Statistics and Probability Letters 82 (2012) 473–477

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A theory for the multiset sampler

Yuguo Chen *

Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street, Champaign, IL 61820, USA

ARTICLE INFO

Article history:
Received 2 August 2011
Received in revised form 24 September 2011
Accepted 26 September 2011
Available online 25 November 2011

Keywords:
Data augmentation
Evolutionary forest algorithm
Forward operator
Gibbs sampler
Markov chain Monte Carlo

ABSTRACT

The multiset sampler (MSS) can be viewed as a new data augmentation scheme and it has been applied successfully to a wide range of statistical inference problems. The key idea of the MSS is to augment the system with a multiset of the missing components, and construct an appropriate joint distribution of the parameters of interest and the missing components to facilitate the inference based on Markov chain Monte Carlo. The standard data augmentation strategy corresponds to the MSS with multiset size one. This paper provides a theoretical comparison of the MSS with different multiset sizes. We show that the MSS converges to the target distribution faster as the multiset size increases. This explains the improvement in convergence rate for the MSS with large multiset sizes over the standard data augmentation scheme.

1. Introduction

In statistical inference, if the target distribution $p(\theta)$ is complicated, we often need to use Monte Carlo methods to draw samples from $p(\theta)$ to make inference about certain features of the distribution, such as the mean of $p(\theta)$. This occurs very often in Bayesian inference where the posterior distribution of the unknown parameters θ is of interest. If drawing samples directly from $p(\theta)$ is not straightforward, sometimes we can consider an augmented system (θ, X), such that the marginal distribution of $p(\theta, X)$ is the same as $p(\theta)$ and drawing samples from $p(\theta, x)$ is easier. In particular, if the augmented system facilitates iterative conditional sampling from $p(\theta|x)$ and $p(x|\theta)$, the Gibbs sampler can be used to sample from the joint distribution $p(\theta, x)$, and we only need to keep the samples of θ to make inference about the target distribution $p(\theta)$. This is an application of the data augmentation scheme proposed by Tanner and Wong (1987) and further studied by Meng and van Dyk (1999) and van Dyk and Meng (2001).

One of the areas that data augmentation schemes have been used extensively is population genetics. In this context θ usually denotes the population parameters we want to estimate, such as the mutation rate, the population size of a species, etc. A Bayesian approach seeks to characterize the posterior distribution of θ given the observed DNA sequence data. However, this posterior distribution typically is not available in analytical forms, because it involves summing over a huge number of possible underlying genealogical trees. One way to avoid the summation is to consider the joint posterior distribution $p(\theta, t)$ of the parameters and the tree, whose marginal distribution $p(\theta)$ is the target distribution. This is exactly the data augmentation idea. If the conditional distributions $p(\theta|t)$ and $p(t|\theta)$ are difficult to sample from, one can use a Metropolis-within-Gibbs algorithm ((Robert and Casella, 1999), pp. 322–323) to sample from $p(\theta, t)$. That is, one samples iteratively from the distributions $p(\theta|t)$ and $p(t|\theta)$, but a Metropolis–Hastings step is used in each iteration to generate θ and t.

Leman et al. (2007) notice that when they apply the data augmentation scheme to study the time since divergence between two closely related species, the Metropolis-within-Gibbs algorithm for sampling (θ, t) converges very slowly,

* Tel.: +1 217 244 3932; fax: +1 217 244 7190.
E-mail address: yuguo@illinois.edu.

0167-7152/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2011.09.025

especially in the tree space. The acceptance rate for a new tree is extremely low, and the Markov chain can easily get stuck in a local mode. Exploration of the tree space seems to be a common challenging problem for all genealogically-based data augmentation approaches. Such convergence problems may also occur in applying the data augmentation idea to other statistical inference problems besides population genetics.

To overcome this difficulty, Leman et al. (2007) develop a new data augmentation scheme, called the evolutionary forest (EF) algorithm, which leads to an efficient Markov chain Monte Carlo (MCMC) scheme. The EF algorithm is also based on the data augmentation idea, but it augments the system with multiple trees (i.e., a *forest*) rather than a single tree. Therefore the updates of the population parameters are made on the basis of their likelihood over multiple trees. The EF algorithm greatly increases the mixing of the Markov chain. Leman et al. (2007) show that their new EF algorithm converges to the target distribution much faster than the standard MCMC algorithms on simulated and real data.

Leman et al.'s (2007) original EF algorithm is proposed in the context of population genetics, but the evolutionary forest idea is a quite general way to design efficient algorithms. In fact, Leman et al. (2009) develop a general-purpose algorithm, the multiset sampler (MSS), based on the EF algorithm, and apply the MSS to other statistical inference problems besides population genetics.

In this paper, we provide a theoretical justification for the MSS and compare the convergence rate of the MSS with different multiset sizes. We prove that the MSS with a larger multiset size converges to the target distribution faster. This explains why the MSS performs better than the MCMC algorithm based on the standard data augmentation scheme which corresponds to the MSS with multiset size one. In Section 2 we briefly review the MSS. The main theorems on the comparison of the convergence rate are given in Section 3. Section 4 concludes the paper with a discussion.

2. The multiset sampler

Here we review the basic framework of the multiset sampler in Leman et al. (2009). Suppose our goal is to draw samples from the target distribution $p(\theta)$. Assume that an augmented system (θ, X) is available to facilitate the sampling, and the marginal distribution of θ under $p(\theta, x)$ is the target distribution $p(\theta)$. Sometimes X is referred to as the nuisance parameters (Leman et al., 2009). For simplicity, assume X takes values in a finite set Ω with $|\Omega| = n$. The standard data augmentation scheme would sample from $p(\theta, x)$ and use the draws of θ to make inference about $p(\theta)$.

In the MSS, a new random variable Y_K is introduced which takes values in the set

$$\Omega_K = \left\{ \biguplus_{i=1}^{K} x_i : \ x_i \in \Omega \right\}, \tag{1}$$

where K a fixed positive integer. Here $\biguplus_{i=1}^{K} x_i$ is a multiset, which is an unordered collection that allows repeated elements. For example, $x_1 \uplus x_2 \uplus x_1 = \{x_1, x_1, x_2\}$. Each $y \in \Omega_K$ is called a multiset of size K. Obviously there are totally $\binom{n+K-1}{K}$ multisets in Ω_K. Define the joint distribution of (θ, Y_K) as

$$q_K(\theta, y) = C_K \sum_{x \in y} p(\theta, x), \tag{2}$$

where

$$C_K = \frac{n}{\binom{n+K-1}{K} K}$$

is the normalizing constant. On the right hand side of (2), we sum over all components in y. For example, if $K = 3$ and $y = \{x_1, x_1, x_2\}$, then $\sum_{x \in y} p(\theta, x) = 2p(\theta, x_1) + p(\theta, x_2)$. It is easy to verify that the marginal distribution of θ under $q_K(\theta, y)$ is still the same as the target distribution $p(\theta)$. Therefore we could sample from $q_K(\theta, y)$ and keep the samples of θ as draws from $p(\theta)$. This is the multiset sampler.

Note that when $K = 1$, the distribution $q_1(\theta, y)$ is the same as $p(\theta, x)$, which means the standard data augmentation scheme can be viewed as a special case of the multiset sampler with multiset size $K = 1$. Leman et al.'s (2007) numerical results show that the MSS with $K > 1$ converges to the target distribution $p(\theta)$ faster than the MSS with $K = 1$. In the next section, we provide a theoretical explanation for the performance of the MSS with different multiset sizes.

3. Convergence of multiset samplers

Assume that the Gibbs sampler can be used to draw samples from $p(\theta, x)$ and $q_K(\theta, y)$ which are defined on the augmented systems (θ, X) and (θ, Y_K) in Section 2. In other words, we assume the corresponding conditional distributions are easy to sample from. Then a systematic scan Gibbs sampler for sampling from $p(\theta, x)$ constructs a Markov chain $\{(\theta^{(i)}, x^{(i)}), i = 0, 1, \ldots\}$ with transition function

$$K[(\theta^{(i)}, x^{(i)}), (\theta^{(i+1)}, x^{(i+1)})] = p(\theta^{(i+1)}|x^{(i)}) p(x^{(i+1)}|\theta^{(i+1)}). \tag{3}$$