Effective sample size approximations as entropy measures

L. Martino*, V. Elvira$^{\top}$,

* Universitá degli Studi di Catania, Italy.

$^{\top}$ University of Edinburgh, UK.

Abstract

In this work, we analyze alternative effective sample size (ESS) measures for importance sampling algorithms. We show the relationship between the ESS expressions used in the literature and two entropy families, the Rényi and Tsallis entropy. The Rényi entropy is connected to the Huggins-Roy's ESS family introduced in [12]. We prove that that all the ESS functions included in the Huggins-Roy's family fulfill all the desirable theoretical conditions. Moreover, we show that the Gini impurity index can be converted in a proper ESS formula. We also highlight its connection with the Tsallis entropy. Finally, by numerical simulations, we study the performance of different ESS expressions contained in the previous ESS families in term of approximation of the theoretical ESS definition.

Keywords: Importance Sampling; Effective Sample Size; Entropy; Diversity measure; Gini impurity; Resampling.

The effective sample size (ESS) measure is an important concept in order to quntify the efficiency of different Monte Carlo methods, such as Markov Chain Monte Carlo (MCMC) [10, 17] and Importance Sampling (IS) techniques [2, 4]. In an IS context, heuristically speaking, we can assert that ESS measures how many independent identically distributed (i.i.d.) samples, drawn directly from the target distribution $\bar{\pi}(\mathbf{x}) = \frac{1}{Z}\pi(\mathbf{x})$, are equivalent *in some sense* to the N weighted samples, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, drawn from a proposal distribution $q(\mathbf{x})$ and weighted according to the ratio $w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}$ [20]. This consideration is represented in the first box of Figure 1, referred as "abstract ESS concept".

The theoretical definition [10, 14] considers the ESS function proportional to the ratio between the variance of the ideal Monte Carlo estimator (drawing samples directly from the target) over the variance of the estimator obtained by the IS technique, using with the same number of samples in both estimators. This definition presents some drawbacks (see [19, 9] for an exhaustive discussion) and is useless for practical purposes since it cannot be computed in general. Hence, approximations of this theoretical formula are required. In Figure 1, this theoretical definition is represented by the second box. Within a IS context, the most common choice in literature to approximate this theoretical ESS definition is $\text{ESS} = \frac{1}{\sum_{n=1}^{M} \bar{w}_n^2}$, which involves (only) the normalized importance weights $\bar{w}_n = \frac{w_n}{\sum_{j=1}^{N} w_j}$, $n = 1, \ldots, N$ [6, 7, 15, 20]. This expression has been widely used in particle filtering in order to apply the resampling steps adaptively [7, 6, 11]. However, it presents

different weaknesses since it has been obtained after several approximations of the theoretical definition. For instance, it just depends on the normalized weights, but it is not dependent on particle locations and from the particular integral to approximate (see [19] for further details).
Several other alternatives have been studied in literature and applied in order to perform adaptive resampling within sequential Monte Carlo (SMC) methods [12, 19]. For instance, another measure called perplexity, involving the discrete entropy [5] of the normalized weights has been also proposed in [3]; see also [20, Chapter 4], [8, Section 3.5]. Another expression is defined as the inverse of the maximum of the normalized weights $\bar{w}_n$ [19].
In this work, we recall the definition of the Generalized ESS (G-ESS) functions given in [19]. We stress and show that the G-ESS functions can be considered *diversity indices* [13] (see third box in Figure 1). The actual reason for the success of the ESS expressions introduced in the literature, is related to the fact they are discrepancy measure and/or can considered as diversity indices. Indeed, we show that the G-ESS functions can be associate to different entropy families [5]. Given an entropy measure of the probability mass function (pmf) defined by the normalized weights $\bar{w}_n$, $n = 1, \ldots, N$, we can obtain a G-ESS formula by taking the exponential transformation of the entropy expression (in some cases, some additional translation and scaling are needed).
More specifically, we analyze the Rényi and Tsallis entropy families, converting them in G-ESS functions. The ESS formulas corresponding to the Rényi entropy coincides with the Huggins-Roy's ESS family introduced and studied independently in [12],

$$\text{ESS} \quad = \quad \left(\sum_{n=1}^{N} \bar{w}_n^{\beta} \right)^{\frac{1}{1-\beta}}, \qquad \beta \geq 0.$$

We show that all the G-ESS expressions belonging to this family satisfy all the desired requirements, being all *proper and stable*. Moreover, all the main formulas previously proposed in the literature are contained in the Huggins-Roy's family. Using the Tsallis entropy, we obtain another ESS family which contain the *Gini impurity index* as special case, that is widely employed in machine learning within decision tree algorithms [1, 16]. We also discuss the connection to another ESS family provided in [19]. However, generally the Tsallis ESS formulas are not proper and stable. Hence, we focus the numerical studies to the Huggins-Roy's family.
Furthermore, by numerical simulations, we obtain the G-ESS function within Huggins-Roy's family which provides the best approximation the theoretical ESS definition, in two specific scenarios. We also study linear combinations of G-ESS functions in order to enhance the approximation of the theoretical definition. The results of our numerical simulations suggest the use of the formulas of type $\text{ESS} = \left(\frac{1}{\sum_{n=1}^{M} \bar{w}_n^4} \right)^{1/3}$ and $\text{ESS} = \left(\frac{1}{\sum_{n=1}^{M} \bar{w}_n^8} \right)^{1/7}$. Both expression differ from the classical formula $\text{ESS} = \frac{1}{\sum_{n=1}^{M} \bar{w}_n^2}$, which is contained in Huggins-Roy's family with $\beta = 2$. Our study suggest the use of $\beta > 2$. These considerations can be also relevant clues for future applications and studies.

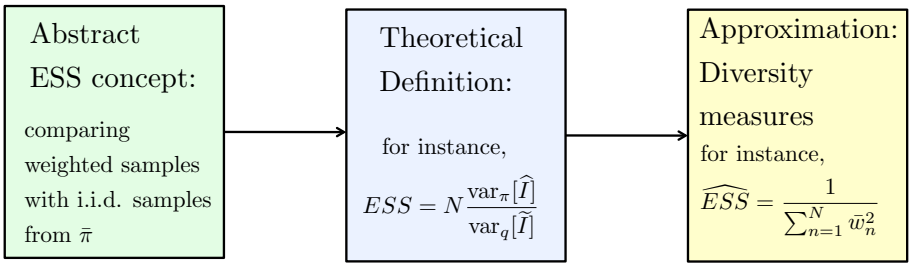


Figure 1: Graphical representation of the development of the approximated ESS formulas for importance sampling. The abstract concept of Effective Sample Size has been translated in a mathematical formulation providing a first attempt of theoretical definition. Since this definition cannot compute, several approximations have been proposed (based only in the information provided by the normalized IS weights). The expression $\mathrm{ESS} = \frac{1}{\sum_{n=1}^M \bar{w}_n^2}$ is the most applied so far in the literature.

1 Effective sample size (ESS) for importance sampling

Let us denote the target probability density function (pdf) as $\bar{\pi}(\mathbf{x}) \propto \pi(\mathbf{x})$ (known up to a normalizing constant) with $\mathbf{x} \in \mathcal{X}$. Moreover, we consider the following integral involving $\bar{\pi}(\mathbf{x})$ and a square-integrable function $h(\mathbf{x})$,

$$I = \int_{\mathcal{X}} h(\mathbf{x})\bar{\pi}(\mathbf{x})d\mathbf{x}, \tag{1}$$

which we desire to approximate using a Monte Carlo approach. If we are able to draw N independent samples $\mathbf{x}_1, \dots, \mathbf{x}_N$ from $\bar{\pi}(\mathbf{x})$, then the Monte Carlo estimator of I is

$$\widehat{I} = \frac{1}{N}\sum_{n=1}^{N} h(\mathbf{x}_n) \approx I, \tag{2}$$

where $\mathbf{x}_n \sim \bar{\pi}(\mathbf{x})$, with $n = 1, \dots, N$. However, in general, generating samples directly from the target, $\bar{\pi}(\mathbf{x})$, is impossible. Alternatively, we can draw N samples $\mathbf{x}_1, \dots, \mathbf{x}_N$ from a (simpler) proposal pdf $q(\mathbf{x})$,[1] and then assign a weight to each sample, $w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}$, with $n = 1, \dots, N$, according to the importance sampling (IS) approach. Defining the normalized weights,

$$\bar{w}_n = \frac{w_n}{\sum_{i=1}^N w_i}, \quad n = 1, \dots, N, \tag{3}$$

then the self-normalized IS estimator is

$$\widetilde{I} = \sum_{n=1}^{N} \bar{w}_n h(\mathbf{x}_n) \approx I, \tag{4}$$

with $\mathbf{x}_n \sim q(\mathbf{x})$, $n = 1, \dots, N$. In general, the estimator $\widetilde{I}$ is less efficient than $\widehat{I}$, since the samples are not directly drawn from $\bar{\pi}(\mathbf{x})$. In several applications [6, 7], it is necessary to measure the loss

[1] We assume that $q(\mathbf{x}) > 0$ for all $\mathbf{x}$ where $\bar{\pi}(\mathbf{x})) \neq 0$, and $q(\mathbf{x})$ has heavier tails than $\bar{\pi}(\mathbf{x})$.

of the efficiency using $\widetilde{I}$ instead of $\widehat{I}$. The idea is to define the Effective Sample Size (ESS) as the ratio of the variances of the estimators [14],

$$\mathrm{ESS}_{\texttt{teo}}(h) = N\frac{\mathrm{var}_\pi[\widehat{I}]}{\mathrm{var}_q[\widetilde{I}]}. \tag{5}$$

Note the dependence on the function $h(\mathbf{x})$ corresponding to a specific integral. The theoretical value above $\mathrm{ESS}_{\texttt{teo}}(h)$ is always positive, could be smaller than 1 and, in some situations, bigger than N as well [9].

2 Practical ESS formulas

2.1 ESS expressions in the literature

Finding a useful expression of ESS derived analytically from the theoretical definition is not straightforward. Then, different derivations [14, 15], [7, Chapter 11], [20, Chapter 4] proceed using several approximations and assumptions for yielding an expression useful from a practical point of view. A well-known rule of thumb, widely used in literature [7, 18, 20], is

$$\mathrm{ESS}_N(\bar{\mathbf{w}}) \;=\; \frac{1}{\sum_{i=1}^N \bar{w}_n^2}, \tag{6}$$

where we have used the the normalized weights

$$\bar{\mathbf{w}} = [\bar{w}_1, \ldots, \bar{w}_N],$$

in the first equality, and the unnormalized ones in the second equality.[2] Another similar measure, called *perplexity*, has been proposed in literature [3, 20] based only on the normalized importance weights,

$$\mathrm{ESS}_N(\bar{\mathbf{w}}) = \exp\{H(\bar{\mathbf{w}})\} \tag{7}$$

where

$$H(\bar{\mathbf{w}}) = -\sum_{n=1}^N \bar{w}_n \log \bar{w}_n,$$

is the discrete entropy of the vector $\bar{\mathbf{w}}$ [5]. Another proposed in the literature and easy to use is the following formula

$$\mathrm{ESS}_N(\bar{\mathbf{w}}) = \frac{1}{\max \bar{w}_n}. \tag{8}$$

An interesting property of all the three expressions above in Eqs. (6)-(7)-(8) is

$$1 \leq \mathrm{ESS}_N(\bar{\mathbf{w}}) \leq N. \tag{9}$$

[2] Due to the several approximations which have been applied to obtain the final formula, P_N does not depend on the particles $\mathbf{x}_n$, $n = 1, \ldots, N$, which is obviously a drawback (for further considerations see [19]).

They are only based on the normalized weights $\bar{\mathbf{w}}$ and do not consider the information of the generated samples $\mathbf{x}_n$, which is clearly a drawback [9, 19]. Moreover, the theoretical value above $\mathrm{ESS}_{\texttt{teo}}(h)$ could be smaller than 1 and, in some scenario, bigger than N as well (see [9, Section 3.3]). Therefore, all of them are quite rough approximations of $\mathrm{ESS}_{\texttt{teo}}(h)$ but are quite use in practice. The reason is perhaps explained below: they are actually discrepancy/diversity measures.

2.2 Discrepancy w.r.t. the uniform pmf.

All the formulas above can be considered *diversity indices* or *discrepancy measures* [13, 19]. Indeed, consider the discrepancy between two pmfs: the pmf defined by the weights $\bar{\mathbf{w}} = [\bar{w}_1, \dots, \bar{w}_N]$ and the discrete uniform pmf defined by $\bar{\mathbf{w}}^* = \left[\frac{1}{N}, \dots, \frac{1}{N}\right]$. The ESS formula in Eq. (6) is related to the Euclidean distance between these two pmfs, i.e.,

$$\begin{aligned}
\left\|\bar{\mathbf{w}} - \bar{\mathbf{w}}^*\right\|_2 &= \sqrt{\sum_{n=1}^{N}\left(\bar{w}_n - \frac{1}{N}\right)^2} \\
&= \sqrt{\left(\sum_{n=1}^{N}\bar{w}_n^2\right) + N\left(\frac{1}{N^2}\right) - \frac{2}{N}\sum_{n=1}^{N}\bar{w}_n} \\
&= \sqrt{\left(\sum_{n=1}^{N}\bar{w}_n^2\right) - \frac{1}{N}} \\
&= \sqrt{\frac{1}{\mathrm{ESS}_N(\bar{\mathbf{w}})} - \frac{1}{N}},
\end{aligned}$$

where we have used $\mathrm{ESS}_N(\bar{\mathbf{w}}) = \frac{1}{\sum_{i=1}^{N}\bar{w}_n^2}$ in Eq. (6). Hence, maximizing the expression in Eq. (6) is equivalent to minimizing the Euclidean distance $\left\|\bar{\mathbf{w}} - \bar{\mathbf{w}}^*\right\|_2$. Note that this behavior is also typical of discrete entropy measures, as we stress in the next sections. Indeed, if the weights are more "diverse" to each other, the distance w.r.t. the discrete uniform pmf $\bar{\mathbf{w}}^*$ is higher, the ESS and the entropy of $\bar{w}$ are smaller. On the other hand, if the normalized weights are more similar to each other, they are all closer to the value $1/N$, so that the distance w.r.t. the discrete uniform pmf $\bar{\mathbf{w}}^*$ is smaller. As a consequence, the corresponding ESS and the entropy of $\bar{w}$ would be greater. Hence, it appears natural to consider the possibility of using other discrepancy and/or entropy measures to design alternative ESS expressions.

In the following, we describe five conditions that a generic ESS approximation based only on the information of the normalized weights must satisfy. Then we show that the family of functions proposed in [12] fulfills these five conditions. Furthermore, we link this G-ESS family with the Rényi entropy providing also some theoretical results.

3 Generalized ESS functions

Considering the practical approach employed above for defining ESS formulas as discrepancy-diversity measures, here we describe the five properties that a generalized ESS measure (G-ESS) should satisfy, based only on the information of the normalized weights. Here, first of all, note that any possible G-ESS is a function of the vector of normalized weights $\bar{\mathbf{w}} = [\bar{w}_1, \ldots, \bar{w}_N]$,

$$\mathrm{ESS}_N(\bar{\mathbf{w}}) = \mathrm{ESS}_N(\bar{w}_1, \ldots, \bar{w}_N) : \mathcal{S}_N \to [1, N], \tag{10}$$

where $\mathcal{S}_N \subset \mathbb{R}^N$ represents the *unit simplex* in $\mathbb{R}^N$. Namely, the variables $\bar{w}_1, \ldots, \bar{w}_N$ are subjected to the constrain

$$\bar{w}_1 + \bar{w}_2 + \ldots + \bar{w}_N = 1. \tag{11}$$

Moreover, we denoted

$$\bar{\mathbf{w}}^* = \left[\frac{1}{N}, \ldots, \frac{1}{N}\right], \tag{12}$$

and the vertices of the simplex $\mathcal{S}_N$ are denoted as

$$\bar{\mathbf{w}}^{(j)} \quad = \quad [\bar{w}_1 = 0, \ldots, \bar{w}_j = 1, \ldots, \bar{w}_N = 0], \tag{13}$$

i.e., $\bar{w}_j = 1$ and $\bar{w}_n = 0$ (it can occurs only if $\pi(\mathbf{x}_n) = 0$), for $n \neq j$ with $j \in \{1, \ldots, N\}$.

Below we list the five conditions that $\mathrm{ESS}_N(\bar{\mathbf{w}})$ should fulfill:

C1. **Symmetry:** ESS_N must be invariant under any permutation of the weights, i.e.,

$$\mathrm{ESS}_N(\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_N) = \mathrm{ESS}_N(\bar{w}_{j_1}, \bar{w}_{j_2}, \ldots, \bar{w}_{j_N}), \tag{14}$$

for any possible set of indices $\{j_1, \ldots, j_N\} = \{1, \ldots, N\}$.

C2. **Maximum condition:** A maximum value is N and it is reached at $\bar{\mathbf{w}}^*$ (see Eq. (12)), i.e.,

$$\mathrm{ESS}_N\left(\bar{\mathbf{w}}^*\right) = N \geq \mathrm{ESS}_N(\bar{\mathbf{w}}). \tag{15}$$

C3. **Minimum condition**: the minimum value is 1 and it is reached (at least) at the vertices $\bar{\mathbf{w}}^{(j)}$ of the unit simplex in Eq. (13),

$$\mathrm{ESS}_N(\bar{\mathbf{w}}^{(j)}) = 1 \leq \mathrm{ESS}_N(\bar{\mathbf{w}}). \tag{16}$$

for all $j \in \{1, \ldots, N\}$.

C4. **Unicity of extreme values:** The maximum at $\bar{\mathbf{w}}^*$ is unique and the the minimum value 1 is reached *only* at the vertices $\bar{\mathbf{w}}^{(j)}$, for all $j \in \{1, \ldots, N\}$.

C5. **Stability of the rate** ESS_N/N**:** Consider the vector of weights $\bar{\mathbf{w}} \in \mathbb{R}^N$ and the vector $\bar{\mathbf{v}} = [\bar{v}_1, \ldots, \bar{v}_{MN}] \in \mathbb{R}^{MN}$, $\quad M \geq 1$, obtained repeating and scaling by $\frac{1}{M}$ the entries of $\bar{\mathbf{w}}$, i.e.,

$$\bar{\mathbf{v}} = \frac{1}{M}[\underbrace{\bar{\mathbf{w}}, \bar{\mathbf{w}}, \ldots, \bar{\mathbf{w}}}_{M-times}]. \tag{17}$$

The invariance condition is expressed as

$$\mathrm{ESS}_N(\bar{\mathbf{w}}) \;\; = \;\; \frac{1}{M}\mathrm{ESS}_{MN}(\bar{\mathbf{v}}), \tag{18}$$

for all $M \in \mathbb{N}^+$.

This last requirement can be interpreted as an adjustment of the well-known *homogeneity* (scale-invariance) condition for real functions.[3] Note that, given conditions C2 and C3, we always have

$$1 \leq \mathrm{ESS}_N(\bar{\mathbf{w}}) \leq N. \tag{19}$$

On the condition C5. For clarifying this condition, consider the vector $\bar{\mathbf{v}} = [0, 1, 0]$ with $N = 3$, and the two additional vectors obtained repeating $\bar{\mathbf{v}}$ two or three times,

$$\bar{\mathbf{v}}' = \left[0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0\right] = \frac{1}{2}[\bar{\mathbf{v}}, \bar{\mathbf{v}}],$$
$$\bar{\mathbf{v}}'' = \left[0, \frac{1}{3}, 0, 0, \frac{1}{3}, 0, 0, \frac{1}{3}, 0\right] = \frac{1}{3}[\bar{\mathbf{v}}, \bar{\mathbf{v}}, \bar{\mathbf{v}}],$$

We would like to obtain $\mathrm{ESS}_N(\bar{\mathbf{v}}) = 1, \mathrm{ESS}_{2N}(\bar{\mathbf{v}}') = 2$ and $\mathrm{ESS}_{3N}(\bar{\mathbf{v}}'') = 3$, i.e., the ratio $\frac{\mathrm{ESS}_N}{N}$ is constant, i.e.,

$$\frac{\mathrm{ESS}_N(\bar{\mathbf{v}})}{N} = \frac{\mathrm{ESS}_{2N}(\bar{\mathbf{v}}')}{2N} = \frac{\mathrm{ESS}_{3N}(\bar{\mathbf{v}}'')}{3N} = \frac{1}{3}.$$

Classification of G-ESS. Table 1 classifies the G-ESS functions in different families depending on the conditions fulfilled. Recall that the first three conditions are strictly required, to be considered an ESS function. For instance, all the G-ESS functions which satisfy at least the first four conditions, i.e., from C1 to C4, are called proper functions. If all the conditions are fulfilled they are called proper and stable. We are interested in this kind of G-ESS expressions, proper and stable.

4 Huggins-Roy's ESS family

The Huggins-Roy's ESS family introduced in [12] is defined as

$$\text{ESS-H}_N^{(\beta)}(\bar{\mathbf{w}}) \;\; = \;\; \left(\frac{1}{\sum_{n=1}^N \bar{w}_n^\beta}\right)^{\frac{1}{\beta-1}}, \tag{20}$$

$$\;\; = \;\; \left(\sum_{n=1}^N \bar{w}_n^\beta\right)^{\frac{1}{1-\beta}}, \qquad \beta \geq 0. \tag{21}$$

[3] A function $f(\mathbf{x})$ is said to be homogeneous of degree k if $f(c\,\mathbf{x}) = c^k f(\mathbf{x})$ where c is a non-zero constant value.

Table 1: Classification of G-ESS expressions.

Class of G-ESS	C1	C2	C3	C4	C5
Degenerate	✓	✓	✓	x	x
Proper	✓	✓	✓	✓	x
Degenerate and Stable	✓	✓	✓	x	✓
Proper and Stable	✓	✓	✓	✓	✓

Table 2 shows that the Huggins-Roy's family contains all the main G-ESS functions introduced in literature. The special cases with $\beta = 0$ and $\beta = 1$ bring to two undetermined expressions that will be solved and clarified below (when the relationship with Rényi entropy is shown). We can easily note that $1 \leq \text{ESS-H}_N^{(\beta)}(\bar{\mathbf{w}}) \leq N$ for all $\beta \geq 0$. More generally, it is straightforward to observe that the conditions C1, C2, C3 and C4 are fulfilled (with the exception of $\beta = 0$ that does not satisfy C4). Furthermore, the condition C5 is also satisfied as we show below.

Proof. In order to prove that C5 is satisfied, for simplicity let us consider a vector $\bar{\mathbf{v}} = \frac{1}{2}[\bar{\mathbf{w}}, \bar{\mathbf{w}}]$, defined repeating twice the vector $\bar{\mathbf{w}}$ (i.e., $M = 2$). In this case, we have

$$\begin{aligned}
\text{ESS-H}_{2N}^{(\beta)}(\bar{\mathbf{v}}) &= \left(\frac{1}{2^\beta}\sum_{n=1}^{N}\bar{w}_n^\beta + \frac{1}{2^\beta}\sum_{n=1}^{N}\bar{w}_n^\beta\right)^{\frac{1}{1-\beta}}, \\
&= \left(\frac{1}{2^{\beta-1}}\sum_{n=1}^{N}\bar{w}_n^\beta\right)^{\frac{1}{1-\beta}}, \\
&= 2\left(\sum_{n=1}^{N}\bar{w}_n^\beta\right)^{\frac{1}{1-\beta}}, \\
&= 2\ \text{ESS-H}_N^{(\beta)}(\bar{\mathbf{w}}), \quad \forall\beta,
\end{aligned} \tag{22}$$

which is exactly the condition in Eq. (18). The proof can be easily repeated for a value $M > 2$.

Remark. Hence, all G-ESS functions (except for $\beta \to 0$) belonging to the Huggins-Roy's ESS family are *proper and stable*. For $\beta \to 0$, the corresponding ESS is degenerate and stable. Moreover, some specific cases, provided in Table 2, coincide with other proper and stable G-ESS formulas proposed in [19].

Table 2: Relevant special cases contained in the Huggins-Roy's family.

$\beta \to 0$	$\beta = 1/2$	$\beta \to 1$	$\beta = 2$	$\beta = \infty$
$N - N_Z$ *where* N_Z *is the number of zeros in* $\bar{\mathbf{w}}$	$\left(\sum_{n=1}^{N} \sqrt{\bar{w}_n}\right)^2$	$\exp\left(-\sum_{n=}^{N} \bar{w}_n \log \bar{w}_n\right)$ *Perplexity - Eq.* (7) [3, 20]	$\frac{1}{\sum_{n=1}^{N} \bar{w}_n^2}$ *Standard formula in Eq.* (6) - [14]	$\frac{1}{\max[\bar{w}_1,\ldots,\bar{w}_N]}$ *In Eq.* (8) [19]

5 Relationship with the entropy measures

5.1 Relationship with the Rényi entropy

The Rényi entropy [5] is defined as

$$R_N^{(\beta)}(\bar{\mathbf{w}}) \;=\; \frac{1}{1-\beta} \log \left[\sum_{n=1}^{N} \bar{w}_n^{\beta} \right], \qquad \beta > 0, \tag{23}$$

Then, it is straightforward to note that

$$\text{ESS-H}_N^{(\beta)}(\bar{\mathbf{w}}) = \exp\left(R_N^{(\beta)}(\bar{\mathbf{w}})\right) = \left(\sum_{n=1}^{N} \bar{w}_n^{\beta} \right)^{\frac{1}{1-\beta}}, \qquad \beta > 0. \tag{24}$$

i.e., the Huggins-Roy's family contains *diversity indices* derived by the Rényi entropy [5, 13]. For $\beta = 0$, we have $R_N^{(0)}(\bar{\mathbf{w}}) = \log(N - N_Z)$ [5] where $N_Z = \#\big\{\text{all } \bar{w}_n\text{:} \quad \bar{w}_n = 0, \quad \forall n = 1, \ldots, N\big\}$, so that $\text{ESS-H}_N^{(0)}(\bar{\mathbf{w}}) = N - N_Z$. For $\beta = 1$, we have $R_N^{(0)}(\bar{\mathbf{w}}) = -\sum_{n=}^{N} \bar{w}_n \log \bar{w}_n$ [5] then

$$\text{ESS-H}_N^{(1)}(\bar{\mathbf{w}}) = \exp\left(-\sum_{n=}^{N} \bar{w}_n \log \bar{w}_n \right), \tag{25}$$

that is the perplexity in Eq. (7) [3, 20]. The connection with the Rényi entropy shows that the G-ESS functions contained in the Huggins-Roy's family are diversity indices [13]. Moreover, this observation allow us to obtain some theoretical results about $\text{ESS-H}_N^{(\beta)}$. Indeed, for instance, it is well-known that [5]

$$R_N^{(0)}(\bar{\mathbf{w}}) \geq R_N^{(1)}(\bar{\mathbf{w}}) \geq R_N^{(2)}(\bar{\mathbf{w}}) \geq \ldots R_N^{(\beta')}(\bar{\mathbf{w}}) \ldots \geq R_N^{(\infty)}(\bar{\mathbf{w}}),$$

with $\beta' \geq 2$, Then, since $\text{ESS-H}_N^{(\beta)}$ is an increasing monotonic function of $R_N^{(\beta)}$, we can also assert

$$\text{ESS-H}_N^{(0)}(\bar{\mathbf{w}}) \geq \text{ESS-H}_N^{(1)}(\bar{\mathbf{w}}) \geq \text{ESS-H}_N^{(2)}(\bar{\mathbf{w}}) \geq \ldots \text{ESS-H}_N^{(\beta')}(\bar{\mathbf{w}}) \ldots \geq \text{ESS-H}_N^{(\infty)}(\bar{\mathbf{w}}). \tag{26}$$

with $\beta' > 2$. Moreover, since from [5]

$$R_N^{(2)}(\bar{\mathbf{w}}) \leq 2R_N^{(\infty)}(\bar{\mathbf{w}}),$$

we also have

$$\text{ESS-H}_N^{(2)}(\bar{\mathbf{w}}) \leq 2\text{ESS-H}_N^{(\infty)}(\bar{\mathbf{w}}). \tag{27}$$

5.2 Relationship with the Tsallis entropy

Another famous entropy family is the so-called Tsallis entropy [21], defined as

$$T_N^{(\alpha)}(\bar{\mathbf{w}}) = \frac{1}{\alpha - 1} \log \left[1 - \sum_{n=1}^{N} \bar{w}_n^{\alpha} \right], \qquad \alpha > 0. \tag{28}$$

We can obtain a corresponding G-ESS family based on the Tsallis entropy, taking the exponential of $T_N^{(\alpha)}(\bar{\mathbf{w}})$ and after some additional simple operations of translation and scaling, i.e.,

$$\text{ESS-T}_N^{(\alpha)}(\bar{\mathbf{w}}) = \frac{N(N-1)}{(N^{\alpha-1}-1)^{\frac{1}{\alpha-1}}} \exp\left(T_N^{(\alpha)}(\bar{\mathbf{w}}) \right) + 1, \tag{29}$$

$$= \frac{N(N-1)}{(N^{\alpha-1}-1)^{\frac{1}{\alpha-1}}} \left[1 - \sum_{n=1}^{N} \bar{w}_n^{\alpha} \right]^{\frac{1}{\alpha-1}} + 1, \qquad \alpha > 0. \tag{30}$$

Note that

$$1 \leq \text{ESS-T}_N^{(\alpha)}(\bar{\mathbf{w}}) \leq N.$$

Setting $\alpha = 2$, we have

$$\begin{aligned} \text{ESS-T}_N^{(2)}(\bar{\mathbf{w}}) &= N \left(1 - \sum_{n=1}^{N} \bar{w}_n^2 \right) + 1, \\ &= N \text{ Gini-impurity}(\bar{\mathbf{w}}) + 1, \end{aligned} \tag{31}$$

where we have used the definition of the function below,

$$\text{Gini-impurity}(\bar{\mathbf{w}}) = 1 - \sum_{n=1}^{N} \bar{w}_n^2,$$

that is the so-called *Gini impurity* or *Gini's diversity index* or also known as *Gini-Simpson index* in biodiversity field, that is widely used in machine learning within decision tree algorithms [1, 16]. Furthermore, it is interesting to remark that the final form of $\text{ESS-T}_N^{(\alpha)}(\bar{\mathbf{w}})$ resembles the G-ESS family $\text{ESS-S}_N^{(r)}(\bar{\mathbf{w}})$ introduced in [19],

$$\text{ESS-S}_N^{(r)}(\bar{\mathbf{w}}) = \frac{N-1}{N^{\frac{1-r}{r}} - 1} \left[\sum_{n=1}^{N} \bar{w}_n^r \right]^{\frac{1}{r}} + 1 - \frac{N-1}{N^{\frac{1-r}{r}} - 1}, \qquad r > 0,$$

that contains $\text{ESS-S}_N^{(1/2)}(\bar{\mathbf{w}}) = \left(\sum_{n=1}^{N} \sqrt{\bar{w}_n} \right)^2$ for $r = 1/2$, that is a proper and stable ESS formula. However, generally the rest of ESS expressions contained in $\text{ESS-S}_N^{(r)}(\bar{\mathbf{w}})$ and $\text{ESS-T}_N^{(\alpha)}(\bar{\mathbf{w}})$ are not stable. For this reason, in the rest of work we focus on Huggins-Roy ESS family.

6 Numerical experiments

Since all the ESS functions in the Huggins-Roy family are proper and stable and contains all the relevant formulas, we focus our study on this family. First of all, we recall the theoretical definition of ESS in Eq. (5),

$$\mathrm{ESS}_{\texttt{teo}}(h) = N\frac{\mathrm{var}_\pi[\widehat{I}]}{\mathrm{var}_q[\widetilde{I}]}. \tag{32}$$

where, for simplicity, we consider a scalar $x \in \mathbb{R}$ the use of the integrand $h(x) = x$ (in the definition above, we have clarified the dependence on the function h). Namely, $\widehat{I}$ and $\widetilde{I}$ are estimators of the expected value of a random variable X with a target pdf $\bar{\pi}(x)$ (defined below). In this numerical example, we compute approximately via Monte Carlo the theoretical definition $\mathrm{ESS}_{\texttt{teo}}$, and compare them with the G-ESS functions ESS-H$_N^{(\beta)}$. More specifically, we consider a univariate standard Gaussian density as target pdf,

$$\bar{\pi}(x) = \mathcal{N}(x; 0, 1), \tag{33}$$

and also a Gaussian proposal pdf,

$$q(x) = \mathcal{N}(x; \mu_p, \sigma_p^2), \tag{34}$$

with mean μ_p and variance σ_p^2. In all the experiments, we consider $N = 1000$.

6.1 Varying the proposal mean μ_p

In a first analysis, we keep fixed $\sigma_p = 1$ and vary $\mu_p \in [0, 2]$. Figures 2(a)-2(b) depict two scenarios in this experimental setup, corresponding to two specific values of μ_p, 0.5 and 1.5. Clearly, for $\mu_p = 0$ we have the ideal Monte Carlo case, $q(x) \equiv \bar{\pi}(x)$. As μ_p increases, the proposal becomes more different from $\bar{\pi}$. We recall that $N = 1000$. Figure 3(a) shows the theoretical $\mathrm{ESS}_{\texttt{teo}}/N$ curves (solid line) ESS-H$_N^{(2)}/N$ (circles) and ESS-H$_N^{(\infty)}/N$ (squares), averaged over 10^5 independent runs. Note that $\frac{1}{N} \leq \frac{ESS}{N} \leq 1$.

Optimal linear combination of ESS-H$_N^{(2)}$ and ESS-H$_N^{(\infty)}$. The functions ESS-H$_N^{(2)}$ and ESS-H$_N^{(\infty)}$ are the most used and suggested formulas in different studies [12, 19]. Moreover, at least in this simulation scenario, they seem to play the role of upper bound and lower bound of the true value, as shown by Figure 3(a). For this reason, we also consider the linear combination of the G-ESS formulas ESS-H$_N^{(2)}$ and ESS-H$_N^{(\infty)}$,

$$\text{Comb-ESS}_N(\bar{\mathbf{w}}) = a_1\text{ESS-H}_N^{(2)}(\bar{\mathbf{w}}) + a_2\text{ESS-H}_N^{(\infty)}(\bar{\mathbf{w}}). \tag{35}$$

This example suggests the use of

$$\begin{aligned} a_1 &= 0.6245, \\ a_2 &= 0.4289, \end{aligned} \tag{36}$$

obtained using a Least Squares (LS) regression in order to obtain an expression Comb-ESS$_N(\bar{\mathbf{w}})$ as close as possible to the theoretical ESS curve.

Optimal β for ESS-$\mathrm{H}_N^{(\beta)}(\bar{\mathbf{w}})$. Furthermore, we have computed the curves (as function β) of ESS-$\mathrm{H}_N^{(\beta)}(\bar{\mathbf{w}})$ for different values of β, considering a thin grid of β values from 0.2 to 50 with a step of 0.01 (i.e., $\beta \in \mathcal{G}$ denoting $\mathcal{G}$ the thin grid). We consider a L_1 distance between each ESS-$\mathrm{H}_N^{(\beta)}(\bar{\mathbf{w}})$ curve and the theoretical ESS curve,[4], i.e., $|\text{ESS-H}_N^{(\beta)} - \text{ESS}_{\texttt{teo}}|$, and compute

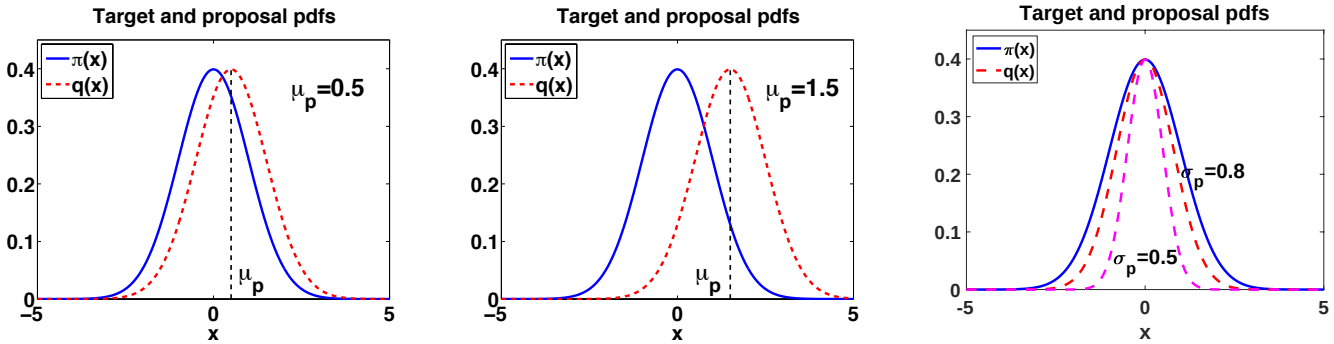
$$\beta^* = \arg\min_{\beta \in \mathcal{G}} |\text{ESS-H}_N^{(\beta)} - \text{ESS}_{\texttt{teo}}|. \tag{37}$$

With this procedure, we obtain

$$\beta^* \approx 4.$$

Discussion of the results. Figure 3(b) shows the curves of the ESS rates corresponding to the theoretical ESS curve (solid line), the best linear combination corresponding to the Eqs. (35)-(39) (squares) and the curve corresponding to ESS-$\mathrm{H}_N^{(\beta^*)}$ (dashed line). First of all, we can note that the linear combination can return values greater than 1 (recall that we are considering ESS/N). Moreover, we can see that the curve corresponding to ESS-$\mathrm{H}_N^{(4)}(\bar{\mathbf{w}})$ fits particularly well in this numerical setup, providing a very close to the theoretical ESS curve. Observe that the approximation provided by ESS-$\mathrm{H}_N^{(4)}$ is virtually perfect for $\mu_p \leq 1$. Hence, in this kind of scenario, we would suggest the use of the expression

$$\text{ESS-H}_N^{(4)}(\bar{\mathbf{w}}) \quad = \quad \left(\frac{1}{\sum_{n=1}^{N} \bar{w}_n^4} \right)^{\frac{1}{3}}. \tag{38}$$



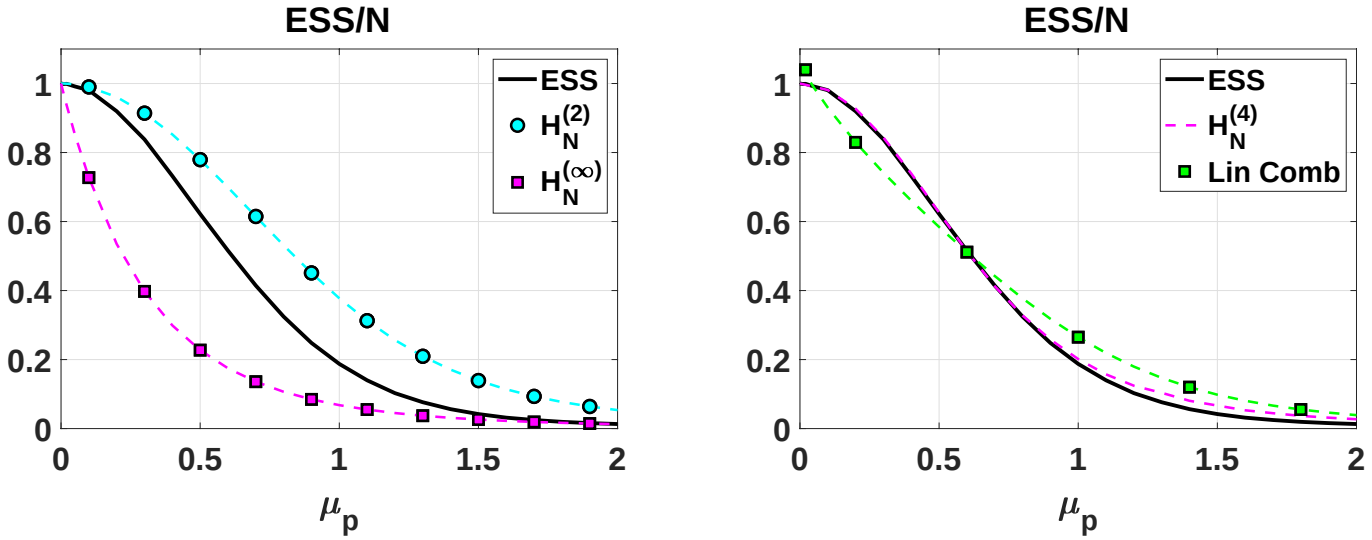
(a) (b) (c)

Figure 2: Target and proposal pdfs: **(a)-(b)** with $\mu_p \in \{0.5, 1.5\}$. The variances in both is set to 1. **(c)** here $\mu_p = 0$ and $\sigma_p \in \{0.5, 0.8\}$.

6.2 Varying the proposal standard deviation σ_p

Now, we keep fixed $\mu_p = 0$ and vary the standard deviation of the proposal $\sigma_p \in [0.5, 1]$. Figure 2(c) depicts the target density and the proposal density for two specific values of σ_p, 0.5 and 0.8,

[4]Recall that these curves are functions of μ_p and are averaged over 10^5 independent runs.



(a) (b)

Figure 3: Ratio of ESS values over N (with $N = 1000$) versus μ_p. The curve corresponding to theoretical ESS value, i.e., $\text{ESS}_{\texttt{teo}}/N$ is shown in black solid line in both figures. In **(a)** the curves of ESS-H$_N^{(2)}/N$ (circles) and ESS-H$_N^{(\infty)}/N$ (squares) are also depicted. In **(b)** we show the curves ESS-H$_N^{(4)}/N$ (dashed line) and the linear combination in Eq. (35)-(39) (squares), as well. The approximation provided by ESS-H$_N^{(4)}$ is virtually perfect for $\mu_p \leq 1$.

used in this experimental setup. We recall that $N = 1000$ and the results have been averaged over 10^5 independent runs. In Figure 4(a), we can observe the results of $\text{ESS}_{\texttt{teo}}/N$ versus σ_p (in solid line), jointly with the curves ESS-H$_N^{(2)}/N$ (given with circles) and ESS-H$_N^{(\infty)}/N$ (shown with squares).

Optimal linear combination of ESS-H$_N^{(2)}$ and ESS-H$_N^{(\infty)}$. Since the formulas ESS-H$_N^{(2)}$ and ESS-H$_N^{(\infty)}$ are the most used in practice, again we consider the linear combination of the G-ESS formulas ESS-H$_N^{(2)}$ and ESS-H$_N^{(\infty)}$,

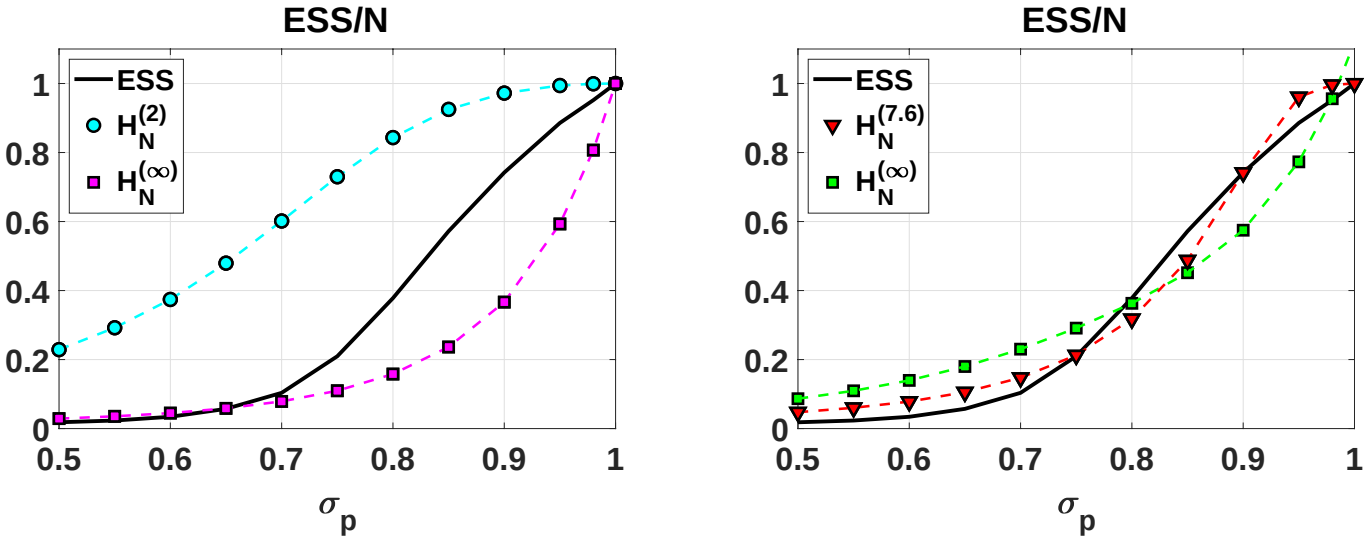
$$\text{Comb-ESS}_N(\bar{\mathbf{w}}) = a_1 \text{ESS-H}_N^{(2)}(\bar{\mathbf{w}}) + a_2 \text{ESS-H}_N^{(\infty)}(\bar{\mathbf{w}}), \tag{39}$$

where in this scenario we get by LS solution

$$\begin{aligned} a_1 &= 0.2715, \\ a_2 &= 0.8483, \end{aligned} \tag{40}$$

hence ESS-H$_N^{(\infty)}$ takes more importance in this scenario. Figure 4(b) provides the curve corresponding to $\text{Comb-ESS}_N(\bar{\mathbf{w}})/N$ with a dashed line and green squares.

Optimal β for ESS-H$_N^{(\beta)}(\bar{\mathbf{w}})$. Furthermore, we have computed the curves (as function β) of ESS-H$_N^{(\beta)}(\bar{\mathbf{w}})$ for different values of β, considering a grid of values of β denoted as $\mathcal{G}$. We consider



(a) (b)

Figure 4: Ratio of ESS values over N (with $N = 1000$) versus σ_p. The curve corresponding to theoretical ESS value, i.e., $\text{ESS}_{\texttt{teo}}/N$ is shown in black solid line in both figures. In **(a)** the curves of ESS-$\text{H}_N^{(2)}/N$ (circles) and ESS-$\text{H}_N^{(\infty)}/N$ (squares) are also depicted. In **(b)** we show the curves ESS-$\text{H}_N^{(7.6)}/N$ (dashed line) and the linear combination in Eq. (40)-(**??**) (squares), as well.

a L_1 distance between each ESS-$\text{H}_N^{(\beta)}(\bar{\mathbf{w}})$ curve and the theoretical ESS curve, and compute

$$\beta^* = \arg\min_{\beta \in \mathcal{G}} |\text{ESS-H}_N^{(\beta)} - \text{ESS}_{\texttt{teo}}|. \tag{41}$$

In this scenario, we obtain

$$\beta^* \approx 7.6.$$

The corresponding curve is depicted in Figure 4(b) with a dashed line and red triangles. We can see that we obtain a very good approximation of $\text{ESS}_{\texttt{teo}}/N$, but slightly worse than in the case described in the previous section. Moreover, here the optimal β^* is ≈ 7.6 whereas, in the previous section, was β^* is ≈ 4.

Discussion of the results. Figure 4(b) shows the curves of the ESS rates corresponding to the theoretical ESS curve (solid line), the best linear combination corresponding to the Eqs. (35)-(39) (green squares) and the curve corresponding to ESS-$\text{H}_N^{(\beta^*)}$ (red triangles). Again the linear combination can return values greater than 1 (recall that we are considering ESS/N). This behavior could be exploited in future works since actually $\text{ESS}_{\texttt{teo}}/N$ can exceed 1 (see [9, Section 3.3]). Moreover, we can see that ESS-$\text{H}_N^{(7.6)}(\bar{\mathbf{w}})$ performs particularly well in this scenario, providing a close to the theoretical ESS curve. Hence, in this setup, we would suggest the use of ESS-$\text{H}_N^{(7.6)}(\bar{\mathbf{w}})$. Only for simplicity in computation and comparison, one could consider the closest

integer and use $\beta = 8$,

$$\text{ESS-H}_N^{(8)}(\bar{\mathbf{w}}) \;=\; \left(\frac{1}{\sum_{n=1}^{N} \bar{w}_n^8}\right)^{\frac{1}{7}}. \tag{42}$$

Finally, it is important to remark that even if the optimal $\beta^* \approx 7.6$ (or 8) is different from the value $\beta^* \approx 4$ suggested in the previous section, however both values differ from 2 (that corresponds to the typical formula employed in the literature) and both values are bigger than 2. The expression with $\beta \to \infty$, i.e., $\text{ESS-H}_N^{(\infty)} = \frac{1}{\max \bar{w}_n}$ seems that can be employed as a lower bound for the theoretical value $\text{ESS}_{\texttt{teo}}$, in both setups. These considerations can be relevant clues for future applications and studies.

7 Conclusions

In this work, we have analyzed alternative effective sample size (ESS) measures for Monte Carlo algorithms based on the importance sampling techniques. We have remarked the connection to the practical ESS formulas used in the literature and entropy families [5]. We have shown that all the ESS functions included in the Huggins-Roy's ESS family fulfill all the required theoretical conditions described in [19], and we have also highlighted the relationship of this family with the Rényi entropy [5]. We have also shown the application of the Gini impurity index as ESS formula and its connection to the Tsallis entropy.
Furthermore, we have studied the performance of different Huggins-Roy's ESS formulas by numerical simulations, introducing also an optimal linear combination of the most promising ESS indices. In two numerical examples, we have obtained the best ESS approximations within the Huggins-Roy's family in two different setups, $\text{ESS} = \left(\frac{1}{\sum_{n=1}^{M} \bar{w}_n^4}\right)^{1/3}$ and $\text{ESS} = \left(\frac{1}{\sum_{n=1}^{M} \bar{w}_n^8}\right)^{1/7}$. These formulas provide a good approximation (and in the first case almost a perfect match) of the theoretical ESS values, in two different considered experimental scenarios. Moreover, the expression $\text{ESS} = \frac{1}{\max \bar{w}_n}$, which corresponds to $\beta \to \infty$, also provides good performance in some specific cases (and playing the role of lower bound of the ESS measures in other cases). All these considerations suggest us that the use of a $\beta > 2$ can more adequate in practical applications, e.g., in order to fight the sample degeneracy and impoverishment within a particle filtering algorithm.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[2] M. F. Bugallo, L. Martino, and J. Corander. Adaptive importance sampling in signal processing. *Digital Signal Processing*, 47:36–49, 2015.

[3] O. Cappé, R. Douc, A. Guillin, J. M. Marin, and C. P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18:447–459, 2008.

[4] O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert. Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, 2004.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York (USA), 1991.

[6] P. M. Djurić, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Míguez. Particle filtering. *IEEE Signal Processing Magazine*, 20(5):19–38, September 2003.

[7] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

[8] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: fifteen years later. *technical report*, 2008.

[9] V. Elvira, L. Martino, and C. P. Robert. Rethinking the Effective Sample Size. *International Statistical Review*, 90(3):525–550, 2022.

[10] D. Gamerman and H. F. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference,*. Chapman & Hall/CRC Texts in Statistical Science, 2006.

[11] N. Gordon, D. Salmond, and A. F. M. Smith. Novel approach to nonlinear and non-Gaussian Bayesian state estimation. *IEE Proceedings-F Radar and Signal Processing*, 140:107–113, 1993.

[12] J. H Huggins and D. M Roy. Convergence of sequential Monte Carlo based sampling methods. *arXiv:1503.00966*, 2015.

[13] L. Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[14] A. Kong. A note on importance sampling using standardized weights. *Technical Report 348, Department of Statistics, University of Chicago*, 1992.

[15] A. Kong, J. S. Liu, and W. H. Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994.

[16] M. Krzywinski and N. Altman. Classification and regression trees. *Nature Methods*, 14(8):757–758, 2017.

[17] F. Liang, C. Liu, and R. Caroll. *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. Wiley Series in Computational Statistics, England, 2010.

[18] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2004.

[19] L. Martino, V. Elvira, and F. Louzada. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401, 2017.

[20] C. P. Robert and G. Casella. *Introducing Monte Carlo Methods with R*. Springer, 2010.

[21] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1):479–487, Jul 1988.