Group Metropolis Sampling

Luca Martino⋆, Víctor Elvira†, Gustau Camps-Valls⋆
⋆ Image Processing Laboratory, Universitat de València (Spain).
† IMT Lille Douai CRISTAL (UMR 9189), Villeneuve d'Ascq (France).

***Abstract*—Monte Carlo (MC) methods are widely used for Bayesian inference and optimization in statistics, signal processing and machine learning. Two well-known class of MC methods are the Importance Sampling (IS) techniques and the Markov Chain Monte Carlo (MCMC) algorithms. In this work, we introduce the Group Importance Sampling (GIS) framework where different sets of weighted samples are properly summarized with one summary particle and one summary weight. GIS facilitates the design of novel efficient MC techniques. For instance, we present the Group Metropolis Sampling (GMS) algorithm which produces a Markov chain of sets of weighted samples. GMS in general outperforms other multiple try schemes as shown by means of numerical simulations.**

Keywords: Bayesian inference, Importance Sampling, Markov Chain Monte Carlo (MCMC), Gaussian Processes (GP)

I. Introduction

Many applications in statistical signal processing, machine learning and statistics, require the computation of a-posteriori estimators induced by complicated posterior probability distributions [1], [2]. The approximation of these estimators needs often the use of Monte Carlo methods [3]–[5]. The most popular MC approaches are the Importance Sampling (IS) methods and the Markov chain Monte Carlo (MCMC) algorithms [1], [4]. IS schemes produce a random discrete approximation of the posterior distribution by a population of weighted samples [4], [6], [7]. MCMC techniques generate a Markov chain (i.e., a sequence of correlated samples) with a pre-established target probability density function (pdf) as invariant density [5], [8].

In this work, we introduce the Group Importance Sampling (GIS) framework where different sets of weighted samples can be properly summarized with one summary particle and one summary weight. This idea has been indirectly and implicitly employed in different Monte Carlo schemes: parallel particle filters [9], [10], particle island and related methods [11]–[13], tracking and model selection algorithms [14], nested sequential Monte Carlo schemes [15], [16] are some examples.

Furthermore, we also show that the GIS theory facilitates the design of novel efficient Monte Carlo techniques. As an example, we present the Group Metropolis Sampling (GMS) algorithm that generates a Markov chain of sets of weighted samples. All these resulting sets of samples are jointly employed obtaining a unique particle approximation of the target distribution. On the one hand, GMS can be considered as an MCMC method since it produces a Markov chain of sets of samples. On the other hand, the GMS can be also considered as an iterated importance sampler where different estimators are finally combined in order to build a unique IS estimator. This combination is obtained *dynamically* through random repetitions given by MCMC-type acceptance tests. GMS is closely related to Multiple Try Metropolis (MTM) techniques and Particle Metropolis-Hastings (PMH) algorithms [17]–[22], as we discuss below. The GMS algorithm can be also seen as an extension of the method in [23], for recycling auxiliary samples in a MCMC method.

The paper has the following structure. Section II recalls some background material. The GIS theory is introduced in Section III. In Section IV, we present the GMS algorithm. Section V provides some numerical results and in Section VI we discuss some conclusions.

II. Problem statement and background

In many applications the goal is to infer a variable of interest, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$, given a set of related observations or measurements, $\mathbf{y} \in \mathbb{R}^{d_y}$. The statistical information is summarized in the posterior probability density function (pdf) given by

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

where $\ell(\mathbf{y}|\mathbf{x})$ is the likelihood function, $g(\mathbf{x})$ is the prior pdf and $Z(\mathbf{y})$ is the marginal likelihood (a.k.a., Bayesian evidence). In general Z is unknown and often impossible to compute, so we only assume to be able to evaluate the unnormalized target function,[1]

$$\pi(\mathbf{x}) = \ell(\mathbf{y}|\mathbf{x})g(\mathbf{x}). \qquad (2)$$

The computation of integrals involving $\bar{\pi}(\mathbf{x}) = \frac{1}{Z}\pi(\mathbf{x})$ are often intractable. We consider the problem of approximating via Monte Carlo a complicated integral involving the target $\bar{\pi}(\mathbf{x})$ and an integrable function $h(\mathbf{x})$ with respect to $\bar{\pi}$, i.e.,

$$I = E_{\bar{\pi}}[h(\mathbf{X})] = \int_{\mathcal{X}} h(\mathbf{x})\bar{\pi}(\mathbf{x})d\mathbf{x}, \qquad \mathbf{X} \sim \bar{\pi}(\mathbf{x}). \qquad (3)$$

A. Importance Sampling

Let us consider a proposal density $q(\mathbf{x})$,[2] The importance sampling (IS) method consists of drawing N samples, $\mathbf{x}_1, \ldots, \mathbf{x}_N$, from $q(\mathbf{x})$ (also called particles in this work),

[1] We drop the dependence on $\mathbf{y}$ of the marginal likelihood, i.e. $Z \equiv Z(\mathbf{y})$ for the ease of notation.

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

ISBN 978-0-9928626-7-1 © EURASIP 2017

and then assign to each sample the following unnormalized weights

$$w_n = w(\mathbf{x}_n) = \frac{\pi(\mathbf{x}_n)}{q(\mathbf{x}_n)}, \quad n = 1, \dots, N. \tag{4}$$

If Z is known, a possible (unbiased) IS estimator [4], [5] is given by $\widehat{I}_N = \frac{1}{ZN}\sum_{n=1}^{N} w_n h(\mathbf{x}_n)$. If Z is unknown, defining the normalized weights, $\bar{w}_n = \frac{w_n}{\sum_{i=1}^{N} w_i}$ with $n = 1, \dots, N$, an alternative self-normalized (biased) IS estimator is

$$\overline{I}_N = \sum_{n=1}^{N} \bar{w}_n h(\mathbf{x}_n). \tag{5}$$

Both $\widehat{I}_N$ and $\overline{I}_N$ are consistent estimators of I in Eq. (3) [4], [5]. Moreover, an unbiased estimator of marginal likelihood, $Z = \int_{\mathcal{X}} \pi(\mathbf{x})d\mathbf{x}$, is given by $\widehat{Z} = \frac{1}{N}\sum_{i=1}^{N} w_i$. More generally, the pairs $\{\mathbf{x}_i, w_i\}_{i=1}^{N}$ represents a particle approximation of the posterior distribution,

$$\widehat{\pi}(\mathbf{x}|\mathbf{x}_{1:N}) = \sum_{n=1}^{N} \frac{w_n}{N\widehat{Z}} \delta(\mathbf{x} - \mathbf{x}_n) = \sum_{n=1}^{N} \bar{w}_n \delta(\mathbf{x} - \mathbf{x}_n). \tag{6}$$

B. Concept of proper weighting

The standard IS weights in Eq. (4) are broadly used in the literature. However the definition of *proper weighted sample* can be extended as suggested in [4, Section 14.2], [5, Section 2.5.4], and [24]. More specifically, given a set of samples, they are proper weighted with respect to the target π if, for any square integrable function h,

$$E_q[w(\mathbf{x}_n)h(\mathbf{x}_n)] = cE_{\bar{\pi}}[h(\mathbf{x}_n)], \quad \forall n = \{1, \dots, N\}, \tag{7}$$

where c is a constant value, also independent from the index n, and the expectation of the left hand side is performed, in general, w.r.t. the joint pdf of $w(\mathbf{x})$ and $\mathbf{x}$, i.e., $q(w, \mathbf{x})$. Namely, the weight $w(\mathbf{x})$, (for a given value of $\mathbf{x}$), could even be considered a random variable. Thus, in order to obtain consistent estimators, one can design any joint $q(w, \mathbf{x})$ as long as the restriction of Eq. (7) is fulfilled.

C. The Independent Metropolis-Hastings (IMH) algorithm

The Metropolis-Hastings (MH) method [4], [5], [25] is one of the most popular MCMC algorithm. It generates a Markov chain $\{\mathbf{x}_t\}_{t=1}^{\infty}$ with $\bar{\pi}(\mathbf{x})$ as stationary density. Considering a proposal pdf $q(\mathbf{x})$ independent from the previous state $\mathbf{x}_{t-1}$, the independent MH method is given in Table I.

Observe that $\alpha(\mathbf{x}_{t-1}, \mathbf{v}') = \min\left[1, \frac{w(\mathbf{v}')}{w(\mathbf{x}_{t-1})}\right]$ in Eq. (8) involves the ratio between the importance weight of the proposed samples $\mathbf{v}'$ and the importance weight of the previous state $\mathbf{x}_{t-1}$. Note that at each iteration only one new sample $\mathbf{v}'$ is generated to be compared with the previous state $\mathbf{x}_{t-1}$ by the acceptance probability $\alpha(\mathbf{x}_{t-1}, \mathbf{v}')$.

Table I
The IMH algorithm

```
Initialization: Choose an initial state $\mathbf{x}_0$.
For $t = 1, \dots, T$ :
    1) Draw a sample $\mathbf{v}' \sim q(\mathbf{x})$.
    2) Accept the new state, $\mathbf{x}_t = \mathbf{v}'$, with probability
           $\alpha(\mathbf{x}_{t-1}, \mathbf{v}') = \min\left[1, \frac{\pi(\mathbf{v}')q(\mathbf{x}_{t-1})}{\pi(\mathbf{x}_{t-1})q(\mathbf{v}')}\right]$   (8)
           $= \min\left[1, \frac{w(\mathbf{v}')}{w(\mathbf{x}_{t-1})}\right]$,   (9)
       where $w(\mathbf{x}) = \frac{\pi(\mathbf{x})}{q(\mathbf{x})}$ (importance weight). Otherwise,
       set $\mathbf{x}_t = \mathbf{x}_{t-1}$.
Return: $\{\mathbf{x}_t\}_{t=1}^{T}$.
```

III. Group Importance Sampling (GIS)

Let us consider the M disjoint sets of weighted samples (a.k.a., particles)

$$\mathcal{S}_m = \{\mathbf{x}_{m,n}, w_{m,n}\}_{n=1}^{N_m}, \quad m = 1, \dots, M,$$

where $\mathbf{x}_{m,n} \sim q_m(\mathbf{x})$ i.e., a different proposal pdf for each set $\mathcal{S}_m$. In the most general case we consider that $N_i \neq N_j$, $\forall i \neq j$ with $i, j \in \{1, ..., M\}$. We can summarize the statistical information of each set using a pair of summary sample, $\widetilde{\mathbf{x}}_m$, and summary weight, W_m, $m = 1, \dots, M$, in such a way that the following estimator

$$\widetilde{I}_M = \frac{1}{\sum_{j=1}^{M} W_m} \sum_{m=1}^{M} W_m h(\widetilde{\mathbf{x}}_m), \tag{10}$$

is a consistent estimator of I. We denote the importance weight of the n-th sample in the m-th group as $w_{m,n} = w(\mathbf{x}_{m,n}) = \frac{\pi(\mathbf{x}_{m,n})}{q_m(\mathbf{x}_{m,n})}$, the m-th marginal likelihood estimator

$$\widehat{Z}_m = \frac{1}{N_m} \sum_{i=1}^{N_m} w_{m,n}, \tag{11}$$

and the normalized weights within a set as $\bar{w}_{m,n} = \frac{w_{m,n}}{N_m \widehat{Z}_m}$, for $n = 1, \dots, N$ and $m = 1, \dots, M$.

Definition 1. *A summary particle $\widetilde{\mathbf{x}}_m$ for the m-group is a resampled particle,*

$$\widetilde{\mathbf{x}}_m \sim \widehat{\pi}_m(\mathbf{x}|\mathbf{x}_{m,1:N_m}) = \sum_{n=1}^{N_m} \bar{w}_{m,n} \delta(\mathbf{x} - \mathbf{x}_{m,n}), \tag{12}$$

i.e., $\widetilde{\mathbf{x}}_m$ is selected within $\{\mathbf{x}_{m,1}, \dots, \mathbf{x}_{m,N_m}\}$ according to the probability mass function (pmf) defined by $\bar{w}_{m,n}$, $n = 1, \dots, N_m$.

It is possible to use the Liu's definition in order to assign a proper importance weight to a resampled particle [26], as stated in the following theorem.

Theorem 1. *Let us consider a resampled particle $\widetilde{\mathbf{x}}_m \sim \widehat{\pi}_m(\mathbf{x})$. A proper unnormalized weight following the Liu's definition in Eq. (7) for this resampled particle is $\widetilde{w}_m = \widetilde{w}(\widetilde{\mathbf{x}}_m) = \widehat{Z}_m$.*

The proof of this theorem is given in [27] and further discussions in [26].

Definition 2. *The summary weight for the m-th group of* *samples is $W_m = N_m \widetilde{w}_m = N_m \widehat{Z}_m$, defined in Eq.* (11).

Given the M summary pairs $\{\widetilde{\mathbf{x}}_m, W_m\}_{m=1}^M$ in a common computational node, we can obtain the following particle approximation of $\bar{\pi}(\mathbf{x})$, i.e.,

$$\widehat{\pi}(\mathbf{x}|\widetilde{\mathbf{x}}_{1:M}) = \frac{1}{\sum_{j=1}^M N_j \widehat{Z}_j} \sum_{m=1}^M N_m \widehat{Z}_m \delta(\mathbf{x} - \widetilde{\mathbf{x}}_m), \qquad (13)$$

involving M weighted samples in this case. For a given function $h(\mathbf{x})$, the corresponding specific GIS estimator in Eq. (10) is

$$\widetilde{I}_M \;=\; \frac{1}{\sum_{j=1}^M N_j \widehat{Z}_j} \sum_{m=1}^M N_m \widehat{Z}_m h(\widetilde{\mathbf{x}}_m). \qquad (14)$$

It is a consistent estimator of I. Indeed, the expression in Eq. (14) can be interpreted as a standard IS estimator (then consistent) since $\widetilde{w}(\widetilde{\mathbf{x}}_m) = \widehat{Z}_m$ is a proper weight of a resampled particle [26]. Moreover, we are giving more importance to the resampled particle belonging to a set with more cardinality. The joint use of the concepts of summary particle and summary weight is not strictly needed. In some application, both are required whereas in other applications only one of them is employed as we shown in the next section.

IV. Group Metropolis Sampling (GMS)

In this section, we show how GIS facilitates the design of novel efficient techniques. More specifically, we use the concept of summary weight associated to a set of samples in order to generalize the IMH algorithm in Table I. Unlike in the IMH scheme, GMS produces a sequence of sets of weighted samples. The Group Metropolis Sampling (GMS) is shown in Table II. Note that the GMS algorithm uses the idea of summary weight for comparing sets.

Table II
Group Metropolis Sampling (GSM)

```
Initialization: Choose an initial set $\mathcal{S}_0 = \{\mathbf{x}_n, \rho_{n,0}\}_{n=1}^N$
and $\widehat{Z}_0 = \frac{1}{N}\sum_{n=1}^N \rho_{n,0}$.
For $t = 1, \ldots, T$:
    1) Draw $N$ samples, $\mathbf{v}_1, \ldots, \mathbf{v}_N \sim q(\mathbf{x})$.
    2) Weight them $w_n = \frac{\pi(\mathbf{v}_n)}{q(\mathbf{v}_n)}$, $n = 1, \ldots, N$, define $\mathcal{S}' =$
       $\{\mathbf{v}_n, w_n\}_{n=1}^N$ and compute $\widehat{Z}' = \frac{1}{N}\sum_{n=1}^N w_n$.
    3) Set $\mathcal{S}_t = \{\mathbf{x}_{n,t} = \mathbf{v}_n, \rho_{n,t} = w_n\}_{n=1}^N$ (i.e., $\mathcal{S}_t = \mathcal{S}'$),
       and $\widehat{Z}_t = \widehat{Z}'$, with probability
           $\alpha(\mathcal{S}_{t-1}, \mathcal{S}') = \min\left[1, \frac{\widehat{Z}'}{\widehat{Z}_{t-1}}\right]$.   (15)
       Otherwise, set $\mathcal{S}_t = \mathcal{S}_{t-1}$ and $\widehat{Z}_t = \widehat{Z}_{t-1}$.
Return: $\{\mathcal{S}_t\}_{t=1}^T$.
```

Given the generated sets $\mathcal{S}_t = \{\mathbf{x}_{n,t}, \rho_{n,t}\}_{n=1}^N$, for $t = 1, \ldots, T$, GMS provides the global particle approximation

$$\begin{aligned}\widehat{\pi}(\mathbf{x}|\mathbf{x}_{1:N,1:T}) &= \frac{1}{T}\sum_{t=1}^T \sum_{n=1}^N \frac{\rho_{n,t}}{\sum_{i=1}^N \rho_{i,t}} \delta(\mathbf{x} - \mathbf{x}_{n,t}), \\ &= \frac{1}{T}\sum_{t=1}^T \sum_{n=1}^N \bar{\rho}_{n,t} \delta(\mathbf{x} - \mathbf{x}_{n,t}), \qquad (16)\end{aligned}$$

Relationship with IMH. The acceptance probability α in Eq. (15) is the extension of the acceptance probability of IMH in Eq. (9), considering the proper GIS weighting of a set of weighted samples (note that, in GMS, all the sets are the same number of samples).

Relationship with multiple try methods. GMS is strictly related to Multiple Try Metropolis (MTM) schemes [19], [21], [28], [29] and Particle Metropolis Hastings (PMH) techniques [17], [29]. The difference between GMS and the PMH and MTM methods is that GMS does not use resampling steps at each iteration for generating summary samples, indeed GMS uses the entire set. Another difference with PMH is that PMH generates sequentially the set of N candidates using a Sequential Importance Resampling (SIR) procedure (so that N candidates are correlated in PMH, in general) [29]. However, considering a sequential of a batch procedure for generating the N tries at each iteration, we can recover a MTM (or PMH) chain by the GMS output applying T resampling steps,

$$\widetilde{\mathbf{x}}_t = \begin{cases} \widetilde{\mathbf{v}}_t \sim \sum_{n=1}^N \bar{\rho}_{n,t} \delta(\mathbf{x} - \mathbf{x}_{n,t}), & \text{if} \quad \mathcal{S}_t \neq \mathcal{S}_{t-1}, \\ \widetilde{\mathbf{x}}_{t-1}, & \text{if} \quad \mathcal{S}_t = \mathcal{S}_{t-1}, \end{cases} \qquad (17)$$

Namely, $\{\widetilde{\mathbf{x}}_t\}_{t=1}^T$ is the chain obtained by one run of the MTM (or PMH) technique.

Ergodicity. As also discussed above, the acceptance probabilities and the dynamics of GMS exactly coincides with the PMH or MTM steps (with a sequential or batch particle generation, respectively), so that the ergodicity of the chain is ensured [17], [19], [21], [29]. Indeed, we can recover the MTM (or PMH) chain as shown in Eq. (17).

Recycling samples. The GMS algorithm can be seen as a method of recycling auxiliary weighted samples in PMH and MTM schemes. In [23], the authors show how recycling and including the rejected samples in a MH run into a unique consistent estimator. GMS can be considered an extension of this technique where, unlike in [23] N samples are generated at each iteration.

Iterated IS. GMS can be also interpreted as an iterative importance sampling scheme where an IS approximation of N samples is built at each iteration and compared with the previous IS approximation. This procedure is iterated T times and all the accepted IS estimators are finally combined for providing a unique global approximation of NT samples. Note that, the temporal combination of the IS estimators is obtained *dynamically* by the random repetitions due to the rejections in the MH test.

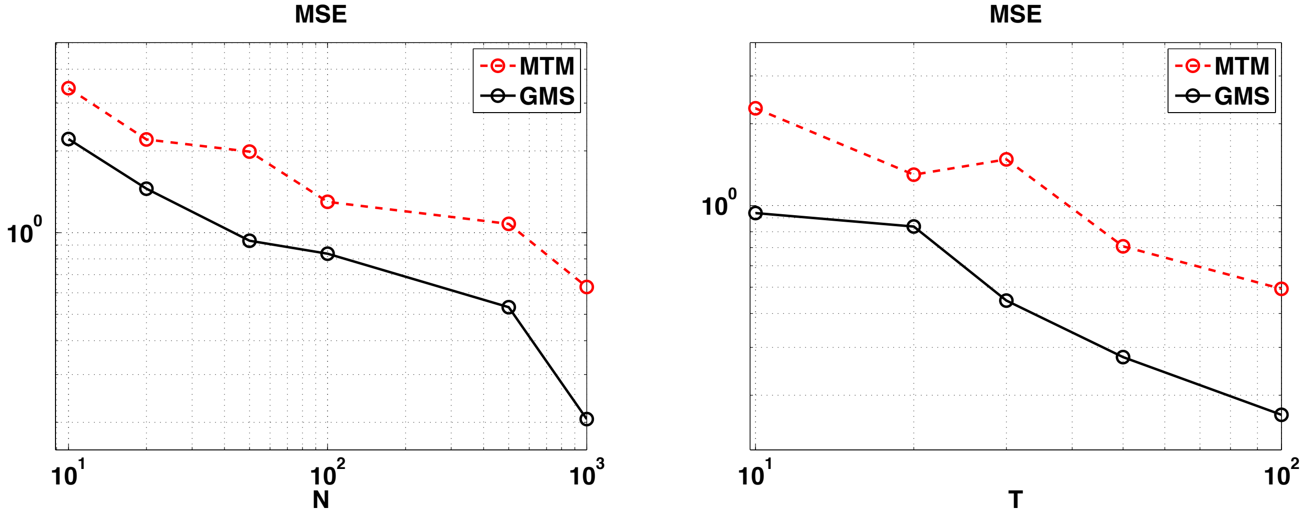

Figure 1. MSE (loglog-scale; averaged over 10^3 independent runs) obtained with the MTM and GMS algorithms (left) as function of N fixing $T = 20$ and (right) as function of T setting $N = 100$.

V. Numerical Simulations

We test the proposed GMS approach for the estimation of hyperparameters of a Gaussian process (GP) regression model [30], [31]. Let us assume observed data pairs $\{y_j, \mathbf{z}_j\}_{j=1}^P$, with $y_j \in \mathbb{R}$ and $\mathbf{z}_j \in \mathbb{R}^L$. We also denote the corresponding $P \times 1$ output vector as $\mathbf{y} = [y_1, \dots, y_P]^\top$ and the $L \times P$ input matrix as $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_P]$. We address the regression problem of inferring the unknown function f which links the variable y and $\mathbf{z}$. Thus, the assumed model is $y = f(\mathbf{z}) + e$, where $e \sim N(e; 0, \sigma^2)$, and that $f(\mathbf{z})$ is a realization of a GP [31]. Hence $f(\mathbf{z}) \sim \mathcal{GP}(\mu(\mathbf{z}), \kappa(\mathbf{z}, \mathbf{r}))$ where $\mu(\mathbf{z}) = 0$, $\mathbf{z}, \mathbf{r} \in \mathbb{R}^L$, and we consider the kernel function

$$\kappa(\mathbf{z}, \mathbf{r}) = \exp\left(-\sum_{\ell=1}^{L} \frac{(z_\ell - r_\ell)^2}{2\delta^2}\right), \tag{18}$$

Given these assumptions, the vector $\mathbf{f} = [f(\mathbf{z}_1), \dots, f(\mathbf{z}_P)]^\top$ is distributed as $p(\mathbf{f}|\mathbf{Z}, \delta, \kappa) = \mathcal{N}(\mathbf{f}; \mathbf{0}, \mathbf{K})$, where $\mathbf{0}$ is a $P \times 1$ null vector, and $\mathbf{K}_{ij} := \kappa(\mathbf{z}_i, \mathbf{z}_j)$, for all $i, j = 1, \dots, P$, is a $P \times P$ matrix. Therefore, the vector containing all the hyper-parameters of the model is $\boldsymbol{\theta} = [\delta, \sigma]$, i.e., all the parameters of the kernel function in Eq. (18) and standard deviation σ of the observation noise. In this experiment, we focus on the marginal posterior density of the hyperparameters [31], $\bar{p}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa) \propto p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa) = p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{Z}, \kappa)p(\boldsymbol{\theta})$, which can be evaluated analytically, but we cannot compute integrals involving it. Considering a uniform prior within $[0, 20]^2$ and since $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{Z}, \kappa) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \mathbf{K} + \sigma^2\mathbf{I})$, we have

$$\begin{aligned}\log\left[p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa)\right] = &-\frac{1}{2}\mathbf{y}^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y} \\ &-\frac{1}{2}\log\left[\det\left(\mathbf{K} + \sigma^2\mathbf{I}\right)\right],\end{aligned}$$

where clearly $\mathbf{K}$ depends on δ [31]. The moments of this marginal posterior cannot be computed analytically. Then, in order to compute the Minimum Mean Square Error (MMSE) estimator $\widehat{\boldsymbol{\theta}} = [\widehat{\delta}, \widehat{\sigma}]$, i.e., the expected value $\mathbb{E}[\boldsymbol{\Theta}]$ with $\boldsymbol{\Theta} \sim p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{Z}, \kappa)$, we approximate $\mathbb{E}[\boldsymbol{\Theta}]$ via Monte Carlo quadrature. More specifically, we apply a the novel GMS technique and compare with an MTM sampler.

We generated $P = 200$ pairs of data, $\{y_j, \mathbf{z}_j\}_{j=1}^P$, according to the GP model setting $\delta^* = 3$, $\sigma^* = 10$. $L = 1$, and drawing $z_j \sim \mathcal{U}([0, 10])$. Keeping fixed the generated data for each scenario, we then computed the ground-truth $\widehat{\boldsymbol{\theta}} \approx [\widehat{\delta} \approx 3.5200, \widehat{\sigma} \approx 9.2811]$ using an exhaustive and costly grid approximation, in order to compare the different techniques. For both GMS and MTM schemes, we consider the same adaptive Gaussian proposal pdf $q_t(\mathbf{x}|\boldsymbol{\mu}_t, \lambda^2\mathbf{I}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_t, \lambda^2\mathbf{I})$, with $\lambda = 5$ and $\boldsymbol{\mu}_t$ is adapted considering the arithmetic mean of the outputs after a training period, $t \geq 0.2T$, in the same fashion of [32], [33] ($\boldsymbol{\mu}_0 = [1, 1]^\top$). First, we test both techniques fixing $T = 20$ and varying the number of tries N. Then, we set $N = 100$ and vary the number of iterations T. Figure 1 (log-log plot) shows the Mean Square Error (MSE) in the approximation of $\widehat{\boldsymbol{\theta}}$ averaged over 10^3 independent runs. Observe that always GMS outperforms the corresponding MTM scheme.

VI. Conclusions

In this work, we introduce the Group Importance Sampling (GIS) theory which facilitates the design of novel Monte Carlo algorithms. For instance, we present the Group Metropolis Sampling (GMS) method that outperforms the corresponding benchmark Monte Carlo techniques without any extra computational cost, as we have shown in an hyperparameter estimation problem for GP regression models.

Acknowledgements

This work has been supported by the European Research Council (ERC) through the ERC Consolidator Grant SEDAL ERC-2014-CoG 647423.

References

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning*, vol. 50, no. 1, pp. 5–43, 2003.

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

[3] M. F. Bugallo, L. Martino, and J. Corander, "Adaptive importance sampling in signal processing," *Digital Signal Processing*, no. 47, pp. 36–49, 2015.

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

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

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

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

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

[9] M. Bolić, P. M. Djurić, and S. Hong, "Resampling algorithms and architectures for distributed particle filters," *IEEE Transactions Signal Processing*, vol. 53, no. 7, pp. 2442–2450, July 2005.

[10] J. Miguez and M. A. Vazquez, "A proof of uniform convergence over time for a distributed particle filter," *Signal Processing*, vol. 122, pp. 152–163, 2016.

[11] C. Verg, C. Dubarry, P. D. Moral, and E. Moulines, "On parallel implementation of sequential Monte Carlo methods: the island particle model," *Statistics and Computing*, vol. 25, no. 2, pp. 243–260, 2015.

[12] C. Verg, P. D. Moral, E. Moulines, and J. Olsson, "Convergence properties of weighted particle islands with application to the double bootstrap algorithm," *arXiv:1410.4231*, pp. 1–39, 2014.

[13] N. Whiteley, A. Lee, and K. Heine, "On the role of interaction in sequential Monte Carlo algorithms," *Bernoulli*, vol. 22, no. 1, pp. 494–529, 2016.

[14] L. Martino, J. Read, V. Elvira, and F. Louzada, "Cooperative parallel particle filters for on-line model selection and applications to urban mobility," *Digital Signal Processing*, vol. 60, no. 3, pp. 172–185, 2017.

[15] C. A. Naesseth, F. Lindsten, and T. B. Schon, "Nested Sequential Monte Carlo methods," *Proceedings of theInternational Conference on Machine Learning*, vol. 37, pp. 1–10, 2015.

[16] R. B. Stern, "A statistical contribution to historical linguistics," *Phd Thesis*, 2015.

[17] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *J. R. Statist. Soc. B*, vol. 72, no. 3, pp. 269–342, 2010.

[18] M. Bédard, R. Douc, and E. Mouline, "Scaling analysis of multiple-try MCMC methods," *Stochastic Processes and their Applications*, vol. 122, pp. 758–786, 2012.

[19] R. Casarin, R. V. Craiu, and F. Leisen, "Interacting multiple try algorithms with different proposal distributions," *Statistics and Computing*, vol. 23, no. 2, pp. 185–200, 2013.

[20] R. V. Craiu and C. Lemieux, "Acceleration of the Multiple Try Metropolis algorithm using antithetic and stratified sampling," *Statistics and Computing*, vol. 17, no. 2, pp. 109–120, 2007.

[21] L. Martino and J. Read, "On the flexibility of the design of multiple try Metropolis schemes," *Computational Statistics*, vol. 28, no. 6, pp. 2797–2823, 2013.

[22] L. Martino and F. Louzada, "Issues in the Multiple Try Metropolis mixing," *Computational Statistics*, vol. 32, no. 1, pp. 239–252, 2017.

[23] G. Casella and C. P. Robert, "Rao-blackwellisation of sampling schemes," *Biometrika*, vol. 83, no. 1, pp. 81–94, 1996.

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

[25] L. Martino and V. Elvira, "Metropolis Sampling," *Wiley StatsRef: Statistics Reference Online*, pp. 1–23, 2017.

[26] L. Martino, V. Elvira, and F. Louzada, "Weighting a resampled particle in Sequential Monte Carlo," *IEEE Statistical Signal Processing Workshop, (SSP)*, pp. 1–5, 2016.

[27] ——, "Weighting a resampled particle in Sequential Monte Carlo (extended preprint)," *Techinical report, viXra:1602.0333*, 2016.

[28] L. Martino, V. P. D. Olmo, and J. Read, "A multi-point Metropolis scheme with generic weight functions," *Statistics & Probability Letters*, vol. 82, no. 7, pp. 1445–1453, 2012.

[29] L. Martino, F. Leisen, and J. Corander, "On multiple try schemes and the Particle Metropolis-Hastings algorithm," *Technical report, viXra:1409.0051*, 2014.

[30] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[31] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[32] D. Luengo and L. Martino, "Fully adaptive Gaussian mixture Metropolis-Hastings algorithm," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[33] H. Haario, E. Saksman, and J. Tamminen, "An adaptive Metropolis algorithm," *Bernoulli*, vol. 7, no. 2, pp. 223–242, April 2001.

ISBN 978-0-9928626-7-1 © EURASIP 2017