Safe importance sampling based on partial posteriors and neural variational approximations

Fernando Llorente[†], Ernesto Curbelo[†], Luca Martino[⋆], Pablo Olmos[†], David Delgado[†],

[†] Universidad Carlos III de Madrid, Léganes, Madrid, Spain.

[⋆] Universidad Rey Juan Carlos I, Móstoles, Madrid, Spain.

***Abstract*—In this work, we present two novel importance sampling (IS) methods, which can be considered *safe* in the sense that they avoid catastrophic scenarios where the IS estimators could have infinite variance. This is obtained by using a population of proposal densities where each one is wider than the posterior distribution. In fact, we consider partial posterior distributions (i.e., considering a smaller number of data) as proposal densities. Neuronal variational approximations are also discussed. The experimental results show the benefits of the proposed schemes.**

***Index Terms*—Bayesian inference, Markov Chain Monte Carlo (MCMC), multiple importance sampling (MIS), variational approximations, neural networks, partial posteriors.**

I. Introduction

Bayesian inference requires efficient Monte Carlo schemes, such as Markov chain Monte Carlo (MCMC) and importance sampling (IS) for the computation of *a posteriori* quantities [1]–[5]. In this work, we focus on IS schemes. Below, we describe the four main factors, denoted with **(a)**, **(b)**, **(c)** and **(d)**, that are contained in the benchmark and advanced IS methods (currently in the literature), and explain their success.

As in the rest of Monte Carlo methods, in IS, the choice of a suitable proposal density function (pdf) is key to obtain efficient estimators for such quantities [6]–[9]. For this reason, a large body of literature has devoted to the design of good proposal pdfs in IS [10], [11]. In order to increase the robustness, **(a)** the use and the adaptation of a population of proposal pdfs is often employed. In this scenario, **(b)** an optimal construction of the importance weights is available which considers all the mixture of proposal pdfs. This technique is usually referred as multiple IS (MIS) [12]–[16]. Moreover, since the high probability mass represented by the posterior density is often concentrated in a very small space of the domain, many Monte Carlo methods (including IS) often struggle to find the high probability regions. In order to foster the space exploration, **(c)** *tempered*, hence *wider* posterior versions are usually considered. This can be obtained by adding an additional scale parameter (called *temperature*) to the posterior, or considering a smaller number of data (this is called *data tempering*) [17]–[21].

The last main characteristic is related to the adaptation process. In fact, in the last decades, many very efficient IS schemes have been designed by combining the IS idea with the benefits provided by the MCMC algorithms [22]–[27]. For instance, **(d)** the adaptation of the population of proposal pdfs can be obtained by performing MCMC steps. This is the case in the Layered Adaptive IS (LAIS) and its extensions [27], [28]. The MCMC steps are employed in order to locate the mean of the proposal pdfs in high-posterior probability regions of the posterior. LAIS combines the desired exploratory behavior of MCMC methods with the advantages of an IS scheme (easier theoretical validation, easier estimation of the marginal likelihood, etc.).

In this work, we propose novel IS schemes which involve all the four reasons of success **(a)**, **(b)**, **(c)** and **(d)**, described above. Specifically, we describe two different methods. Both are cheaper than LAIS in terms of total number of posterior evaluations (at least, they require half evaluations with respect to LAIS). In the first proposed scheme, partial posteriors (i.e., data-tempered posteriors) are used as a population of proposal densities in MIS approach. We called this framework as *Partial Posteriors IS* (PAPIS). The combination of using multiple proposals and the data-tempering effect (which provides wider partial posteriors than the full-posterior) yields a robust scheme that avoids catastrophic scenarios where the IS estimators could have infinite variance [1], [21]. In this sense, the proposed approach is a *safe* IS scheme. However, PAPIS require the efficient computation of the partial marginal likelihoods of each partial posterior, which is not a straightforward task. In order to overcome this issue, we introduce the *neural* PAPIS, where, we fit neural variational approximations to the partial posteriors and use them as proposal pdfs. Variational approximations can be crude approximations to the true distributions, but they are cheap and fast to compute, and also fulfill the requirements of being easy to sample and evaluate. The experiments show the benefits of the proposed methods.

Other related works. The combination of IS and variational approximations has already being considered in the literature [29]–[32]. Broadly speaking, one can either use variational methods to improve the construction of IS algorithms [31], [33], [34]; or employ IS ideas to improve the computation of variational bounds, e.g., [32]. Our work makes a contribution to the former group of approaches. In this context, in [34], variational algorithms are used for building a good initial set of proposal pdfs that are then evolved using an AIS algorithm. Furthermore, the authors in [33] propose an adaptive IS (AIS) algorithm where variational autoencoders are employed for generating the samples. In [31], we can find possibly the closest approach to our proposed methotologies. There, the authors propose to obtain a mean-field approximation of the posterior pdf for use in IS. In this sense, we improve on their work by proposing to learn instead a robust population of proposals, each one approximating a different partial posterior pdf (rather than the complete posterior). PAPIS presents also some connections with the Recycling LAIS (RLAIS) presented in [28]. If, instead of MCMC samplers, other IS schemes are employed to approximate the partial posteriors, PAPIS can be interpreted as a special case of Deep IS, described in [35].

II. Problem statement and background

A. Problem statement

Let $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ be the variable of interest and consider a matrix of observed measurements, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_L] \subset \mathbb{R}^{d_y \times L}$. In the Bayesian framework, one complete model is formed by a likelihood function $\ell(\mathbf{Y}|\mathbf{x})$ and an a-priori probability density function (pdf) $g(\mathbf{x})$. All the statistical information is summarized by the posterior pdf,

$$\bar{\pi}(\mathbf{x}|\mathbf{Y}) = \frac{\ell(\mathbf{Y}|\mathbf{x})g(\mathbf{x})}{\int_{\mathcal{X}} \ell(\mathbf{Y}|\mathbf{x}')g(\mathbf{x}')d\mathbf{x}'}. \quad (1)$$

The denominator plays the role of a normalizing constant, $Z = p(\mathbf{Y}) = \int_{\mathcal{X}} \ell(\mathbf{Y}|\mathbf{x})g(\mathbf{x})d\mathbf{x}$. The quantity Z is called marginal likelihood (a.k.a., Bayesian evidence) and is important for model selection [21]. Generally, we are interested in computing integrals with respect to (w.r.t.) the posterior pdf $I = \int_{\mathcal{X}} f(\mathbf{x})\bar{\pi}(\mathbf{x}|\mathbf{Y})d\mathbf{x} = \frac{1}{Z}\int_{\mathcal{X}} f(\mathbf{x})\pi(\mathbf{x}|\mathbf{Y})d\mathbf{x}$, where $\bar{\pi}(\mathbf{x}|\mathbf{Y}) \propto \pi(\mathbf{x}|\mathbf{Y}) = \ell(\mathbf{Y}|\mathbf{x})g(\mathbf{x})$ and $f(\mathbf{x})$ is some integrable function.

B. Multiple Importance sampling

In IS, the use of a population of proposal pdfs (instead of only one proposal) often increases the robustness of the results. In the so-called multiple IS (MIS) [12], [13], we have a population of M normalized proposal pdfs, $q_m(\mathbf{x})$ for $m = 1, \ldots, M$, and $\{\mathbf{x}_i^{(m)}\}_{i=1}^N$ denotes the sets of N samples drawn from each of them. By assigning the following full-deterministic mixture weights [12]

$$w_i^{(m)} = \frac{\pi(\mathbf{x}_i^{(m)}|\mathbf{Y})}{\frac{1}{M}\sum_{s=1}^M q_s(\mathbf{x}_i^{(m)})}, \quad (2)$$

for all $i = 1, \ldots, N$ and $m = 1, \ldots, M$, we can compute the self-normalized estimator of I, with $\widehat{I} = \frac{\sum_{m=1}^M \sum_{i=1}^N w_i^{(m)} f(\mathbf{x}_i^{(m)})}{\sum_{s=1}^M \sum_{j=1}^N w_j^{(s)}}$. Furthermore, the marginal likelihood can be approximated by $\widehat{Z} = \frac{1}{NM}\sum_{m=1}^M \sum_{i=1}^N w_i^{(m)}$. There exist different ways of choosing and adapting the population of proposal pdfs [10], [11].

III. Partial posterior densities

The posterior distribution is often highly concentrated, making difficult to discover the regions of high-probability in the state-space, and hence jeopardizing the construction of good proposal pdfs. The underlying idea of using a tempering effect is to consider *wider posteriors* in the first iterations of the applied Monte Carlo algorithm. This effect can be obtained by adding an additional scale parameter to the target density or considering a smaller number of data [17]–[20]. The MIS scheme and the tempering effect are included in sophisticated and benchmark techniques, so far in the literature. For this reason, in this work, we suggest a new methodology for efficiently combining both approaches. For this purpose, we need to define the partial posterior densities, i.e., posteriors considering only subsets of data [36]–[38]. Let $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_b\}$ denote a partition of the data in B (possibly non-overlapping) subsets, i.e.,

$$\mathbf{Y} = \bigcup_{b=1}^{B} \mathbf{Y}_b, \quad B \leq L, \quad (3)$$

where

$$\mathbf{Y}_b \in \mathbb{R}^{d_y \times L_b}, \quad L_b \leq L, \quad \sum_{b=1}^{B} L_b = L. \quad (4)$$

For each subset $\mathbf{Y}_b$, we consider its posterior pdf (i.e. a *partial posterior*)

$$\bar{\pi}_b(\mathbf{x}) = p(\mathbf{x}|\mathbf{Y}_b) = \frac{1}{Z_b}\ell(\mathbf{Y}_b|\mathbf{x})g_b(\mathbf{x}), \quad b = 1, \ldots, B, \quad (5)$$

where $\ell(\mathbf{Y}_b|\mathbf{x})$ denotes the likelihood of the subset $\mathbf{Y}_b$,[1] $Z_b = \int_{\mathcal{X}} \ell(\mathbf{Y}_b|\mathbf{x})g_b(\mathbf{x})d\mathbf{x}$, and $g_b(\mathbf{x})$ denotes a partial prior pdf. When disjoint sets are considered $\mathbf{Y}_b \cap \mathbf{Y}_{b'} = \emptyset$ and each subset $\mathbf{Y}_b$ has the same number of observations (i.e. $L_b = \frac{L}{B}$ for all b), a typical choice is $g_b(\mathbf{x}) \propto g(\mathbf{x})^{\frac{1}{B}}$ [36]–[40]. The partial posteriors $\bar{\pi}_b(\mathbf{x})$ is a data-tempered (hence wider) version of the full posterior $\bar{\pi}(\mathbf{x})$ [17], [18].

IV. Partial posterior importance sampling (PAPIS)

In this section, we describe the scheme *partial posterior importance sampling* (PAPIS). The main idea is to use the population of B partial posteriors as proposal densities to perform MIS on the full posterior. This is particularly appealing since the partial posteriors are in some sense wider versions of the full posterior, and hence they are robust choices of proposal pdfs (avoiding the catastrophic scenario of infinite variance). Since in general we are not able to draw samples from the partial posteriors, in PAPIS, we suggest running B different MCMC chains on the partial posteriors $\bar{\pi}_b$, and then use the obtained chains as samples in a MIS scheme.[2] The algorithm is outlined in Table I.

Important considerations. PAPIS uses the set of partial posteriors, $\bar{\pi}_b(\mathbf{x}|\mathbf{Y}_b)$ for $b = 1, \ldots, B$, as proposal pdfs. This ensures to avoid catastrophic IS scenario with estimators with infinite variance. Another advantage of PAPIS with respect to the similar LAIS algorithm (which employs also MCMC samples) is that the overall number of target evaluations is $E = NT$ whereas in LAIS is at least double, $E = 2NT$

[1] Here, we assume one can compute the likelihood of subsets of data.

[2] Clearly, this is an approximate strategy but, if the chains are long enough and after a burn-in period, the samples can be considered distributed as the corresponding partial posterior.

TABLE I
PARTIAL POSTERIOR IMPORTANCE SAMPLING (PAPIS).

- **Sampling:** Generate BT samples, $\mathbf{x}_{b,1}, \dots, \mathbf{x}_{b,T}$ for $b = 1, \dots, B$, using B (independent or interacting) MCMC chains, each one addressing a different invariant density $\bar{\pi}_b(\mathbf{x}|\mathbf{y}_b)$.
- **Normalize:**
  1) Obtain $\widehat{Z}_b$, an estimation of the partial marginal likelihood Z_b of the partial posterior $\bar{\pi}_b$, via Reverse Importance Sampling (RIS) or using any other suitable method for estimating marginal likelihoods via MCMC samples [21], [41], for all $b = 1, \dots, B$.
  2) Set

$$\widehat{\pi}_b(\mathbf{x}|\mathbf{Y}_b) = \frac{1}{\widehat{Z}_b}\pi_b(\mathbf{x}|\mathbf{Y}_b), \tag{6}$$

for $b = 1, \dots, B$.

- **Weighting:** Assign to $\mathbf{x}_{b,t}$ the weights

$$w_{b,t} = \frac{\pi(\mathbf{x}_{b,t}|\mathbf{Y})}{\frac{1}{B}\sum_{b=1}^{B}\widehat{\pi}_b(\mathbf{x}_{b,t}|\mathbf{Y}_b)}. \tag{7}$$

- **Output:** Return all the pairs $\{\mathbf{x}_{b,t}, w_{b,t}\}$.

[27], [28]. However, the use of partial posteriors as proposals entails two difficulties in practice: (i) we cannot sample them directly, (ii) we cannot evaluate them in closed form, due to the unknown normalizing constants $Z_b = \int_{\mathcal{X}} \ell(\mathbf{Y}_b|\mathbf{x})g_b(\mathbf{x})d\mathbf{x}$, $b = 1, \dots, B$. Therefore, we consider an approximated sampling using MCMC algorithms and an estimation $\widehat{Z}_b$ can be obtained by any suitable procedure (e.g., *reverse importance sampling*) using the same states of the generated chain $\mathbf{x}_{b,1}, \dots, \mathbf{x}_{b,T}$ [21], [41]. Obtaining good estimations of all $Z_b' s$ is the trickier part of the PAPIS algorithm: this step can jeopardize the performance of the resulting PAPIS estimators. In the next section, we try to overcome these issues by building neural variational approximations to each partial posterior. The variational approximations are cheap and powerful approximations, that allow us to sample and evaluate them in direct form.

V. NEURAL PAPIS

In this section, we introduce our proposed approach, consisting in two separated stages. At stage 1, we divide data $\mathbf{Y}$ into B subsets, $\mathbf{Y}_1, \dots, \mathbf{Y}_b$, and run parallel variational inference algorithms in order to learn the optimal parameters ϕ_b^* for the variational approximations to each partial posterior

$$q_{\phi_b^*}(\mathbf{x}|\mathbf{Y}_b) \approx \bar{\pi}(\mathbf{x}|\mathbf{Y}_b), \quad b = 1, \dots, B.$$

At stage 2, we apply MIS to the population $\{q_{\phi_b^*}(\mathbf{x}|\mathbf{Y}_b)\}_{b=1}^{B}$, that is, we sample N times from each $q_{\phi_b^*}^{(b)}(\mathbf{x})$ and weigh the final NB samples according to Eq. (2). This scheme is summarized in Table II. Note that the construction and sampling of the $q_{\phi_b^*}(\mathbf{x}|\mathbf{Y}_b)$ can be done in parallel, and only in the weighting step one needs to communicate all the proposals $q_{\phi_b^*}(\mathbf{x}|\mathbf{Y}_b)$ and samples $\{\mathbf{x}_n^{(b)}\}_{n=1}^{N}$ to a central node.

Compared to PAPIS, with this neuro-variational solution (denoted as Neural-PAPIS), we can draw independent samples from the partial posterior approximations $q_{\phi_b^*}(\mathbf{x}|\mathbf{Y}_b)$ as well as evaluate them exactly. On the other hand, a potential difficulty of the proposed approach is that the variational proposals could represent only "crude" approximations to the partial posteriors (especially when the chosen variational family is very simple), and hence produce final estimators with low efficiency [31].

A. Building the partial posterior approximation

In this section, we aim to learn an approximation of the partial posterior $\bar{\pi}(\mathbf{x}|\mathbf{Y}_b)$, given our knowledge of the likelihood function and the prior. Assume a family of pdfs, $q_\phi(\mathbf{x})$, parameterized by parameter vector ϕ. We consider building a variational approximation to each partial posterior $\bar{\pi}(\mathbf{x}|\mathbf{Y}_b)$,

$$q_{\phi_b^*}(\mathbf{x}|\mathbf{Y}_b) \approx \bar{\pi}(\mathbf{x}|\mathbf{Y}_b) \propto \ell(\mathbf{Y}_b|\mathbf{x})g_b(\mathbf{x}), \quad b = 1, \dots, B, \tag{8}$$

where the optimal parameters are determined by maximizing the so-called *Evidence-Lower-Bound* (ELBO) [42]

$$\phi_b^* = \arg\max_{\phi} \mathcal{L}(\phi, \mathbf{Y}_b), \tag{9}$$

where

$$\mathcal{L}(\phi, \mathbf{Y}_b) = \mathbb{E}_{q_\phi(\mathbf{x}|\mathbf{Y}_b)}\left[\log \ell(\mathbf{Y}_b|\mathbf{x})\right] - KL\left(q_\phi(\mathbf{x}|\mathbf{Y}_b)||g_b(\mathbf{x})\right). \tag{10}$$

For instance, we can take the variational family to be Gaussian (which is the one that we tested in the experiments),

$$q_\phi(\mathbf{x}|\mathbf{Y}_b) = \mathcal{N}\left[\mathbf{x}|\boldsymbol{\mu}_\phi(\mathbf{Y}_b), \boldsymbol{\Sigma}_\phi(\mathbf{Y}_b)\right], \tag{11}$$

where

$$\boldsymbol{\mu}_\phi(\mathbf{Y}_b) : \mathbb{R}^{d_y \times L_b} \to \mathbb{R}^{d_x} \tag{12}$$

$$\boldsymbol{\Sigma}_\phi(\mathbf{Y}_b) : \mathbb{R}^{d_y \times L_b} \to \mathbb{R}^{d_x \times d_x}, \tag{13}$$

are the outputs of a neural network (NN) with weights and biases given by ϕ. For simplicity, we assume $\boldsymbol{\Sigma}_\phi$ is a diagonal covariance matrix, and that we actually learn the logarithm of its diagonal.

VI. NUMERICAL SIMULATIONS

For testing our proposed approach, we address the problem of positioning a target on a two dimensional space from the measurements this produces in a wireless net of sensors.

Let us consider the random vector $\mathbf{x} \in \mathbb{R}^2$ as the target position. In the net, we have L sensors whose positions are known and denoted as $\mathbf{s}_1, \dots, \mathbf{s}_L$. Given that we observe M measurements from every sensor, these follow the distribution

$$y_{m,l} \sim \mathcal{N}\left(-A\log\left(\parallel \mathbf{x} - \mathbf{s}_l \parallel_2^2\right), \sigma^2\right), \quad m = 1, \dots, M \tag{14}$$

where the parameter A is a constant related to how fast the signal degrades with the distance and it can depend on many conditions such as environmental or manufacturing ones. We assume that the variance σ^2 is known and equal for all sensors.

TABLE II
Neural PAPIS

- **Initialization:** Partition $\{\mathbf{Y}_1, \dots, \mathbf{Y}_b\}$ and prior pdfs $g_b(\mathbf{x})$ $(b = 1, \dots, B)$.
- **For** $b = 1, \dots, B$**:** (in parallel)
    1) *Proposals contruction.* Obtain $q_{\phi_b^*}(\mathbf{x}|\mathbf{Y}_b)$ by maximizing $\mathcal{L}(\phi, \mathbf{Y}_b)$ in Eq. (10).
    2) *Sampling.* Draw $\{\mathbf{x}_n^{(b)}\}_{n=1}^N \sim q_{\phi_b^*}(\mathbf{x}|\mathbf{Y}_b)$.
- **Weighting:** For all $n = 1, \dots, N$ and $b = 1, \dots, B$, compute

$$w_n^{(b)} = \frac{\pi(\mathbf{x}_n^{(b)}|\mathbf{Y})}{\frac{1}{B}\sum_{s=1}^B q_{\phi_s^*}(\mathbf{x}_n^{(b)}|\mathbf{Y}_s)}, \quad \bar{w}_n^{(b)} = \frac{w_n^{(b)}}{\sum_{s=1}^B \sum_{j=1}^N w_j^{(s)}}.$$

- **Outputs:** The NB weighted samples $\{\mathbf{x}_n^{(b)}, \bar{w}_n^{(b)}\}$, $n = 1, \dots, N$, $b = 1, \dots, B$.

We set $A = 10$ and $\sigma^2 = 5$. We consider only three sensors with positions $\mathbf{s}_1 = [0.5, 1]^T$, $\mathbf{s}_2 = [3.5, 1]^\top$, $\mathbf{s}_3 = [2, 3]^\top$. We took the target position at $\mathbf{x}_{\text{true}} = [-1, 2]^\top$ for generating the observations (i.e. the ground-truth). We simulate 15 observations per sensor from the model in Eq. (14), i.e., $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_{15}\}$, where $\mathbf{y}_m = [y_{m,1}, y_{m,2}, y_{m,3}] \in \mathbb{R}^3$.
We use a standard Gaussian as prior distribution over $\mathbf{x}$. In this toy example, we can compute the (diagonal of the) moments numerically using a thin grid over the parametric space, to get $\boldsymbol{\mu} = [-1.01, 1.84]^\top$, and $\boldsymbol{\mu}^{(5)} = [-1.11, 22.31]^\top$.
The previous quantities were estimated using the proposed approach. In order to apply Neural-PAPIS, we run B parallel variational inference algorithms. Specifically, for each batch $\mathbf{Y}_b$, we consider an architecture with one hidden layer of 50 neurons, and used the Adam optimizer for training during 1000 epochs [43]. Several number of batches B were tested: $B \in \{1, 3, 5, 15\}$, and in every case the total number of samples drawn in the IS phase was fixed at 300 and 600 (i.e. $N = \frac{300}{B}$ or $N = \frac{600}{B}$ samples per proposal, respectively).
For obtaining the batches, we always partition the data in B batches with even number of observations. We calculate the mean square error (MSE) in the estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^{(5)}$ over 500 simulations. The results were compared with PAPIS using 5000 and 10000 samples, namely, a much larger number of evaluations than Neural-PAPIS. The results of the comparison are presented in Figure 1. In the y-axis we used the log scale for a better visualization.
In Figure 1-above, we see how both approaches are valid to solve the location problem, the Neural-PAPIS obtaining a very low MSE in the estimation of the posterior mean $\boldsymbol{\mu}$ when $B = 3$, close to PAPIS with $E = 10000$ total evaluations. We stress that we deliberately chose PAPIS with much more number of evaluations ($E \in \{5000, 10000\}$) than Neural-PAPIS ($E \in \{300, 600\}$) in order to show the benefits of the proposed approach. In fact, regarding the estimation of $\boldsymbol{\mu}^{(5)}$, in Figure 1-below, Neural-PAPIS with only 600 evaluations significantly outperforms PAPIS with considerable more evaluations. Furthermore, as expected, both Figures show that the use of partial posteriors improve the results compared to using a full posterior, which corresponds to $B = 1$. We also recall that Neural-PAPIS with $B = 1$ correspond to using the variational approximation to the full posterior as a single proposal (as in [31]). We see that better results are obtained if we use a population of *wider* proposals instead, i.e., $B > 1$. Interestingly, the curves feature a minimum MSE around $B = 3$, suggesting that there possibly exists an optimal number of batches, B^*, which is worth investigating in the future. We have also compared the results obtained by PAPIS schemes, with a standard Population Monte Carlo (PMC) [44]. The best result of PMC with the bigger number of evaluations $E = 10^5$ has been an MSE of order 10^{-2} for the first moment and an MSE of order 10^2 for the fifth moment. Hence, all the PAPIS schemes outperform the PMC.

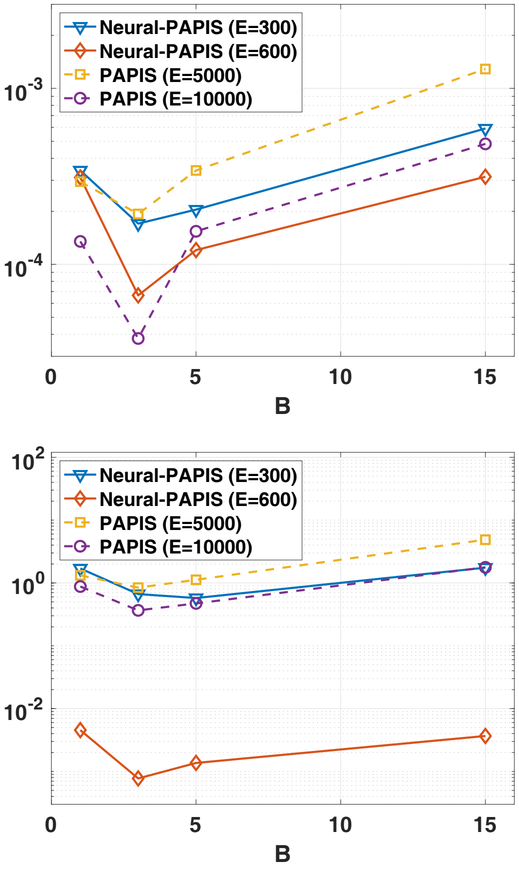

Fig. 1. MSE of Neural-PAPIS and PAPIS when estimating the first moment (above) and the fifth moment (below). The number of evaluations is denoted by E.

VII. Conclusions

In this work, we presented a framework for robust and safe IS based on the use of multiple partial posteriors (i.e., posteriors of subsets of data) within a MIS scheme, which we called PAPIS. This framework combines two powerful ideas that feature in many IS algorithms, namely, (data) tempering and proposals population. However, this simple framework requires exact sampling and evaluation of partial posteriors. We bypass this problem by building variational approximations to each partial posterior and then use them as proposals. The (neural) variational approximations are cheap and fast to compute (parallelization is also possible), and give results that outperform the naive application of PAPIS, as we showed in the experiment.

Acknowledgments

FLL acknowledges support from Spanish government via grant FPU19/00815, LM and FLL from Agencia Estatal de Investigación AEI (project SP-GRAPH, ref. num. PID2019-105032GB-I00), and PO from Spanish government (AEI/MCI) under grants RTI2018-099655-B-100, PID2021-123182013-100 and TED2021-13182313-100, Comunidad de Madrid under grants IND2018/TIC-9649 and Y2018/TCS-4705, and the European Union (FEDER) and the European Research Council (ERC) through the European Union's Horizon 2020 research and innovation program under Grant 714161.

References

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

[2] D. Luengo, L. Martino, M. Bugallo, V. Elvira, and S. Särkkä, "A survey of Monte Carlo methods for parameter estimation," *EURASIP J. Adv. Signal Process.*, vol. 25, pp. 1–62, 2020.

[3] A. Owen, *Monte Carlo theory, methods and examples*, http://statweb.stanford.edu/∼owen/mc/, 2013.

[4] W. J. Fitzgerald, "Markov chain Monte Carlo methods with applications to signal processing," *Signal Processing*, vol. 81, no. 1, pp. 3–18, January 2001.

[5] L. Martino, D. Luengo, and J. Miguez, *Independent Random Sampling Methods*, Springer, 2018.

[6] G.R. Douc, J.M. Marin, and C. Robert, "Convergence of adaptive mixtures of importance sampling schemes," *Annals of Statistics*, vol. 35, pp. 420–448, 2007.

[7] G.R. Douc, J.M. Marin, and C. Robert, "Minimum variance importance sampling via population Monte Carlo," *ESAIM: Probability and Statistics*, vol. 11, pp. 427–447, 2007.

[8] O. D. Akyildiz and J. Miguez, "Convergence rates for optimised adaptive importance samplers," *Statistics and Computing*, vol. 31, no. 2, pp. 1–17, 2021.

[9] O. D. Akyildiz, "Global convergence of optimized adaptive importance samplers," *arXiv:2201.00409*, pp. 1–16, 2022.

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

[11] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric, "Adaptive importance sampling: The past, the present, and the future," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 60–79, 2017.

[12] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Generalized Multiple Importance Sampling," *Statistical Science*, vol. 34, no. 1, 2019.

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

[14] J. M. Marin, P. Pudlo, and M. Sedki, "Consistency of the adaptive multiple importance sampling," *arXiv:1211.2548*, 2012.

[15] J. M. Cornuet, J. M. Marin, A. Mira, and C. P. Robert, "Adaptive multiple importance sampling," *Scandinavian Journal of Statistics*, vol. 39, no. 4, pp. 798–812, December 2012.

[16] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Heretical multiple importance sampling," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1474–1478, 2016.

[17] N. Chopin, "A sequential particle filter for static models," *Biometrika*, vol. 89, pp. 539–552, 2002.

[18] L. Martino, F. Llorente, E. Curbelo, J. Lopez-Santiago, and J. Miguez, "Automatic tempered posterior distributions for Bayesian inversion problems," *Mathematics*, vol. 9, no. 7, 2021.

[19] S. K. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, May 1983.

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

[21] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago, "Marginal likelihood computation for model selection and hypothesis testing: an extensive review," *(to appear) SIAM review - (extended version arXiv:2005.08334)*, pp. 1–91, 2022.

[22] Z. I. Botev, P. L' Ecuyer, and B. Tuffin, "Markov chain importance sampling with applications to rare event probability estimation," *Statistics and Computing*, vol. 23, pp. 271–285, 2013.

[23] X. Yuan, Z. Lu, and C. Z. Yue, "A novel adaptive importance sampling algorithm based on Markov chain and low-discrepancy sequence," *Aerospace Science and Technology*, vol. 29, pp. 253–261, 2013.

[24] E. F. Mendes, M. Scharth, and R. Kohn, "Markov Interacting Importance Samplers," *arXiv:1502.07039*, 2015.

[25] I. Schuster and I. Klebanov, "Markov Chain Importance Sampling?a highly efficient estimator for MCMC," *Journal of Computational and Graphical Statistics*, pp. 1–9, 2020.

[26] D. Rudolf and B. Sprungk, "On a Metropolis-Hastings importance sampling estimator," *Electronic Journal of Statistics*, vol. 14, no. 1, pp. 857–889, 2020.

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

[28] F. Llorente, E. Curbelo, L. Martino, V. Elvira, and D. Delgado, "MCMC-driven importance samplers," *arXiv:2105.02579*, pp. 1–28, 2021.

[29] J. Hernandez, J. Capdevila, and J. Cerquides, "Variational importance sampling: initial i findings," in *22nd International Conference of the Catalan Association for Artificial Intelligence*. IOS Press, 2019, vol. 319, p. 95.

[30] G. Jerfel, S. Wang, C. Wong-Fannjiang, K. A. Heller, Y. Ma, and M. Jordan, "Variational refinement for importance sampling using the forward kullback-leibler divergence," in *Uncertainty in Artificial Intelligence*, 2021, pp. 1819–1829.

[31] X. Su and Y. Chen, "Variational approximation for importance sampling," *Computational Statistics*, vol. 36, no. 3, pp. 1901–1930, 2021.

[32] O. Kviman, H. Melin, H. Koptagel, V. Elvira, and J. Lagergren, "Multiple importance sampling elbo and deep ensembles of variational approximations," *arXiv preprint arXiv:2202.10951*, 2022.

[33] H. Wang, M. F. Bugallo, and . M. Djurić, "Adaptive importance sampling supported by a variational auto-encoder," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2019, pp. 619–623.

[34] M. Dowling, J. Nassar, P. M. Djurić, and M. F. Bugallo, "Improved adaptive importance sampling based on variational inference," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1632–1636.

[35] F. Llorente, L. Martino, D. Delgado, and G. Camps-Valls, "Deep importance sampling based on regression for model inversion and emulation," *Digital Signal Processing*, vol. 116, pp. 103104, 2021.

[36] S. L. Scott, "Comparing consensus Monte Carlo strategies for distributed Bayesian computation," *Braz. J. Probab. Stat.*, vol. 31, no. 4, pp. 668–685, 2017.

[37] W. Neiswanger, C. Wang, and E. Xing, "Asymptotically exact, embarrassingly parallel MCMC," *arXiv:1311.4780*, 2013.

[38] D. Luengo, L. Martino, V. Elvira, and M. Bugallo, "Efficient linear fusion of partial estimators," *Digital Signal Processing*, vol. 78, pp. 265–283, 2018.

[39] S. L. Scott et al., "Bayes and big data: The consensus Monte Carlo algorithm," *International Journal of Management Science and Engineering Management*, vol. 11, no. 2, pp. 78–88, 2016.

[40] L. Martino and V. Elvira, "Compressed Monte Carlo for distributed Bayesian inference," *viXra:1811.0505*, pp. 1–14, 2018.

[41] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago, "On the computation of marginal likelihood via MCMC for model selection and hypothesis testing," in *28th European Signal Processing Conference (EUSIPCO)*, 2020, pp. 2373–2377.

[42] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[44] 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.