ALTERNATIVE EFFECTIVE SAMPLE SIZE MEASURES FOR IMPORTANCE SAMPLING

L. Martino[◇]*, V. Elvira*[†]*, F. Louzada*[◇]

[†] Dep. of Signal Theory and Communic., Universidad Carlos III de Madrid, Leganés (Spain).
[◇] Institute of Mathematical Sciences and Computing, Universidade de São Paulo, São Carlos (Brazil).

ABSTRACT

The Effective Sample Size (ESS) is an important measure of efficiency in the Importance Sampling (IS) technique. A well-known approximation of the theoretical ESS definition, involving the inverse of the sum of the squares of the normalized importance weights, is widely applied in literature. This expression has become an essential piece within Sequential Monte Carlo (SMC) methods, using adaptive resampling procedures. In this work, first we show that this ESS approximation is related to the Euclidean distance between the probability mass function (pmf) described by the normalized weights and the uniform pmf. Then, we derive other possible ESS functions based on different discrepancy measures. In our study, we also include another ESS measure called *perplexity*, already proposed in literature, that is based on the discrete entropy of the normalized weights. We compare all of them by means of numerical simulations.

Index Terms— Effective Sample Size; Importance Sampling; Perplexity measure; Resampling; Sequential Monte Carlo.

1. INTRODUCTION

The Effective Sample Size (ESS) is a widely used concept for measuring the efficiency of different Monte Carlo methods, such as Markov Chain Monte Carlo (MCMC) [11, 15, 20] and Importance Sampling (IS) techniques [1, 4, 18, 16, 21]. ESS is theoretically defined as the equivalent number of independent samples generated directly from the target distribution, which yields the same efficiency in the estimation obtained by the MCMC or IS algorithms. Thus, a possible mathematical definition [11, 13] 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 MCMC or IS techniques, using with the same number of samples in both estimators.

The most common choice in literature to approximate this theoretical ESS definition is $\widehat{ESS} \approx \frac{1}{\sum_{n=1}^{M} \bar{w}_n^2}$, which involves (only) the normalized importance weights $\bar{w}_n$, $n = 1, \ldots, N$ [6, 7, 14, 21]. This expression presents different weaknesses since it has been obtained after several approximations of the theoretical definition (see [17] for further details). Another measure called *perplexity*, involving the discrete entropy [5] of the normalized weights has been also proposed in [2]; see also [21, Chapter 4], [9, Section 3.5].

However, the approximation $\widehat{ESS}$ is widely used in different Sequential Monte Carlo (SMC) methods (a.k.a., particle filtering algorithms) [7, 8, 6, 12, 19]. A key point for the success of a SMC method is the use of resampling procedures, that are applied for avoiding the particle degeneracy [6, 7]. However, the application of resampling increases the variance of the Monte Carlo estimators so that one desire to employ resampling steps parsimoniously, only when it is strictly required. This adaptive implementation of the resampling procedure needs the use of an approximation of the ESS [6, 16, 21]. We show that $\widehat{ESS}$ is related to Euclidean distance between the multinomial probability mass function (pmf) defined by the normalized weights $\bar{w}_n$, $n = 1, \ldots, N$, and the discrete uniform pmf. When the pmf defined by $\bar{w}_n$ is close to the discrete uniform pmf, $\widehat{ESS}$ provides high values otherwise, when the pmf defined by $\bar{w}_n$ is concentrated mainly in one weight, $\widehat{ESS}$ provides small values. We deduce other possible ESS functions based on different distances. We compare them by means of numerical simulations. This analysis shows that (at least) one novel ESS expression, defined as the inverse of the maximum of the normalized weights, $\frac{1}{\max[\bar{w}_1, \ldots, \bar{w}_N]}$, presents interesting features from a theoretical and practical point of view and it can be considered a valid alternative to the standard formula $\frac{1}{\sum_{n=1}^{M} \bar{w}_n^2}$.

2. EFFECTIVE SAMPLE SIZE 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 $h(\mathbf{x})$,

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

This work has been supported by the ERC grant 239784 and AoF grant 251170, the Spanish government through the OTOSiS (TEC2013-41718-R), by the Grant 2014/23160-6 of the São Paulo Research Foundation (FAPESP) and by the Grant 305361/2013-3 of the National Council for Scientific and Technological Development (CNPq).

which we desire to approximate using a Monte Carlo approach. If we are able to draw N independent samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ 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 $n = 1, \ldots, N$. However, in general, generating samples directly from the target, $\bar{\pi}(\mathbf{x})$, is impossible. Alternatively, we can draw N samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ 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 $n = 1, \ldots, N$, according to the importance sampling (IS) approach. Defining the normalized weights,

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

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 $\mathbf{x}_n \sim q(\mathbf{x})$, $n = 1, \ldots, N$. 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, 12, 19], it is necessary to measure the loss 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 [13],

$$ESS = N \frac{\text{var}_{\pi}[\widehat{I}]}{\text{var}_{q}[\widetilde{I}]}. \tag{5}$$

2.1. Approximations of ESS

Finding a useful expression of ESS derived analytically from the theoretical definition above is not straightforward. Then, different derivations [13, 14], [7, Chapter 11], [21, 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, 16, 21], is

$$\widehat{ESS} = P_N(\bar{\mathbf{w}}), \tag{6}$$

$$= \frac{1}{\sum_{i=1}^{N} \bar{w}_n^2} = \frac{\left(\sum_{i=1}^{N} w_n\right)^2}{\sum_{i=1}^{N} w_n^2}, \tag{7}$$

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. An interesting property of the expression (7) is

$$1 \leq P_N(\bar{\mathbf{w}}) \leq N. \tag{8}$$

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 [17]). Another similar measure, called *perplexity*, has been proposed in literature [2, 21] based only on the normalized importance weights,

$$\widehat{ESS} = \text{Per}_N(\bar{\mathbf{w}}) = 2^{H(\bar{\mathbf{w}})} \tag{9}$$

where

$$H(\bar{\mathbf{w}}) = -\sum_{n=2}^{N} \bar{w}_n \log \bar{w}_n$$

is the discrete entropy of the vector $\bar{\mathbf{w}}$ [5].[2] Note that, $1 \leq \text{Per}_N(\bar{\mathbf{w}}) \leq N$.

3. ESS BASED ON DISCREPANCY MEASURES

Many population Monte Carlo (PMC) [3, 10] or sequential Monte Carlo (SMC) methods [7, 12], employ resampling steps for updating the parameters of the used proposal functions. On the one hand, PMC and SMC suffer the so-called particle degeneracy, i.e., after some iterations only one sample is statistically relevant in terms of importance weights. This problem could be solved by applying resampling procedures. However, on the other hand, the application of resampling yields loss of diversity in the set of samples (and incorporating additional variance in the Monte Carlo estimators). Therefore, one often attempts to apply resampling steps only in certain specific iterations, when it is considered strictly required.

In the standard multinomial resampling, the indices of the particles used in the next generation are drawn according to a multinomial probability mass function (pmf) defined by the normalized weights $\bar{w}_n = \frac{w_n}{\sum_{i=1}^{N} w_i}$, with $n = 1, \ldots, N$. Ideally, if the samples were drawn directly from the target distribution all the weights w_n would be equal, so that $\bar{w}_n = \frac{1}{N}$, $n = 1, \ldots, N$. We denote

$$\bar{\mathbf{w}}^* = \left[\frac{1}{N}, \ldots, \frac{1}{N}\right], \tag{10}$$

that is the vector with equal components $\bar{w}_n = \frac{1}{N}$, $n = 1, \ldots, N$. It important to note that the inverse is not always true: namely the scenario $\bar{w}_n = \frac{1}{N}$, $n = 1, \ldots, N$, could occur even if the proposal density is different from the target. Hence, in general, in this case we can assert $ESS \leq N$

[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 $\bar{\pi}(\mathbf{x})$.

[2] Different ESS approximations involving the discrete entropy can be designed. However, the perplexity satisfies certain important properties [17].

(considering independent samples). The other extreme case is

$$\bar{\mathbf{w}}^{(j)} = [\bar{w}_1 = 0, \ldots, \bar{w}_j = 1, \ldots, \bar{w}_N = 0], \quad (11)$$

i.e., $\bar{w}_j = 1$ and $\bar{w}_n = 0$ (it occurs only if $\pi(\mathbf{x}_n) = 0$), for $n \neq j$ with $j \in \{1, \ldots, N\}$. In general, in this case $ESS \leq 1$. Both approximations P_N and Per_N, based only on the information given by the vector $\bar{\mathbf{w}}$, apply an *optimistic approach* setting $\widehat{ESS} = N$ and $\widehat{ESS} = 1$ in the extreme scenarios described above. The underlying idea behind both formulas $P_N(\bar{\mathbf{w}})$, $\text{Per}_N(\bar{\mathbf{w}})$, is that if the pmf $\bar{w}_n$, $n = 1, \ldots, N$, is reasonably close to the discrete uniform pmf $\mathcal{U}\{1, 2, \ldots, N\}$ then the resampling is not needed. Otherwise, the resampling is applied. Below, we show that the formula P_N is related to the Euclidean distance between these two pdfs. We derive some alternative ESS functions employing other kind of distances between the pmf represented by weights $\bar{w}_n$ and the discrete uniform pmf. For further information see [17].

Euclidean distance L_2. Let us consider the Euclidean distance L_2 between the discrete uniform pmf $\mathcal{U}\{1, 2, \ldots, N\}$ and the pmf given by the normalized weights $\bar{w}_n$, i.e,

$$\begin{aligned} ||\bar{\mathbf{w}} - \bar{\mathbf{w}}^*||_2 &= \sqrt{\sum_{n=1}^{N} \left(\bar{w}_n - \frac{1}{N}\right)^2} \\ &= \sqrt{\left(\sum_{n=1}^{N} \bar{w}_n^2\right) + N\left(\frac{1}{N^2}\right) - \frac{2}{N}\sum_{n=1}^{N} \bar{w}_n} \\ &= \sqrt{\left(\sum_{n=1}^{N} \bar{w}_n^2\right) - \frac{1}{N}} \\ &= \sqrt{\frac{1}{P_N(\bar{\mathbf{w}})} - \frac{1}{N}}. \end{aligned} \quad (12)$$

Therefore, maximizing P_N is equivalent to minimizing the Euclidean distance between the pmf $\bar{w}_n$ and the discrete uniform pmf.

Distance L_1. Given the previous observations, we can attempt to obtain other suitable ESS formulas, employing other distances. Let us define two disjoint sets of weights

$$\{\bar{w}_1^+, \ldots, \bar{w}_{N^+}^+\} = \{\text{all } \bar{w}_n\text{:} \quad \bar{w}_n \geq 1/N, \quad \forall n = 1, \ldots, N\},$$
$$\{\bar{w}_1^-, \ldots, \bar{w}_{N^-}^-\} = \{\text{all } \bar{w}_n\text{:} \quad \bar{w}_n < 1/N, \quad \forall n = 1, \ldots, N\},$$

where $N^+ = \#\{\bar{w}_1^+, \ldots, \bar{w}_{N^+}^+\}$ and $N^- = \#\{\bar{w}_1^-, \ldots, \bar{w}_{N^+}^-\}$. Clearly, $N^- + N^+ = N$ and $\sum_{i=1}^{N^+} \bar{w}_i^+ + \sum_{i=1}^{N^-} \bar{w}_i^- = 1$.

Considering the L_1 distance, we can write

$$\begin{aligned} ||\bar{\mathbf{w}} - \bar{\mathbf{w}}^*||_1 &= \sum_{n=1}^{N} \left|\bar{w}_n - \frac{1}{N}\right| \\ &= \sum_{i=1}^{N^+} \left(\bar{w}_i^+ - \frac{1}{N}\right) + \sum_{j=1}^{N^-} \left(\frac{1}{N} - \bar{w}_j^-\right) \\ &= \sum_{i=1}^{N^+} \bar{w}_i^+ - \sum_{i=1}^{N^-} \bar{w}_i^- - \frac{N^+ - N^-}{N} \end{aligned} \quad (13)$$

and replacing the relationships $\sum_{i=1}^{N^-} \bar{w}_i^- = 1 - \sum_{i=1}^{N^+} \bar{w}_i^+$ and $N^- = N - N^+$,

$$\begin{aligned} ||\bar{\mathbf{w}} - \bar{\mathbf{w}}^*||_1 &= 2\left[\sum_{i=1}^{N^+} \bar{w}_i^+ - \frac{N^+}{N}\right], \\ &= 2\frac{N\sum_{i=1}^{N^+} \bar{w}_i^+ - N^+}{N} = 2\left[\frac{N - Q_N(\bar{\mathbf{w}})}{N}\right] + 2, \end{aligned}$$

where

$$\widehat{ESS} = Q_N(\bar{\mathbf{w}}) = -N\sum_{i=1}^{N^+} \bar{w}_i^+ + N^+ + N, \quad (14)$$

Note that $1 \leq Q_N(\bar{\mathbf{w}}) \leq N$, with $Q_N(\bar{\mathbf{w}}^*) = N$ and $Q_N(\bar{\mathbf{w}}^{(i)}) = 1$ for all $i \in \{1, \ldots, N\}$. Maximizing Q_N is equivalent to minimizing the L_1 distance between the pmf $\bar{w}_n$ and the uniform pmf.

Norm L_∞. Considering the distance between the vector $\bar{\mathbf{w}}$ and the vector of null entries (i.e. the norm of $\bar{\mathbf{w}}$), we can obtain other suitable ESS formulas. For instance, we can consider the norm L_∞, i.e.,

$$||\bar{\mathbf{w}}||_\infty = \max\left[|\bar{w}_1|, \ldots, |\bar{w}_N|\right] = \frac{1}{D_N(\bar{\mathbf{w}})}, \quad (15)$$

where

$$\widehat{ESS} = D_N(\bar{\mathbf{w}}) = \frac{1}{\max\left[\bar{w}_1, \ldots, \bar{w}_N\right]}, \quad (16)$$

is another valid ESS measure. We have also $1 \leq D_N(\bar{\mathbf{w}}) \leq N$, with $D_N(\bar{\mathbf{w}}^*) = N$ and $D_N(\bar{\mathbf{w}}^{(i)}) = 1, \forall i \in \{1, \ldots, N\}$.

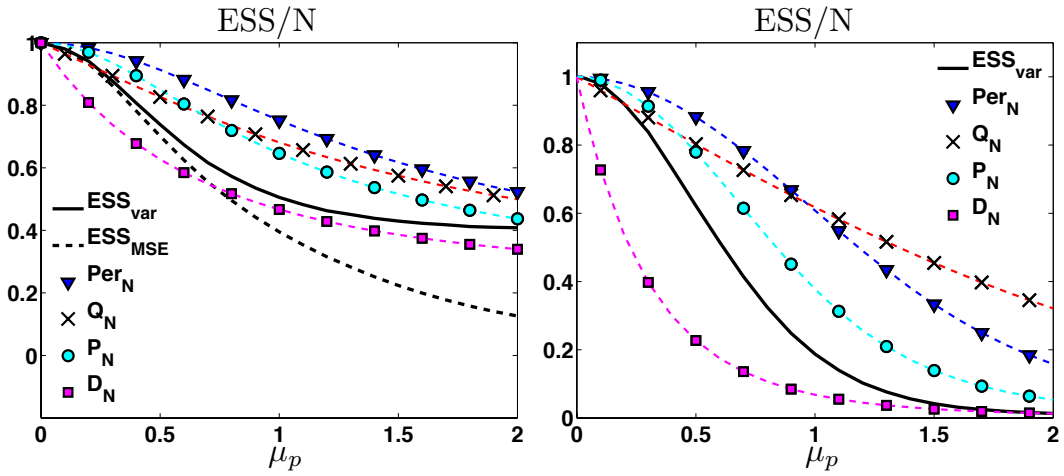
4. NUMERICAL SIMULATIONS

Let us recall the theoretical definition of ESS in Eq. (5),

$$ESS_{var} = N\frac{\text{var}_\pi[\widehat{I}]}{\text{var}_q[\widetilde{I}]}. \quad (17)$$

Since $\widetilde{I}$ is biased and its bias is not negligible for small N, a more convenient definition is

$$ESS_{MSE} = N\frac{\text{MSE}_\pi[\widehat{I}]}{\text{MSE}_q[\widetilde{I}]} = N\frac{\text{var}_\pi[\widehat{I}]}{\text{MSE}_q[\widetilde{I}]}. \quad (18)$$



(a) $N = 5$. (b) $N = 1000$.

Fig. 1. Rates $\frac{ESS}{N}$ as function of μ_p, corresponding to the theoretical values ESS_{var} (solid line), ESS_{MSE} (dashed line; shown only in (a)), P_N (circles), D_N (squares), Q_N (x-marks), and Per_N (triangles down).

considering the Mean Square Error (MSE) of the estimators, instead of only the variance. For large values of N the difference between the two definitions is negligible since the bias of $\widetilde{I}$ tends to zero. In this section, we compute approximately via Monte Carlo the theoretical definitions ESS_{var}, ESS_{MSE}, and we compare them with the ESS functions presented in the previous section. More specifically, we consider a univariate standard Gaussian density as target pdf,

$$\bar{\pi}(x) = \mathcal{N}(x; 0, 1), \tag{19}$$

and also a Gaussian proposal pdf,

$$q(x) = \mathcal{N}(x; \mu_p, \sigma_p^2), \tag{20}$$

with mean μ_p and variance σ_p^2. We set $\sigma_p = 1$ and vary $\mu_p \in [0, 2]$. 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 test $N \in \{5, 1000\}$. Figure 1 shows the (approximated) theoretical ESS curves and the curves corresponding to different ESS formulas, averaged over 10^5 independent runs. More specifically, we provide the rates $\frac{ESS}{N}$. Note that $\frac{1}{N} \leq \frac{ESS}{N} \leq 1$. For $N = 1000$, the difference between ESS_{var} and ESS_{MSE} is negligible, so that we only show ESS_{var}.

Figure 1(a) shows the results for $N = 5$. First of all, we can observe that ESS_{var} and ESS_{MSE} are very close when $\mu_p \approx 0$ (i.e., $q(x) \approx \bar{\pi}(x)$) but they differ substantially when the bias increases. Moreover, P_N and D_N also provide good approximations of ESS_{var}. Note that ESS_{var} is always contained between D_N and P_N. Figure 1(b) shows the results for $N = 1000$. The formula P_N provides the closest curve to ESS_{var}. The ESS function D_N gives a good approximation when μ_p increases, i.e., the scenario becomes worse from a Monte Carlo point of view. Again, ESS_{var} is always contained between D_N and P_N.

We can conclude that in general when the proposal differs substantially from the target, D_N provides the best results (i.e., closer to the theoretical values), whereas in better scenarios and large N, P_N seems to be the best approximations. The ESS function D_N seems to perform better than P_N when the number of particles N is small. The ESS function Q_N provides the best approximation when the proposal is very similar to the target in both cases $N = 5$ and $N = 1000$. In the analyzed scenarios, the perplexity does not seem a suitable approximation of the theoretical ESS values.

5. CONCLUSIONS

In this work, we have introduced alternative ESS functions for sampling algorithms based on IS. They are derived based on discrepancy measures between the multinomial pmf defined by the normalized weights $\bar{w}_n$, $n = 1, \ldots, N$, and the discrete uniform pmf (or the vector of null entries). We have shown that the standard ESS approximation P_N is related to the Euclidean distance L_2. Another measure called *perplexity* [2, 21] can be included in this class of ESS approximations based on discrepancy measure, since it is based on the discrete entropy of the normalized weights $\bar{w}_n$, $n = 1, \ldots, N$. We have tested and compared all them by numerical simulations.

At least one of them, $D_N(\bar{\mathbf{w}}) = \frac{1}{\max[\bar{w}_1, \ldots, \bar{w}_N]}$, presents interesting features and some benefit, compared to the rest of ESS formulas, when the proposal function differs substantially from the target distribution. Moreover, $D_N(\bar{\mathbf{w}})$ seems to behave as a "lower bound" for the theoretical ESS definition, in our simulations.

References

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

[2] 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.

[3] 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.

[4] S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *arXiv:1511.01437*, 2015.

[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, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York (USA), 2001.

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

[10] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Improving population Monte Carlo: Alternative weighting and resampling schemes. *viXra:1601.0174*, 2015.

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

[12] 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.

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

[14] 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.

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

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

[17] L. Martino, V. Elvira, and M. F. Louzada. Effective Sample Size for importance sampling based on the discrepancy measures. *arXiv:1602.03572*, 2016.

[18] L. Martino, V. Elvira, D. Luengo, and J. Corander. An adaptive population importance sampler: Learning from the uncertanity. *IEEE Transactions on Signal Processing*, 63(16):4422–4437, 2015.

[19] L. Martino, J. Read, V. Elvira, and F. Louzada. Cooperative parallel particle filters for on-line model selection and applications to urban mobility. *viXra:1512.0420*, 2015.

[20] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

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

View publication stats