A COMPARISON OF CLIPPING STRATEGIES FOR IMPORTANCE SAMPLING

L. Martino[†], *V. Elvira*[⋆], *J. Míguez*[†], *A. Artés-Rodríguez*[†]. *P. M. Djurić*[*],

[†] Universidad Carlos III de Madrid, Leganés (Spain).
[⋆] IMT Lille Douai & CRIStAL (UMR 9189), Villeneuve d'Ascq (France).
[*] Stony Brook University, Stony Brook, New York (USA).

ABSTRACT

Importance Sampling (IS) methods approximate a targeted distribution with a set of weighted samples, drawn from a proposal distribution. Unfortunately, a mismatch between the proposal and the targeted distribution may endanger the performance of the estimators. In this paper, we focus on the so-called nonlinear IS (NIS) framework, where a nonlinear function is applied to the standard importance weights (IWs). The aim of this transformation is to mitigate the well-known problem of the degeneracy of the IWs by controlling the weight variability. We consider the clipping transformation and test its robustness with respect to the choice of the clipping value. We also propose a novel NIS methodology, where not only a subset of weights is modified a posteriori, but also the corresponding samples are moved. We compare these NIS schemes with standard IS and Monte Carlo methods by means of illustrative numerical examples.

***Index Terms*—** Monte Carlo methods, Importance Sampling, Bayesian Inference, Parameter Estimation, Variance Reduction methods

1. INTRODUCTION

Many relevant problems in statistical signal processing involve the numerical approximation of complicated multidimensional integrals which have neither a closed-form expression nor an accurate analytic approximation. Importance sampling (IS) is a well-known class of Monte Carlo methods, widely applied in signal processing [1, 2, 3, 4, 5]. IS methods approximate moments of a random variable of interest by sets of N samples, drawn from a proposal probability density function (pdf) different from the targeted one, and weights, assigned to the samples in order to measure their adequacy in approximating the target pdf.

In the literature, several advanced and adaptive IS schemes can be found [6, 7, 8, 9, 10, 11, 12, 13]. In this work, we focus on a recent improvement of the standard IS approach, the so-called *nonlinear IS* method (NIS) [5, 14, 15]. NIS applies a nonlinear transformation of the importance weights in order to prevent the degeneracy problem of conventional importance samplers and thereby avert their degradation in performance.

We consider two different "clipping" transformations that consist in clipping a certain number, $N_C \leq N$, of highest importance weights. In the first one, the N_C highest weights are set equal to their empirical mean. In the second one, the N_C highest weights are set equal to their minimum value. It is possible to show that if $N_C \leq \sqrt{N}$, the resulting estimators are consistent [5, 15] (i.e., the estimator converge to the true value as $N \to \infty$). We point out that a scheme related to the clipping strategy is the so-called *truncated IS* technique [10] (see also [9] and [16, 17, 12] for further related discussions). Furthermore, we extend the NIS approach combining it with another recent contribution, the Group Importance Sampling (GIS) technique [18]. In GIS, a single weighted sample is used for compressing the information contained in a population of weighted samples. The NIS scheme based on GIS consists in clipping the N_C highest weights (using their empirical mean) and move all the corresponding samples (particles) to a suitable summary sample. In this way, the estimators of the normalizing constant of the target and of the expected value of the target are the same as that of the standard IS approach. In this work, we compare all the previous schemes by numerical simulations in order to clarify the performance, strengths and drawbacks of the different methodologies.

The paper is organized as follows. In the next section we formulate the problem and provide some background on importance sampling. In Section 3, we propose two novel NIS schemes and in Section 4 we present numerical results that show their performances. We make our concluding remarks in Section 5.

2. PROBLEM STATEMENT AND BACKGROUND

2.1. Bayesian Inference

In many applications, the interest lies in obtaining the posterior density function (pdf) of set of unknown parameters given the observed data. More specifically, denoting the vector of unknowns as $\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^{d_x}$ and the observed data as

$\mathbf{y} \in \mathbb{R}^{d_y}$, the posterior pdf is defined as

$$\bar{\pi}(\mathbf{x}|\mathbf{y}) = \frac{\ell(\mathbf{y}|\mathbf{x})g(\mathbf{x})}{Z(\mathbf{y})} \propto \pi(\mathbf{x}|\mathbf{y}) = \ell(\mathbf{y}|\mathbf{x})g(\mathbf{x}), \quad (1)$$

where $\ell(\mathbf{y}|\mathbf{x})$ is the likelihood function, $g(\mathbf{x})$ is the prior pdf, and $Z(\mathbf{y})$ is the so-called marginal likelihood.[1] Generally, we need to compute the expected value of some function of $\mathbf{x}$, $\mathbf{f}(\mathbf{x})$, with respect to the posterior pdf of $\mathbf{x}$, $\bar{\pi}(\mathbf{x})$, i.e.,

$$\mathbf{I}(\mathbf{f}) = \frac{1}{Z}\int_{\mathcal{D}} \mathbf{f}(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}, \quad (2)$$

where $\mathbf{f}(\mathbf{x}) : \mathcal{D} \to \mathbb{R}^{d_x}$ can be any integrable transformation of $\mathbf{x}$. As an example, if $\mathbf{f}(\mathbf{x}) = \mathbf{x}$, then the integral $\mathbf{I}(\mathbf{x})$ represents the expected value of the target pdf; since $\bar{\pi}(\mathbf{x})$ is a posterior pdf, then its expected value is the so-called Minimum Mean Square Error (MMSE) estimator.

2.2. The Monte Carlo method

In many practical scenarios, we cannot obtain an analytical solution of Eq. (2) and Monte Carlo methods are used to approximate it [2, 3]. If we are able to draw samples directly from $\bar{\pi}(\mathbf{x})$, we can apply the ideal Monte Carlo approach shown in Table 1. When drawing directly from $\bar{\pi}(\mathbf{x})$ is not possible, the Importance Sampling (IS) approach given in Table 2 can be applied.[2] In this case, the samples are generated from a proposal pdf $q(\mathbf{x})$, and then suitable importance weights (IWs) are assigned to each sample in order to provide a final consistent estimator. Moreover, the approach IS provides an unbiased estimator, $\widehat{Z} = \frac{1}{N}\sum_{n=1}^{N} w_n$, of the marginal likelihood Z.

Table 1: Standard Monte Carlo method

- **Initialization:** Choose $N \in \mathbb{N}^+$.
  1. Draw $\mathbf{x}_n \sim \bar{\pi}(\mathbf{x})$, $\quad n = 1, ..., N$.
  2. Return $\{\mathbf{x}_n\}_{n=1}^N$ or $\widehat{\mathbf{I}}(\mathbf{f}) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{f}(\mathbf{x}_n)$.

3. NONLINEAR IMPORTANCE SAMPLING (NIS)

The key feature of NIS approach is to compute a set of *transformed importance weights* (TIWs) by applying a nonlinear function to the standard IWs [14] (see also [10]). The aim of this transformation is to mitigate the well-known problem of degeneracy of the IWs (common to many IS methods, see [14]) by controlling the weight variability. Several possibilities exist for the choice of the nonlinear transformation. Here we present two.

[1] From now on, we remove the dependence on $\mathbf{y}$ to simplify the notation.

[2] Differents sampling algorithms are also available, such as rejection sampling schemes and Markov Chain Monte Carlo (MCMC) methods [3, 19]. However, in this work, we focus on importance sampling.

Table 2: Importance Sampling

- **Initialization:** Choose $q(\mathbf{x})$ and $N \in \mathbb{N}^+$.
  1. Draw $\mathbf{x}_n \sim q(\mathbf{x})$, $\quad n = 1, ..., N$.
  2. Assign the weights $w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}$, and normalize them as
  $$\bar{w}_n = \frac{w_n}{\sum_{j=1}^{N} w_j}, \quad (3)$$
  for $n = 1, \ldots, N$.
  3. Return $\{\mathbf{x}_n, \bar{w}_n\}_{n=1}^N$ or $\widehat{\mathbf{I}}(\mathbf{f}) = \sum_{n=1}^{N} \bar{w}_n \mathbf{f}(\mathbf{x}_n)$.

3.1. NIS based on Clipping

First we consider a clipping transformation that consists in clipping the standard weights that are above a certain threshold. More specifically, the weights with highest values are modified, setting all these weights to a constant value η. As a consequence, a sufficient number of "flat" (modified) weights is guaranteed in the regions of the space of $\mathbf{x}$ where the standard weights were larger. The clipping strategy automatically increases the approximated Effective Sample Size (ESS) function [20].

In this work, instead of choosing a threshold, we decide a certain number $N_C \leq N$ of weights to be clipped. It is possible to show that if $N_C \leq \sqrt{N}$, then the resulting estimators are consistent, i.e., they converge to $\mathbf{I}$ as $N \to \infty$ [5, 15]. The NIS algorithm based on clipping is shown Table 3.

We consider two possibilities for the choice of η (see step 4): either the empirical mean or the minimum value of the set weights to be transformed. With the former choice, NIS keeps unaltered the estimator $\widehat{Z}$ of the marginal likelihood with respect to the standard IS approach [15]. In the latter, the transformed weights introduce a bias in $\widehat{Z}$ (see [14] for more details). Note that the positions of the particles $\mathbf{x}_n$ remains unchanged with respect to the standard IS method.

3.2. NIS based on Group Importance Sampling

As a second NIS scheme, we consider the application of NIS jointly with the so-called Group Importance Sampling (GIS) [18]. GIS considers the assignation of a single weighted sample which compresses the information contained in a population of weighted samples. Namely, the information contained in different sets of weighted samples is compressed by using only one, yet properly selected, sample (particle), and one suitable weight. In this case, we consider only one group, containing the particles corresponding to the N_C greatest IWs. The resulting algorithm is shown in Table 4. As in NIS of Table 3, the highest weights are set equal to $\eta = \frac{1}{N_C}\sum_{i=1}^{N_C} w_{j_i}$ where $w_{j_1} \geq w_{j_2} \geq ... \geq w_{j_N}$. The corresponding particles are moved, setting $\widetilde{\mathbf{x}}_{j_i} = \boldsymbol{\mu}$ where

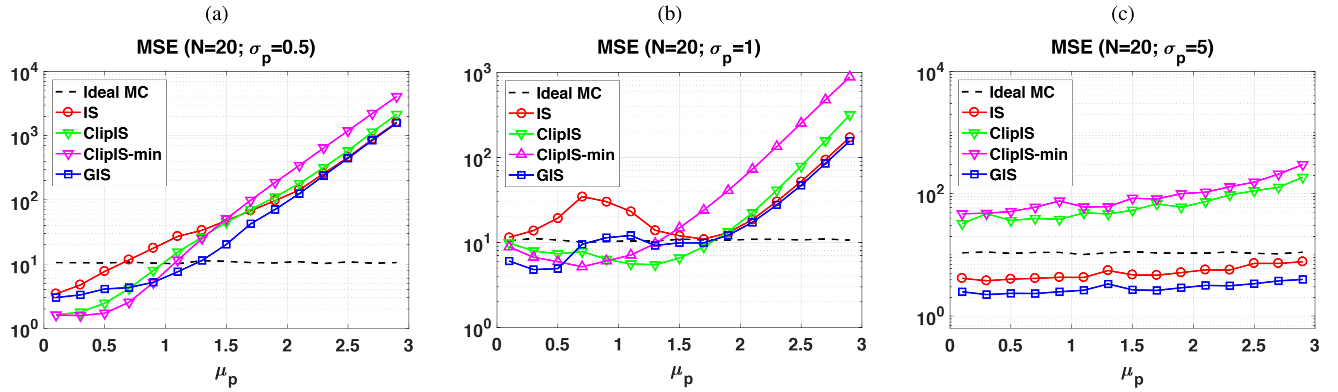

Fig. 1: MSE in estimation of the first 5 non-central moments of the target pdf (shown in semilog scale), averaged over $5 \cdot 10^5$ independent runs. We vary the mean of the proposal pdf $\mu_p \in [0.1, 3]$, set $N = 20$ and consider different values of $\sigma_p\{0.5, 1, 5\}$: in **(a)** $\sigma_p = 0.5$, in **(b)** $\sigma_p = 1$, and in **(c)** $\sigma_p = 5$.

Table 3: Nonlinear Importance Sampling (NIS) based on Clipping

- **Initialization:** Choose $q(\mathbf{x})$, $N \in \mathbb{N}^+$ and $N_C \leq N$.

1. Draw $\mathbf{x}_n \sim q(\mathbf{x})$, $\quad n = 1, ..., N$.
2. Assign the weights $w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}$, for $n = 1, \dots, N$.
3. Sort the weights in descending order,
$$w_{j_1} \geq w_{j_2} \geq ... \geq w_{j_N}. \quad (4)$$
4. Compute $\eta = \frac{1}{N_C}\sum_{i=1}^{N_C} w_{j_i}$ or, alternatively, $\eta = \min\{w_{j_1}, \dots, w_{j_{N_C}}\}$.
5. Set $\widetilde{w}_{j_i} = \eta$ for $i = 1, \dots, N_C$, and set $\widetilde{w}_{j_i} = w_{j_i}$ for $i = N_C + 1, \dots, N$.
6. Normalize the clipped weights
$$\bar{w}_n = \frac{\widetilde{w}_n}{\sum_{j=1}^N \widetilde{w}_j}, \quad (5)$$
for $n = 1, \dots, N$.
7. Return $\{\mathbf{x}_n, \bar{w}_n\}_{n=1}^N$ or $\widehat{\mathbf{I}}(\mathbf{f}) = \sum_{n=1}^N \bar{w}_n \mathbf{f}(\mathbf{x}_n)$.

$\boldsymbol{\mu} = \frac{1}{\sum_{k=1}^{N_C} w_{j_k}} \sum_{i=1}^{N_C} w_{j_k}\mathbf{x}_{j_i}$. The positions of the remaining samples is unchanged. Furthermore, note that NIS based on GIS keeps unaltered the estimator $\widehat{Z}$ (i.e., the estimator of the marginal likelihood) and $\widehat{\mathbf{I}}(\mathbf{x})$ (i.e., the estimator of the expected value of the target) with respect to the standard IS approach [15]. It is possible to show that if $N_C \leq \sqrt{N}$, then the resulting estimators are consistent exactly as in NIS based on clipping (e.g., see [5]). Indeed, the rate $\frac{N_C}{N} = \frac{\sqrt{N}}{N} \to 0$ as $N \to +\infty$, so that the NIS scheme approaches the performance of a standard IS method if $N_C \leq \sqrt{N}$.

Table 4: NIS based on GIS

- **Initialization:** Choose $q(\mathbf{x})$, $N \in \mathbb{N}^+$ and $N_C \leq N$.

1. Draw $\mathbf{x}_n \sim q(\mathbf{x})$, $\quad n = 1, ..., N$.
2. Assign the weights $w_n = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}$, for $n = 1, \dots, N$.
3. Sort the weights in descending order,
$$w_{j_1} \geq w_{j_2} \geq ... \geq w_{j_N}. \quad (6)$$
4. Compute $\eta = \frac{1}{N_C}\sum_{i=1}^{N_C} w_{j_i}$ and
$$\boldsymbol{\mu} = \frac{1}{\sum_{k=1}^{N_C} w_{j_k}} \sum_{i=1}^{N_C} w_{j_k}\mathbf{x}_{j_i}.$$
5. Set $\widetilde{w}_{j_i} = \eta$, $\widetilde{\mathbf{x}}_{j_i} = \boldsymbol{\mu}$, for $i = 1, \dots, N_C$, and set $\widetilde{w}_{j_i} = w_{j_i}$, $\widetilde{\mathbf{x}}_{j_i} = \mathbf{x}_{j_i}$, for $i = N_C + 1, \dots, N$.
6. Normalize the clipped weights
$$\bar{w}_n = \frac{\widetilde{w}_n}{\sum_{j=1}^N \widetilde{w}_j}, \quad (7)$$
for $n = 1, \dots, N$.
7. Return $\{\widetilde{\mathbf{x}}_n, \bar{w}_n\}_{n=1}^N$ or $\widehat{\mathbf{I}}(\mathbf{f}) = \sum_{n=1}^N \bar{w}_n \mathbf{f}(\widetilde{\mathbf{x}}_n)$.

4. NUMERICAL RESULTS

We test the different Monte Carlo techniques on a toy example where the ground truth is known. We consider a univariate

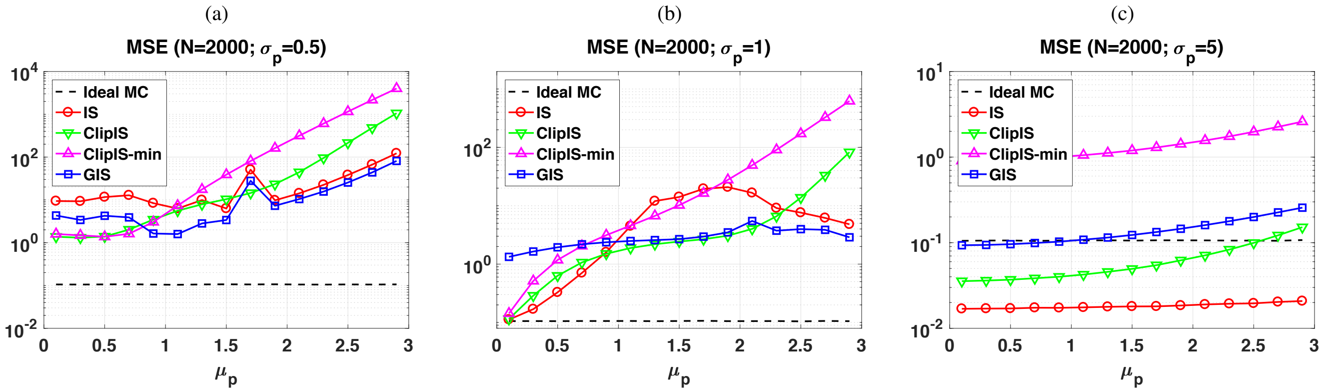

Fig. 2: MSE in estimation of the first 5 non-central moments of the target pdf (shown in semilog scale), averaged over $5 \cdot 10^5$ independent runs. We vary the mean of the proposal pdf $\mu_p \in [0.1, 3]$, set $N = 2000$ and consider different values of $\sigma_p\{0.5, 1, 5\}$:in **(a)** $\sigma_p = 0.5$, in **(b)** $\sigma_p = 1$ and in **(c)** $\sigma_p = 5$,

standard Gaussian density as target pdf,

$$\bar{\pi}(x) = \mathcal{N}(x; 0, 1), \tag{8}$$

and also a Gaussian proposal pdf,

$$q(x) = \mathcal{N}(x; \mu_p, \sigma_p^2), \tag{9}$$

with mean μ_p and variance σ_p^2. We address the problem of estimating the first 5 non-central moments of $\bar{\pi}(x)$, i.e., $f(x) = x^r$ with $r = 1, 2, 3, 4, 5$. In this case, the true values of $I(x^r) = \int_{\mathbb{R}} x^r \bar{\pi}(x)dx$ are $I(x) = 0$, $I(x^2) = 1$, $I(x^3) = 0$, $I(x^4) = 3$, and $I(x^5) = 0$ (we know analytically the ground-truth since the target is Gaussian). We test the different IS schemes and the standard Monte Carlo for approximating the five integrals $I(x^r)$ using N samples. We compute the Mean Square Error (MSE) in the estimation, averaged over the five integrals and over $5 \cdot 10^5$ independent runs. We compare

- the standard Monte Carlo scheme of Table 1, drawing samples directly from $\bar{\pi}(x)$ (denoted as Ideal MC),
- the standard IS technique in Table 2 (denoted as IS),
- the NIS method based on clipping using the value $\eta = \frac{1}{N_C}\sum_{i=1}^{N_C} w_{j_i}$ for the transformed weights (denoted as ClipIS),
- the NIS method based on clipping with $\eta = \min w_{j_i}$ with $i = 1, ..., N_C$ the transformed weights (denoted as ClipIS-min).
- the NIS method based on GIS (denoted as GIS).

We fix the value of $\sigma_p \in \{0.5, 1, 5\}$ and $N \in \{20, 2000\}$ and vary $\mu_p \in [0.1, 3]$, computing the MSE in each scenario. We set $N_C = \frac{N}{5}$. Note that if $\mu_p = 0$ and $\sigma_p = 1$, we have the ideal Monte Carlo case, $q(x) \equiv \bar{\pi}(x)$. As μ_p increases, the proposal becomes more different from $\bar{\pi}$. The results are shown in Figures 1-2 (in each figure, we have a specific value of σ_p and N).

The clipping schemes provide good performance when there is a small discrepancy between the proposal and the target functions. Moreover, they work better with high number of samples N. ClipIS outperforms ClipIS-min in most of the cases. GIS seems the better scheme with smaller N. Note that the standard IS method provides the best results only when $N = 2000$ and $\sigma_p = 5$. We can also observe that, in several cases, the IS strategies provide smaller MSE than the ideal Monte Carlo method. The IS schemes have always a bias (the ideal MC estimator is unbiased) but, in those cases, present a considerable reduction on the variance of the estimators (much smaller than the ideal MC estimator).

5. CONCLUSIONS

In this work, we have introduced a novel NIS method based on the GIS approach. Furthermore, we have compared different advanced IS schemes by means of numerical simulations. We have tested three different NIS schemes, jointly with the classical IS technique and the standard Monte Carlo methods. The NIS technique using the marginal likelihood estimation as clipped value outperforms the corresponding NIS scheme using the minimum value of the weights. The NIS method based on GIS seems to work better than the other IS schemes, when a smaller number of the samples, N, is used and when the discrepancy between proposal and target is high. In some scenario, the IS algorithms are even able to outperform the standard MC scheme due to a sensible reduction of the variance of the final estimators. As future line, we consider the design of extended NIS strategies applying the clipping idea to different group of samples. Another possible future study consists in generalizing the GIS technique for keeping unaltered further moment approximations with respect to the standard IS method.

6. REFERENCES

[1] A. Doucet and X. Wang, "Monte Carlo methods for signal processing," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 152–170, Nov. 2005.

[2] J. Geweke, "Bayesian inference in econometric models using Monte Carlo integration," *Econometrica*, vol. 24, pp. 1317–1399, 1989.

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

[4] L. Martino and J. Míguez, "A novel rejection sampling scheme for posterior probability distributions," *Proc. of the 34th IEEE ICASSP*, April 2009.

[5] J. Míguez, I. P. Mariño, and M. A. Vazquez, "Analysis of a nonlinear importance sampling scheme for Bayesian parameter estimation in state-space models," *Signal Processing*, vol. 142, pp. 281 – 291, 2018.

[6] O. Cappé, A. Guillin, J. M. Marin, and C. P. Robert, "Population Monte Carlo," *Journal of Computational and Graphical Statistics*, vol. 13, no. 4, pp. 907–929, 2004.

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

[8] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Míguez, and P. M. Djuric, "Adaptive importance sampling: The past, the present, and the future," *IEEE Signal Processing Magazine*, vol. 34, pp. 60–79, 2017.

[9] A. Vehtari, A. Gelman, and J. Gabry, "Pareto Smoothed Importance Sampling," *ArXiv e-prints*, 2015.

[10] E. L. Ionides, "Truncated Importance Sampling," *Journal of Computational and Graphical Statistics*, vol. 17, no. 2, pp. 295–311, 2008.

[11] R. M. Neal, "Annealed importance sampling," *Statistics and Computing*, vol. 11, no. 2, pp. 125–139, 2001.

[12] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Generalized multiple importance sampling," *arXiv:1511.03095*, 2015.

[13] L. Martino, V. Elvira, D. Luengo, and J. Corander, "Layered adaptive importance sampling," *Statistics and Computing*, vol. 27, no. 3, pp. 599–623, 2017.

[14] E. Koblents and J. Míguez, "A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models," *Statistics and Computing*, vol. 25, no. 2, pp. 407–425, 2015.

[15] J. Míguez, "On the performance of nonlinear importance samplers and population Monte Carlo schemes," *22nd International Conference on Digital Signal Processing (DSP)*, pp. 1–5, 2017.

[16] A. Owen and Y. Zhou, "Safe and effective importance sampling," *Journal of the American Statistical Association*, vol. 95, no. 449, pp. 135–143, 2000.

[17] E. Veach and L. Guibas, "Optimally combining sampling techniques for Monte Carlo rendering," *In SIGGRAPH 1995 Proceedings*, pp. 419–428, 1995.

[18] L. Martino, V. Elvira, and G. Camps-Valls, "Group importance sampling for Particle Filtering and MCMC," *arXiv:1704.02771*, pp. 1–39, 2017.

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

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