Digital Signal Processing 74 (2018) 1–13

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

The Recycling Gibbs sampler for efficient learning

Luca Martino [a,b,*], Víctor Elvira [c], Gustau Camps-Valls [a]

[a] Image Processing Laboratory, Universitat de València, 46980 Paterna, València, Spain
[b] Signal Processing Group, Universidad Carlos III de Madrid, Spain
[c] IMT Lille Douai CRISTAL (UMR 9189), Villeneuve d'Ascq, France

ARTICLE INFO

Article history:
Available online 5 December 2017

Keywords:
Bayesian inference
Markov Chain Monte Carlo (MCMC)
Gibbs sampling
Gaussian Processes (GP)
Automatic relevance determination (ARD)

ABSTRACT

Monte Carlo methods are essential tools for Bayesian inference. Gibbs sampling is a well-known Markov chain Monte Carlo (MCMC) algorithm, extensively used in signal processing, machine learning, and statistics, employed to draw samples from complicated high-dimensional posterior distributions. The key point for the successful application of the Gibbs sampler is the ability to draw efficiently samples from the full-conditional probability density functions. Since in the general case this is not possible, in order to speed up the convergence of the chain, it is required to generate auxiliary samples whose information is eventually disregarded. In this work, we show that these auxiliary samples can be recycled within the Gibbs estimators, improving their efficiency with no extra cost. This novel scheme arises naturally after pointing out the relationship between the standard Gibbs sampler and the chain rule used for sampling purposes. Numerical simulations involving simple and real inference problems confirm the excellent performance of the proposed scheme in terms of accuracy and computational efficiency. In particular we give empirical evidence of performance in a toy example, inference of Gaussian processes hyperparameters, and learning dependence graphs through regression.

1. Introduction

'Reduce, Reuse, Recycle'
The Greenpeace motto

Monte Carlo algorithms have become very popular over the last decades [1,2]. Many practical problems in statistical signal processing, machine learning and statistics, demand fast and accurate procedures for drawing samples from probability distributions that exhibit arbitrary, non-standard forms [3,4], [5, Chapter 11]. One of the most popular Monte Carlo methods are the families of Markov chain Monte Carlo (MCMC) algorithms [3,2] and particle filters [6,7]. The 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 [1,8].

The Gibbs sampling technique is a well-known MCMC algorithm, extensively used in the literature in order to generate samples from multivariate target densities, drawing each component of the samples from the full-conditional densities [9–14].[1] In order to draw samples from a multivariate target distribution, the key point for the successful application of the standard Gibbs sampler is the ability to draw efficiently from the univariate conditional pdfs [1,2]. The best scenario for Gibbs sampling occurs when specific direct samplers are available for each full-conditional, e.g. inversion method or, more generally, some transformation of a random variable [16,2]. Otherwise, other Monte Carlo techniques, such as rejection sampling (RS) and different flavors of the Metropolis–Hastings (MH) algorithms, are typically used *within* the Gibbs sampler to draw from the complicated full-conditionals. The performance of the resulting Gibbs sampler depends on the employed *internal* technique, as pointed out for instance in [17–20].

In this context, some authors have suggested to use more steps of the MH method within the Gibbs sampler [21–23]. Moreover, other different algorithms have been proposed as alternatives to the MH technique [17,10,24]. For instance, several automatic and self-tuning samplers have been designed to be used primarily

* Corresponding author at: Image Processing Laboratory, Universitat de València, 46980 Paterna, València, Spain.
E-mail address: lmartino@uc3m.es (L. Martino).

[1] In general these full-conditionals are univariate. Nevertheless, block-wise Gibbs sampling approaches where several random variables are updated simultaneously, have been proposed to speed up the convergence of the Gibbs sampler [15]. However, unless direct sampling from the multi-variate full-conditionals is feasible, these approaches still result in an increased difficulty of drawing samples and a higher computational cost per iteration. Furthermore, the performance of the overall algorithm can decrease if the blocks are not properly chosen, especially when direct sampling from the multivariate full-conditionals is unfeasible [1,8,2]. The novel recycling scheme can also be used in the block approach.

https://doi.org/10.1016/j.dsp.2017.11.012
1051-2004/

within-Gibbs: the adaptive rejection sampling (ARS) [25,26], the griddy Gibbs sampler [27], the FUSS sampler [20], the Adaptive Rejection Metropolis Sampling (ARMS) method [18,28,29,14], and the Independent Doubly Adaptive Rejection Metropolis Sampling (IA^{2}RMS) technique [19], just to name a few.

Most of the previous solutions require performing several MCMC steps for each full-conditional in order to improve the performance, although only one of them is considered to produce the resulting Markov chain because the rest of samples play the mere role of auxiliary variables. Strikingly, they require an increase in the computational cost that is not completely paid off: several samples are drawn from the full-conditionals, but only a subset of these generated samples is employed in the final estimators. In this work, we show that the rest of generated samples can be directly incorporated within the corresponding Gibbs estimator. We call this approach the *Recycling Gibbs (RG) sampler* since *all* the samples drawn from each full-conditional can be used also to provide a better estimation, instead of discarding them.

The consistency of the proposed RG estimators is guaranteed, as will be noted after considering the connection between the Gibbs scheme and the chain rule for sampling purposes [16,2]. In particular, we show that the standard Gibbs approach is equivalent (after the burn-in period) to the standard chain rule, whereas RG is equivalent to an alternative version of the chain rule presented in this work as well. RG fits particularly well combined with adaptive MCMC schemes where different internal steps are performed also for adapting the proposal density, see e.g. [18,19,29,14]. The novel RG scheme allows us to obtain better performance without adding any extra computational cost. This will be shown through intensive numerical simulations. First, we test RG in a simple toy example with a bimodal bivariate target. We also include experiments for hyper-parameter estimation in Gaussian Processes (GPs) regression problems with the so-called *automatic relevance determination* (ARD) kernel function [5]. Finally, we apply the novel scheme in real-life geoscience problems of dependence estimation among bio-geo-physical variables from satellite sensory data. The MATLAB code of the numerical examples is provided at http://isp.uv.es/code/RG.zip.

The remainder of the paper is organized as follows. Section 2 fixes notation and recalls the problem statement of Bayesian inference. The standard Gibbs sampler and the chain rule for sampling purposes are summarized in Section 3, highlighting their connections. In the same section, we then introduce an alternative chain rule approach, which is useful for describing the novel scheme. The RG technique proposed here is formalized in Section 4. Sections 5 provides empirical evidence of the benefits of the proposed scheme, considering different multivariate posterior distributions. Finally, Section 6 concludes and outlines further work.

2. Bayesian inference

Machine learning, statistics, and signal processing often face the problem of inference through density sampling of potentially complex multivariate distributions. In particular, Bayesian inference is concerned about doing inference about a variable of interest exploiting the Bayes' theorem to update the probability estimates according to the available information. Specifically, in many applications, the goal is to infer a variable of interest, $\mathbf{x} = [x_1, \ldots, x_D] \in \mathbb{R}^D$, given a set of observations or measurements, $\mathbf{y} \in \mathbb{R}^P$. In Bayesian inference all the statistical information is summarized by means of the posterior pdf, i.e.,

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

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, $Z(\mathbf{y})$ is unknown and difficult to estimate in general, so we assume to be able to evaluate the unnormalized target function,

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

The analytical study of the posterior density $\bar{\pi}(\mathbf{x}) \propto \pi(\mathbf{x})$ is often unfeasible and integrals involving $\bar{\pi}(\mathbf{x})$ are typically intractable. For instance, one might be interested in the estimation of

$$I = \int_{\mathbb{R}^D} f(\mathbf{x})\bar{\pi}(\mathbf{x})d\mathbf{x}, \tag{3}$$

where $f(\mathbf{x})$ is a squared integrable function (with respect to $\bar{\pi}$). In order to compute the integral I numerical approximations are typically required. Our goal here is to approximate this integral by using Monte Carlo (MC) quadrature [1,2]. Namely, considering T independent samples from the posterior target pdf, i.e., $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)} \sim \bar{\pi}(\mathbf{x})$, we can write

$$\widehat{I}_T = \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}^{(t)}) \xrightarrow{p} I. \tag{4}$$

This means that for the weak law of large numbers, $\widehat{I}_T$ converges in probability to I: that is, for any positive number $\epsilon > 0$, we have $\lim_{T\to\infty} \Pr(|\widehat{I}_T - I| > \epsilon) = 0$. In general, a direct method for drawing independent samples from $\bar{\pi}(\mathbf{x})$ is not available, and alternative approaches, e.g., MCMC algorithms, are needed. An MCMC method generates an ergodic Markov chain with invariant density $\bar{\pi}(\mathbf{x})$ (a.k.a., stationary pdf). Even though, the generated samples $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}\}$ are then correlated in this case, $\widehat{I}_T$ is still a consistent estimator.

Within the MCMC framework, we can consider a block approach working directly into the D-dimensional space, e.g., using a Metropolis–Hastings (MH) algorithm [2], or a component-wise approach [30–32] working iteratively in different uni-dimensional slices of the entire space, e.g., using a Gibbs sampler [1,8].[2] In many applications, and for different reasons, the component-wise approach is the preferred choice. For instance, this is the case when the full-conditional distributions are directly provided or when the probability of accepting a new state with a complete block approach becomes negligible as the dimension of the problem D increases. In the following section, we outline the standard Gibbs approach, and remark its connection with the chain rule method. The main notation and acronyms of the work are summarized in Table 1.

3. Gibbs sampling and the chain rule method

This section reviews the fundamentals about the standard Gibbs sampler, reviews the recent literature on Gibbs sampling when complicated full-conditional pdfs are involved, and points out the connection between GS and the chain rule. A variant of the chain rule is also described, which is related to the novel scheme introduced in the next section.

3.1. The Standard Gibbs (SG) sampler

The Gibbs sampler is perhaps the most widely used algorithm for inference in statistics and machine learning [9,10,12,2]. Let us define $\mathbf{x}_{\neg d} := [x_1, \ldots, x_{d-1}, x_{d+1}, \ldots, x_D]$ and introduce the following equivalent notations

[2] There also exist intermediate strategies where the same subset of variables are jointly updated, which is often called the Blocked Gibbs Sampler.