t) \geq \pi(F \mid T = t).$$

Proof. The existence a measurable complement (Theorem B.1) implies *a fortiori*, the existence of a measurable function $S : E \to [0, 1]$ such that, for $T\sharp\pi$-almost all t, $S\sharp\pi(\ . \mid T = t)$ is the Lebesgue distribution. Let us consider the pair $(y_\varepsilon(t), z_\varepsilon(t)) \in \mathbb{R} \times [0, 1]$ defined by

$$\pi_t\left(\{F > y_\varepsilon(t)\} \cup \{F = y_\varepsilon(t)\,\&\,S \geq z_\varepsilon(t)\}\right) = \varepsilon;$$

The above equality being required for t in a measurable set of $T\sharp\pi$-measure 1 on which $S\sharp\pi(\ . \mid T = t)$ is the Lebesgue distribution. We conventionally set $(y_\varepsilon(t), z_\varepsilon(t)) = (0, 0)$ elsewhere. By Lemma B.3, $t \mapsto (y_\varepsilon(t), z_\varepsilon(t))$ is a uniquely defined measurable map.

We can then consider the measurable set

$$A_{f,h} \stackrel{\text{def}}{=} \{x \in E :\, F(x) > y_\varepsilon(T(x))\} \cup \{x \in E :\, F(x) = y_\varepsilon(T(x)) \,\&\, S(x) \geq z_\varepsilon(T(x))\}\,.$$

Let us now remark that for any subset $B \subset E$ the function F takes greater values on the set

$$A \stackrel{\text{def}}{=} \{F > y\} \cup (\{F = y\} \cap B)$$

than on its complementary A^c. Now for any probability η on E

$$\int F \mathrm{d}\eta = \eta(A)\mathrm{d}\eta(F \mid A) + (1 - \eta(A))\eta(F \mid A^c),$$

which yields $\eta(F \mid A) \geq \eta(F)$. One can apply this argument in our case to $\eta = \pi(\,.\mid T = t)$ and $A = A_{\varepsilon,F}$ in order to obtain $\pi(F \mid A_{f,h} \,\&\, T = t) \geq \pi(F \mid T = t)$ for $T\sharp\pi$-almost all t. □

Finally, we give an elementary lemma which proves the measurability of the set $A_{f,h}$ constructed in the proof of Lemma B.2 above.

Lemma B.3. *Let E and F denotes two measurable spaces, and let $t \in F \mapsto \pi_t \in \mathcal{P}(E)$ denotes a measurable probability kernel ($t \mapsto \pi_t(A)$ is measurable for any measurable subset $A \subset E$). Assume given two measurable bounded functions $F : E \to \mathbb{R}$ and $G \to [0,1]$ such that the push-forward probability $G\sharp\pi_t$ is the Lebesgue measure for all $t \in F$. For each $t \in F$ and each $\varepsilon \in (0,1)$, there exists a unique pair $(y_\varepsilon(t), z_\varepsilon(t)) \in \mathbb{R} \times [0,1]$ such that*

$$\pi_t\left(\{F > y_\varepsilon(t)\} \cup \{F = y_\varepsilon(t) \,\&\, S \geq z_\varepsilon(t)\}\right) = \varepsilon;$$

moreover, the map $t \mapsto (y_\varepsilon(t), z_\varepsilon(t))$ is measurable.

Proof. Let us denote

$$p_t(y,z) \stackrel{\text{def}}{=} \pi_t\left(\{F > y\} \cup \{F = y \,\&\, S \geq z\}\right).$$

By a routine argument of measure theory, $(t,z,y) \mapsto p_t(x,y)$ is a measurable map as a bounded pointwise limit of simple functions in tensorial form. The map $(y,z) \mapsto p_t(y,z)$ is decreasing for the lexicographic order, and $z \mapsto p_t(y,z)$ is continuous and strictly decreasing since, by assumption, $S\sharp\pi_t$ is atomless. $y_\varepsilon(t)$ can thus be defined as the unique y such that $\lim_{y+} p_t(\,.\,,1) > \varepsilon \leq p_t(y,1)$, and $z_\varepsilon(t)$ the unique z such that $p_t(y_\varepsilon(t), z) = \varepsilon$. This shows existence and uniqueness.

In order to prove the measurability of $y_\varepsilon, z_\varepsilon$, it suffices to show that they are the monotone limit of measurable simple functions.

For each $k \in N$, let us denote the unique pair $(y_\varepsilon^k(t), z_\varepsilon^k(t))$ in $2^{-k}\mathbb{Z} \times 2^{-k}[0, \dots, 2^k]$ defined by

$$p_t(y_\varepsilon^k(t), 1) \leq \varepsilon < p_t(y_\varepsilon^k(t) - 2^{-k}, 1),$$

as well as

$$p_t(y_\varepsilon(t), z_\varepsilon^k) \leq \varepsilon < p_t(y_\varepsilon(t), z_\varepsilon^k(t) - 2^{-k}).$$

By construction, $k \mapsto y_\varepsilon^k(t)$ and $k \mapsto z_\varepsilon^k(t)$ are decreasing maps, and by σ-additivity, their respective infima are given by $y_\varepsilon(t)$ and $z_\varepsilon(t)$.

We now claim that $t \mapsto y_\varepsilon^k(t)$ is measurable. Since the latter takes its values in a countable space, it suffices to show that the set $A_{y_0} \stackrel{\text{def}}{=} \left\{ t : \, y_\varepsilon^k(t) = y_0 \right\}$ is measurable for each $y_0 \in \mathbb{R}$. But by definition

$$A_{y_0} = \left\{ t : \, p_t(y_0, 1) \leq \varepsilon < p_t(y_0 - 2^{-k}, 1) \right\},$$

which is indeed measurable since $t \mapsto p_t(y, z)$ is measurable. As a consequence, $t \mapsto y_t^\varepsilon$ is measurable as a decreasing limit of simple measurable functions.

We finally claim that $t \mapsto z_\varepsilon^k(t)$ is also measurable. The argument works as in the paragraph above, except that we now use the measurability of the map $t \mapsto p_t(y_\varepsilon(t), z)$.

□

Acknowledgement

This work has been partially supported by ANR SINEQ, ANR-21-CE40-0006.

References

[1] Siu-Kui Au and JL Beck. Important sampling in high dimensions. *Structural safety*, 25(2):139–163, 2003.

[2] V.I. Bogachev. *Measure Theory, vol.* 2. Springer Berlin Heidelberg, 2006.

[3] Sourav Chatterjee and Persi Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

[4] Imre Csiszár and Frantisek Matus. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003.

[5] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

[6] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

[7] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2009.

[8] Randal Douc, Arnaud Guillin, J-M Marin, and Christian P Robert. Convergence of adaptive mixtures of importance sampling schemes. *The Annals of Statistics*, 35(1):420–448, 2007.

[9] Víctor Elvira, Luca Martino, and Christian P Robert. Rethinking the effective sample size. *International Statistical Review*, 90(3):525–550, 2022.

[10] Jonathan H Huggins and Daniel M Roy. Sequential monte carlo as approximate sampling: bounds, adaptive resampling via infinity-ess, and an application to particle gibbs. *Bernoulli*, 25(1):584–622, 2019.

[11] J.S. Liu. *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[12] Luca Martino, Víctor Elvira, and Francisco Louzada. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401, 2017.

[13] Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

[14] Reuven Y Rubinstein and Dirk P Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, volume 133. Springer, 2004.

[15] Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

[16] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.